UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

ESTIMACIÓN DE POBLACIONES SUMERGIDAS EN REDES SOCIALES MEDIANTE ENCUESTAS INDIRECTAS

Por: Ignacio Ortega Aviles

Tesis presentada a la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Concepción para optar al título de Ingeniera Civil Matemática

Agosto 2023
Concepción, Chile

Profesor Guía: Dr. Christopher Thraves Caro

A mis padres

AGRADECIMIENTOS

Quiero expresar mi profundo agradecimiento a mis padres, pues yo no estaría en la universidad sin la ayuda de ellos, si que les agradezco de todo corazón por todo lo que han hecho por mí. Por otro lado, también agradezco a mis hermanos y a mi primo que me ayudan a distraerme y a crear buenos momentos, y a mi hermana con la cuál siempre terminando conversando cosas entretenidas.

Además de mi familia, quisiera agradecer a los profesores de la universidad que me han hecho clases. Quiero agradecer a mi profesor guía Christopher Thraves, quien fue el que me guió en este trabajo y me dio su apoyo. Por último y no menos importante, quiero dar gracias a la comisión evaluadora por aceptar evaluarme en este trabajo.

Muchas gracias a todos.

Índice general

Índice de figuras

Capítulo 1

Introducción

Un problema grave que comenzó en los años 80 es el virus del SIDA, ya que han fallecido 32 millones de personas en el mundo producto de enfermedades relacionadas con el SIDA. En Chile, según la ONUSIDA 70 mil personas viven con VIH y 5 mil personas se infectan al año [5]. Además considerando que el SIDA aún no tiene cura entonces se considera una pandemia que va aumentando la cantidad de personas infectadas año tras año. Para mantener el virus bajo control es necesario realizar estimaciones de la cantidad total de personas infectadas, aunque realizar esta operación es difícil. Lamentablemente, el ser portador del virus del SIDA esta estigmatizado, lo que implica que una persona que tenga VIH positivo es difícil que admita que tiene la enfermedad, lo que complica aún más las estimaciones. Por otro lado, las personas con más riesgos de contraer el virus del SIDA son aquellas que consumen heroína, las que tienen múltiples parejas sexuales y las que mantienen relaciones sexuales sin la protección adecuada. Debido a esto, también nos interesa estimar la cantidad de personas que realizan estas actividades para poder frenar la propagación del VIH. De esta forma, una vez realizadas las estimaciones correspondientes se podrían realizar diversas estrategias para reducir los contagios de VIH como charlas que incentiven el uso de condón, prevenir el compartir agujas entre los consumidores de heroína e incentivar la reducción de la cantidad de personas con la que tienen intimidad. Además de los mecanismos de prevención, se podría administrar de mejor manera los tratamientos y personal para esta enfermedad.

Esta memoria tiene relación con la estimación de grupos dentro de una población, así para definir el problema de esta memoria, primero debemos realizar una definición.

Definición 1 (Encuesta)**.** *Dada una población y un grupo desconocido de la población*

entonces la encuesta es un conjunto de las siguientes preguntas ¿ Cuántas personas conoce?¿ Cuántas personas conoce en el grupo desconocido ? y ¿ Usted pertenece o no al grupo desconocido? Estas preguntas son realizadas sólo a una porción de la población.

De esta forma se define el siguiente problema.

Problema 1. *Dada una población cualquiera T de tamaño conocido, en donde los individuos de esta pueden o no estar relacionados, las relaciones entre individuos son simétricas y no se conocen, sin embargo se pueden preguntar por ellas a través de una encuesta hecha a una porción de la población. Además, existe un grupo desconocido de individuos de la población, entonces el problema consiste en estimar la cantidad de individuos de este grupo desconocido a través de una encuesta.*

El Problema de la Definición 1 es motivado por diversas razones. Una de ellas es el interés en conocer cuantas personas en un lugar determinado comparten cierta característica, como por ejemplo VIH positivo, homosexuales, personas en situación de calle, drogadictos y fallecidos a causas de desastres naturales.

Debido a esta necesidad, Bernard, H Russell and Johnsen, Eugene C and Killworth, Peter D en1987 publicaron [2], en donde comenzaron el desarrollo de un método para realizar aquellas estimaciones, estos métodos aún en nuestros tiempos siguen mejorándose. La idea base de estos métodos de estimación es conseguir información de una pequeña parte de la población, y a partir de esta información y en base a suposiciones se realizan las estimaciones correspondientes a nivel global, estos métodos son llamados **métodos escalables**. En la fase de recopilación de datos se realizan varias preguntas al encuestado sobre el mismo, la cantidad de conocidos que tiene y sus características. Cabe aclarar que el enfoque que veremos en esta tesis, son estimaciones a partir de métodos escalables que utilicen una encuesta para recopilar información de la población general, y por lo tanto no se verán otros métodos, como por ejemplo el método Capture-recapture, que se centra en capturar información directamente del grupo que se desea estimar, y los métodos en donde se usa estimación bayesiana, que consiste en conocer en una probabilidad a priori, para luego determinar distribución a posteriori, que no es el tamaño del grupo de interés sino que es la distribución de la cantidad de personas del grupo de interés que es conocido por los integrantes de la población, Laga Ian, Bao Le y Niu Xiaoyue muestra más modelos de este tipo en el artículo [9].

En el artículo [9] se menciona que los métodos escalables empezaron su desarrollo en 1987 con el modelo propuesto por Bernard, H Russell et al. en el artículo [2], en donde

propusieron un método para acotar por arriba y por abajo el tamaño del grupo de interés. Luego, Killworth, Peter D; Johnsen, Eugene C; McCarty, Christopher; Shelley, Gene Ann; Bernard, H Russell crea el modelo PIMLE y lo presenta en el artículo [7], además Killworth, Peter D; McCarty, Christopher; Bernard, H Russell; Shelley, Gene Ann; Johnsen, Eugene C en el artículo [8] presentan el modelo MLE. Los modelos PINMLE y MLE sirven para estimar un grupo en la población. A grandes rasgos, esto se realiza suponiendo que la proporción entre la cantidad media de conocidos en el grupo de interés y la cantidad media de conocidos por los integrantes de la población es igual a la proporción entre el tamaño del grupo de interés y el tamaño de la población. Salganik, Matthew J; Heckathorn, Douglas D presentaron en [16] el modelo RDS I, además Feehan, Dennis M; Salganik, Matthew J en [3] desarrollan el modelo GNSUM. Los modelos RDS I y GNSUM modelan la población con un grafo y se estiman propiedades locales de él a través del estimador Hansen-Hurwitz y el estimador Horvitz–Thompson respectivamente, para luego estimar propiedades globales, es decir el tamaño del grupo de interés. Por último,Volz, Erik; Heckathorn, Douglas D en [18] propone el modelo RDS II con el propósito de estimar el tamaño del grupo de interés utilizando el estimador Hansen-Hurwitz.

Debido a la amplia cantidad de formas para estimar el tamaño de un grupo de interés dentro de la población utilizando distintos conceptos, vale la pena determinar que métodos son mejores que otros en una determinada situación. Para esto, podemos utilizar medidas estadísticas, de esta forma el objetivo de esta tesis está relacionado con el rendimiento de los estimadores de grupo utilizando varios tipos de encuestas. Pero, debido a la complejidad que se puede dar en cada situación y los distintos problemas que se pueden dar, la evaluación de los métodos se realizará bajo el supuesto que la información obtenida por la encuesta es correcta, es decir **los encuestados responden verazmente la encuesta**. En base al contexto antes mencionado, se realizarán simulaciones que pondrán a prueba los métodos estimativos en variados contextos y se compararán. Para comparar los métodos se estimarán medidas estadísticas y se compararán sus gráficas correspondiente, lo que permitirá deducir información útil para realizar las conclusiones correspondientes.

Para lograr el modelado del problema se utilizará un grafo que representará a la población, en la cuál los nodos son los individuos de la población y las aristas son las conexiones entre los individuos. Se tendrá un grupo desconocido de nodos H, el cuál se deberá estimar su tamaño a través de la recopilación de información de una porción de

la población y la aplicación de los métodos estimadores. Por lo tanto esta tesis consta de varios capítulos, el Capítulo 2 involucra el marco teórico que explicará las propiedades básicas de grafos, conceptos de muestreo, y la definición formal y los objetivos de esta tesis. En el Capítulo 3 de definirá algunos tipos de muestreos aleatorios. Para cada tipo de muestreo aleatorio se definirán estimadores de propiedades de grafos. En el Capítulo 4 se presentan los estimadores de tamaño de grupos utilizando propiedades de grafos y los muestreos aleatorios del Capítulo 3. Para el Capítulo 5 se definirán los experimentos que pondrán a prueba los estimadores de grupo. Por último, en el Capítulo 6 se presentarán las conclusiones de esta tesis, más algunas recomendaciones para trabajos futuros.

Capítulo 2

Marco Teórico

En este Capítulo se hará una contextualización teórica de esta tesis y haremos la definición del problema objetivo. Este capítulo consta de tres secciones. En la Sección 2.1 se presentan conceptos básicos de teoría de grafos, de este modo se presentan definiciones de grafos y digrafos. En la Sección 2.2 se muestran conceptos básicos de estadística como muestra aleatoria y probabilidad de selección. Por último, en la Sección 2.3 se presenta el problema a resolver y los objetivos de esta tesis.

2.1. Conceptos básicos de teoría de grafos.

Para modelar las relaciones entre individuos de una población se utiliza un grafo, entonces la idea de esta sección es mostrar definiciones y propiedades de grafos y digrafos. Primero comenzaremos definiendo un grafo, que es un objeto matemático base para esta tesis.

Definición 2 (Grafo)**.** *Un grafo G es un par ordenado $G = (V, E)$, en donde $V \neq \emptyset$ es el conjunto de vértices y E es un conjunto de pares de vértices no ordenados denotados por $\{u, v\}$, con $u, v \in V$. Por lo tanto denotaremos por $V(G)$ a los vértices del grafo G y por $E(G)$ a las aristas del grafo G.*

En la Figura 2.1.1 se muestra una representación gráfica de un grafo.

Para cualquier $v \in V(G)$ se tiene que la vecindad de v son los nodos con que v comparte arista en G. Por lo tanto, si denotamos por $\mathcal{N}(v)$ a la vecindad de v, entonces se tiene que:

$$\mathcal{N}(v) = \{u \in V(G) : \{u, v\} \in E(G)\}.$$

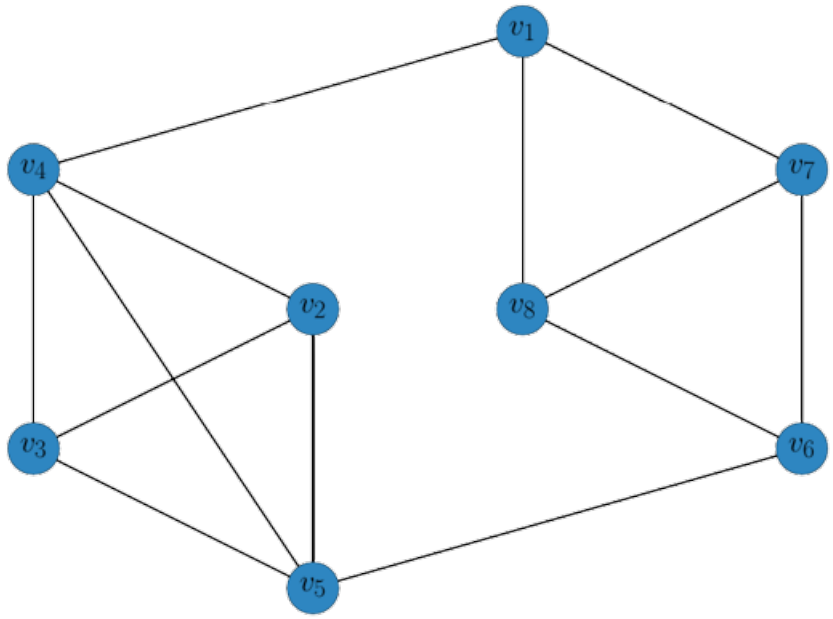


Figura 2.1.1: La figura muestra un la representación gráfica de un grafo $G = (V, E)$, en donde $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$ y $E = \{\{v_1, v_4\}, \{v_1, v_7\}, \{v_1, v_8\}, \{v_2, v_3\}, \{v_2, v_4\}, \{v_2, v_5\}, \{v_3, v_4\}, \{v_3, v_5\}, \{v_4, v_5\}, \{v_5, v_6\}, \{v_6, v_7\}, \{v_7, v_8\}\}$.

Además, el tamaño de $\mathcal{N}(v)$ recibe el nombre de grado de v para cualquier v en $V(G)$ y se denota por $d(v)$. Otra cualidad importante de un grafo es si está conectado o no. Para definir esto debemos formalizar qué es un camino.

Definición 3 (Camino)**.** *Sea $G = (V, E)$ un grafo. Un camino es una sucesión de nodos $\{x_1, x_2, ..., x_k\}$ y aristas $e_1, e_2, ..., e_k$ tal que $e_j = \{x_j, x_{j+1}\} \in E(G)$ para $j = 1, ..., k-1$.*

De aquí podemos obtener la definición de uv-camino.

Definición 4 (uv-Camino)**.** *Dado un grafo $G = (V, E)$ y los nodos $u, v \in V(G)$. Entonces un uv-camino es un camino en donde su sucesión de nodos $\{x_1, x_2, ..., x_k\}$ cumple que $x_1 = u$ y $x_k = v$.*

De esta forma se tendrá que dos nodos $u, v \in V(G)$ están conectados si existe un uv-camino, además si para cualquier par de vértices de $V(G)$ resulta que están conectados entonces el grafo G se dice que es conexo o está conectado.

Ahora se definirá un tipo de grafo llamado grafo bipartito.

Definición 5 (Grafo bipartito)**.** *Un grafo $G = (V, E)$ es llamado bipartito si existe A*

y B tal que A, B son no vacíos, $A \cap B = \emptyset$, $A \cup B = V(G)$ y todas las aristas de G tienen un nodo en A y otro en B.

En la Figura 2.1.2 se muestra una representación gráfica de un grafo bipartito.

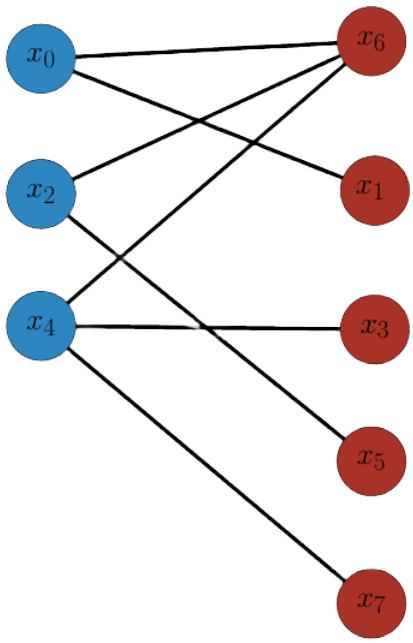


Figura 2.1.2: Representación gráfica de un grafo bipartito, en donde el conjunto A son los nodos de color azul y el conjunto B son los nodos de color rojo.

También se utilizará la noción de digrafo, por lo tanto se formaliza a continuación.

Definición 6 (Digrafo)**.** *Un digrafo es un par ordenado $D = (V, F)$ en donde, $V \neq \emptyset$ es un conjunto de nodos y $F \subseteq V \times V$ es un conjunto de aristas, siendo F un conjunto de pares de nodos ordenados. Por lo tanto denotaremos por $V(D)$ a los nodos del digrafo D y por $F(D)$ a las aristas del digrafo D.*

En la figura 2.1.3 se muestra una representación gráfica de un digrafo. De esta forma, la diferencia entre un grafo y un digrafo es el conjunto de aristas, pues en uno sus elementos no son ordenados y en el otro si. Por lo tanto sea un digrafo $D = (V, F)$, entonces para cualquier $u, v \in V(D)$ tal que $(u, v) \in F(D)$ se denota por una flecha que inicia en u y apunta hacia v.

Un concepto relacionado con lo anterior, es hacia donde apuntan las fechas y de donde salen, por lo tanto se definen las vecindades entrantes y salientes a continuación.

Definición 7 (Vecindad entrante)**.** *Sea un digrafo $D = (V, F)$, entonces para cualquier $v \in V(D)$ se tiene que la vecindad entrante de v en el digrafo D son los nodos que forman aristas del tipo $(\cdot, v)$ en el digrafo D, es decir si denotamos por $\mathcal{N}^-(v)$ a la vecindad entrante de v entonces se tiene que:*

$$\mathcal{N}^-(v) = \{u \in V(D) : (u, v) \in F(D)\}.$$

Por otro lado, se define la vecindad saliente.

Definición 8 (Vecindad saliente)**.** *Sea un digrafo $D = (V, F)$, entonces para cualquier $v \in V(D)$ se tiene que la vecindad saliente de v en el digrafo D son los nodos que forman aristas del tipo $(v, \cdot)$ en el digrafo D, es decir si denotamos por $\mathcal{N}^+(v)$ a la vecindad saliente de v entonces se tiene que:*

$$\mathcal{N}^+(v) = \{u \in V(D) : (v, u) \in F(D)\}.$$

Análogo a los grafos, los nodos de los grafos dirigidos tienen el concepto de grado entrante y grado saliente, que se denota por $d^-(v)$ y $d^+(v)$ para cualquier nodo v en $V(G)$ respectivamente, y se definen como $d^-(v) = |\mathcal{N}^-(v)|$ y $d^+(v) = |\mathcal{N}^+(v)|$.

Por último, se puede relacionar un grafo G con un digrafo D, esta relación se define a continuación.

Definición 9 (Grafo subyacente de un digrafo)**.** *Sea un digrafo $D = (V, F)$, entonces el grafo subyacente asociado al digrafo D es definido como $G = (V, E)$ en donde $E = \{\{v, u\} \subseteq V(G) : (v, u) \in F(D)$ o $(u, v) \in F(D)\}$.*

2.2. Conceptos básicos estadísticos.

Para el trabajo de esta memoria necesitaremos de herramientas estadísticas, por lo tanto se define una muestra aleatoria como propone Arnab, Raghunath en [1].

Definición 10 (Muestra aleatoria)**.** *Una muestra aleatoria es una selección de elementos de la población T utilizando cierta regla o método y se denota por $s = \{x_1, ..., x_m\}$, en donde cada x_j pertenece a T con $j = 1, ..., m$.*

En la definición de muestra aleatoria podemos notar que podríamos obtener una muestra aleatoria utilizando distintos métodos o reglas, que se conocen como diseño muestral.

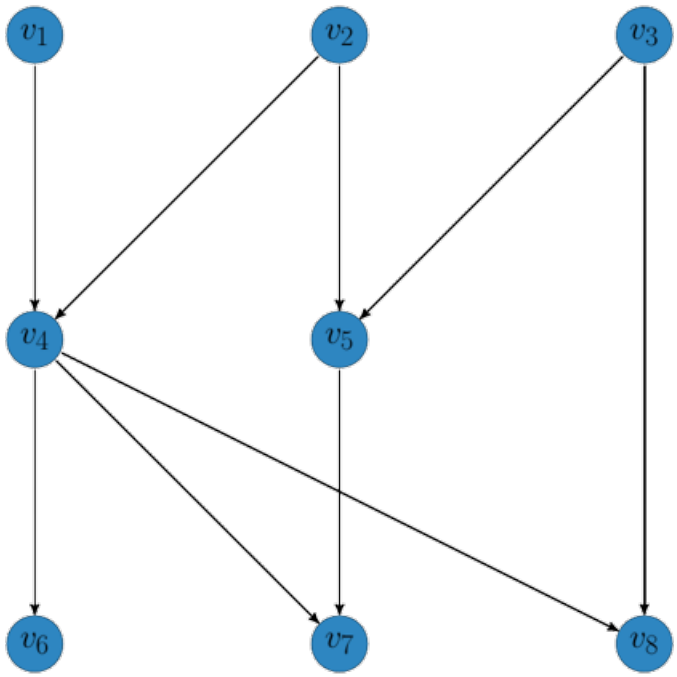


Figura 2.1.3: Representación gráfica de un digrafo $D = (V, F)$, en donde $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$ y $F = \{(v_1, v_4), (v_2, v_4), (v_2, v_5), (v_3, v_5), (v_3, v_8), (v_4, v_6), (v_4, v_7), (v_4, v_8), (v_5, v_7)\}$.

Según la definición de una muestra aleatoria, se puede notar que los elementos de una muestra aleatoria se pueden repetir o no dentro de ella [1].

Definición 11 (Muestra aleatoria con repetición)**.** *Una muestra aleatoria con repetición es una selección de elementos de la población T utilizando cierta regla o método (Diseño muestral) en donde los elementos de la muestra aleatoria pueden repetir.*

Por lo tanto, se puede definir a una muestra aleatoria sin repetición. De esta forma, se utiliza la definición de [1] para muestra aleatoria sin repetición.

Definición 12 (Muestra aleatoria sin repetición)**.** *Una muestra aleatoria sin repetición es una selección de elementos de la población T utilizando cierta regla o método (Diseño muestral) en donde los elementos de la muestra aleatoria no pueden repetir.*

Si suponemos que tenemos una muestra aleatoria $s = \{x_1, ..., x_m\}$ de tamaño m y en donde cada elemento de s está en la población T, sería útil conocer con que probabilidad cualquier elemento $v \in T$ es seleccionado en la j-ésima posición de la muestra s con $j = 1, ..., m$. De esta forma, se utiliza la definición de probabilidad de selección de [1].

Definición 13 (Probabilidad de selección)**.** *Dada una población $T = \{v_1, ..., v_N\}$ y un muestreo aleatorio de tamaño m en que seleccionamos elementos de la población T. Entonces se define la probabilidad de selección $p_j(v)$ asociada al muestreo aleatorio, como la probabilidad que se tiene para seleccionar el elemento v de la población T en la j-ésima extracción de una muestra aleatoria con $v \in T, j = 1, ..., m$, y además se cumple que*

$$0 \leq p_j(v) \leq 1 \quad y \quad \sum_{v \in T} p_j(v) = 1 \quad j = 1, \ldots, m, \ \forall v \in T.$$

Una cosa interesante a notar es que si la probabilidad de selección no depende de j, es decir de la posición en la muestra aleatoria, entonces la probabilidad de selección del individuo v se podrá denotar por $p(v)$.

Los artículos en donde son presentados la mayoría de los estimadores de grupo, los presentan utilizando muestreos aleatorios con repetición. Por lo tanto, en esta tesis, se realizarán principalmente muestreos con repetición para muestrear elementos de la población, de esta forma a continuación se definirá el muestreo aleatorio uniforme con repetición como es definido en [1].

Definición 14 (Muestreo aleatorio uniforme con repetición)**.** *Dada una población $T = \{v_1, ..., v_N\}$ y un diseño muestral en donde la muestra aleatoria es de tamaño fijo $m \in \mathbb{N}$ y conocido. Entonces, para cualquier $j = 1, ..., m$, la j-ésima extracción de la muestra aleatoria s es seleccionada desde la población T de forma uniforme, es decir para cualquier $j = 1, ..., m$, cada elemento de T tiene $\frac{1}{N}$ probabilidad de ser seleccionada en la j-ésima extracción.*

Así por ejemplo, si el muestreo es de tipo uniforme con repetición de tamaño 3 y se está muestreando el conjunto $T = \{1, 2, 3, 4, 5, 6, 7\}$ con $N = 7$, entonces la probabilidad de selección del elemento 2 en T en la j-ésima extracción con $j = 1, 2, 3$ es $p_j(2) = 1/7$, pues el elemento 2 de la población tienen probabilidad $1/7$ de ser extraído en la posición j-ésima de la muestra para $j = 1, 2, 3$. Más adelante, veremos que la probabilidad de selección puede depender del elemento de la población de la que se calcula la probabilidad de selección. Por otro lado, la probabilidad de selección $p_j(v)$ también puede depender de j.

A continuación se hará una breve introducción de variables aleatorias. Para esto, debemos definir qué es un experimento aleatorio.

Definición 15 (Experimento aleatorio)**.** *Un experimento aleatorio es cualquier acción o proceso cuyo resultado está sujeto a la incertidumbre.*

Ahora existen diferentes tipos de experimentos aleatorios, uno de ellos es el siguiente.

Definición 16 (Experimento binomial)**.** *Un experimento aleatorio se llama experimento binomial si cumple lo siguiente:*

- *el experimento consta de una secuencia de m experimentos más pequeños llamados ensayos, donde m se fija antes del experimento*
- *cada ensayo tiene dos posibles resultados, los cuales se denotan como éxito (S) y como falla (F)*
- *los ensayos son independientes, de modo que el resultado en un ensayo particular no influye en el resultado de cualquier otro ensayo*
- *la probabilidad de éxito $Pr(S)$ es constante de un ensayo a otro; esta probabilidad se denota por p*

A partir de este experimento aleatorio se define una variable aleatoria binomial.

Definición 17 (Variable aleatoria binomial X asociada a un experimento binomial)**.** *Sea un experimento binomial de m ensayos y probabilidad de éxito p fija, y se define la variable aleatoria X como la cantidad de éxitos que tiene el experimento binomial, entonces X se llama variable aleatoria binomial X asociada a un experimento binomial o simplemente variable aleatoria binomial de parámetros $(\hat{m}, \hat{p})$.*

Además, se tiene que la probabilidad de $X = x$ es:

$$Pr(\ X = x\ ) = \begin{cases} \binom{m}{x} p^x (1-p)^{m-x} & x = 0, 1, 2, \ldots m \\ 0 & si\ x \notin \{0, ..., m\} \end{cases}$$

Otro tipo de experimento es el hipergeométrico que se define a continuación.

Definición 18 (Experimento hipergeométrico)**.** *Un experimento aleatorio se llama experimento hipergeométrico si cumple lo siguiente:*

- *la población o conjunto que se va a muestrear se compone de N individuos, objetos o elementos (una población finita)*

- *la población está divida por dos clases de individuos, los de tipo A y los de tipo B, en donde la cantidad de individuos del tipo A son N_A*
- *se selecciona una muestra uniforme sin repetición s de la población de tamaño m*

A partir de este experimento aleatorio se define una variable aleatoria hipergeométrica.

Definición 19 (Variable aleatoria hipergeométrica X asociada a un experimento hipergeométrico)**.** *Sea un experimento hipergeométrico en donde la población es de tamaño N, la cantidad de individuos de clase A son N_A y la muestra uniforme sin repetición s de la población es de tamaño m. Se define X como una variable aleatoria que indica la cantidad de individuos del tipo A en la muestra s, entonces X se llama variable aleatoria hipergeométrica asociada a un experimento hipergeométrico o simplemente variable aleatoria hipergeométrica de parámetros (N, N_A, m).*

Además la probabilidad de $X = x$, con $x \in [\ \max\{0, m-(N-N_A)\}, \min\{N_A, m\}\ ]$ es:

$$Pr(\ X = x\ ) = \frac{\begin{pmatrix} N_A \\ x \end{pmatrix}\begin{pmatrix} N - N_A \\ m - x \end{pmatrix}}{\begin{pmatrix} N \\ m \end{pmatrix}}.$$

Si definimos un experimento aleatorio que consiste en realizar un muestreo aleatorio uniforme con repetición, entonces se puede obtener lo siguiente.

Proposición 1. *Sea un grafo $G = (V, E)$ en donde $V(G) = \{v_1, ..., v_N\}$ y si consideramos un experimento aleatorio que consiste en realizar un muestreo aleatorio uniforme con repetición de nodos $s = \{x_1, ..., x_m\}$, entonces la fracción esperada de nodos de cierto grado d en la muestra s es igual la fracción de nodos de grado d en el grafo.*

Demostración. Sea un grafo $G = (V, E)$ con $V(G) = \{v_1, ..., v_N\}$ y si definimos un experimento aleatorio que consiste en realizar un muestreo aleatorio uniforme con repetición de nodos $s = \{x_1, ..., x_m\}$, entonces la probabilidad de selección para cualquier nodo del grafo es $\frac{1}{N}$. Luego, la cantidad esperada de veces que aparezca un nodo $v \in V(G)$ en la muestra es $\frac{m}{N}$. Si consideramos $\nu(V(G), d)$ como la cantidad de nodos de grado d en el grafo G, entonces la cantidad esperada de nodos de grado d en la muestra es

$\nu(V(G), d) \cdot \frac{m}{N}$. Por lo tanto, la fracción esperada de nodos de grado en la muestra s es:

$$\frac{\nu(V(G), d) \cdot \frac{m}{N}}{m} = \frac{\nu(V(G), d)}{N}.$$

Por otro lado, la fracción de nodos de grado d del grafo G es $\frac{\nu(V(G),d)}{N}$, es decir que la fracción esperada de nodos de grado d de una muestra s es igual a la fracción de nodos de grado del grafo. □

2.3. Modelado del problema y objetivos de la tesis.

En esta sección se presentará un problema que modela el Problema 1 y los objetivos de esta tesis. De esta forma se define el siguiente problema.

Problema 2. *Dado un grafo $G = (V, E)$ desconocido con $|V(G)| = N$ conocido y un subconjunto $H \subset V(G)$ desconocido. Entonces realizando una muestra aleatoria de los nodos de G para consultar su información local, se debe estimar el tamaño del subconjunto H*

A continuación se mostrará que el Problema 2 puede modelar el Problema 1 a estudiar. Una población cualquiera, en donde los individuos de esta pueden o no estar relacionados y estas relaciones son simétricas se modela a través de un grafo $G = (V, E)$ en donde sólo se conoce el tamaño $|V(G)|$. Además el grupo a estimar su tamaño en la población es representado como $H \subset V(G)$ en el grafo G. Por lo tanto, una selección de personas a través de una muestra aleatoria en la población puede ser representada por una selección de nodos del mismo tipo en el grafo G. Podemos notar que la encuesta también puede ser modelada en el 2. Por lo tanto el 2 puede modelar el Problema 1.

Existen varios tipos de encuestas, estas se definen a continuación.

Definición 20 (Encuesta indirecta o directa)**.** *Dada una población, un grupo desconocido de la población y una encuesta. Si en la encuesta no sé pregunta al encuestado por la cantidad de conocidos en el grupo desconocido entonces la encuesta es de tipo directa, en caso contrario se dice que es una encuesta de tipo indirecta.*

También puede darse este otro caso.

Definición 21 (Encuesta indirecta exclusiva)**.** *Dada una población, un grupo desconocido de la población y una encuesta indirecta. Si en la encuesta no sé pregunta al*

encuestado si está o no en el grupo desconocido entonces la encuesta es de tipo indirecta exclusiva.

Existen una variedad de modelos para resolver el Problema 2. La idea de esta tesis es comparar los estimadores en función del tipo de encuesta que utilicen.

Objetivos del proyecto .
Si consideramos el experimento aleatorio asociado a un método de estimación como la realización de la muestra aleatoria para luego realizar la estimación del tamaño de grupo a través del método de estimación correspondiente, entonces podemos inferir que la estimación del método del experimento aleatorio es una variable aleatoria, por lo tanto el error relativo del método de estimación también es una variable aleatoria.

De esta forma, fijando una población T y un grupo desconocido $H \subset T$ a estimar su tamaño, se puede extraer una muestra del error relativo del método de estimación realizando varias veces el experimento antes definido. De esta forma se puede estudiar la variable aleatoria del error relativo de cada método a través de una muestra de este y utilizando inferencia, es decir utilizando la media, desviación estándar e intervalos de confianza de la muestra.

Así, los objetivos concretos de este proyecto son:

O1: Buscar, implementar y comprender distintos de métodos escalables para la estimación de tamaños de grupos dentro de una población que utilicen distintos tipos de encuestas.

O2: Evaluar los métodos propuestos a través de un análisis experimental que estudie la media, desviación estándar e intervalo de confianza de una muestra aleatoria del error relativo de cada método implementado en distintos contextos.

O3: Aplicar los métodos propuestos en grafos reales y evaluar los resultados utilizando las medidas antes descritas.

Capítulo 3

Métodos de Muestreo y sus estimadores

En esta sección se mostrarán definiciones y propiedades que serán necesarias para utilizar los estimadores de tamaño de grupo. Primero se definirá el estimador de Hansen-Hurwitz, que es el estimador base de la mayoría de los estimadores de la sección siguiente. Luego se verá cómo estimar propiedades de un grafo dado a partir de una muestra uniforme con repetición, y por último se definirán otras formas de realizar muestreos aleatorios y cómo utilizarlos para estimar propiedades de un grafo dado.

3.1. Estimador de Hansen-Hurwitz

En este informe se deberán estimar propiedades de los grafos que son necesarias para poder utilizar algunos de los estimadores de subgrupo de nodos de los grafos. Para realizar la tarea anterior, se utilizará el estimador de Hansen-Hurwitz, pues debido a que es muy preciso y además se puede aplicar en contextos muy amplios. A continuación se muestra la definición del estimador de Hansen-Hurwitz hecha por S. G. Prabhu-Ajgaonkar en [13].

Definición 22 (Estimador de Hansen-Hurwitz)**.** *Sea una población finita* $T = \{v_1, v_2, ..., v_N\}$*, para* $i = 1, ..., N$ *se tiene que* $v_i \in T$ *tiene asociado dos números* $R(v_i)$ *y* $Y(v_i)$*, en donde* $R(v_i)$ *es conocido e* $Y(v_i)$ *es desconocido. Entonces, para estimar* $Y = \sum_{i=1}^{N} Y(v_i)$ *se usa un muestreo con repetición* s *de tamaño* m *en donde la probabilidad de selección de* v_i *es proporcional a la medida* $R(v_i)$ *con*

$i = 1, .., N$, así el estimador queda como

$$Y = \sum_{v \in T} Y(v) \approx \frac{1}{m} \sum_{x \in s} \frac{y(x)}{p(x)}$$

en donde $p(x_j) = \frac{R(x_j)}{\sum_{i=1}^{N} R(v_i)}$ e $y(x_j)$ representa los valores de $Y(x_j)$ asociado a los x_j muestreados $s = \{x_1, ..., x_m\}$, con $j = 1, .., m$.

El estimador de Hansen-Hurwitz puede ser usado de distintas formas, un caso particular es dado una muestra $s = \{x_1, ..., x_m\}$, entonces el valor $y(x_j)$ puede estar compuesta por la función indicatriz con $j = 1, ..., m$, por lo tanto se define a continuación.

Definición 23 (Función indicatriz)**.** *Sea un espacio X y un subconjunto A de X entonces se tiene que la función $\mathcal{X}_A$ es la función indicatriz de A y se define como $\mathcal{X}_A : X \longrightarrow \{0, 1\}$ y tiene valor:*

$$\mathcal{X}_A(x) = \begin{cases} 1 \text{ si } x \in A \\ 0 \text{ si } x \notin A. \end{cases}$$

De esta forma utilizando el estimador de Hansen-Hurwitz y la función indicatriz se puede estimar medidas de sólo una parte de la población. La utilidad de la función indicatriz dentro del estimador de Hansen-Hurwitz se mostrará más adelante, pero por ejemplo si consideramos una población de personas $T = \{v_1, ..., v_N\}$, entonces podemos definir para todo $i \in \{1, ..., N\}$ a $Y(v_i) = f(v_i)$, en donde $f(v_i)$ es la cantidad de hijos que tiene la persona v_i. Ahora si hacemos $X(v_i) = f(v_i)\ \mathcal{X}_A(v_i)$ en donde $A \subseteq T$ es el conjunto de mujeres en T y $\mathcal{X}_A$ es la función indicatriz de A tal que $\mathcal{X}_A : T \to \{0, 1\}$, entonces si v_i es mujer $X(v_i)$ indicaría la cantidad de hijos que tiene v_i, y cero si v_i no es mujer. Por lo tanto, $\sum_{i=1}^{N} X(v_i)$ sería la cantidad de hijos que tienen las mujeres de la población y considerando una muestra aleatorio con repetición $s = \{x_1, ..., x_m\}$ entonces se podría realizar la siguiente estimación

$$\sum_{i=1}^{N} X(v_i) \approx \frac{1}{m} \sum_{x \in s} \frac{X(x)}{p(x)}.$$

3.2. Propiedades de muestreo aleatorio uniforme con repetición

En esta sección, se mostrará algunas propiedades o valores que serán requeridos más adelante para realizar las estimaciones de tamaños de grupos. Para utilizar el estimador Hansen-Hurwitz primero se necesita mostrar que existe una medida R relacionada a los integrantes de la población, debe ser conocida y debe cumplir que es directamente proporcional a la probabilidad de selección del método de muestreo. En la siguiente proposición se muestra esa medida R para el caso del muestreo uniforme con repetición.

Estimador 1. *Sea un grafo $G = (V, E)$ con $V = \{v_1, ..., v_N\}$ y una muestra aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$. Si para cada $i \in \{1, .., N\}$ se tiene que la probabilidad de selección del nodo v_i es $p(v_i)$ y tenemos una medida $R(v_i) = 1$, entonces para cada $i \in \{1, .., N\}$ $p(v_i)$ es proporcional a $R(v_i)$.*

Justificación: Dado un grafo $G = (V, E)$ con $V(G) = \{v_1, ..., v_N\}$, en donde para cada $i \in \{1, ..., N\}$ se tiene $R(v_i) = 1$. Si tenemos una muestra aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$, entonces por definición de muestreo aleatorio uniforme con repetición se tiene que la probabilidad de selección $p(v_i) = \frac{1}{N}$ para cualquier $i \in \{1, ..., N\}$. De lo planteado podemos ver que para cualquier $i \in \{1, ..., N\}$ se tiene que

$$\frac{p(v_i)}{R(v_i)} = \frac{\frac{1}{N}}{1} = \frac{1}{N} = \text{constante.}$$

Es decir la probabilidad de selección es proporcional a la medida R. □

Podemos utilizar el estimador de Hansen-Hurwitz para el estimar el grado medio del grafo, este resultado se muestra en la siguiente proposición.

Estimador 2. *Sea un grafo $G = (V, E)$, donde $V = \{v_1, ..., v_N\}$, y sea una muestra aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$. Entonces, el grado medio del grafo G se puede estimar a través de la expresión:*

$$\frac{1}{N} \sum_{v \in V} d(v) \approx \frac{1}{m} \sum_{x \in s} d(x).$$

Justificación: Dado un grafo $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$, y para cada $i \in \{1, ..., N\}$ se tiene $R(v_i) = 1$ y $Y(v_i) = \frac{d(v_i)}{N}$. Si tenemos una muestra aleatoria

uniforme con repetición $s = \{x_1, ..., x_m\}$, entonces por la Proposición 1 se tiene que para cualquier $i \in \{1, ..., N\}$ $p(v_i)$ es proporcional a $R(v_i)$. De esta forma, podemos usar el estimador de Hansen-Hurwitz y obtener

$$\frac{1}{N}\sum_{v\in V} d(v) \approx \frac{1}{m}\sum_{x\in s} d(x).$$

□

De aquí podemos estimar $2|E|$ de un grafo $G = (V, E)$ y se muestra en la siguiente proposición.

Estimador 3. *Sea un grafo $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$ y una muestra aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$. Entonces, el doble de la cantidad de aristas del grafo G se puede estimar a través de la expresión*

$$2\ |E| = \sum_{v\in V} d(v) \approx \frac{N}{m}\sum_{x\in s} d(x).$$

También podemos estimar la suma de los grados de un subconjunto de $V(G)$.

Estimador 4. *Sea un grafo $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$, una muestra aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$ y un conjunto $A \subset V(G)$. Si $\mathcal{X}_A : V(G) \to \{0, 1\}$ es la función indicatriz de A, entonces se puede estimar la suma de los grados del conjunto A como*

$$\sum_{v\in A} d(v)\ = \sum_{v\in V} \mathcal{X}_A(v)\ d(v) \approx \frac{N}{m}\sum_{x\in s} \mathcal{X}_A(x)\ d(x).$$

Justificación: Dado un grafo $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$, y para cada $i \in \{1, ..., N\}$ se tiene $R(v_i) = 1$ y $Y(v_i) = \mathcal{X}_A(v_i)\ d(v_i)$. Si tenemos una muestra aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$, entonces usando la Proposición 1 se tiene que para cualquier $i \in \{1, ..., N\}$ $p(v_i)$ es proporcional con $R(v_i)$. De esta forma, podemos usar el estimador de Hansen-Hurwitz y obtener:

$$\sum_{v\in V} \mathcal{X}_A(v)\ d(v) \approx \frac{N}{m}\sum_{x\in s} \mathcal{X}_A(x)\ d(x).$$

Además podemos estimar el grado medio del subconjunto A usando el estimador de Hajek, que no es más que la división de dos estimadores de Hansen-Hurwitz. De esta forma podemos formar una proposición que estime el grado medio de los nodos en $A \subset V(G)$.

Estimador 5. *Sea un grafo $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$, una muestra aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$ y un conjunto $A \subset V(G)$. Si $\mathcal{X}_A : V(G) \to \{0, 1\}$ es la función indicatriz de A, entonces el grado medio del conjunto A se puede estimar como:*

$$\frac{\sum_{v\in A} d(v)}{\sum_{v\in A} 1} = \frac{\sum_{v\in V} \mathcal{X}_A(v)\ d(v)}{\sum_{v\in V} \mathcal{X}_A(v)} \approx \frac{\sum_{x\in s} \mathcal{X}_A(x)\ d(x)}{\sum_{x\in s} \mathcal{X}_A(x)}.$$

Justificación: Para obtener este estimador debemos utilizar dos estimadores de Hansen-Hurwitz. Dado un grafo $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$ y para cada $i \in \{1, ..., N\}$ se tiene $R(v_i) = 1$, $Y(v_i) = \mathcal{X}_A(v_i)\ d(v_i)$ y $Z(v_i) = \mathcal{X}_A(v_i)$. Si tenemos una muestra aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$, entonces usando la Proposición 1 se tiene que para cualquier $i \in \{1, ..., N\}$ $p(v_i)$ es proporcional con $R(v_i)$. De esta forma, para cualquier $i \in \{1, .., N\}$ se tiene que $R(v_i) = 1$ y $Y(v_i) = \mathcal{X}_A(v_i)\ d(v_i)$, así utilizando un estimador de Hansen-Hurwitz asociado a estas valores podemos obtener:

$$\sum_{v\in V} \mathcal{X}_A(v)\ d(v) \approx \frac{1}{m} \sum_{x\in s} \frac{\mathcal{X}_A(x)\ d(x)}{p(x)} = \frac{N}{m} \sum_{x\in s} \mathcal{X}_A(x)\ d(x).$$

Para el otro estimador de Hansen-Hurwitz usamos las variables asociadas $R(v_i) = 1$ y $Z(v_i) = \mathcal{X}_A(v_i)$ para cualquier $i \in \{1, ..., N\}$ y la muestra aleatorio uniforme con repetición $s = \{x_1, ..., x_m\}$, así

$$\sum_{v\in V} \mathcal{X}_A(v) \approx \frac{1}{m} \sum_{x\in s} \frac{\mathcal{X}_A(x)}{p(x)} = \frac{N}{m} \sum_{x\in s} \mathcal{X}_A(x).$$

Por lo tanto, dividiendo ambos estimadores se obtiene que:

$$\frac{\sum_{v\in V} \mathcal{X}_A(v)\ d(v)}{\sum_{v\in V} \mathcal{X}_A(v)} \approx \frac{\frac{1}{m}\sum_{x\in s} \frac{\mathcal{X}_A(x)\ d(x)}{p(x)}}{\frac{1}{m}\sum_{x\in s} \frac{\mathcal{X}_A(x)}{p(x)}} = \frac{\sum_{x\in s} \mathcal{X}_A(x)\ d(x)}{\sum_{x\in s} \mathcal{X}_A(x)}.$$

Obteniendo el estimador del grado medio del subconjunto A de $V(G)$. □

3.3. Otros tipos de muestreos y sus estimadores asociados

En esta sección se definirán tres nuevos tipos de muestreos asociados a la caminata aleatoria, el proceso viral probabilístico y el proceso viral constante. Además, se mostrarán algunas propiedades de estos tipos de muestreos que serán utilizadas para estimadores propiedades para un grafo dado. A continuación se mostrará una descripción de la caminata aleatoria, proceso viral probabilístico y el proceso viral constante, y se citaran los artículos relacionados en donde se recomiendan utilizar estos mecanismos para tomar una muestra aleatoria.

Caminata aleatoria.
Dado un grafo $G = (V, E)$ y un valor $m \in \mathbb{N}$, la caminata aleatoria de longitud m consiste en:

- Elegir aleatoriamente un elemento de $V(G)$ en donde cada elemento tiene la misma probabilidad de ser seleccionado y se define como el nodo del paso 1.
- Luego, se elige aleatoriamente un nodo vecino del nodo del paso 1, donde cada nodo vecino tiene la misma probabilidad de ser elegido. Este nodo vecino seleccionado se define como el nodo del paso 2. El proceso se repite hasta obtener el nodo del paso m.

Una vez que se tienen m nodos, la secuencia de nodos formará la caminata aleatoria, denotada por $s = \{x_1, ..., x_m\}$. Los nodos de s también se pueden describir como los nodos visitados por la caminata aleatoria.

Notar que el valor m que indica el tamaño de la caminata aleatoria, no necesariamente debe ser menor o igual al tamaño de $V(G)$ pues el método puede repetir nodos, es decir puede visitar a un nodo más de una vez. De hecho, Rodero-Merino, Luis; Anta, Antonio Fernández; López, Luis; Cholvi, Vicent publicaron en [14] una estimación de la cantidad de nodos diferentes visitados por el caminata aleatoria.

En [14] se afirma que una caminata aleatoria puede ser utilizada para crear una muestra aleatoria. Ahora como tenemos una muestra aleatoria asociada a una caminata aleatoria, entonces en la Figura 3.3.1 se muestra la relación entre esta muestra aleatoria y su probabilidad de selección. En la Figura 3.3.1 se muestra un ejemplo de un experimento aleatorio en donde se ejecuta una caminata aleatoria de 6 pasos. Por este motivo se

ha indexado a cada paso una variable aleatoria X_j que modele el valor del paso k con $k = 1, ..., 6$. Entonces, para esta caminata aleatoria en particular, se tiene que la secuencia de nodos es $s = \{X_1, ..., X_6\}$ y su probabilidad de selección es:

$$Pr(X_k = v_i) = p_k(v_i) \qquad \text{con } i = 1, ..., N; k = 1, ..., 6 \text{ si } V(G) = \{v_1, ..., v_N\}.$$

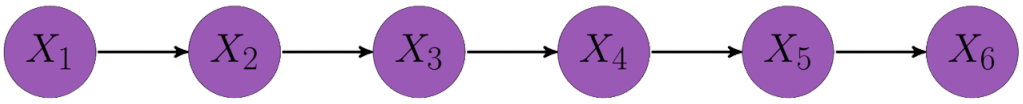


Figura 3.3.1: Ejemplo de una caminata aleatoria.

Proceso viral constante.
Dado un grafo $G = (V, E)$ y valores $c, m \in \mathbb{N}$, entonces el proceso viral constante de tamaño m y velocidad propagación c consiste en:

- Se debe elegir al azar un nodo de $V(G)$ con probabilidad uniforme a la cual se le llamará la ola 1 y se revisa si la cantidad de nodos seleccionados es mayor o igual a m.
- Luego, se eligen de forma aleatoria uniforme sin repetición c nodos vecinos del nodo de la ola 1. Estos nodos elegidos pertenecerán a la ola 2 y se revisa si el total de los nodos seleccionados es mayor o igual a m.
- Más tarde, se eligen de forma aleatoria uniforme sin repetición c nodos vecinos de cada nodo de la ola 2, en donde un nodo puede ser seleccionado más de una vez en esta ola. Estos nodos elegidos pertenecerán a la ola 3, y se revisa si el total de los nodos seleccionados es mayor o igual a m.
- Esto persiste hasta que el tamaño de la selección alcance o supere el valor de m. Si se da el caso en que se supere el valor m, entonces se eliminarán de forma aleatoria uniforme los nodos de la última generación para obtener m nodos en la selección.

Si consideramos a cada ola del proceso anterior como un conjunto de nodos, entonces la concatenación de los conjuntos de las olas formará la secuencia de nodos del proceso viral constante denotada por $s = \{x_1, ..., x_m\}$. Notar que en el caso que en una iteración, un nodo tenga menos que c vecinos, entonces se seleccionan todos los vecinos de ese

nodo. Además, se puede ver que si $c = 1$ entonces el proceso viral constante es una caminata aleatoria, por tanto el proceso viral constante es una generalización de la caminata aleatoria.

Salganik, Matthew J, Heckathorn, Douglas D publican en [16] que un proceso viral constante puede ser utilizado para crear una muestra aleatoria. Ahora como tenemos una muestra aleatoria asociada a un proceso viral constante, entonces en la Figura 3.3.2 se muestra la relación entre esta muestra aleatoria y su probabilidad de selección. En la Figura 3.3.2 se muestra un ejemplo de un experimento aleatorio en donde se ejecuta un proceso viral constante de tamaño 15. Por este motivo se ha indexado a cada nodo de la selección a una variable aleatoria X_k que modele el nodo que podría seleccionar, con $k = 1, ..., 15$. Entonces, para este proceso viral constante en particular, se tiene que la secuencia de nodos es $s = \{X_1, ..., X_{15}\}$ y su probabilidad de selección es:

$$Pr(X_k = v_i) = p_k(v_i) \qquad \text{con } i = 1, ..., N; k = 1, ..., 15 \text{ si } V(G) = \{v_1, ..., v_N\}.$$

Notar que en la Imagen 3.3.2 se considero un grafo en donde todos los nodos tienen grado mayor a $c = 2$.

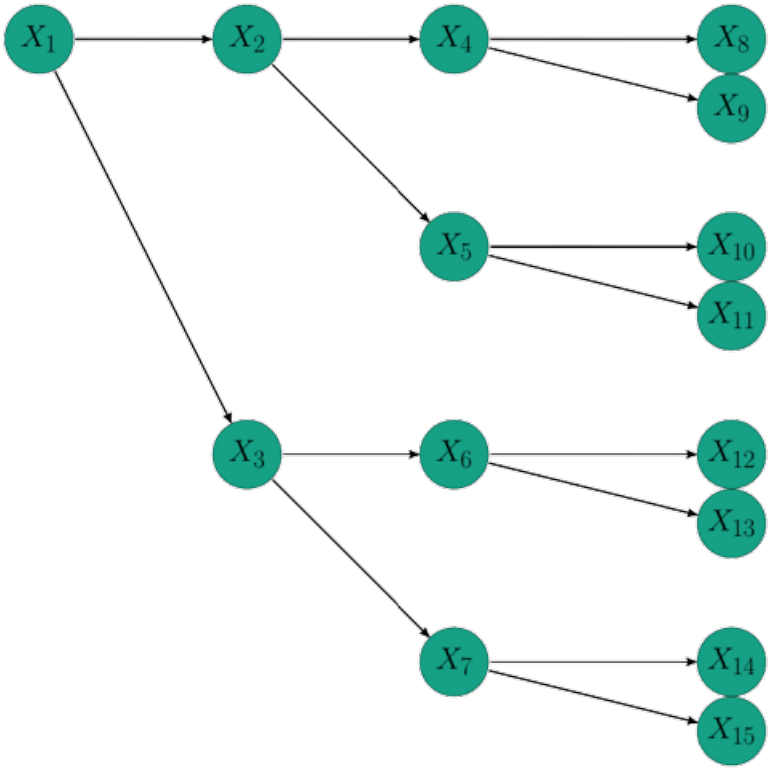


Figura 3.3.2: Ejemplo de un proceso viral constante.

Proceso viral probabilístico.
Dado un grafo $G = (V, E)$, y valores $m \in \mathbb{N}$ y $p \in [0, 1]$, entonces el proceso viral probabilístico de tamaño m consiste en:

- Se debe elegir al azar un nodo de $V(G)$ con probabilidad uniforme a la cual se le llamará la ola 1 y se revisa si la cantidad de nodos seleccionados es mayor o igual a m.
- Luego, cada vecino del nodo de la ola 1 tiene probabilidad p de ser seleccionado. Estos nodos seleccionados pertenecerán a la ola 2 y se revisa si el total de los nodos seleccionados es mayor o igual a m.
- Más tarde, para cada nodo de la ola 2, se seleccionan sus nodos vecinos con probabilidad p, en donde un nodo puede ser seleccionado más de una vez y la vecindad de un nodo se considera tantas veces como el nodo esté en la ola 2. Estos nodos elegidos pertenecerán a la ola 3 y se revisa si el total de los nodos seleccionados es mayor o igual a m.
- Esto persiste hasta que el tamaño de la selección alcance o supere el valor de m. Si se da el caso en que se supere el valor m entonces, se eliminarán de forma aleatoria uniforme los nodos de la última generación para obtener m nodos en la selección.

Si consideramos a cada ola del proceso anterior como un conjunto de nodos, entonces la concatenación de los conjuntos de las olas formará la secuencia de nodos del proceso viral probabilístico denotada por $s = \{x_1, ..., x_m\}$.

Debido al parecido entre proceso viral probabilístico y proceso viral constante, y que el proceso viral constante es recomendado en [16] para realizar muestreos aleatorios, entonces pensamos que sería buena idea utilizarlo para crear muestreos aleatorios.

Ahora como tenemos una muestra aleatoria asociada a un proceso viral probabilístico, entonces en la Figura 3.3.3 se muestra la relación entre esta muestra aleatoria y su probabilidad de selección. En la Figura 3.3.3 se muestra un ejemplo de un experimento aleatorio en donde se ejecuta un proceso viral probabilístico de tamaño 14. Por este motivo se ha indexado a cada nodo de la selección a una variable aleatoria X_k que modele el nodo que podría seleccionar, con $k = 1, ..., 14$. Entonces, para este proceso viral probabilístico en particular, se tiene que la secuencia de nodos es $s = \{X_1, ..., X_{14}\}$

y su probabilidad de selección es:

$$Pr(X_k = v_i) = p_k(v_i) \qquad \text{con } i = 1, ..., N; k = 1, ..., 14 \text{ si } V(G) = \{v_1, ..., v_N\}.$$

Notar que el primer nodo en una caminata aleatoria, en un proceso viral constante y en un proceso viral probabilístico es elegido de forma uniforme y por este motivo tomaremos el consejo establecido en [16] en donde se recomienda eliminarlo. Por lo tanto de ahora en adelante denotaremos $s = \{x_1, ..., x_m\}$ a la secuencia de nodos hecha por los mecanismos mencionados anteriormente pero la muestra aleatoria asociada será $s - \{x_1\}$.

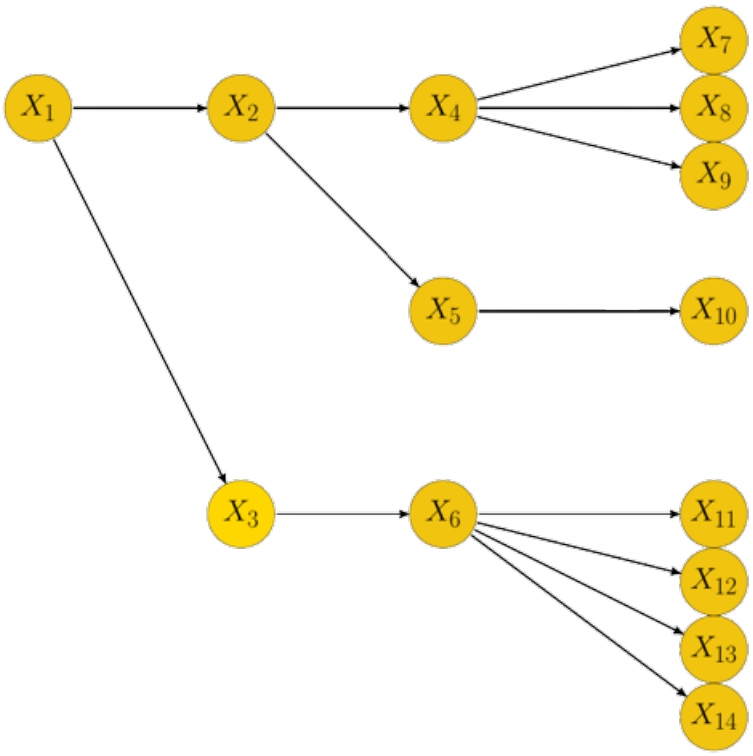


Figura 3.3.3: Ejemplo de un proceso viral probabilístico.

Ya que presentamos algunos métodos para crear muestreos aleatorios, a continuación se verán algunas propiedades de estos. Primero comenzaremos con el método más simple y particular que es la caminata aleatoria, y luego ramificaremos sus estimadores y propiedades a los métodos más complejos.

En el caso de utilizar la caminata aleatoria como método para crear una muestra aleatoria de los nodos de un grafo G, debemos obtener la probabilidad de selección de

cada nodo de $V(G)$ para poder realizar las estimaciones de los tamaños de grupo. Como mencionamos anteriormente, para este algoritmo la probabilidad de selección varía en función del grado de los nodos, y se determinará a través de algunas propiedades de cadena de Markov. Por lo tanto primero definiremos qué es una cadena de Markov.

Definición 24 (Cadena de Markov)**.** *Un proceso estocástico $X_1, X_2, X_3, \ldots$ es una cadena de Markov si:*

$$\begin{aligned}\Pr\left(X_t = a_t \mid X_{t-1} = a_{t-1}, X_{t-2} = a_{t-2}, \ldots, X_1 = a_1\right) &= \Pr\left(X_t = a_t \mid X_{t-1} = a_{t-1}\right) \\ &= P_{a_{t-1},a_t}.\end{aligned}$$

Dado un grafo $G = (V, E)$ y $m \in \mathbb{N}$, si ejecutamos sobre el grafo G una caminata aleatoria, podemos notar que la caminata aleatoria es una cadena de Markov, en donde los estados de la cadena de Markov son los nodos del grafo G. Debido a lo anterior, podemos usar las propiedades de la teoría hecha para cadenas de Markov, así podemos obtener la probabilidad de cada salto de la caminata aleatoria a través de la matriz de transición de cadena de Markov asociada:

$$\mathbf{P} = \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,j} & \cdots \\ P_{2,1} & P_{2,2} & \cdots & P_{2,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ P_{i,1} & P_{i,2} & \cdots & P_{i,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \end{pmatrix}.$$

El valor de la componente $P_{i,j}$ de la matriz $\mathbf{P}$ indica la probabilidad de ir del estado i al estado j, es decir pasar del nodo v_i al nodo v_j en la caminata aleatoria con $i, j = 1, ..., N$ y $v_i, v_j \in V(G)$. Para $i = 1, ..., N$ se denota por $p_i(t)$ a la probabilidad de estar en el estado i en el tiempo t, así podemos definir $\bar{p}(t)$ el vector que almacena todos los $p_i(t)$ posibles, es decir $\bar{p}(t) = (p_1(t), p_2(t), \ldots, p_N(t))$, y además se cumple que:

$$\bar{p}(t) = \bar{p}(t-1)\mathbf{P}.$$

Con esto hemos obtenido la probabilidad de selección de la caminata aleatoria, pero si además se cumplen ciertas condiciones, entonces $\bar{p}(t)$ tiende a una distribución estacionaria y se llama distribución estacionaria de la cadena de Markov. Por lo tanto, se define a continuación el término distribución estacionaria.

Definición 25 (Distribución estacionaria)**.** *Una distribución estacionaria o también*

llamada distribución de equilibrio de una cadena de Markov es una distribución de probabilidad $\bar{\pi}$ tal que:

$$\bar{\pi} = \bar{\pi}\mathbf{P}.$$

Entonces la idea es definir ciertas propiedades y usarlas para conseguir una distribución estacionaria de la cadena de Markov que modela a la caminata aleatoria. En [11] escrito por Mitzenmacher, Michael and Upfal, Eli enuncian el siguiente resultado.

Teorema 1. *Sea una cadena de Markov finita, irreducicle y ergódica, entonces la cadena tiene una única distribución estacionaria $\bar{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$.*

Es decir con este teorema podemos decir que una cadena de Markov finita, irreducicle y ergódica, a partir de cierta cantidad de pasos de la caminata aleatoria, $\bar{p}(t)$ no sufre cambios significativos.

Como la caminata aleatoria es un cadena de Markov, en donde el conjunto de estados es $V(G)$, así la cantidad de estados de la cadena de Markov es finita. Si suponemos que el grafo $G = (V, E)$ es conexo, entonces la caminata aleatoria es un proceso de Markov irreducible. Si el grafo G no es bipartito, entonces por el lema 7,12 de [11, p. 215] presentado por Mitzenmacher, Michael et al., que se presenta a continuación, se puede concluir que la cadena de Markov asociada a la caminata aleatoria es aperiódica.

Lema 1. *Una caminata aleatoria sobre un grafo G es aperiódica si y sólo si G no es un grafo bipartito.*

Hasta ahora hemos obtenido que la caminata aleatoria sobre G tiene finitos estados, hemos supuesto que G es conexo y por lo tanto la caminata aleatoria es una cadena de Markov irreducible, y por último que el grafo G no es bipartito. Entonces por el corolario 7,6 que aparece en el libro [11, p. 181] escrito por Mitzenmacher, Michael et al., que se muestra a continuación, se tiene que la caminata aleatorio es una cadena de Markov ergódica.

Corolario 1. *Si una cadena de Markov tiene una cantidad finita de estados, es irreducicle y es aperiodica, entonces la cadena de Markov es ergódica.*

Ahora, con el Teorema 1 se tiene que la caminata aleatoria puede converger el estado estacionario si se cumplen los supuestos. Con el teorema presentado en el libro [11, p. 190] de autores Mitzenmacher, Michael et al., se tiene que la distribución estacionaria de una caminata aleatoria es:

$$\pi_v = \frac{d(v)}{2|E|}.$$

En resumen, lo descrito anteriormente lo formalizaremos en una proposición.

Proposición 2. *Dado un grafo $G = (V, E)$ conexo y no bipartito en donde $V = \{v_1, ..., v_N\}$. Si sobre el grafo G se ejecuta una caminata aleatoria, la probabilidad de que la caminata aleatoria pase nuevamente por v converge a:*

$$\pi_v = \frac{d(v)}{2|E|} \qquad \forall v \in V(G)$$

en donde $d(v)$ es el grado de v y $|E|$ es la cantidad de aristas de G.

Esto quiere decir, que para cualquier valor $k \in \mathbb{N}$ tal que $k \geq \bar{t}$, la probabilidad de selección puede aproximarse como $p_k(v) \approx \frac{d(v)}{2|E|}$. Para simplificar los cálculos se supondrá que los grafos con que trabajaremos $G = (V, E)$ cumplen que $|V(G)|$ es finito, son grafos conexo y no son bipartitos. Además, en el artículo [16] se utiliza $\bar{t} = 2$, y si eliminamos el primer nodo de la secuencia la caminata aleatoria, entonces la probabilidad de selección muestreando a través de una caminata aleatoria en un grafo G se puede aproximar como $p(v) \approx \frac{d(v)}{2|E|}$ para $V(G) = \{v_1, ..., v_N\}$.

Debido a la proposición anterior se obtiene una aproximación de la probabilidad de selección del muestreo aleatorio asociado a una caminata aleatoria y se presenta a continuación.

Estimador 6. *Dado un grafo $G = (V, E)$ conexo y no bipartito en donde $V = \{v_1, ..., v_N\}$, $|V(G)|$ es finito. Además, si aproximamos el proceso de la caminata aleatoria a su estado estacionario desde el segundo salto y se elimina el primer nodo de la secuencia hecha por este, entonces para cualquier $i \in \{1, ..., N\}$ la probabilidad de selección del nodo v_i se puede aproximar como:*

$$p(v_i) \approx \frac{d(v_i)}{2|E|}.$$

Hasta este punto, hemos visto una forma de aproximar la probabilidad de selección si muestreamos con una caminata aleatoria. En el artículo [16] propuso que la probabilidad de selección del muestreo aleatorio asociado al proceso viral constante es proporcional al grado del nodo. Además, como el proceso viral constante es muy parecido al proceso

viral probabilístico, entonces nosotros proponemos realizar la misma suposición para el proceso viral probabilístico, es decir que la probabilidad de selección del muestreo aleatorio asociado a un proceso viral probabilístico es proporcional al grado del nodo.

Feld, Scott L en [4] mostró que para una población en donde sus individuos tenían relaciones simétricas, entonces se debía considerar lo siguiente:

a) Se debe reconocer la diferencia entre la distribución del número de amigos de las personas y la distribución del número de amigos de los amigos. La distribución de los amigos de los individuos es la distribución habitual del número de amigos que solemos examinar, pero la distribución de los amigos de los amigos incluye a algunos de los mismos individuos una y otra vez.

b) Cuando cada individuo se compara con el número medio de amigos de sus amigos, lo hace con una muestra del número de amigos de amigos, que es una distribución distinta del número de amigos entre individuos.

Para notar bien la diferencia entre la distribución del número de amigos de las personas y la distribución del número de amigos de los amigos se mostrará un ejemplo de población en donde las relaciones entre los individuos son simétricas, esta población ejemplo se ve en la Figura 3.3.4.

De esta forma, el conjunto de personas en la población y la cantidad de amigos de cada persona en la población mostrada en la Figura 3.3.4 es $\{v_1, v_2, v_3, v_4, v_5, v_6\}$ y $\{3, 5, 2, 4, 2, 2\}$ respectivamente. Por otro lado, el conjunto de los amigos de los amigos de la población y la de cantidad de amigos de los amigos en la población es $\{\underbrace{v_2, v_4, v_6}_{\mathcal{N}(v_1)}, \underbrace{v_1, v_3, v_4, v_5, v_6}_{\mathcal{N}(v_2)}, \underbrace{v_2, v_4}_{\mathcal{N}(v_3)}, \underbrace{v_1, v_2, v_3, v_5}_{\mathcal{N}(v_4)}, \underbrace{v_2, v_4}_{\mathcal{N}(v_5)}, \underbrace{v_1, v_2}_{\mathcal{N}(v_6)}\}$ y $\{\underbrace{5, 4, 2}_{\mathcal{N}(v_1)}, \underbrace{3, 2, 4, 2, 2}_{\mathcal{N}(v_2)}, \underbrace{5, 4}_{\mathcal{N}(v_3)}, \underbrace{3, 5, 2, 2}_{\mathcal{N}(v_4)}, \underbrace{5, 4}_{\mathcal{N}(v_5)}, \underbrace{3, 5}_{\mathcal{N}(v_6)}\}$ respectivamente.

Debido al punto *b*) podemos suponer que el muestreo aleatorio asociado a un proceso viral probabilístico y a un proceso viral constante es una muestra aleatorio uniforme con repetición realizada sobre la población $\mathcal{N}(V(G))$, por lo tanto la probabilidad de selección de cierto nodo con estos muestreos aleatorios es directamente proporcional al grado del nodo. Esto último es debido a que si un nodo tiene grado d y otro tiene grado uno, entonces el nodo de grado d estará repetido d veces en la población $\mathcal{N}(V(G))$, mientras que el de grado uno sólo estará una vez, por lo tanto la probabilidad de selección del nodo de grado d en una muestra aleatoria uniforme con repetición de la

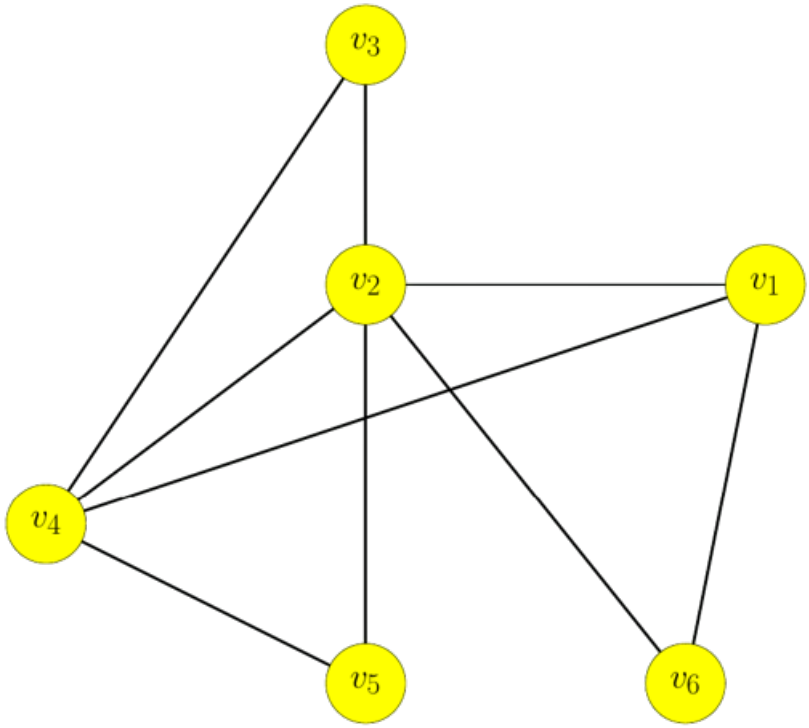


Figura 3.3.4: Ejemplo de una población mostrada a través de un grafo.

población $\mathcal{N}(G)$ es d veces la probabilidad de selección del nodo de grado uno. Por lo tanto, hemos justificado la suposición del artículo [16]. Es importante señalar que en los trabajos de Feld, Scott L [4], Newman, Mark EJ [12], y Salganik, Matthew J et al. [16], no se mencionan explícitamente las condiciones que los grafos deben cumplir para obtener las propiedades mencionadas. Sin embargo, se observa que la probabilidad de selección asociada a una caminata aleatoria se puede aproximar a la distribución estacionaria que es proporcional al grado del nodo, y el proceso viral constante se considera una generalización de la caminata aleatoria. Por lo tanto, asumiremos que los grafos son conexos, no bipartitos y finitos, ya que estas condiciones se aplican a los muestreos aleatorios mediante una caminata aleatoria.

En [12] se afirma que la distribución de grados de los nodos vecinos es proporcional a $d \cdot p(d)$, en donde $q(\cdot)$ es la distribución de grados de los nodos vecinos, $p(\cdot)$ es la distribución de grados de la nodos y d es un grado fijo particular arbitrario. Si lo anterior ocurre, podemos deducir que la la distribución de grados de los nodos vecinos es:

$$q(d) = \frac{d \cdot p(d)}{\sum_{d \in \mathcal{D}} d \cdot p(d)} \tag{3.3.1}$$

en donde $\sum_{d\in\mathcal{D}} d\cdot p(d)$ es el factor que normaliza la expresión de $q(\cdot)$ y $\mathcal{D}$ son los distintos grados que tienen los nodos de la población. El resultado anterior fue mejorado en [16], pues este escribió $p_A(\cdot)$ en función de $q_A(\cdot)$ como sigue:

$$p_A(d) = \frac{\frac{1}{d}\cdot q_A(d)}{\sum_{d\in\mathcal{D}} \frac{1}{d}\cdot q_A(d)} \qquad \forall d \in \mathcal{D}_A$$

en donde A es un subconjunto de nodos, $p_A(\cdot)$ representa la distribución de grados de los nodos en A y $q_A(\cdot)$ es la distribución de grados de los nodos vecinos en A. Si consideramos A como el conjunto que tiene todos los nodos del grafo, entonces se obtiene

$$p(d) = \frac{\frac{1}{d}\cdot q(d)}{\sum_{d\in\mathcal{D}} \frac{1}{d}\cdot q(d)} \qquad \forall d \in \mathcal{D}. \tag{3.3.2}$$

Además, se afirma en [16] que una aproximación de $q_A(\cdot)$ a través de una muestra $s-\{x_1\}$ hecha por una caminata aleatoria, proceso viral constante o proceso viral probabilístico como:

$$\widehat{q}_A(d) = \frac{\nu(A\cap s-\{x_1\},d)}{m_A} \qquad \forall\ d\in\mathcal{D}_{A\cap s-\{x_1\}}$$

en donde m_A es el cantidad de nodos del conjunto A en la muestra $s-\{x_1\}$, $\nu(A\cap s-\{x_1\},d)$ es la cantidad de nodos de grado d en el conjunto $A\cap s-\{x_1\}$ y $\mathcal{D}_{A\cap s-\{x_1\}}$ son los distintos grados de los nodos en el conjunto $A\cap s-\{x_1\}$. Además, se cumple que:

$$\widehat{p_A}(d) = \frac{\frac{1}{d}\cdot \widehat{q_A}(d)}{\sum_{d\in\mathcal{D}_A} \frac{1}{d}\cdot \widehat{q}_A(d)} \quad \forall d\in\mathcal{D}_{A\cap s-\{x_1\}}$$

en donde $\widehat{p_A}(\cdot)$ representa la estimación de la distribución de grados de la población.

Notar que si consideramos como A al conjunto de todos los nodos, entonces la distribución $\widehat{p}(\cdot)$ se puede considerar como la distribución de grados de una muestra aleatoria uniforme con repetición hecha sobre la selección de grados de los nodos de la población y que $\widehat{q}(\cdot)$ se puede considerar como la distribución de grados de una muestra aleatoria uniforme con repetición hecha sobre los grados de los vecinos de los nodos de la población.

Al tener el ajuste de la distribución de densidad de grados de una muestra aleatoria asociada a una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante, se puede realizar la siguiente estimación, hecha en [16].

Estimador 7. *Dado un grafo $G=(V,E)$ conexo y no bipartito en donde*

$V = \{v_1, ..., v_N\}$. Si tenemos la secuencia de nodos $s = \{x_1, ..., x_m\}$ hecha por una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante y un subconjunto $A \subset V$, considerando una muestra aleatoria $s - \{x_1\}$ cuya probabilidad de selección de cierto nodo es proporcional al grado del nodo, entonces se puede estimar el grado medio del subconjunto A denotado por D_A como:

$$D_A = \frac{m_A}{\sum_{x \in A \cap s - \{x_1\}} \frac{1}{d(x)}}$$

con m_A es la cantidad de nodos de A que hay en la muestra $s - \{x_1\}$.

Justificación: Dado un grafo $G = (V, E)$ conexo y no bipartito en donde $V = \{v_1, ..., v_N\}$. Sea un subconjunto $A \subseteq V(G)$, entonces tenemos que $p_A(\cdot)$ es la distribución de grados de los nodos del conjunto A y cumple que:

$$p_A(d) = \frac{\frac{1}{d} \cdot q_A(d)}{\sum_{d \in \mathcal{D}} \frac{1}{d} \cdot q_A(d)} \qquad \forall d \in \mathcal{D}_A$$

en donde $q_A(\cdot)$ es la distribución de grados de los nodos vecinos en A y $\mathcal{D}_A$ es el conjunto de los distintos grados posibles en A. Por definición de esperanza de una variable aleatoria discreta, podemos estimar el grado medio de del conjunto de nodos en A como:

$$\sum_{d \in \mathcal{D}_A} d \cdot p_A(d) = \sum_{d \in \mathcal{D}_A} d \, \frac{\frac{1}{d} \cdot q_A(d)}{\sum_{\hat{d} \in \mathcal{D}_A} \frac{1}{\hat{d}} \cdot q_A(\hat{d})} = \sum_{d \in \mathcal{D}_A} \frac{q_A(d)}{\sum_{\hat{d} \in \mathcal{D}_A} \frac{1}{\hat{d}} \cdot q_A(\hat{d})}.$$

Como $q_A(\cdot)$ es la distribución de los grados de los nodos vecinos de la población, entonces $\sum_{d \in \mathcal{D}_A} q_A(d) = 1$, de esta forma:

$$\sum_{d \in \mathcal{D}_A} d \cdot p_A(d) = \frac{1}{\sum_{d \in \mathcal{D}_A} \frac{1}{d} \, q_A(d)} \cdot \sum_{d \in \mathcal{D}_A} q_A(d) = \frac{1}{\sum_{d \in \mathcal{D}_A} \frac{1}{d} \, q_A(d)}.$$

Si tenemos la secuencia de nodos $s = \{x_1, ..., x_m\}$ hecha por una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante, considerando una muestra aleatoria $s - \{x_1\}$ cuya probabilidad de selección de cierto nodo es proporcional al grado del nodo y definiendo $r = A \cap s - \{x_1\}$, entonces se tiene la siguiente aproximación para $q_A(\cdot)$:

$$\widehat{q_A}(d) = \frac{\nu(r, d)}{m_A} \qquad \forall \; d \in \mathcal{D}_r$$

donde m_A es la cantidad de nodos de A que hay en la muestra $s - \{x_1\}$, $\nu(r, d)$ es la

cantidad de nodos de grado d en el conjunto r, y $\mathcal{D}_r$ son los distintos grados de los nodos en el conjunto r. Por lo tanto se obtiene que:

$$\begin{aligned}\sum_{d\in\mathcal{D}_A} d\cdot p_A(d) &\approx \frac{1}{\sum_{d\in\mathcal{D}_r}\frac{1}{d}\,\widehat{q_A}(d)}\\ &= \frac{1}{\sum_{d\in\mathcal{D}_r}\frac{1}{d}\,\frac{\nu(r,d)}{m_A}}\\ &= \frac{m_A}{\sum_{d\in\mathcal{D}_r}\frac{1}{d}\,\nu\left(r,d\right)}.\end{aligned}$$

Para cualquier $d\in\mathcal{D}$ se define $M_d\subseteq V$ como $M_d=\{x_i\in V : d(x_i)=d\}$, entonces $\nu\left(r,d\right)=\sum_{x\in r}\mathcal{X}_{M_d}(x)$ y se sigue que:

$$\begin{aligned}\sum_{d\in\mathcal{D}} d\cdot p_A(d) &\approx \frac{m_A}{\sum_{d\in\mathcal{D}_r}\frac{1}{d}\,\nu\left(r,d\right)}\\ &= \frac{m_A}{\sum_{d\in\mathcal{D}_r}\frac{1}{d}\,\sum_{x\in r}\mathcal{X}_{M_d}(x)}\\ &= \frac{m_A}{\sum_{x\in r}\sum_{d\in\mathcal{D}_r}\frac{1}{d}\,\mathcal{X}_{M_d}(x)}\\ &= \frac{m_A}{\sum_{x\in r}\frac{1}{d(x)}}.\end{aligned}$$

Así el grado medio de A se puede estimar como:

$$D_A=\frac{m_A}{\sum_{x\in r}\frac{1}{d(x)}}$$

con $r=A\cap s-\{x_1\}$ y m_A es la cantidad de nodos de A que hay en la muestra $s-\{x_1\}$.

□

Con este resultado se puede obtener la siguiente proposición.

Estimador 8. *Dado un grafo $G=(V,E)$ conexo y no bipartito en donde $V=\{v_1,...,v_N\}$. Si tenemos la secuencia de nodos $s=\{x_1,...,x_m\}$ hecha por una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante, considerando una muestra aleatoria $s-\{x_1\}$ cuya probabilidad de selección de cierto nodo es proporcional al grado del nodo, entonces se puede estimar el doble de la cantidad de aristas del grafo G a través de la expresión:*

$$2\,|E|\approx\frac{N}{\sum_{d\in\mathcal{D}_s}\frac{1}{d}\,\widehat{q}(d)}=N\cdot\frac{m-1}{\sum_{x\in s-\{x_1\}}\frac{1}{d(x)}}.$$

Justificación: Haciendo $A = V(G)$ y utilizar la proposición anterior para obtener el grado medio del grafo. Luego multiplicar el estimador del grado medio por el tamaño de $V(G)$. □

Debido a que la probabilidad de selección es proporcional al grado del nodo en los muestreos aleatorio asociados a la caminata aleatoria, proceso viral probabilístico y proceso viral constante, entonces se obtiene la siguiente proposición.

Estimador 9. *Dado un grafo $G = (V, E)$ conexo y no bipartito en donde $V = \{v_1, ..., v_N\}$. Si tenemos una secuencia de nodos $s = \{x_1, ..., x_m\}$ hecha por una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante, considerando una muestra aleatoria $s - \{x_1\}$ cuya probabilidad de selección cumple que para todo $i \in \{1, .., N\}$ se tiene que:*

$$cte = \frac{p(v_i)}{d(v_i)},$$

entonces para cualquier $i \in \{1, ..., N\}$ la probabilidad de selección para cualquiera de los métodos de muestreos antes mencionados es:

$$p(v_i) = \frac{d(v_i)}{2\ |E|}.$$

Justificación: Dado un grafo $G = (V, E)$ conexo y no bipartito en donde $V = \{v_1, ..., v_N\}$. Si tenemos la secuencia de nodos $s = \{x_1, ..., x_m\}$ hecha por una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante, considerando una muestra aleatoria $s - \{x_1\}$, sea un $i \in \{1, ..., N\}$ y debido a que la probabilidad de selección de cualquier nodo es proporcional al grado, entonces

$$c\ = \frac{p(v_i)}{d(v_i)}$$

en donde c es una constante, luego por propiedades de proporciones se tiene que

$$c\ = \frac{p(v_1) + p(v_2) + \ldots + p(v_N)}{d(v_1) + d(v_2) + \ldots + d(v_N)} = \frac{\sum_{v \in V} p(v)}{\sum_{v \in V} d(v)}$$

por definición de probabilidad de selección se cumple que $\sum_{v \in V} p(v) = 1$, así

$$c = \frac{1}{2\ |E|} = \frac{\sum_{v \in V} p(v)}{\sum_{v \in V} d(v)}.$$

Por lo tanto, la probabilidad de selección de un muestreo aleatorio a través de una caminata aleatoria, un proceso viral constante o un proceso viral probabilístico es:

$$p(v_i) = \frac{d(v_i)}{2\ |E|} \quad i \in \{1, .., N\} \text{ con } V(G) = \{v_1, ..., v_N\}\,.$$

□

Si definimos un experimento aleatorio que consiste en realizar un muestreo aleatorio asociado a una caminata aleatoria, un proceso viral probabilísstico o un proceso viral constante, entonces se puede obtener lo siguiente.

Proposición 3. *Sea un grafo $G = (V, E)$ en donde $V(G) = \{v_1, ..., v_N\}$ y si consideramos un experimento aleatorio que consiste en realizar un muestreo aleatorio asociado a una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante $s - \{x_1\} = \{x_2, ..., x_m\}$, entonces la fracción esperada de nodos de cierto grado d en la muestra es igual a la fracción de nodos de grado d en la población $\mathcal{N}(V(G))$.*

Demostración. Sea un grafo conexo y no bipartito $G = (V, E)$ con $V(G) = \{v_1, ..., v_N\}$ y si definimos un experimento aleatorio que consiste en realizar una muestra aleatoria asociada a una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante de nodos $s - \{x_1\} = \{x_2, ..., x_m\}$. Si suponemos que la probabilidad de selección de la muestra $s - \{x_1\}$ es proporcional al grado del nodo, entonces para cualquier nodo $v \in V(G)$ se tiene que la cantidad esperada de veces que aparezca el nodo v en la muestra $s - \{x_1\}$ es:

$$(m-1)\ \frac{d(v)}{2|E|} = (m-1)\ \frac{d}{2|E|}.$$

Si consideramos $\nu(V(G), d)$ como la cantidad de nodos de grado d en el grafo G, entonces la cantidad esperada de nodos de grado d en la muestra es:

$$\nu(V(G), d) \cdot (m-1) \cdot \frac{d}{2|E|}.$$

Por lo tanto, la fracción esperada de nodos de grado d en la muestra $s - \{x_1\}$ es:

$$\frac{\nu(V(G), d) \cdot (m-1) \cdot \frac{d}{2|E|}}{(m-1)} = \nu(V(G), d) \cdot \frac{d}{2|E|}.$$

Por otro lado, si consideramos la población $\mathcal{N}(V(G))$ como la concatenación de todas

las vecindades del grafo G, entonces la cantidad de nodos de grado d en $\mathcal{N}(V(G))$ es $d \cdot \nu(V(G), d)$. Además, el tamaño de $\mathcal{N}(V(G))$ es $2|E|$, por lo tanto, la fracción de nodos de grado d en $\mathcal{N}(V(G))$ es:

$$\frac{d \cdot \nu(V(G), d)}{2|E|}.$$

De esta forma, la fracción esperada de nodos de grado d de una muestra s es igual a la fracción de nodos de grado d de la población $\mathcal{N}(V(G))$. □

De forma análoga al caso en donde se muestrea de forma uniforme con repetición, se puede estimar el grado medio de un grupo de nodos de $V(G)$, a través de un muestreo aleatorio asociada a una caminata aleatoria, un proceso viral probabilístico o a un constante. En [16] se propone una forma, que es utilizando dos estimadores de Hansen-Hurwitz, uno para estimar la suma de los grados de los nodos del grupo , y el otro para estimar la cantidad de nodos del grupo. Con esto en mente definimos la siguiente proposición.

Estimador 10. *Dado un grafo conexo y no bipartito $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$. Si tenemos la secuencia de nodos $s = \{x_1, ..., x_m\}$ hecha por una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante y un subconjunto $A \subset V$, considerando una muestra aleatoria $s - \{x_1\}$ cuya probabilidad de selección de cierto nodo es proporcional al grado del nodo, entonces se puede estimar el grado medio del subconjunto A denotado por D_A como:*

$$D_A \approx \frac{\sum_{x \in s - \{x_1\}} \mathcal{X}_A(x)}{\sum_{x \in s - \{x_1\}} \frac{\mathcal{X}_A(x)}{d(x)}}.$$

Justificación: Dado un grafo conexo y no bipartito $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$. Considerando una muestra aleatoria $s - \{x_1\}$ cuya probabilidad de selección de cierto nodo es proporcional al grado del nodo, entonces para cada $i \in \{1, ..., N\}$ se define $R(v_i) = d(v_i)$ y obtenemos que la probabilidad de selección es $p(v_i) = \frac{d(v_i)}{2\ |E|}$.

Para obtener este estimador debemos utilizar dos estimadores de Hansen-Hurwitz, entonces para cada $i \in \{1, ..., N\}$ se define $Y(v_i) = \mathcal{X}_A(v_i)\ d(v_i)$ y $Z(v_i) = \mathcal{X}_A(v_i)$. Así, utilizando un estimador de Hansen-Hurwitz asociado a las variables $R(v_i) = d(v_i)$ y

$Y(v_i) = \mathcal{X}_A(v_i)d(v_i)$ para cualquier $i \in \{1, ..., N\}$, se obtiene:

$$\begin{aligned}\sum_{v\in V} \mathcal{X}_A(v)\ d(v) &\approx \frac{1}{m-1} \sum_{x\in s-\{x_1\}} \frac{\mathcal{X}_A(x)\ d(x)}{p(x)} \\ &= \frac{2\ |E|}{m-1} \sum_{x\in s-\{x_1\}} \frac{\mathcal{X}_A(x)\ d(x)}{d(x)} \\ &= \frac{2\ |E|}{m-1} \sum_{x\in s-\{x_1\}} \mathcal{X}_A(x).\end{aligned}$$

Para el otro estimador de Hansen-Hurwitz usamos las variables asociadas $R(v_i) = d(v_i)$ y $Z(v_i) = \mathcal{X}_A(v_i)$ para cualquier $i \in \{1, ..., N\}$, así

$$\begin{aligned}\sum_{v\in V} \mathcal{X}_A(v) &\approx \frac{1}{m-1} \sum_{x\in s-\{x_1\}} \frac{\mathcal{X}_A(x)}{p(x)} \\ &= \frac{2\ |E|}{m-1} \sum_{x\in s-\{x_1\}} \frac{\mathcal{X}_A(x)}{d(x)}.\end{aligned}$$

Por lo tanto, dividiendo ambos estimadores se obtiene que:

$$\begin{aligned}\frac{\sum_{v\in V(G)} \mathcal{X}_A(v)\ d(v)}{\sum_{v\in V(G)} \mathcal{X}_A(v)} &\approx \frac{\frac{2\ |E|}{m-1} \sum_{x\in s-\{x_1\}} \mathcal{X}_A(x)}{\frac{2\ |E|}{m-1} \sum_{x\in s-\{x_1\}} \frac{\mathcal{X}_A(x)}{d(x)}} \\ &= \frac{\sum_{x\in s-\{x_1\}} \mathcal{X}_A(x)}{\sum_{x\in s-\{x_1\}} \frac{\mathcal{X}_A(x)}{d(x)}}.\end{aligned}$$

Obteniendo el estimador del grado medio del subconjunto A de $V(G)$. □

Una forma de estimar el doble de la cantidad de aristas de un grafo G a través de un muestreo por caminata aleatoria, proceso viral constante o por el proceso viral probabilístico es usando un estimador de Hansen-Hurwitz para estimar la cantidad de nodos de G, y se muestra en la siguiente proposición.

Estimador 11. *Dado un grafo conexo y no bipartito $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$. Si tenemos la secuencia de nodos $s = \{x_1, ..., x_m\}$ hecha por una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante, considerando una muestra aleatoria $s - \{x_1\}$ cuya probabilidad de selección de cierto nodo es proporcional al grado del nodo, entonces se puede estimar el doble de la cantidad de aristas del grafo G a través de la expresión:*

$$2\ |E| \approx N\ \frac{m-1}{\sum_{x \in s-\{x_1\}} \frac{1}{d(x)}}.$$

Justificación: Dado un grafo conexo y no bipartito $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$, considerando una muestra aleatoria $s - \{x_1\}$ cuya probabilidad de selección de cierto nodo es proporcional al grado del nodo, entonces para cada $i \in \{1, ..., N\}$ se define $R(v_i) = d(v_i)$ y $Y(v_i) = 1$, y obtenemos que la probabilidad de selección es $p(v_i) = \frac{d(v_i)}{2\ |E|}$.

De esta forma, podemos usar el estimador de Hansen-Hurwitz y obtener:

$$|V| = \sum_{v \in V} 1 \approx \frac{1}{m-1} \sum_{x \in s-\{x_1\}} \frac{1}{p(x)} = \frac{1}{m-1} \sum_{x \in s-\{x_1\}} \frac{2\ |E|}{d(x)},$$

despejando $2\ |E|$

$$2\ |E| \approx N\ \frac{m-1}{\sum_{x \in s-\{x_1\}} \frac{1}{d(x)}}.$$

Otra forma es haciendo $A = V(G)$ y utilizar la proposición anterior para obtener el grado medio del grafo. Luego, multiplicar el estimador del grado medio por el tamaño de $V(G)$. □

Notar que las estimaciones de $2|E|$ pueden ser utilizadas para estimar la probabilidad de selección de las muestras aleatorias asociadas a una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante.

Hasta el momento tenemos una variedad de estimadores de las propiedades de un grafo dado, en la siguiente sección se mostrarán los estimadores de tamaño de un subconjunto de la población utilizando las propiedades y definiciones hasta ahora, que es el tema principal de esta tesis.

Capítulo 4

Métodos de estimadores de grupo

4.1. Modelamiento del problema de estimación del tamaño de una subpoblación

En esta sección se presentarán los modelos para estimar el tamaño de un grupo dentro de la población. La idea es que, conociendo el tamaño de la población involucrada y a través de suposiciones se pueda encontrar una estimación para el tamaño de un grupo en la población.

Consideraciones comunes de todos los estimadores.
Los estimadores que se mostrarán en esta sección sirven para resolver el Problema 2 definido anteriormente en la Sección 2.3. Por lo tanto, los métodos de estimación de esta sección son estimadores para determinar el tamaño del subconjunto de nodos H.

A continuación se define la notación a utilizar en esta sección. Dado un grafo $G = (V, E)$ desconocido en donde el tamaño de $V(G) = \{v_1, ..., v_N\}$ es conocido por definición del problema. Además, se tienen los subconjuntos $H, J \subseteq V(G)$ desconocidos tal que $V(G) = J \cup H$ y $H \cap J = \emptyset$, en donde el tamaño de H y J son N_H y N_J respectivamente.

Ahora, proseguiremos con la explicación de los métodos estimativos principales de este trabajo.

4.2. Estimador PIMLE

El primer estimador que se presentará se llama PIMLE y fue propuesto por Killworth, Peter D et al. en el articulo [7]. El modelo PIMLE esta bajo los siguientes supuestos:

a) Los nodos del grafo tienen las mismas posibilidades de ser vecino de algún nodo en H.

b) Los nodos del grafo tienen grados no aleatorio, pero desconocido.

c) Para cualquier nodo del grafo se pueden obtener los nodos vecinos y cada etiqueta de estos.

d) La selección de la muestra aleatoria es independiente de los nodos del grafo.

Utilizando algunas de las suposiciones anteriores se obtiene la siguiente proposición.

Estimador 12. *Sea un grafo $G = (V, E)$ en donde $V(G) = \{v_1, ..., v_N\}$ y sea un subconjunto $H \subseteq V(G)$. Si suponemos que todos los nodos tienen las mismas posibilidades de conectarse con cualquier nodo en H, entonces para un nodo aleatorio v se tiene la siguiente estimación:*

$$\frac{m_v}{d(v)} \approx \frac{N_H}{N}$$

en donde m_v son los nodos vecinos de v que están en H, $d(v)$ es el grado de v, N_H y N son los tamaños de los conjuntos H y $V(G)$ respectivamente.

Justificación: Dado un grafo $G = (V, E)$ en donde $V(G) = \{v_1, ..., v_N\}$ y un subconjunto $H \subseteq V(G)$. Si $v \in V(G)$ es un nodo aleatorio, entonces tenemos que $\mathcal{N}(v)$ es un conjunto de nodos aleatorios. Utilizando ambas suposiciones de la proposición se tiene que $\mathcal{N}(v)$ es una muestra uniforme sin repetición del conjunto $V(G)$, por lo tanto se obtiene la estimación:

$$\frac{m_v}{d(v)} \approx \frac{N_H}{N}$$

en donde $d(v)$ es el grado del nodo v, m_v la cantidad de aristas que comparte el nodo v con algún nodo de H. □

De la proposición, tenemos que la proporción de nodos de etiqueta positiva a nivel de la red personal de v es igual a la proporción de nodos de etiqueta positiva a nivel de la red.

Dada una muestra $s = \{x_1, .., x_m\}$ del conjunto de vértices V y utilizando la proposición

anterior, se puede obtener la siguiente estimación para $j = 1, ..., m$:

$$N_H \approx \hat{N}_{H_j} = \frac{m_{x_j}}{d(x_j)} \cdot N$$

en donde $d(x_j)$ es el grado del nodo x_j, m_{x_j} la cantidad de aristas que comparte el nodo x_j con algún nodo de H, N_H es la cantidad de nodos que tiene el conjunto H y N es la cantidad de nodos de V. Por lo tanto, para cada $j \in \{1, .., m\}$ tal que $x_j \in s$ se tiene un estimador de N_H distinto, luego podemos definir un estimador más general como:

$$N_H \approx \hat{N}_H^{PIMLE} = \frac{1}{|s|} \sum_{j=1}^{m} \hat{N}_{H_j} = \frac{N}{m} \cdot \sum_{x \in s} \frac{m_x}{d(x)}.$$

Debido a la información solicitada por el estimador PIMLE, este estimador utiliza encuestas de tipo indirecta exclusiva.

4.3. Estimador MLE

Nuestro segundo método de estimación, llamado MLE, es presentado por Killworth, Peter D; McCarty, Christopher; Bernard, H Russell; Shelley, Gene Ann; Johnsen, Eugene C en el articulo [8].

Este modelo tiene los mismos supuestos que el modelo PIMLE y agrega que N_H es muy pequeño en comparación con N. De la proposición de la sección anterior se tiene que si elegimos un nodo $v \in V(G)$ de forma aleatoria, entonces $\mathcal{N}(v)$ es una muestra uniforme sin repetición de tamaño $d(v)$ conocido. Por lo tanto, se tiene la siguiente proposición.

Proposición 4. *Sea un grafo $G = (V, E)$ en donde $V(G) = \{v_1, ..., v_N\}$ y sea un subconjunto $H \subseteq V(G)$. Si suponemos que todos los nodos tienen las mismas posibilidades de conectarse con cualquier nodo en H y los grados de los nodos en $V(G)$ son desconocidos y deterministicos, entonces para un nodo aleatorio $v \in V(G)$, la cantidad de conocidos de v en H, es decir m_v, distribuye de forma hipergeométrica con parámetros $(d(v), N_H, N)$.*

Demostración. Dado un grafo $G = (V, E)$ en donde $V(G) = \{v_1, ..., v_N\}$ y un subconjunto $H \subseteq V(G)$. Si $v \in V(G)$ es un nodo aleatorio, entonces tenemos que $\mathcal{N}(v)$ es un conjunto de nodos aleatorios. Utilizando ambas suposiciones de esta proposición se tiene que $\mathcal{N}(v)$ es una muestra uniforme sin repetición del conjunto $V(G)$.

Como tenemos una población $V(G)$ de tamaño N en donde existen dos tipos de nodos, los que pertenecen a H y los que no, en donde el tamaño de H es N_H desconocido. Además, si definimos $s = \mathcal{N}(v)$ donde s es una muestra uniforme sin repetición de tamaño $d(v)$, entonces utilizando la definición de una variable aleatoria de tipo hipergeométrica 19 tenemos que la cantidad de nodos de s que pertenecen a H, es decir m_v distribuye de forma hipergeométrica con parámetros $(d(v), N_H, N)$. □

En el articulo [8] se supone que si N_H es muy pequeño en comparación con N, entonces la distribución de m_v se puede aproximar a una variable de tipo binomial con parámetros $(d(v), \frac{N_H}{N})$, es decir aproximaron una distribución hipergeométrica con parámetros $(d(v), N_H, N)$ a través de una distribución binomial con parámetros $(\hat{m}, \hat{p}) = (d(v), \frac{N_H}{N})$. En el trabajo [17] hecho por Sandiford, Peter J, se utiliza esta aproximación como punto de referencia, dando resultados aceptables en el caso antes mencionado.

Para un nodo aleatorio $v \in V(G)$ y utilizando la función de densidad de la distribución Binomial de parámetros $(\hat{m}, \hat{p}) = (d(v), \frac{N_H}{N})$, se calcula la probabilidad de que v tenga m_v vecinos en el grupo H con $d(v)$ desconocido pero no aleatorio:

$$Pr(\ v \text{ tenga } m_v \text{ vecinos en } H\ ) = \binom{d(v)}{m_v} \hat{p}^{m_v} (1-\hat{p})^{d(v)-m_v}.$$

Considerando que la selección de la muestra aleatoria es independiente de los nodos del grafo y si denotamos una muestra aleatoria como $s = \{x_1, ..., x_m\}$, entonces la probabilidad de que para cualquier $x \in s$ tenga m_x vecinos en H es:

$$Pr(\ \forall x \in s, \text{se cumple } m_x\ ) = \prod_{x \in s} \binom{d(x)}{m_x} \hat{p}^{m_x} (1-\hat{p})^{d(x)-m_x}. \tag{4.3.1}$$

Recordamos que $\hat{p} = \frac{N_H}{N}$, con N_H desconocido, por lo cual $\hat{p}$ es desconocido. Utilizando el método de máxima verosimilitud, que tiene como idea maximizar la probabilidad antes mencionada y notamos que maximizar la Ecuación 4.3.1 es lo equivalente a maximizar:

$$Pr(\ \forall x \in s, \text{se cumple } m_x\ ) = \prod_{x \in s} \hat{p}^{m_x} (1-\hat{p})^{d(x)-m_x}$$

$$Pr(\ \forall x \in s, \text{se cumple } m_x\ ) = \hat{p}^{\sum_{x \in s} m_x} (1-\hat{p})^{\sum_{x \in s} (d(x)-m_x)}. \tag{4.3.2}$$

Así, el máximo de una función de tipo $f(x) = x^m(1-x)^{n-m}$ con $x \in [0,1]$, $n, m \in \mathbb{N}$ y $n > m$, ocurre cuando $x = \frac{m}{n}$, así la Ecuación 4.3.2 se maximiza cuando $\hat{p} = \frac{\sum_{x \in s} m_x}{\sum_{x \in s} d(x)}$, y como $\hat{p} = \frac{N_H}{N}$, entonces se puede estimar el valor de N_H, con el estimador $\hat{N}_H^{MLE}$:

$$\hat{N}_H^{MLE} = N \cdot \frac{\sum_{x \in s} m_x}{\sum_{x \in s} d(x)}.$$

Debido a la información solicitada por el estimador MLE, este estimador utiliza encuestas de tipo indirecta exclusiva.

4.4. Estimador GNSUM

4.4.1. Teoría del estimador GNSUM

El Estimador GNSUM fue presentado por Feehan, Dennis M y Salganik, Matthew J en [3]. Este estimador se basa principalmente en una propiedad de digrafos, esta propiedad permite establecer una relación, y a través de esta relación se estima el tamaño del grupo deseado utilizando estimadores de Hansen-Hurwitz.

Para poder definir la propiedad antes mencionada se debe realizar una definición.

Definición 26. *Dado un digrafo $D = (V, F)$ con $V = \{v_1, ..., v_N\}$ y un subconjunto $B \subseteq V(G)$. Entonces, para cada $v \in V(D)$ se define $\mathcal{Y}_{v,B}$ como la cantidad de nodos de B que pertenece a la vecindad saliente de v en el digrafo D, por otro lado, $\mathcal{V}_{v,B}$ es la cantidad de vecinos entrantes de v que pertenecen a B. Por lo tanto:*

$$\mathcal{Y}_{v,B} = |N^+(v) \cap B| \qquad \mathcal{V}_{v,B} = |N^-(v) \cap B|$$

con $N^-(v)$ y $N^+(v)$ son las vecindades entrante y salientes de v respectivamente.

Además, si $A, B \subseteq V(D)$, entonces se define $\mathcal{Y}_{A,B}$ y $\mathcal{V}_{A,B}$ como:

$$\mathcal{Y}_{A,B} = \sum_{v \in A} y_{v,B} \qquad \mathcal{V}_{A,B} = \sum_{v \in A} v_{v,B}.$$

Esta definición está asociada al siguiente teorema.

Teorema 2. *Sea un digrafo $D = (V, F)$ y un subconjunto $H \subseteq V(G)$. Además el digrafo D cumple que para cualquier nodo $u \in H^c$ se tiene que $\mathcal{V}_{u,V} = 0$, entonces se cumple la*

siguiente relación:

$$\mathcal{Y}_{V,H} = \mathcal{V}_{H,V}.$$

Demostración. Dado un digrafo $D = (V, F)$ y un subconjunto $H \subseteq V(G)$. Si $|F|$ es la cantidad de aristas del digrafo D, entonces por un lado se tiene que las aristas entrantes de un digrafo cumplen:

$$\sum_{v \in H} \mathcal{V}_{v,V} + \sum_{v \in H^C} \mathcal{V}_{v,V} = \sum_{v \in V} \mathcal{V}_{v,V} = |F|.$$

Luego, usando la hipótesis que para cualquier nodo $v \in H^c$ se tiene que $\mathcal{V}_{v,V} = 0$

$$\sum_{v \in H} \mathcal{V}_{v,V} = |F|.$$

Por otro lado, para cualquier $v \in H^C$ se cumple que $\mathcal{V}_{v,V} = 0$, entonces $\mathcal{Y}_{u,v} = 0$ para cualquier $u \in V$. Una forma más clara de ver esto, es pensar que las aristas que apuntan a algún nodo $v \in H^C$ no existen, entonces cualquier arista que salga de algún nodo, esta arista no puede llegar a algún nodo $v \in H^C$. Por lo tanto, se tiene que:

$$|F| = \sum_{v \in V} \mathcal{Y}_{v,V} = \sum_{v \in V} \mathcal{Y}_{v,H} + \sum_{v \in V} \mathcal{Y}_{v,H^C} = \sum_{v \in V} \mathcal{Y}_{v,H}.$$

Luego,

$$\mathcal{V}_{H,V} = \sum_{v \in H} \mathcal{V}_{v,V} = |F| = \sum_{v \in V} \mathcal{Y}_{v,H} = \mathcal{Y}_{V,H}.$$

Obteniendo que

$$\mathcal{V}_{H,V} = \mathcal{Y}_{V,H}.$$

□

En palabras simples este teorema nos dice que la suma de las aristas que salen de algún nodo y llegan a H es igual a la suma de las aristas que llegan a H, véase la Imagen 4.4.1 para comprender mejor esta idea. La Figura 4.4.1 muestra que al nodo v_5 tiene tres aristas dirigidas que llegan a él y por lo tanto en el digrafo bipartito asociado también tiene tres aristas dirigidas que llegan a él, análogo para v_7 sólo que a que este nodo son cuatro aristas dirigidas. Ahora, a partir del teorema anterior podemos obtener la

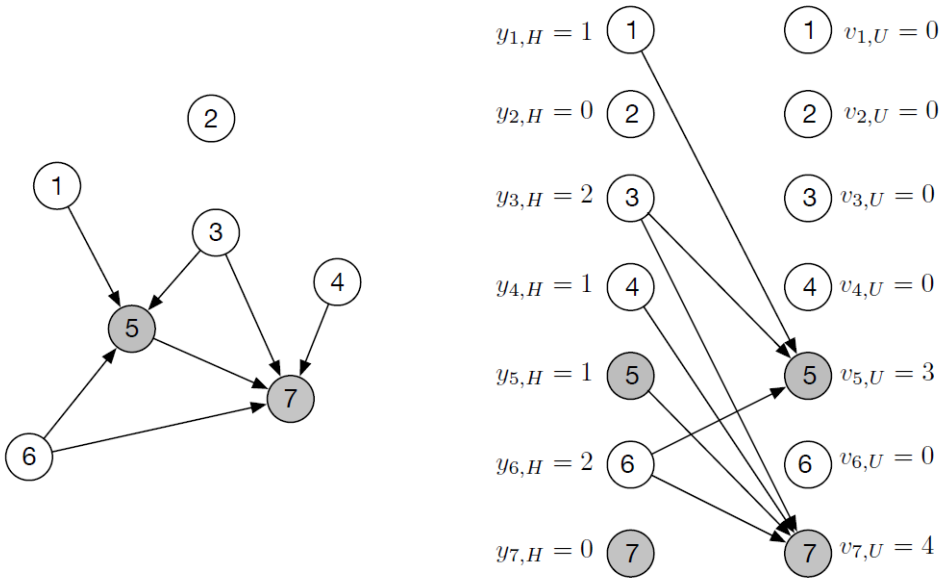


Figura 4.4.1: Imagen que refleja la idea del Teorema 2.

relación:

$$N_H = \frac{\mathcal{Y}_{V,H}}{\frac{\mathcal{V}_{H,V}}{N_H}} = \frac{\mathcal{Y}_{V,H}}{\bar{\mathcal{V}}_{H,V}}$$

en donde $\bar{\mathcal{V}}_{H,V} = \frac{\mathcal{V}_{H,V}}{N_H}$.

Lo anterior se puede utilizar para resolver el problema principal de esta tesis, entonces definimos lo siguiente.

Definición 27 (Reportar)**.** *Dada una población V y un grupo H en V, en donde los individuos de la población están relacionados de alguna forma. Para cualquier par de personas u, v que estén relacionadas, entonces u reporta a v si u sabe que v está en H.*

Con la definición anterior, podemos explicar como modelar el problema principal a través del método GNSUM.

Proposición 5. *Dada una población V y un grupo H en V, en donde los individuos de la población están relacionados de alguna forma. Se define el digrafo $D = (V, F)$, en donde V indexa a cada integrante de la población y F se define como:*

$$F = \{(u, v) \in V \times V : u \text{ reporta a } v\}.$$

Entonces, utilizando el Teorema 2 podemos determinar el tamaño de H como:

$$N_H = \frac{\mathcal{Y}_{V,H}}{\frac{\mathcal{V}_{H,V}}{N_H}} = \frac{\mathcal{Y}_{V,H}}{\bar{\mathcal{V}}_{H,V}}.$$

El método de estimación GNSUM sólo supone que no existan falsos positivos, es decir, que no exista algún vecino de un nodo muestreado que sea considerado en H cuando en realidad no está en H.

En la Proposición 5, se puede ver que el digrafo D cumple que si $u, v \in H$, entonces pueden suceder 3 cosas que se enumeran a continuación:

- Las aristas (u, v) y (v, u) están en $F(D)$.
- Sólo una aristas de las (u, v) y (v, u) está en $F(D)$.
- Ninguna de las aristas (u, v) y (v, u) están en $F(D)$.

Por lo tanto, la principal diferencia de modelar el problema con un grafo y con un digrafo es que el grafo puede sólo indexar dos casos.

En nuestros experimentos sólo se consideran relaciones simétricas pues los otros modelos no permiten relaciones asimétricas, por lo tanto se incluirá esta condición en el modelo y veremos como cambia. Por lo tanto, se define formalmente que la población tenga sólo relaciones simétricas.

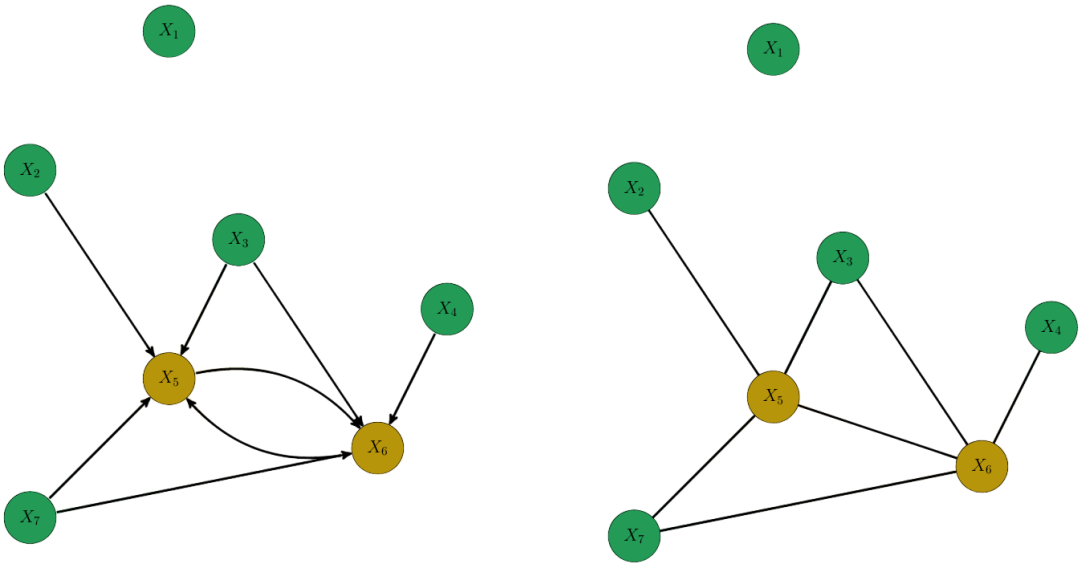
Definición 28. *Un población tiene relaciones simétricas cuando para cualquier par de integrantes a, b de la población cumplen las dos siguientes proposiciones:*

- *a se relaciona con b si y sólo b se relaciona con a.*
- *a conoce el estado de b si y sólo b conoce el estado de a.*

Si la población en cuestión cumple la definición anterior, entonces tenemos que se cumple lo siguiente.

Proposición 6. *Dada una población V con relaciones simétricas y un grupo de la población H, entonces se cumple que para cualquier par de personas a, b de la población sólo se cumple una de las siguientes situaciones:*

- *Si a y b no pertenecen a H, entonces a no reporta b y b no reporta a.*



(a) Digrafo **(b)** Grafo

Figura 4.4.2: Figura que muestra modelado GNSUM con relaciones asimétricas y simétricas, en donde $H = \{X_5, X_6\}$.

- *Si a no pertenece a H, b pertenece a H, y a y b se conocen, entonces a reporta b y b no reporta a.*
- *Si a no pertenece a H, b pertenece a H, y a y b no se conocen, entonces a no reporta b y b no reporta a.*
- *Si a y b pertenecen a H, y a y b se conocen, entonces a reporta b y b reporta a.*
- *Si a y b no pertenecen a H, y a y b no se conocen, entonces a no reporta b y b no reporta u.*

Con lo anterior podemos obtener lo siguiente.

Proposición 7. *Dado un digrafo $D = (V, F)$ en donde $V = \{v_1, ..., v_N\}$ y $F = \{(u, v) \in V \times V : u \text{ reporta a } v\}$. Si el digrafo D representa una población con relaciones simétricas y sean $u, v \in H$, entonces el digrafo D cumple que si $(u, v) \in F(D)$ si y sólo $(v, u) \in F(D)$.*

De esta forma, la Figura 4.4.2 muestra un ejemplo en donde se tiene un digrafo en la Figura 4.4.2a y el grafo subyacente al él en la Figura 4.4.2b. Sin la suposición de relaciones simétricas puede darse el caso en que sólo la arista (x_5, x_6) esté en el digrafo de la Figura 4.4.2a y no (x_6, x_5), pero con la suposición existen las dos aristas o no existe ninguna.

De esta forma, podemos obtener lo siguiente.

Proposición 8. *Dada una población con relaciones simétricas y el digrafo $D = (V, F)$ que modela los reportes de la población, entonces el digrafo D que resuelve el problema está ligado a único grafo subyacente G y viceversa.*

A partir de lo anterior, se tiene la siguiente proposición.

Proposición 9. *Dada una población con relaciones simétricas y el digrafo $D = (V, F)$ que modela los reportes de la población. Si el grafo subyacente del digrafo D lo denotamos como $G = (V, E)$, entonces podemos estimar el tamaño de H a través de la siguiente expresión:*

$$N_H = \frac{\sum_{v \in V} |\mathcal{N}(v) \cap H|}{\frac{1}{N_H} \cdot \sum_{v \in H} d(v)}.$$

Demostración. Sea el digrafo $D = (V, F)$ que modela los reportes de la población, entonces $D = (V, F)$ cumple que $V = \{v_1, ..., v_N\}$ indexa a cada integrante de la población y $F = \{(u, v) \in V \times V : u \text{ reporta a } v\}$. Utilizando la Proposición 8 se tiene que el digrafo D está asociado a un único grafo subyacente $G = (V, E)$ y este grafo G está ligado únicamente a este digrafo D. Luego, el grafo G cumple que

$$\begin{aligned}
\sum_{v \in V} |\mathcal{N}(v) \cap H| &= \sum_{v \in V} \sum_{u \in \mathcal{N}(v)} \mathcal{X}_H(u) \\
&= \sum_{v \in V} \sum_{u \in V} \mathcal{X}_H(u) \mathcal{X}_{\mathcal{N}(v)}(u) \\
&= \sum_{u \in V} \mathcal{X}_H(u) \sum_{v \in V} \mathcal{X}_{\mathcal{N}(v)}(u) \\
&= \sum_{u \in V} \mathcal{X}_H(u) d(u) \\
&= \sum_{u \in H} d(u)
\end{aligned}$$

es decir

$$\sum_{v \in V} |\mathcal{N}(v) \cap H| = \sum_{v \in H} d(v).$$

Así obtenemos que

$$N_H = \frac{\sum_{v \in V} |\mathcal{N}(v) \cap H|}{\frac{1}{N_H} \sum_{v \in H} d(v)}.$$

□

Notar que si trabajamos con el grafo subyacente del digrafo que modela los reportes de la población, entonces debemos encontrar los valores requeridos de la proposición anterior. De esta forma se presenta la siguiente proposición:

Proposición 10. *Dada una población con relaciones simétricas y el digrafo $D = (V, F)$ que modela los reportes de la población. Si el grafo subyacente del digrafo D lo denotamos como $G = (V, E)$, entonces podemos obtener las siguientes relaciones:*

- *Para cualquier $v \in H$ se tiene que $d^-(v) = d(v)$.*
- *Para cualquier $v \in V$ se tiene que $\mathcal{N}^+(v) \cap H = \mathcal{N}(v) \cap H$.*

Notar que la información que puede ser recopilada por las encuestas es la del digrafo D y no la del grafo G.

Demostración. Sea el digrafo $D = (V, F)$ que modela los reportes de la población, entonces $D = (V, F)$ cumple que $V = \{v_1, ..., v_N\}$ indexa a cada integrante de la población y $F = \{(u, v) \in V \times V : u \text{ reporta a } v\}$. Utilizando la Proposición 8 se tiene que el digrafo D está asociado a un único grafo subyacente $G = (V, E)$ y este grafo G está ligado únicamente a este digrafo D.

Se tiene que para cualquier $u \in H$ se cumple que $\mathcal{N}^-(u) = \mathcal{N}(u)$, en donde $\mathcal{N}^-(u)$ corresponde a la vecindad entrante del digrafo D y $\mathcal{N}(u)$ corresponde a la vecindad del grafo G, por lo tanto $d^-(u) = d(u)$.

Por otro lado, para cualquier $v \in H^C$ se tiene que $\mathcal{N}^+(v) = \mathcal{N}(v)$, en donde $\mathcal{N}^+(v)$ son las aristas salientes del nodo v en el digrafo D y $\mathcal{N}(v)$ es la vecindad de v en el grafo G, así $\mathcal{N}^+(v) \cap H = \mathcal{N}(v) \cap H$.

Si $v \in H$, suponiendo que las relaciones son simétricas y por definición del digrafo D se tiene que $\mathcal{N}^+(v) \subseteq H$, entonces $\mathcal{N}^+(v) \cap H = \mathcal{N}^-(v) \cap H$. Anteriormente, dijimos que si $v \in H$, entonces $\mathcal{N}^-(v) = \mathcal{N}(v)$, por lo tanto $\mathcal{N}^+(v) \cap H = \mathcal{N}^-(v) \cap H = \mathcal{N}(v) \cap H$.

□

A partir del Teorema 9 podemos estimar el tamaño de H utilizando el grafo subyacente del digrafo $D = (V, F)$ denotado por $G = (V, E)$ a través de la siguiente ecuación:

$$N_H = \frac{\sum_{v \in V} |\mathcal{N}(v) \cap H|}{\frac{1}{N_H} \cdot \sum_{v \in H} d(v)}.$$

En la siguiente sección veremos como realizar las estimaciones de $\sum_{v\in V}|\mathcal{N}(v)\cap H|$ y $\frac{1}{N_H}\cdot\sum_{v\in H}d(v)$.

4.4.2. Estimador de $\mathcal{Y}_{V,H}$

Ya que trabajaremos con poblaciones en donde las relaciones son simétricas y como $\mathcal{Y}_{V,H}$ está asociado al digrafo $D=(V,F)$ que modela los reportes de la población, entonces se puede obtener el grafo $G=(V,E)$ asociado al digrafo D y se cumple que $\mathcal{Y}_{V,H}=\sum_{v\in V}|\mathcal{N}(v)\cap H|$.

Para estimar $\mathcal{Y}_{V,H}$ usaremos el estimador de Hansen-Hurwitz.

$$\mathcal{Y}_{V,H}=\sum_{v\in V}\mathcal{Y}_{v,H}=\sum_{v\in V}|\mathcal{N}(v)\cap H|\approx\frac{1}{m}\sum_{x\in s}\frac{|\mathcal{N}(x)\cap H|}{p(x)}$$

en donde $\frac{1}{m}\sum_{x\in s}\frac{|\mathcal{N}(x)\cap H|}{p(x)}$ es el estimador de Hansen-Hurwitz para $\mathcal{Y}_{V,H}$, con $s=\{x_1,...,x_m\}$ una muestra con reemplazo, m es el tamaño de la muestra y $p(x)$ es la probabilidad de selección de x en la muestra. Así para los siguientes tipos de muestreos queda como:

- Caso: Muestreo uniforme con reemplazo.

$$\mathcal{Y}_{V,H}=\sum_{v\in V}|\mathcal{N}(v)\cap H|\approx\frac{N}{m}\sum_{x\in s}|\mathcal{N}(x)\cap H|.$$

- Caso: Muestreo por caminata aleatoria, proceso viral probabilístico y constante. Si $s=\{x_1,..,x_m\}$ es la secuencia de nodos obtenida por caminata aleatoria, proceso viral probabilístico o constante, entonces la muestra asociada a la caminata aleatoria, proceso viral probabilístico ó constante es $s-\{x_1\}$ y por lo tanto el estimador queda como:

$$\mathcal{Y}_{V,H}=\sum_{v\in V}\mathcal{Y}_{v,H}\approx\frac{2|E|}{m-1}\sum_{x\in s-\{x_1\}}\frac{|\mathcal{N}(x)\cap H|}{d(x)}.$$

4.4.3. Estimador de $\bar{\mathcal{V}}_{H,V}$

Ya que trabajaremos con poblaciones en donde las relaciones son simétricas y como $\bar{\mathcal{V}}_{H,V}$ está asociado al digrafo $D=(V,F)$ que modela los reportes de la población,

entonces se puede obtener el grafo $G = (V, E)$ asociado al digrafo D y se cumple que $\bar{\mathcal{V}}_{H,V} = \frac{\sum_{v \in H} d(v)}{N_H}$.

Dada una muestra, para estimar $\bar{\mathcal{V}}_{H,V}$ tomamos la idea utilizada por Volz, Erik et al. en [18, p. 86], que es utilizar dos estimadores de Hansen-Hurwitz como sigue:

$$\mathcal{V}_{H,V} = \sum_{v \in V} \mathcal{X}_H(v) d(v) = \sum_{v \in V} f(v) \approx \frac{1}{m} \sum_{x \in s} \frac{f(x)}{p(x)} = \frac{1}{m} \sum_{x \in s} \frac{\mathcal{X}_H(x) d(x)}{p(x)}$$

$$N_H = \sum_{v \in V} \mathcal{X}_H(v) = \sum_{v \in V} g(v) \approx \frac{1}{m} \sum_{x \in s} \frac{g(x)}{p(x)} = \frac{1}{m} \sum_{x \in s} \frac{\mathcal{X}_H(x)}{p(x)}$$

en donde para cualquier $v \in V$ las funciones utilizadas se definen como $f(v) = \mathcal{X}_H(v) d(v)$ y $g(v) = \mathcal{X}_H(v)$. Luego, dividiendo ambas estimaciones podemos encontrar la nueva estimación deseada:

$$\bar{\mathcal{V}}_{H,V} \approx \frac{\frac{1}{m} \sum_{x \in s} \frac{\mathcal{X}_H(x) d(x)}{p(x)}}{\frac{1}{m} \sum_{x \in s} \frac{\mathcal{X}_H(x)}{p(x)}} = \frac{\sum_{x \in s} \frac{\mathcal{X}_H(x) d(x)}{p(x)}}{\sum_{x \in s} \frac{\mathcal{X}_H(x)}{p(x)}}.$$

La probabilidad de selección de cualquier nodo dependen del tipo de muestreo y estos deben de ser con reemplazo pues el estimador Hansen-Hurwitz esta definido para ellas.

- Caso: Muestreo uniforme con reemplazo.

  $$\bar{\mathcal{V}}_{H,V} \approx \frac{\sum_{x \in s} \frac{\mathcal{X}_H(x) d(x)}{p(x)}}{\sum_{x \in s} \frac{\mathcal{X}_H(x)}{p(x)}} = \frac{\sum_{x \in s} \mathcal{X}_H(x) d(x)}{\sum_{x \in s} \mathcal{X}_H(x)}$$

  notar que en este caso $p(v) = 1/N$ para cualquier $v \in V(G)$.

- Caso: Muestreo por caminata aleatoria, proceso viral probabilístico y constante. Si $s = \{x_1, .., x_m\}$ es la secuencia de nodos obtenida por caminata aleatoria, proceso viral probabilístico o constante, entonces la muestra asociada a la caminata aleatoria, proceso viral probabilístico o constante es $s - \{x_1\}$ y por lo tanto el estimador queda como:

  $$\bar{\mathcal{V}}_{H,V} \approx \frac{\sum_{x \in s} \frac{\mathcal{X}_H(x) d(x)}{p(x)}}{\sum_{x \in s} \frac{\mathcal{X}_H(x)}{p(x)}} = \frac{\sum_{x \in s} \mathcal{X}_H(x)}{\sum_{x \in s} \frac{\mathcal{X}_H(x)}{d(x)}}.$$

En resumen se tiene que el estimador GNSUM se basa en:

$$N_H = \frac{\mathcal{Y}_{V,H}}{\bar{\mathcal{V}}_{H,V}}.$$

En donde $\mathcal{Y}_{V,H}$ se puede estimar como:

$$\mathcal{Y}_{V,H} \approx \begin{cases} \frac{N}{M} \sum_{x \in s} |\mathcal{N}(x) \cap H| & \text{Muestreo uniforme con reemplazo.} \\ \frac{2|E|}{M-1} \sum_{x \in s - \{x_1\}} \frac{|\mathcal{N}(x) \cap H|}{d(x)} & \text{Muestreo por caminata aleatoria, proceso viral constante o proceso viral probabilístico.} \end{cases}$$

Por último, $\bar{\mathcal{V}}_{H,U}$ se puede estimar como:

$$\bar{\mathcal{V}}_{H,V} \approx \begin{cases} \frac{\sum_{x \in s} \mathcal{X}_H(x) d(x)}{\sum_{x \in s} \mathcal{X}_H(x)} & \text{Muestreo uniforme con reemplazo.} \\ \frac{\sum_{x \in s} \mathcal{X}_H(x)}{\sum_{x \in s} \frac{\mathcal{X}_H(x)}{d(x)}} & \text{Muestreo por caminata aleatoria, proceso viral constante o proceso viral probabilístico.} \end{cases}$$

Debido a la información solicitada por el estimador GNSUM, este estimador utiliza encuestas de tipo indirecta.

4.5. Estimador RDS I

4.5.1. Teoría del estimador RDS I

Este modelo se presentado por Salganik, Matthew J; Heckathorn, Douglas D en el articulo [16], el cual utiliza propiedades de las vecindades de los nodos de la muestra para estimar propiedades generales de la red.

Para poder ilustrar como funciona el estimador, primero debemos hacer algunas definiciones.

Definición 29. *Sea el digrafo $D = (V, F)$ y un subconjunto A de V de tamaño N_A, entonces se define el grado medio saliente de A denotado por D_A^+ como:*

$$D_A^+ = \frac{\sum_{v \in A} d^+(v)}{N_A}.$$

De aquí obtenemos la siguiente definición.

Definición 30. *Sea el digrafo $D = (V, F)$ en donde $V = \{v_1, ..., v_N\}$. Sean los subconjuntos A, B de V, entonces se define la probabilidad de que una arista con un nodo inicial en el subconjunto A se conecte con un nodo del subconjunto B en el digrafo D como:*

$$\tilde{C}_{A,B} = \frac{\mathcal{Y}_{A,B}}{\mathcal{Y}_{A,V}}.$$

Ahora, podemos enunciar la siguiente proposición.

Proposición 11. *Sea un digrafo $D = (V, F)$ en donde $V = \{v_1, ..., v_N\}$, D representa una población con relaciones simétricas y $F = \{(u, v) \in V \times V : u \text{ conoce a } v\}$. Sean los subconjuntos H, J de V de tamaño N_H y N_J respectivamente, entonces obtenemos lo siguiente:*

$$N_H = N \ \frac{D_J^+ \cdot \tilde{C}_{J,H}}{D_H^+ \cdot \tilde{C}_{H,J} + D_J^+ \cdot \tilde{C}_{J,H}}$$

y

$$N_J = N \ \frac{D_H^+ \cdot \tilde{C}_{H,J}}{D_H^+ \cdot \tilde{C}_{H,J} + D_J^+ \cdot \tilde{C}_{J,H}}.$$

Demostración. Dado un digrafo $D = (V, F)$ y los subconjuntos H, J de $V(D)$, considerando que D representa una población con relaciones simétricas y como $\mathcal{Y}_{H,J}$ es la cantidad de aristas del digrafo D que tienen un nodo de inicio en H y un nodo de llegada en J, entonces se puede obtener la relación $\mathcal{Y}_{H,J} = \mathcal{Y}_{J,H}$.

Por otro lado, despejando $\mathcal{Y}_{H,J}$ de la relación $\tilde{C}_{H,J} = \frac{\mathcal{Y}_{H,J}}{\mathcal{Y}_{H,V}}$ se obtiene:

$$\mathcal{Y}_{H,J} = \tilde{C}_{H,J} \mathcal{Y}_{H,V}.$$

Realizando lo mismo para $\mathcal{Y}_{J,H}$ y utilizando la primera relación:

$$\mathcal{Y}_{H,V} \cdot \tilde{C}_{H,J} = \mathcal{Y}_{H,J} = \mathcal{Y}_{J,H} = \mathcal{Y}_{J,V} \cdot \tilde{C}_{J,H}$$

como $N_H \cdot D_H^+ = \mathcal{Y}_{H,V}$, entonces

$$N_H \cdot D_H^+ \cdot \tilde{C}_{H,J} = N_J \cdot D_J^+ \cdot \tilde{C}_{J,H}.$$

Por un lado, como $N = N_H + N_J$, entonces hacemos $N_J = N - N_H$

$$N_H \cdot D_H^+ \cdot \tilde{C}_{H,J} = (N - N_H) \cdot D_J^+ \cdot \tilde{C}_{J,H}$$

$$N_H = N \cdot \frac{D_J^+ \cdot \tilde{C}_{J,H}}{D_H^+ \cdot \tilde{C}_{H,J} + D_J^+ \cdot \tilde{C}_{J,H}}.$$

Por otro lado, hacemos $N_H = N - N_J$

$$(N - N_J) \cdot D_H^+ \cdot \tilde{C}_{H,J} = N_J \cdot D_J^+ \cdot \tilde{C}_{J,H}$$

$$N \cdot \frac{D_H^+ \cdot \tilde{C}_{H,J}}{D_J^+ \cdot \tilde{C}_{J,H} + D_H^+ \cdot \tilde{C}_{H,J}} = N_J.$$

Obteniendo lo pedido. □

Podemos ver que la Proposición 11 presenta una ecuación para calcular la cantidad de individuos de un grupo utilizando un digrafo que indexa a cada individuo de la población a través de un nodo y que modela todas las relaciones de la población a través de su conjunto de aristas.

Ahora que hemos entendido la idea de este estimador podemos simplificar los cálculos utilizando un grafo en vez de un dígrafo. Para esto se utilizará el grafo subyacente del digrafo que resuelve el problema, de este modo se define el grado medio de un grafo.

Definición 31. *Sea el grafo $G = (V, E)$ y un subconjunto A de V de tamaño N_A, entonces se define el grado medio de A denotado por D_A como:*

$$D_A = \frac{\sum_{v \in A} d(v)}{N_A}.$$

Además se define la siguiente medida.

Definición 32. *Sea el grafo $G = (V, E)$ en donde $V = \{v_1, ..., v_N\}$. Sean los subconjuntos A, B de V, entonces se define $C_{A,B}$ como:*

$$C_{A,B} = \frac{\sum_{v \in A} |\mathcal{N}(v) \cap B|}{\sum_{v \in A} d(v)}.$$

De estas definiciones podemos obtener la siguiente proposición.

Proposición 12. *Sea un digrafo $D = (V, F)$ en donde $V = \{v_1, ..., v_N\}$, D representa una población con relaciones simétricas y $F = \{(u, v) \in V \times V : u \text{ conoce a } v\}$. Sean los subconjuntos H, J de V de tamaño N_H y N_J respectivamente y el grafo subyacente del digrafo D denotado por $G = (V, E)$, entonces obtenemos lo siguiente:*

$$N_H = N \ \frac{D_J \cdot C_{J,H}}{D_H \cdot C_{H,J} + D_J \cdot C_{J,H}}$$

y

$$N_J = N \ \frac{D_H \cdot C_{H,J}}{D_H \cdot C_{H,J} + D_J \cdot C_{J,H}}.$$

Demostración. Para demostrar esta proposición sólo deberemos utilizar la Proposición 11. Así, dado un digrafo $D = (V, F)$ en donde $V = \{v_1, ..., v_N\}$, D representa una población con relaciones simétricas y $F = \{(u, v) \in V \times V : u \text{ conoce a } v\}$. Además, sean los subconjuntos H, J de V de tamaño N_H y N_J respectivamente, entonces utilizando la Proposición 11 se tiene que:

$$N_H = N \ \frac{D_J^+ \cdot \tilde{C}_{J,H}}{D_H^+ \cdot \tilde{C}_{H,J} + D_J^+ \cdot \tilde{C}_{J,H}} \tag{4.5.1}$$

y

$$N_J = N \ \frac{D_H^+ \cdot \tilde{C}_{H,J}}{D_H^+ \cdot \tilde{C}_{H,J} + D_J^+ \cdot \tilde{C}_{J,H}}. \tag{4.5.2}$$

Como $G = (V, E)$ es el grafo subyacente del digrafo $D = (V, F)$ entonces $\mathcal{N}^+(v) = \mathcal{N}(v)$ para cualquier $v \in V$.

Así podemos obtener que:

$$D_J^+ = \frac{\sum_{v \in J} d^+(v)}{N_J} = \frac{\sum_{v \in J} |\mathcal{N}^+(v)|}{N_J}$$

utilizando que $N^+(v) = N(v)$ para cualquier $v \in V$:

$$D_J^+ = \frac{\sum_{v \in J} |\mathcal{N}(v)|}{N_J} = \frac{\sum_{v \in J} d(v)}{N_J} = D_J.$$

Análogo para D_H^+.

Además, se tiene que $\tilde{C}_{H,J}$:

$$\tilde{C}_{H,J} = \frac{\mathcal{Y}_{H,J}}{\mathcal{Y}_{H,V}} = \frac{\sum_{v \in H} |\mathcal{N}^+(v) \cap J|}{\sum_{v \in H} |\mathcal{N}^+(v) \cap V|} = \frac{\sum_{v \in H} |\mathcal{N}^+(v) \cap J|}{\sum_{v \in H} d^+(v)}$$

utilizando que $N^+(v) = N(v)$ para cualquier $v \in V$:

$$\tilde{C}_{H,J} = \frac{\sum_{v \in H} |\mathcal{N}(v) \cap J|}{\sum_{v \in H} d(v)}$$

luego por definición de $C_{H,J}$ tenemos que:

$$\tilde{C}_{H,J} = C_{H,J}.$$

Análogo para $\tilde{C}_{J,H}$.

Ahora, reemplazando lo anterior en la Ecuación 4.5.1 y 4.5.2:

$$N_H = N \, \frac{D_J \cdot C_{J,H}}{D_H \cdot C_{H,J} + D_J \cdot C_{J,H}}$$

y

$$N_J = N \, \frac{D_H \cdot C_{H,J}}{D_H \cdot C_{H,J} + D_J \cdot C_{J,H}}.$$

□

La Proposición 12 es la ecuación fundamental del modelo RDS I. Por lo tanto, teniendo en cuenta la Proposición 12 podemos ver que el modelo RDS I estima la cantidad de individuos de un grupo utilizando un grafo que indexa a cada individuo de la población a través de un nodo y que modela todas las relaciones de la población a través de su conjunto de aristas.

4.5.2. Estimación $C_{H,J}$ y $C_{J,H}$

Para estimar los valores $C_{H,J}$ y $C_{J,H}$ debemos realizar una división entre los distintos tipos de muestreo de nodos a realizar.

En particular si tenemos una secuencia de nodos $s = \{x_1, ..., x_m\}$ de grafo a través de una caminata aleatoria, un proceso viral constante o proceso viral probabilístico, entonces se puede generar una muestra aleatoria de nodos $s - \{x_1\}$. Considerando el caso de una caminata aleatoria, se puede ver a partir de la muestra $s - \{x_1\}$ se puede generar una muestra aleatoria de aristas dirigidas de digrafo asociado $s_e = \{(x_2, x_3), (x_3, x_4), ..., (x_{m-1}, x_m)\}$ de tamaño $(m-2)$. Ahora, realizaremos el análogo para el proceso viral constante y probabilístico. Si aplicamos alguno de los algoritmos de procesos virales, tendremos que para cierto nodo aleatorio x en cierta ola, serán seleccionados ciertos vecinos de x en la siguiente ola, digamos el conjunto $\{u_1, ..., u_\alpha\}$ con $\alpha \in \mathbb{N}$. Esto quiere decir que existen las aristas dirigidas $\{(x, u_1), (x, u_2), ..., (x, u_\alpha)\}$ en el digrafo simétrico asociado. Por lo tanto, si realizamos lo anterior para todas las oleadas y los nodos de $s - \{x_1\}$ se puede generar una muestra aleatoria de aristas dirigidas $s_e = \{e_1, ..., e_{m-2}\}$ que llamaremos muestreo aleatorio de arista inducido por el muestreo de nodos realizado.

Ahora definiremos una función que permite caracterizar las aristas dirigidas.

Definición 33. *Sea un digrafo $D = (V, F)$ y sean los subconjuntos $A, B \subseteq V(D)$, entonces se define la función $\mathcal{X}_{A,B}$ como $\mathcal{X}_{A,B} : F \longrightarrow \{0, 1\}$ y tiene valor:*

$$\mathcal{X}_{A,B}(\tilde{e}) = \begin{cases} 1 \text{ si } \tilde{e} \text{ tiene nodo inicial en } A \text{ y tiene nodo terminal en } B \\ 0 \text{ cualquier otro caso.} \end{cases}$$

En [16] hecho por Salganik, Matthew J; Heckathorn, Douglas D se puede ver que las aristas dirigidas por el método anteriormente mencionado tienen probabilidad de selección constante para cualquier arista dirigida, es decir que todas las aristas dirigidas tienen igual probabilidad de ser elegidas en cierta posición de la muestra y no depende de la posición en la muestra. Como la probabilidad de selección de las aristas es constante para cada posición de la muestra de aristas s_e, significa que podemos realizar estimaciones convencionales en base a s_e, pues es un muestreo aleatorio uniforme.

Estimador 13. *Dado un digrafo conexo y no bipartito $D = (V, F)$ en donde $V = \{v_1, ..., v_N\}$ y los subconjuntos A, B de $V(D)$. Si tenemos la secuencia de nodos*

$s = \{x_1, ..., x_m\}$ hecha por una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante, considerando una muestra aleatoria $s - \{x_1\}$ cuya probabilidad de selección de cierto nodo es proporcional al grado del nodo, entonces podemos generar la muestra aleatoria de aristas dirigidas $s_e = \{e_1, ..., e_{m-2}\}$ asociadas a $s - \{x_1\}$ para estimar $C_{A,B}$ como:

$$C_{A,B} \approx \frac{\sum_{j=1}^{m-2} \mathcal{X}_{A,B}(e_j)}{\sum_{j=1}^{m-2} \mathcal{X}_{A,A}(e_j) + \sum_{j=1}^{m-2} \mathcal{X}_{A,B}(e_j)}.$$

En el caso de realizar un muestreo uniforme con repetición al digrafo $D = (V, F)$, para obtener la muestra $s = \{x_1, ..., x_m\}$ y siguiendo el espíritu de los autores del método RDS I, entonces para cada nodo de la muestra podemos elegir aleatoriamente una arista dirigida que comienza en el nodo mencionado del digrafo D, es decir para $j = 1, ..., m$ se puede elegir una arista dirigida (x_j, u), en donde u es un nodo elegido de forma aleatoria uniforme del conjunto $\mathcal{N}^+(x_j)$, consiguiendo un muestreo aleatorio de aristas $s_e = \{e_1, e_2, ..., e_m\}$ que lo llamaremos muestreo aleatorio de aristas inducido por el muestreo uniforme.

Para calcular la probabilidad de selección asociado al muestreo de aristas antes mencionado, debemos considerar $v \in V(D)$ y $u \in \mathcal{N}^+(v)$ y definir los eventos.

- E_1 : Elegir la arista (v, u) en cierta posición de la muestra de aristas.
- E_2 : Elegir v en cierta posición de la muestra de la muestra de nodos.

Luego, podemos calcular lo siguiente:

$$\Pr(E_1) = \Pr(E_2) \Pr(E_1 \mid E_2)$$

$$\Pr(E_1) = \frac{1}{N} \frac{1}{d^+(v)}.$$

Es decir, elegir cierta arista dirigida en cierta posición de la muestra s_e sólo depende del nodo en que comienza, por lo tanto se puede obtener la siguiente proposición que estima $C_{A,B}$.

Estimador 14. *Dado un digrafo simétrico $D = (V, F)$ en donde $V = \{v_1, ..., v_N\}$ y los subconjuntos A y B de $V(D)$. Si tenemos la muestra aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$, entonces podemos obtener el muestreo aleatorio de aristas inducido por el muestreo uniforme $s_e = \{e_1, ..., e_m\}$, con el cual se puede estimar $C_{A,B}$ como:*

$$C_{A,B} \approx \frac{\sum_{j=1}^{m} d(x_j)\mathcal{X}_{A,B}(e_j)}{\sum_{j=1}^{m} d(x_j)\mathcal{X}_{A}(x_j)}.$$

Justificación: Sea el digrafo simétrico $D = (V, F)$ en donde $V = \{v_1, ..., v_N\}$ y los subconjuntos A y B de $V(D)$. Si tenemos que $F = \{\tilde{e}_1, ..., \tilde{e}_L\}$, entonces por definición de $C_{A,B}$ se tiene que:

$$C_{A,B} = \frac{\sum_{i=1}^{L} \mathcal{X}_{A,B}(\tilde{e}_i)}{\sum_{i=1}^{L} \mathcal{X}_{A,A}(\tilde{e}_i) + \mathcal{X}_{A,B}(\tilde{e}_i)}.$$

Sea $\tilde{p}((v, u))$ la probabilidad de selección de cualquier arista dirigida (v, u), con $v, u \in V(D)$, entonces por lo visto anteriormente se tiene que $\tilde{p}((v, u)) = \frac{1}{N}\frac{1}{d^+(v)}$.

Luego, utilizando el estimador Hansen-Hurwitz para estimar por separado el numerador y el denominador:

$$\frac{\sum_{i=1}^{L} \mathcal{X}_{A,B}(\tilde{e}_i)}{\sum_{i=1}^{L} \mathcal{X}_{A,A}(\tilde{e}_i) + \mathcal{X}_{A,B}(\tilde{e}_i)} \approx \frac{\frac{1}{m}\sum_{j=1}^{m} \frac{d(x_j)}{N}\mathcal{X}_{A,B}(e_j)}{\frac{1}{m}\sum_{j=1}^{m} \frac{d(x_j)}{N}\mathcal{X}_{A,A}(e_j) + \frac{d(x_j)}{N}\mathcal{X}_{A,B}(e_j)}.$$

Por lo tanto,

$$C_{A,B} \approx \frac{\sum_{j=1}^{m} d(x_j)\mathcal{X}_{A,B}(e_j)}{\sum_{j=1}^{m} d(x_j)\mathcal{X}_{A}(x_j)}.$$

□

Obteniendo una forma de estimar $C_{A,B}$ si realizamos un muestreo uniforme con repetición. En resumen, en esta sección hemos propuesto una forma para estima $C_{A,B}$ en función del tipo de muestreo aleatorio.

Por otro lado, la estimación del grado medio de un conjunto de nodos para los distintos tipos de muestreos aleatorios se ha visto varias veces, por lo tanto se recomienda revisarlos en caso de dudas.

En resumen se tiene que el estimador RDS I se basa en:

$$N_H = N\ \frac{D_J \cdot C_{J,H}}{D_H \cdot C_{H,J} + D_J \cdot C_{J,H}}.$$

En donde $C_{H,J}$ se puede estimar como:

$$C_{H,J} \approx \begin{cases} \frac{\sum_{j=1}^{m} d(x_j)\mathcal{X}_{H,J}(e_j)}{\sum_{j=1}^{m} d(x_j)\mathcal{X}_H(x_j)} & \text{Muestreo uniforme con reemplazo.} \\ \frac{\sum_{j=1}^{m-2} \mathcal{X}_{H,J}(e_j)}{\sum_{j=1}^{m-2} \mathcal{X}_{H,H}(e_j)+\sum_{j=1}^{m-2} \mathcal{X}_{H,J}(e_j)} & \text{Muestreo por caminata aleatoria, proceso viral constante o proceso viral probabilístico.} \end{cases}$$

Por último, D_H se puede estimar como:

$$D_H \approx \begin{cases} \frac{\sum_{x\in s} \mathcal{X}_H(x)d(x)}{\sum_{x\in s} \mathcal{X}_H(x)} & \text{Muestreo uniforme con reemplazo.} \\ \frac{\sum_{x\in s} \mathcal{X}_H(x)}{\sum_{x\in s} \frac{\mathcal{X}_H(x)}{d(x)}} & \text{Muestreo por caminata aleatoria, proceso viral constante o proceso viral probabilístico.} \end{cases}$$

Debido a la información solicitada por el estimador RDS I, este estimador utiliza encuestas de tipo indirecta.

4.6. Estimador RDS II

El estimador RDS II es presentado por Volz, Erik; Heckathorn, Douglas D en el articulo [18]. El estimador RDS II simplemente se define a través de un estimador de Hansen-Hurwitz para estimar el tamaño del grupo deseado. Por lo tanto, si las relaciones de la población son simétricas, utilizando un grafo $G = (V, E)$ que modele las relaciones de la población y si H es el subconjunto de $V(G)$ que queremos calcular su tamaño, entonces se tiene que:

$$N_H = \sum_{v\in V} \mathcal{X}_H(v) \approx \frac{1}{m} \sum_{x\in s} \frac{\mathcal{X}_H(x)}{p(x)} = \frac{1}{m} \sum_{x\in s} \frac{f(x)}{p(x)}$$

en donde $p(v)$ es la probabilidad de selección del muestreo del nodo $v \in V(G)$, $s = \{x_1, ..., x_m\}$ es una muestra aleatoria con repetición, por último para cada $x \in s$ se tiene que $f(x) = \mathcal{X}_H(x)$ es la variable de la definición del estimador de Hansen-Hurwitz, en donde se desea estimar el tamaño del conjunto H. Así, podemos obtener la siguiente proposición en el caso de realizar un muestreo uniforme con repetición.

Estimador 15. *Sea un grafo $G = (V, E)$ en donde $V(G) = \{v_1, ..., v_N\}$ y un subconjunto desconocido de $V(G)$ que denotaremos por H de tamaño N_H. Si tenemos una muestra*

aleatoria uniforme con repetición $s = \{x_1, ..., x_m\}$*, entonces utilizando el estimador Hansen-Hurwitz podemos estimar* N_H *como:*

$$N_H \approx \frac{N}{m} \sum_{x \in s} \mathcal{X}_H(x).$$

También podemos obtener realizar un muestreo aleatorio por caminata aleatoria, un proceso viral probabilístico o un proceso viral constante y obtener la siguiente proposición.

Estimador 16. *Dado un grafo conexo y no bipartito* $G = (V, E)$ *en donde* $V = \{v_1, ..., v_N\}$*. Si tenemos la secuencia de nodos* $s = \{x_1, ..., x_m\}$ *hecha por una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante y un subconjunto* $H \subset V$*, considerando una muestra aleatoria* $s - \{x_1\}$ *cuya probabilidad de selección de cierto nodo es proporcional al grado del nodo, entonces podemos estimar* N_H *como:*

$$N_H \approx \frac{N}{\sum_{x \in s - \{x_1\}} \frac{1}{d(x)}} \cdot \sum_{x \in s - \{x_1\}} \frac{\mathcal{X}_H(x)}{d(x)}$$

en donde $N \cdot \frac{m-1}{\sum_{x \in s - \{x_1\}} \frac{1}{d(x)}}$ *es el estimador de* $2|E|$*.*

Debido a la información solicitada por el estimador RDS II, este estimador utiliza encuestas de tipo directa.

Capítulo 5

Experimentación y análisis de métodos estimadores de grupos

En este capítulo se pondrán a prueba los distintos estimadores bajo distintas circunstancias con el objetivo de encontrar propiedades de ellos. De este modo, este capítulo se dividirá en las siguientes secciones:

- Creación de stock de grafos.
- Creación de stock de grupos a estimar y muestras.
- Análisis de distribución grados de muestreos aleatorios
- Estimación del grupo desconocido en circunstancias normales.
- Estimación de tamaño de grupos reales.

5.1. Creación de stock de grafos

En esta etapa se crearon grafos utilizando las funciones barabasi_albert_graph y erdos_renyi_graph de la libreria networkx de python presentada en [6] creada por Aric A. Hagberg and Daniel A. Schult and Pieter J. Swart.

La función barabasi_albert_graph tiene como entrada n y m, en donde n es la cantidad de nodos del grafo a crear y m es la cantidad de arista que genera en cada iteración de un total de $n - m$ iteraciones. El grafo creado por esta función es de tipo conexión preferencial, un tipo especial de grafo escala libre. Los grafos escala libre tienen una

distribución de grados exponencial, es decir pocos nodos tienen un alto grado y muchos nodos tienen un bajo grado. La función barabasi_albert_graph crea grafos con $(n-m)\,m$ aristas. Notar que si queremos encontrar los parámetros para generar un grafo de cierta cantidad de aristas y cantidad de nodos podemos utilizar la ecuación cuadrática $(n-m)\,m = |E|$ para encontrar el valor de m. Debemos recordar que se elegirá el valor m con menor valor y recordar que m debe ser natural. En la Figura 5.1.1 se muestran un ejemplo de la distribución de grados de un grafo creado con la función barabasi_albert_graph de parámetros $(n, m) = (10^4, 101)$ a través de dos histogramas. El histograma de la izquierda está en escala log log mostrando que su distribución de grados en escala log log se puede ajustar a una linea recta, propiedad característica de este tipo de grafos. Por otro lado, el histograma de la derecha tiene una escala lineal que muestra una vista más general de la distribución de grados del grafo.

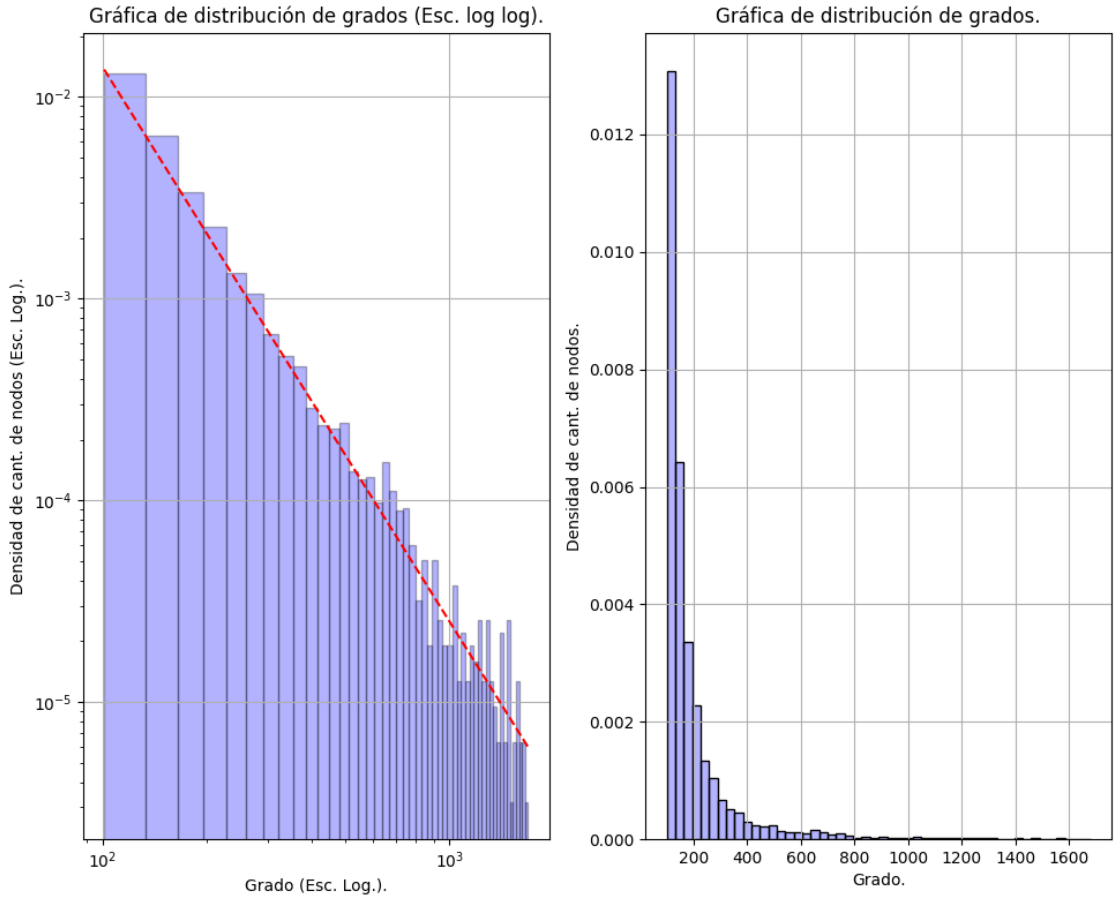


Figura 5.1.1: Distribución de grados de un grafo creado con la función barabasi_albert_graph y parámetros $(n, m) = (10^4, 101)$.

Por otro lado, la función erdos_renyi_graph tiene como entrada n y p, en donde n es la cantidad de nodos del grafo a crear y p es la probabilidad de que exista una arista para

cualquier par de nodos, el grafo creado por esta función es de tipo aleatorio. Los grafos aleatorios tienen una distribución de grados en forma de campana de Gauss, por lo tanto los nodos del grafo tienen grados parecidos. De la definición de erdos_renyi_graph se puede ver que la cantidad esperada de aristas del grafo a crear es $\frac{1}{2}p|V|(|V|-1)$ pero la podemos aproximarla a $\frac{p}{2}|V|^2$. Notar que si queremos encontrar los parámetros para generar un grafo de cierta cantidad de aristas y cantidad de nodos podemos utilizar la ecuación $\frac{p}{2}|V|^2 = |E|$ para encontrar el valor de p. En la Figura 5.1.2 se muestra un ejemplo de histograma de la distribución de grados los grafos creados con la función erdos_renyi_graph de parámetros $(n, p) = (10^4, 2 \cdot 10^{-2})$.

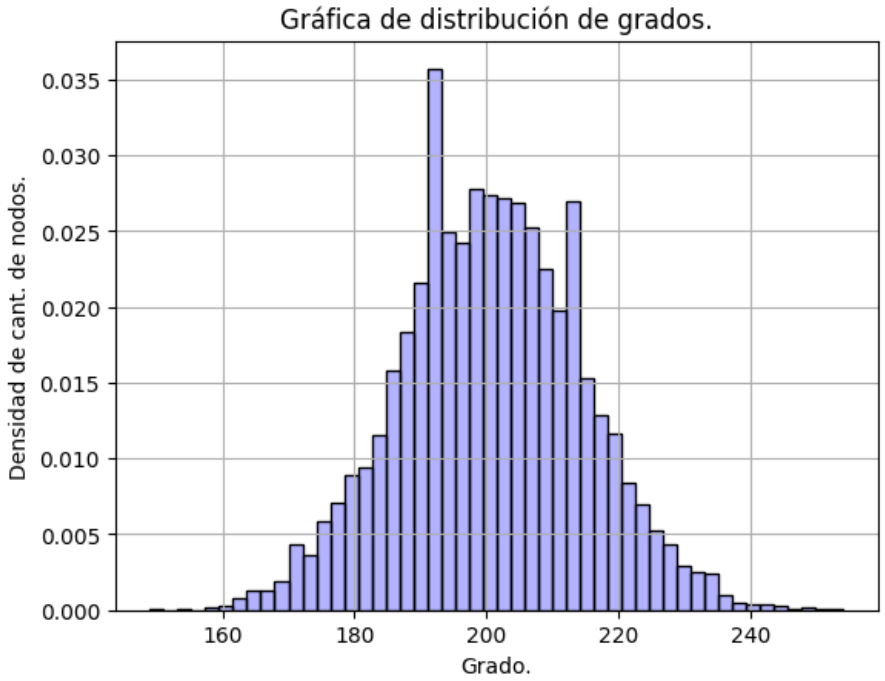


Figura 5.1.2: Distribución de grados de un grafo creado con la función erdos_renyi_graph y parámetros $(n, p) = (10^4, 2 \cdot 10^{-2})$.

Si consideramos la densidad de un grafo como el porcentaje de $\frac{|E(G)|}{|V(G)|^2} \cdot 100\,\%$, entonces podemos crear grafos con densidad de:

- 25 % y 10 % de $|V(G)|^2$ que llamaremos grafos pesados,
- 1 % y 0,1 % de $|V(G)|^2$ que llamaremos grafos medios,
- 0,075 % y 0,05 % de $|V(G)|^2$ que llamaremos grafos ligeros.

El notebook utilizado para crear los grafos tiene 16 gb de ram y windows 11, es decir tiene aproximadamente 8 gb de ram para crear grafos. Realizando una serie de cálculos

y considerando la ram del notebook utilizado hemos generado 100 grafos para las siguientes entradas de las funciones barabasi_albert_graph y erdos_renyi_graph.

Caso: barabasi_albert_graph.

- Para crear los grafos pesados utilizamos $(n, m) = (6000, 3000)$ y $(n, m) = (6000, 676)$.
- Para crear los grafos medios utilizamos $(n, m) = (10^4, 101)$ y $(n, m) = (10^4, 10)$.
- Para crear los grafos ligeros utilizamos $(n, m) = (5 \cdot 10^4, 38)$ y $(n, m) = (5 \cdot 10^4, 25)$.

Para encontrar el valor de m se fijó el valor n y se utilizaron las ecuaciones anteriormente descritas.

Caso: erdos_renyi_graph.

- Para crear los grafos pesados utilizamos $(n, p) = (6000, 5 \cdot 10^{-1})$ y $(n, p) = (6000, 2 \cdot 10^{-1})$.
- Para crear los grafos medios utilizamos $(n, p) = (10^4, 2 \cdot 10^{-2})$ y $(n, p) = (10^4, 2 \cdot 10^{-3})$.
- Para crear los grafos ligeros utilizamos $(n, p) = \left(5 \cdot 10^4, \frac{3}{2} \cdot 10^{-3}\right)$ y $(n, p) = (5 \cdot 10^4, 10^{-3})$.

Para encontrar el valor de p se fijo el valor n y se utilizaron las ecuaciones anteriormente descritas.

El peso total de los grafos almacenados es 41 Gb y se tarda en generarse un aproximado de 16 horas en un notebook ryzen 5600h de procesador $3{,}3 - 4{,}2$ GHz. Además, cabe resaltar que todos los grafos creados son conexos y no bipartitos.

5.2. Creación de stock de grupos a estimar y muestras

La idea de esta sección es generar conjuntos H y muestras aleatorias del conjunto $V(G)$ de distintos tamaños a través varios de métodos y para cada grafo creado en la sección anterior. Para generar los conjuntos H utilizamos los algoritmos definidos anteriormente con ciertas alteraciones, es decir los muestreos aleatorios con repetición como el muestreo uniforme con repetición, muestreo asociado a una caminata aleatoria, a un proceso viral constante y a un proceso viral probabilístico, hicimos que generaran

nodos aleatorios hasta que tuvieran una cierta cantidad de nodos distintos y se eliminaron los nodos repetidos. Además, se utilizó el módulo numpy que tiene una función llamada random.choice, la cuál realiza una selección de elementos de un conjunto, en donde cada elemento puede tener una cierta ponderación para ser seleccionado y los elementos elegidos se van descartando, es decir un elemento no se puede muestrear más de una vez. Utilizamos dos mapeos distintos de ponderación, una fue el grado del nodo y la otra fue el inverso del grado del nodo. Los tamaños de H para cada método descrito anteriormente son $1\,\%, 10\,\%, 50\,\%$ y el $90\,\%$ de la cantidad de nodos del grafo en que se forma. En resumen, para cada grafo del stock de la sección anterior se crearon conjuntos H que varían en método de selección y tamaño.

Por otro lado, para generar los muestreos aleatorios de nodos de cada grafo generado se utilizaron el muestreo uniforme con repetición, muestreo aleatorio asociado a una caminata aleatoria, a un proceso viral constante y a un proceso viral probabilística cuyo tamaño de muestra es del $1\,\%, 10\,\%, 50\,\%$ y el $90\,\%$ de la cantidad de nodos del grafo en que se forma. Para calcular los parámetros de los muestreos a través de los procesos virales se calculó un c (proceso viral constante) tal que tuviera como mínimo una cantidad de oleadas. Salganik, Matthew J, Heckathorn, Douglas D en [16] recomiendan seis oleadas, por lo tanto calculamos un valor c que depende de la cantidad de nodos del grafo y la proporción de nodos a muestrear para que tenga seis oleadas. A partir de ese valor c, se calculó un p (parámetro del proceso viral probabilístico) como se define a continuación $p = \text{mín}(1, c \cdot \frac{N}{2|E|})$, para que tenga aproximadamente seis oleadas cada muestra asociada a un proceso viral probabilístico. Notar que $c \cdot \frac{N}{2|E|}$ es c dividido por el grado medio del grafo.

Así, para cada grafo del stock de la sección anterior se crearon muestras aleatorias que varían en método de selección y tamaño. El peso total entre los subconjuntos y las muestras generadas es de aproximadamente 40 gb y se tarda 40 horas.

5.3. Análisis de distribución grados de muestreos aleatorios

Antes de mostrar las aproximaciones de la distribución del error relativo se hará un breve análisis de la distribución de grados de los muestreos aleatorios. El objetivo de esta subsección es mostrar en que situaciones la muestra aleatoria no se comporta como indica su probabilidad de selección o que aún la muestra es demasiado pequeña.

Si consideramos un grafo $G = (V, E)$ y una muestra aleatoria uniforme con repetición s de tamaño m, la fracción esperada de nodos de grado d de una muestra s es igual a la fracción de nodos de grado d del grafo. Debido a lo anterior, tomamos un grafo elegido al azar para cada tipo de densidad (pesado, medio o ligero) y de cada tipo de grafo (aleatorio o escala) para construir un histograma de distribución de grado de las muestras s de tipo aleatorio uniforme con repetición. Es decir, para cada un cierto grafo, concatenamos todas las muestras uniforme con repetición con una proporción fija. La idea, es graficar el histograma de grados y compararlos con la distribución de grados del grafo. Los histogramas relacionados a esta situación se encuentran en el apéndice A con su respectiva descripción.

En el caso de muestras aleatorias asociada a una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante, se realizó algo análogo, solamente que con respecto a la distribución de grados de la población $\mathcal{N}(V(G))$, en donde $\mathcal{N}(V(G))$ es la concatenación de todas las vecindades del grafo G. Notar que la fracción esperada de nodos de grado d de una muestra s es igual a la fracción de nodos de grado d en $\mathcal{N}(V(G))$. Los histogramas relacionados a esta situación se encuentran en el apéndice A con su respectiva descripción.

Análisis de los histogramas.
De las figuras 5.3.1 y 5.3.2 se puede ver que la distribución de grados de la muestra tiende a ser más parecida a la distribución teórica cuando la muestra aumenta de tamaño para cualquier grafo. Por lo general, el ajuste de las muestras con proporción del 1 % del tamaño de la cantidad de nodos del grafo son muy malas en comparación a las muestras con proporción del 10 %. De hecho, si aumentamos nos fijamos en las muestras con proporción del 50 % no existe mucha diferencia con el ajuste de las muestras al 10 %. Por otro lado, todos los tipos de muestreos se ajustan a un nivel parecido a sus distribuciones teóricas.

Por lo tanto, sólo se puede concluir que una condición necesaria para que la muestra genere buenas estimaciones para los estimadores GNSUM, RDS I y RDS II (sólo estos estimadores utilizan la probabilidad de selección) es que la muestra tenga una proporción igual o superior al 10 % del tamaño de la población. Hemos de notar que las demás gráficas de este tipo están en el Apéndice A.

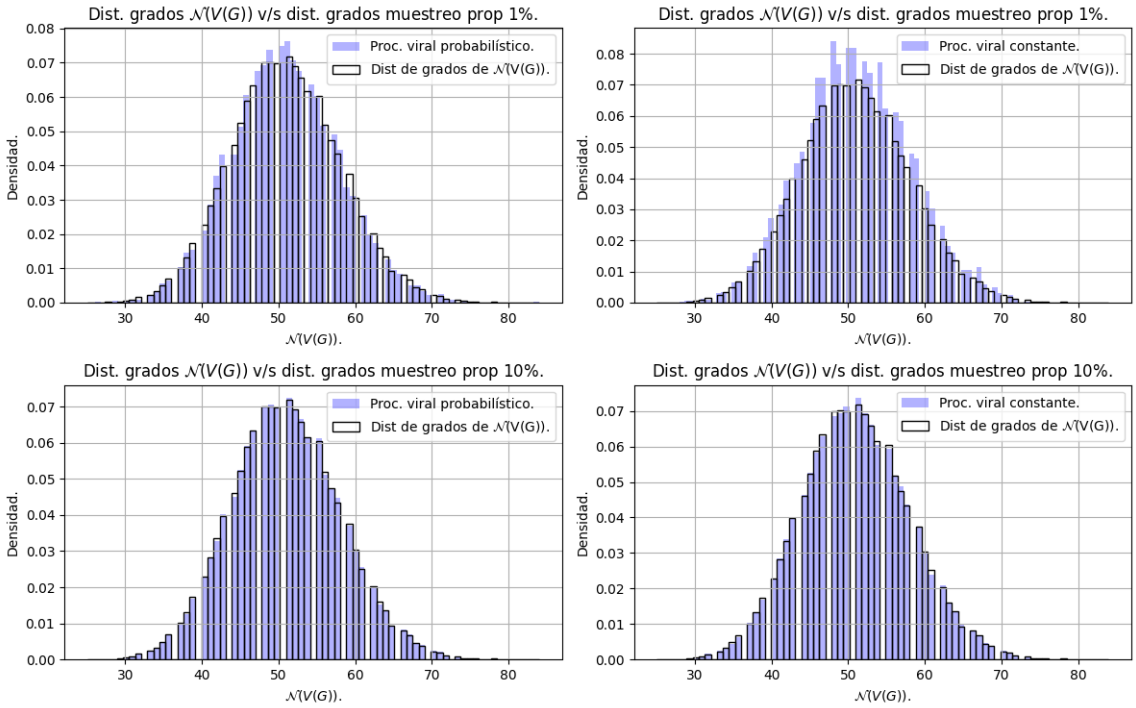


Figura 5.3.1: Gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral probabilístico variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral constante variando su proporción de tamaño de un grafo aleatorio ligero.

5.4. Estimación del grupo desconocido en distintas circunstancias

Como ya tenemos un stock de grafos, grupos para estimar y muestras aleatorias, entonces podemos comenzar a realizar estimaciones de tamaños de grupos con los métodos antes definidos.

Definimos el error relativo de un método como $\frac{N_H - \hat{N}_H}{N_H}$ en donde N_H y $\hat{N}_H$ es el valor exacto y estimado del tamaño de un grupo relacionado en cierta situación, respectivamente. En este trabajo se considerará que el error relativo de cierto método estimador bajo cierta situación sirve para medir el rendimiento del método estimador para dicha circunstancia. Además, si para cada estimador consideramos el experimento aleatorio de realizar una muestra aleatoria, para luego estimar el tamaño de un grupo con el estimador asociado, entonces el rendimiento de un estimador puede ser modelado como una variable aleatoria con distribución desconocida que deseamos estudiar y los

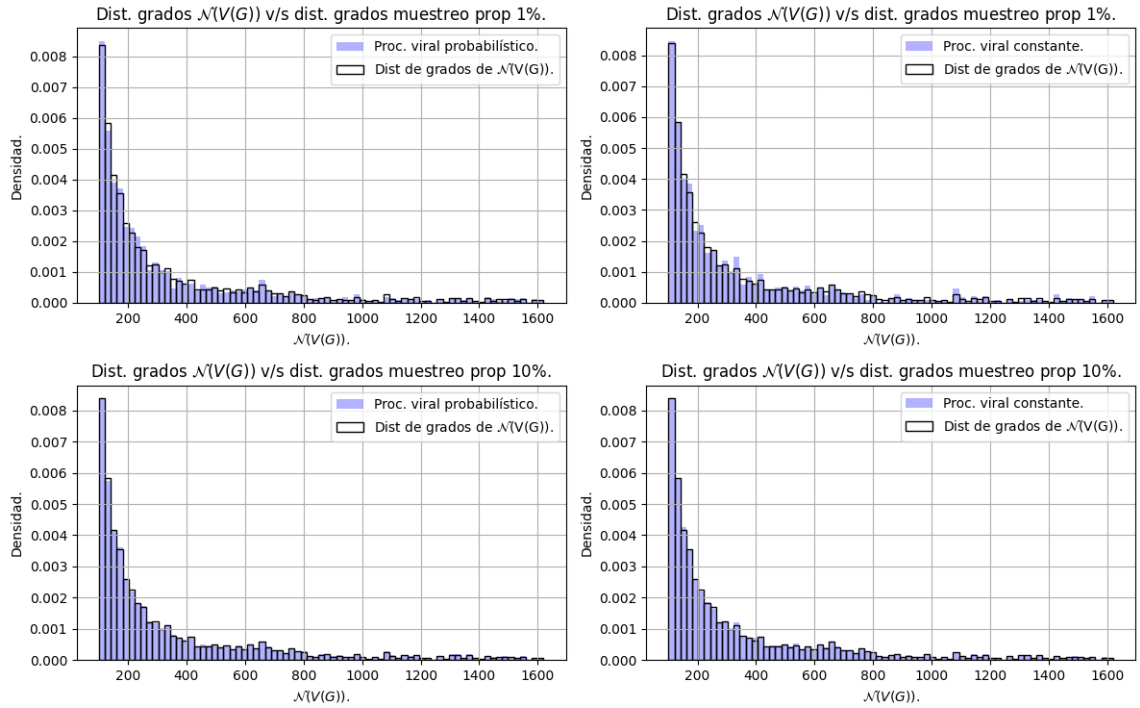


Figura 5.3.2: Gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral probabilístico variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral constante variando su proporción de tamaño de un grafo escala medio.

ensayos son realizaciones o un muestreo de dicha variable aleatoria. La idea es utilizar las realizaciones de la variable aleatoria, que modela el rendimiento de un estimador, para estudiar sus propiedades como la esperanza, la varianza e intervalo de confianza. De lo anterior, dado cierto contexto, podemos considerar que la media de las observaciones puede ser considerada como estimador de la esperanza del rendimiento para dicho contexto, análogo para estimadores de la varianza e intervalo de confianza.

5.4.1. Estimación de la distribución de densidad del error relativo e intervalos de confianza para la esperanza del error relativo

En esta subsección se presentarán varios experimentos, los experimentos consistirán en realizar estimaciones del tamaño de un grupo desconocido utilizando los estimadores antes definidos. La diferencia entre cada experimento es la generación del conjunto H y los resultados de estos serán mostrados a través de dos tipos de figuras, un tipo para

mostrar una aproximación de la distribución de densidad del error relativo, y otra para la media muestral del error relativo y el intervalo de confianza para la esperanza del error relativo. Además, cada figura mostrará una tabla de gráficas asociadas al tipo de figura.

En el caso de las figuras que muestran la distribución del error relativo, la figura está compuesta por una tabla de gráficas, cada gráfica contiene varias aproximaciones de la distribución de densidad del error relativo de los distintas proporciones de nodos muestreados (1 %, 10 %, 50 % y 90 %) fijando un estimador y un tipo de muestra. Además, para cada fila de la tabla de gráficas muestra los resultados de un estimador distinto y en cada columna varia los tipo de muestras.

En el caso de las figuras que muestran la media muestral del error relativo, la figura está compuesta por una tabla de gráficas, cada gráfica contiene la media muestral e intervalo de confianza para la esperanza del error relativo de cada estimador en donde el eje x indica la proporción del tamaño de H con respecto N para una proporción de nodos muestreados y tipo de muestreo fijos. Además, para cada fila de la tabla de gráficas muestra los resultados de una proporción de nodos muestreados distintos(1 %, 10 %, 50 % y 90 %) y en cada columna varia los tipo de muestras.

Ahora procedemos a explicar como se genera el conjunto H de cada experimento y mostrar los resultados asociados.

Experimento 1**.**
En el caso del experimento 1 no consideramos condiciones en la forma de generar H, es decir *los métodos utilizados para generar H son el uniforme con respecto al grado del nodo, caminata aleatoria, proceso viral probabilístico, proceso viral proceso viral constante, selección directamente proporcional al grado del nodo y la selección inversamente proporcional al grado del nodo.*

La primera fila de la Figura 5.4.3 muestra las aproximaciones de la distribución de densidad del error relativo de cada estimador asociado al experimento 1 tomando muestras de tamaño igual al 10 % del tamaño de la población. La Figura 5.4.3 está compuesta por cuatro gráficas, cada una corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden).

Por otro lado, en la Figura 5.4.1 se muestran las estimaciones de la media muestra e intervalo de confianza del error relativo al 95 % de cada estimador asociado al

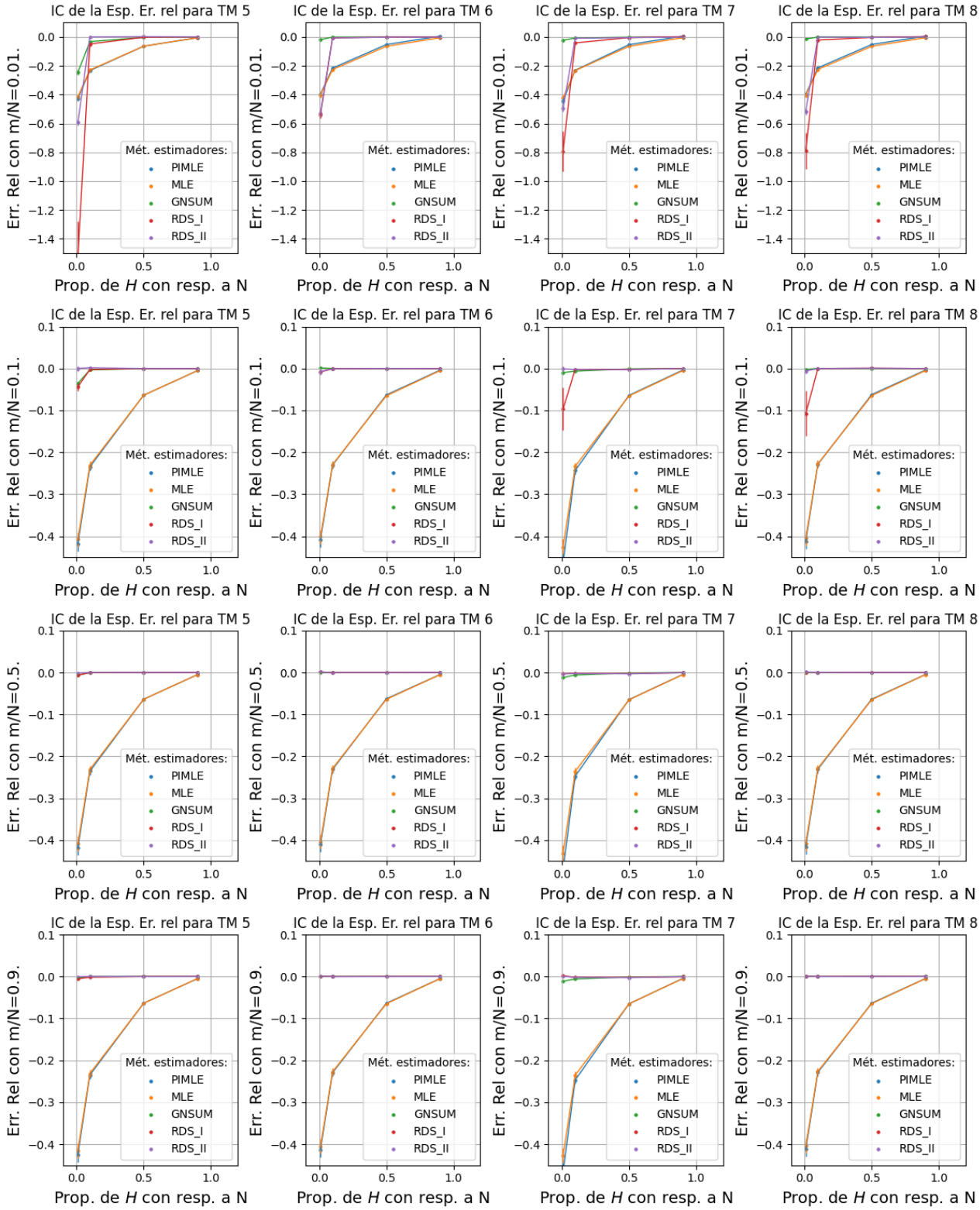


Figura 5.4.1: Experimento 1: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

experimento 1. La Figura 5.4.1 está compuesta por cuatro columnas y cuatro filas, cada columna corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden). Además, cada fila de la Figura 5.4.1 corresponde a una proporción de nodos muestreadas distinta, en donde las proporciones muestreadas son $1\,\%, 10\,\%, 50\,\%$ y $90\,\%$.

Para mayores detalles de las aproximaciones de la distribución de densidad del error relativo se puede ver las figuras B0.1, B0.2 y B0.3 del Apéndice B. En caso de las aproximaciones de la media muestra e intervalo de confianza del error relativo, se pueden mayores detalles en las figuras B0.4, B0.5 y B0.6 del Apéndice B.

Experimento 2.
En el caso del experimento 2, *la generación de los conjuntos H se realiza utilizando el muestreo uniforme sin repetición*, es decir la selección de los nodos de H no depende del grado del nodo.

La segunda fila de la Figura 5.4.3 muestra las aproximaciones de la distribución de densidad del error relativo de cada estimador asociado al experimento 2 tomando muestras de tamaño igual al $10\,\%$ del tamaño de la población.

La Figura 5.4.3 está compuesta por cuatro gráficas, cada una corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden).

Por otro lado, en la Figura 5.4.2 se muestran las estimaciones de la media muestra e intervalo de confianza del error relativo al $95\,\%$ de cada estimador asociado al experimento 2.

La Figura 5.4.2 está compuesta por cuatro columnas y cuatro filas, cada columna corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden). Además, cada fila de la Figura 5.4.2 corresponde a una proporción de nodos muestreadas distinta, en donde las proporciones muestreadas son $1\,\%, 10\,\%, 50\,\%$ y $90\,\%$.

Para mayores detalles de las aproximaciones de la distribución de densidad del error relativo se puede ver las figuras B0.7, B0.8 y B0.9 del Apéndice B. En caso de las aproximaciones de la media muestra e intervalo de confianza del error relativo, se pueden mayores detalles en las figuras B0.10, B0.11 y B0.12 del Apéndice B.

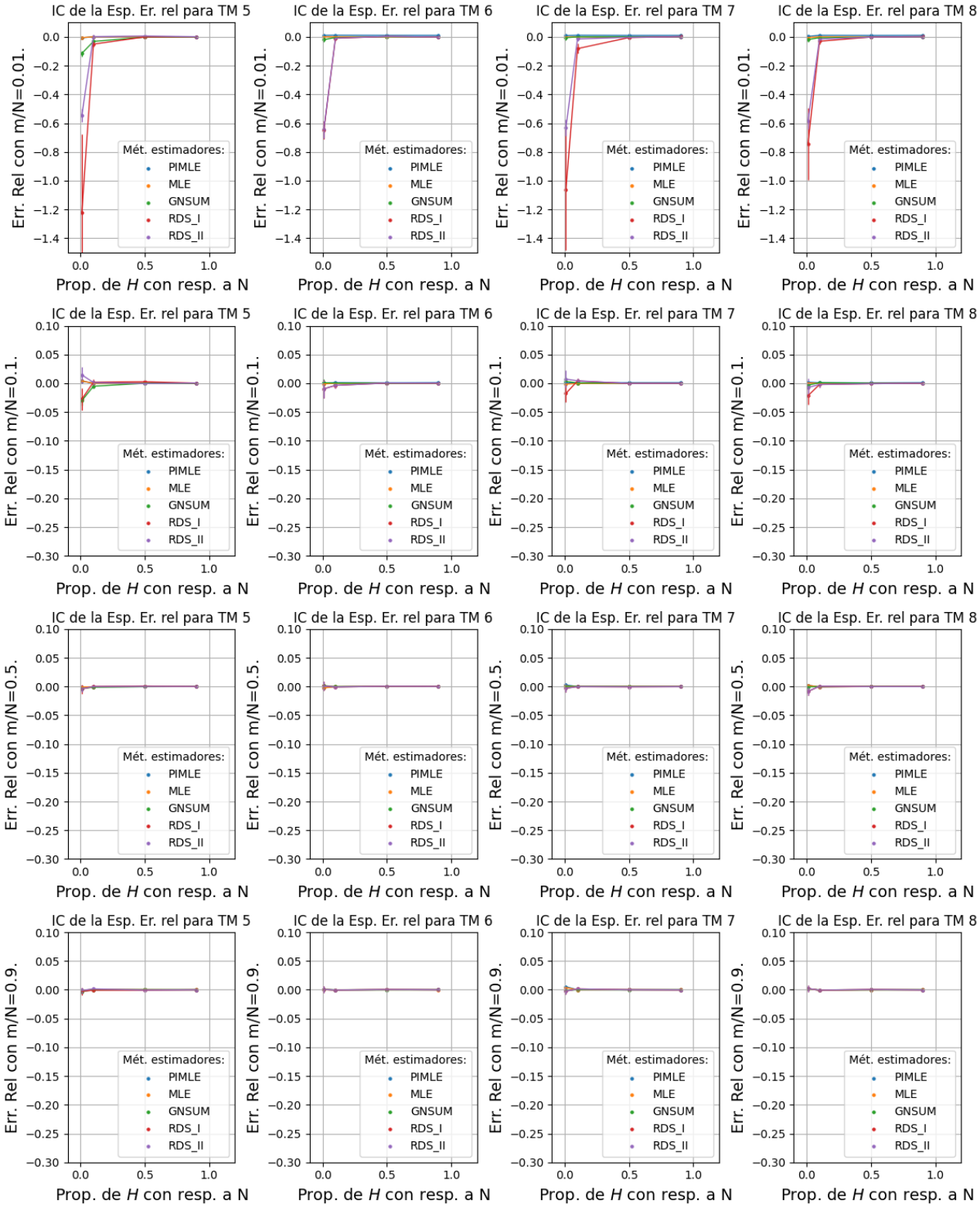


Figura 5.4.2: Experimento 2: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

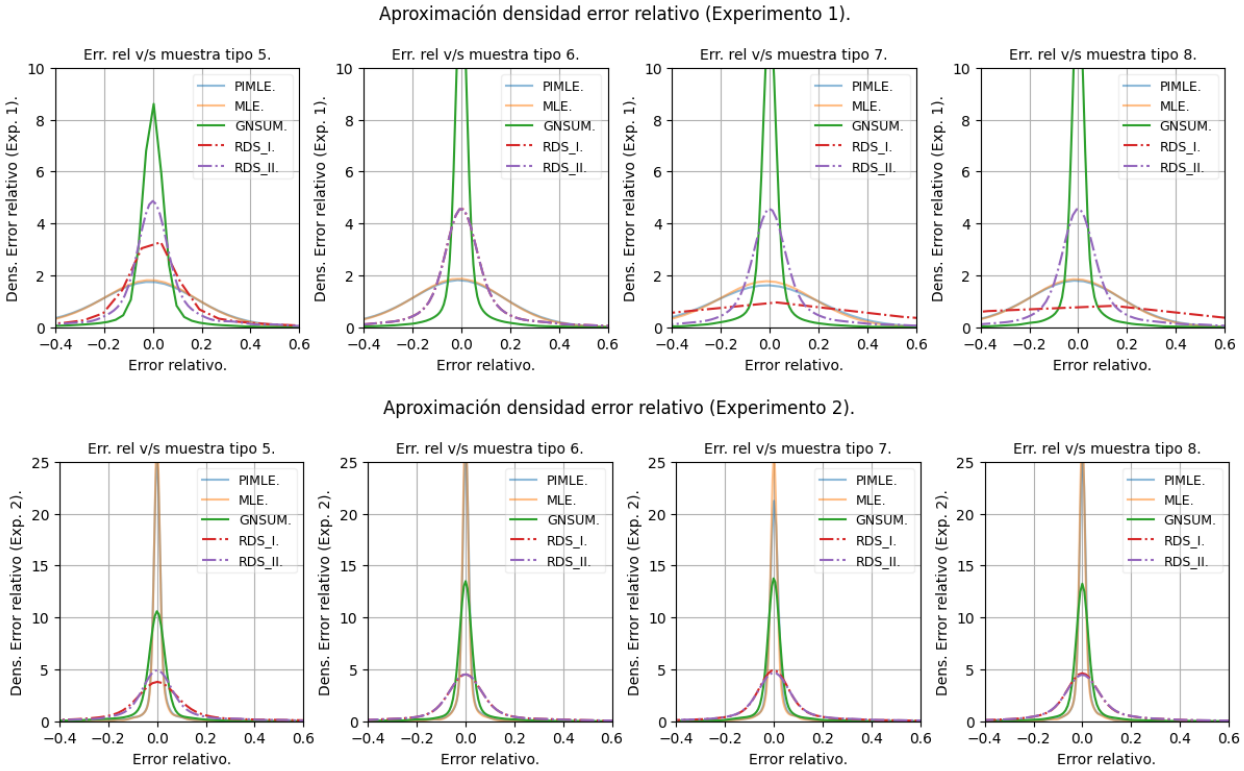


Figura 5.4.3: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población. Los resultados correspondientes al experimento 1 se encuentran en la primera fila, mientras que los del experimento 2 están en la segunda fila. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

Experimento 3**.**

En el caso del experimento 3, *la generación de los conjuntos H se realiza utilizando el proceso viral constante sin repetición.*

La primera fila de la Figura 5.4.6 muestra las aproximaciones de la distribución de densidad del error relativo de cada estimador asociado al experimento 3 tomando muestras de tamaño igual al 10 % del tamaño de la población. La Figura 5.4.6 está compuesta por cuatro gráficas, cada una corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden).

Por otro lado, en la Figura 5.4.4 se muestran las estimaciones de la media muestra e intervalo de confianza del error relativo al 95 % de cada estimador asociado al experimento 3. La Figura 5.4.4 está compuesta por cuatro columnas y cuatro filas, cada

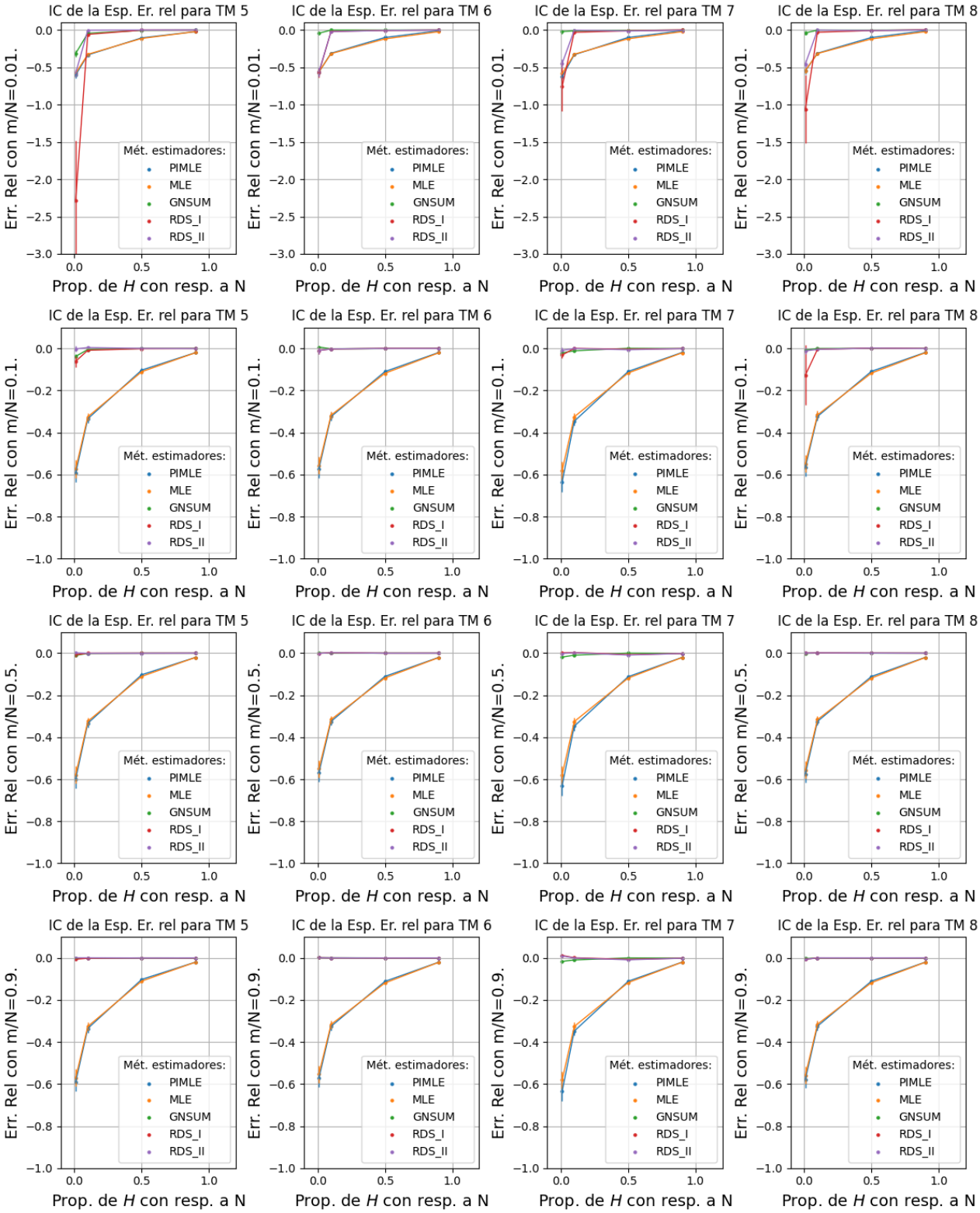


Figura 5.4.4: Experimento 3: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

columna corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden). Además, cada fila de la Figura 5.4.4 corresponde a una proporción de nodos muestreadas distinta, en donde las proporciones muestreadas son 1 %, 10 %, 50 % y 90 %.

Para mayores detalles de las aproximaciones de la distribución de densidad del error relativo se puede ver las figuras B0.13, B0.14 y B0.15 del Apéndice B. En caso de las aproximaciones de la media muestra e intervalo de confianza del error relativo, se pueden mayores detalles en las figuras B0.16, B0.17 y B0.18 del Apéndice B.

Experimento 4.
En el caso del experimento 4, *la generación de los conjuntos H se realiza utilizando una selección de nodos directamente proporcional al grado del nodo*.

La segunda fila de la Figura 5.4.6 muestra las aproximaciones de la distribución de densidad del error relativo de cada estimador asociado al experimento 4 tomando muestras de tamaño igual al 10 % del tamaño de la población. La Figura 5.4.6 está compuesta por cuatro gráficas, cada una corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden).

Por otro lado, en la Figura 5.4.5 se muestran las estimaciones de la media muestra e intervalo de confianza del error relativo al 95 % de cada estimador asociado al experimento 4.

La Figura 5.4.5 está compuesta por cuatro columnas y cuatro filas, cada columna corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden). Además, cada fila de la Figura 5.4.5 corresponde a una proporción de nodos muestreadas distinta, en donde las proporciones muestreadas son 1 %, 10 %, 50 % y 90 %.

Para mayores detalles de las aproximaciones de la distribución de densidad del error relativo se puede ver las figuras B0.19, B0.20 y B0.21 del Apéndice B. En caso de las aproximaciones de la media muestra e intervalo de confianza del error relativo, se pueden mayores detalles en las figuras B0.22, B0.23 y B0.24 del Apéndice B.

Experimento 5.
Por último, para el experimento 5 *la generación de los conjuntos H se realiza utilizando*

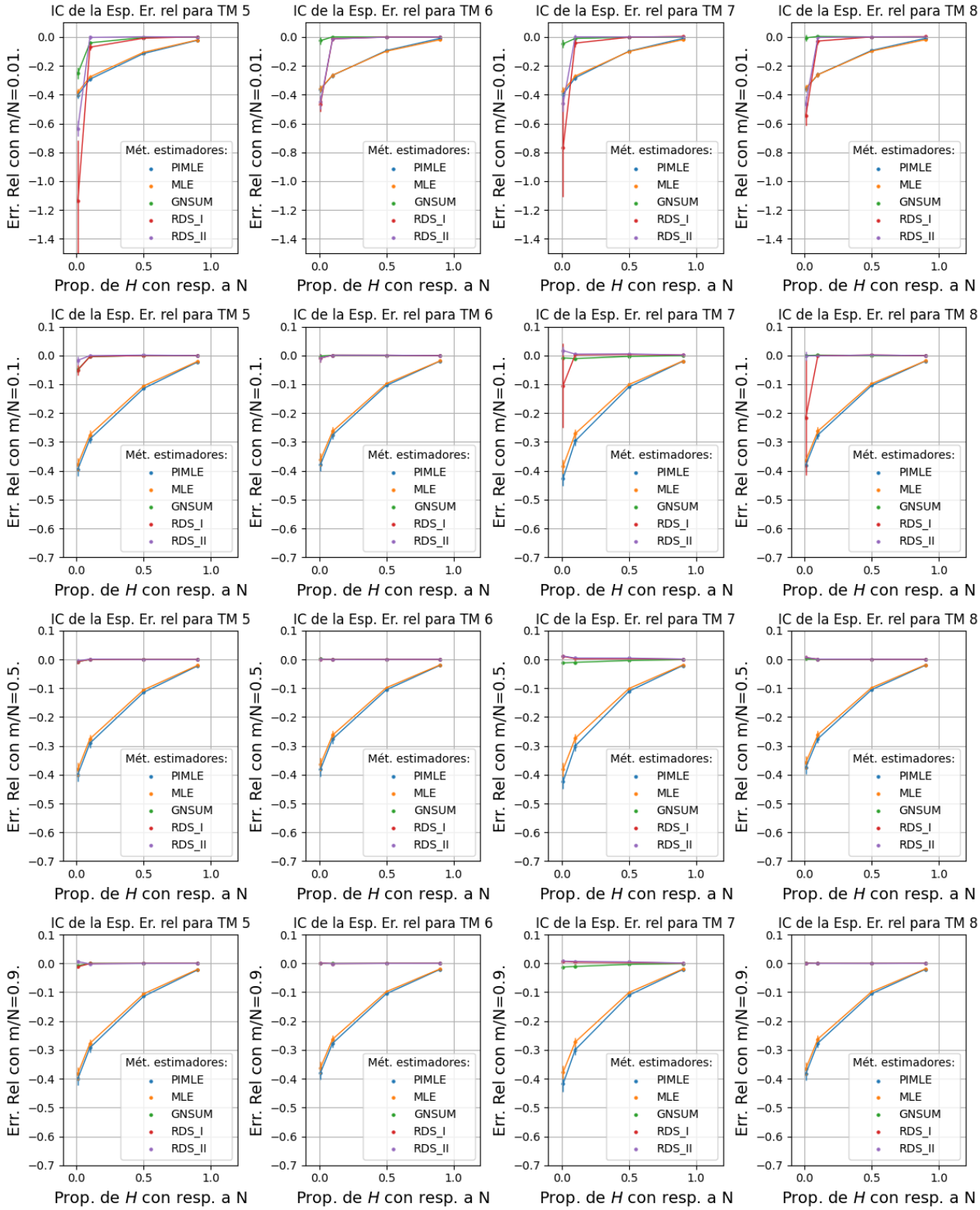


Figura 5.4.5: Experimento 4: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

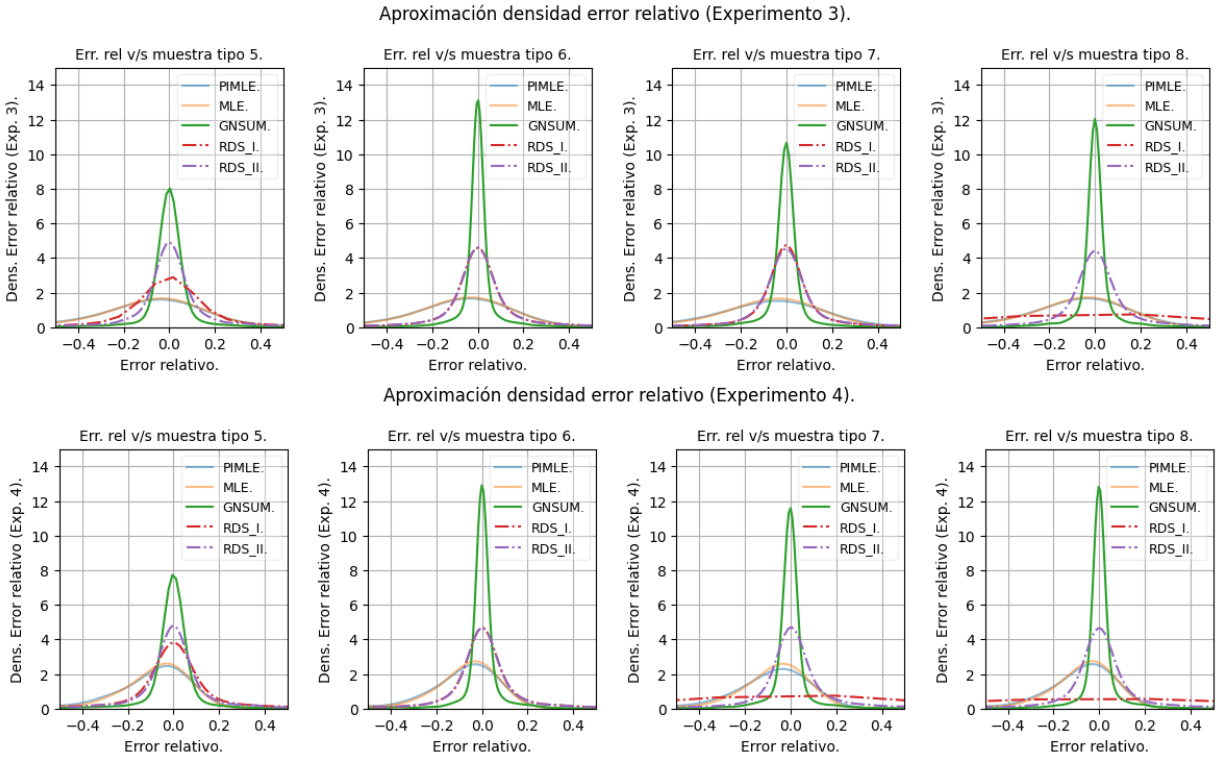


Figura 5.4.6: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población. Los resultados correspondientes al experimento 3 se encuentran en la primera fila, mientras que los del experimento 4 están en la segunda fila. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

una selección de nodos inversamente proporcional al grado del nodo.

La Figura 5.4.8 muestra las aproximaciones de la distribución de densidad del error relativo de cada estimador asociado al experimento 5 tomando muestras de tamaño igual al 10 % del tamaño de la población. La Figura 5.4.8 está compuesta por cuatro gráficas, cada una corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden).

Por otro lado, en la Figura 5.4.7 se muestran las estimaciones de la media muestra e intervalo de confianza del error relativo al 95 % de cada estimador asociado al experimento 5. La Figura 5.4.7 está compuesta por cuatro columnas y cuatro filas, cada columna corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral

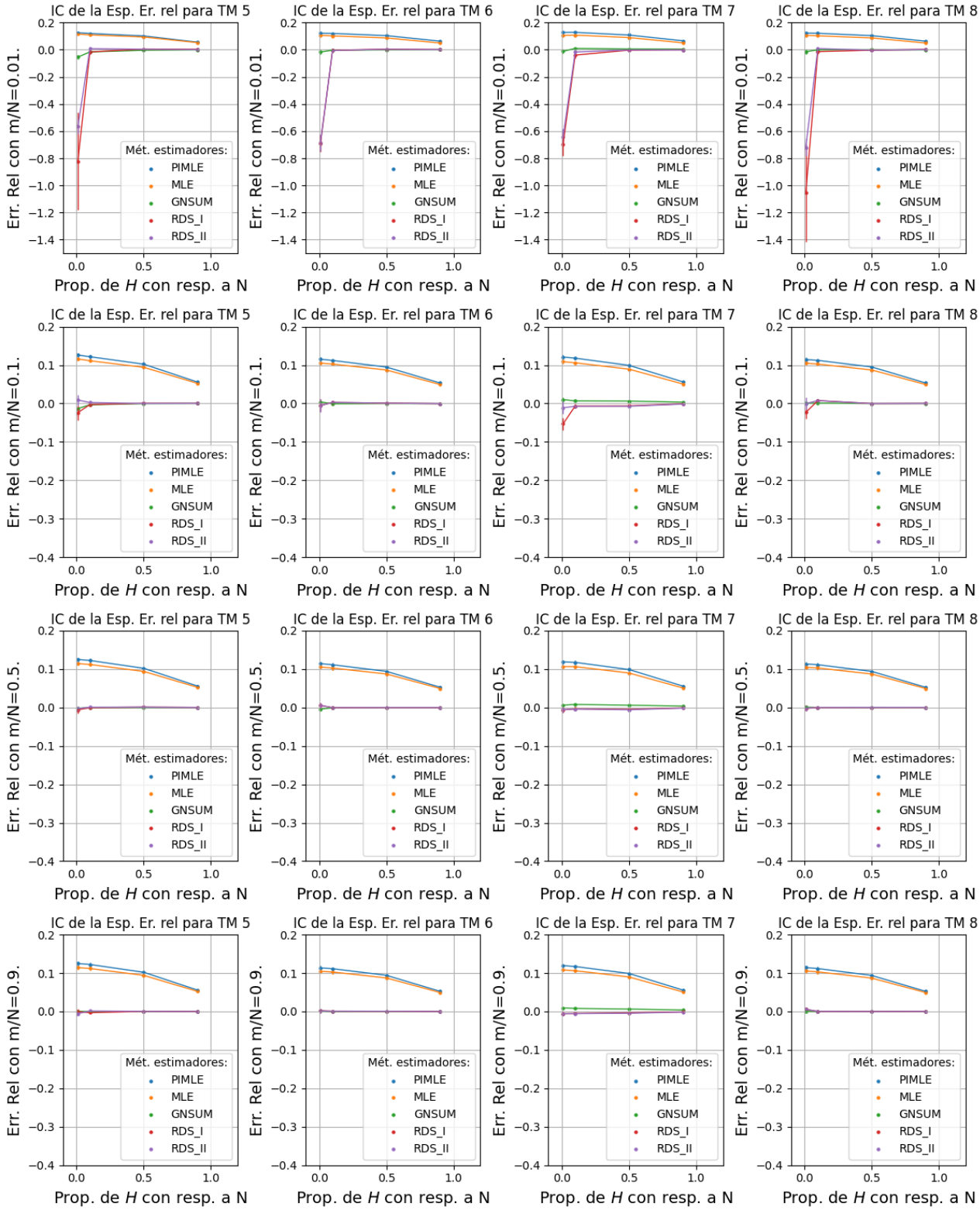


Figura 5.4.7: Experimento 5: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

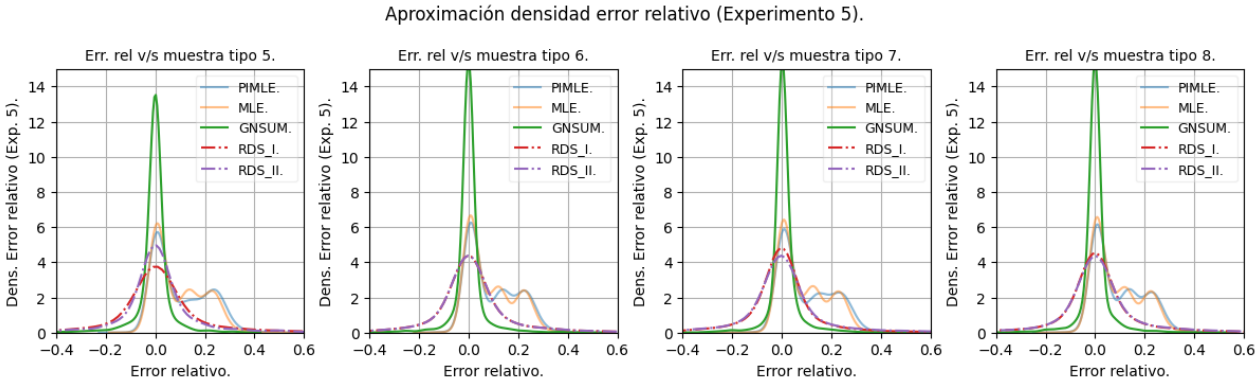


Figura 5.4.8: Gráficas de la distribución del error relativo utilizando un tamaño de muestreo del 10 % del tamaño de la población del experimento 5. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

constante, en ese orden). Además, cada fila de la Figura 5.4.7 corresponde a una proporción de nodos muestreadas distinta, en donde las proporciones muestreadas son 1 %, 10 %, 50 % y 90 %.

Para mayores detalles de las aproximaciones de la distribución de densidad del error relativo se puede ver las figuras B0.25, B0.26 y B0.27 del Apéndice B. En caso de las aproximaciones de la media muestra e intervalo de confianza del error relativo, se pueden mayores detalles en las figuras B0.28, B0.29 y B0.30 del Apéndice B.

Análisis de los resultados de los experimentos.
De forma general, podemos ver que el estimador GNSUM es el con mejor rendimiento, luego le sigue el RDS II , RDS I, MLE y PIMLE. También podemos ver que a medida que el tamaño del conjunto H aumenta (por sobre el 10 % por debajo del 90 %) entonces las estimaciones son mejores para todos los estimadores, esto a pesar que el modelo MLE supone que N_H es mucho más pequeño que N.

Además, podemos ver que en los casos en que H no es uniforme con respecto a los grados de los nodos, entonces el modelo MLE y PIMLE tienden a fallar. Más concretamente, cuando H es generado por un método directamente proporcional al grado del nodo, entonces los métodos sobreestiman la estimación y en caso que H se genere por un método inversamente proporcional al grado de los nodos entonces los métodos MLE y PIMLE subestiman el tamaño de H. Pero si H es generado por método uniforme con respecto al grado entonces los métodos PIMLE y MLE son los que tienen una mejores

estimaciones. Otro hecho importante, es que los métodos PIMLE Y MLE no se ven afectados por el tamaño de la muestra, pues su distribución de error relativo en cada caso no se ve alterada. La varianza de la aproximación del error relativo de los modelos PIMLE y MLE es la más alta de todos los métodos, a excepción cuando H es generada por una secuencia de nodos uniforme con respecto al grado. Cuando H es generada por una secuencia de nodos uniforme con respecto al grado entonces los modelos PIMLE y MLE tienen la menor varianza.

Por otro lado, el modelo que más se ve afectado por el tamaño de muestra es el RDS I pues cuando el tamaño de muestra es del 1 % se obtienen resultados de muy mala calidad pero a medida que aumenta la proporción de la muestra tiene un rendimiento semejante al modelo RDS II. En el único caso que el rendimiento del modelo RDS I es comparable con el estimador RDS II es cuando se muestrea a través de una caminata aleatoria, entonces podemos concluir que a mayor cantidad de oleadas del proceso viral constante, este generará una mejor muestra para el estimador RDS I.

Los estimadores GNSUM y RDS II son los que tienen mejor rendimiento, siendo superior el estimador GNSUM. Además, estos estimadores no muestran variaciones significativas en la aproximación de la distribución del error relativo cuando varia el tipo de muestreo y la proporción del tamaño de muestreo con respecto la población.

La varianza de la aproximación de la distribución del error relativo de los modelos RDS I, RDS II y GNSUM permanecen invariable con respecto el tipo de muestreo, sólo cambian con respecto a la proporción de nodos muestreados y el método de generación de H.

5.5. Estimación de tamaño de grupos reales

En esta sección se mostrará el rendimiento de los estimadores y tipos de muestreo en un contexto real, es decir se utilizará un grafo real en donde las etiquetas de los nodos también son reales para realizar las estimaciones de los distintos grupos de nodos utilizando los distintos tipos de muestreos y métodos.

Para seleccionar el grafo a utilizar se buscó en la página de snap.stanford [10], que contiene una gran cantidad de grafos. El grafo seleccionado fue un grafo de Twitch [15]. Este grafo está compuesto por nodos que representan a los usuarios de la página y las aristas que son las relaciones entre estos. Cada usuario puede utilizar la plataforma para

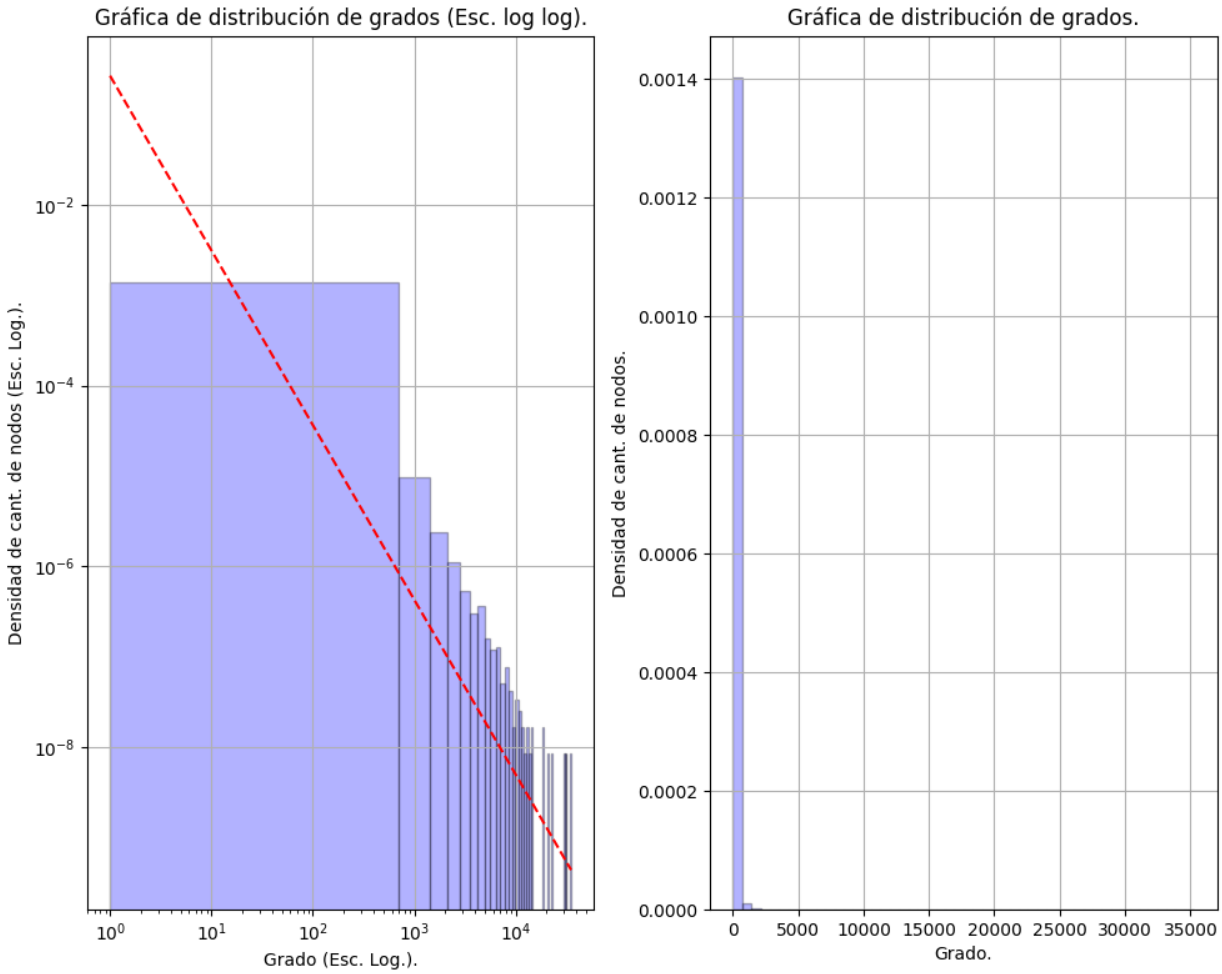


Figura 5.5.1: Gráficas que muestran la distribución de la frecuencia de grados del grafo de Twitch con escala log log (izquieda) y escala lineal (derecha).

ver streaming de otros usuario, realizar el streaming o ambos. El archivo descargado también contiene información de las características de los nodos del grafo. Estas características se muestran a través de una tabla cuyas columnas son: cantidad de espectadores, si es experimentado o no, tiempo como usuario, fecha de creación de la cuenta, id del usuario, si es una cuenta inactiva o no, idioma y si es auspiciado. De estas características sólo vamos a estimar usuarios con menos de 520 espectadores, cantidad de usuarios no experimentados, cantidad de usuarios que tengan menos de 600(medida) de tiempo de vida (la tabla no especifica medida), cantidad de cuentas inactivas, cantidad de usuarios de habla inglés, cantidad de usuarios de habla francés, cantidad de usuarios de habla sur coreano y cantidad de usuarios de habla japonés.

Ahora procedemos a describir el grafo de Twitch [15], el grafo tiene 168114 nodos,

6797557 aristas, densidad $0,024\,\%$ ($\frac{|E(G)|}{|V(G)|^2} \cdot 100\,\%$), su grado medio es 40,42 y el grafo es conexo. A continuación mostraremos su gráfica de frecuencia de grados en escala logarítmica. De la Figura 5.5.1 notamos que la distribución de grado del grafo de Twitch en escala log log se ajusta bien a una linea recta.

Para poder realizar las estimaciones de densidad del error relativo de cada método e intervalos de confianza de la esperanza del error relativo necesitamos un grupo de muestras de un mismo tipo y tamaño. Por lo tanto, se realizaron 100 muestras aleatorias de cada tipo (uniforme con repetición, asociada a una caminata aleatoria, asociada a un proceso viral probabilístico y asociada a un proceso viral constante) y proporciones con respecto la cantidad de nodos del grafo ($1\,\%, 10\,\%, 50\,\%$ y $90\,\%$).

Para cada muestra aleatoria obtenida se ejecutaron los distintos métodos estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II), los resultados de las estimaciones fueron almacenadas y luego utilizadas para generar las aproximaciones de la distribución del error relativo e intervalo de confianza de la esperanza del error relativo. A continuación se definen los distintos experimentos realizados.

Experimento 1**.**
En el caso del experimento 1 se realizaron las estimaciones de tamaño de grupo para todos los grupos antes definidos utilizando las muestras aleatorias almacenadas y los diferentes tipos de estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II).

Los resultados de este experimento se muestran en las figuras 5.5.3 y 5.5.2. En la primera columna de la Figura 5.5.3 se muestran las aproximaciones del intervalo de confianza de la esperanza del error relativo para el experimento 1. Cada fila de la figura 5.5.3 corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden) y en cada gráfica se muestran los resultados de los distintos estimadores. En la Figura 5.5.2 se muestra la aproximación de la distribución del error relativo del experimento 1. Cada Gráfica de la Figura 5.5.2 muestra los resultados de cierto tipo de muestreo distinto cuyo tamaño es el $10\,\%$ del tamaño de la población.(muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden). En las figuras C0.2, C0.3 y C0.4 del Apéndice C se pueden ver más detalles de la Figura 5.5.2.

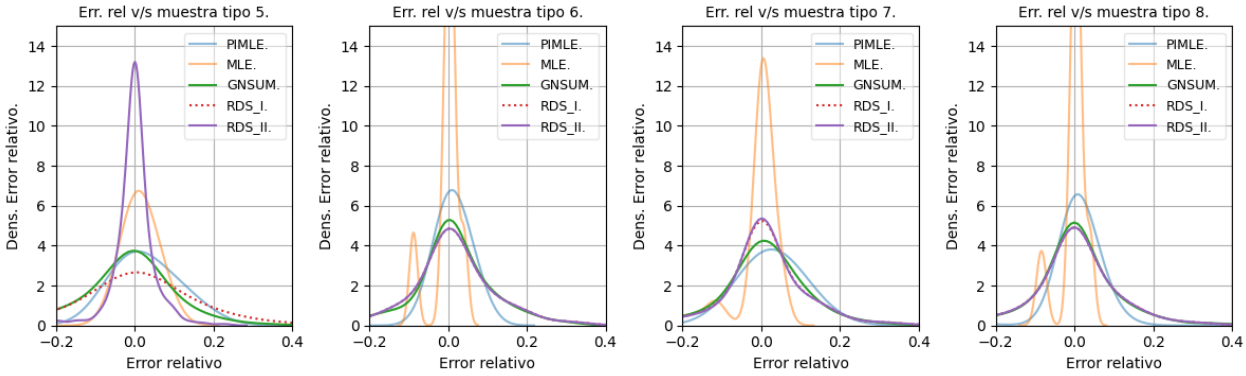


Figura 5.5.2: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población que muestra los resultados del experimento 1. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

Experimento 2**.**

En el caso del experimento 2 se realizaron estimaciones para la cantidad de usuarios con menos de 520 espectadores utilizando las muestras aleatorias almacenadas y los diferentes tipos de estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II).

Los resultados de este experimento se muestran en las figuras 5.5.3 y 5.5.4. En la segunda columna de la Figura 5.5.3 se muestran las aproximaciones del intervalo de confianza de la esperanza del error relativo para el experimento 2. En la Figura 5.5.4 se muestra la aproximación de la distribución del error relativo del experimento 2. Cada Gráfica de la Figura 5.5.4 muestra los resultados de cierto tipo de muestreo distinto cuyo tamaño es el 10 % del tamaño de la población.(muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden).

En las figuras C0.5, C0.6 y C0.7 del Apéndice C se pueden ver más detalles de la Figura 5.5.4.

Experimento 3**.**

En el caso del experimento 3 se realizaron estimaciones para la cantidad de usuarios no experimentados utilizando las muestras aleatorias almacenadas y los diferentes tipos de estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II).

Los resultados de este experimento se muestran en las figuras 5.5.7 y 5.5.5. En la primera columna de la Figura 5.5.7 se muestran las aproximaciones del intervalo de confianza

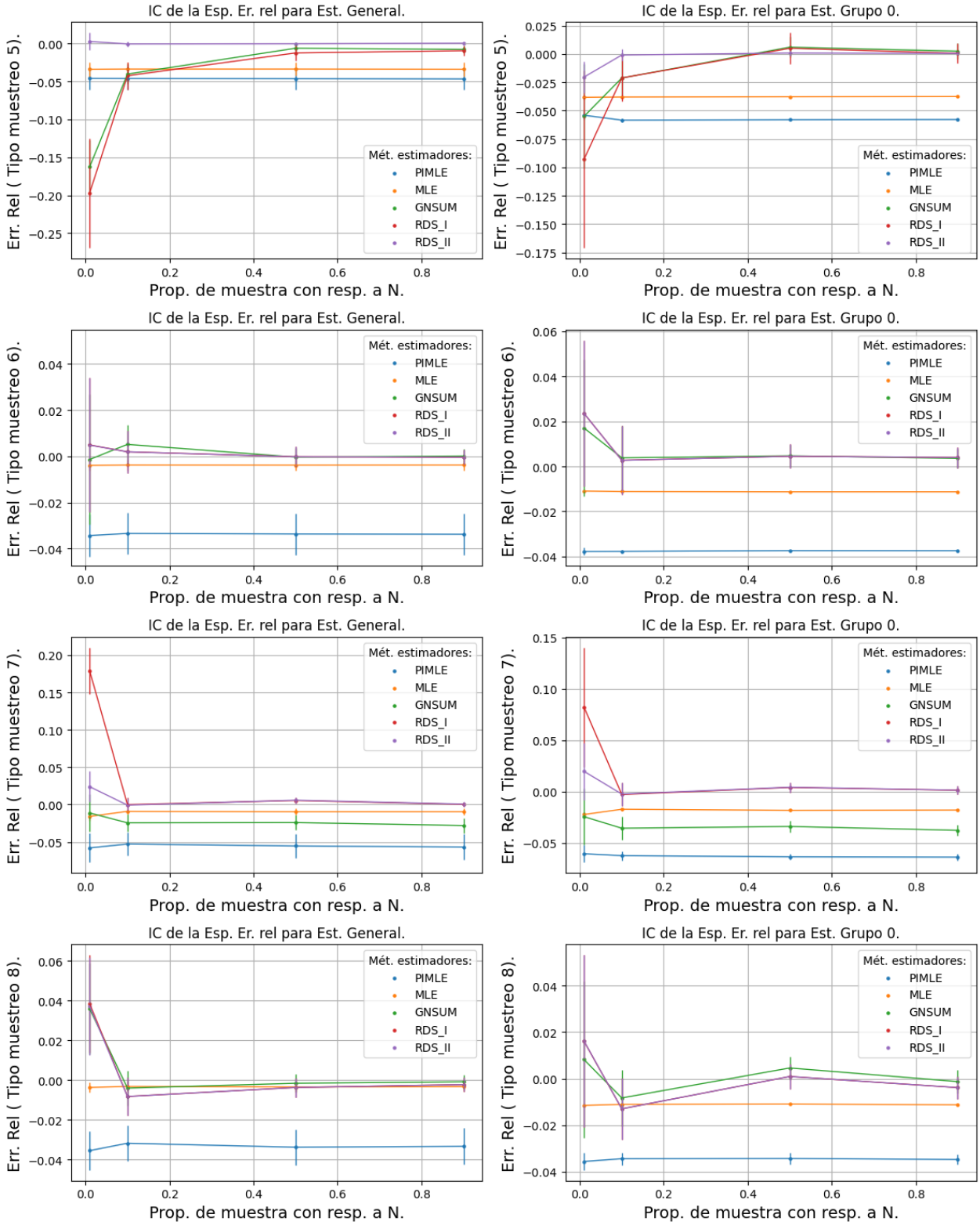


Figura 5.5.3: Resultados experimento 1 y 2: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia el grupo real a estimar.

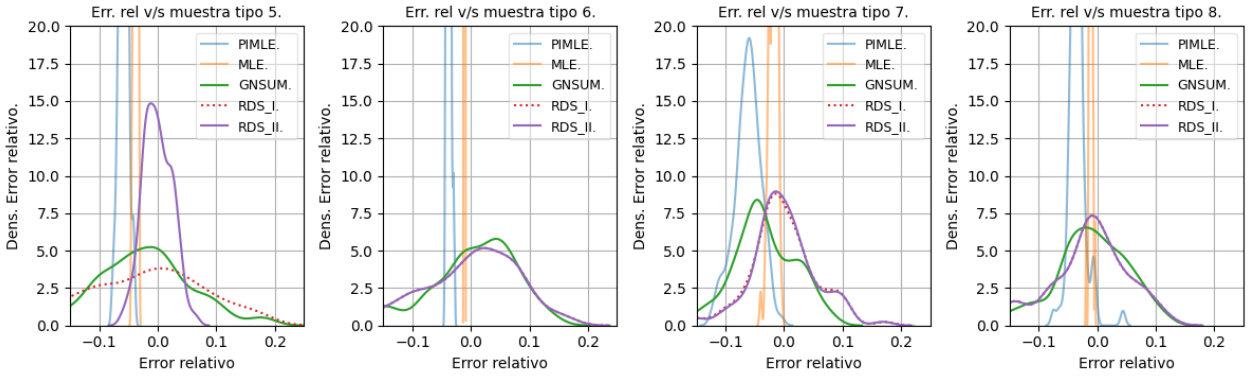


Figura 5.5.4: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población que muestra los resultados del experimento 2. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

de la esperanza del error relativo para el experimento 3. Cada fila de la figura 5.5.7 corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden) y en cada gráfica se muestran los resultados de los distintos estimadores. En la Figura 5.5.5 se muestra la aproximación de la distribución del error relativo del experimento 3. Cada Gráfica de la Figura 5.5.5 muestra los resultados de cierto tipo de muestreo distinto cuyo tamaño es el 10 % del tamaño de la población.(muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden).

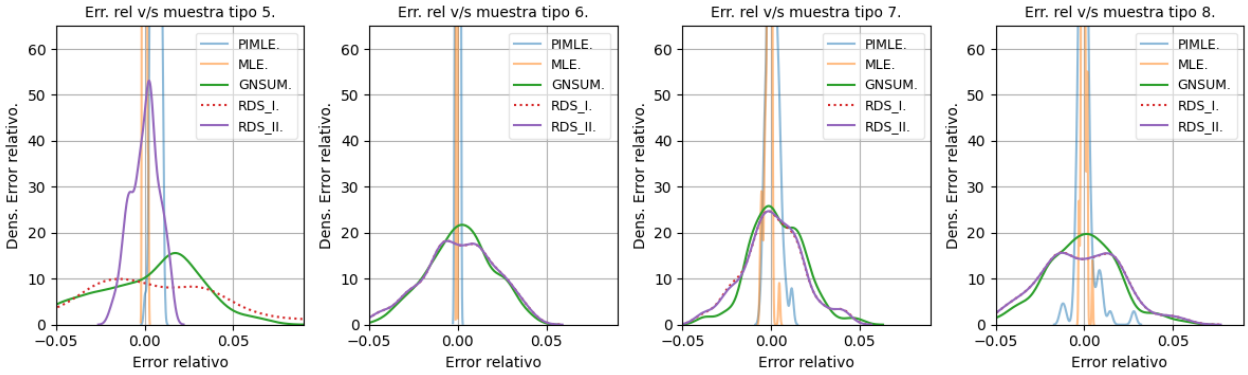


Figura 5.5.5: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población que muestra los resultados del experimento 3. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

En las figuras C0.9, C0.10 y C0.11 del Apéndice C se pueden ver más detalles de la Figura 5.5.5.

Experimento 4.
En el caso del experimento 4 se realizaron estimaciones para la cantidad de usuarios que tengan menos de 600(medida) de tiempo de vida (la tabla no especifica medida) utilizando las muestras aleatorias almacenadas y los diferentes tipos de estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II).

Los resultados de este experimento se muestran en las figuras 5.5.7 y 5.5.6. En la segunda columna de la Figura 5.5.7 se muestran las aproximaciones del intervalo de confianza de la esperanza del error relativo para el experimento 4.

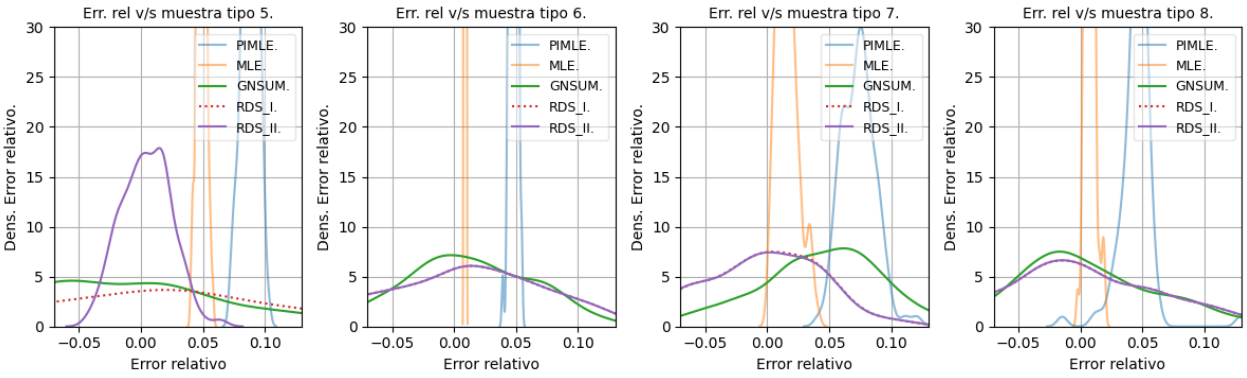


Figura 5.5.6: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población que muestra los resultados del experimento 4. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

En la Figura 5.5.6 se muestra la aproximación de la distribución del error relativo del experimento 4. Cada Gráfica de la Figura 5.5.6 muestra los resultados de cierto tipo de muestreo distinto cuyo tamaño es el 10 % del tamaño de la población.(muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden).

En las figuras C0.12, C0.13 y C0.14 del Apéndice C se pueden ver más detalles de la Figura 5.5.6.

Experimento 5.
En el caso del experimento 5 se realizaron estimaciones para la cantidad de usuarios

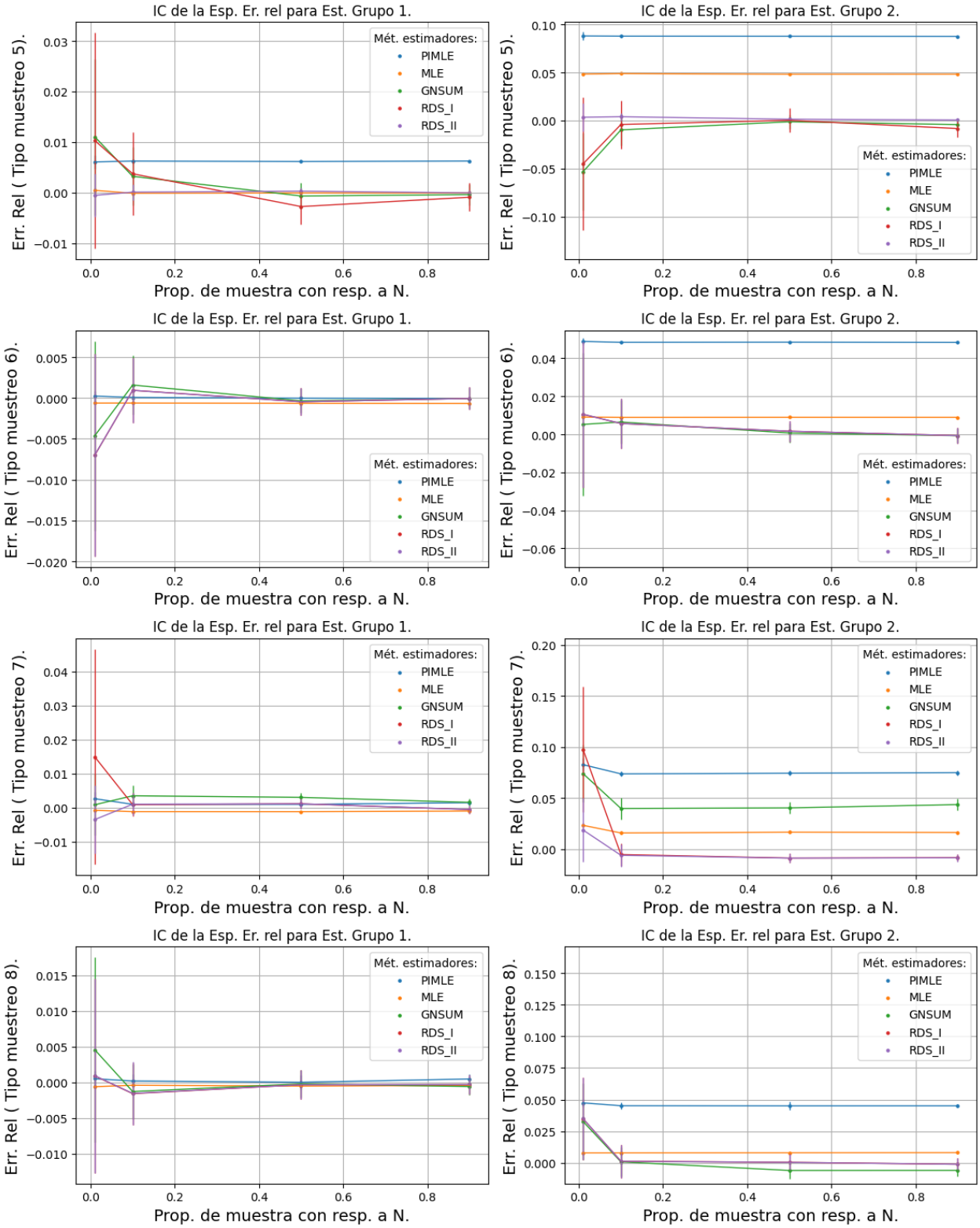


Figura 5.5.7: Resultados experimento 3 y 4: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia el grupo real a estimar.

inactivos utilizando las muestras aleatorias almacenadas y los diferentes tipos de estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II).

Los resultados de este experimento se muestran en las figuras 5.5.9 y 5.5.8. En la primera columna de la Figura 5.5.9 se muestran las aproximaciones del intervalo de confianza de la esperanza del error relativo para el experimento 5. Cada fila de la figura 5.5.9 corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden) y en cada gráfica se muestran los resultados de los distintos estimadores. En la Figura 5.5.8 se muestra la aproximación de la distribución del error relativo del experimento 5. Cada Gráfica de la Figura 5.5.8 muestra los resultados de cierto tipo de muestreo distinto cuyo tamaño es el 10 % del tamaño de la población.(muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden). En las figuras C0.16, C0.17 y C0.18 del Apéndice C se pueden ver más detalles de la Figura 5.5.8.

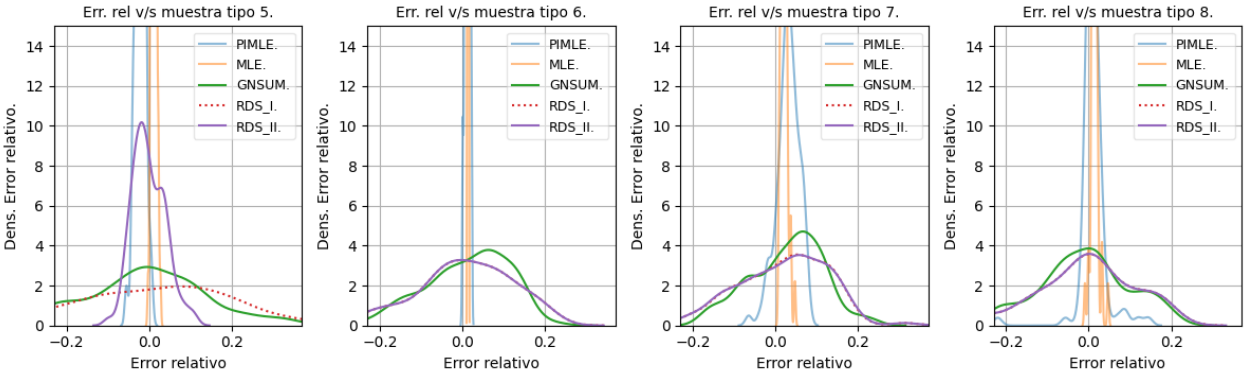


Figura 5.5.8: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población que muestra los resultados del experimento 5. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

Experimento 6.

En el caso del experimento 6 se realizaron estimaciones para la cantidad de usuarios de habla inglés utilizando las muestras aleatorias almacenadas y los diferentes tipos de estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II).

Los resultados de este experimento se muestran en las figuras 5.5.9 y 5.5.10. En la segunda columna de la Figura 5.5.9 se muestran las aproximaciones del intervalo de

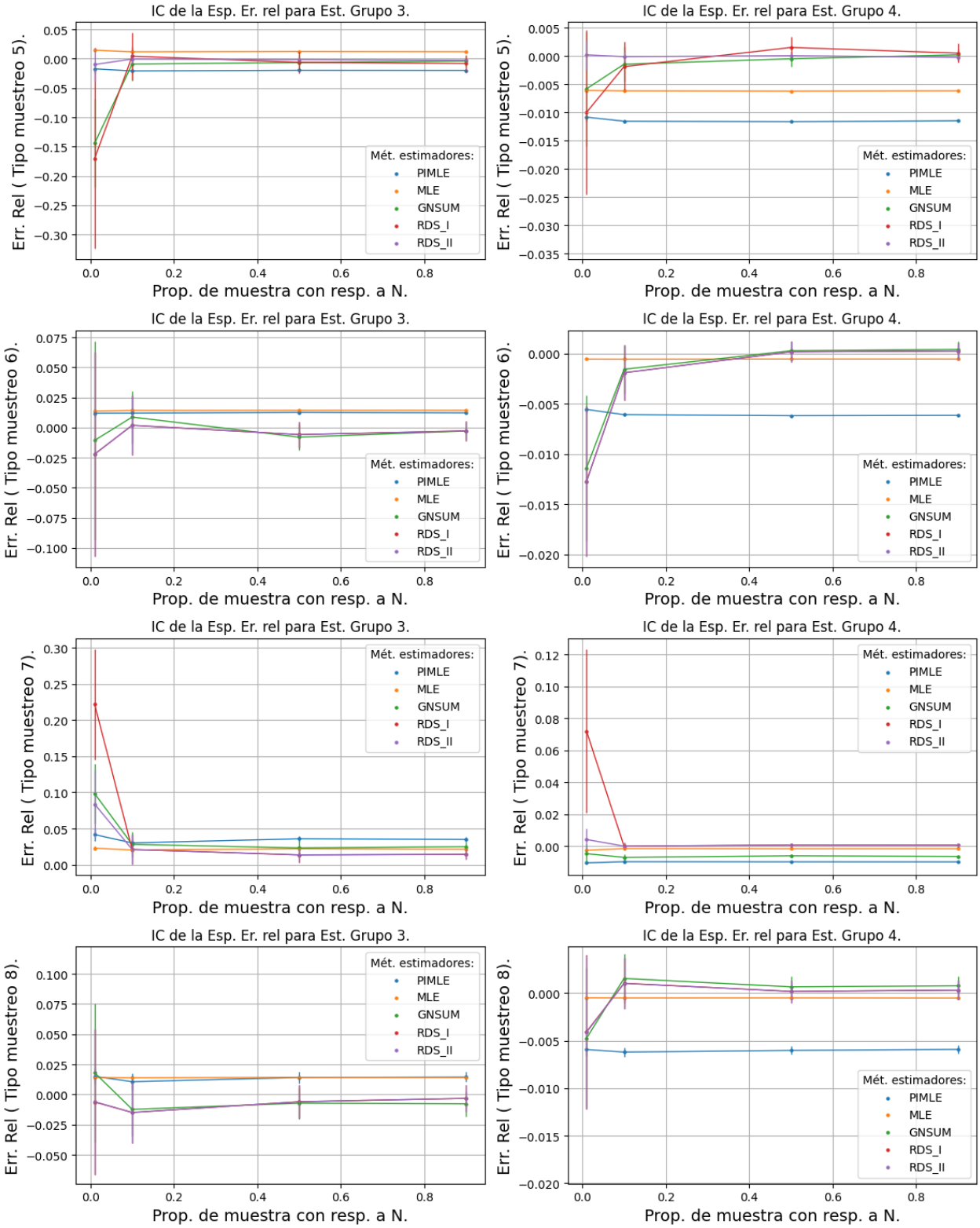


Figura 5.5.9: Resultados experimento 5 y 6: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia el grupo real a estimar.

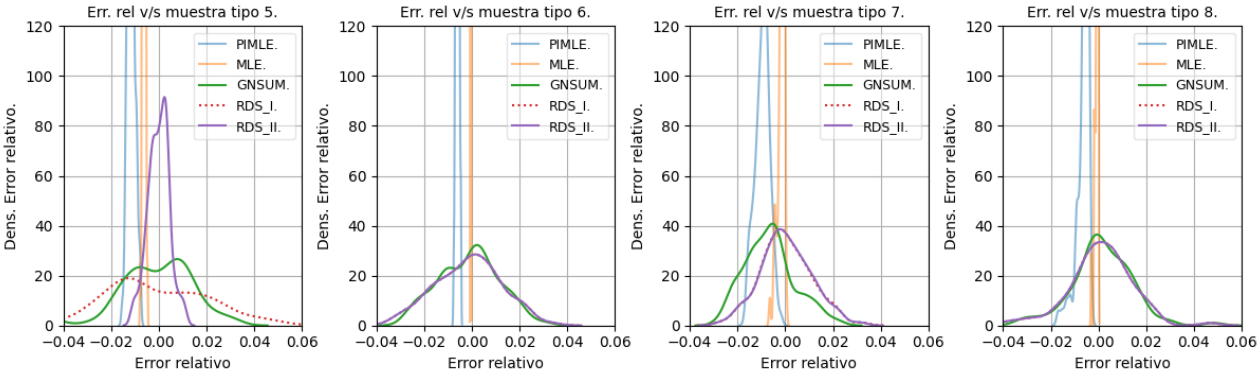


Figura 5.5.10: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población que muestra los resultados del experimento 6. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

confianza de la esperanza del error relativo para el experimento 6. En la Figura 5.5.10 se muestra la aproximación de la distribución del error relativo del experimento 6. Cada Gráfica de la Figura 5.5.10 muestra los resultados de cierto tipo de muestreo distinto cuyo tamaño es el 10 % del tamaño de la población.(muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden). En las figuras C0.19, C0.20 y C0.21 del Apéndice C se pueden ver más detalles de la Figura 5.5.10.

Experimento 7.

En el caso del experimento 7 se realizaron estimaciones para la cantidad de usuarios de habla francés utilizando las muestras aleatorias almacenadas y los diferentes tipos de estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II).

Los resultados de este experimento se muestran en las figuras 5.5.12 y 5.5.11. En la primera columna de la Figura 5.5.12 se muestran las aproximaciones del intervalo de confianza de la esperanza del error relativo para el experimento 7. Cada fila de la figura 5.5.12 corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden) y en cada gráfica se muestran los resultados de los distintos estimadores. En la Figura 5.5.11 se muestra la aproximación de la distribución del error relativo del experimento 7. Cada Gráfica de la Figura 5.5.11 muestra los resultados de cierto tipo de muestreo distinto cuyo tamaño es el 10 % del tamaño de la población.(muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico

y proceso

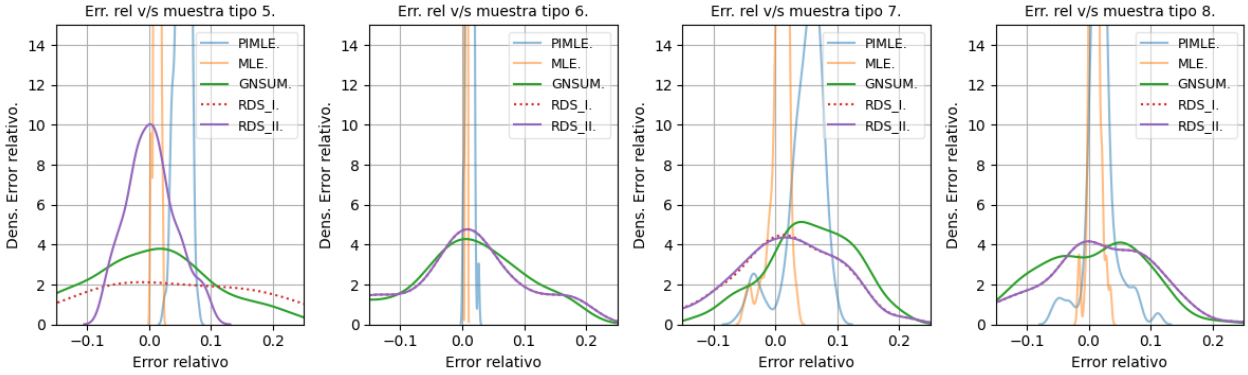


Figura 5.5.11: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población que muestra los resultados del experimento 7. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

viral constante, en ese orden). En las figuras C0.23, C0.24 y C0.25 del Apéndice C se pueden ver más detalles de la Figura 5.5.11.

Experimento 8.

En el caso del experimento 8 se realizaron estimaciones para la cantidad de usuarios de habla sur coreano utilizando las muestras aleatorias almacenadas y los diferentes tipos de estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II).

Los resultados de este experimento se muestran en las figuras 5.5.12 y 5.5.13. En la segunda columna de la Figura 5.5.12 se muestran las aproximaciones del intervalo de confianza de la esperanza del error relativo para el experimento 8. En la Figura 5.5.13 se muestra la aproximación de la distribución del error relativo del experimento 8. Cada Gráfica de la Figura 5.5.13 muestra los resultados de cierto tipo de muestreo distinto cuyo tamaño es el 10 % del tamaño de la población.(muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden). En las figuras C0.26, C0.27 y C0.28 del Apéndice C se pueden ver más detalles de la Figura 5.5.13.

Experimento 9.

En el caso del experimento 9 se realizaron estimaciones para la cantidad de usuarios de habla japonés utilizando las muestras aleatorias almacenadas y los diferentes tipos de estimadores de tamaño de grupo (PIMLE, MLE, GNSUM, RDS I y RDS II).

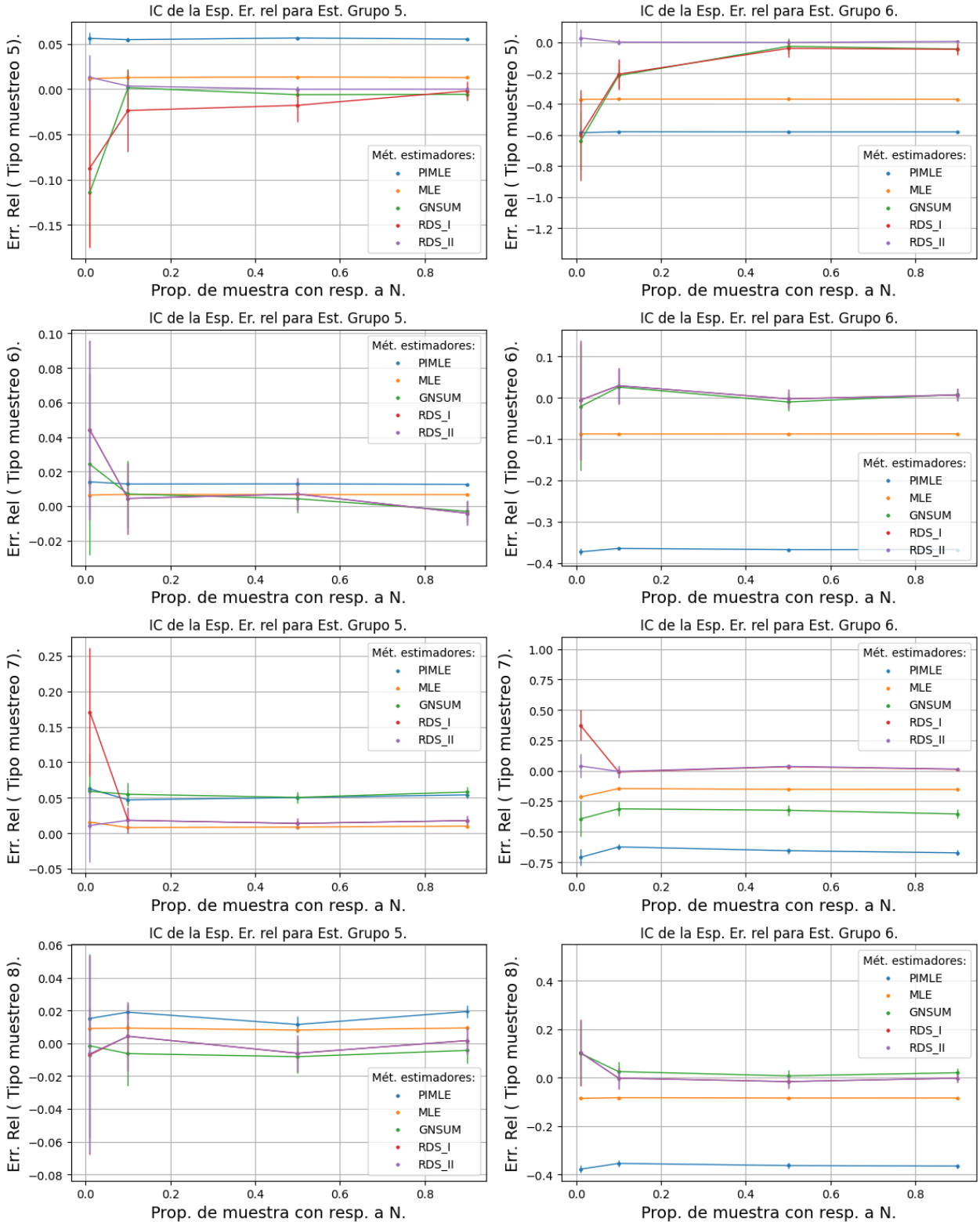


Figura 5.5.12: Resultados experimento 7 y 8: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia el grupo real a estimar.

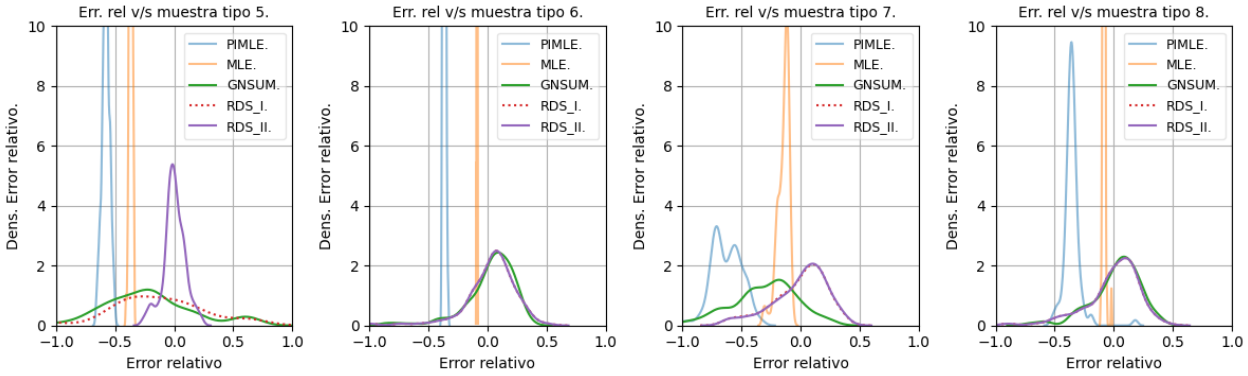


Figura 5.5.13: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población que muestra los resultados del experimento 8. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

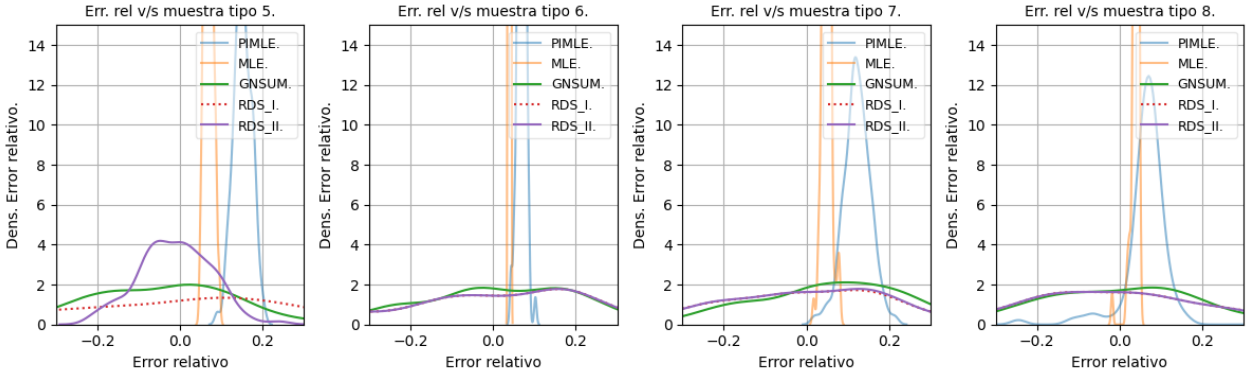


Figura 5.5.14: Gráficas que ilustran la distribución del error relativo, considerando un tamaño de muestra del 10 % del tamaño total de la población que muestra los resultados del experimento 9. Cada columna representa un tipo de muestreo distinto: muestreo uniforme, muestreo basado en caminata aleatoria, muestreo mediante proceso viral probabilístico y muestreo constante, en ese orden.

Los resultados de este experimento se muestran en las figuras 5.5.15 y 5.5.14. En la Figura 5.5.15 se muestran las aproximaciones del intervalo de confianza de la esperanza del error relativo para el experimento 9. Cada fila de la figura 5.5.15 corresponde a un tipo de muestreo distinto (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden) y en cada gráfica se muestran los resultados de los distintos estimadores. En la Figura 5.5.14 se muestra la aproximación de la distribución del error relativo del experimento 9. Cada Gráfica de la Figura 5.5.14 muestra los resultados de cierto tipo de muestreo

distinto cuyo tamaño es el 10 % del tamaño de la población (muestreo uniforme y los muestreos asociados a una caminata aleatoria, proceso viral probabilístico y proceso viral constante, en ese orden).

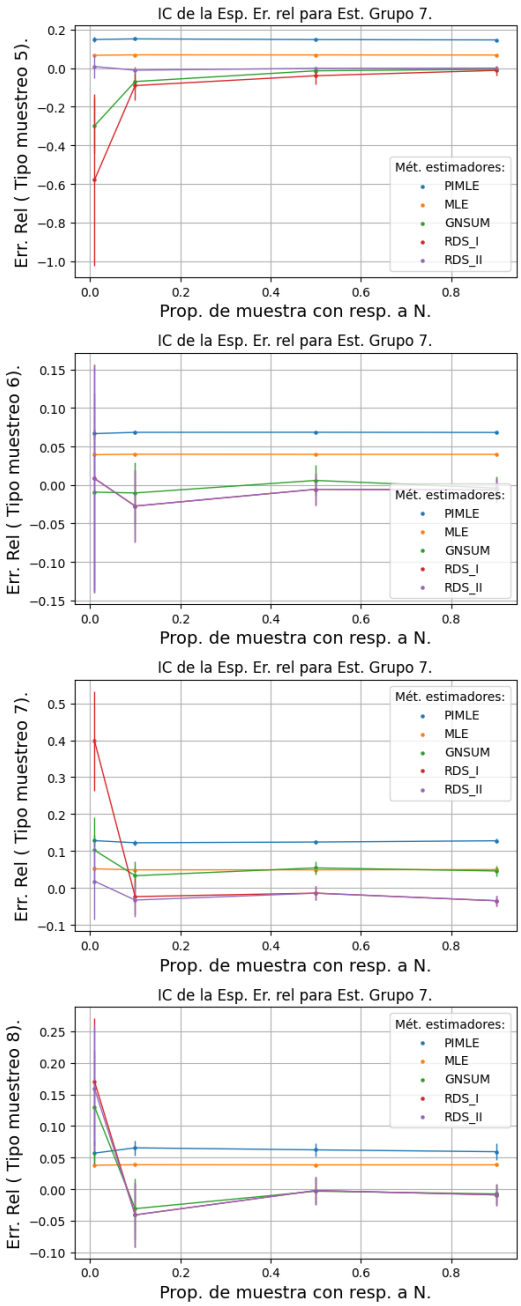


Figura 5.5.15: Resultados experimento 9: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia el grupo real a estimar.

En las figuras C0.30, C0.31 y C0.32 del Apéndice C se pueden ver más detalles de la Figura 5.5.14.

Análisis de los resultados de los experimentos.
De la Figura 5.5.2 se puede ver que los estimadores con mejor rendimiento general es el RDS II y el MLE. Algunas veces los tipos de muestreos aleatorios que sobresalen son el uniforme y otras veces el asociado a una caminata aleatoria, además si sobresale el muestreo aleatorio asociado a una caminata aleatoria entonces también tiene buen rendimiento el muestreo aleatorio asociado a un proceso viral constante. Además, la proporción ideal de la muestra es 10 %, pues centra las aproximaciones de la distribución del error relativo a su media de la muestra de proporción superior. De la Figura 5.5.3 podemos ver que la esperanza del error relativo en el caso de las estimaciones es general son cercanas a cero, pero el intervalo de confianza del estimador RDS I sigue siendo amplio, lo que implica que la estimación de la esperanza es de mala calidad para el estimador RDS I.

Los estimadores PIMLE y MLE en los experimentos del 1 al 9 tienen media no nula pero está cerca de este. De las estimaciones de la distribución de la densidad del error relativo para los estimadores PIMLE y MLE en los experimentos del 1 al 9 se puede ver que la varianza es muy pequeña cuando el tamaño de la muestra aleatorio supera o iguala al 10 % de la población, lo que implica que las estimaciones de estos estimadores son muy parecidos para distintas muestras con igual proporción. Los mejores resultados con estos estimadores se obtienen utilizando un muestreo aleatorio asociado a una caminata aleatoria. Cabe destacar que el rendimiento del estimador MLE es superior al estimador PIMLE, pero comparten muchas características.

El estimador GNSUM tiene una estimación de la distribución de densidad bastante buena para los experimentos del 1 al 9, en estos casos la media se podría considerar nula cuando se utiliza una proporción de muestreo del 10 % o superior. En general, para los experimentos del 1 al 9 el modelo GNSUM tiene los mejores resultados cuando se utiliza un muestreo aleatorio asociado a una caminata aleatoria, luego le sigue el muestreo aleatorio asociado a un proceso viral constante. La aproximación de la distribución de densidad del error relativo para los experimentos del 1 al 9 muestra que no existe una cantidad considerable de varianza en las estimaciones.

El estimador RDS I tiene una estimación de la distribución de densidad bastante buena y la media se podría considerar nula para los experimentos del 1 al 9 cuando la proporción

del tamaño del muestreo es como mínimo del 10 %. En general, para los experimentos del 1 al 9 el muestreo aleatorio asociado a una caminata aleatoria y el asociado a un proceso viral probabilístico tiene los mejores resultados. De la aproximación del intervalo de confianza podemos notar que la varianza de este estimador en los experimentos del 1 al 9 es bastante grande y que la estimación de la esperanza del error relativo podría no ser tan precisa como los estimadores PIMLE, MLE ó GNSUM.

El estimador RDS II tiene una estimación de la distribución de densidad bastante buena para los experimentos del 1 al 9, en estos casos la media es nula cuando se utiliza una proporción de muestreo del 10 % o superior. En general, para los experimentos del 1 al 9 el estimador RDS II cuando utiliza el muestreo aleatorio uniforme con repetición se obtienen los mejores resultados posibles. La aproximación de la distribución de densidad del error relativo para los experimentos del 1 al 9 muestra que no existe una cantidad considerable de varianza en las estimaciones.

Sorprendentemente los estimadores PIMLE y MLE funcionan mucho mejor con el grafo de twitch que con los grafos generados. La precisión de estos estimadores es muy superior a los otros estimadores pero definitivamente su error relativo no tiene media nula. El estimador que les sigue en rendimiento es el RDS II pues tiene más precisión que el GNSUM y el RDS I, y la media del error relativo del estimador RDS II es nula.

Capítulo 6

Conclusiones

De la etapa de experimentación en donde estudiamos la distribución de grados de la muestra con su respectiva distribución de grados esperada, en los experimentos con grafos generados y grafos reales, vimos que la proporción del tamaño de muestra debe ser como mínimo del 10 % del tamaño de población para obtener buenos resultados con los estimadores propuestos. Esto se puede ver considerando el hecho que una muestra aleatoria para que sea representativa a la población muestreada debe tener cierto tamaño, y para cumplir esto, la muestra aleatoria debe por lo menos parecerse a su distribución esperada. Un ejemplo de esto podría ser el hecho obtener una aproximación del promedio de notas de un curso a través de un muestreo aleatorio uniforme con repetición. Si la muestra es muy pequeña podemos sólo considerar alumnos con notas bajas o alumnos con notas altas, lo que implica una mala estimación.

En el caso de la experimentación con grafos generados, el estimador con mejor calidad en estimación es el GNSUM, el que le sigue es el RDS II, luego el RDS I y por últimos los MLE y PIMLE, en un sentido general. Además, en los estimadores PIMLE y MLE su rendimiento no se ve afectado por el tipo de muestreo utilizado. GNSUM, RDS I tiene un rendimiento superior cuando se utiliza con los muestreos de tipo asociados a una caminata aleatorio o un proceso viral constante. Mientras que el estimador RDS II sus mejores estimaciones se obtienen cuando se utiliza un muestreo aleatorio uniforme con repetición.

En el caso de la experimentación con el grafo de Twitch, los estimadores con mejor calidad de estimación son el MLE, el PIMLE y el RDS II, en ese orden. Sin embargo, los estimadores PIMLE y MLE no tienen media nula, pero si muy cercano a este

valor. Lo que hace que los estimadores MLE y PIMLE sean buenos estimadores es su baja varianza de su estimación, es decir para muestras aleatorias distintas de igual proporción sus estimaciones son muy cercanas. El alto rendimiento de los estimadores MLE y PIMLE puede deberse a la distribución de grados de H en comparación con la distribución de grados de la población. Cuando generamos los grafos y los conjuntos H a estimar su tamaño, notamos que cuando H se generaba a partir de un muestreo proporcional al grado e inversamente proporcional al grado la media del error relativo negativo y positivo respectivamente, esto es debido a que el modelo PIMLE y MLE fija una ponderación de estimación más alta para los nodos de alto grado. De esta forma, si las comunidades de Twitch podrían ser formados por una selección de tipo uniforme con respecto al grado, entonces se explicaría el alto rendimiento de los modelos PIMLE y MLE.

El estimador RDS II también obtuvo buenos resultados, pues su media de error relativo en los experimentos con el grafo de Twitch fue nula. El rendimiento superior del RDS II por sobre el GNSUM se puede deber a que la proporción de la cantidad de aristas y cantidad de nodos al cuadrado era mucho más pequeña que en los experimentos con grafos generados. El modelo GNSUM se basa en la extracción de información de los nodos vecinos de los nodos muestreado, como el grafo de Twitch tiene nodos de bajo grado entonces no se puede extraer mucha información de la vecindad de los nodos muestreados, haciendo que la estimación del modelo GNSUM se vea afectada negativamente.

Recordando las definiciones de encuesta directa, encuesta indirecta y encuesta indirecta exclusiva, los estimadores que utilizan encuestas indirectas exclusiva son el PIMLE y el MLE, los modelos GNSUM y RDS I utiliza una encuesta indirecta, y el que utiliza encuesta de tipo directa es el RDS II. Cabe mencionar que el modelo GNSUM necesita la información de toda la vecindad de cada nodo muestreado, mientras que el modelo RDS I sólo necesita conocer la información de un vecino aleatorio a cada nodo muestreado. **En conclusión**, los métodos asociados a encuestas indirectas exclusiva están muy por detrás en rendimiento que los métodos asociados a las encuestas directas e indirecta cuando se trabaja en grafos escala libre. Pero, si trabajamos con un grafo cuya distribución de grados es muy parecida a la de un grafo aleatorio, entonces los modelos indirectos exclusivos son la mejor opción. Por lo tanto, en el caso de tener que elegir un método de estimación primero debemos notar si la distribución de grados del grafo en donde se desea estimar un subconjunto de nodos es como un grafo aleatorio

o parecido. Si el grafo problema es parecido a un grafo aleatorio, entonces debemos utilizar una estimación con el modelo MLE, pero si no es el caso entonces debemos usar el estimador GNSUM o RDS II. Recordar que podemos estudiar la distribución de grados de un grafo a través de una muestra. Si la muestra es uniforme con repetición debemos graficar la distribución de grados de la muestra y está tiende a la distribución de grados del grafo. Si la muestra aleatoria está asociada a una caminata aleatoria, un proceso viral probabilístico o un proceso viral constante, entonces debemos realizar el ajuste mostrado en la Ecuación 3.3.2 y luego graficar la distribución de grados ajustada, obteniendo una aproximación de la distribución de grados del grafo. En el caso que no se pueda realizar un estudio de la distribución de grados del grafo problema, entonces el estimador más confiable para realizar la estimación sería el modelo RDS II.

La elección del tipo de muestreo a utilizar va a depender del modelo que usaremos. Por lo tanto, si realizaremos las estimaciones con el modelo PIMLE ó el MLE entonces debemos utilizar un muestreo asociado a una caminata aleatoria. En el caso del modelo GNSUM se debe usar un muestreo asociado a una caminata aleatoria como primera opción, sino no disponemos de este podemos utilizar un muestreo aleatorio uniforme con repetición o un muestreo aleatorio asociado a un proceso viral constante. Si queremos realizar estimaciones con el modelo RDS I podemos utilizar todos los muestreos aleatorios definidos anteriormente a excepción del muestreo uniforme con repetición, pues con este método se obtienen los peores resultados. Por último, el estimador RDS II obtiene sus mejores resultados con el muestreo aleatorio uniforme con repetición.

Entre los trabajos que se podrían profundizar seria el estudio de los estimadores cuando existan en las encuestas falsos positivos, falsos negativos, o incluso una mezcla de falsos positivos y falsos negativos. También merece la pena realizar un estudio con el objetivo de relajar las condiciones del método estimador GNSUM con respecto a la información requerida de la vecindad cada nodo muestreado. Para esto podría considerarse sólo en solicitar información de un porcentaje o una cantidad fija de la vecindad de cada nodo muestreado. Otra cosa interesante, sería encontrar una mejor aproximación de la probabilidad de selección de los muestreos asociados al proceso viral probabilístico y proceso viral constante. También, se podría considerarse el estudiar a mayor profundidad los contextos en que el modelo MLE y PIMLE obtienen buenas estimaciones y establecer medidas relacionadas al contexto. Por último, encontrar de mejor manera la proporción de nodos ideal para que equilibrar entre la calidad de la estimación y el costo de realizar una muestra aleatoria.

Bibliografía

[1] R. Arnab. *Survey sampling theory and applications*. Academic Press, 2017.

[2] H. R. Bernard, E. C. Johnsen, and P. D. Killworth. Estimating the size of an average personal network and of an event subpopulation. *The Small World*, 1987.

[3] D. M. Feehan and M. J. Salganik. Generalizing the network scale-up method: a new estimator for the size of hidden populations. *Sociological methodology*, 46(1): 153–186, 2016.

[4] S. L. Feld. Why your friends have more friends than you do. *American journal of sociology*, 96(6):1464–1477, 1991.

[5] E. Goldstein B. Estimaciones de la prevalencia y evolución del vih/sida en chile. Technical report, Departamento de Estudios, Extensión y Publicaciones, abril 2018. Informe Técnico.

[6] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using networkx. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[7] P. D. Killworth, E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. A social network approach to estimating seroprevalence in the united states. *Social networks*, 20(1):23–50, 1998.

[8] P. D. Killworth, C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. *Evaluation review*, 22(2):289–308, 1998.

[9] I. Laga, L. Bao, and X. Niu. Thirty years of the network scale-up method. *Journal of the American Statistical Association*, 116(535):1548–1559, 2021.

[10] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[11] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[12] M. E. Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1): 016128, 2002.

[13] S. G. Prabhu-Ajgaonkar. Comparison of the horvitz-thompson strategy with the hansen-hurwitz strategy. *Survey Methodology*, page 221, 1987.

[14] L. Rodero-Merino, A. F. Anta, L. López, and V. Cholvi. Performance of random walks in one-hop replication networks. *Computer Networks*, 54(5):781–796, 2010.

[15] B. Rozemberczki and R. Sarkar. Twitch gamers: a dataset for evaluating proximity preserving and structural role-based node embeddings, 2021.

[16] M. J. Salganik and D. D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, 34(1): 193–240, 2004.

[17] P. J. Sandiford. A new binomial approximation for use in sampling from finite populations. *Journal of the American Statistical Association*, 55(292):718–722, 1960.

[18] E. Volz and D. D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of official statistics*, 24(1):79, 2008.

Apéndice A

Gráficos de muestreos aleatorios

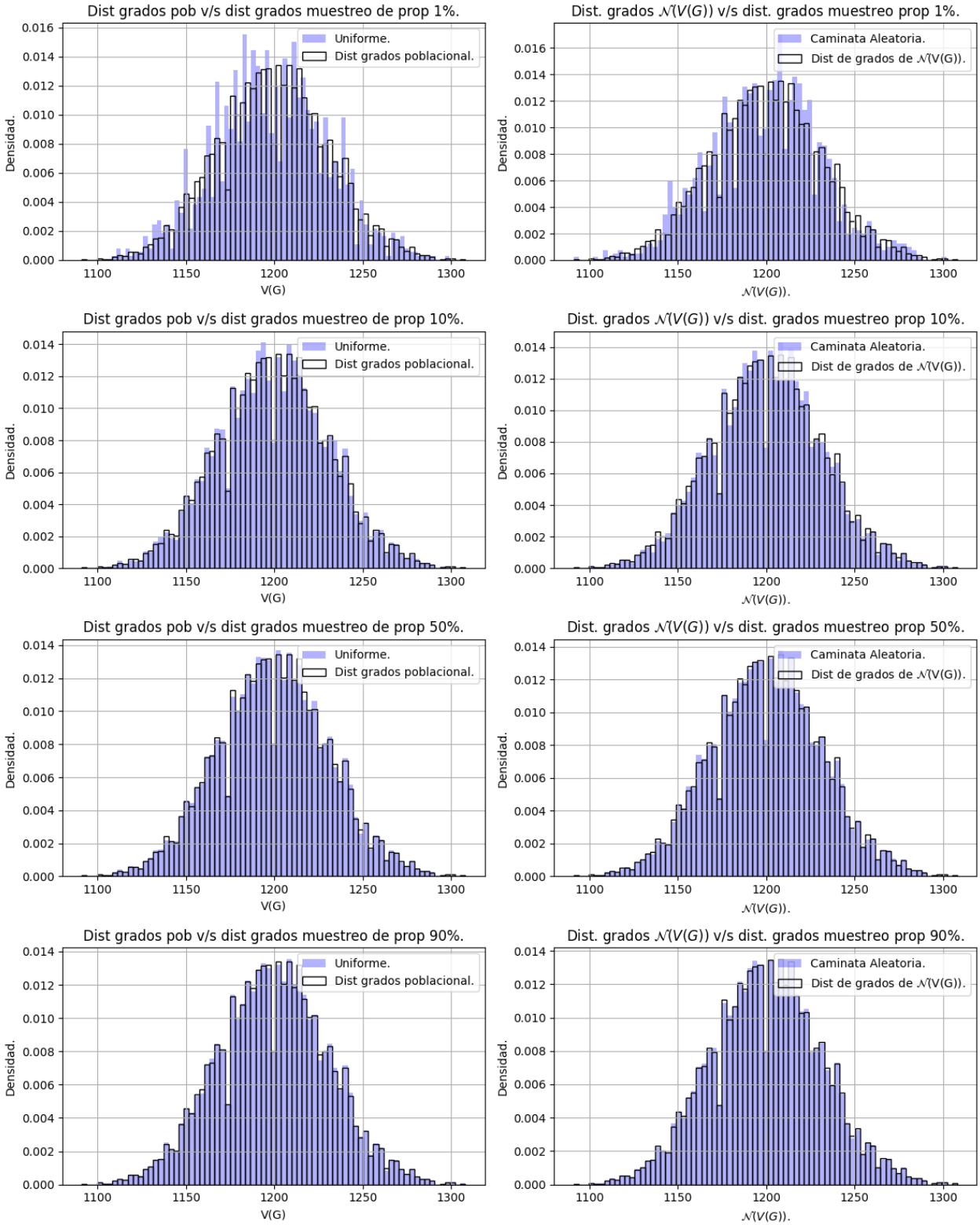


Figura A0.1: Gráficas de distribución de grados de $V(G)$ v/s distribución de grados de un muestreo uniforme variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a una caminata aleatoria variando su proporción de tamaño en un grafo aleatorio pesado.

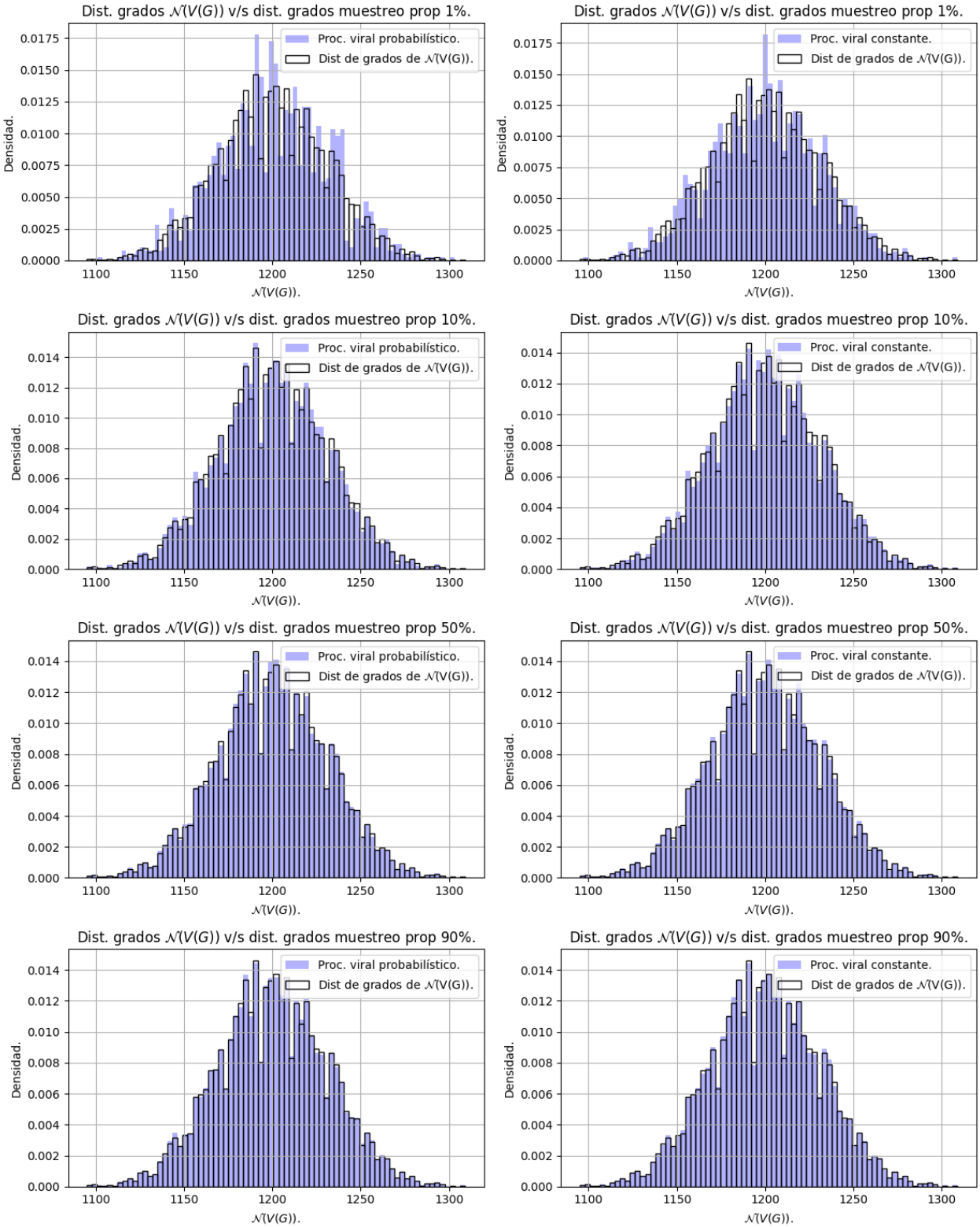


Figura A0.2: Gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral probabilístico variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral constante variando su proporción de tamaño de un grafo aleatorio pesado.

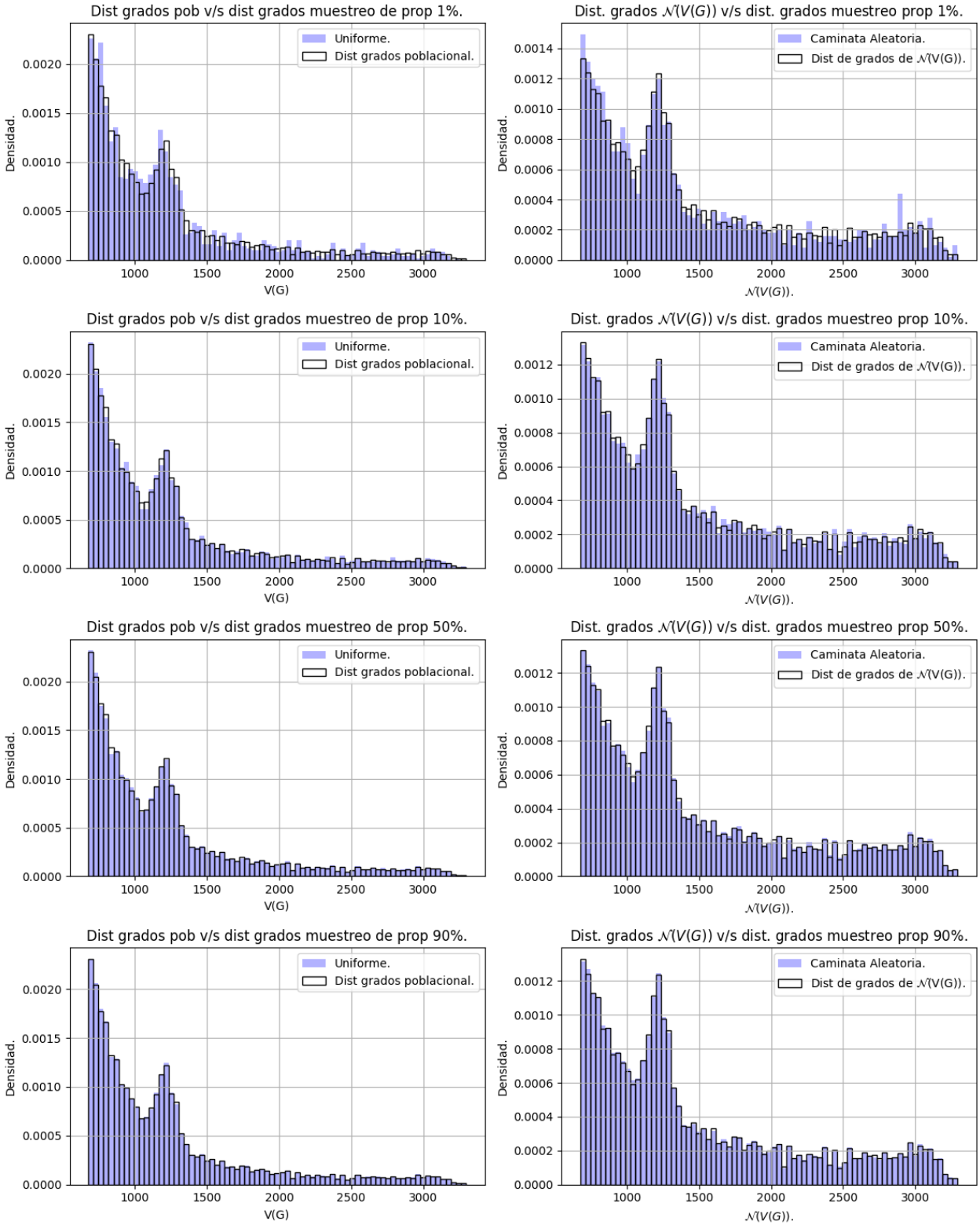


Figura A0.3: Gráficas de distribución de grados de $V(G)$ v/s distribución de grados de un muestreo uniforme variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a una caminata aleatoria variando su proporción de tamaño en un grafo escala pesado.

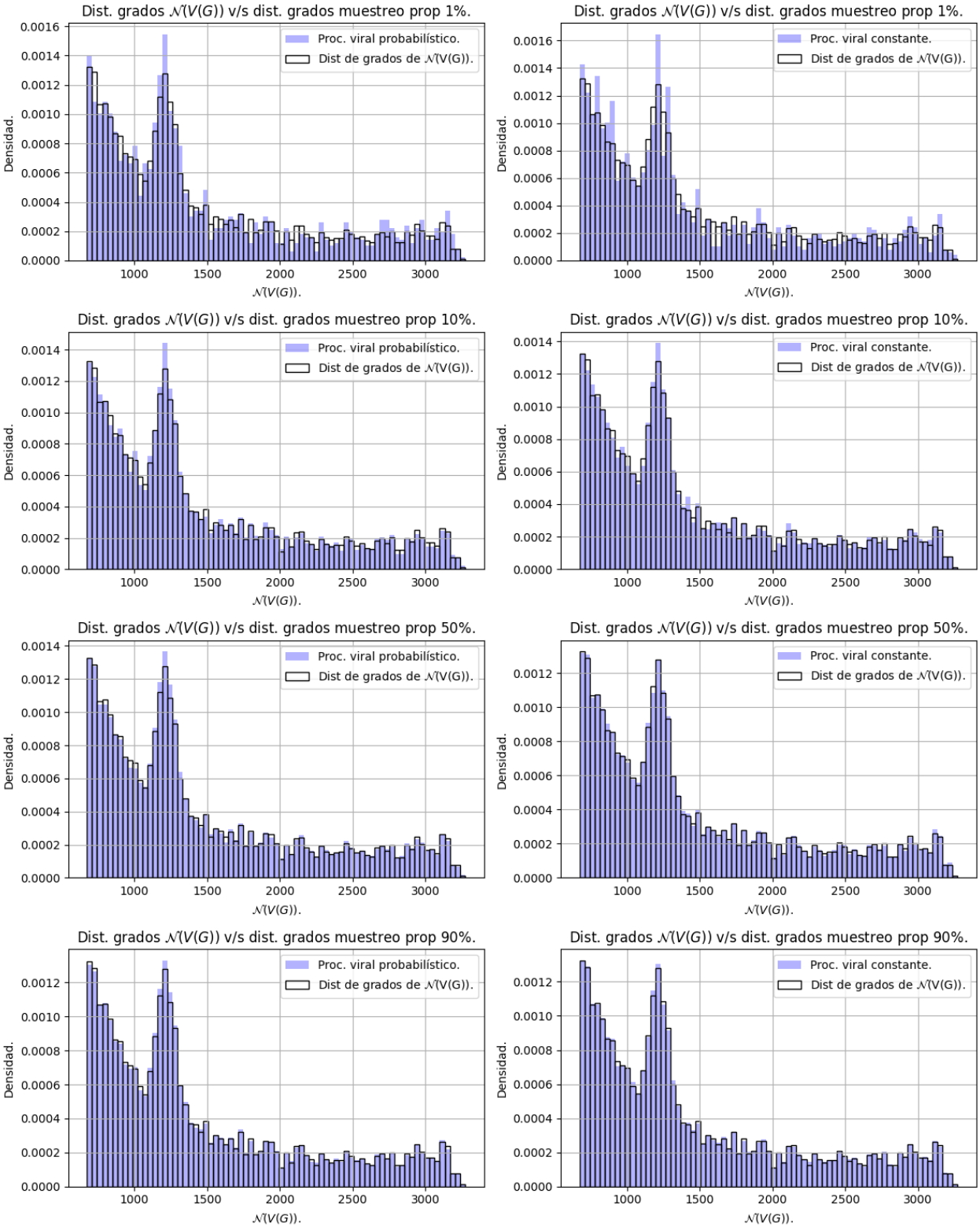


Figura A0.4: Gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral probabilístico variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral constante variando su proporción de tamaño de un grafo escala pesado.

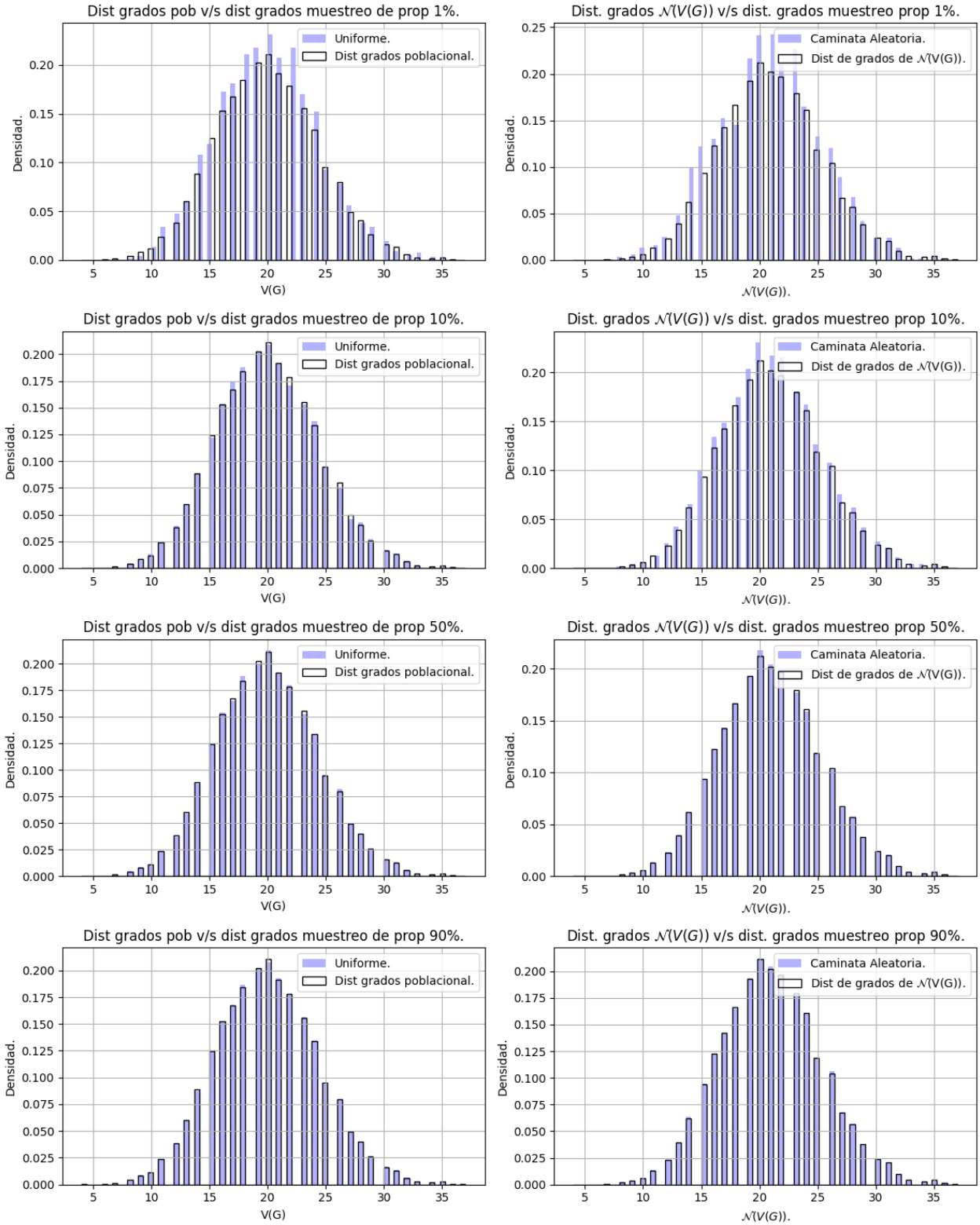


Figura A0.5: Gráficas de distribución de grados de $V(G)$ v/s distribución de grados de un muestreo uniforme variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a una caminata aleatoria variando su proporción de tamaño en un grafo aleatorio medio.

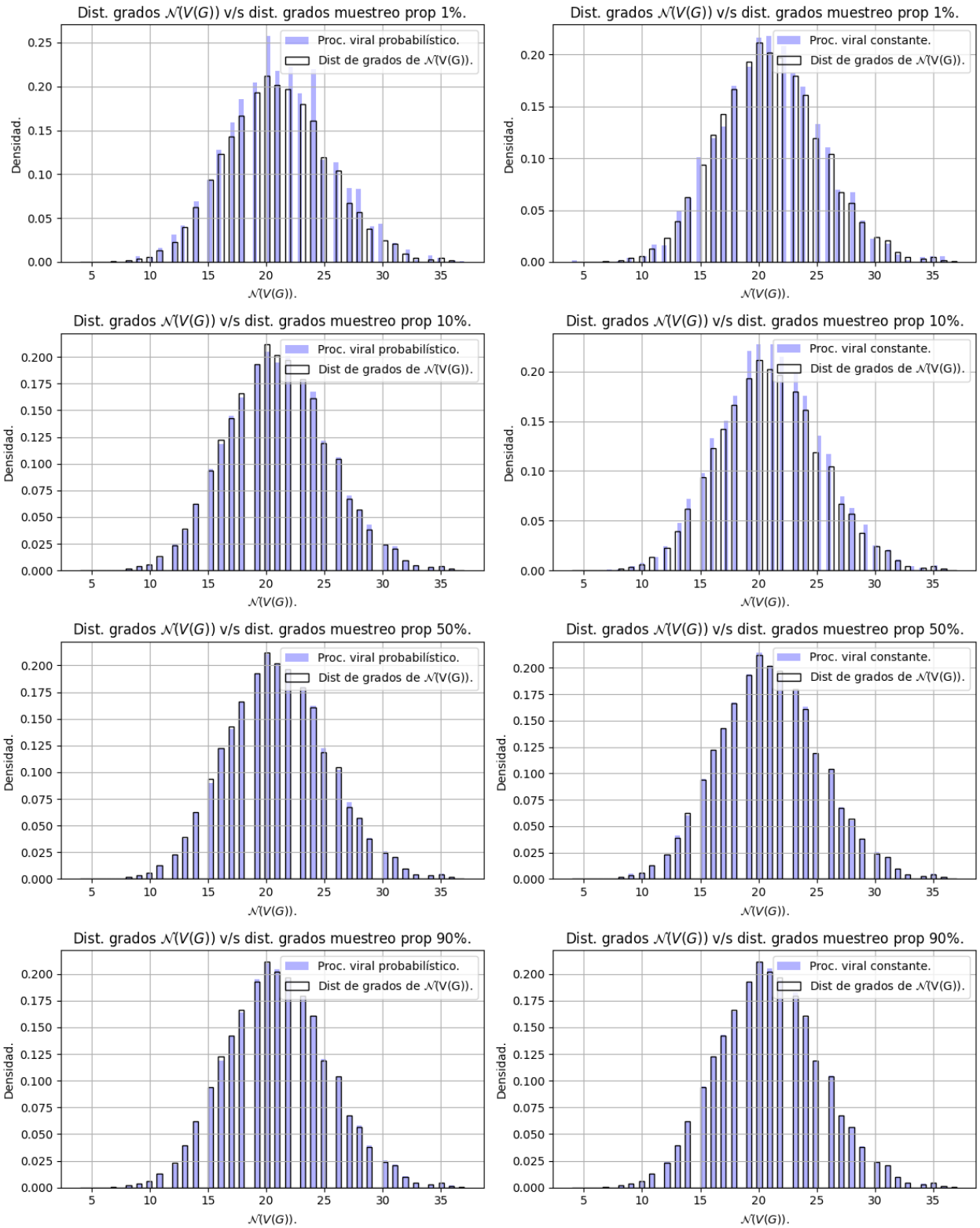


Figura A0.6: Gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral probabilístico variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral constante variando su proporción de tamaño de un grafo aleatorio medio.

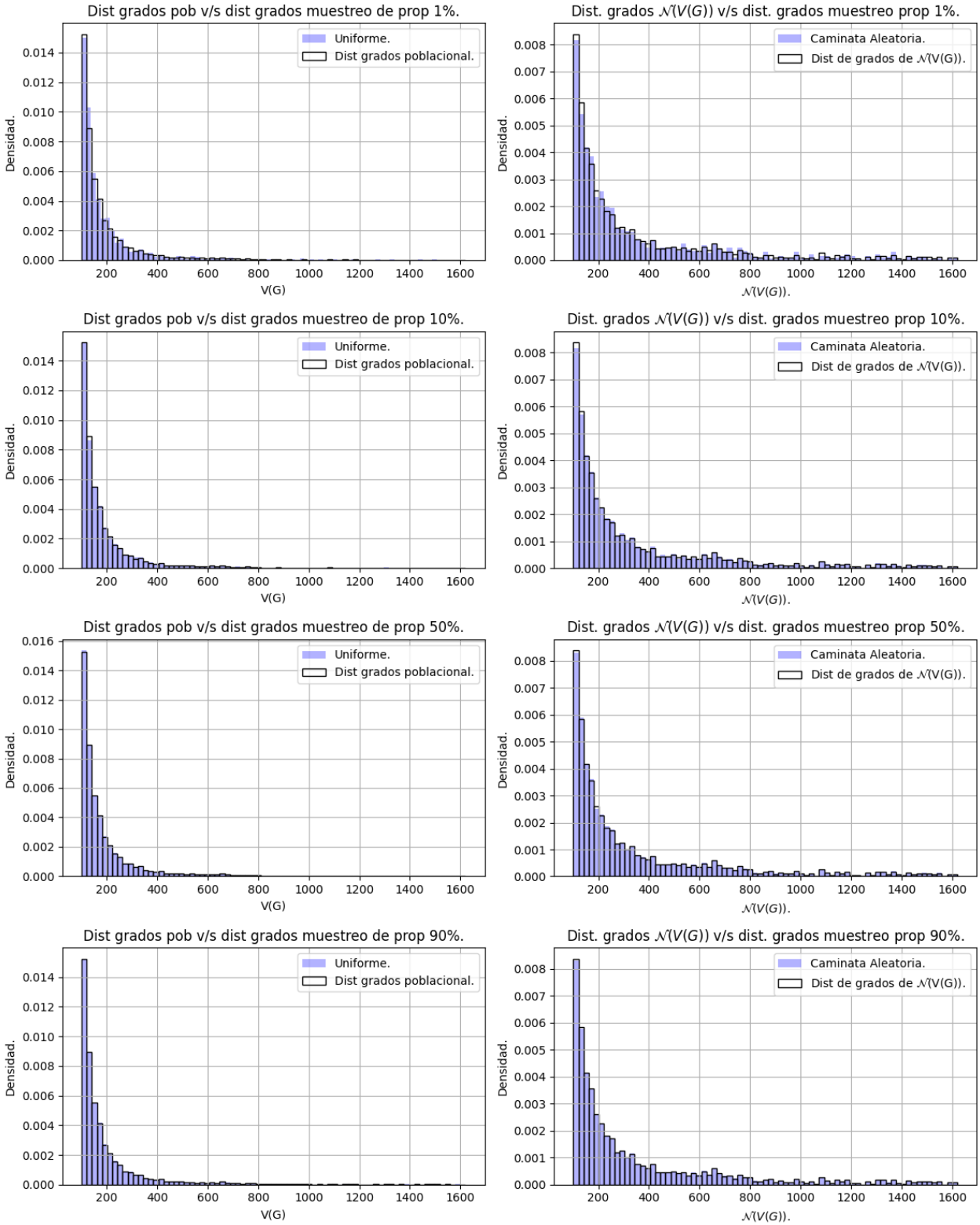


Figura A0.7: Gráficas de distribución de grados de $V(G)$ v/s distribución de grados de un muestreo uniforme variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a una caminata aleatoria variando su proporción de tamaño en un grafo escala medio.

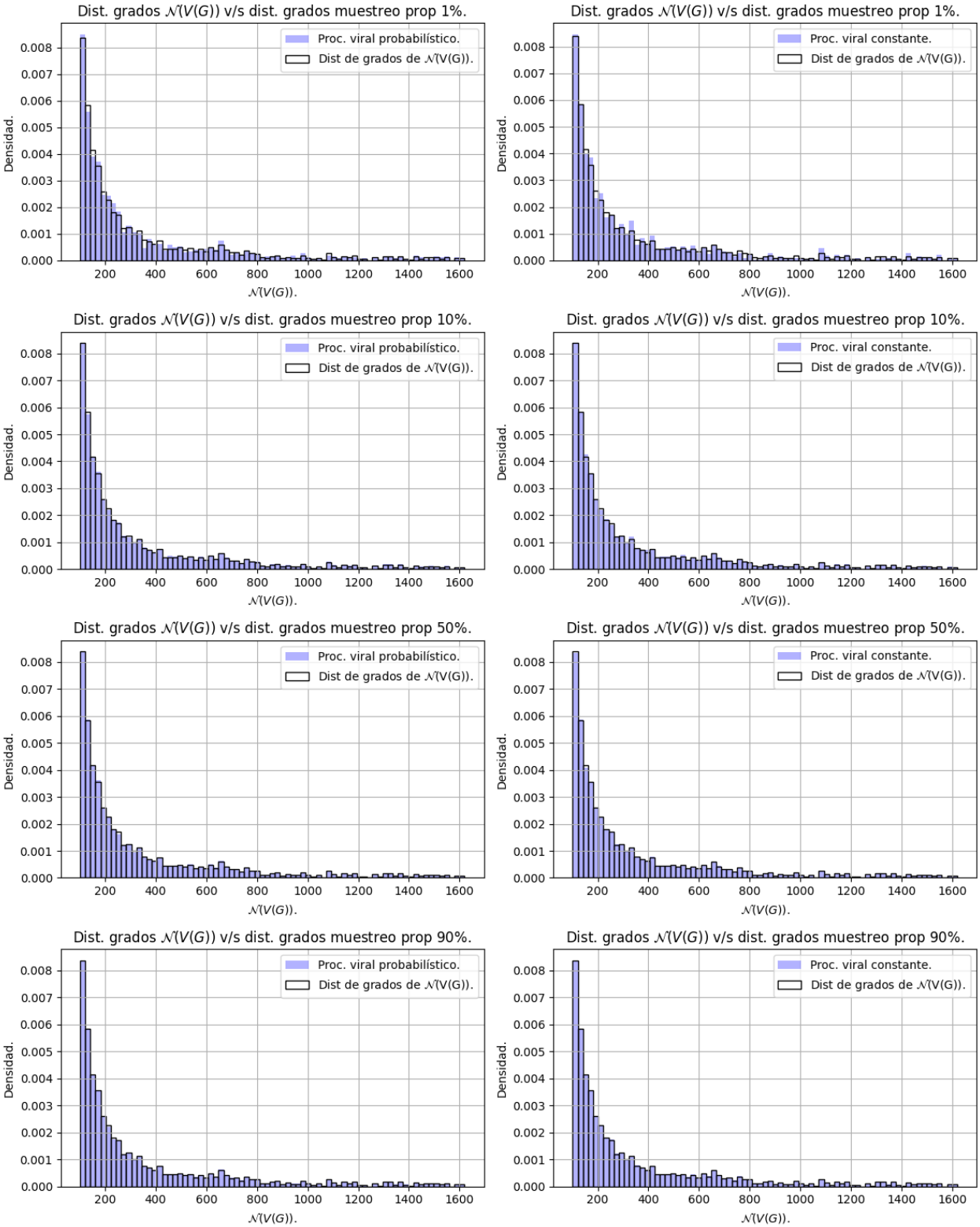


Figura A0.8: Gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral probabilístico variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral constante variando su proporción de tamaño de un grafo escala medio.

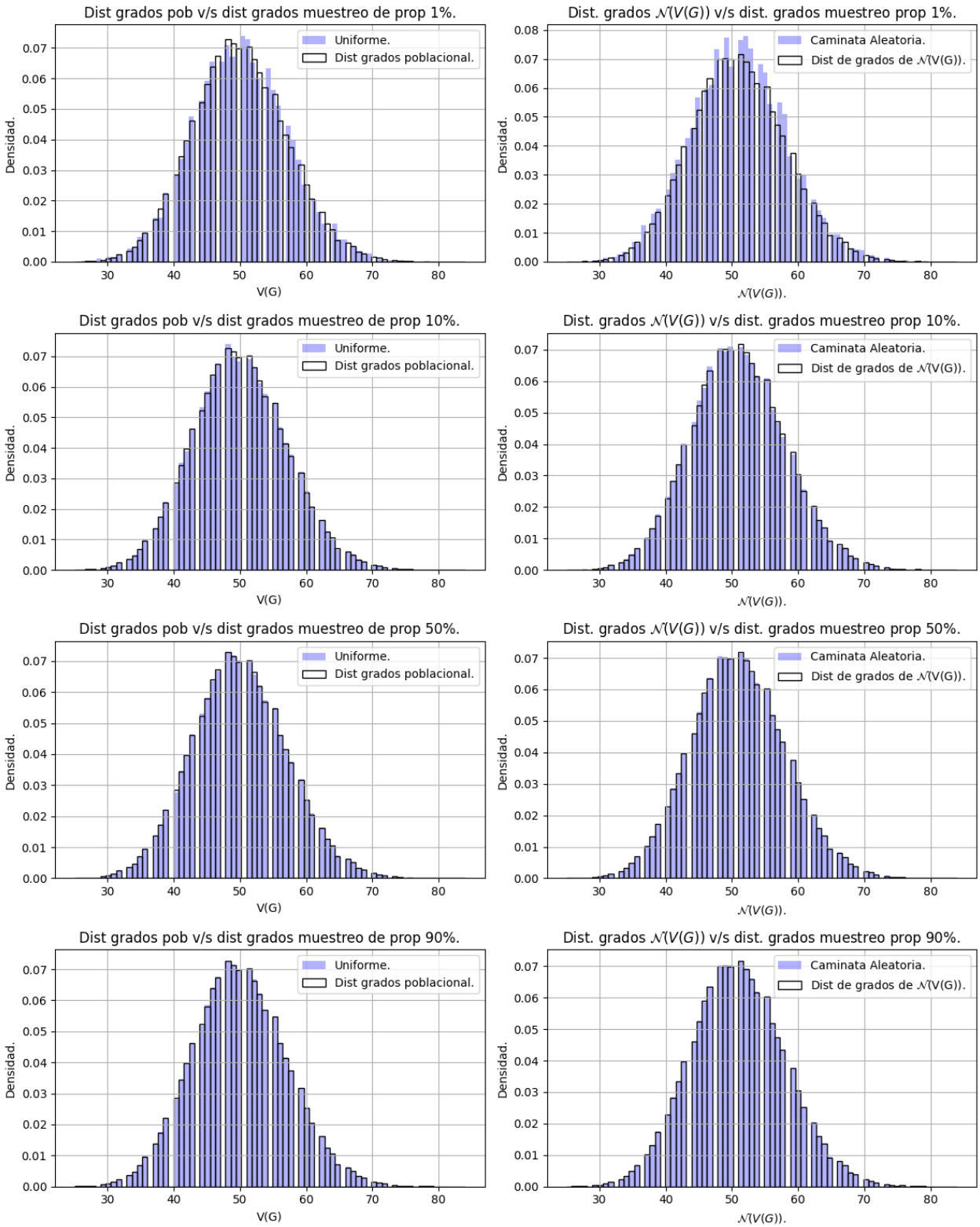


Figura A0.9: Gráficas de distribución de grados de $V(G)$ v/s distribución de grados de un muestreo uniforme variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a una caminata aleatoria variando su proporción de tamaño en un grafo aleatorio ligero.

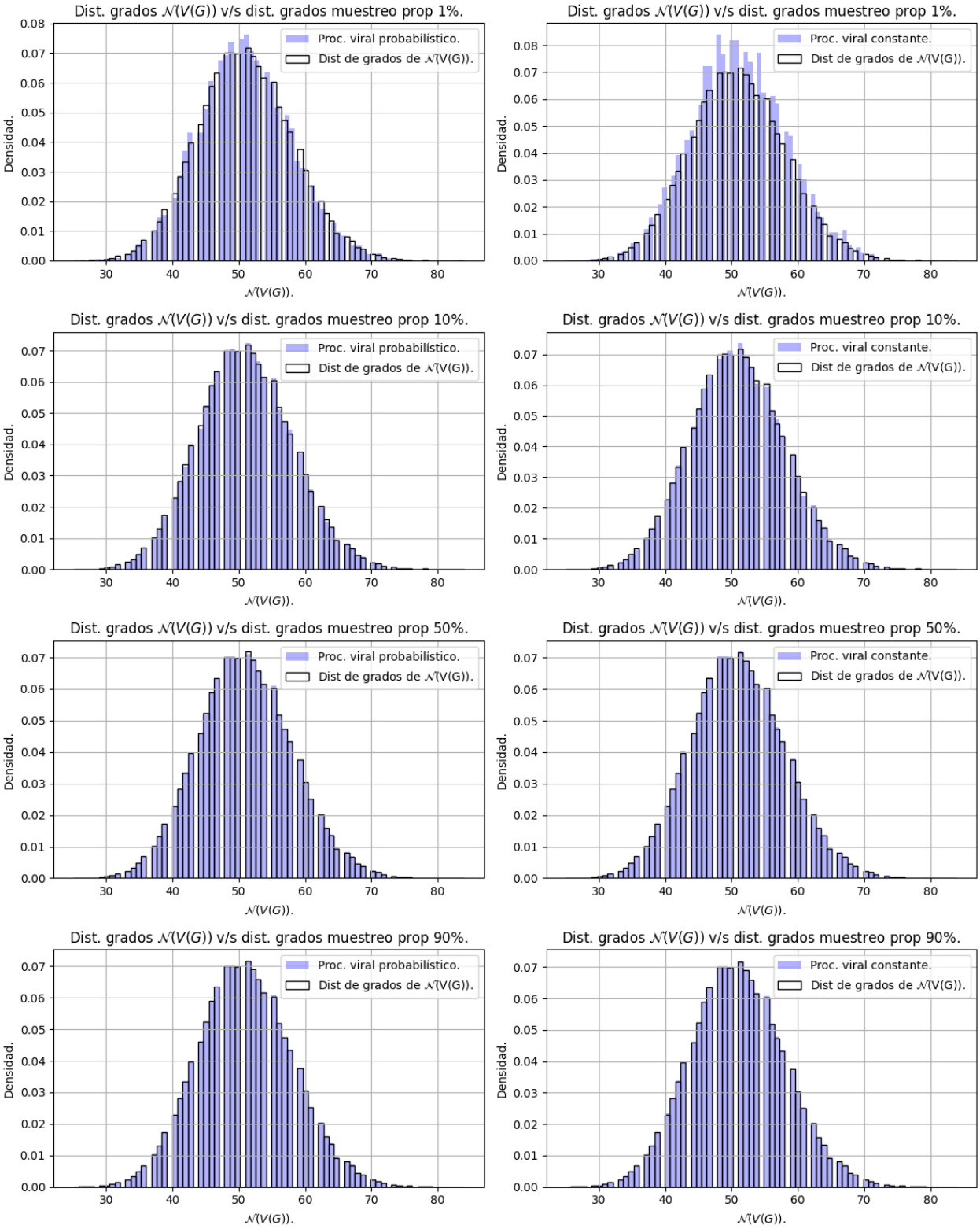


Figura A0.10: Gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral probabilístico variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral constante variando su proporción de tamaño de un grafo aleatorio ligero.

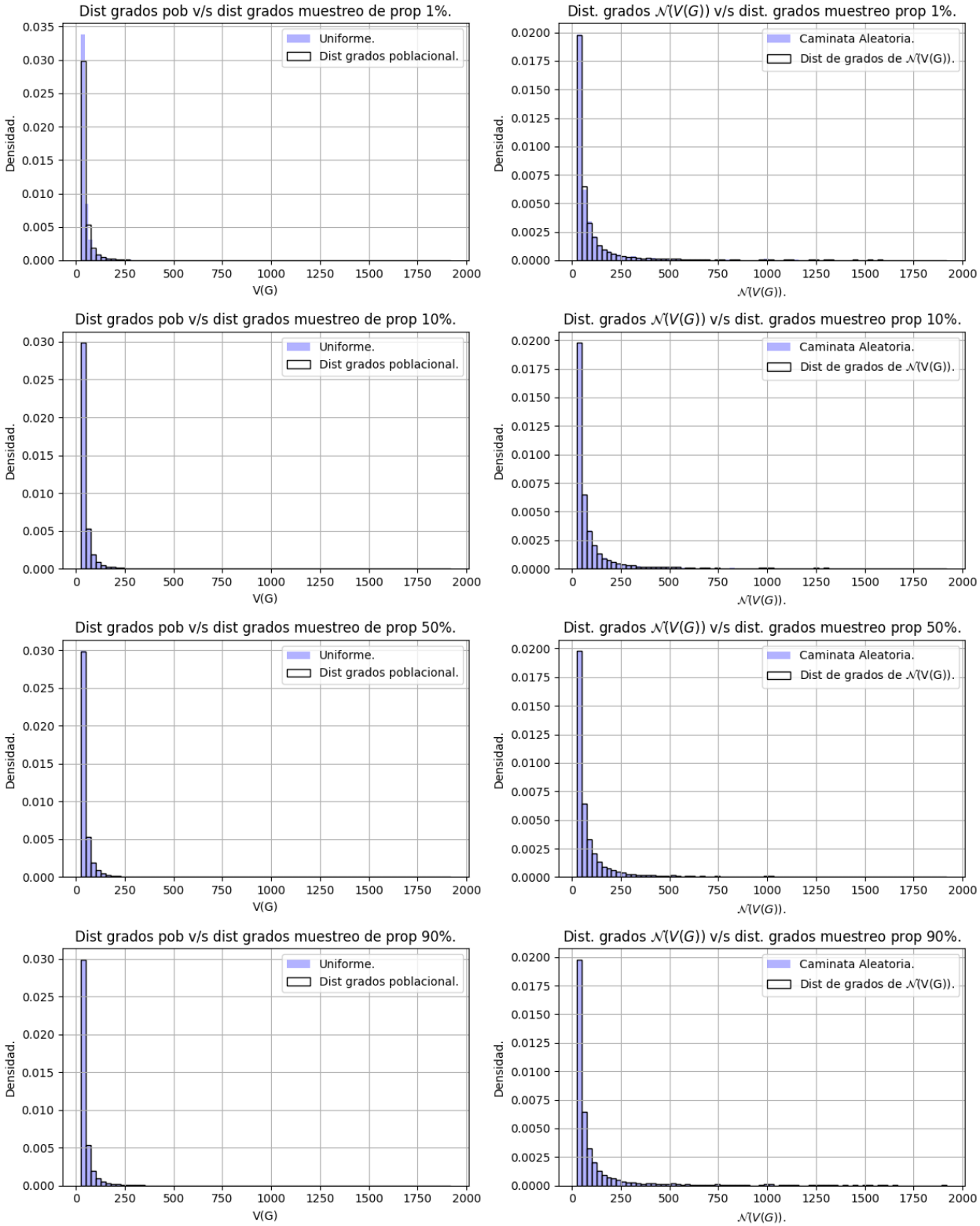


Figura A0.11: Gráficas de distribución de grados de $V(G)$ v/s distribución de grados de un muestreo uniforme variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a una caminata aleatoria variando su proporción de tamaño en un grafo escala ligero.

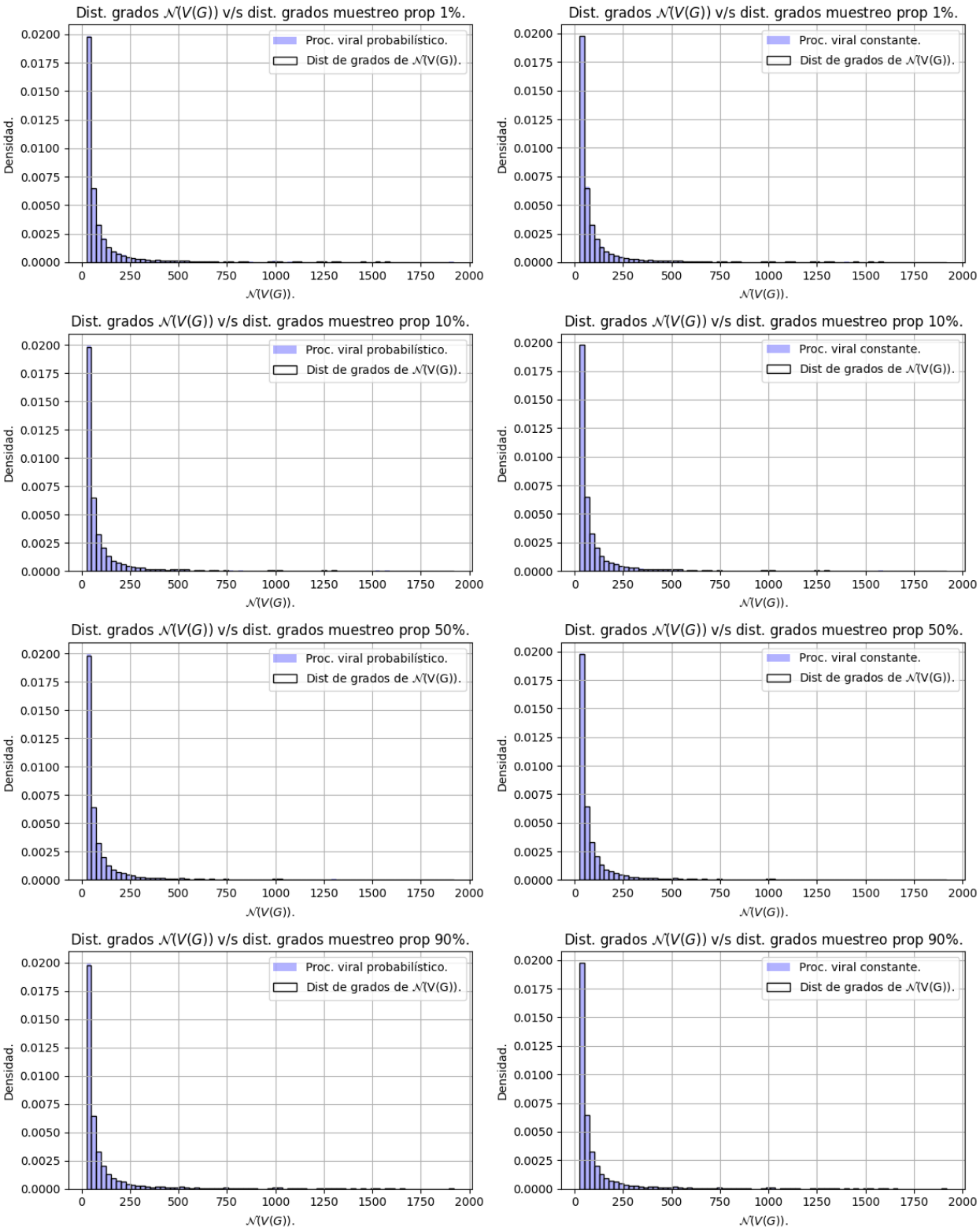


Figura A0.12: Gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral probabilístico variando su proporción de tamaño y gráficas de distribución de grados de $\mathcal{N}(V(G))$ v/s distribución de grados de una muestra asociado a un proceso viral constante variando su proporción de tamaño de un grafo escala ligero.

Apéndice B

Resultados de experimentación con grupos generados

Simbología de tipos de muestreos aleatorios:

- Tipo de muestreo 5 se refiere al muestreo aleatorio uniforme con repetición.
- Tipo de muestreo 6 se refiere al muestreo aleatorio asociado a una caminata aleatoria.
- Tipo de muestreo 7 se refiere al muestreo aleatorio asociado a un proceso viral probabilístico.
- Tipo de muestreo 8 se refiere al muestreo aleatorio asociado a un proceso viral constante.

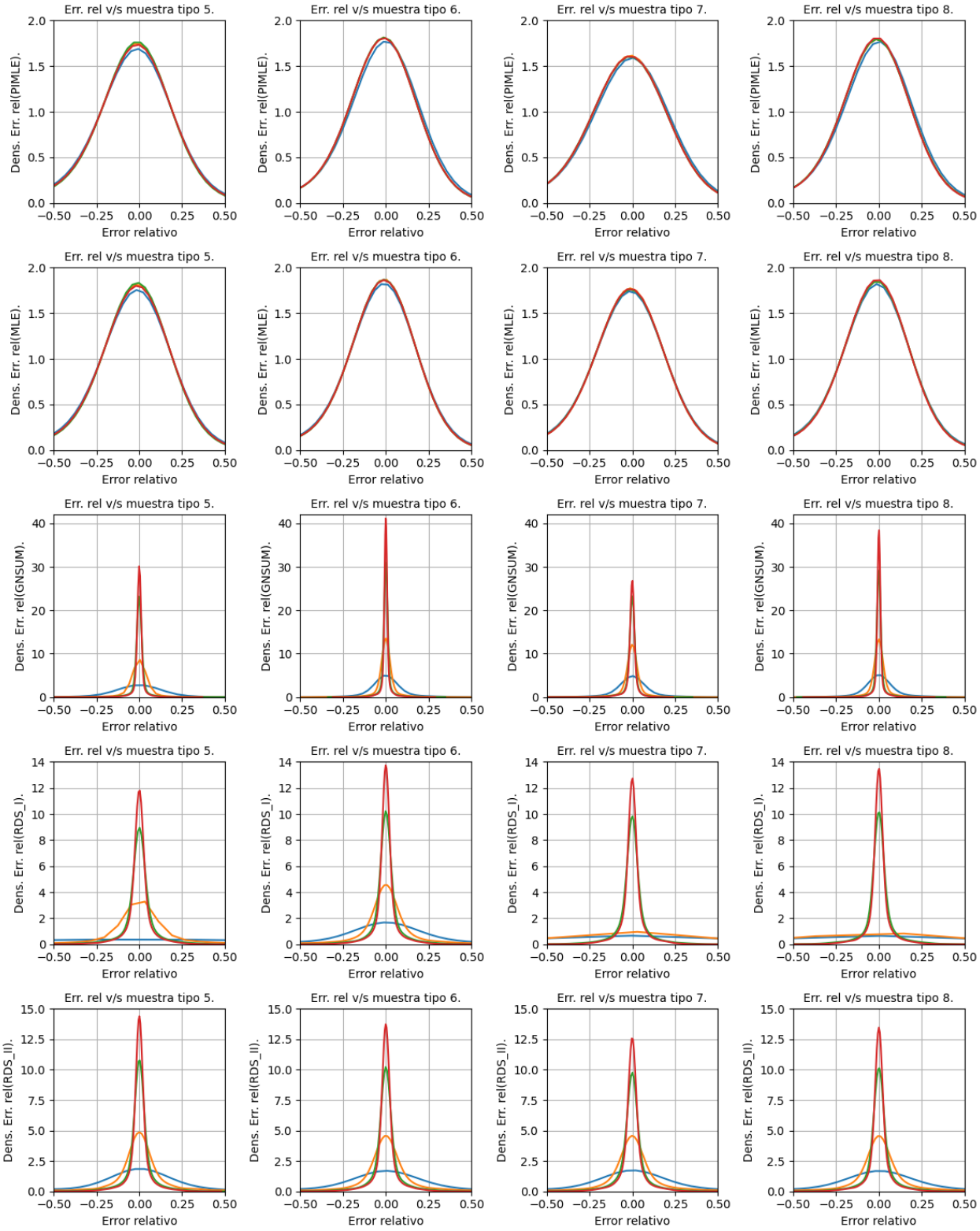


Figura B0.1: Experimento 1: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

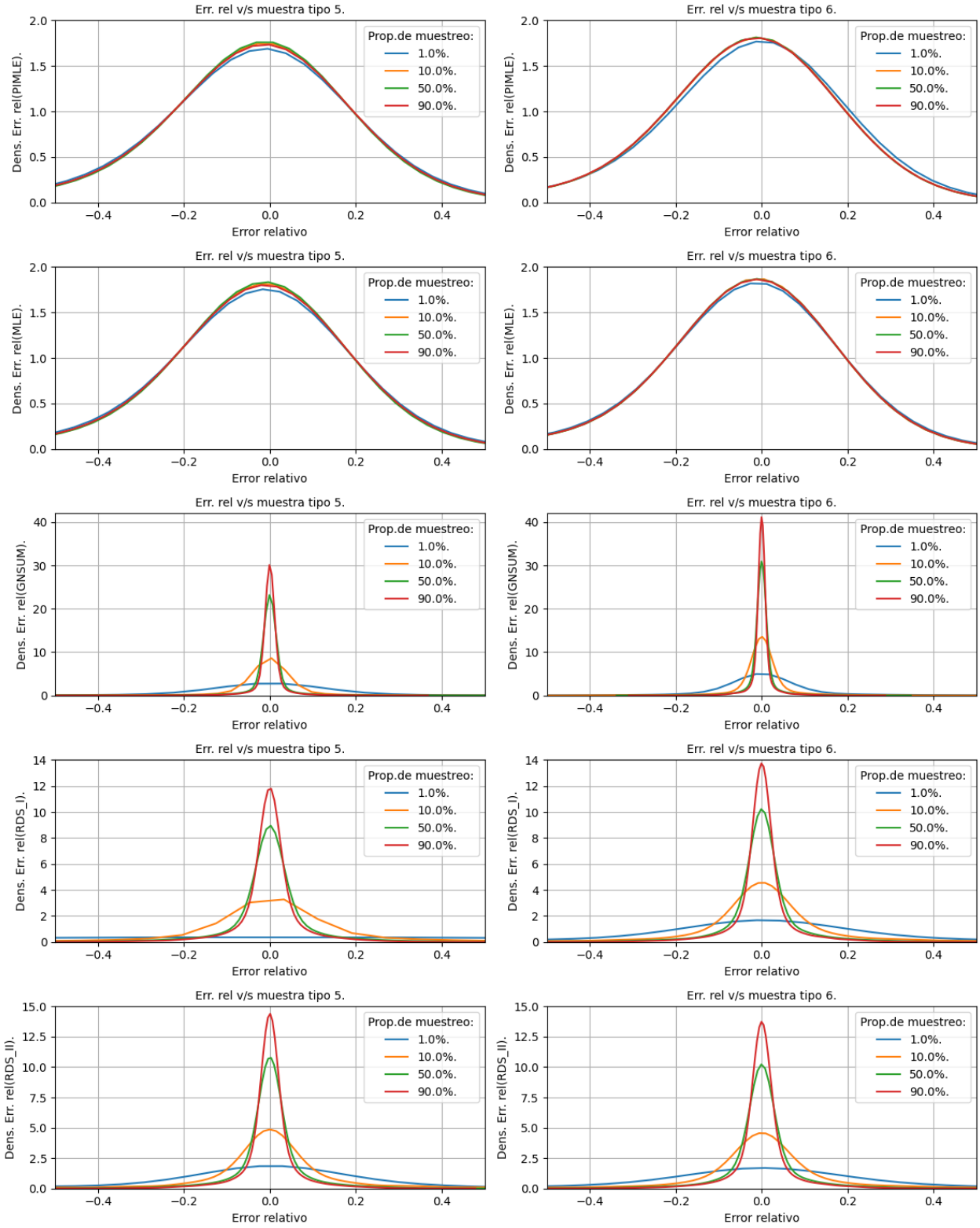


Figura B0.2: Muestra las dos primeras columnas de la figura B0.1.

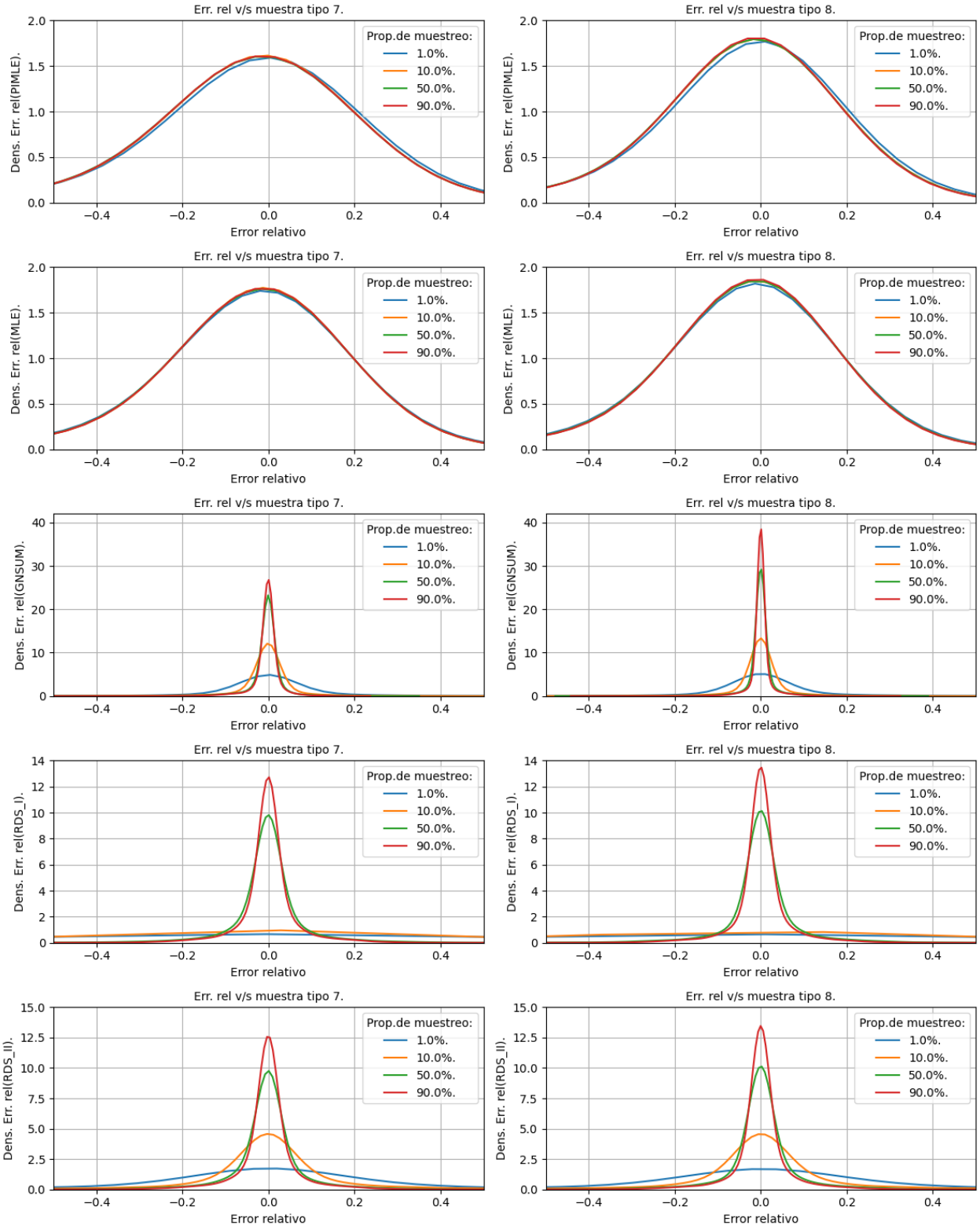


Figura B0.3: Muestra las dos últimas columnas de la figura B0.1.

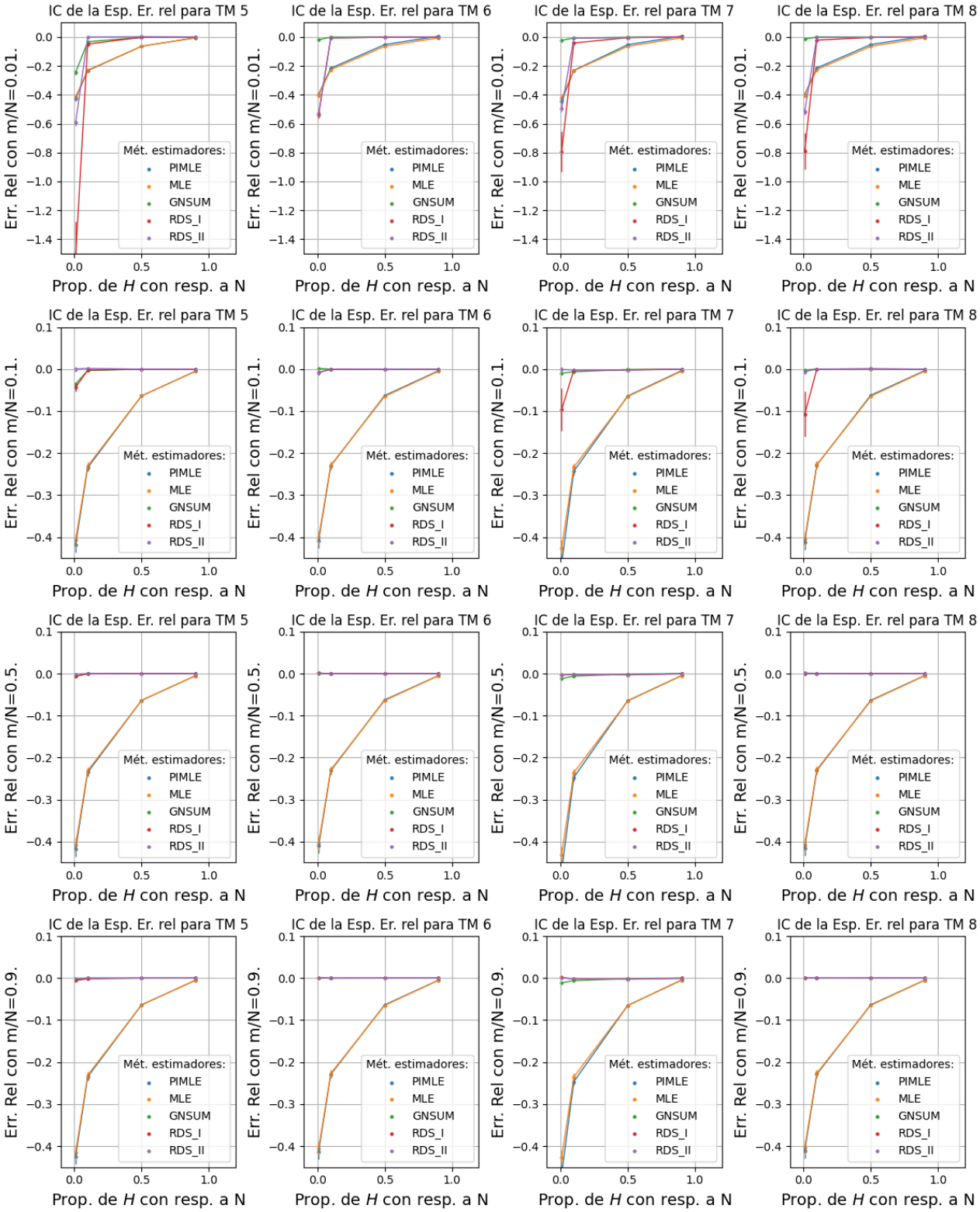


Figura B0.4: Experimento 1: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

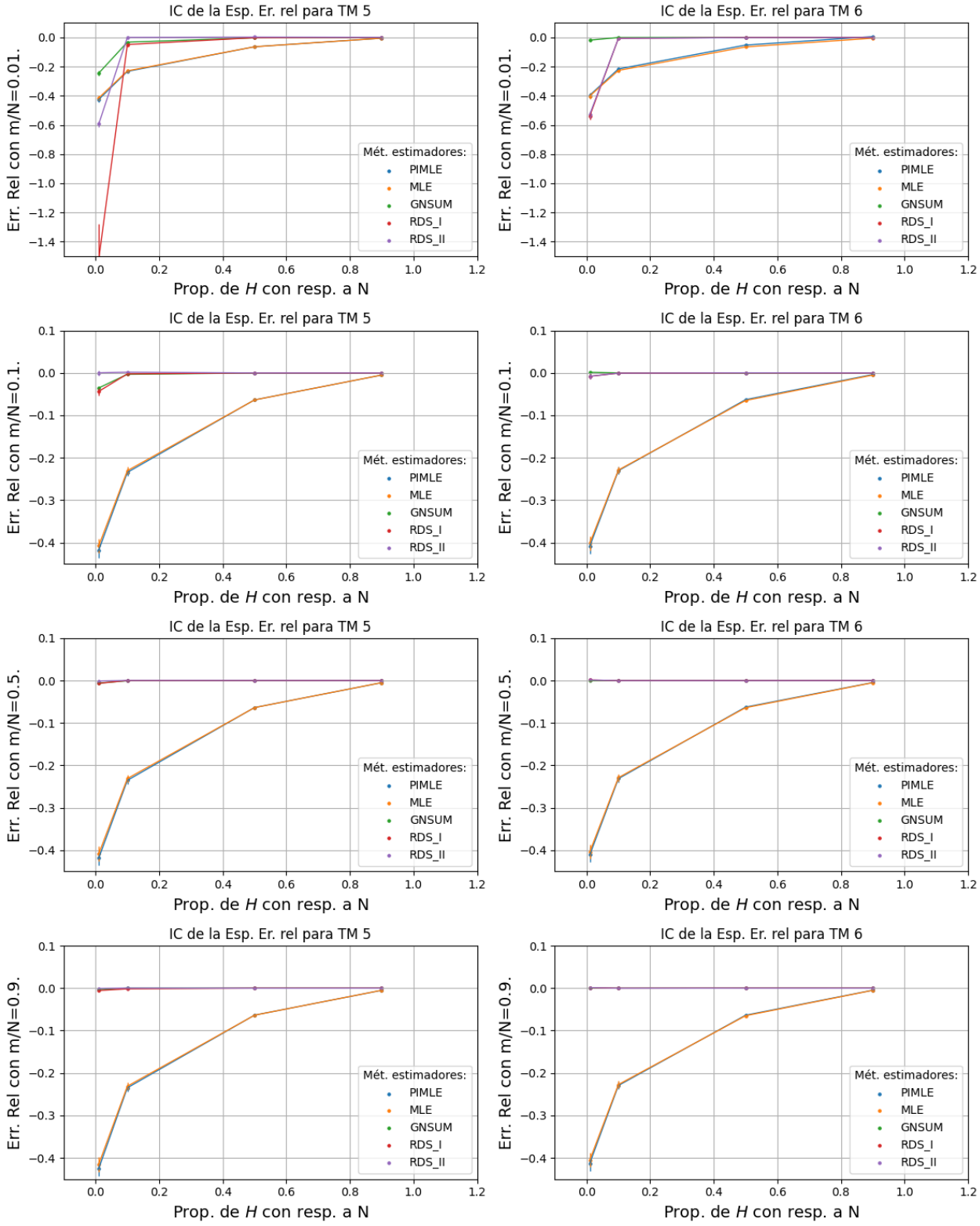


Figura B0.5: Muestra las dos primeras columnas de la figura B0.4.

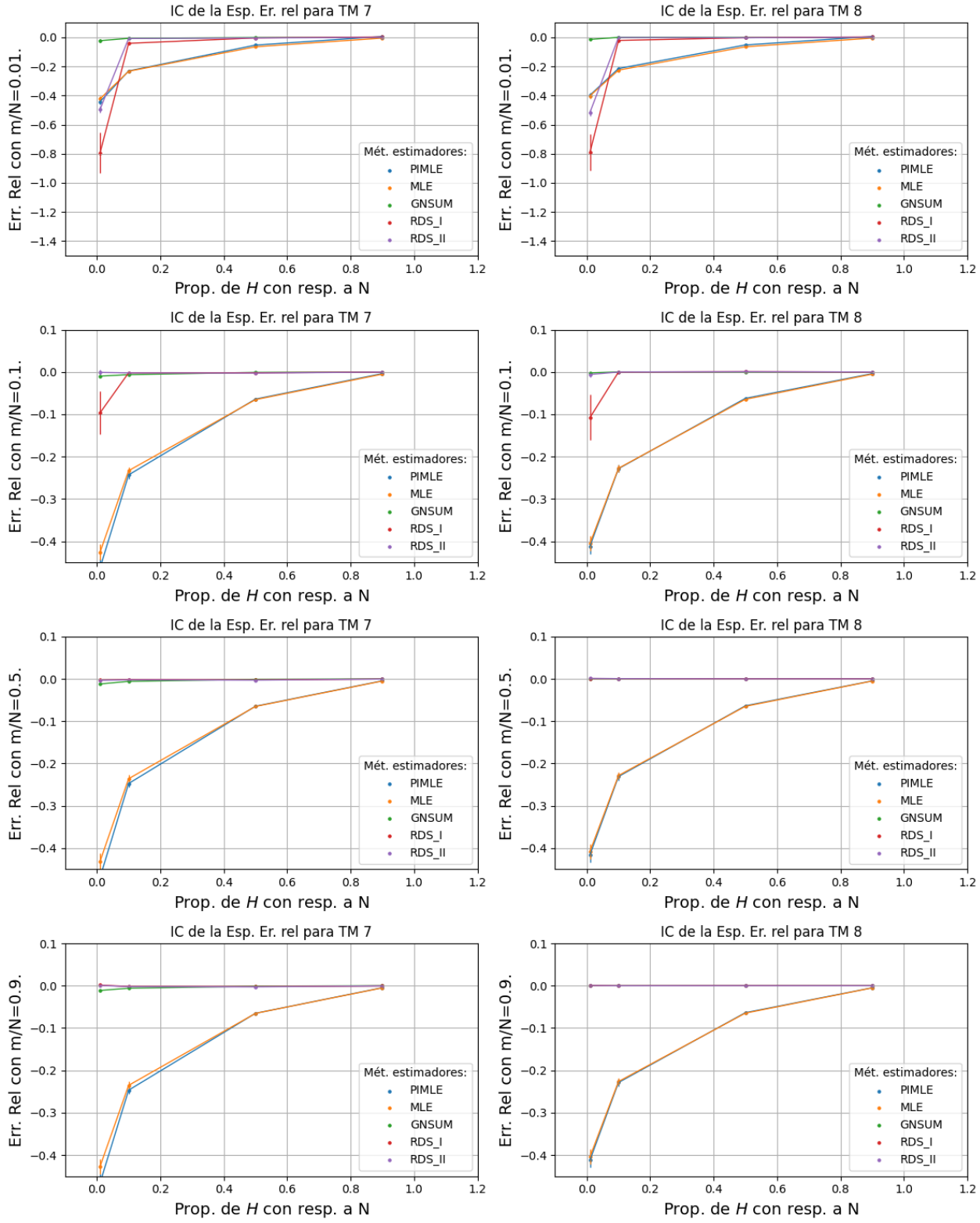


Figura B0.6: Muestra las dos últimas columnas de la figura B0.4.

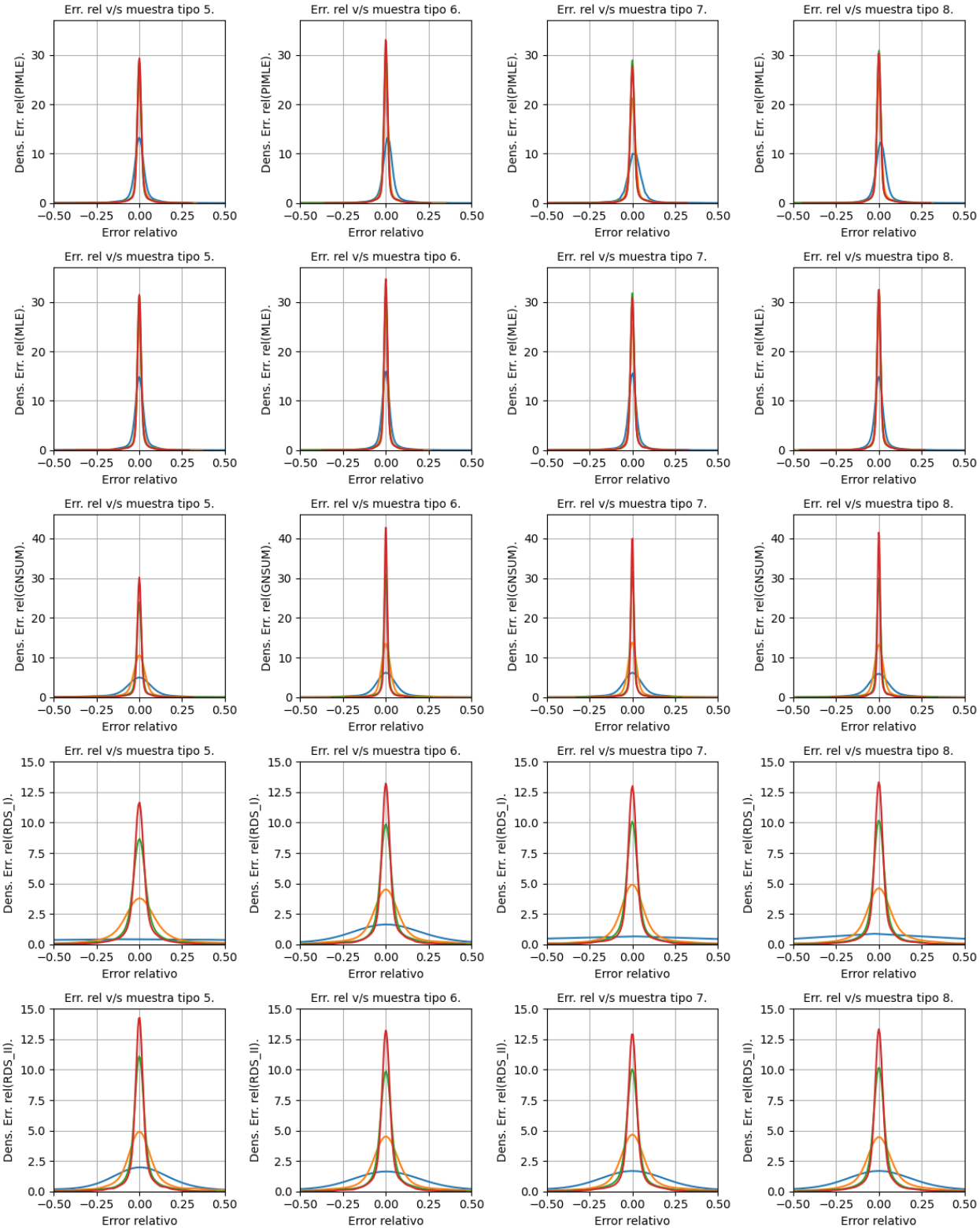


Figura B0.7: Experimento 2: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

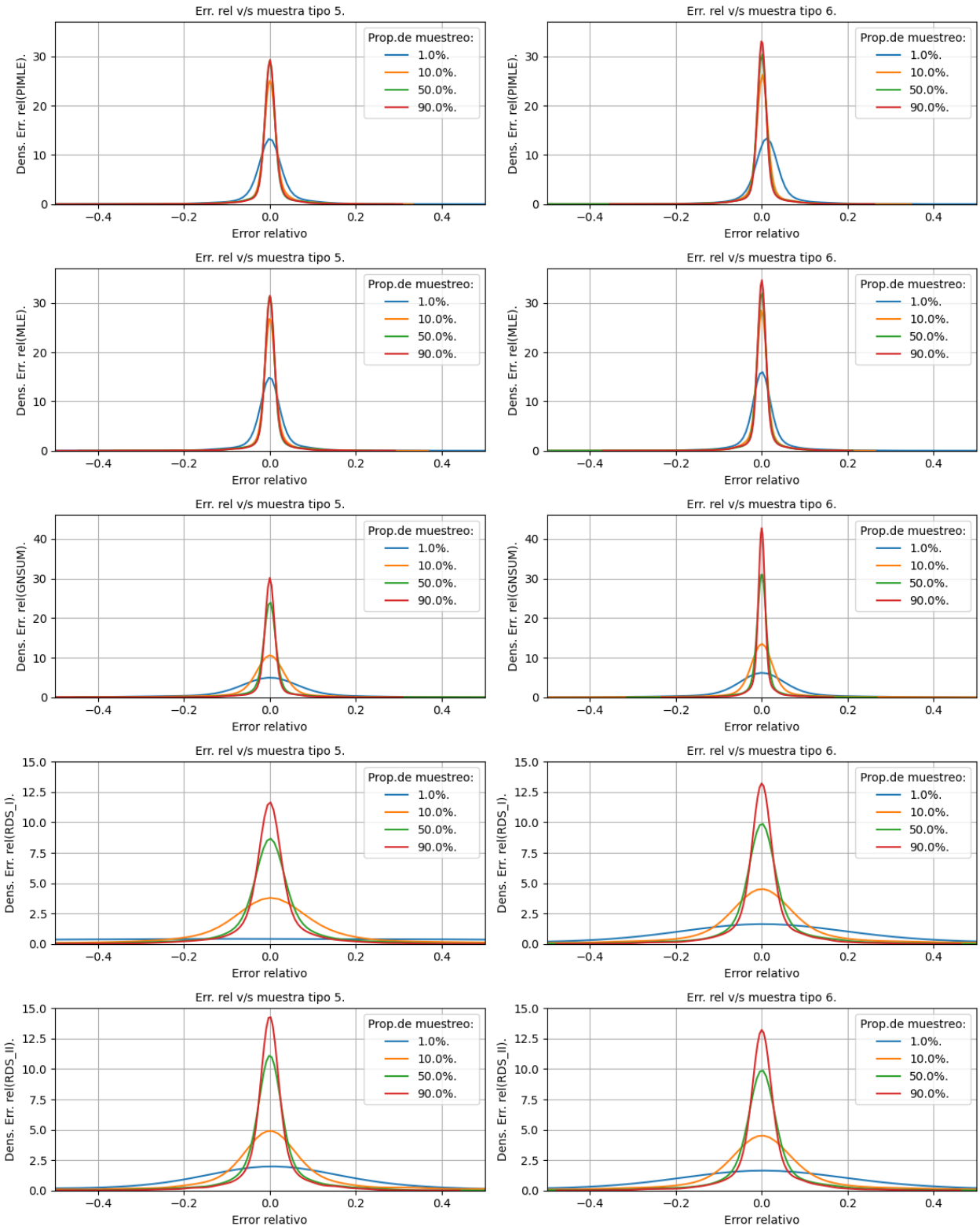


Figura B0.8: Muestra las dos primeras columnas de la figura B0.7.

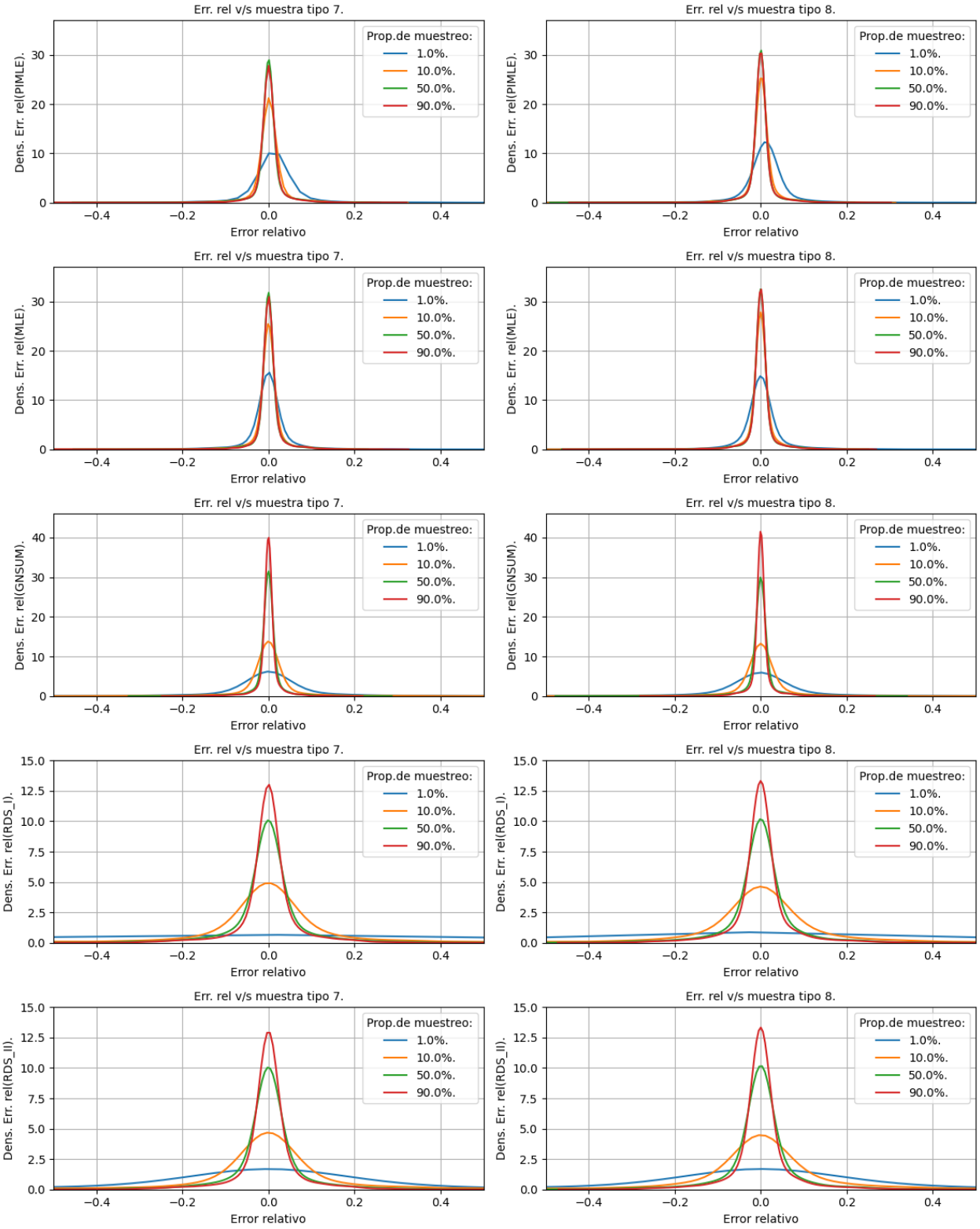


Figura B0.9: Muestra las dos últimas columnas de la figura B0.7.

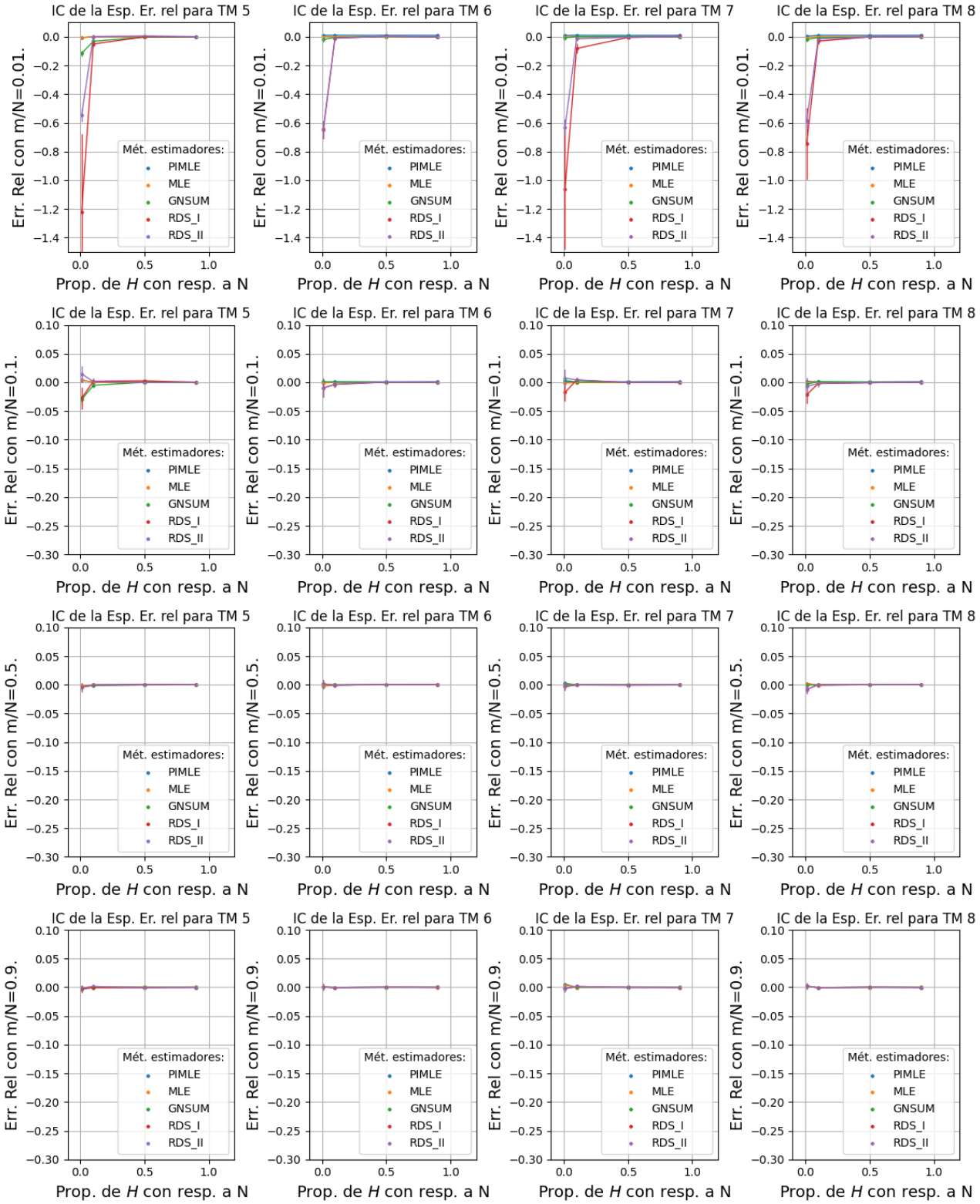


Figura B0.10: Experimento 2: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

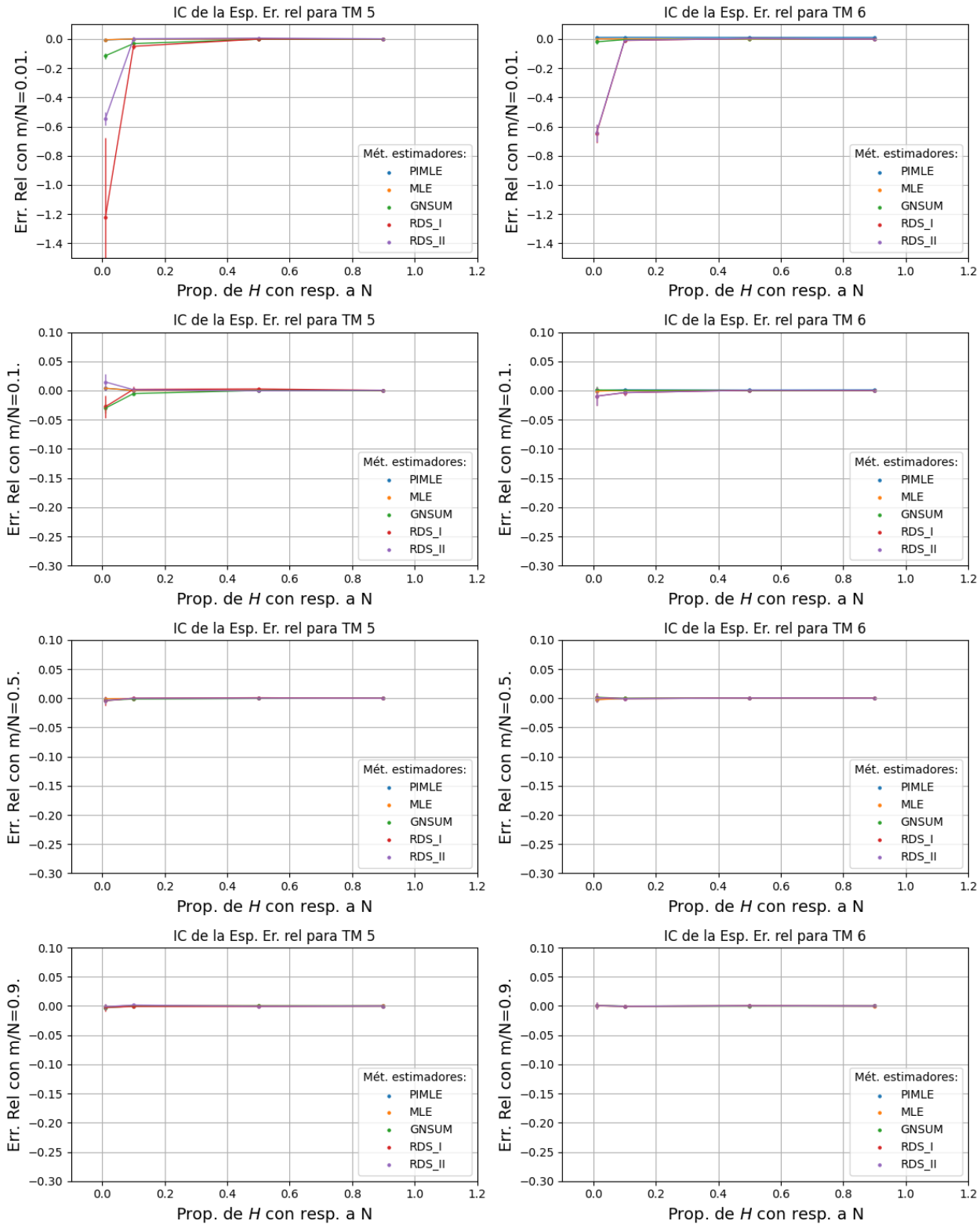


Figura B0.11: Muestra las dos primeras columnas de la figura B0.10.

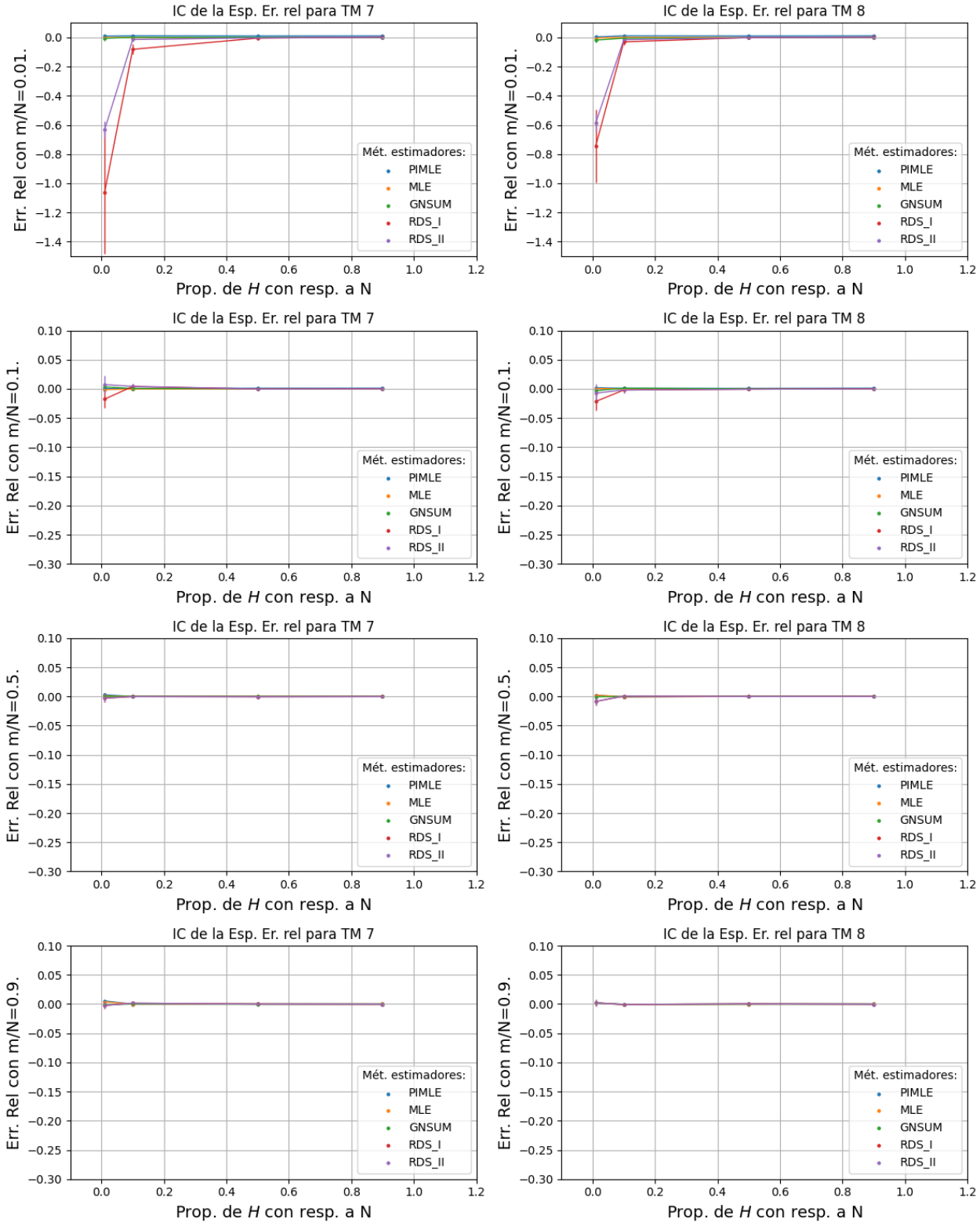


Figura B0.12: Muestra las dos últimas columnas de la figura B0.10.

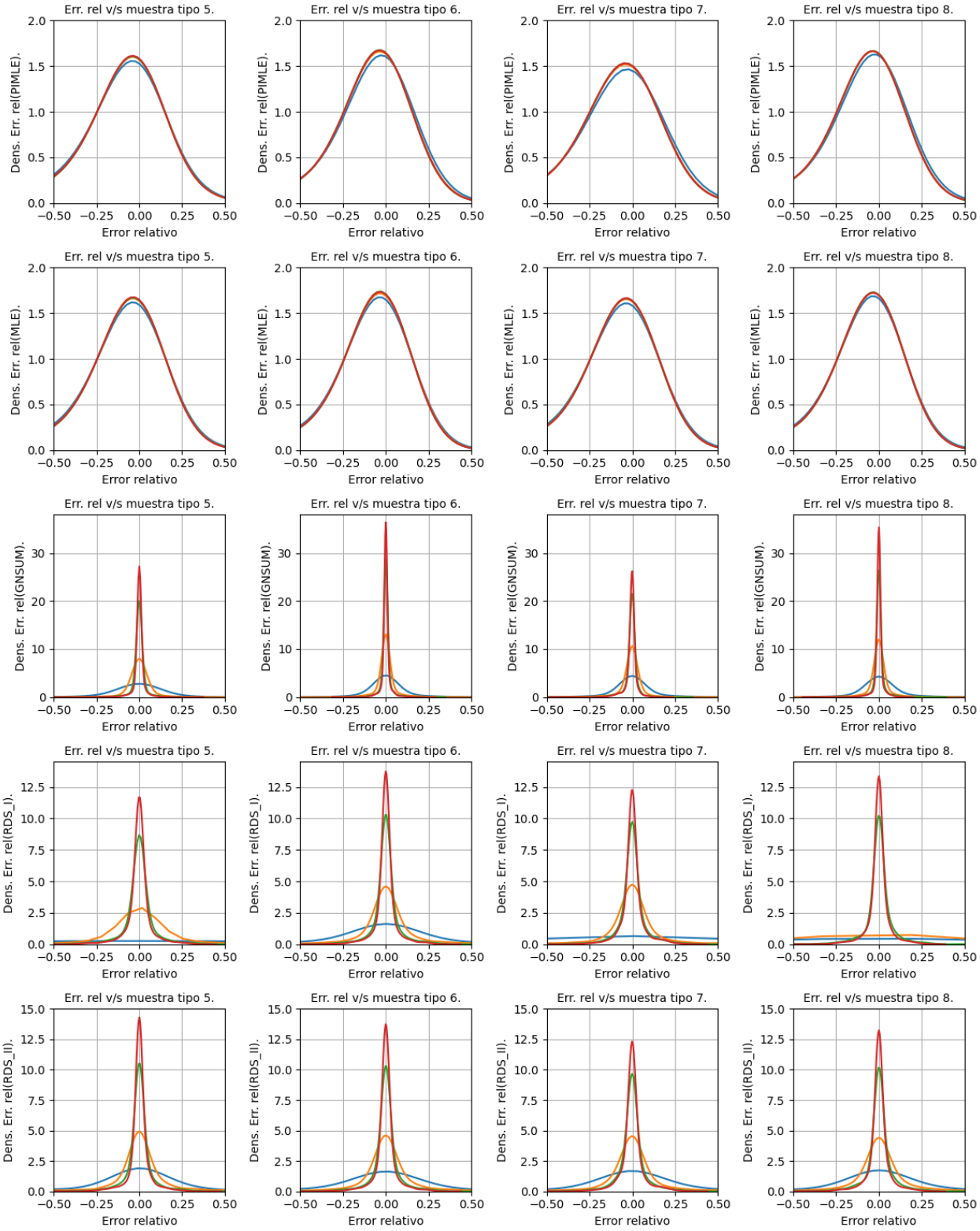


Figura B0.13: Experimento 3: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

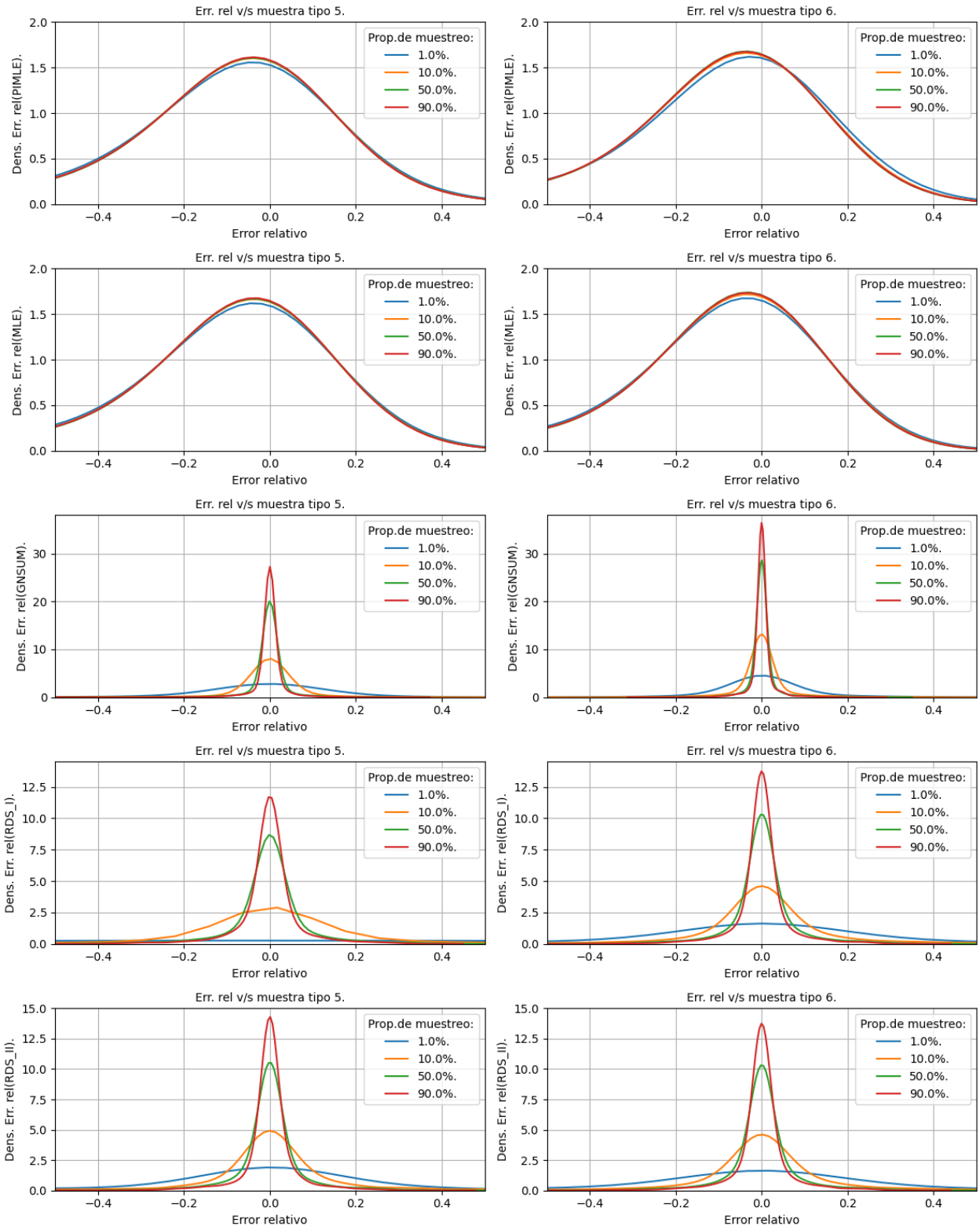


Figura B0.14: Muestra las dos primeras columnas de la figura B0.13.

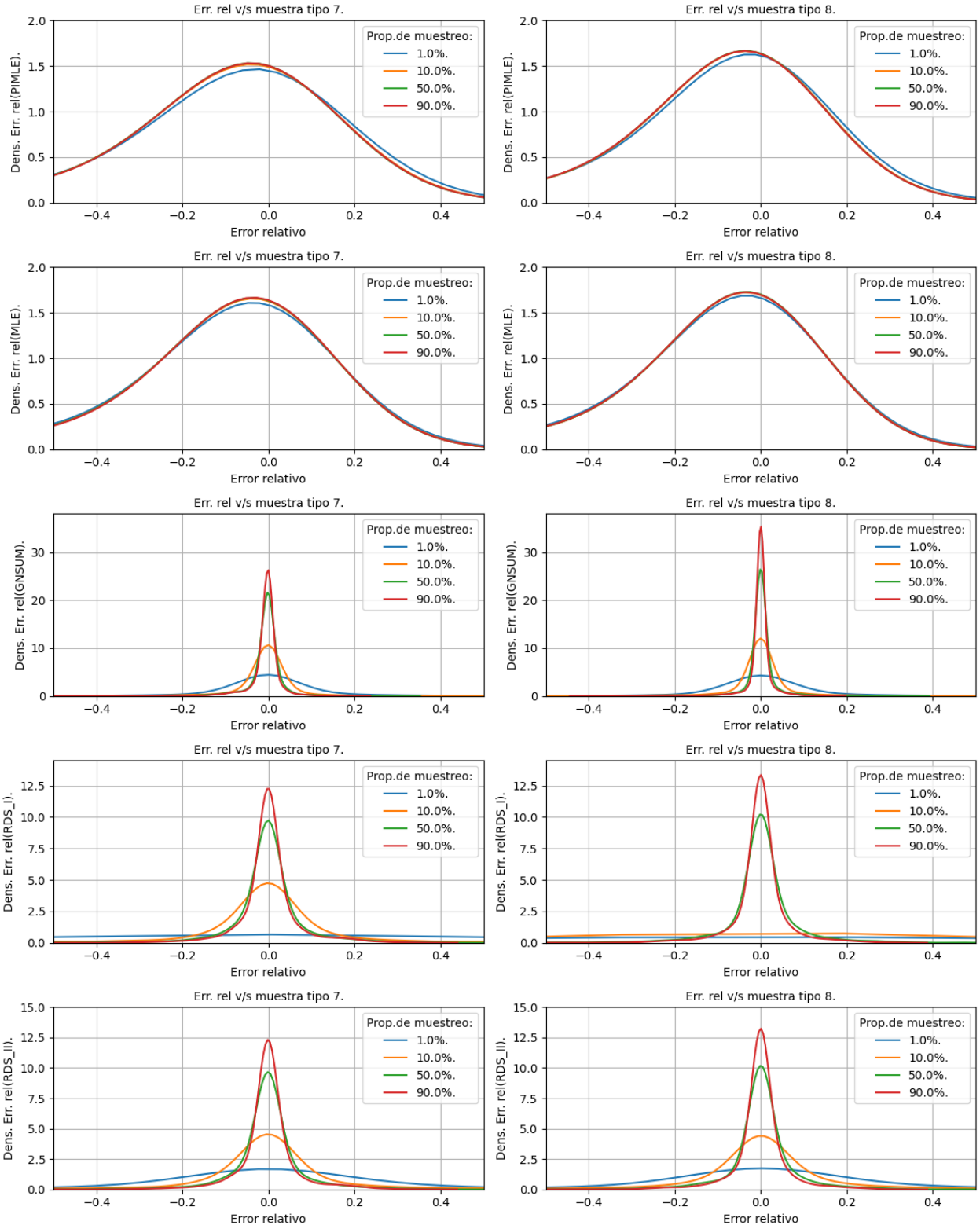


Figura B0.15: Muestra las dos últimas columnas de la figura B0.13.

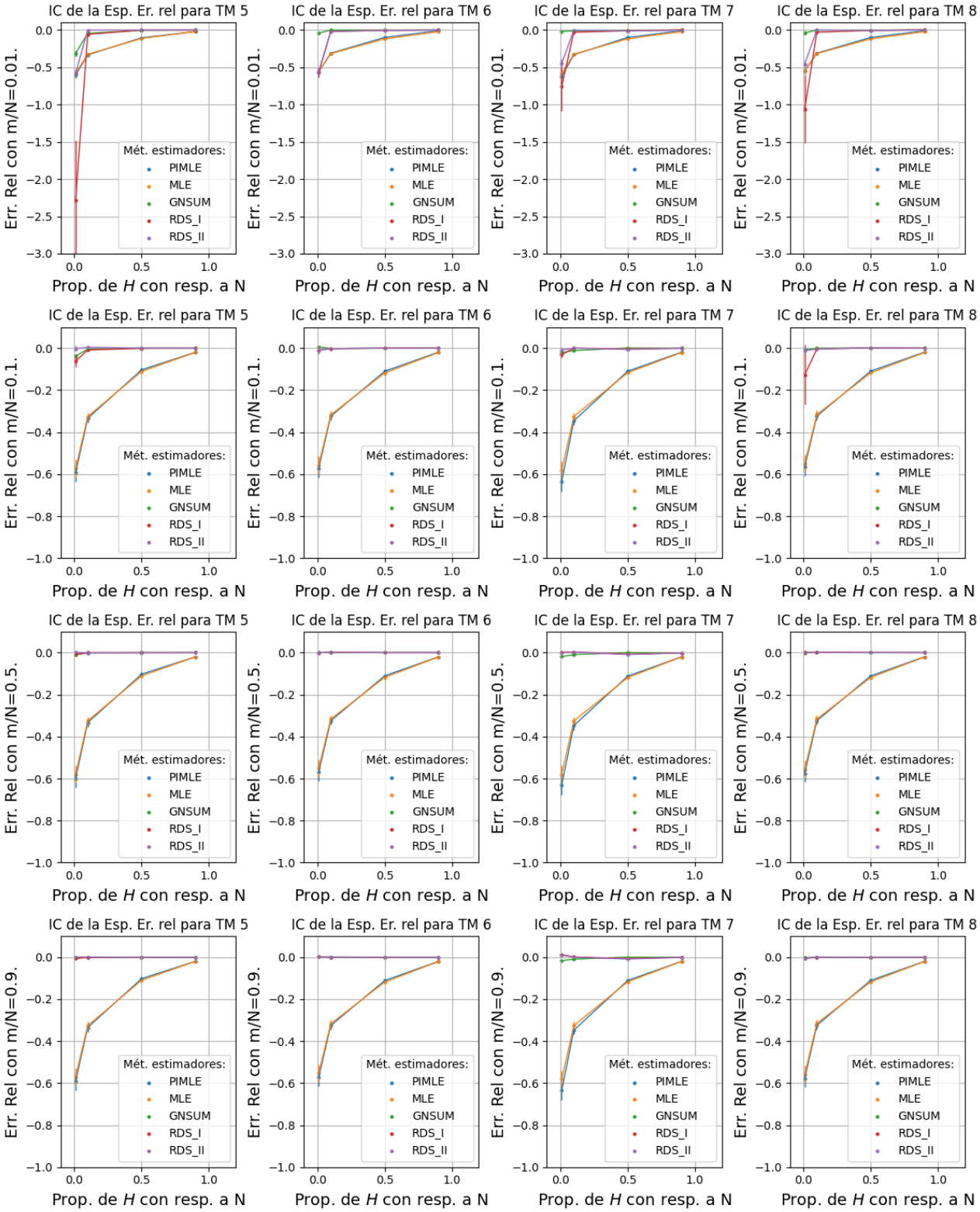


Figura B0.16: Experimento 3: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

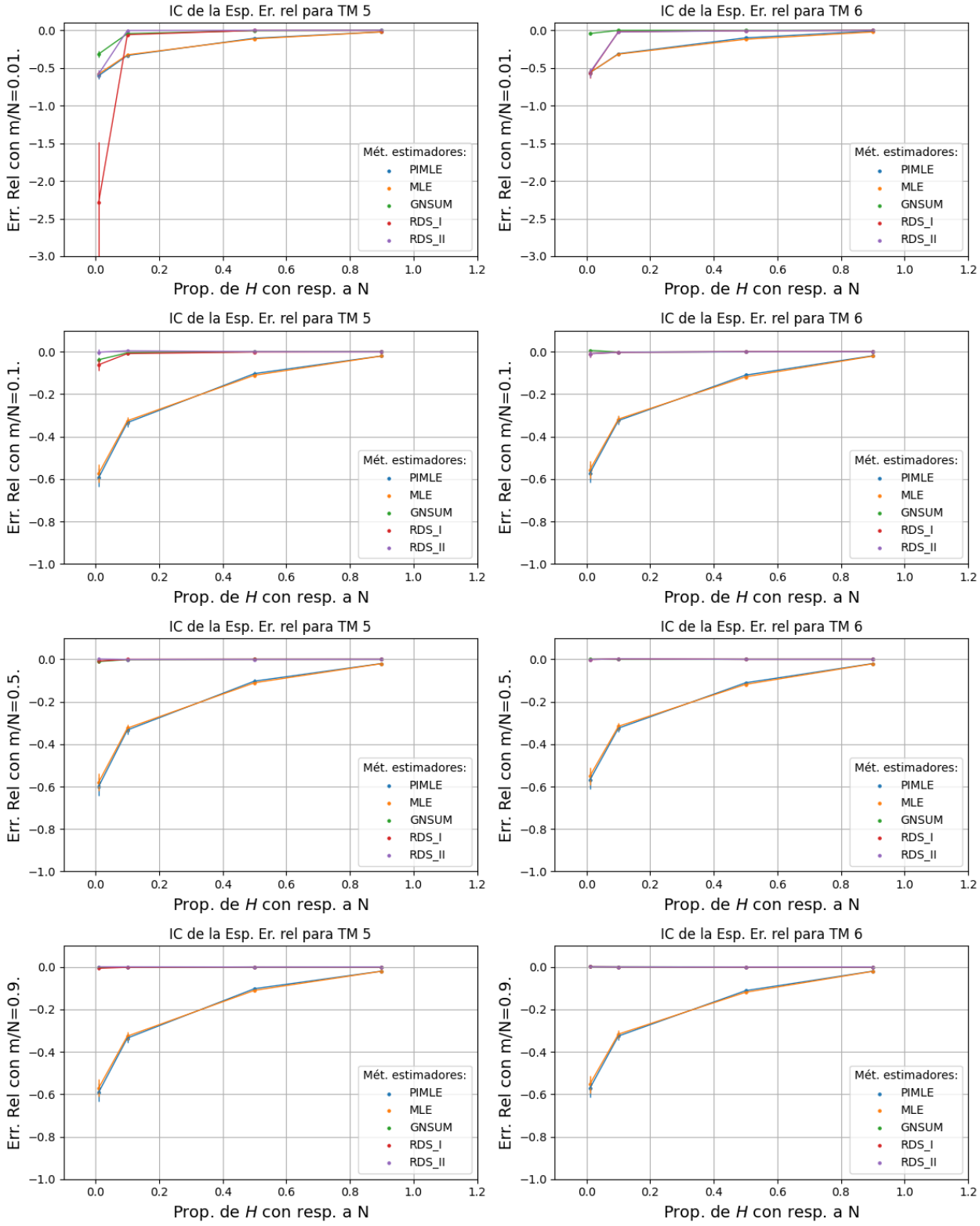


Figura B0.17: Muestra las dos primeras columnas de la figura B0.16.

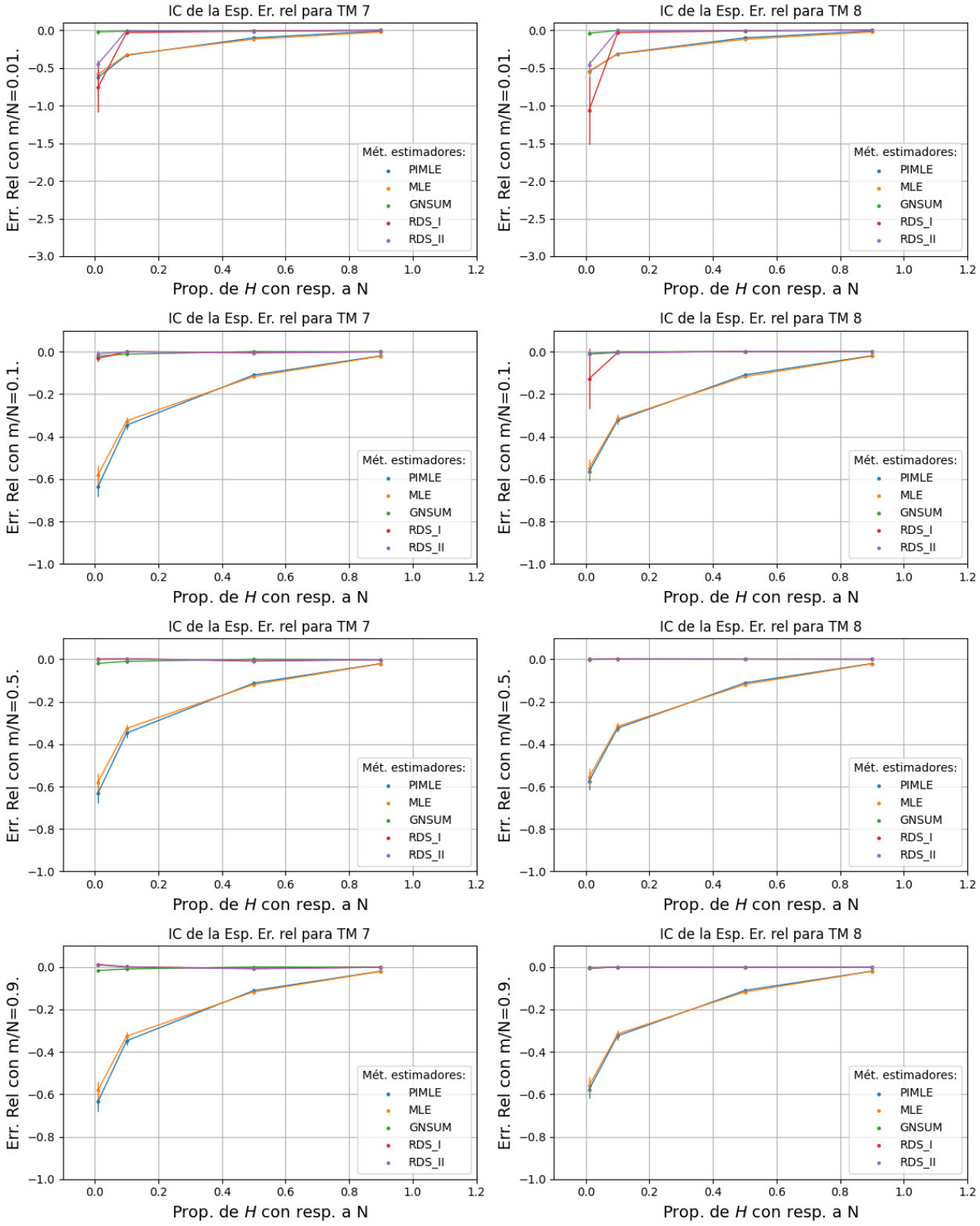


Figura B0.18: Muestra las dos últimas columnas de la figura B0.16.

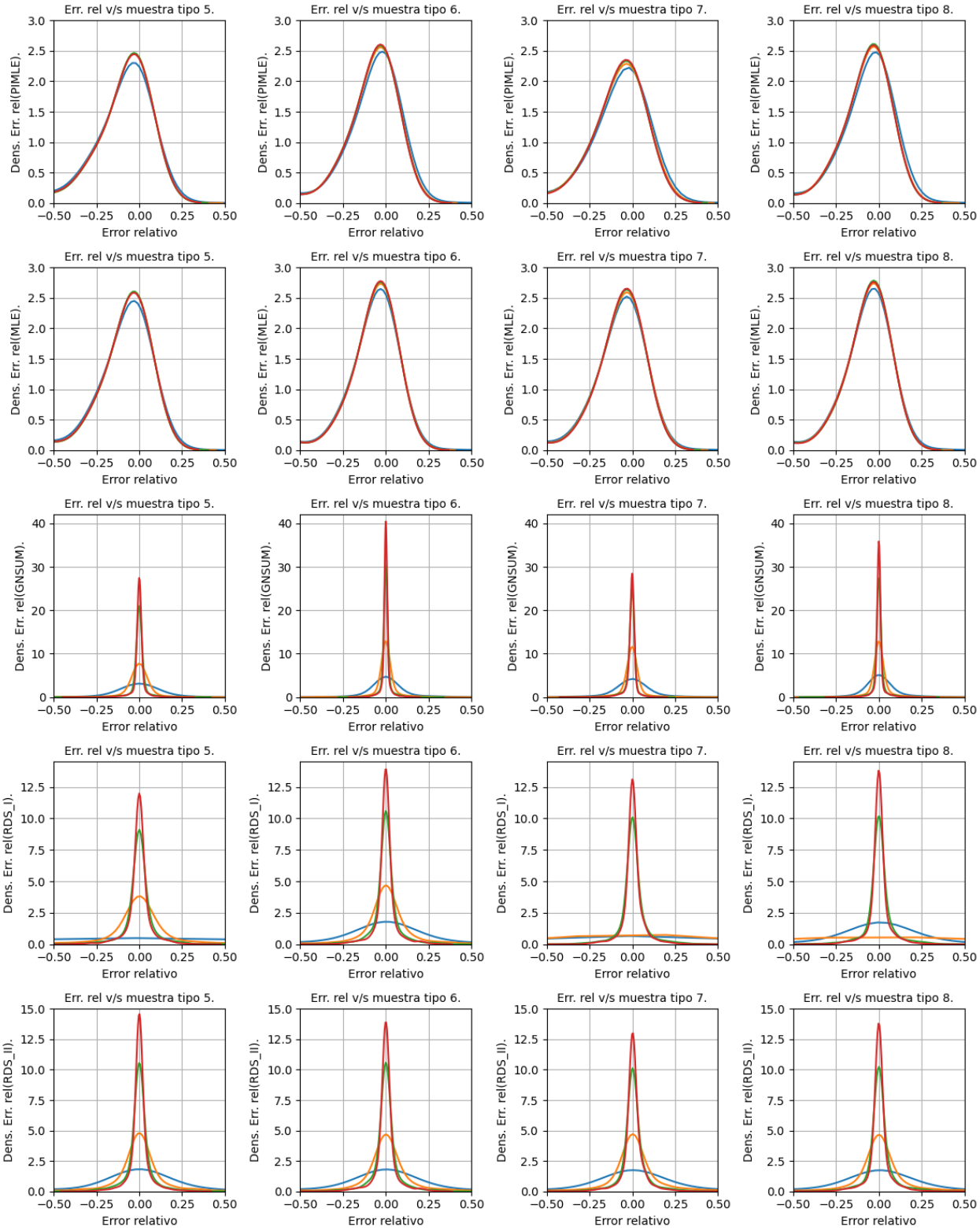


Figura B0.19: Experimento 4: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

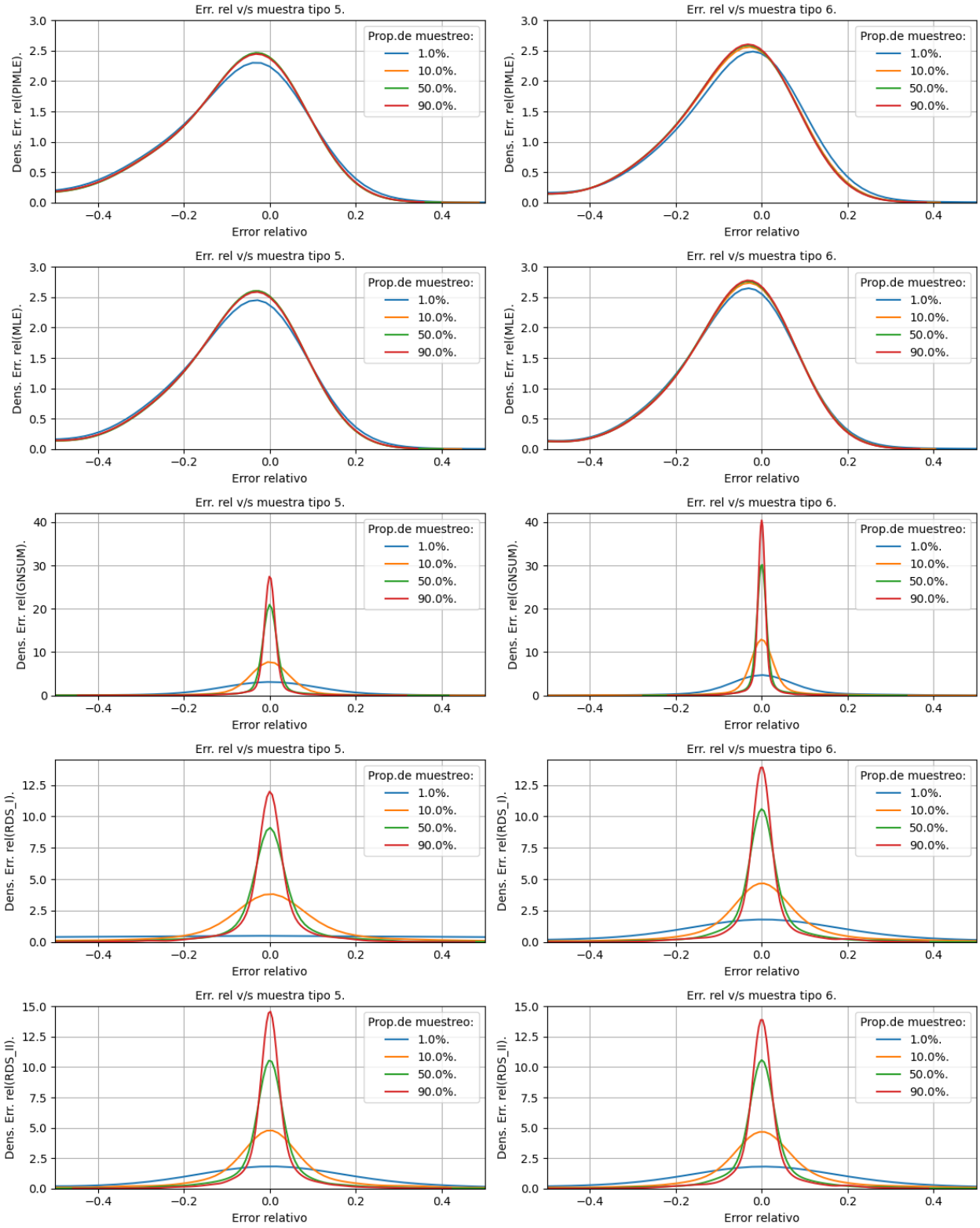


Figura B0.20: Muestra las dos primeras columnas de la figura B0.19.

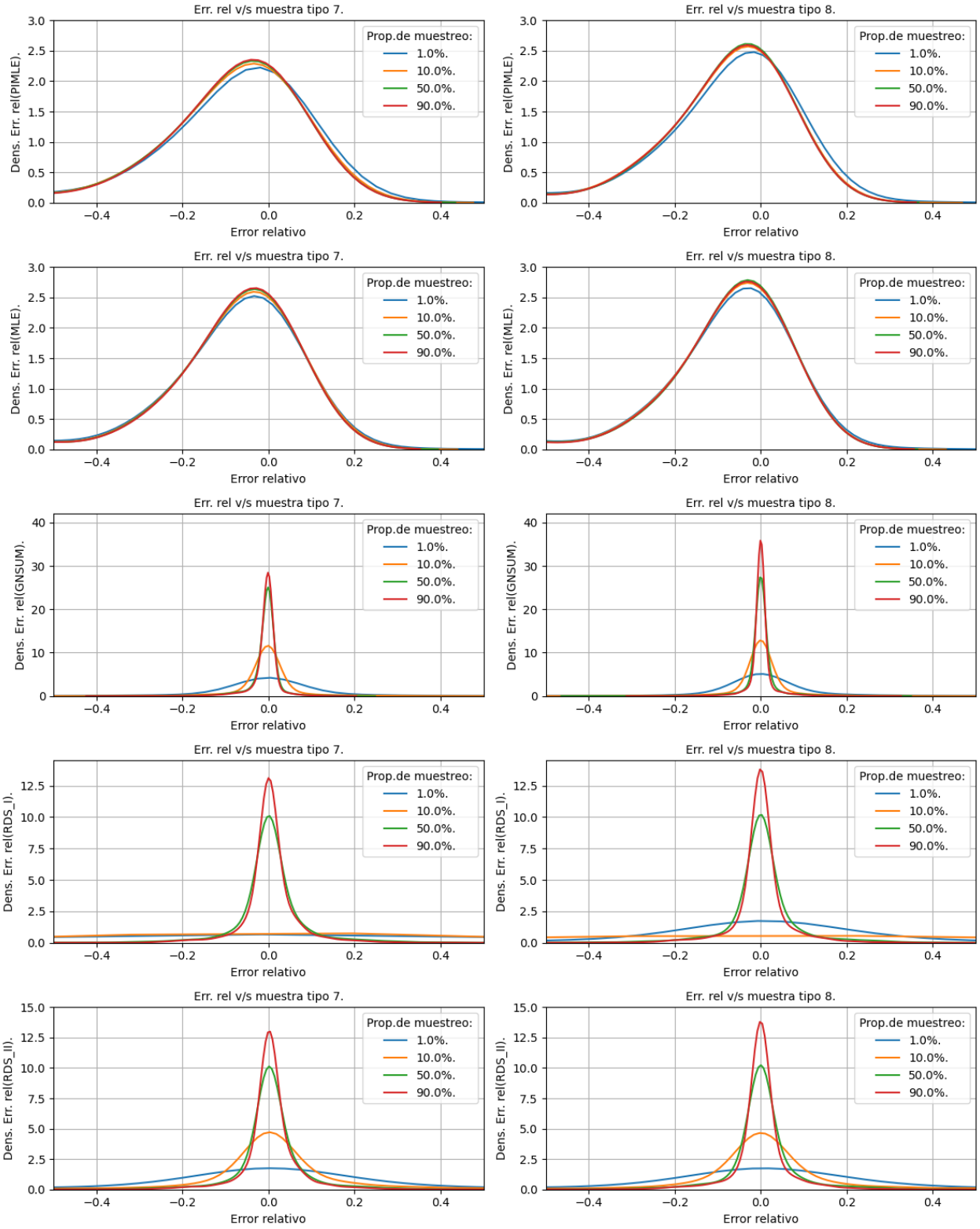


Figura B0.21: Muestra las dos últimas columnas de la figura B0.19.

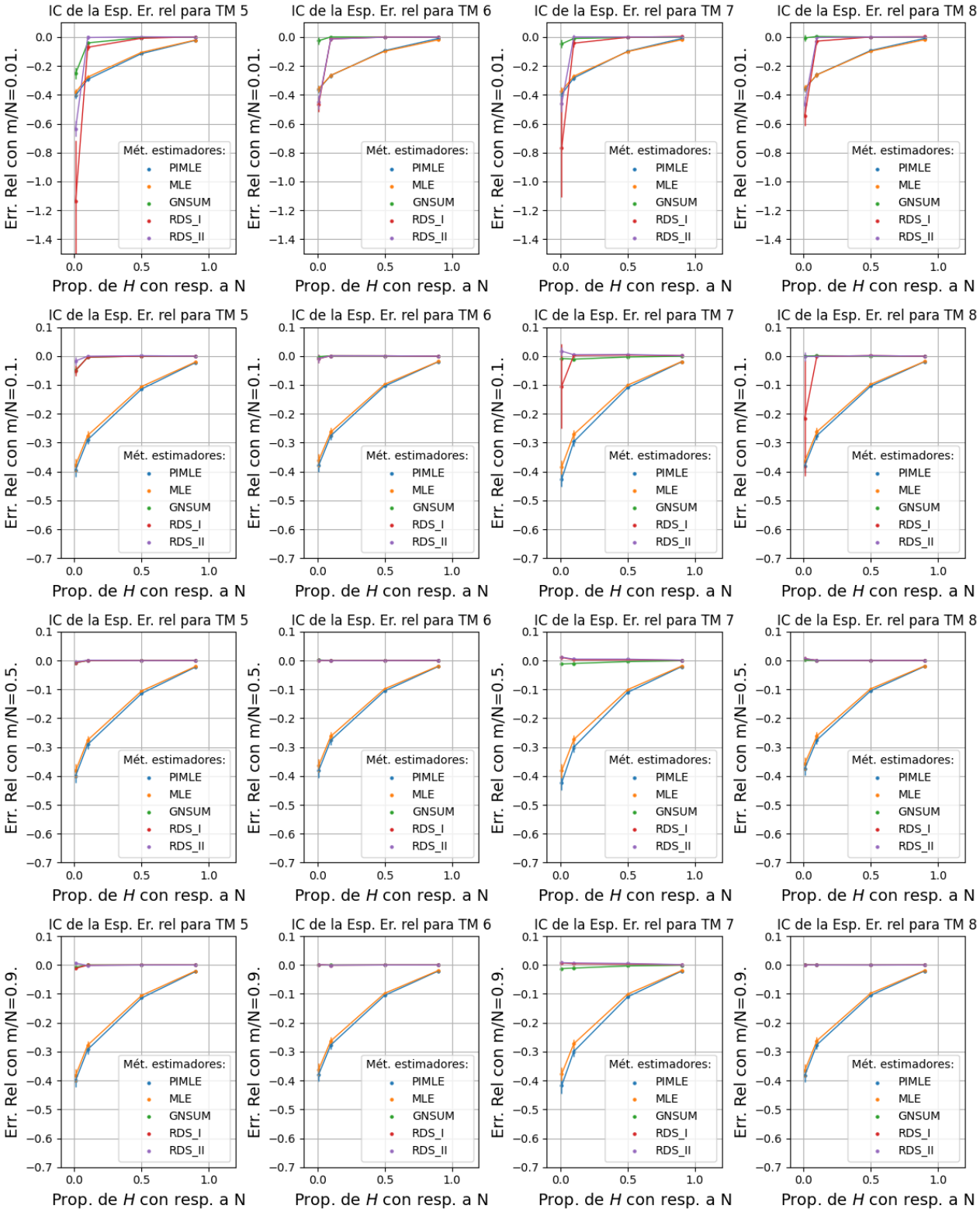


Figura B0.22: Experimento 4: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

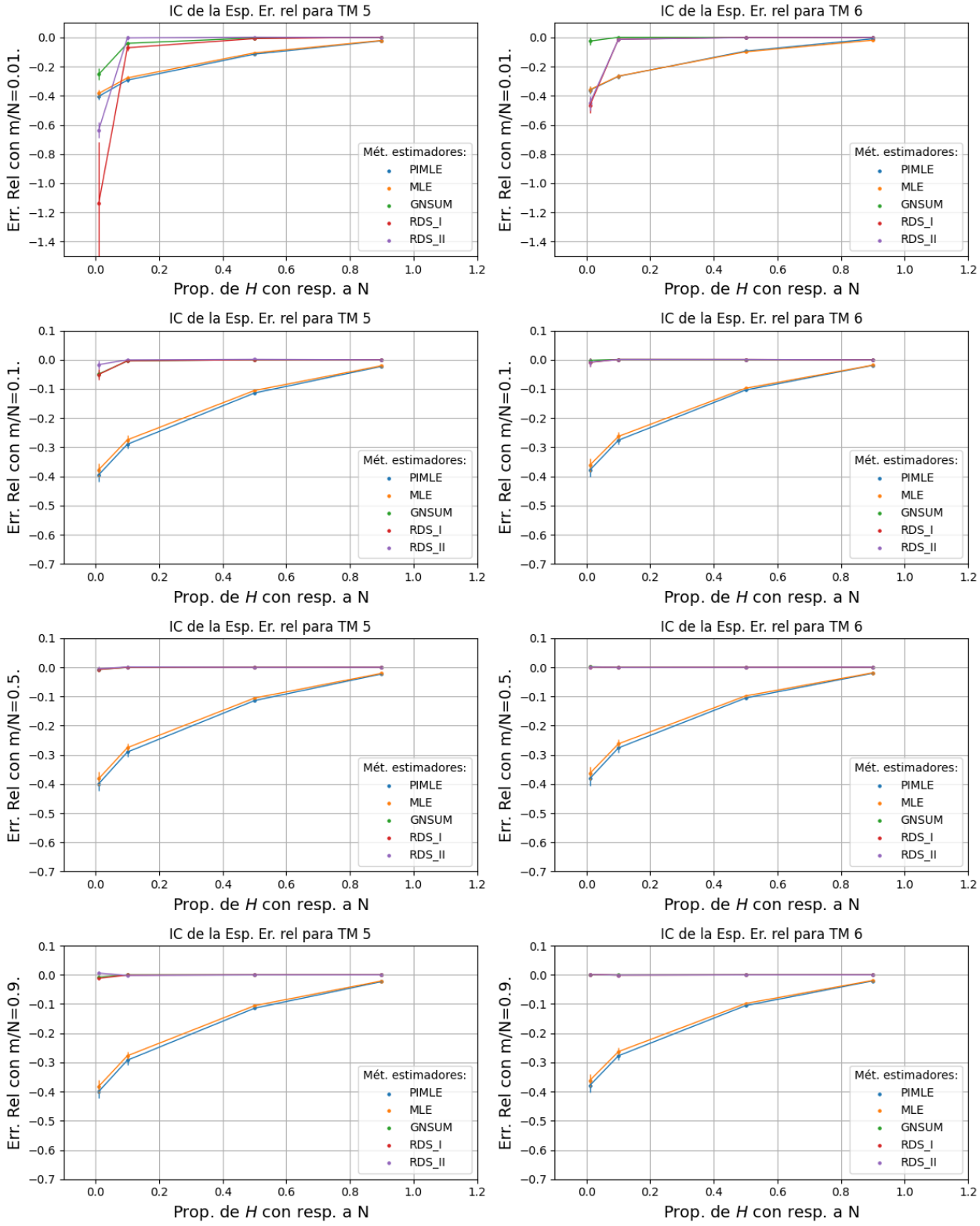


Figura B0.23: Muestra las dos primeras de la figura B0.22.

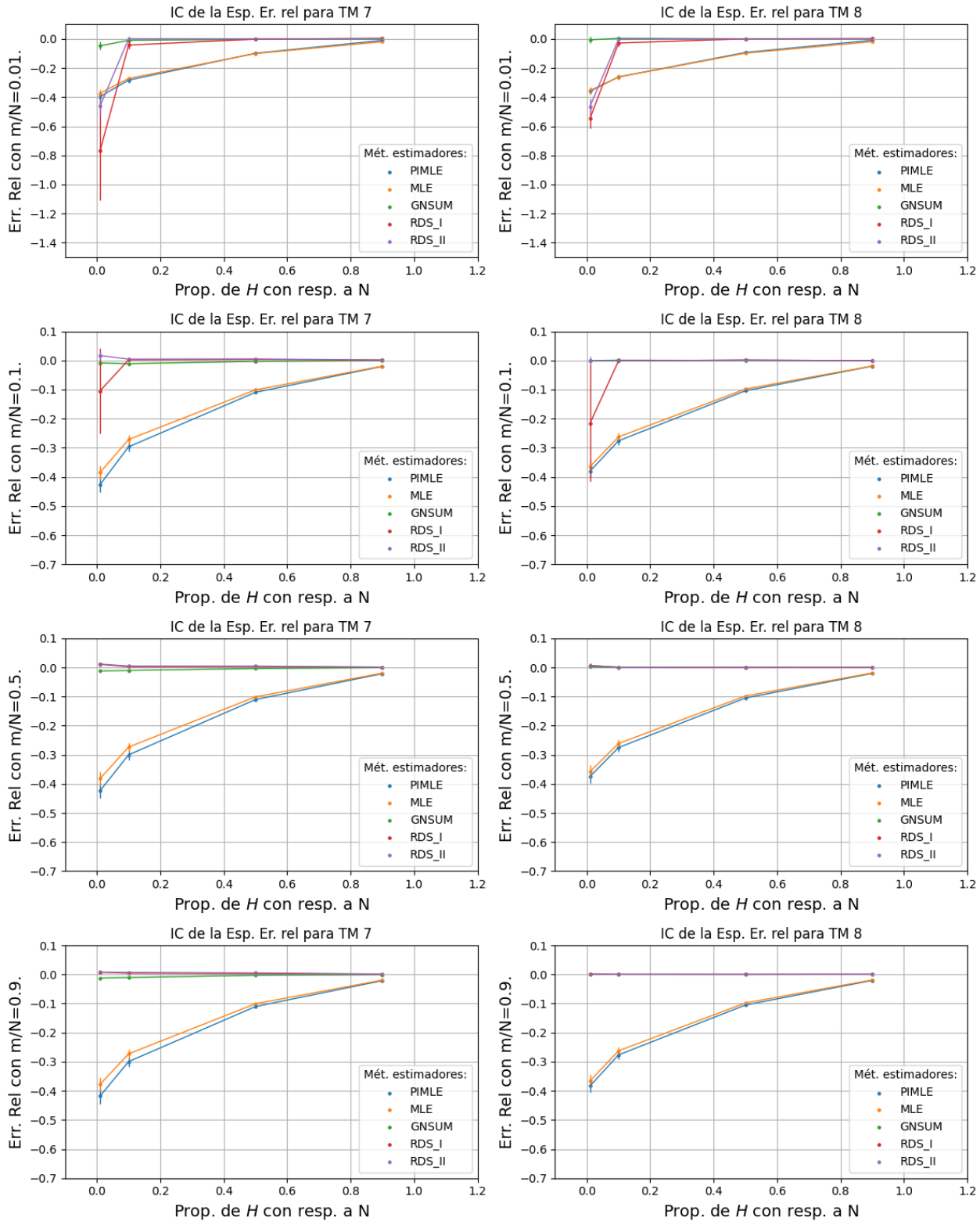


Figura B0.24: Muestra las dos últimas de la figura B0.22.

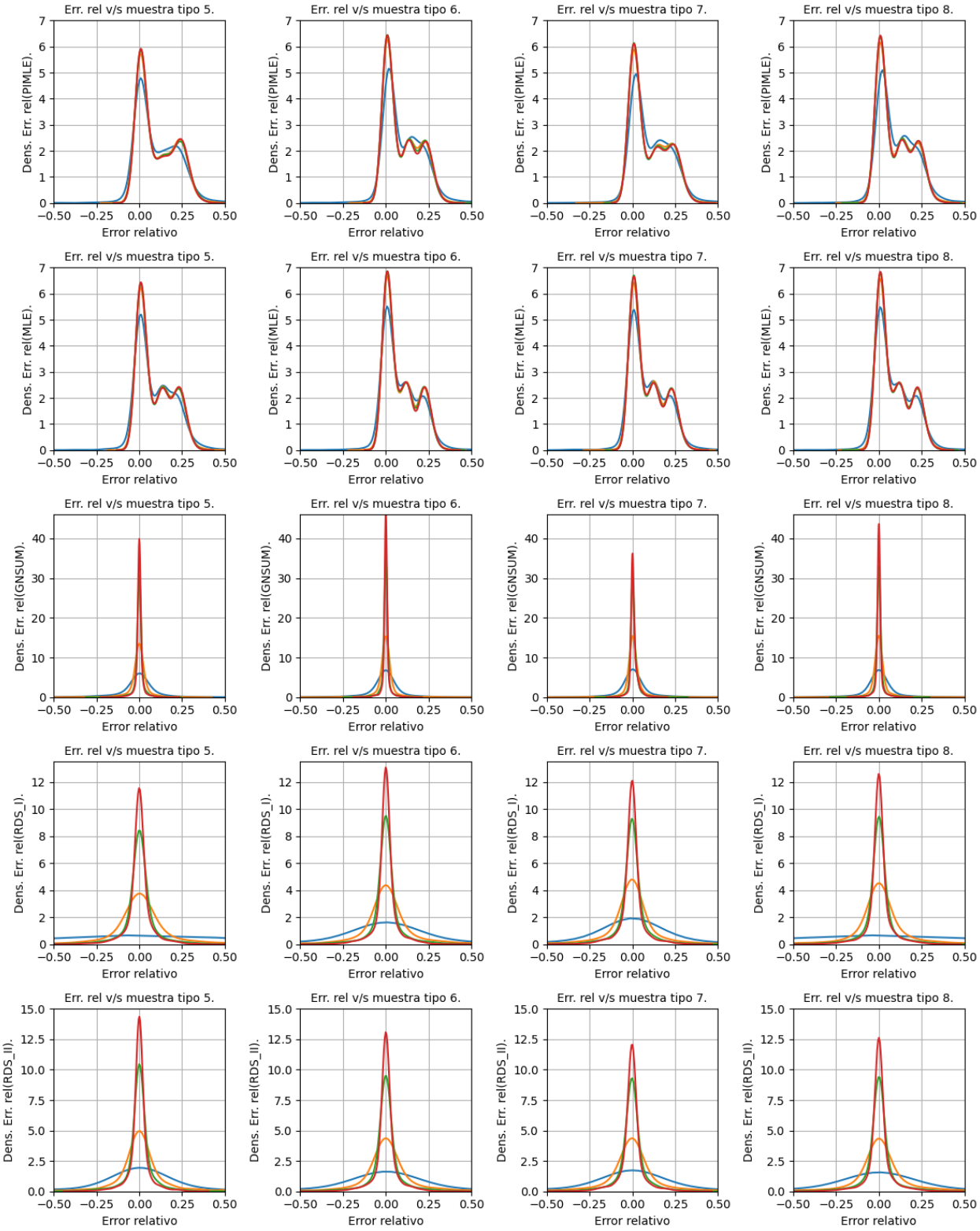


Figura B0.25: Experimento 5: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

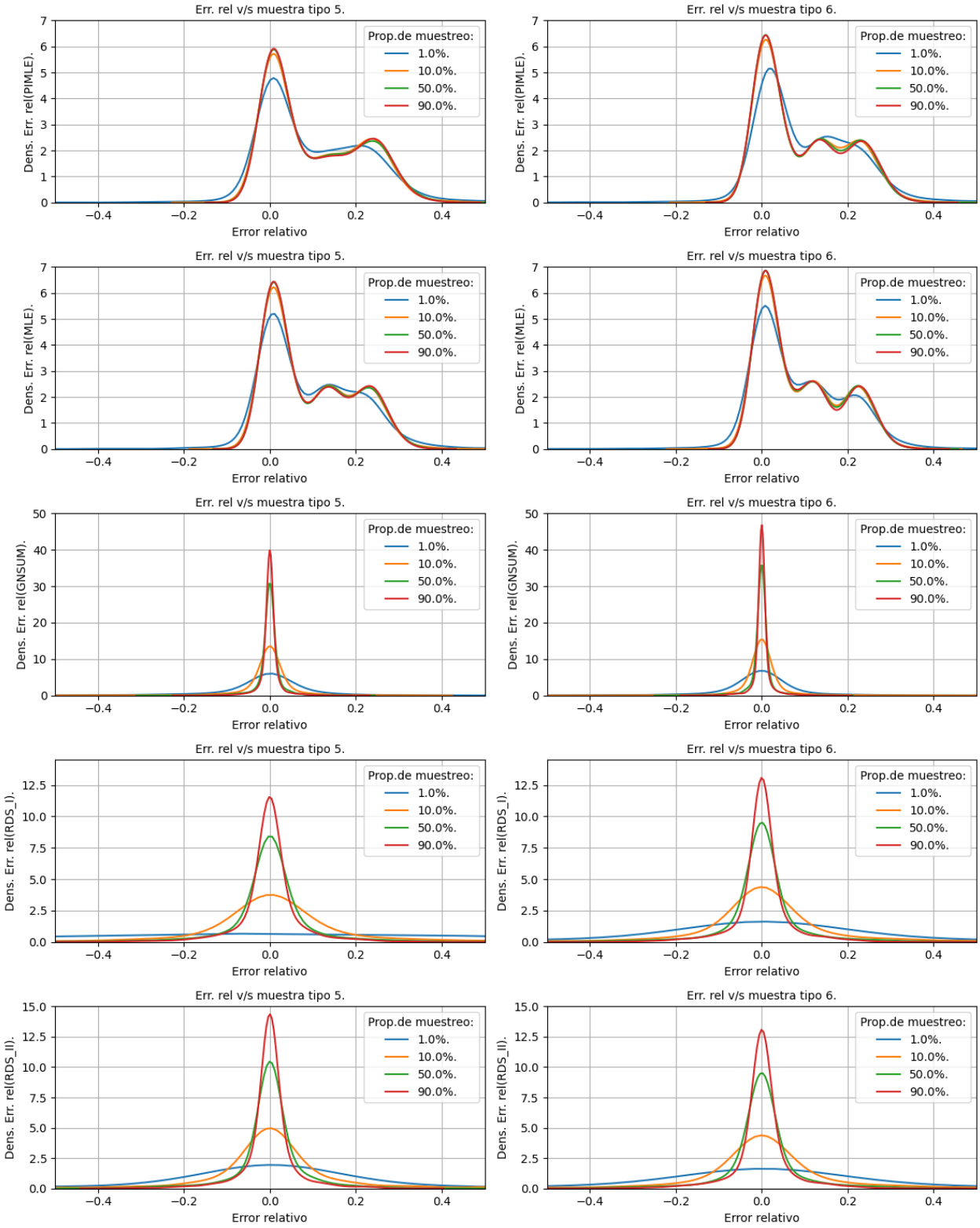


Figura B0.26: Muestra las dos primeras columnas de la figura B0.25.

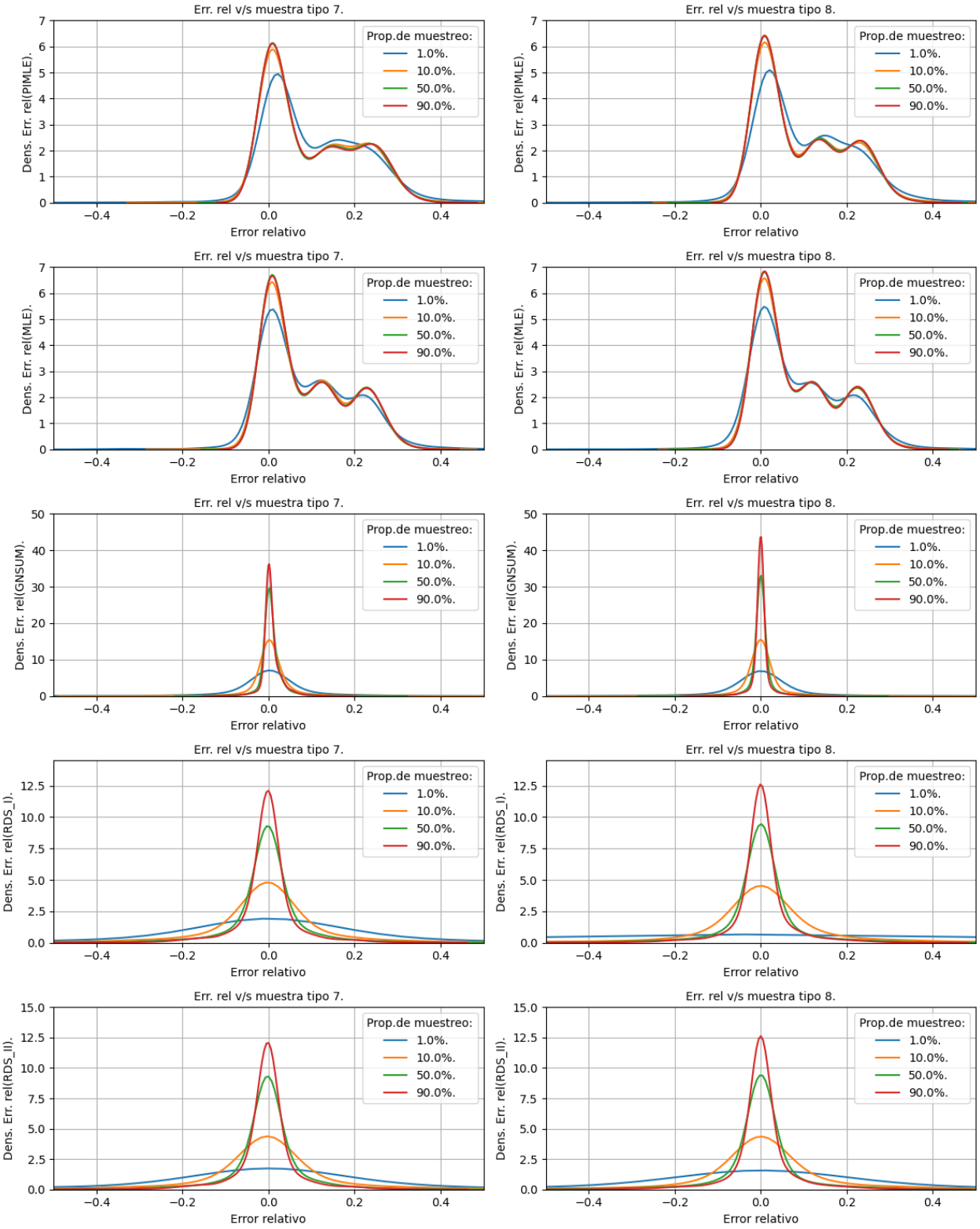


Figura B0.27: Muestra las dos últimas columnas de la figura B0.25.

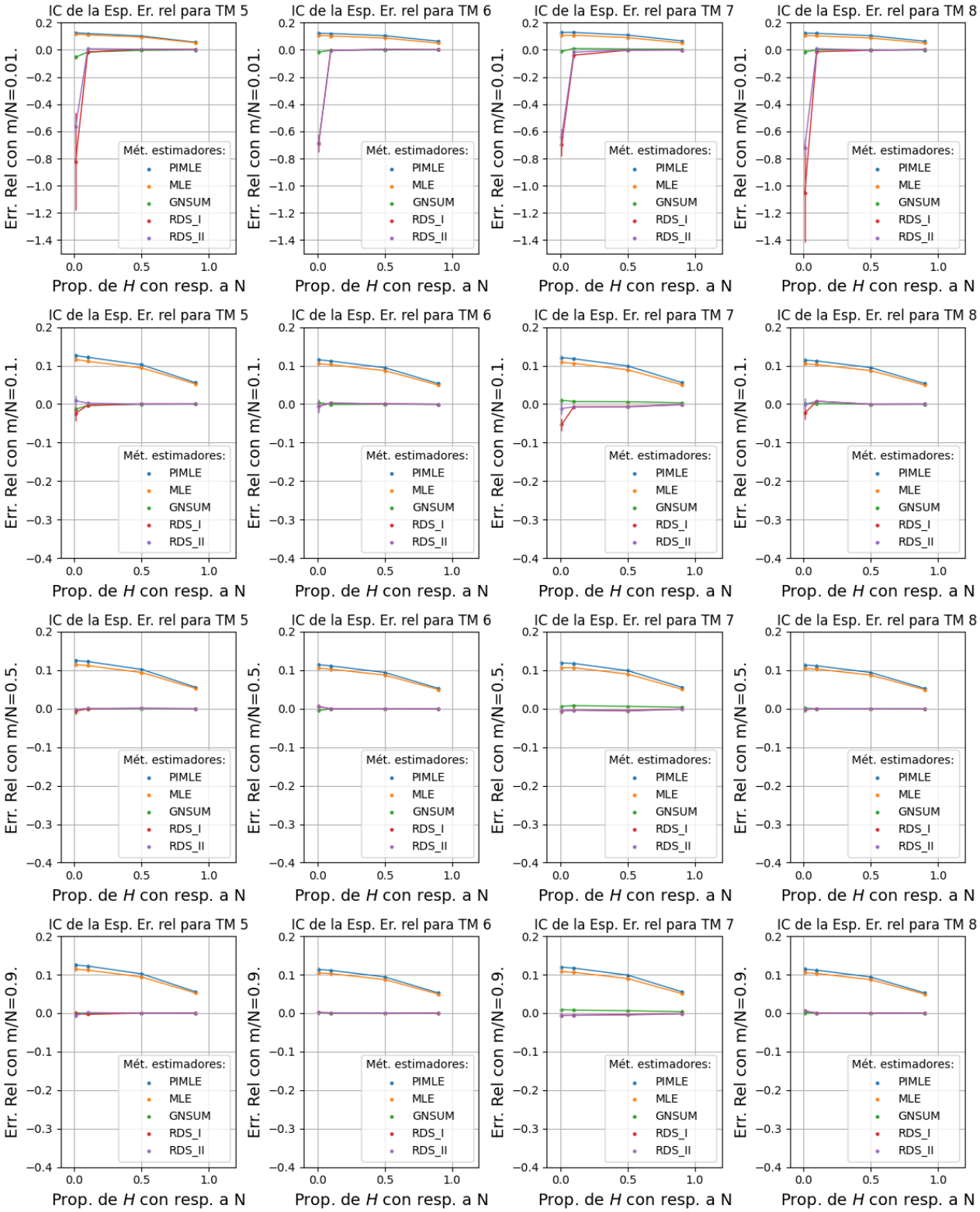


Figura B0.28: Experimento 5: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de H con respecto a N y fijando la proporción de nodos muestreados y el tipo de muestreo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia la proporción de nodos muestreados y cada columna varia el tipo de muestreo.

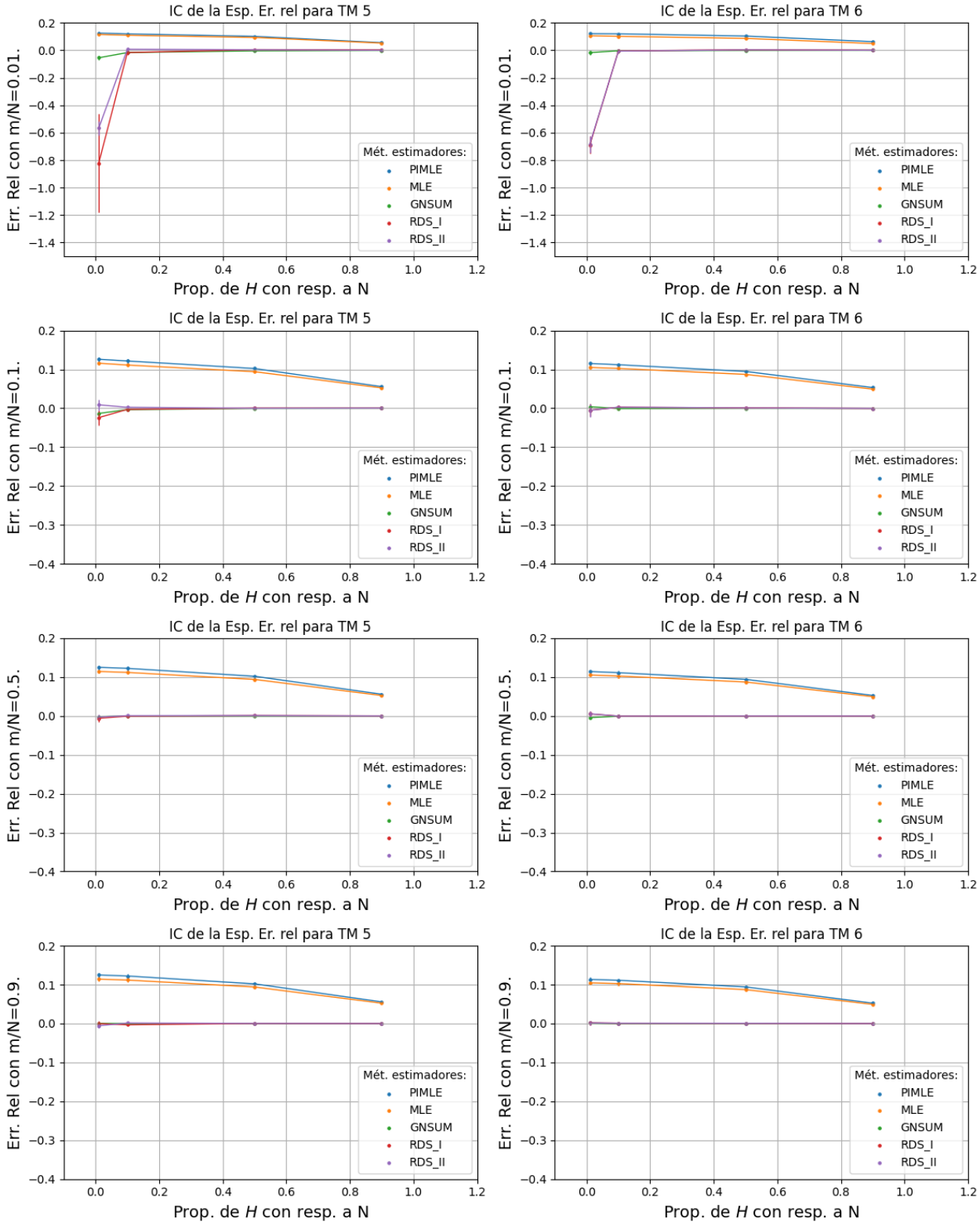


Figura B0.29: Muestra las dos primeras columnas de la figura B0.28.

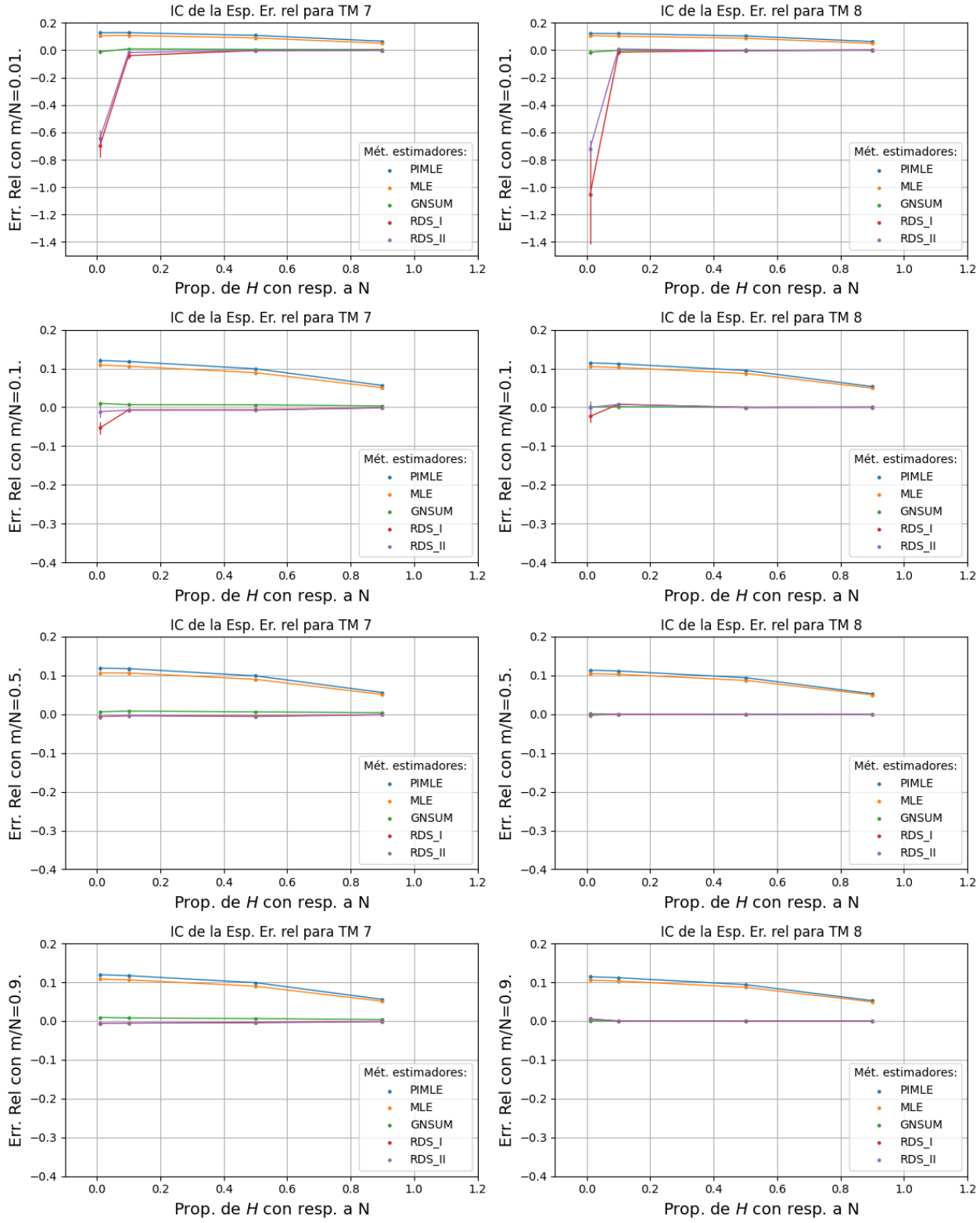


Figura B0.30: Muestra las dos últimas columnas de la figura B0.28.

Apéndice C

Resultados de experimentación con grupos reales

Simbología de tipos de muestreos aleatorios:

- Tipo de muestreo 5 se refiere al muestreo aleatorio uniforme con repetición.
- Tipo de muestreo 6 se refiere al muestreo aleatorio asociado a una caminata aleatoria.
- Tipo de muestreo 7 se refiere al muestreo aleatorio asociado a un proceso viral probabilístico.
- Tipo de muestreo 8 se refiere al muestreo aleatorio asociado a un proceso viral constante.

Simbología de grupos estimados:

- Grupo 0 canales de twitch con menos de 520 suscriptores.
- Grupo 1 canales de twitch que recién comienzan.
- Grupo 2 canales de twitch que tienen tiempo de vida menos de 600 (no especifica medida).
- Grupo 3 canales de twitch que están inactivos.
- Grupo 4 canales de twitch con transmisiones en inglés.
- Grupo 5 canales de twitch con transmisiones en francés.

- Grupo 6 canales de twitch con transmisiones en sur coreano.
- Grupo 7 canales de twitch con transmisiones en sur japonés.

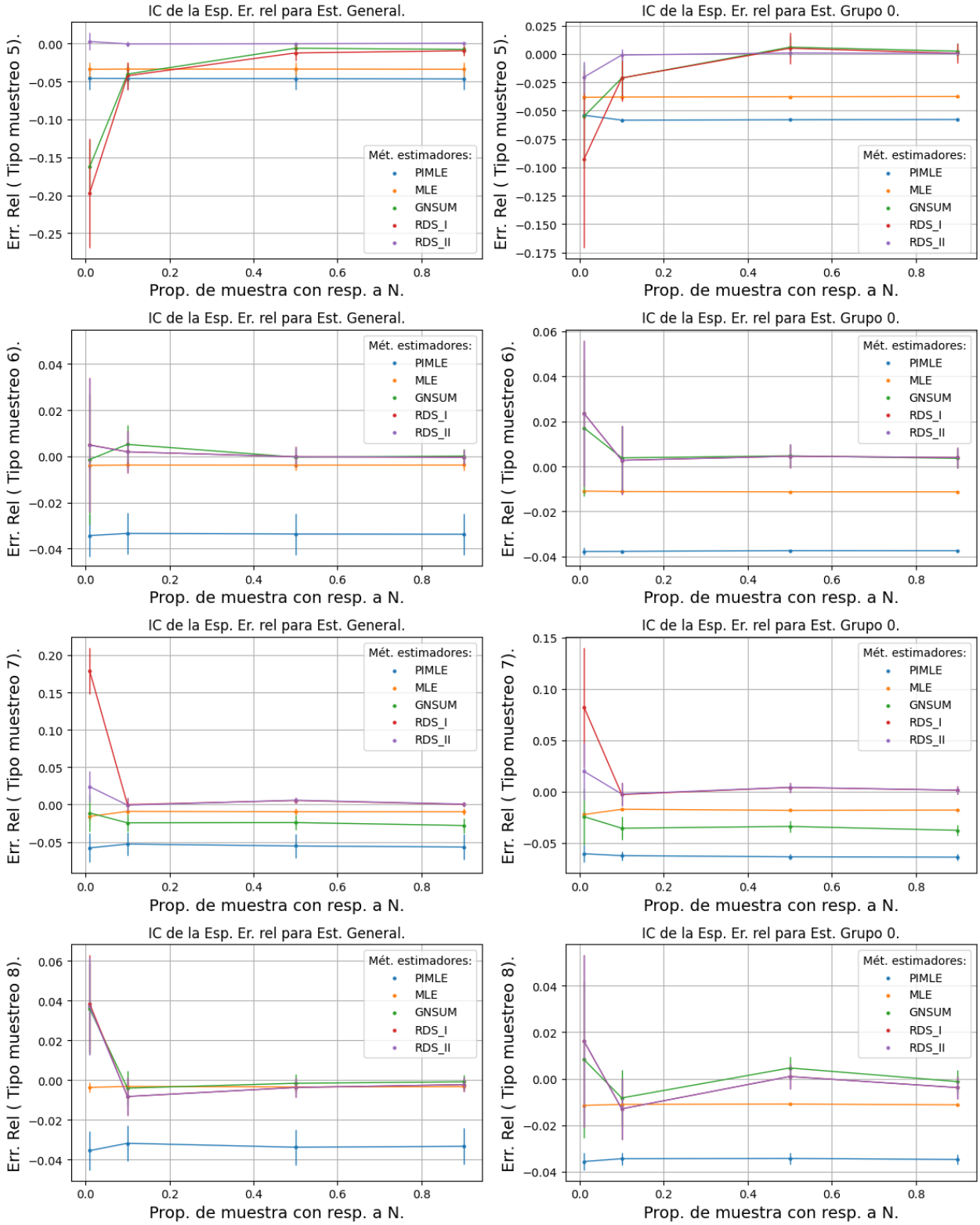


Figura C0.1: Resultados Estimación general y grupo 0: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia el grupo real a estimar.

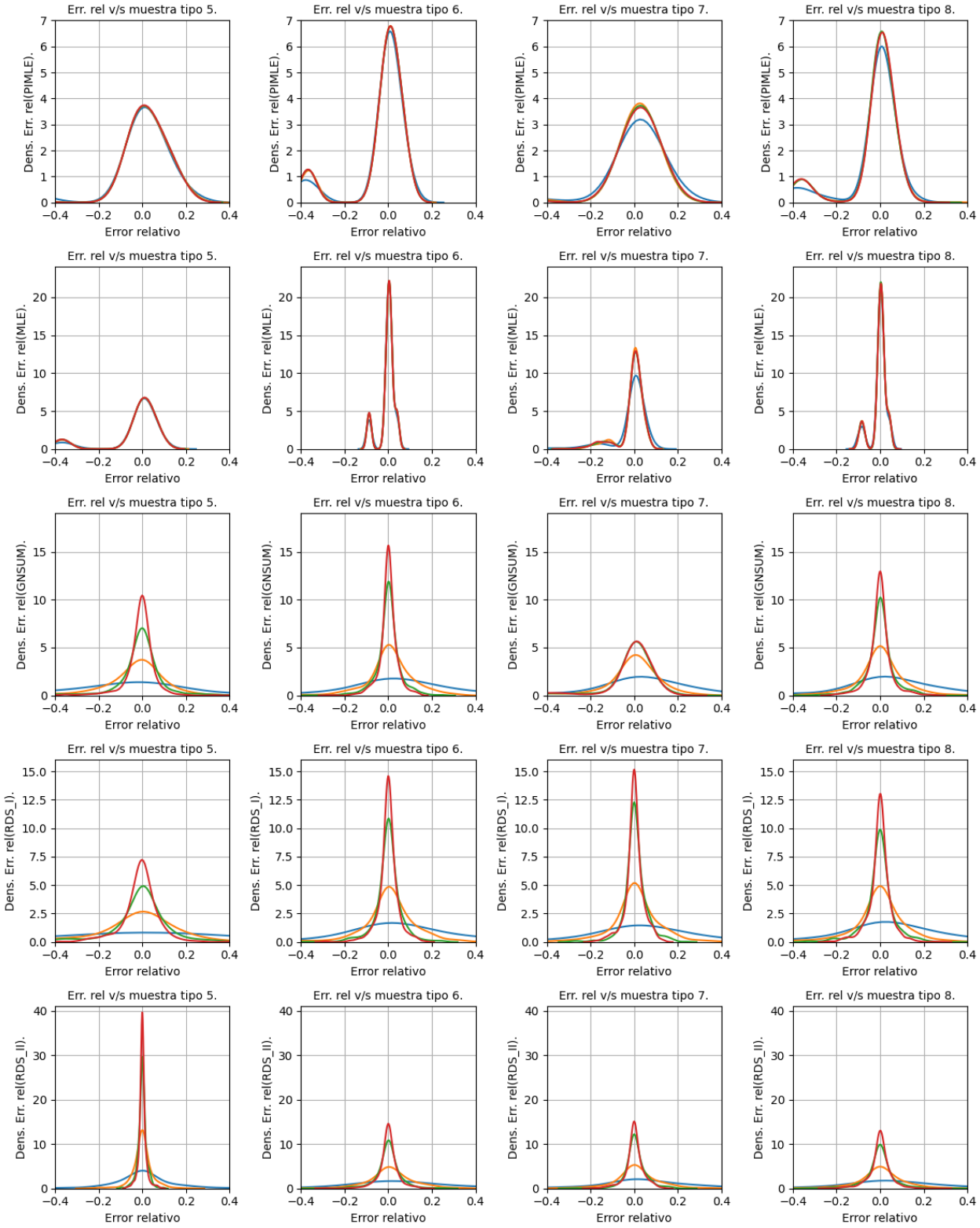


Figura C0.2: Resultados estimación grupo en general: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

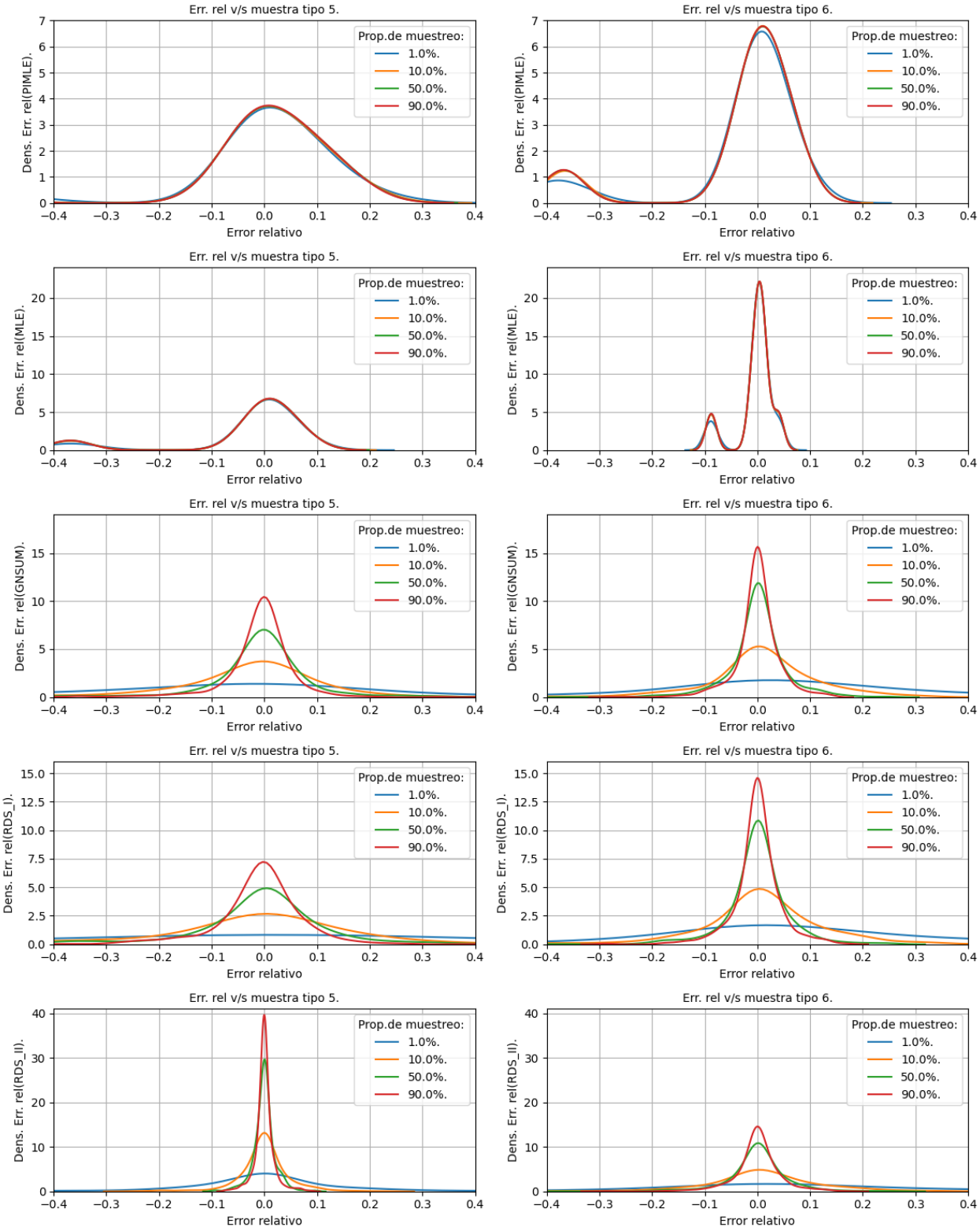


Figura C0.3: Muestra las dos primeras columnas de la figura C0.2.

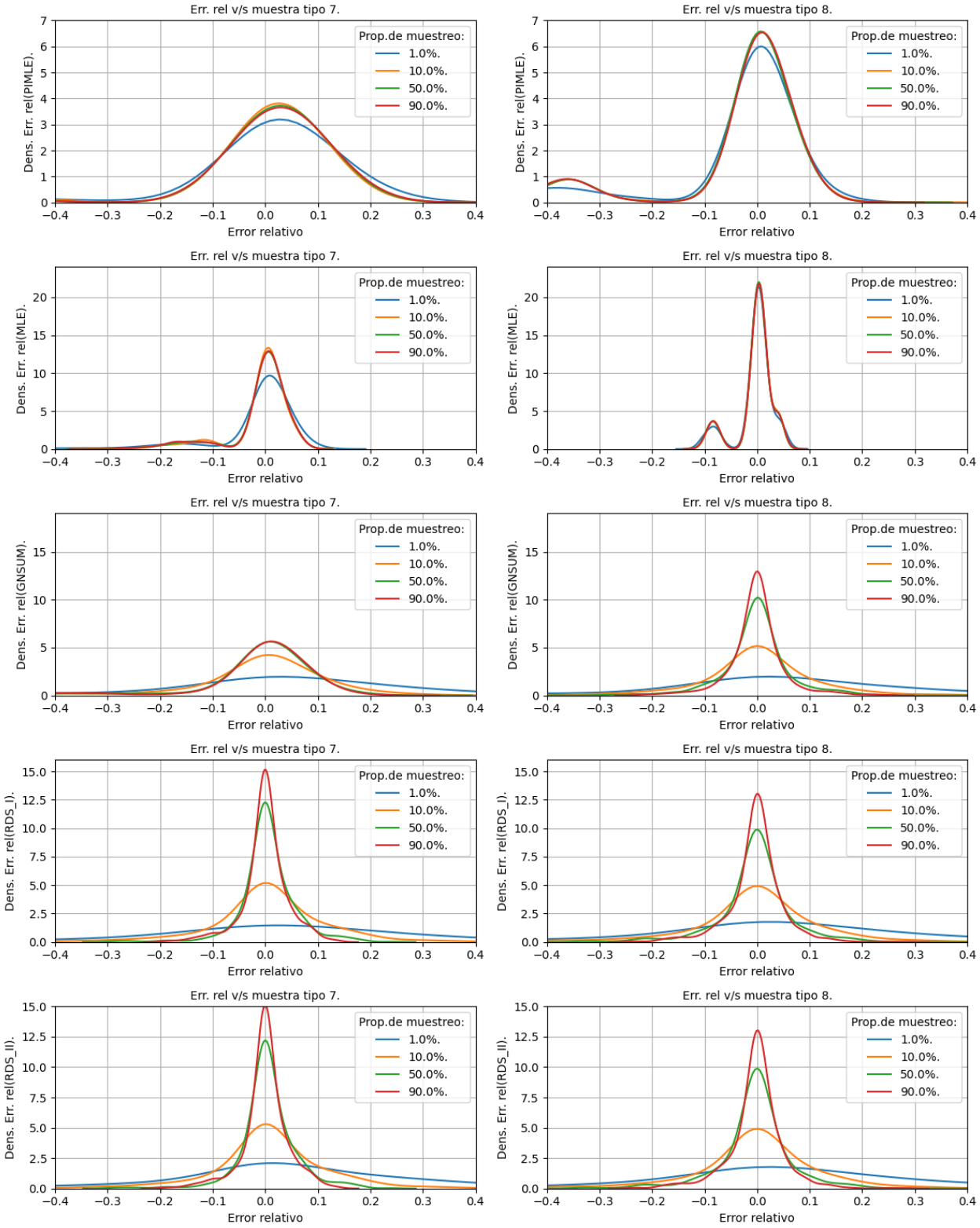


Figura C0.4: Muestra las dos últimas columnas de la figura C0.2.

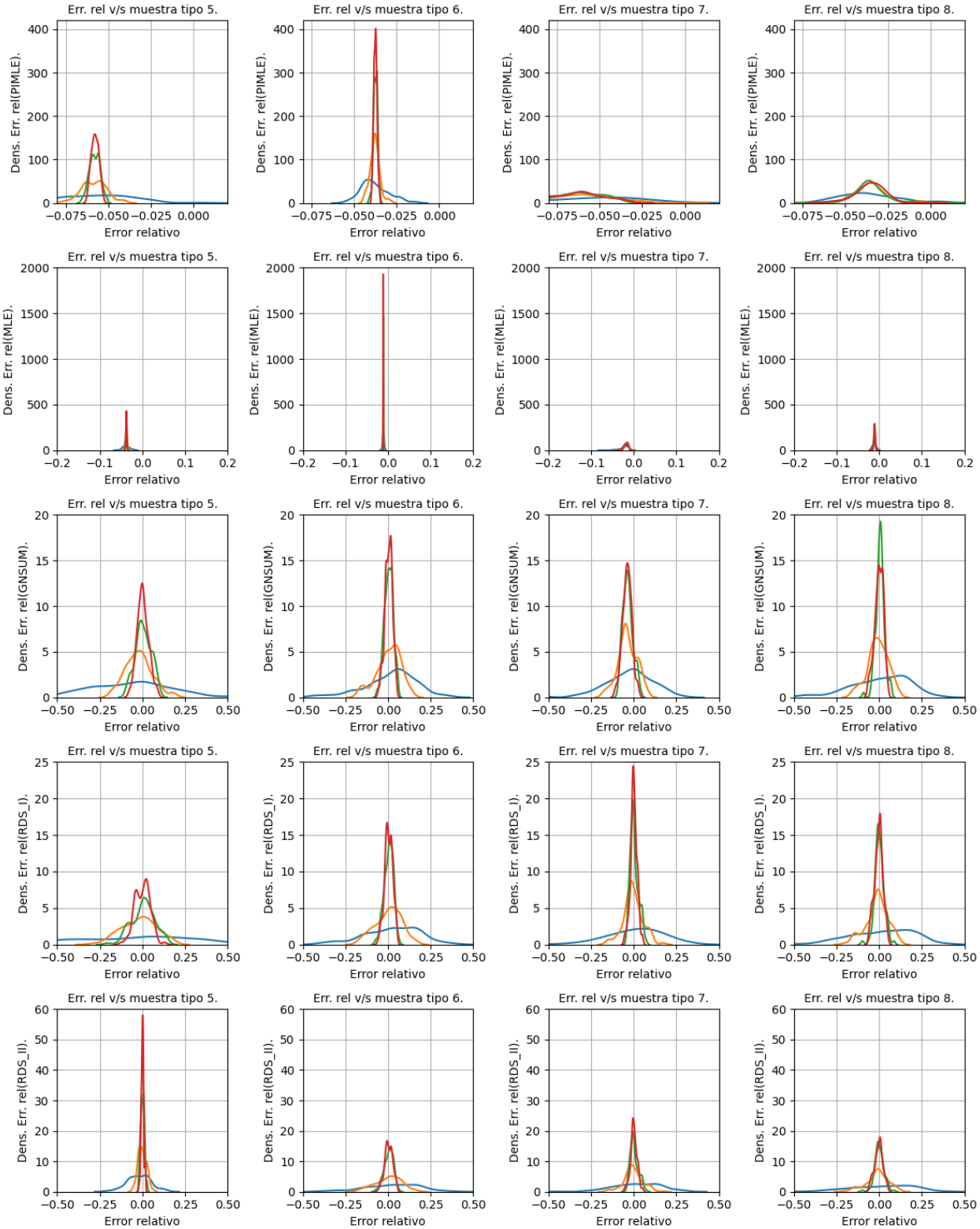


Figura C0.5: Resultados estimación grupo 0: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

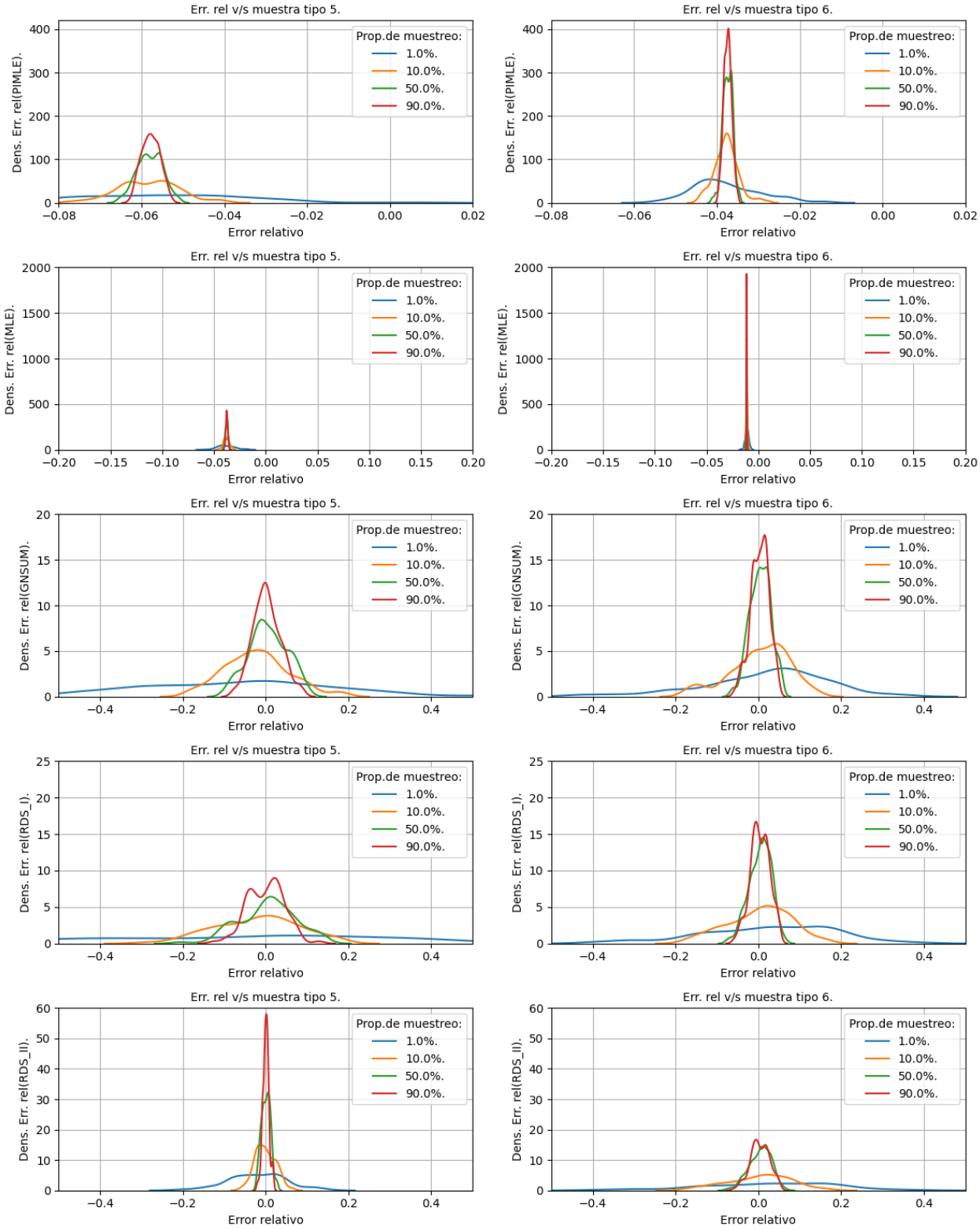


Figura C0.6: Muestra las dos primeras columnas de la figura C0.5.

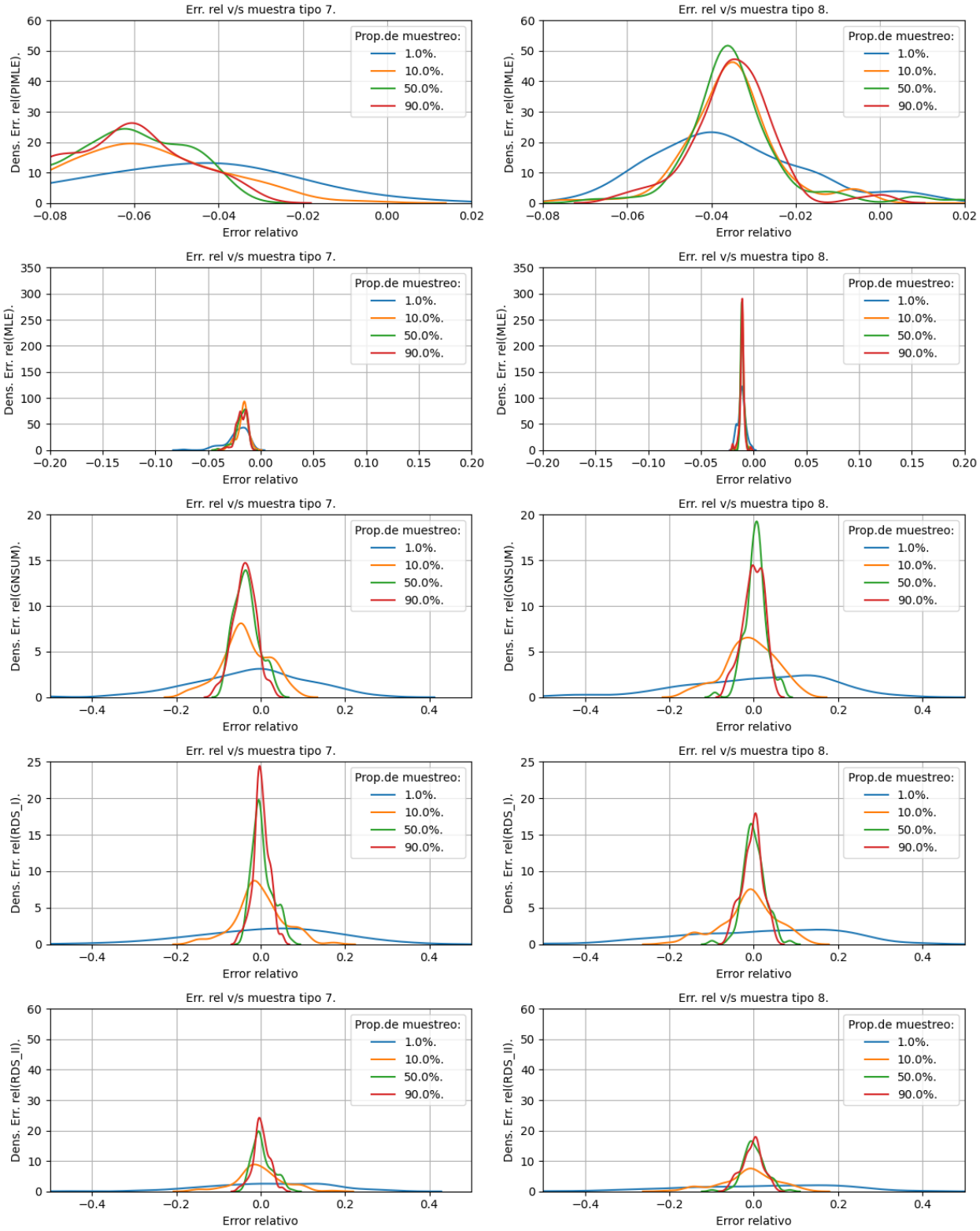


Figura C0.7: Muestra las dos últimas columnas de la figura C0.5.

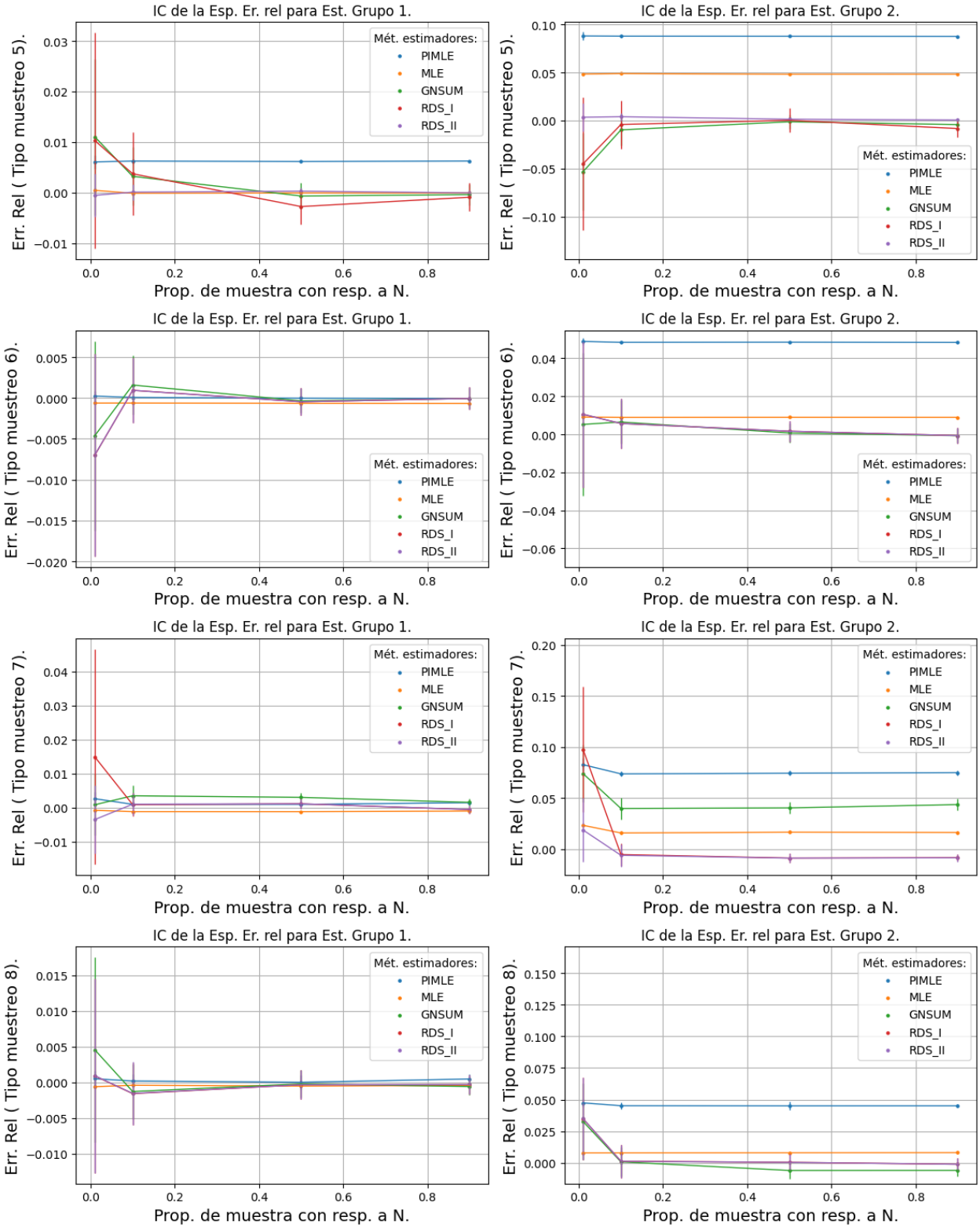


Figura C0.8: Resultados estimaciones de los grupos 1 y 2: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia el grupo real a estimar.

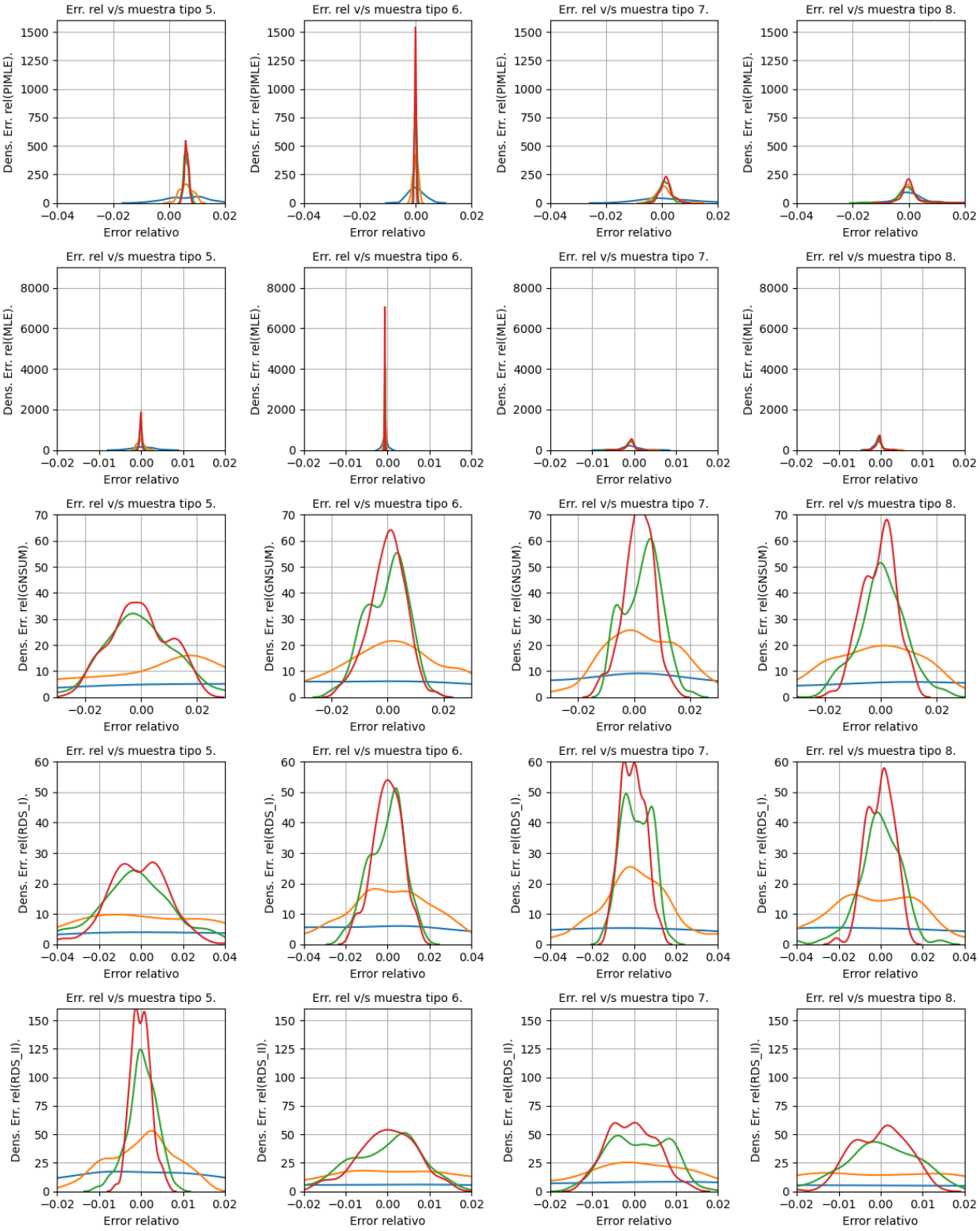


Figura C0.9: Resultados estimación grupo 1: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

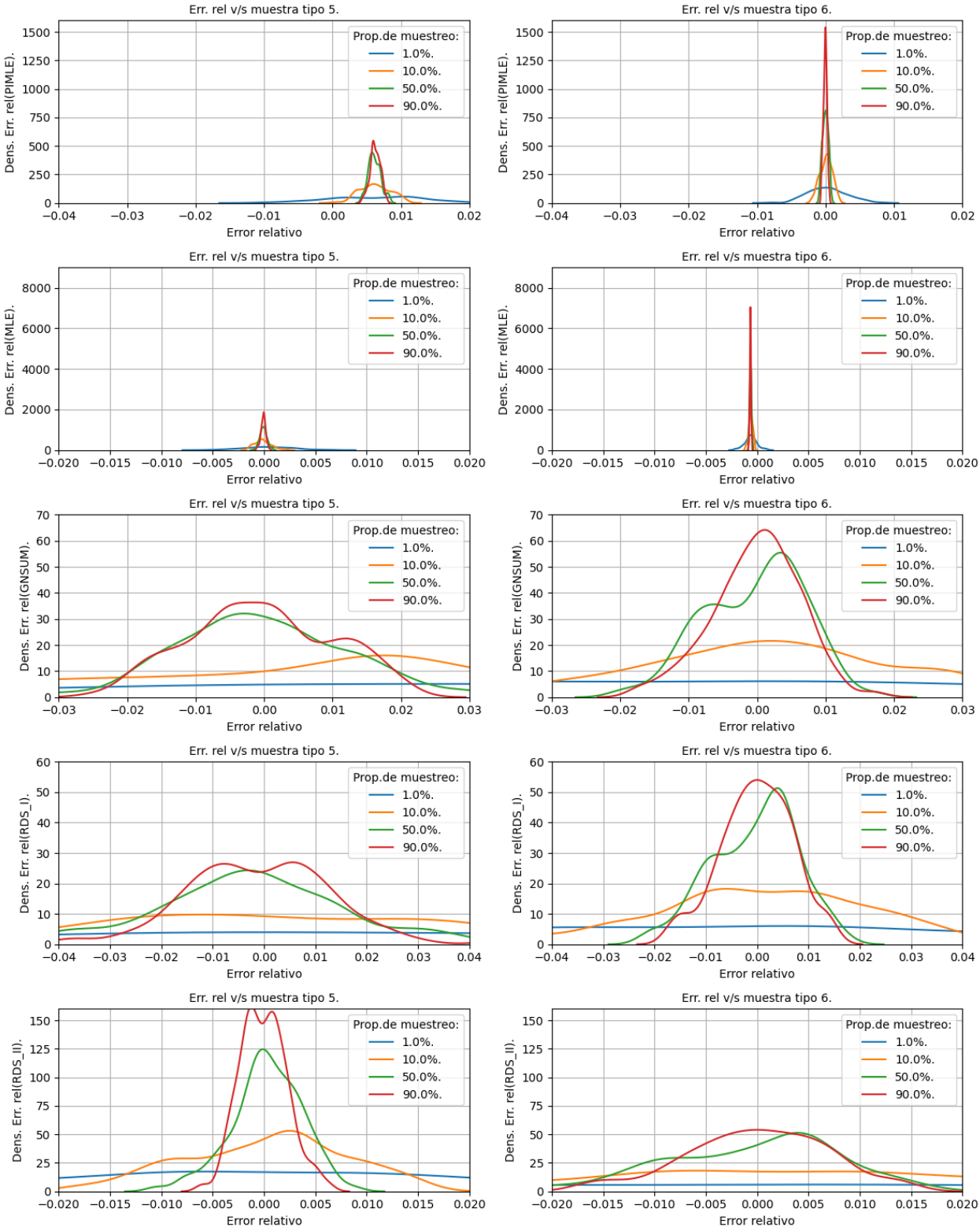


Figura C0.10: Muestra las dos primeras columnas de la figura C0.9.

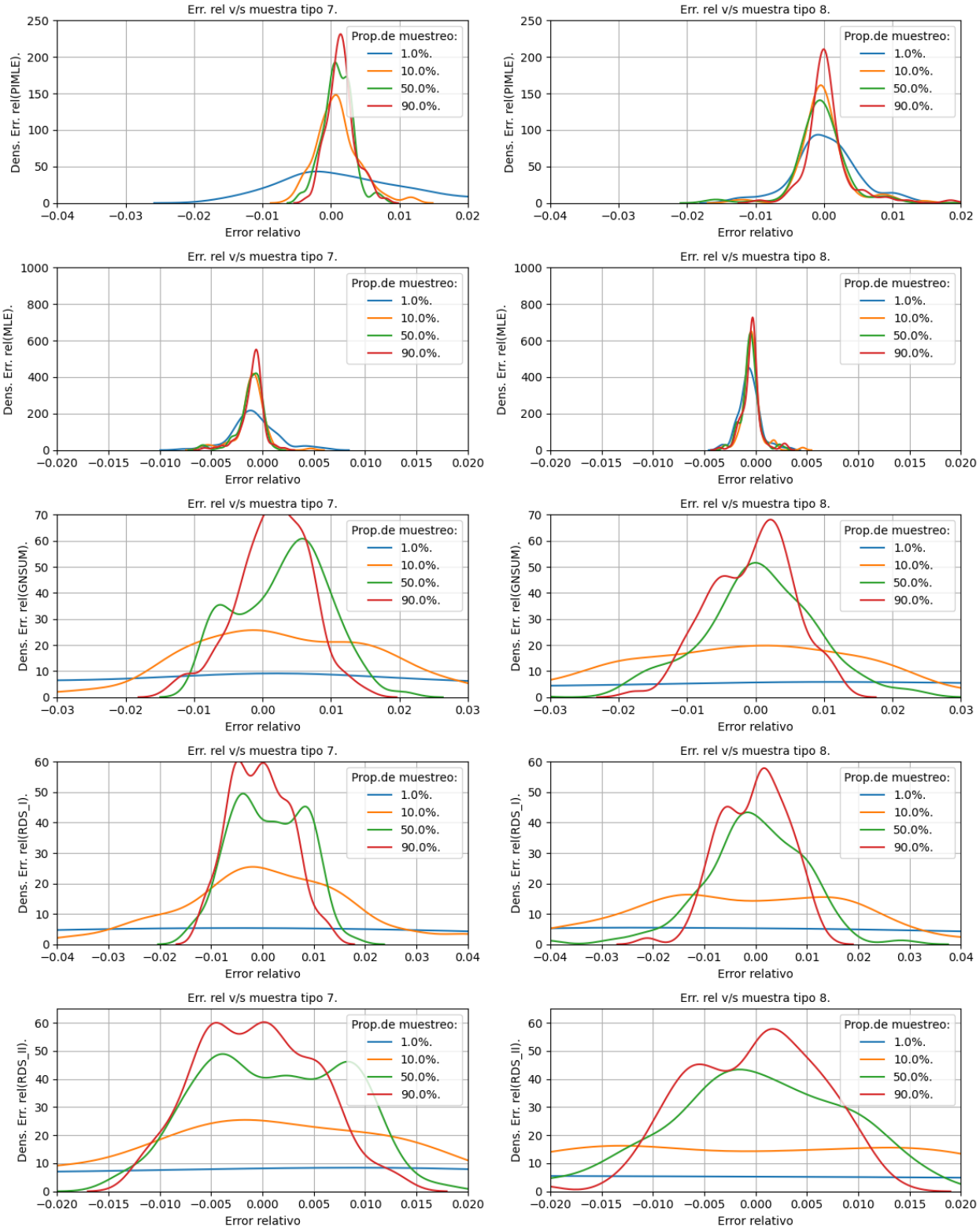


Figura C0.11: Muestra las dos últimas columnas de la figura C0.9.

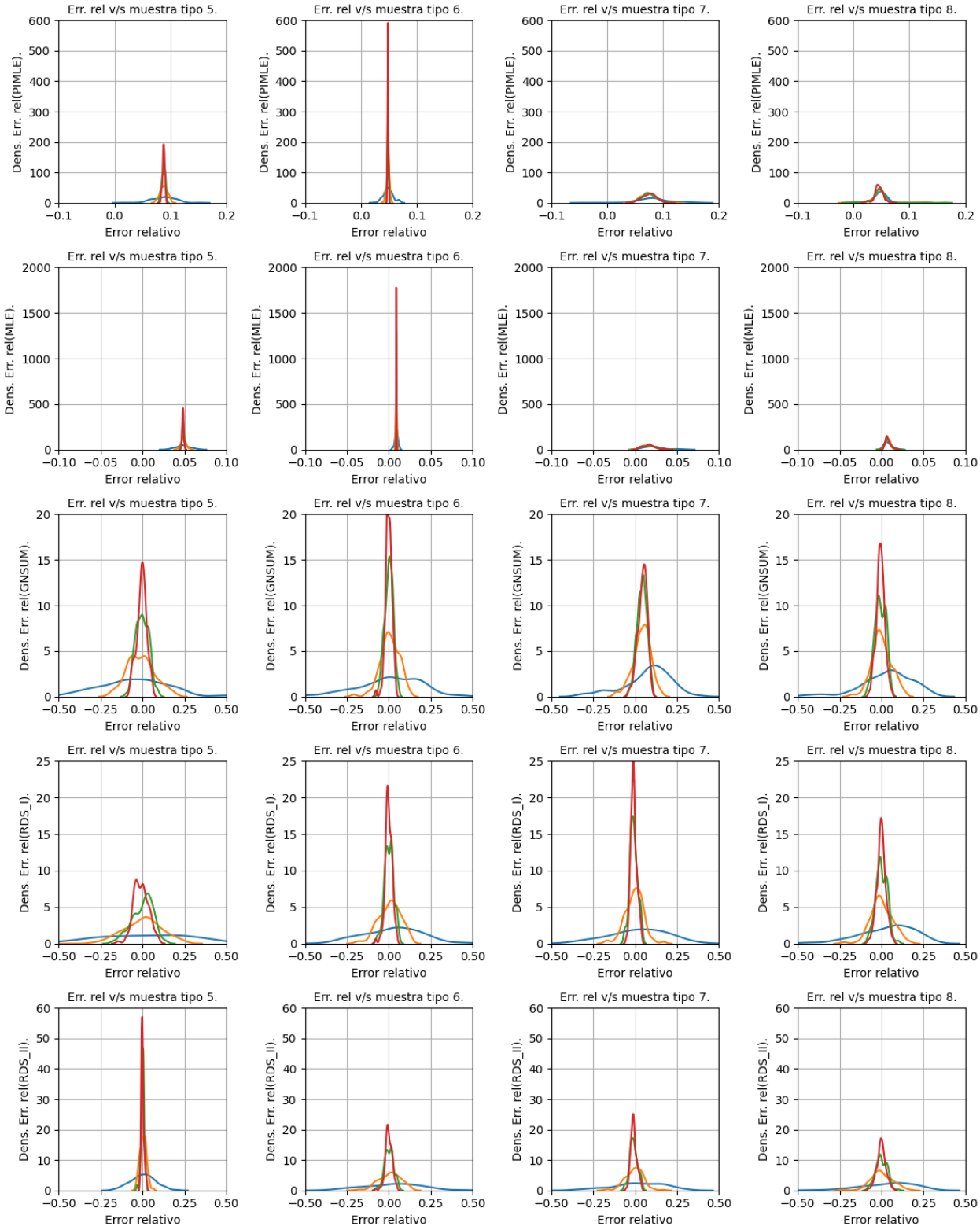


Figura C0.12: Resultados estimación grupo 2: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

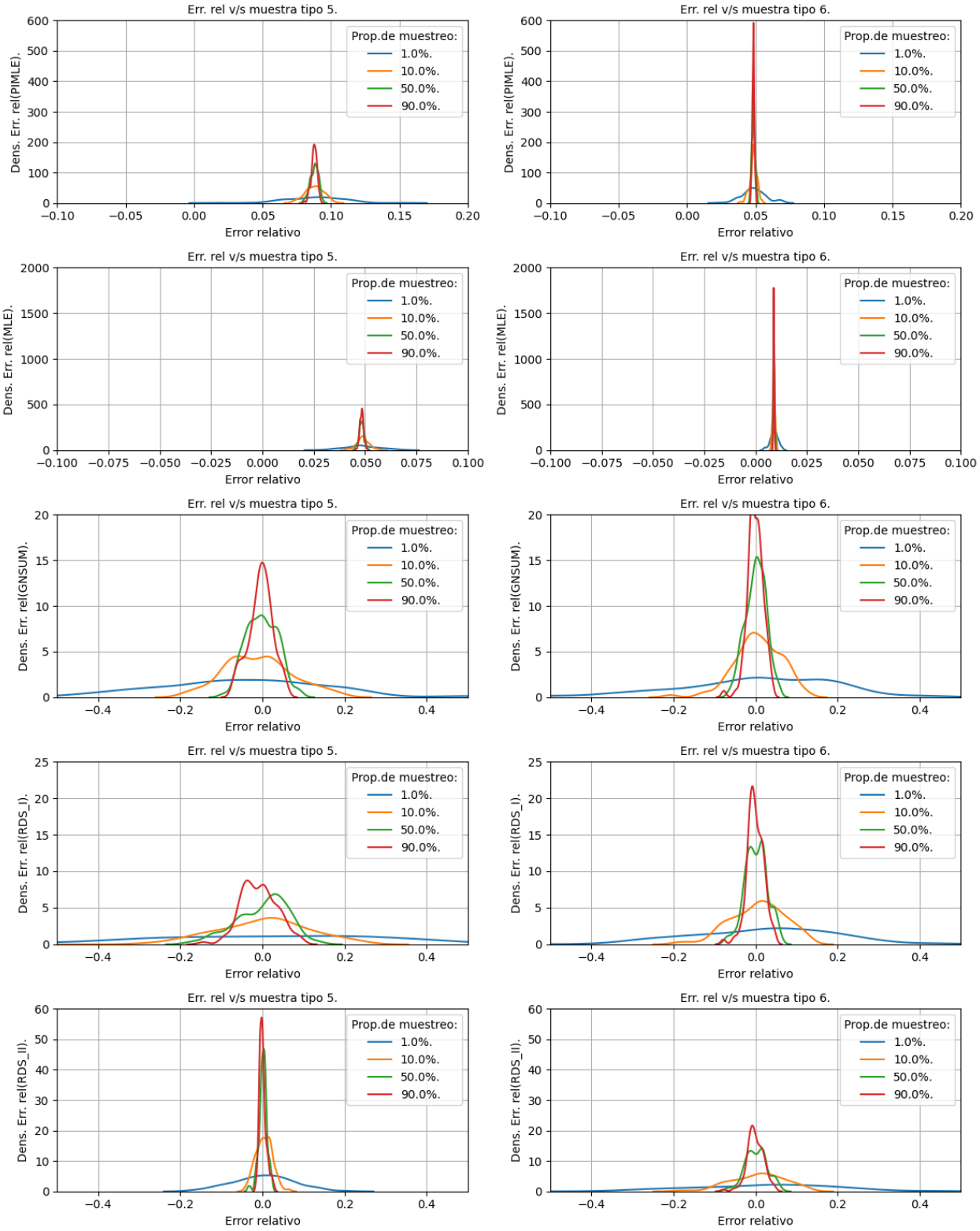


Figura C0.13: Muestra las dos primeras columnas de la figura C0.12.

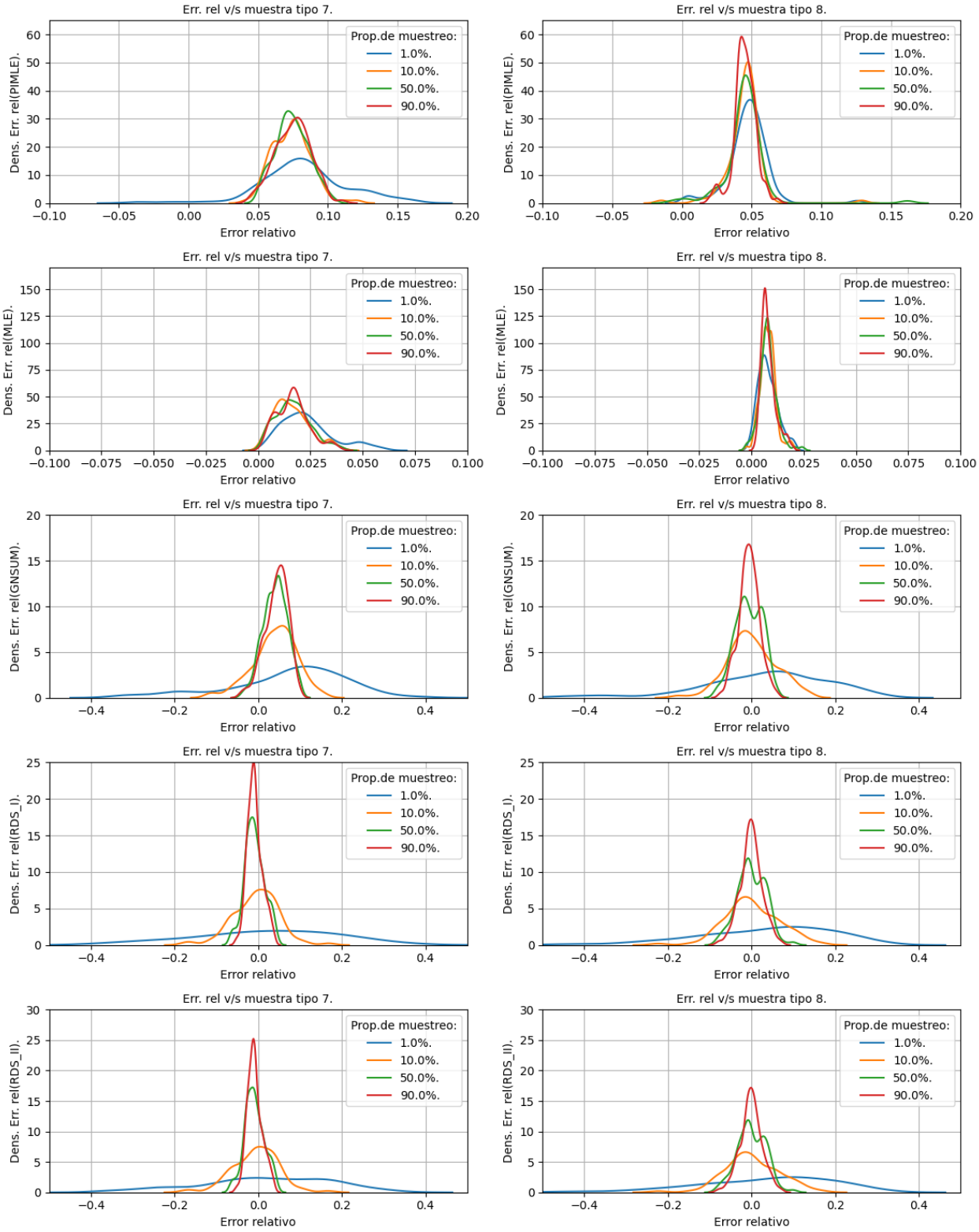


Figura C0.14: Muestra las dos últimas columnas de la figura C0.12.

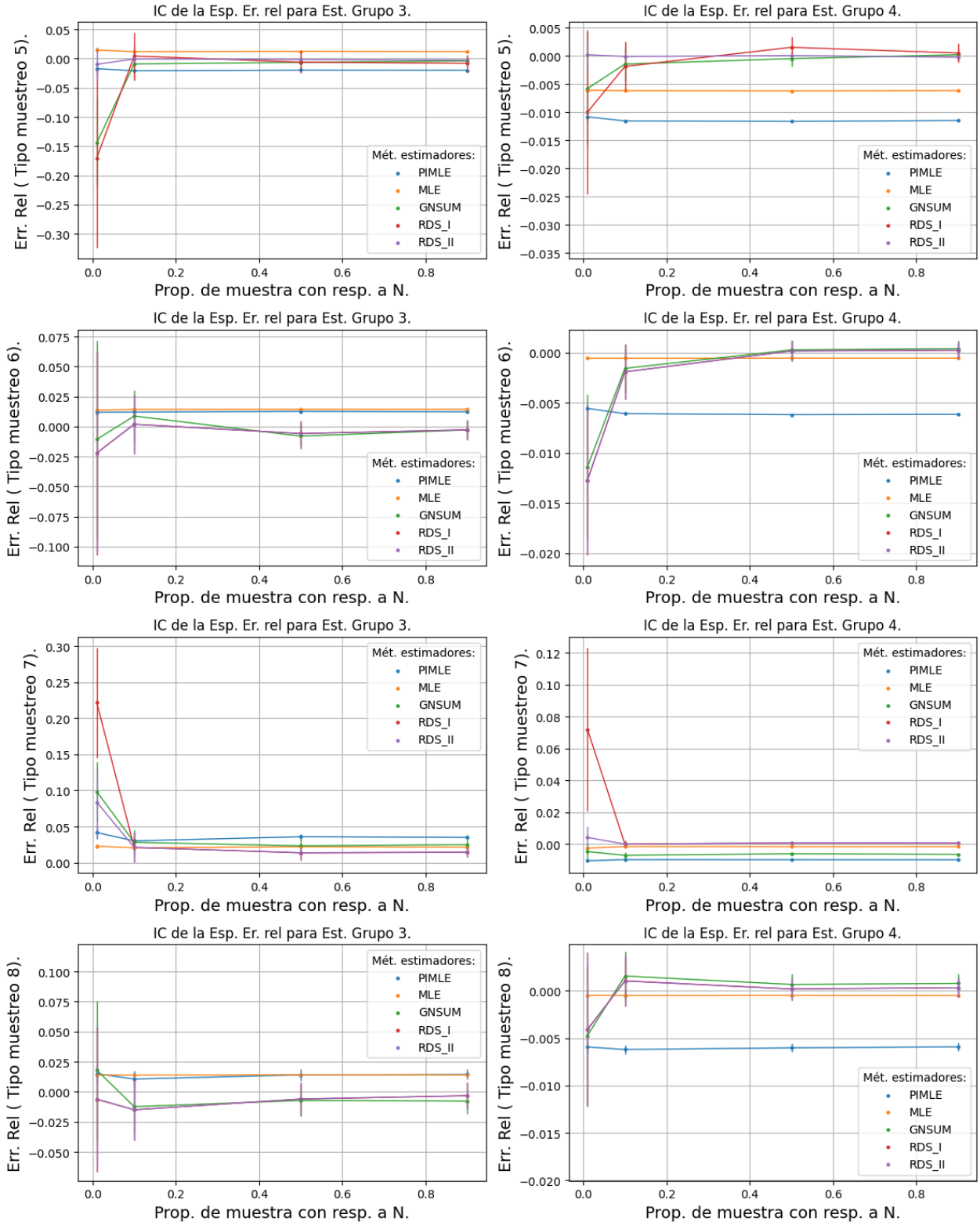


Figura C0.15: Resultados estimaciones de los grupos 3 y 4: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia columna varia el grupo real a estimar.

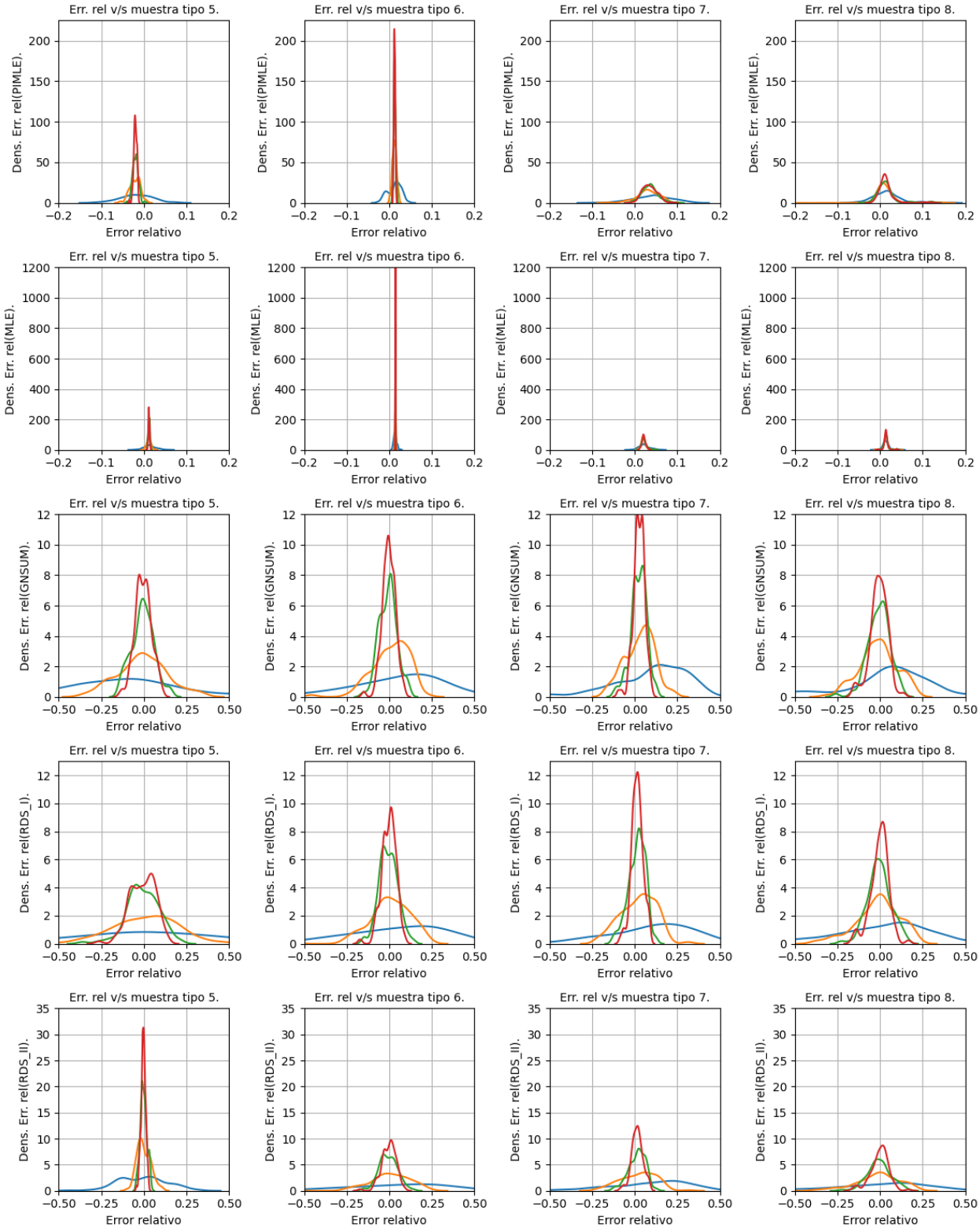


Figura C0.16: Resultados estimación grupo 3: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

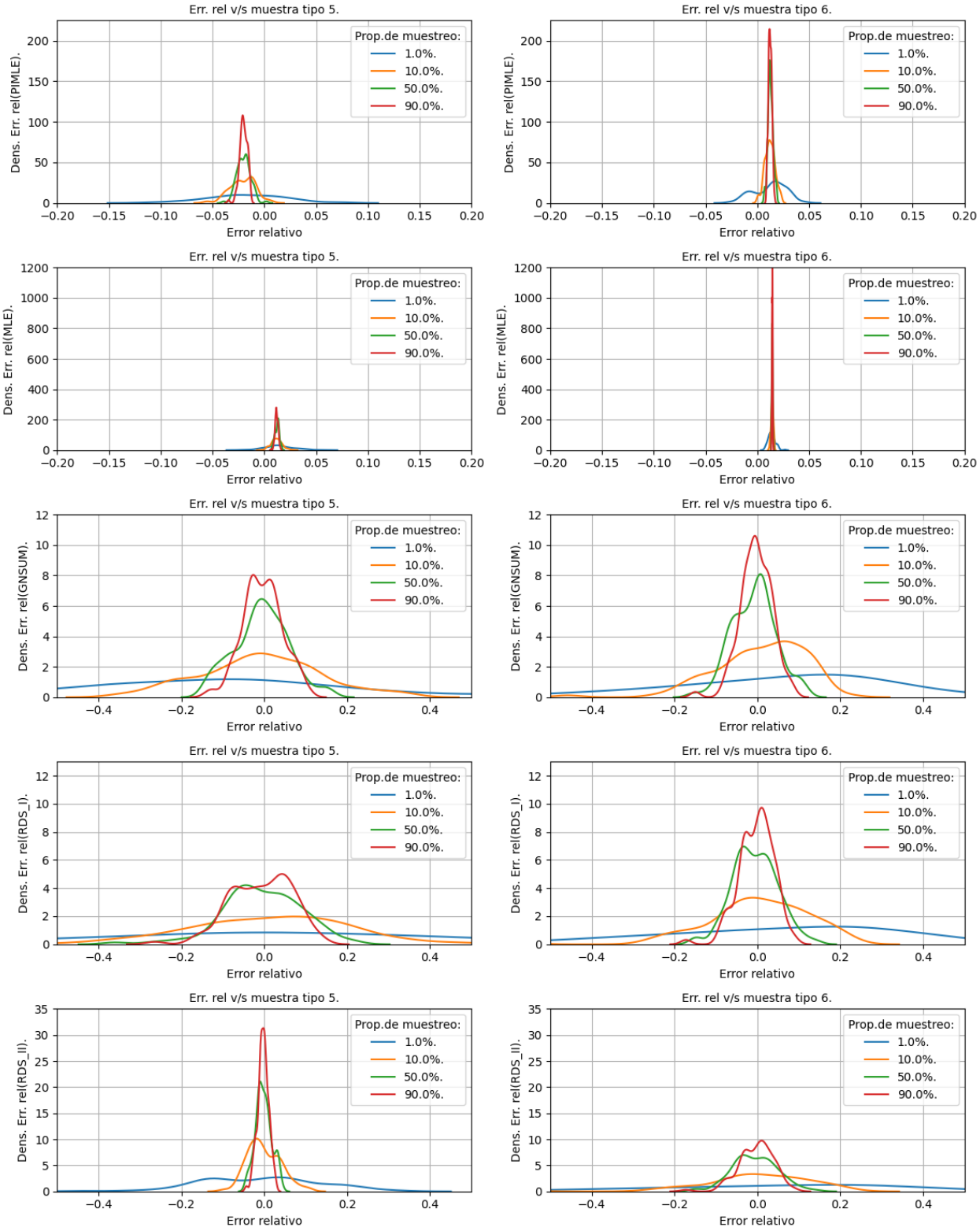


Figura C0.17: Muestra las dos primeras columnas de la figura C0.16.

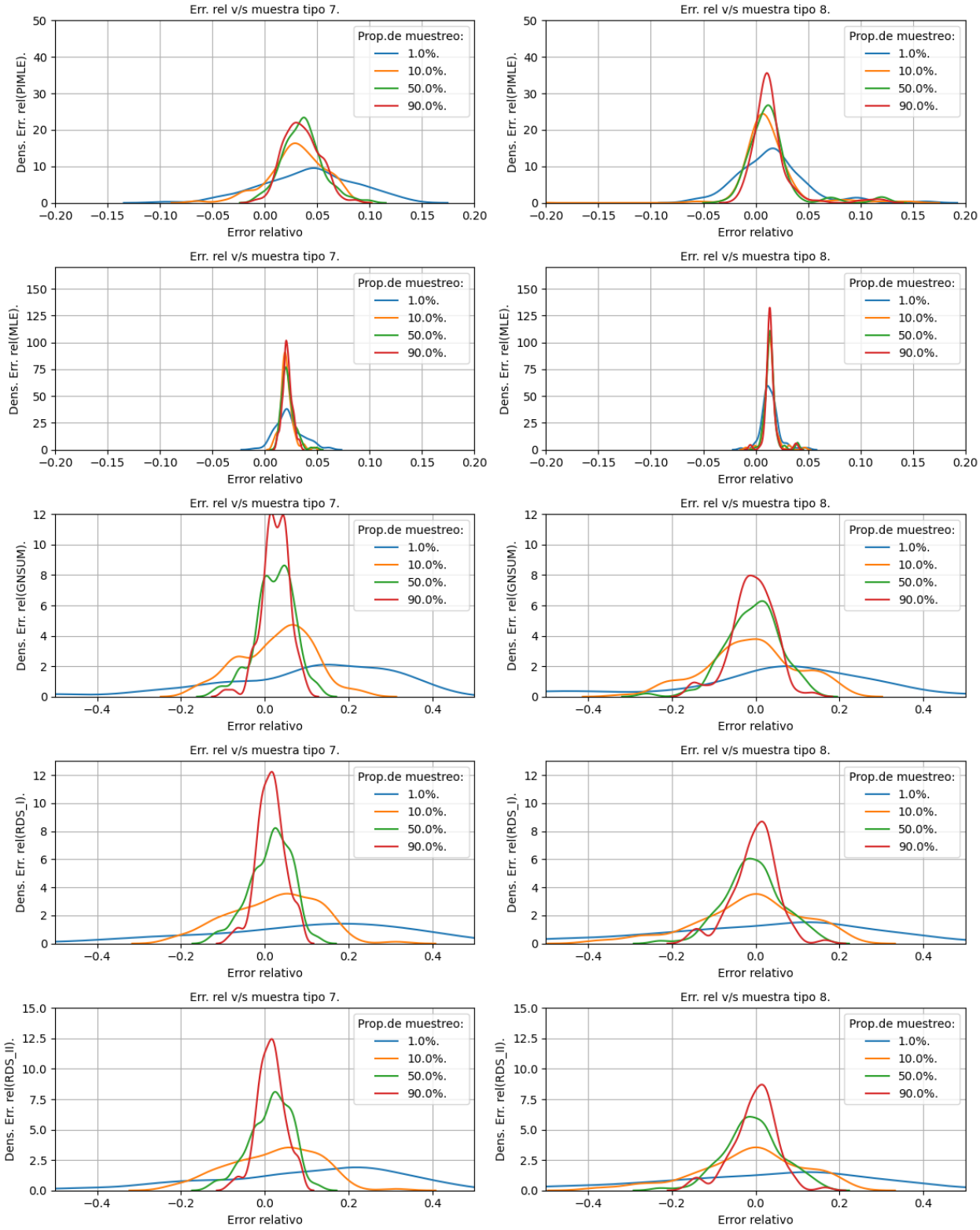


Figura C0.18: Muestra las dos últimas columnas de la figura C0.16.

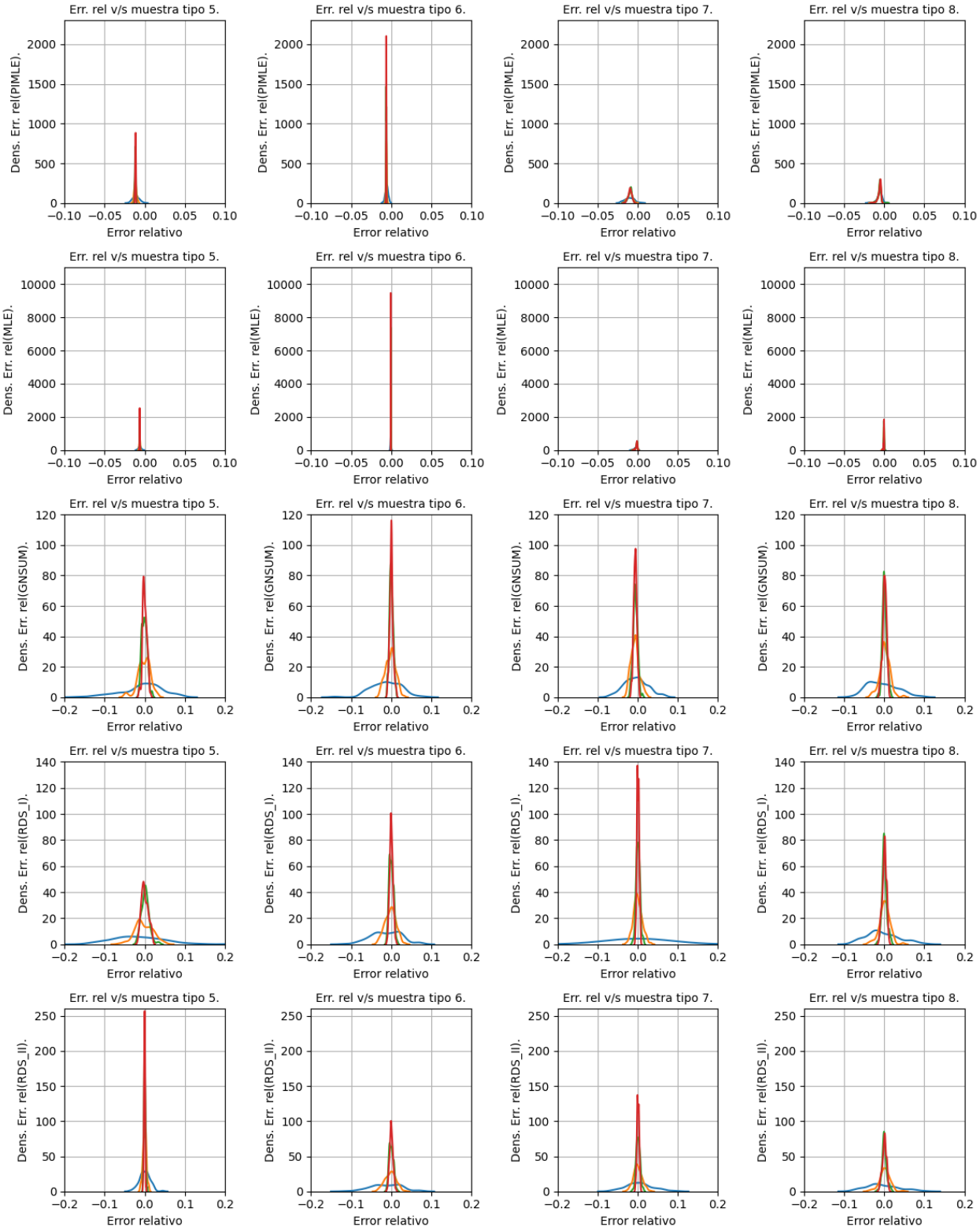


Figura C0.19: Resultados estimación grupo 4: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

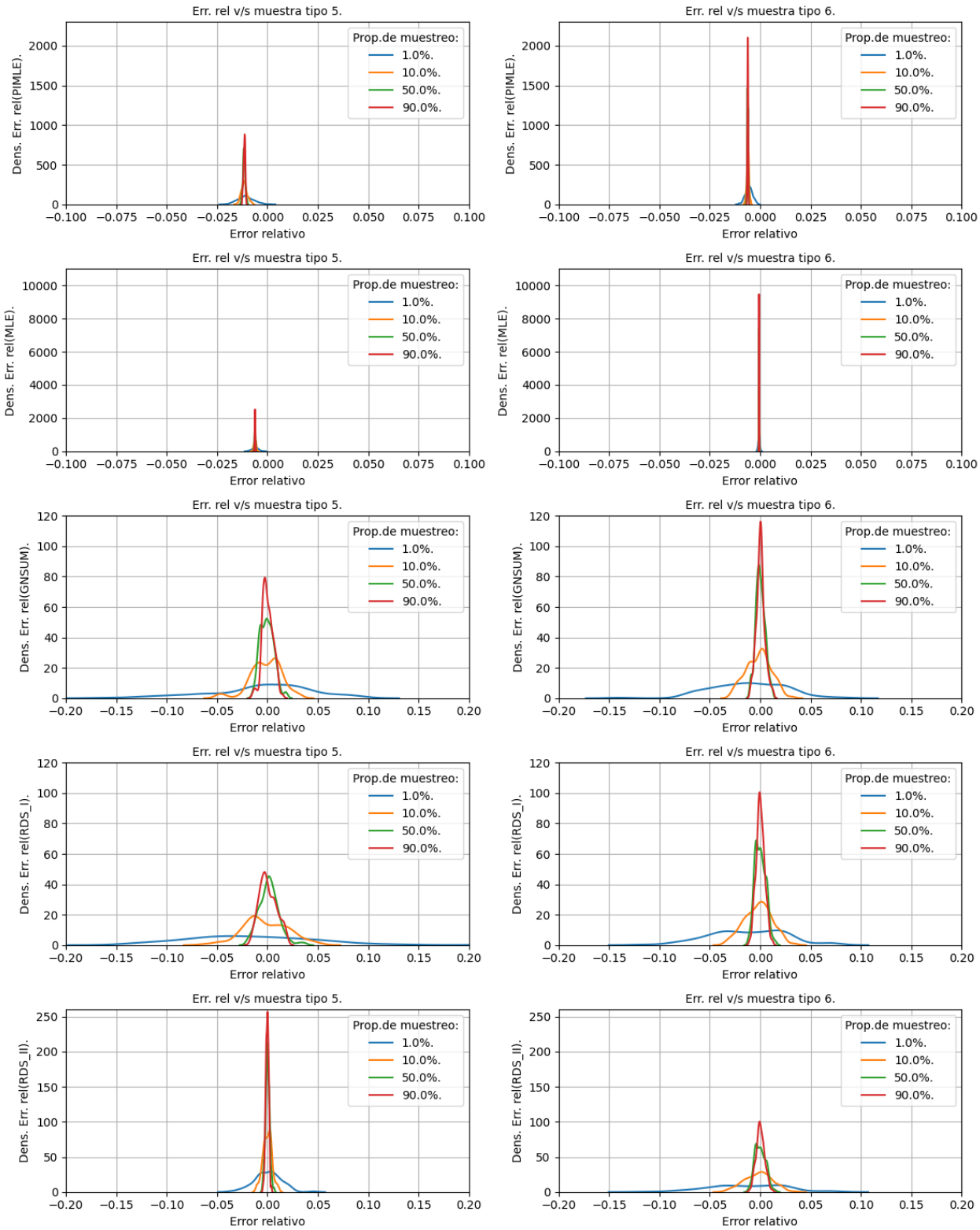


Figura C0.20: Muestra las dos primeras columnas de la figura C0.19.

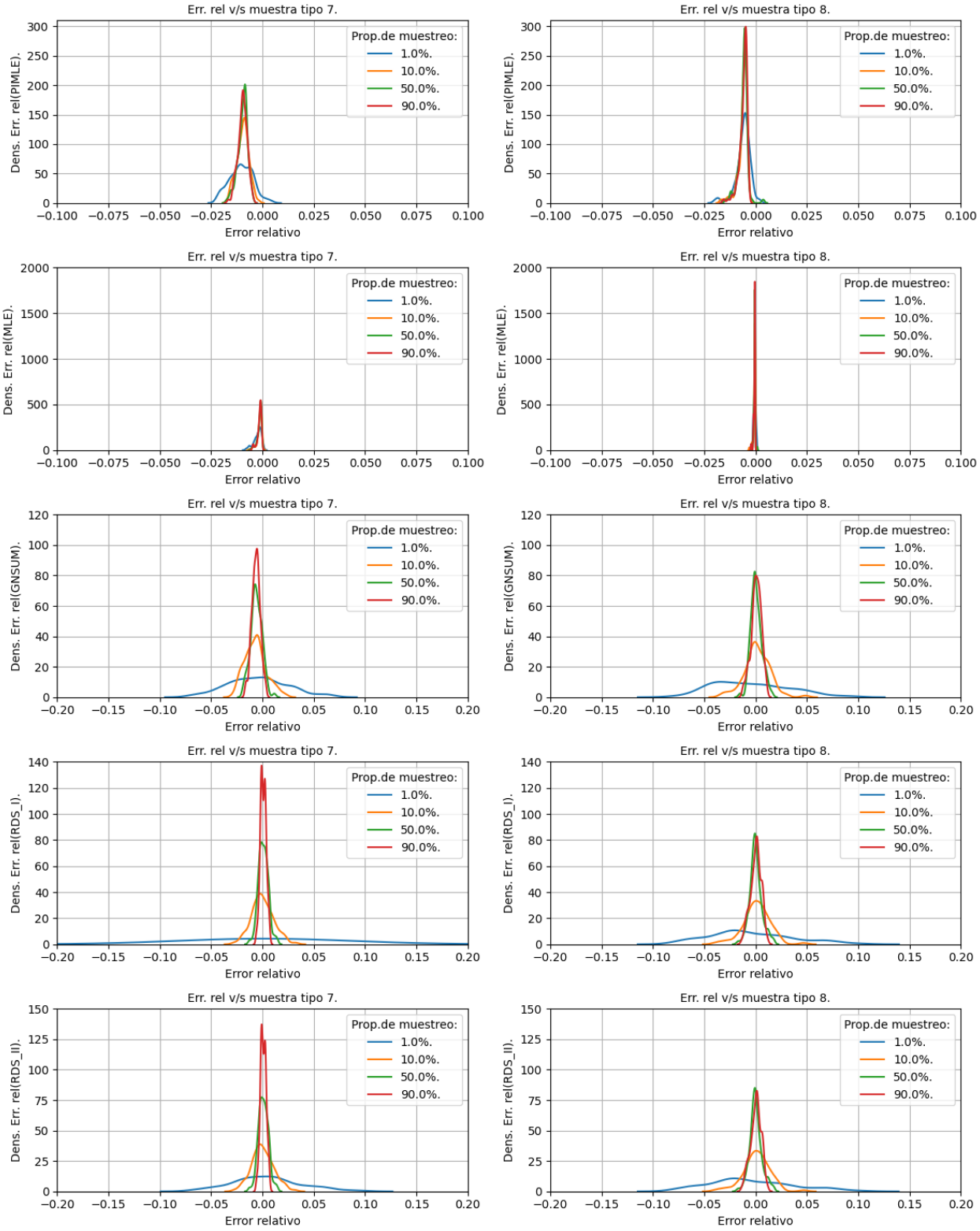


Figura C0.21: Muestra las dos últimas columnas de la figura C0.19.

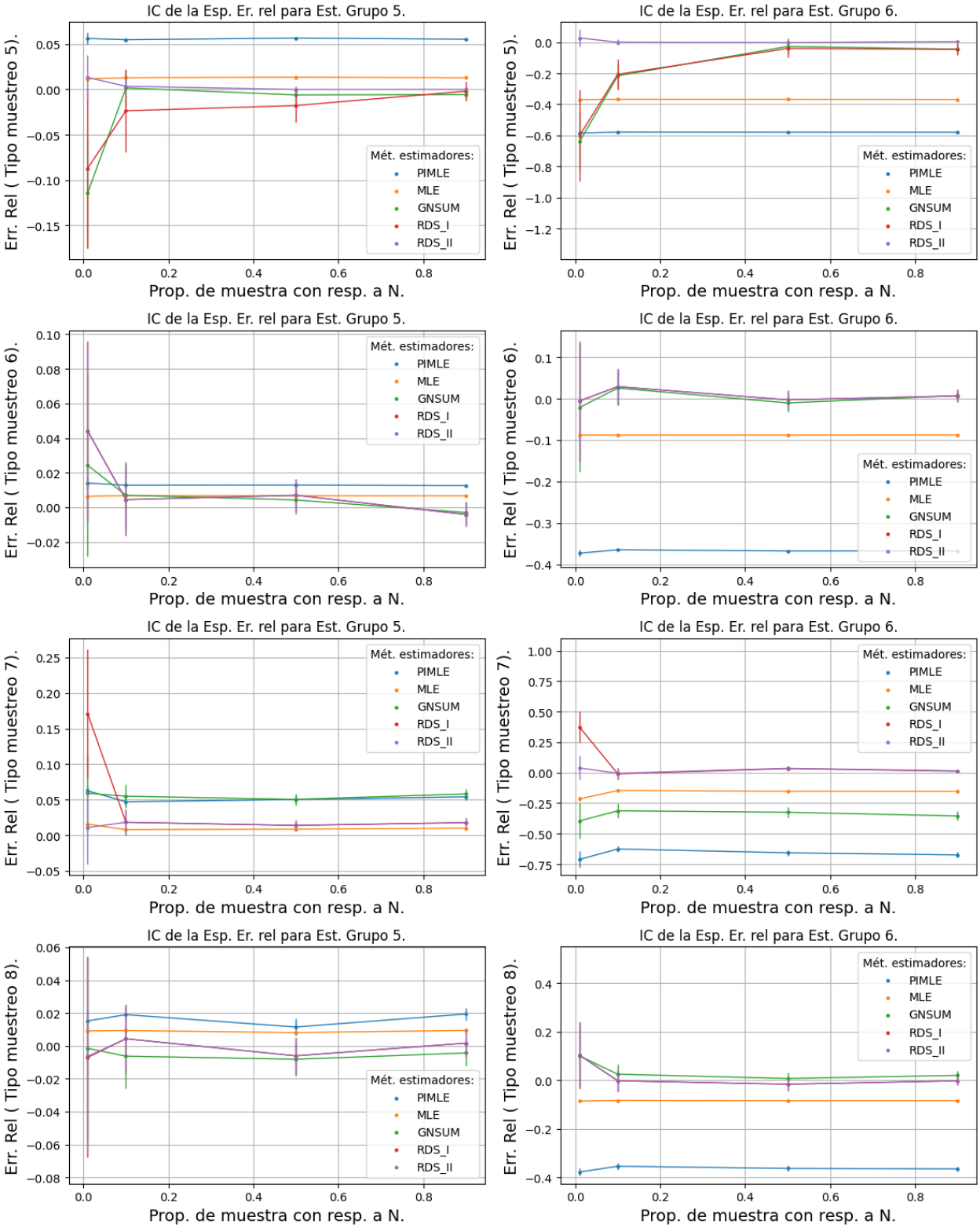


Figura C0.22: Resultados estimaciones de los grupos 5 y 6: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación PIMLE, MLE, GNSUM, RDS I y RDS II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia columna varia el grupo real a estimar.

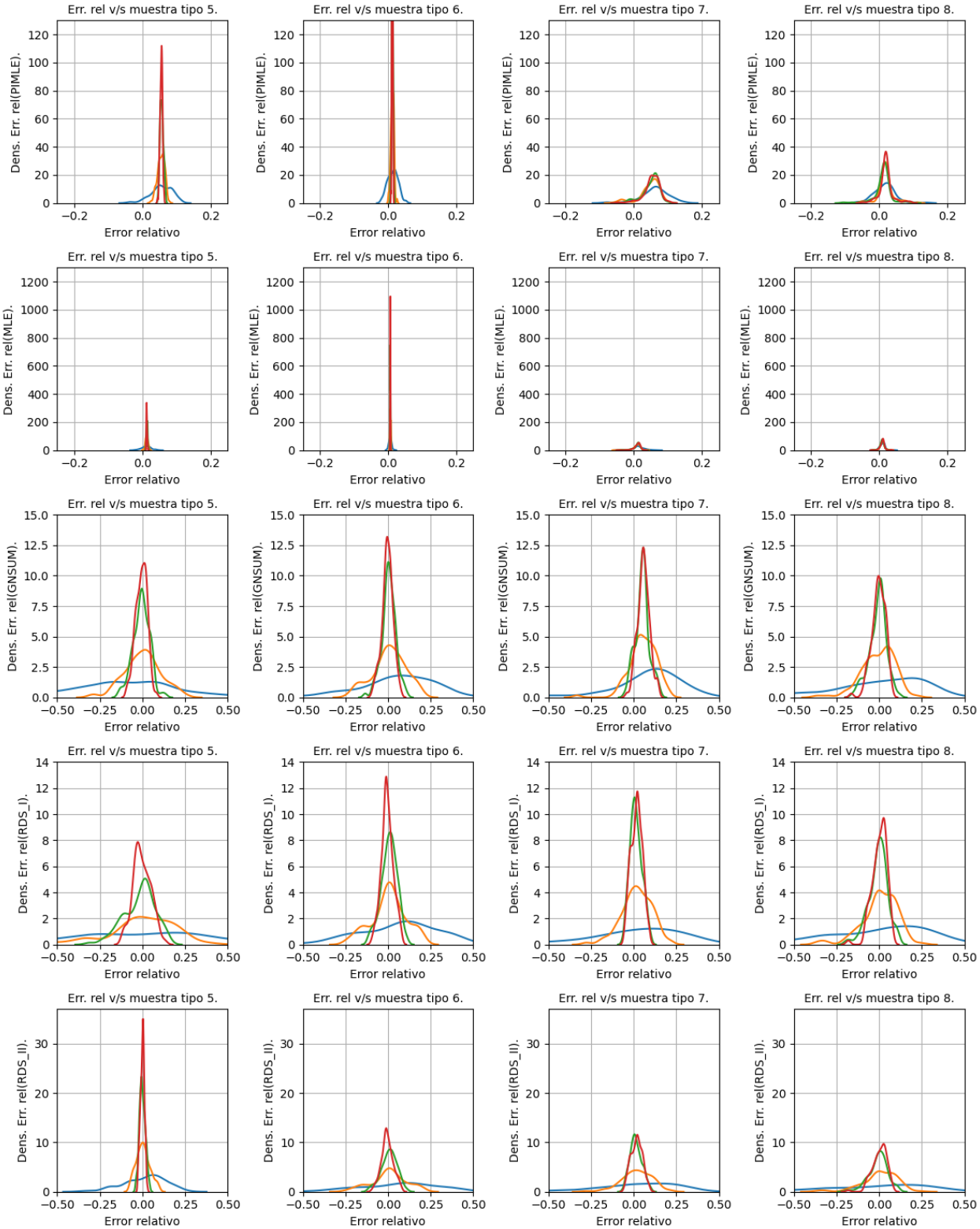


Figura C0.23: Resultados estimación grupo 5: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

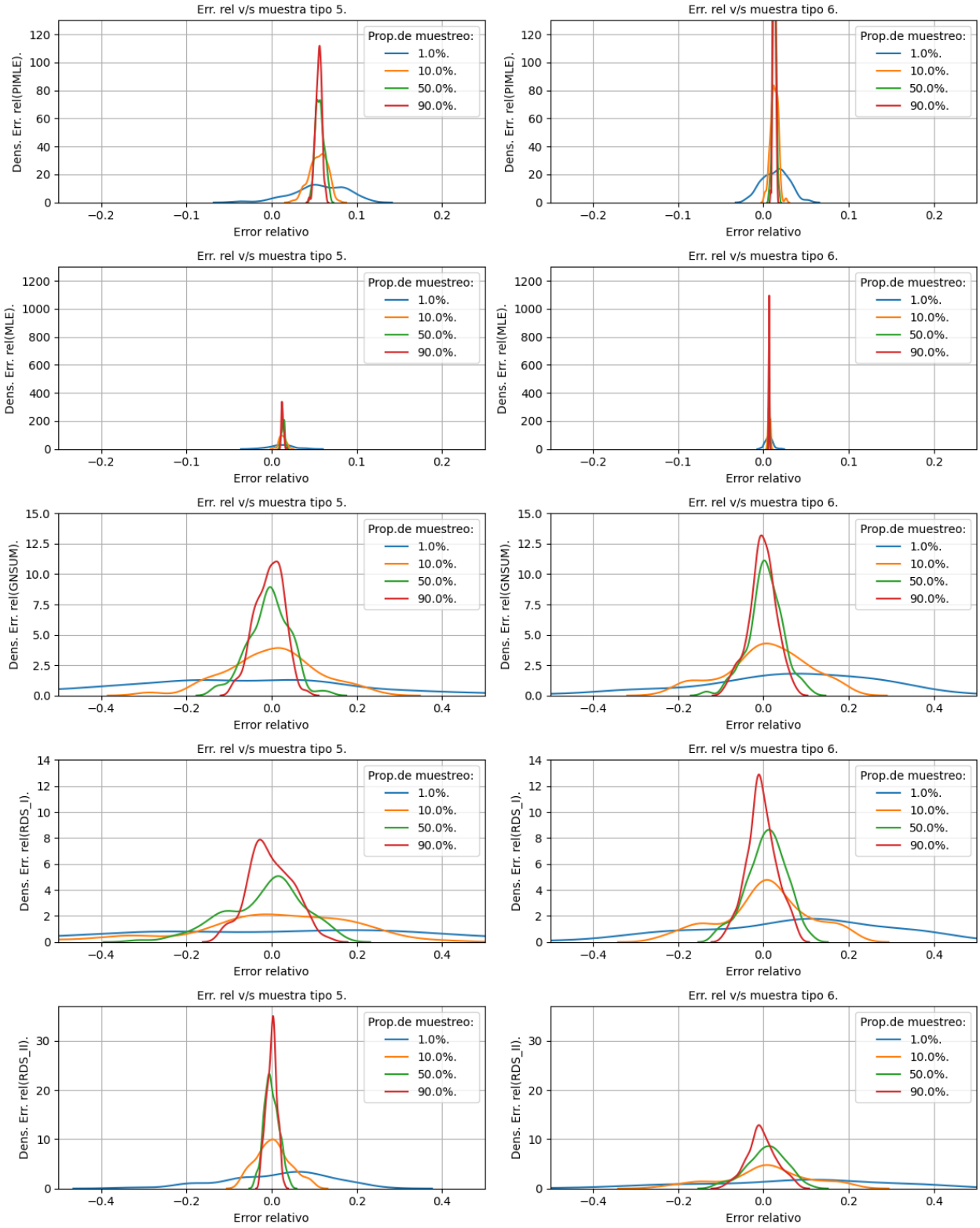


Figura C0.24: Muestra las dos primeras columnas de la figura C0.23.

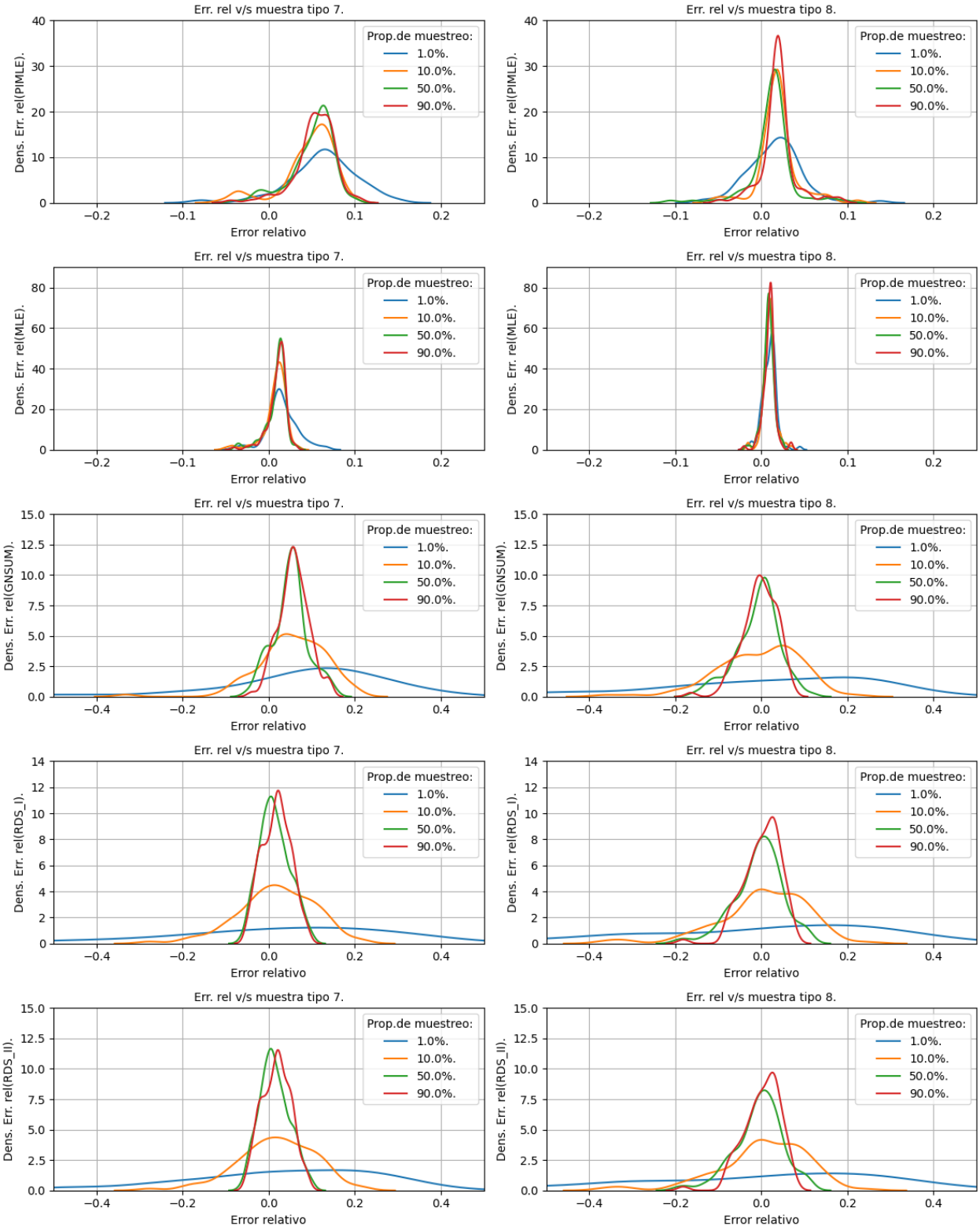


Figura C0.25: Muestra las dos últimas columnas de la figura C0.23.

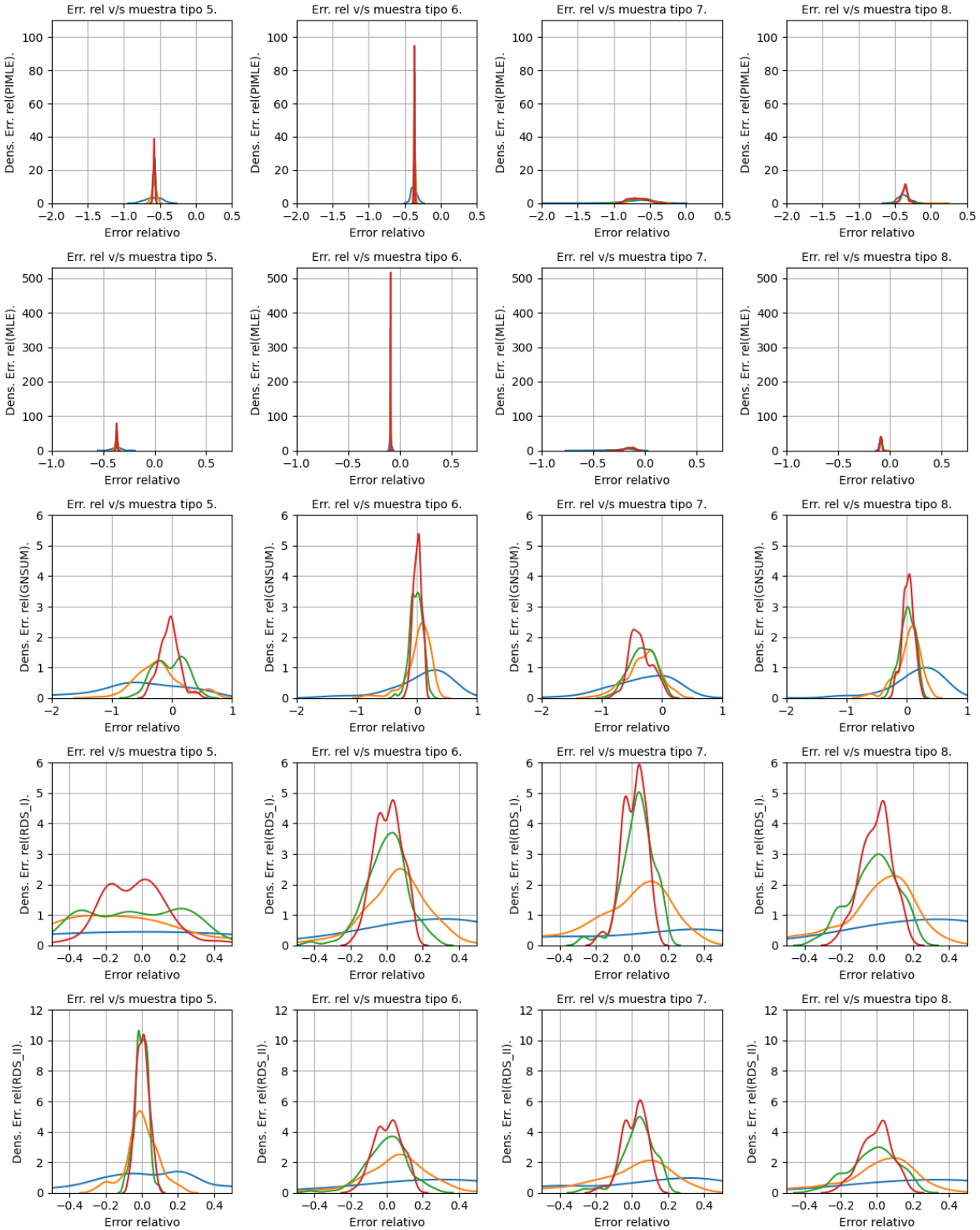


Figura C0.26: Resultados estimación grupo 6: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

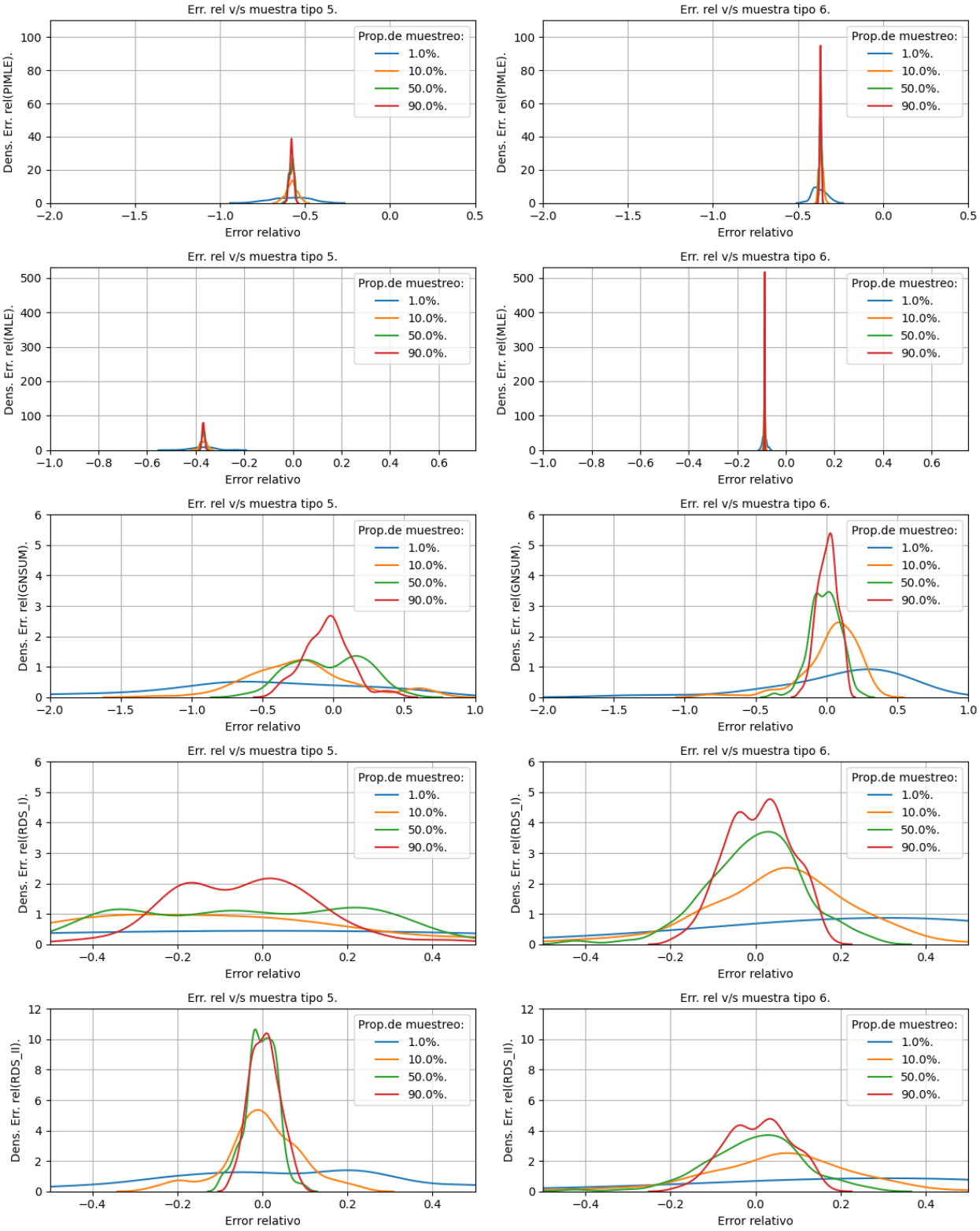


Figura C0.27: Muestra las dos primeras columnas de la figura C0.26.

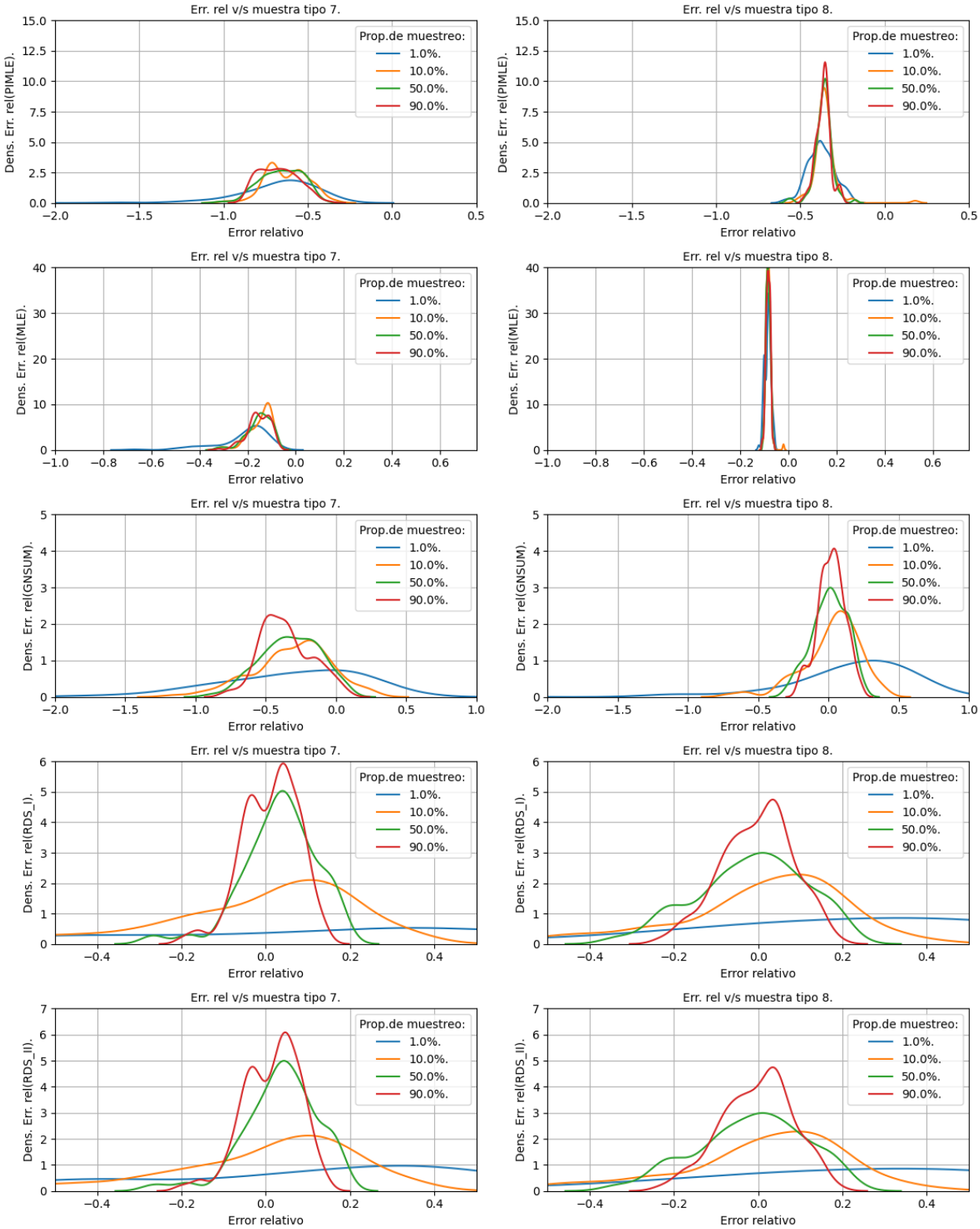


Figura C0.28: Muestra las dos últimas columnas de la figura C0.26.

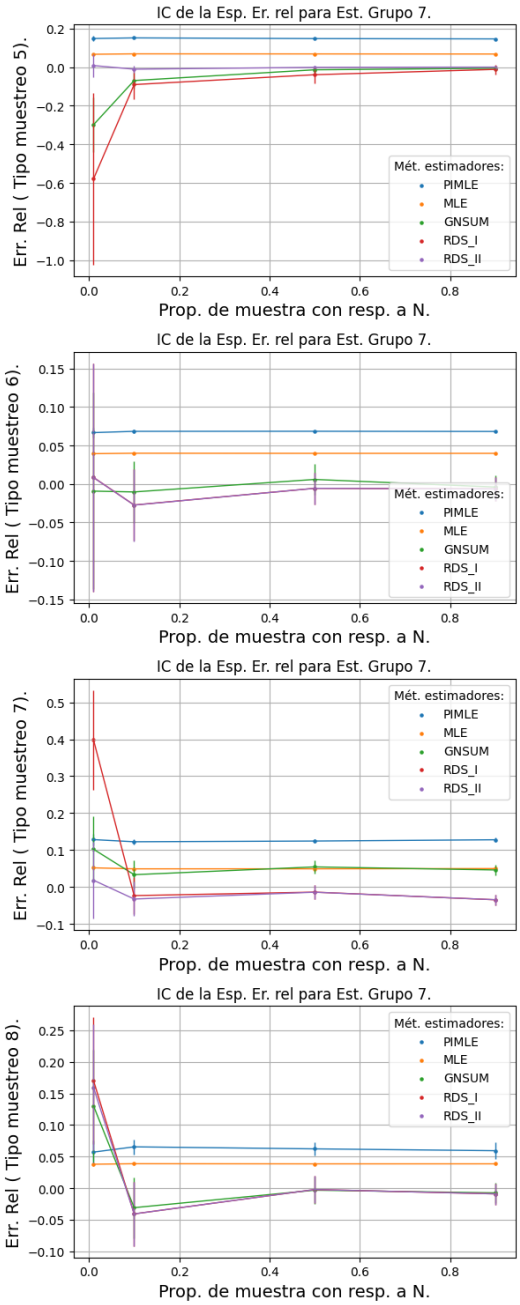


Figura C0.29: Resultados estimación del grupo 7: Gráficas de la media muestral del error relativo e intervalo de confianza de la esperanza de error relativo para cada estimador, en donde el eje x representa la proporción de la muestra con respecto a N para un tipo de muestreo fijo. El color azul, naranjo, verde, rojo y morado indica el método de estimación $PIMLE$, MLE, $GNSUM$, RDS_I y RDS_II respectivamente. Cada fila varia el tipo de muestreo y en cada columna varia el grupo real a estimar.

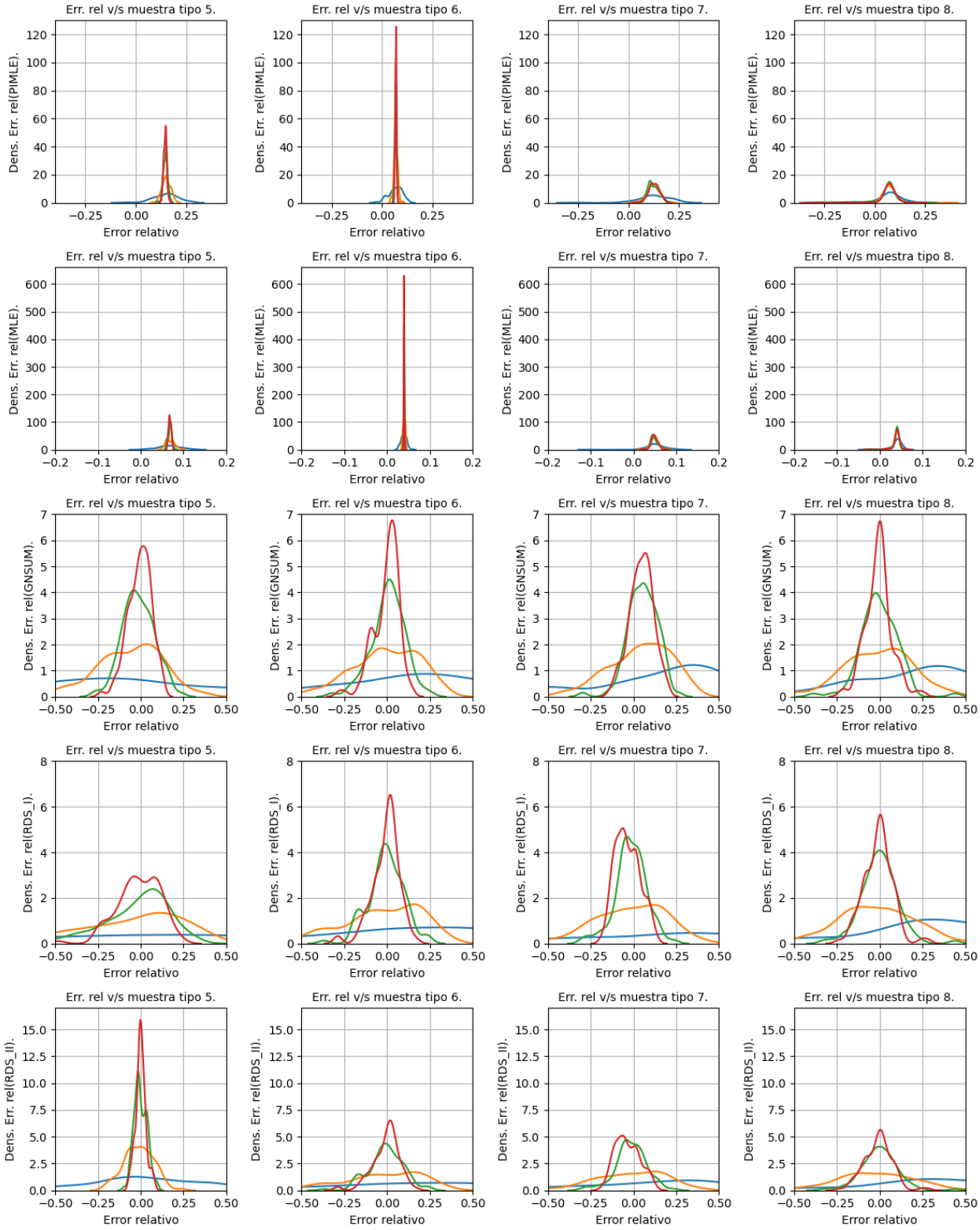


Figura C0.30: Resultados estimación grupo 7: Gráficas de la distribución del error relativo para cada proporción de nodos muestreados fijando el tipo de muestreo y el método estimador. El color azul, naranjo, verde y rojo es para la proporción de nodos muestreados del $1\,\%, 10\,\%, 50\,\%$ y del $90\,\%$ respectivamente. Cada fila varia el método estimador y cada columna varia el tipo de muestreo.

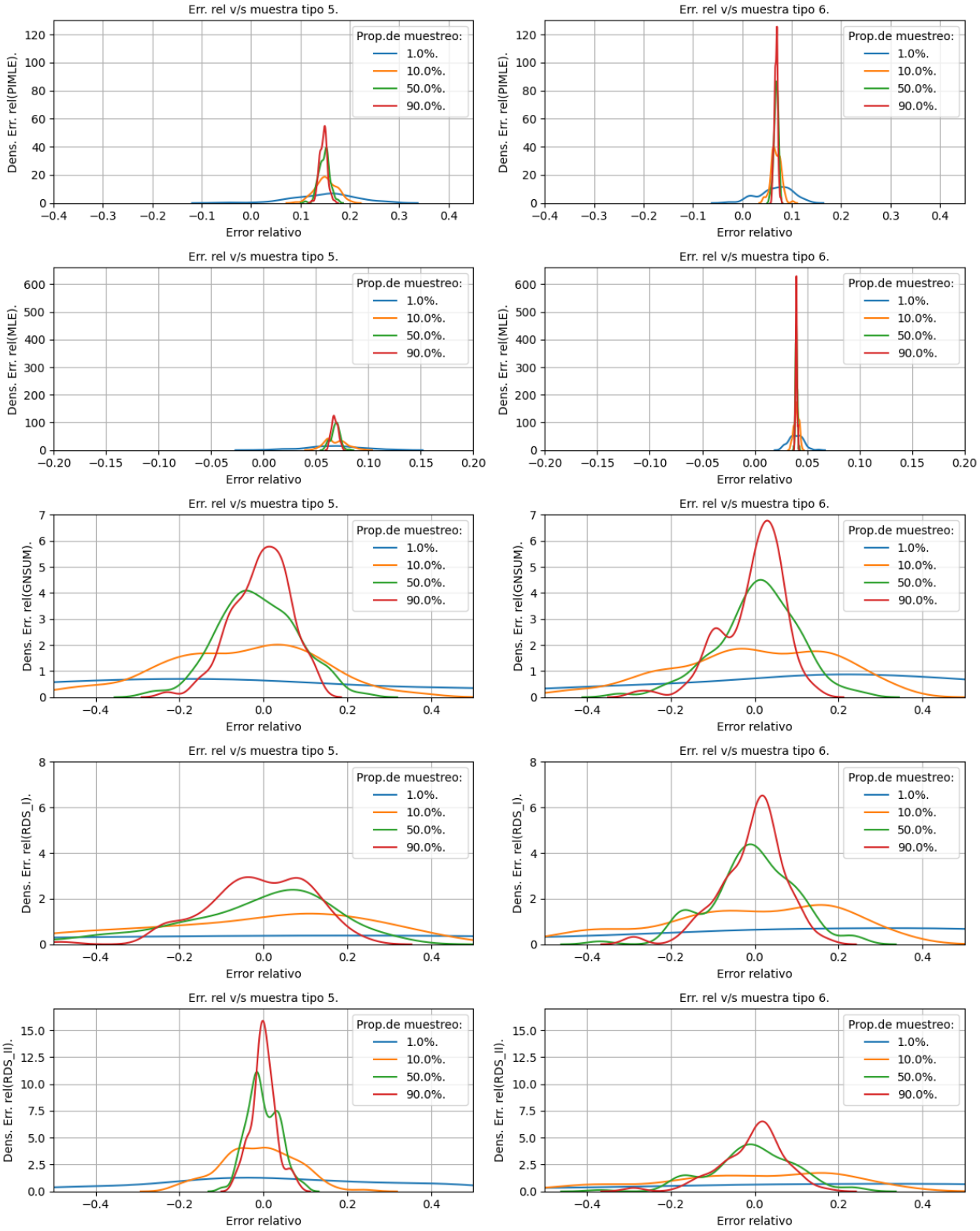


Figura C0.31: Muestra las dos primeras columnas de la figura C0.30.

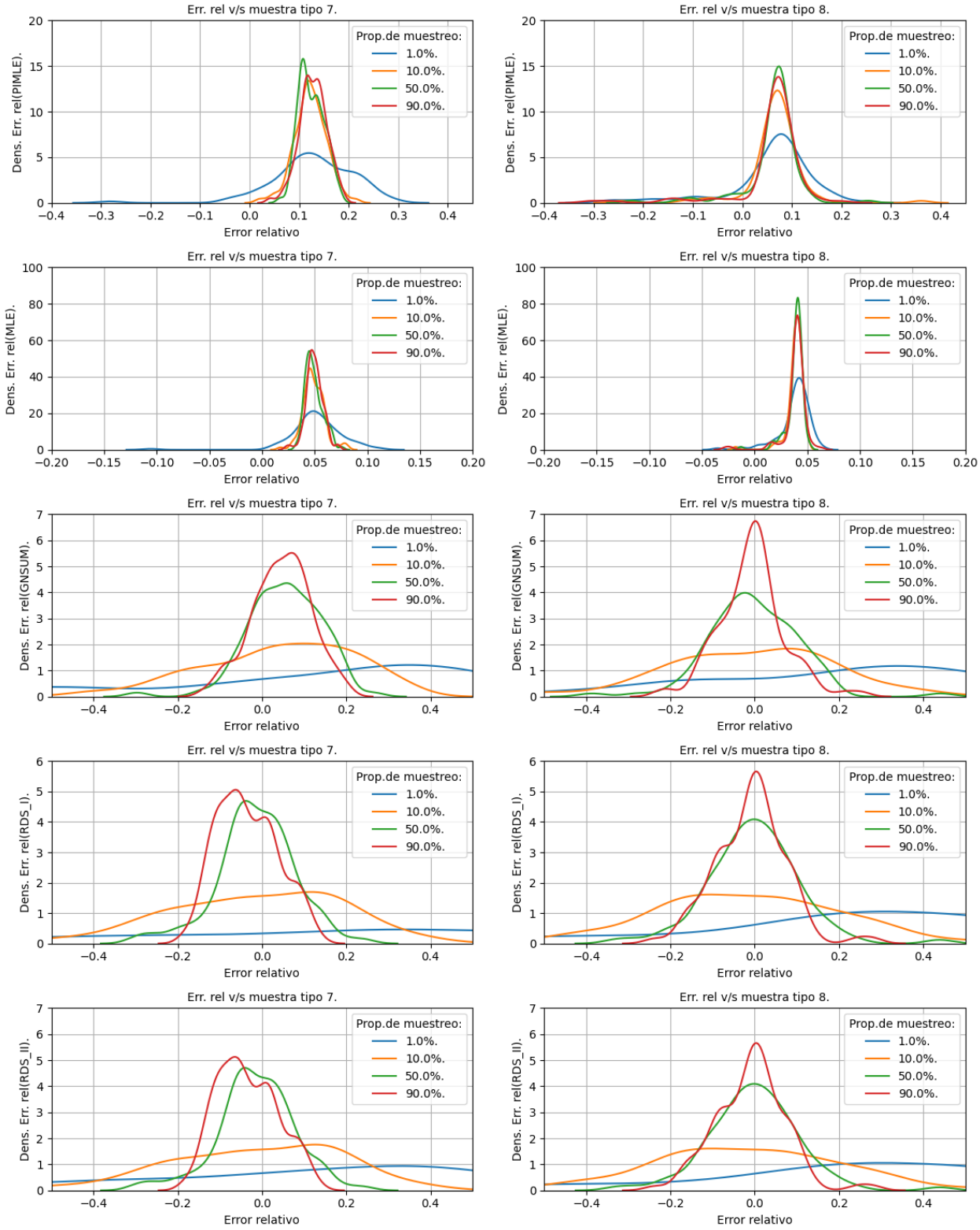


Figura C0.32: Muestra las dos últimas columnas de la figura C0.30.