



UNIVERSIDAD DE BUENOS AIRES  
FACULTAD DE CIENCIAS EXACTAS Y NATURALES

# Propuesta de mejora en la medición del bienestar económico: Consideración de la utilidad que cada individuo obtiene de sus ingresos

Tesis de Licenciatura en Ciencias de Datos

Lautaro Lasorsa

Director: Rodrigo Castro

Codirector: Walter Sosa Escudero

Buenos Aires, Argentina, 2024

## Índice general

1.. Introducción . . . . .	1
2.. Medición del Bienestar Económico . . . . .	2
2.1. Definición y motivación . . . . .	2
2.1.1. Definición . . . . .	2
2.1.2. Motivación . . . . .	2
2.2. Medición actual . . . . .	3
2.2.1. Metodología . . . . .	3
2.2.2. Ventajas . . . . .	3
2.2.3. Desventajas . . . . .	5
2.3. Medición alternativa actual: Corrección por desigualdad . . . . .	5
2.4. Medición alternativa actual: Mediana de los ingresos . . . . .	6
2.5. Propuesta de mejora: Medición ideal . . . . .	6
2.5.1. Metodología . . . . .	6
2.5.2. Ventajas . . . . .	6
2.5.3. Desventajas . . . . .	6
2.6. Propuesta de mejora: Mediciones aproximadas . . . . .	7
2.6.1. Metodología . . . . .	7
2.6.2. Ventajas . . . . .	7
2.6.3. Desventajas . . . . .	7
3.. Técnicas a utilizar . . . . .	8
3.1. Correlación de Pearson . . . . .	8
3.2. Correlaciones no lineales . . . . .	8
3.2.1. Correlación de Spearman . . . . .	8
3.2.2. $\tau$ de Kendall . . . . .	9
3.3. Bootstrap . . . . .	9
4.. Datos sintéticos . . . . .	10
4.1. Generación . . . . .	10
4.2. Datos $LN(0, 1)$ . . . . .	11
4.3. Datos $LN(0, \sigma^2)$ . . . . .	12
4.3.1. Comportamiento condicional . . . . .	12
4.3.2. Interacción con granularidad . . . . .	15
4.3.3. Comportamiento no condicional . . . . .	16
5.. Datos reales: Distribución de los ingresos . . . . .	17
5.1. Datos disponibles . . . . .	17
5.2. Distribución empírica . . . . .	18

## 1. INTRODUCCIÓN

Actualmente se utiliza para estimar el bienestar económico de una población el logaritmo natural del GNI PPA per cápita [1], pero este método tiene el problema de que aplica la utilidad marginal decreciente de los ingresos al promedio de los ingresos y no a los ingresos de cada individuo, y además es sensible a los ingresos más altos, esperables por la distribución lognormal de los ingresos en una sociedad [2].

Alternativas como corregir este indicador por la desigualdad tienen el problema de que la penalizan de forma deliverada y exógena y no como un resultado endógeno de las ineficiencias económicas que genera. Otras alternativas, como la mediana o la proporción de la población que está debajo de un cierto umbral, tienen el problema de sintetizar excesivamente la información y ser insensibles a grandes partes de la distribución.

En el presente trabajo, estudiamos si el reemplazar el  $\ln(\text{GNI PPA per capita})$  por el promedio de los logaritmos de los ingresos de los individuos resulta en un mejor indicador, dado que contempla la utilidad de los ingresos a nivel de cada individuo, y el cómo nos podemos aproximar a este indicador ideal con los datos disponibles.

En el capítulo 2 se comparará la metodología actual para medir el bienestar económico de una sociedad y se propondrá una medición alternativa. En el capítulo 4 se buscará aplicar ambos métodos a un conjunto de datos sintéticos para comparar su comportamiento. En el capítulo 5 se comenta la disponibilidad de datos reales sobre distribución del ingreso, y en el capítulo 6 se busca comparar ambos indicadores (propuesto y actual) mediante su capacidad de predecir otras variables de interés, como la esperanza de vida al nacer. El capítulo 3 explica las distintas técnicas que se utilizarán en este trabajo, para quienes no las conozcan o quieran refrescarlas antes de seguir la lectura, y los últimos 2 capítulos son sobre las conclusiones y posibles trabajos futuros.

## 2. MEDICIÓN DEL BIENESTAR ECÓNOMICO

En este capítulo se presenta la problemática a tratar en la tesis, se motiva su estudio y se comparán las formas actuales de medirlo con las alternativas propuestas.

### 2.1. Definición y motivación

#### 2.1.1. Definición

En este trabajo definiremos el bienestar económico (BE) de un individuo como la utilidad que obtiene de sus ingresos. Es decir, el valor que puede obtener de los recursos económicos, principalmente monetarios, de los que dispone.

Sea  $W$  el nivel de ingresos de un individuo, modelamos la utilidad que obtiene de esos ingresos como  $U = \log(W)$ . Entonces,  $BE = U(W) = \log(W)$

Es importante tener presente que si bien al definirlo como  $BE = \log(W)$  estamos dando una definición constructiva que nos permite calcularlo a partir de los datos recabados, esta es una aproximación que realizamos al concepto en base al comportamiento que esperamos que tenga:

- **Es creciente:** Esperamos que un mayor ingreso cause un mayor bienestar económico en el individuo que lo recibe. Formalmente,  $W_1 \geq W_2 \rightarrow U(W_1) \geq U(W_2)$
- **Ley de utilidades marginales decrecientes:** Esperamos que, a partir de un valor  $W_1$ , la utilidad marginal del ingreso sea decreciente. Formalmente:

$$\exists W_1 / \forall W \geq W_1, \epsilon > 0, U(W + \epsilon) - U(W) > U(W + 2 * \epsilon) - U(W + \epsilon)$$

Este es un comportamiento análogo a la especificación y la implementación de una función en programación, donde las expectativas que tenemos sobre el indicador (su definición conceptual) cumplen el rol de la especificación y la definición objetivamente medible cumple el rol de la implementación.

#### 2.1.2. Motivación

La correcta medición del bienestar económico tiene diversas motivaciones, como por ejemplo:

- Es un objeto de los objetos de estudio elementales de las ciencias económicas.
- Permite medir el impacto de las políticas públicas.
- Es un insumo para otras disciplinas y puede utilizarse como predictor de otras variables de interés.

De estas motivaciones puede deducirse que tiene tanto un valor intrínseco (por sí mismo) como un valor instrumental, y que en conjunto vuelven a la correcta medición de este concepto un asunto de interés.

Notar que al ser un concepto al cual aproximamos con una definición constructiva, hay 2 aspectos independientes:

- La calidad de la implementación que realicemos, es decir, qué tan bien captura la definición objetivamente medible las expectativas que tenemos sobre el concepto.
- La calidad de la medición de esta implementación.

De estos dos puntos, el presente trabajo propone posibles mejoras sobre el segundo basandose en la definición empirica ya explicada. Sin embargo, al estudiar la correlación entre el bienestar económico y otras variables, ambos puntos tendrán impacto aunque solo el segundo se estudie explicitamente.

## 2.2. Medición actual

### 2.2.1. Metodología

La medición actual del *BE*, utilizada por ejemplo por el **Índice de Desarrollo Humano** [1] (HDI por sus siglas en inglés, Human Development Index), es mediante el logaritmo del promedio de los ingresos.

En el caso puntual del HDI utiliza como insumo el GNI (Gross Domestic Income, Ingreso Nacional Bruto) a Paridad de Poder Adquisitivo (PPA, PPP en inglés). Es decir:

$$HDI_{income} = \min(1, \frac{\ln(GNI \text{ per capita}) - \ln(100)}{\ln(75000) - \ln(100)})$$

Generalizandolo, dada una población de  $N$  individuos, cuyos ingresos son  $X_1, X_2, \dots, X_N$ , la medición actual del bienestar económico es:

$$BE(X) = \log\left(\frac{1}{N} * \sum_{i=1}^N X_i\right)$$

### 2.2.2. Ventajas

La principal ventaja de este metodo es a nivel logistico, dado que calcula el *BE* en base a una colección de datos que pueden ser calculados o estimados de forma independiente:

- **GNI PPP:** El ingreso total de los individuos del país, deflactado a la paridad de poder adquisitivo. Inclusive puede ser posible calcular de forma separada el GNI nominal y el deflactor de PPP, pero también hay metodologías que apuntan directamente al valor GNI PPP utilizando cantidades en vez del valor monetario de los bienes y servicios.
- **Población:** La población de esa economía durante ese año. Usualmente se utiliza la cantidad de población a mitad del año.

Al ser estas metricas (GNI, deflactor, población) ya calculadas por tener interes en sí mismas, no supone una dificultad adicional obtener esta metrica derivada. Adicionalmente, es una metodología confiable ya que su confianza se deriva de la confianza que tenemos en la capacidad de medir las metricas base.

Además, las metricas base que se utilizan para calcular esta metrica tienen la ventaja adicional de ser valores aditivos. Es decir, cada uno de estos puede ser medido de forma independiente en subdivisiones de la economía a estudiar (por ejemplo las provincias de un país, y dentro de estas los municipios) y luego simplemente sumar las cantidades medidas por cada subdivisión para obtener la cantidad correspondiente al total de la economía.

A su vez, como el GNI se pueden descomponer en distintos componentes, estos componentes pueden calcularse independientemente.

Se puede hacer esto utilizando que  $GNI = GDP + EX_{net}$ , donde:

- GDP (Gross Domestic Product, Producto Interior Bruto) es la producción total de bienes y servicios en el país
- $EX_{net}$  es el neto de pagos y transferencia (salvo importaciones y exportaciones, que se incluyen en el GDP) hacia el exterior (donde el signo positivo indica que a la economía ingresó más dinero del que salió)

Y a su vez podemos descomponer al principal sumando, el GDP, en sus componentes de dos maneras distintas, utilizando la Perspectiva de los Gastos y la Perspectiva de los ingresos.

### Perspectiva de Gastos

$$GDP = C + I + G + (X - M)$$

Donde:

- C = consumo
- I = Inversión
- G = Gasto del gobierno
- X = Exportaciones
- M = Importaciones

### Perspectiva de los ingresos

$$GDP = S + B + R + I + II + D - SU$$

Donde:

- S = Salarios
- B = Beneficios empresariales
- R = Rentas (alquileres)
- I = Intereses
- II = Impuestos Indirectos (ejemplo: IVA)
- D = Depreciación de bienes de capital
- SU = Subsidios

El tener estos dos enfoques permite:

- Descomponer el cálculo del GNI en muchas variables chicas que se pueden medir de forma especializada y tienen interés en sí mismo. Por tanto, se vuelve en sí mismo una estadística derivada de otras.

- Al tener 2 métodos independientes de calcular el GDP, que es el factor más importante en el computo del GNI, es posible detectar y corregir errores e inconsistencias en las mediciones.

En síntesis, la medición actualmente utilizada supone una gran cantidad de ventajas en materia logística y de ser consecuencia de otra batería de mediciones con interés en sí mismas.

### 2.2.3. Desventajas

Hay un primer problema que podemos observar en esta metodología y es que el GNI incluye también los ingresos empresariales y del gobierno, mientras que el *BE* lo definimos a nivel de individuos y sus ingresos personales. Sin embargo, como las empresas y gobiernos pueden utilizar estos recursos para proveer bienes y servicios a las personas (que a estas no se les imputan dentro de sus ingresos personales, por ejemplo la educación pública no arancelada), este alejamiento de la definición que dimos originalmente puede permitir capturar mejor el bienestar económico de una sociedad.

Sin embargo, hay otro problema que en principio es más importante a tener en cuenta. La utilidad marginal decreciente aplica en los ingresos de cada individuo y no en los ingresos agregados de la sociedad. Por tanto, es importante aplicar la función de utilidad ( $U$ , en este caso  $\log$ ) a los ingresos de cada individuo y no al promedio de los ingresos. Y lo importante es que **el logaritmo del promedio no es el promedio de los logaritmos**

Es esta última desventaja la que este trabajo busca subsanar proponiendo una medición alternativa, y a su vez evaluar qué mejoras ofrece dicha alternativa frente a la metodología actual.

### 2.3. Medición alternativa actual: Corrección por desigualdad

Reconociendo que una de las falencias del método para calcular el HDI es la insensibilidad frente a la desigualdad, el informe de la UNDP (United Nations Development Programme) también incluye el *IHDI* (Índice de Desarrollo Humano ajustado por Desigualdad). En este se define una métrica de penalización para cada indicador:

$$A_X = 1 - \frac{\sqrt[N]{\prod_{i=1}^N X_i}}{\frac{1}{N} * \sum_{i=1}^N X_i}$$

Así, si el PBI PPA per cápita de un país en el año 2024 fue  $X$ , su puntaje de ingresos para el *HDI* sería  $HDI_{income} = \min(1, \frac{(\log(X) - \log(100))}{(\log(75000) - \log(100))})$ , y su puntaje de ingresos para el *IHDI* sería  $IHDI_{income} = (1 - A_{income}) * HDI_{income}$

El principal problema achacable a esta métrica es que la penalización de la desigualdad surge de la propia definición (exógena) de la misma y no como una consecuencia natural (endógena) de las inefficiencias que esta causa en la economía y en el bienestar de los individuos.

Generalización, esta genera una métrica tal que :

$$BE_I(X) = \frac{\sqrt[N]{\prod_{i=1}^N X_i}}{\frac{1}{N} * \sum_{i=1}^N X_i} * \log\left(\frac{1}{N} * \sum_{i=1}^N X_i\right)$$

## 2.4. Medición alternativa actual: Mediana de los ingresos

Una alternativa para aportar robustez es utilizar la mediana de los ingresos en vez de la media. El problema de este índicador es que descarta demasiada información de la muestra. Dada una distribución del ingreso, cualquier modificación en los ingresos de aquellos que están por sobre la mediana de los ingresos será ignorada por el indicador, así como cualquier modificación de aquellos que tienen menos ingresos que no sea lo suficientemente significativa para llevarlos a un nivel de ingresos superior a la mediana actual.

Generalizando, esto resulta en:

$$BE_M(X) = F_X^{-1}\left(\frac{1}{2}\right)$$

Donde  $F_X$  es la función de densidad acumulada de  $X$  (es decir, dado un valor  $a$ ,  $F_X(a)$  nos da la proporción de valores de  $X$  menores a  $a$ ), y  $F_X^{-1}$  es su inversa (de no ser invertible la función, da el mínimo de la preimagen)

## 2.5. Propuesta de mejora: Medición ideal

### 2.5.1. Metodología

Dada una población  $X$  de  $N$  individuos cuyos ingresos son  $X_1, X_2, \dots, X_N$ , definimos el Bienestar Económico (promedio) de esa población como:

$$BE(X) = \frac{1}{N} * \sum_{i=1}^N \log(X_i) = \log\left(\sqrt[N]{\prod_{i=1}^N X_i}\right)$$

### 2.5.2. Ventajas

Es la medición exacta de la definición que propusimos de Bienestar Económico. Por tanto, es la mejor aproximación a dicho concepto. Lo que tiene especial sentido si pensamos que los ingresos son un instrumento para obtener bienestar y no el bienestar en sí mismo, y eso es lo que se busca reflejar al aplicarles una función de utilidad distinta de la identidad.

Adicionalmente, al contemplar que un aumento en los ingresos de individuos que ya tienen altos ingresos les genera un menor beneficio marginal que aumentar en la misma cantidad absoluta los ingresos de individuos de menores ingresos, esta métrica refleja la eficiencia de la distribución del ingreso en una sociedad.

De esta forma, esta métrica premia simultáneamente un aumento de la productividad de una economía como una distribución más eficiente de la misma.

### 2.5.3. Desventajas

La principal desventaja de esta metodología es la contracara de la ventaja de la metodología actual, la logística. En este caso llega al punto de la infactibilidad, puesto que para poder realizar el cálculo propuesto es necesario conocer los ingresos de cada uno de los individuos, algo que es logisticamente imposible.

## 2.6. Propuesta de mejora: Mediciones aproximadas

### 2.6.1. Metodología

Debido a la imposibilidad factica de aplicar la metodología ideal propuesta en este trabajo, es necesario buscar aproximaciones con los datos disponibles.

De esta forma podemos introducir el concepto de **Granularidad** de una medición,  $G$ .

Sea una población  $X$  de  $N$  individuos, cuyos ingresos son  $X_1, X_2, \dots, X_N$ , con  $X_i \leq X_{i+1} \forall 1 \leq i < N$ , y sea la granularidad  $G$ , tal que  $N = G * T$ , definimos  $BE_G(X)$  como:

$$BE_G(X) = \frac{1}{G} * \sum_{i=0}^{G-1} \log\left(\frac{1}{T} \sum_{j=i*T}^{(i+1)*T-1} X_j\right)$$

De esta definición se pueden reescribir la medición actual, utilizando el promedio de los ingresos de los individuos, como  $BE_1$  y la medición ideal antes propuesta como  $BE_N$ . Si en lugar del promedio de los ingresos de los individuos, utilizamos el GNI (que también incluye ingresos de las empresas y del gobierno), lo llamaremos  $BE_{GNI}$

### 2.6.2. Ventajas

Esta metodología tiene, parcialmente, las ventajas tanto de la metodología ideal propuesta como de la metodología actual:

- $\exists G > 1$  para el cuál la metodología es logisitcamente factible. Puede calcularse en base a encuestas de ingresos como la EPH (Encuesta Permanente de Hogares).
- Modera más que  $BE_1$  el impacto de los outliers (ultra ricos) en la medición del bienestar económico, ya que solo impactan de forma líneal en el último de los  $G$  grupos en los que se divide la población.

### 2.6.3. Desventajas

Tiene la desventaja de ser una solución de compromiso, y justamente parte del interés de este trabajo es ver, para valores logisticamente factibles de  $G$ , cuál es la diferencia entre  $BE_G$  y  $BE_N$ .

### 3. TÉCNICAS A UTILIZAR

En este capítulo se explicarán las técnicas que se utilizarán a lo largo del trabajo para analizar los datos, tanto sintéticos como reales, y obtener conclusiones de los mismos.

#### 3.1. Correlación de Pearson

El coeficiente de correlación de Pearson se utiliza para medir la similitud líneal entre 2 variables. Se define como:

$$r(X, Y) = \frac{Cov(X, Y)}{\sigma_X * \sigma_Y}$$

Donde  $Cov(X, Y)$  es la covarianza entre ambas variables, y  $\sigma_X$  y  $\sigma_Y$  son los desvios estandar de las mismas. Este indicador tiene varias propiedades interesantes:

- Si planteamos el modelo  $Y = \alpha * X + \beta$ , para los valores  $\alpha*$  y  $\beta*$  que minimizan el MSE (Mean Square Error):
  - $R^2 = \frac{\sum_{i=1}^N (Y_i - X_i * \alpha * + \beta)^2}{\sum_{i=1}^N (Y_i - \mu_Y)^2}$
  - $R^2 = (r(X, Y))^2$

Es decir, captura la capacidad predictiva (en sentido estadístico) del modelo líneal  $Y = \alpha * X + \beta$

- Podemos tratar a  $L^2$  (variables aleatorias con segundo momento finito) como un espacio vectorial euclídeo donde:
  - $Cov(X, Y)$  es el producto interno entre  $X$  e  $Y$
  - $Var(X) = Cov(X, X)$  es la norma cuadrada de  $X$ . Analogamente,  $std(X) = \sigma_X = \|X\|$
  - $r(X, Y) = \frac{\langle X, Y \rangle}{\|X\| * \|Y\|}$  es el coseno del angulo entre  $X$  e  $Y$ .

Esto reafirma su interpretación como medida de similitud entre las variables.

#### 3.2. Correlaciones no lineales

Notar que la correlación de Pearson solo captura relaciones líneales entre las variables, y si queremos capturar otras posibles relaciones (por ejemplo,  $Y = \log(X)$ ) corresponde utilizar otras métricas, como son la correlación de Spearman y la  $\tau$  de Kendall.

##### 3.2.1. Correlación de Spearman

Sea  $rank_X[x]$  la función que dado un individuo nos dice su índice en la población ordenada,

$$Spearman(X, Y) = 1 - \frac{6}{N * (N^2 - 1)} * \sum_{i=1}^N (rank_X[X_i] - rank_Y[Y_i])^2$$

### 3.2.2. $\tau$ de Kendall

$$\tau(X, Y) = \frac{2}{N * (N - 1)} * (P_C - P_D)$$

Donde

- $P_C = ||\{(i, j) | i < j \wedge ((X_i < X_j) \wedge (Y_i < Y_j)) \vee (X_i > X_j) \wedge (Y_i > Y_j)\}||$  son los pares que tienen la misma relación ordinal en  $X$  y en  $Y$ .
- $P_D = ||\{(i, j) | i < j \wedge ((X_i < X_j) \wedge (Y_i > Y_j)) \vee (X_i > X_j) \wedge (Y_i < Y_j)\}||$  son los pares que tienen distinta relación ordinal en  $X$  que en  $Y$ .

Una forma alternativa de escribirla es:

$$\tau(X, Y) = \frac{2}{N * (N - 1)} * \sum_{i=1}^N \sum_{j=1}^{i-1} (\text{signo}(X_i - X_j) * \text{signo}(Y_i - Y_j))$$

### 3.3. Bootstrap

Bootstrap es una técnica que consiste en utilizar los datos de una muestra para estimar la distribución de la misma, en base a esa estimación generar datos sintéticos y estudiar sobre esos datos sintéticos la distribución del indicador de interés.

En el caso de este trabajo se utilizará bootstrap no paramétrico de la siguiente forma:

Dada una muestra  $M$  de  $N$  individuos  $M_1, M_2, \dots, M_N$ , una nueva muestra se generará tomando  $N$  elementos de  $M$  con reposición (es decir, al seleccionar un elemento de  $M$  para incluirlo en nuestra nueva muestra, no lo eliminamos de  $M$ ), y aplicamos el indicador (por ejemplo, la correlación entre 2 dimensiones de los individuos) a la población simulada.

## 4. DATOS SINTETICOS

En este capítulo se utilizan datos simulados para explorar características de los indicadores (existentes y propuestos) en sí mismos.

### 4.1. Generación

Para la generación de los datos se modela la distribución del ingreso en una sociedad como  $\logNorm(\mu, \sigma^2)$ , basandonos en (Gibrat, 1931)[2]. Es decir, sea  $X$  una población con  $N$  individuos cuyos ingresos son  $X_1, X_2, \dots, X_N$ , las variables  $X_i$  son IID (independientes e identicamente distribuidas) y  $X_i \sim LN(\mu, \sigma^2)$ .

Recordar que

$$X \sim LN(\mu, \sigma^2) \iff \log(X) \sim N(\mu, \sigma^2)$$

Notar que bajo esta distribución:

- El bienestar económico tiene distribución normal  $N(\mu, \sigma^2)$
- $BE_N(X)$  es un estimador insesgado de  $\mu$

La metodología consistió en lo siguiente:

- Generar poblaciones con  $N = 1,000,000$  individuos cada una.
- Ordenar a los individuos de la población, en orden creciente de ingresos.
- Para cada población  $X^i$  y para divisor  $G$  de  $N$ , calcular  $BE_G(X^i)$
- Almacenar para posterio uso los valores  $BE_j^i = BE_{G_j}(X^i)$

Notar que todas las observaciones y todas las poblaciones generadas son independientes entre si.

Como modificar  $\mu$  es lo mismo que multiplicar a todos los  $X_i$  por  $e^{\Delta\mu}$ , utilizaremos  $\mu = 0$  para la generación de datos sintéticos.

Respecto del valor de  $\sigma^2$ , se generaron 2 datasets:

- **Datos  $LN(0, 1)$ :** Un dataset donde  $X_i \sim LN(0, 1)$ , para el que se simularon contiene 20,000 poblaciones
- **Datos  $LN(0, \sigma^2)$ :** Un dataset donde se toman diversos valores de  $\sigma^2$ , en el cuál se simularon 1,000 poblaciones para cada valor entero de  $\sigma^2$  entre 1 y 10, generando 10,000 poblaciones en total.

La generación de datos se paralelizo utilizando GPU mediante CUDA[3].

## 4.2. Datos $LN(0, 1)$

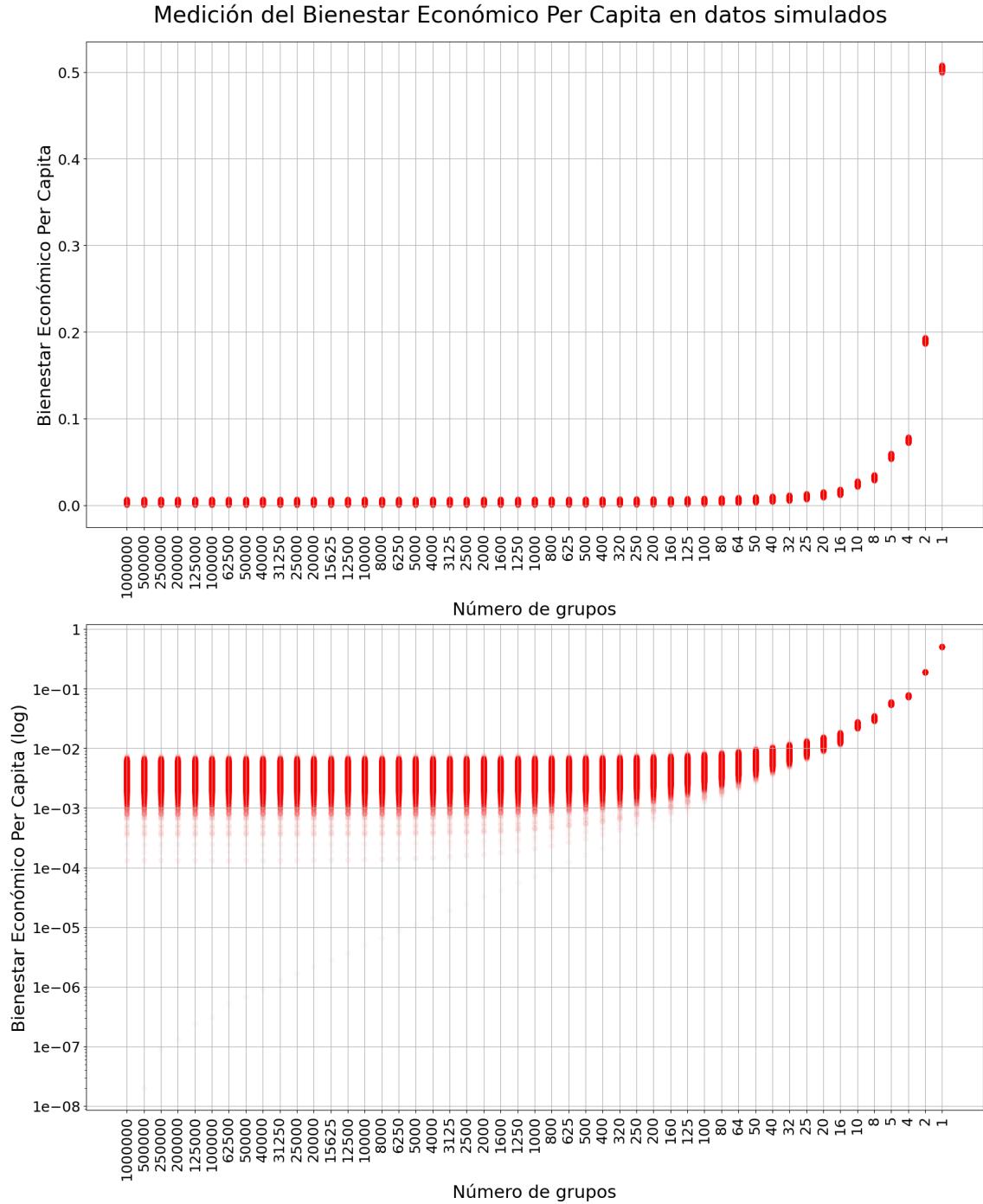


Fig. F4.1: Distribución de  $BE_j^i$ , donde el eje X es la granularidad (cantidad de grupos) de la medición. Para facilitar la visualización, a cada elemento se le resta el mínimo de todo el dataset

Como puede verse en F4.1, al tener más de 200 grupos es difícil distinguir las distintas distribuciones, inclusive en escala logarítmica. A su vez, se nota que cuando la granularidad es baja, cada aumento de granularidad acerca significativamente  $BE_{G_j}$  a  $BE_N$

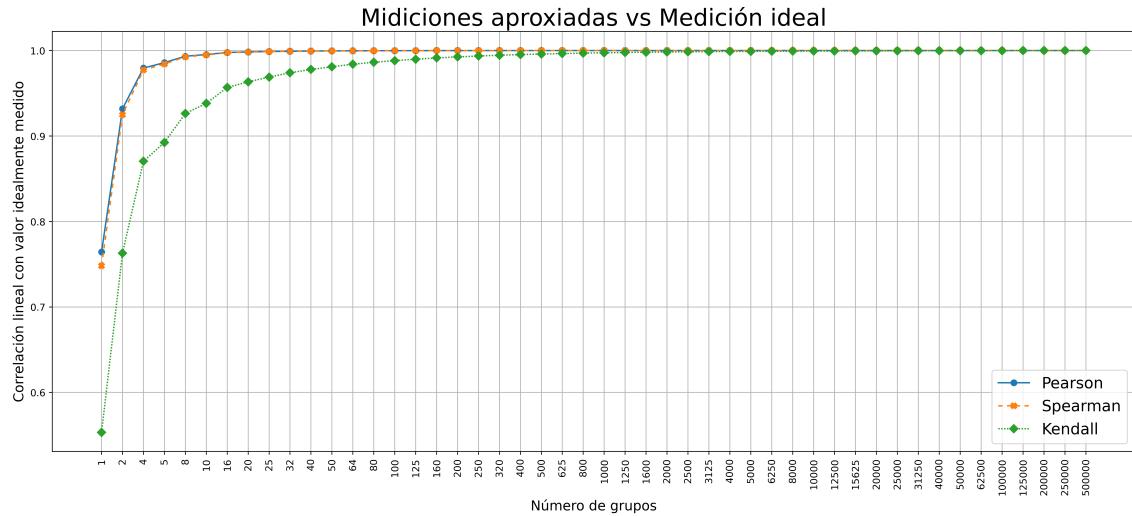
Esto nos permite formular la primera de las hipótesis del presente trabajo:

**Hipótesis H4.1** (Mejora de la granularidad). *Sea  $P$  un conjunto de poblaciones, donde*

cada  $P_i \in P$  es una población de  $N$  individuos  $P_{i,j}$  ( $1 \leq j \leq N$ ), con  $P_{i,j}$  variables IID y  $P_{i,j} \sim LN(0, 1)$ , y sean  $G_1, G_2$  tales que  $G_1|N, G_2|N, G_1 < G_2$ , entonces  $\exists N_0/\forall N > N_0$ ,  $Pearson(BE_{G_1}, BE_N) < Pearson(BE_{G_2}, BE_N)$  para casi toda población  $P$

Para la falsación empírica de esta hipótesis, es necesaria la hipótesis auxiliar de  $N_0 \leq 1,000,000$

Como puede verse en la figura F4.2, la hipótesis H4.1 resulta verosímil con los datos simulados con distribución  $LN(0, 1)$



$\exists N_0 / \forall N > N_0, \text{Pearson}(BE_{G_1}, BE_N) < \text{Pearson}(BE_{G_2}, BE_N)$  para casi toda población  $P$

Para la falsación empírica de esta hipótesis, es necesaria la hipótesis auxiliar de  $N_0 \leq 1,000,000$

Notar que la hipótesis H4.2 es una generalización no trivial de H4.1 ya que modificar el  $\sigma^2$  tiene efectos no lineales en la distribución.

Como puede verse en la figura F4.3, esta hipótesis generalizada también resulta verosímil con los datos generados.

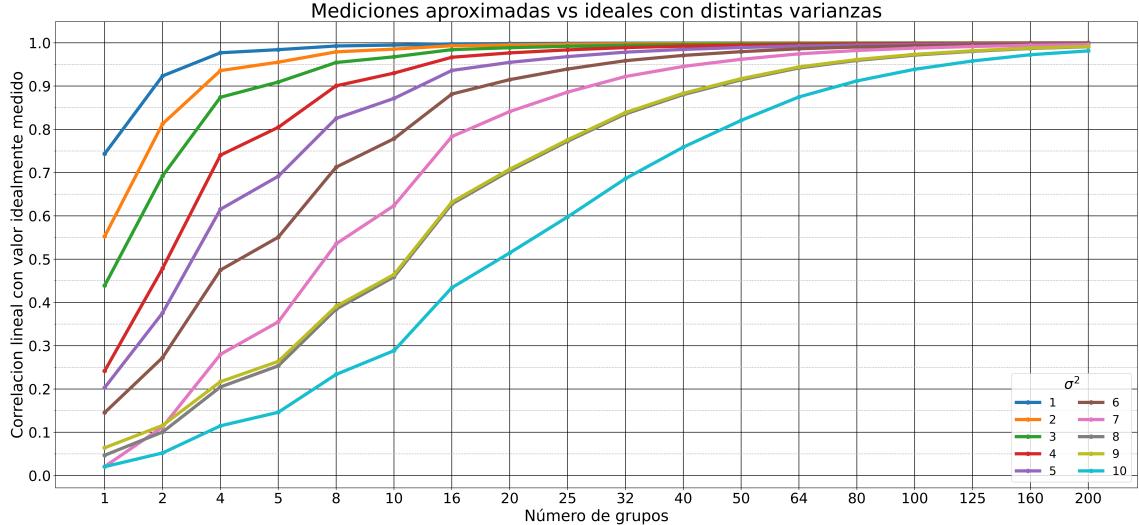


Fig. F4.3: Aplicación de métrica de correlación lineal entre  $BE_G$  y  $BE_N$  para todos los  $G|N$  en muestras  $LN(0, \sigma^2)$  con  $N = 10^6$ , condicional a distintos valores de  $\sigma^2$

Algo que podemos notar observando F4.3 es que para varianzas altas la correlación lineal entre las granularidades bajas y la medición ideal es notablemente baja.

Esto puede apreciarse mejor en F4.4, donde podemos ver que mientras que  $BE_1$  y  $BE_N$  tienen una tendencia marcada para  $\sigma^2 = 1$ , para  $\sigma^2 = 10$  se comportan como variables prácticamente independientes.

### Distribución conjunta con distintas varianzas y cantidades de grupos

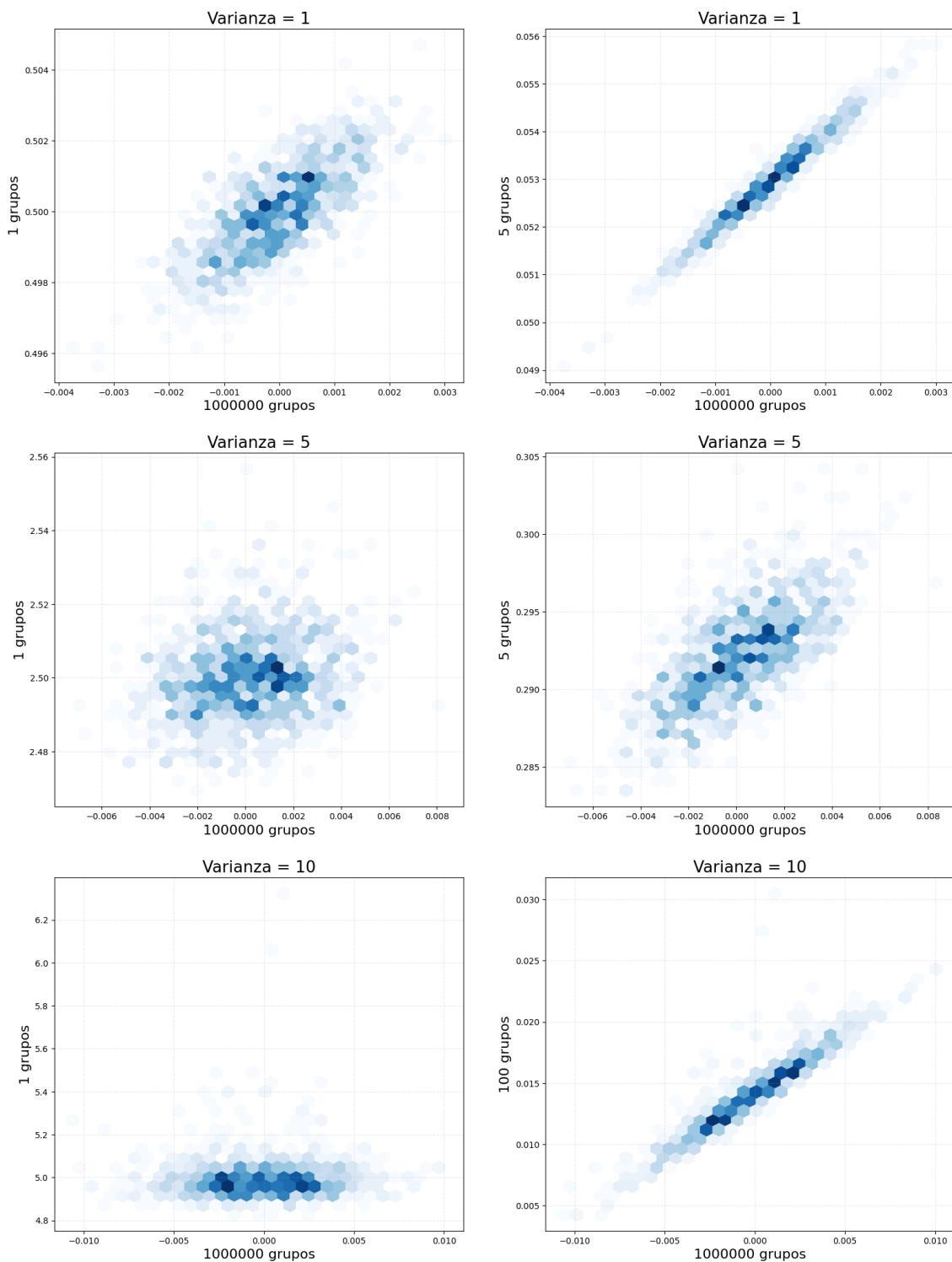


Fig. F4.4: Dispersión conjunta de las métricas entre distintas granularidades, condicional a distintos valores de  $\sigma^2$ . Puede verse como la correlación entre las distintas granularidades (ejes X e Y de cada gráfico) disminuye al aumentar la varianza (indicada en el título de cada cuadro)

### 4.3.2. Interacción con granularidad

Lo observado en F4.3 y F4.4 nos puede llevar a proponer una hipótesis ambiciosa sobre la relación entre la varianza de la población y la granularidad de la medición:

**Hipótesis H4.3** (Relación entre granularidad y varianza). *Sean:*

- $P$  un conjunto de poblaciones, donde cada  $P_i \in P$  es una población de  $N$  individuos  $P_{i,j}$  ( $1 \leq j \leq N$ ), con  $P_{i,j}$  variables IID y  $P_{i,j} \sim LN(0, \sigma_P^2)$
- $Q$  un conjunto de poblaciones, donde cada  $Q_i \in Q$  es una población de  $N$  individuos  $Q_{i,j}$  ( $1 \leq j \leq N$ ), con  $Q_{i,j}$  variables IID y  $Q_{i,j} \sim LN(0, \sigma_Q^2)$ , con  $\sigma_Q^2 > \sigma_P^2$
- $G_1, G_2 \in Divs(N)$ , con  $G_1 \neq G_2$ .

Entonces,  $\exists N_0 / \forall N > N_0$ ,  $Pearson_P(BE_{G_1}, BE_{G_2}) > Pearson_Q(BE_{G_1}, BE_{G_2})$ ,  $\forall G_1, G_2 (G_1 \neq G_2)$  y para casi toda tupla  $(P, Q)$ .

Para la falsación empírica de esta hipótesis, es necesaria la hipótesis auxiliar de  $N_0 \leq 1,000,000$

Sorprendentemente, la hipótesis H4.3 resulta ser inverosímil con los datos, como puede verse en la figura F4.5, y también lo es si cambiamos  $Pearson_P(BE_{G_1}, BE_{G_2}) > Pearson_Q(BE_{G_1}, BE_{G_2})$  por  $Spearmann_P(BE_{G_1}, BE_{G_2}) > Spearmann_Q(BE_{G_1}, BE_{G_2})$ , como puede verse en F4.6

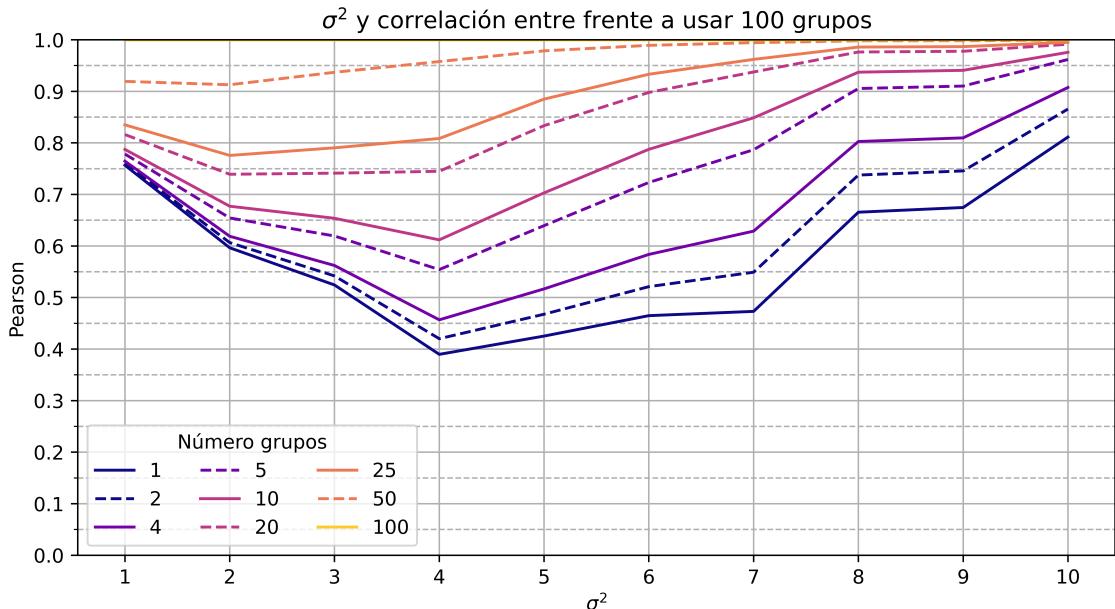


Fig. F4.5: En los datos simulados con diversas varianzas, la  $Pearson(BE_i, BE_{100} | \sigma^2)$ , donde cada serie representa un nivel de granularidad y el eje X los distintos  $\sigma^2$

Observando F4.5 y F4.6 podemos también notar que:

- Para un mismo valor de  $\sigma^2$ , aumentar la granularidad siempre meora la correlación con el caso tomar 100. Esto es verosímil con H4.2
- Las series tienen un comportamiento similar, no en nivel pero si en cambio, frente a los cambios en  $\sigma^2$

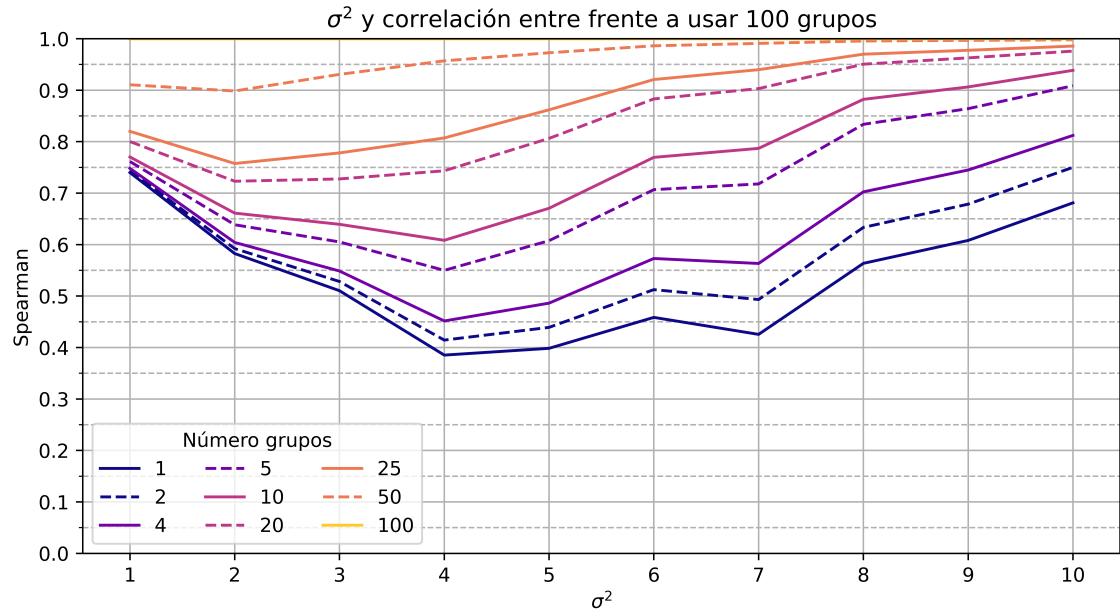


Fig. F4.6: En los datos simulados con diversas varianzas, la  $Spearman(BE_i, BE_{100}|\sigma^2)$ , donde cada serie representa un nivel de granularidad y el eje X los distintos  $\sigma^2$

#### 4.3.3. Comportamiento no condicional

Lo estudiado en H4.2 nos lleva a preguntarnos como se comporta la correlación entre granularidades cuando tenemos diferentes  $\sigma^2$  coexistiendo en la misma muestra.

Lo observado en F4.7 nos indica que tener poblaciones con distintas  $\sigma^2$  aumenta enormemente la correlación entre las distintas granularidades.

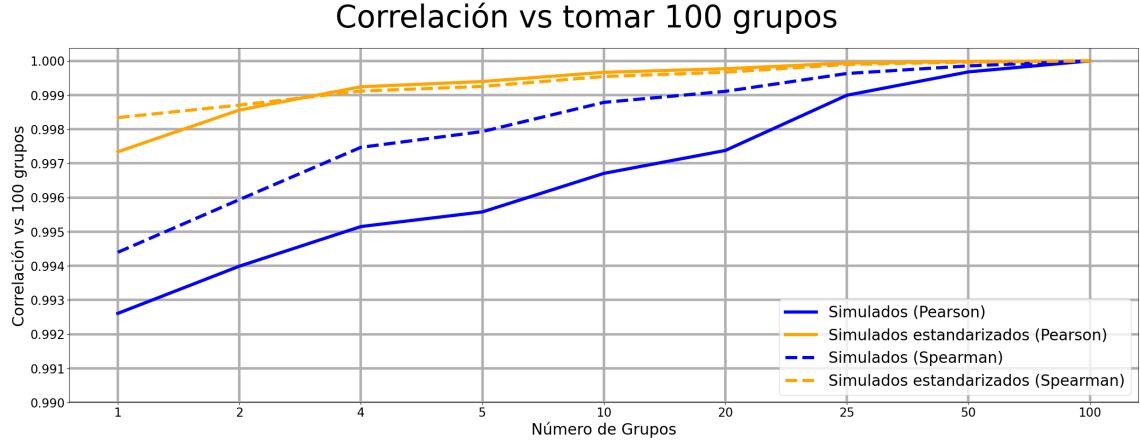


Fig. F4.7: En los datos simulados con diversas varianzas, la  $Spearman(BE_i, BE_{100})$ , donde el eje X son las distintas granularidades (cantidad de grupos). Notar que el eje Y está en el rango [0.99,1]. La serie **simulados estandarizados** corresponde a, para una población  $P_i$ , restarle a cada  $BE_{G_j}(P_i)$  el valor de  $BE_N(P_i)$ , para evitar que medias poblacionales ligeramente distintas (en los logaritmos) influyán en la correlación. Lo notable es que hacer esto aumenta la correlación en vez de reducirla.

Esta característica lleva a pensar que en el caso empírico, como no tenemos a los países agrupados por  $\sigma^2$ , va a haber una correlación fuerte entre las distintas granularidades.

## 5. DATOS REALES: DISTRIBUCIÓN DE LOS INGRESOS

En esta sección se utilizan datos reales de distribución del ingreso de distintos países y años para estudiar su distribución y compararlos con los datos sintéticos antes generados.

### 5.1. Datos disponibles

Se utilizarán datos del Banco Mundial [4] sobre distribución del ingreso en diversos países. Esta base de datos está construida en base a encuestas aportadas por los países, y tiene una granularidad al nivel de los percentiles.

Sobre este dataset, hay que tener en cuenta 2 aspectos importantes:

#### Alcance de la encuesta:

- Nacional
- Urbano
- Rural

En este caso, se utilizarán únicamente los registros de alcance nacional y se ignoraran las encuestas que tienen solo alcance urbano o rural.

#### Tipo de encuesta

- Ingresos
- Consumo

Como estas dos metodologías son incompatibles entre si, todos los análisis se harán por separado para cada uno de los tipos de encuestas.

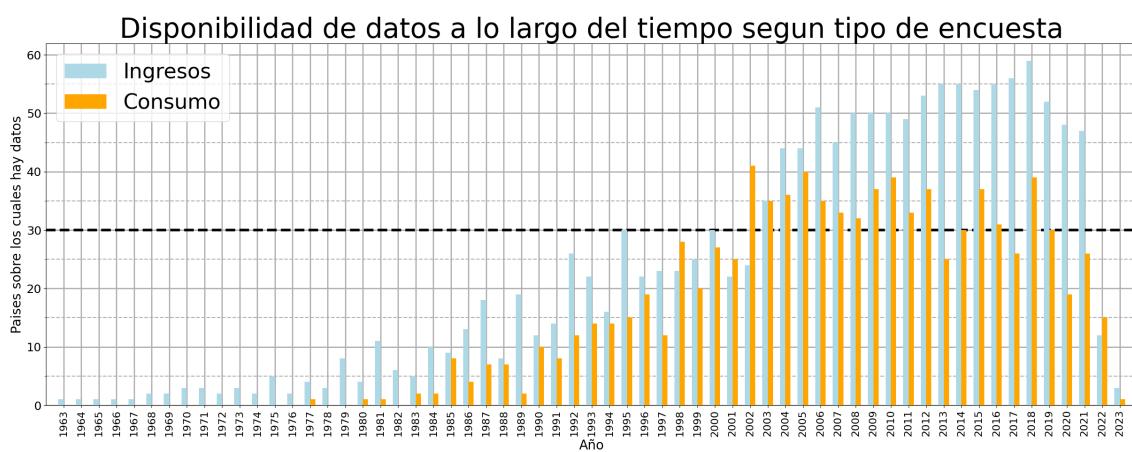


Fig. F5.1: Disponibilidad de las encuestas de ingresos y consumo a lo largo de los años

Cómo puede verse en F5.1, la distribución del ingreso es irregular y ha aumentado en los últimos años. Para que la muestra sea significativa, solo se utilizarán los años para los cuales haya al menos 30 países con datos disponibles para ese tipo de encuesta.

## 5.2. Distribución empírica

Se estudian las distribuciones empíricas disponibles y se las compara con los datos sintéticos generados en el capítulo anterior.

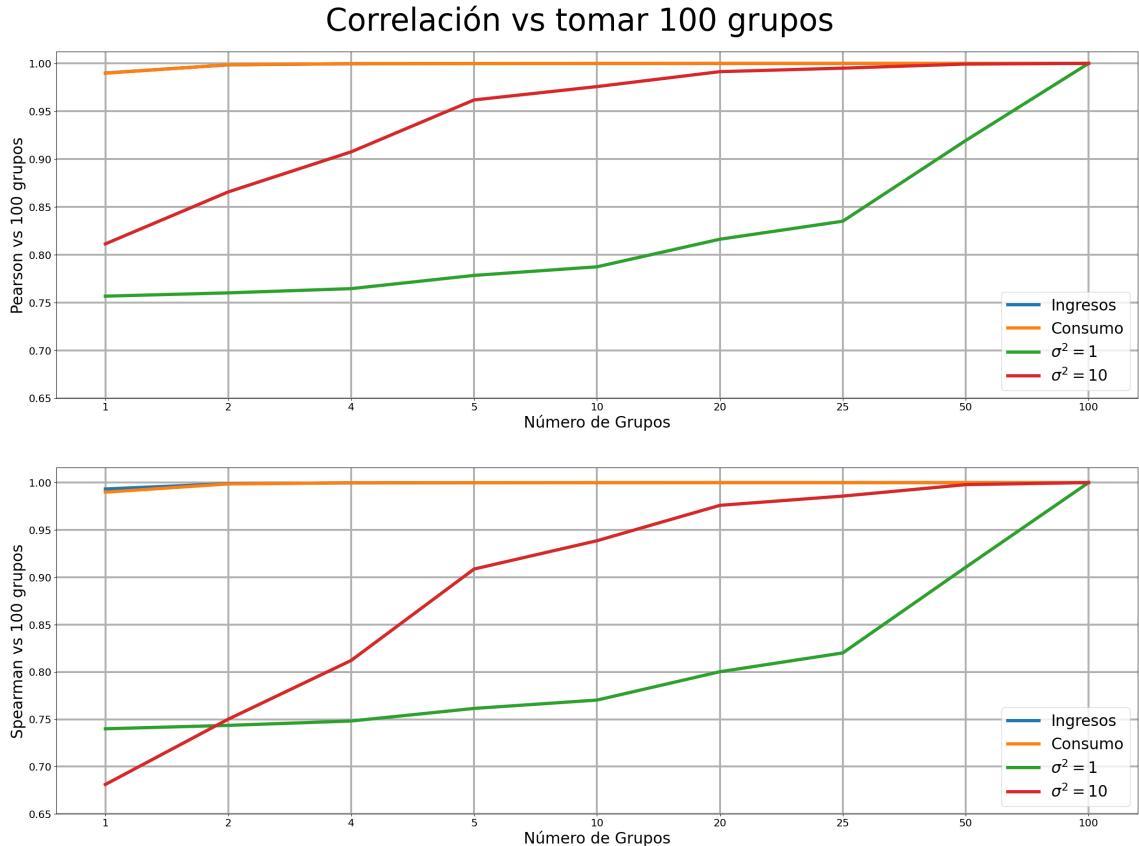


Fig. F5.2: Correlaciones de Pearson y Spearman entre  $BE_1$  y  $BE_{100}$  en los datos reales (separados entre ingresos y consumo) y los datos simulados de  $\sigma^2 = 1$  y  $\sigma^2 = 10$ . Se ve que en los datos reales la correlación es mucho mayor

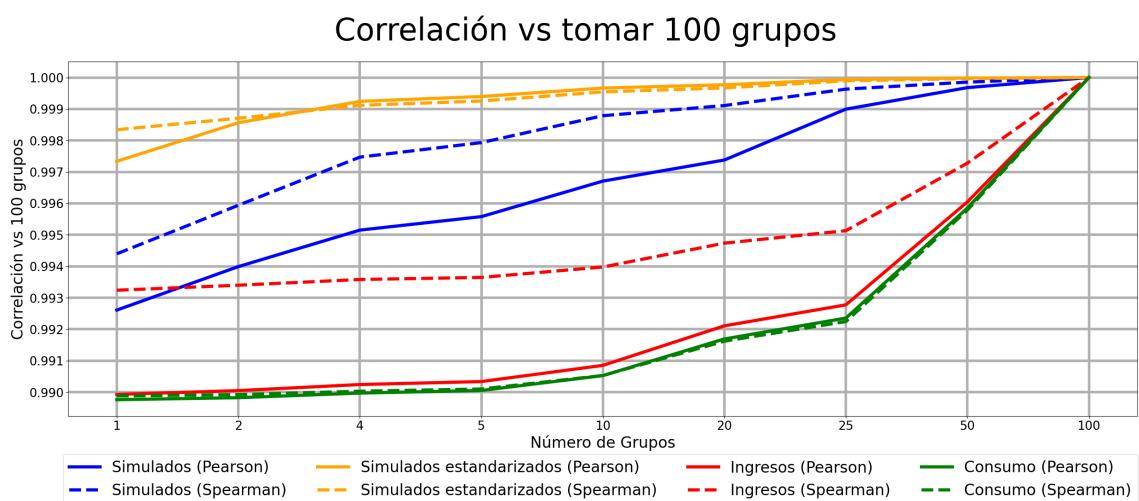


Fig. F5.3: Correlaciones de Pearson y Spearman entre  $BE_1$  y  $BE_{100}$  en los datos reales (separados entre ingresos y consumo) y en los datos simulados sin condicionar por  $\sigma^2$ . Se observa una mayor similitud entre ambas capacidades predictivas

Como puede verse en F5.2 y F5.3, la correlación entre las mediciones del bienestar económico sobre los datos reales condice más con lo observado en los datos sintéticos sin condicionar por  $\sigma^2$  que cuando se condiciona por  $\sigma^2$ . Lo interesante, que puede observarse en F5.4 y F5.5 es que cuando se condiciona por año este fenómeno se mantiene. Esto nos lleva a pensar que esa capacidad predictiva no es debido a una correlación entre el bienestar económico y el año.

### C.C.de Pearson con la medición de granularidad 100

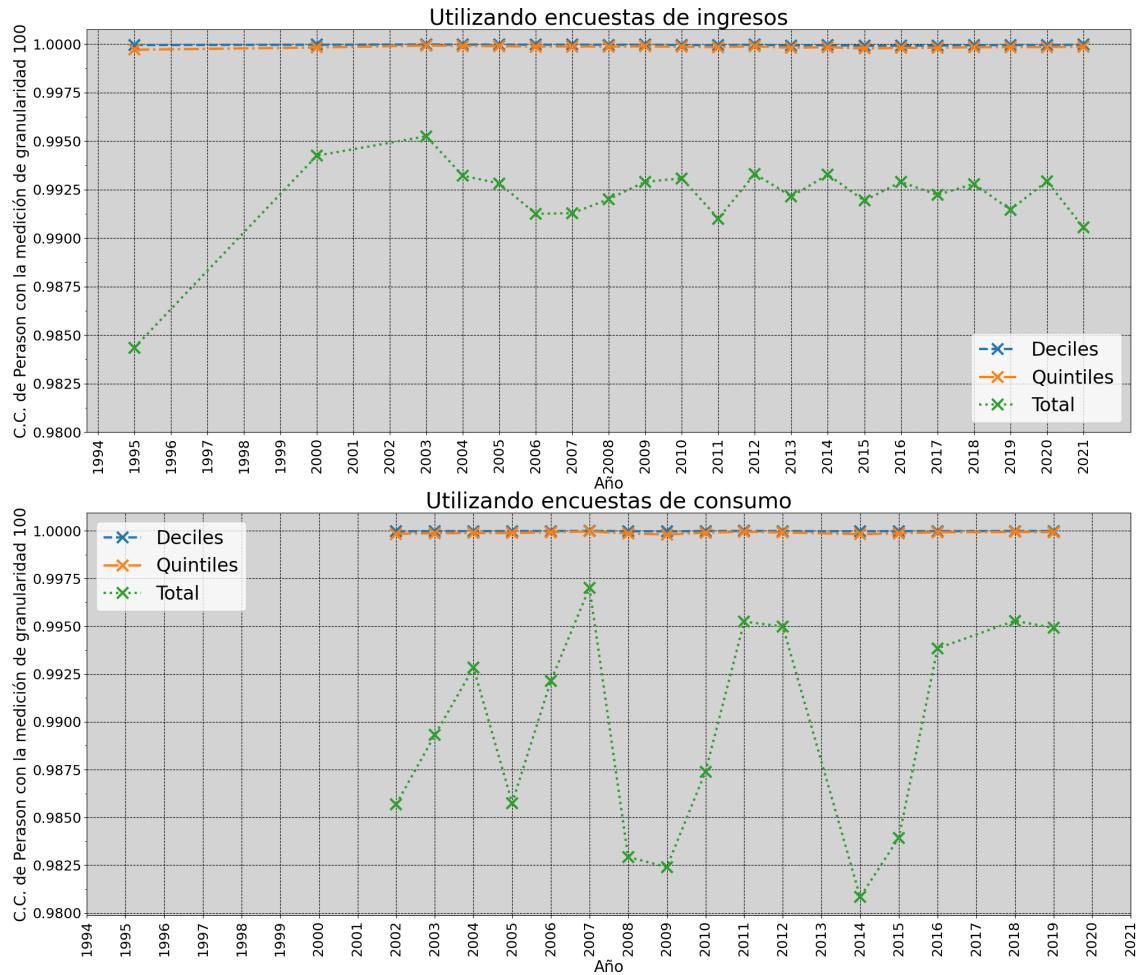


Fig. F5.4: Correlación de Pearson entre utilizar 1, 5 y 10 grupos frente a usar 100 grupos en los datos empíricos, condicionado por año y tipo de encuesta. Solo para los años con al menos 30 países disponibles

### C.C.de Spearman con la medición de granularidad 100

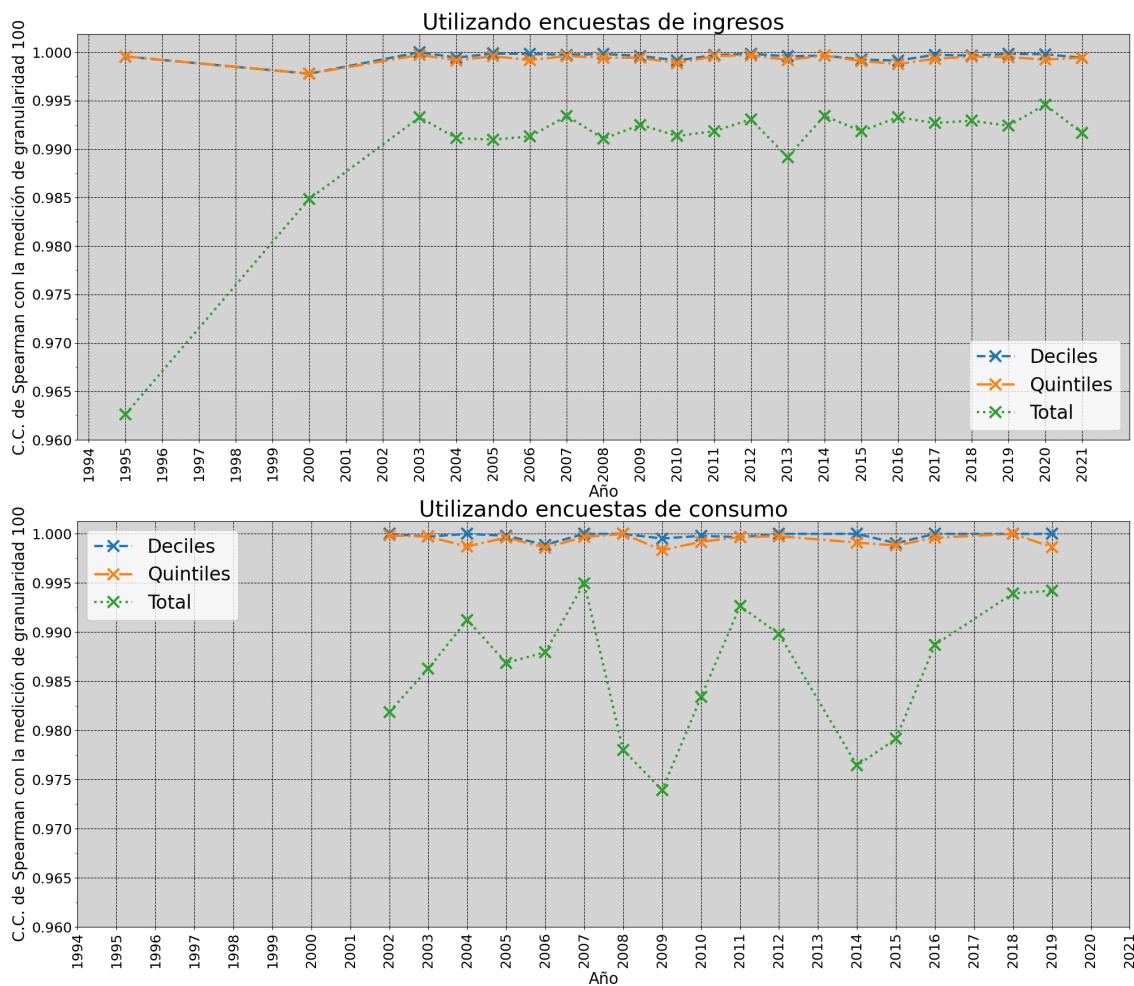


Fig. F5.5: Correlación de Spearman entre utilizar 1, 5 y 10 grupos frente a usar 100 grupos en los datos empíricos, condicionado por año y tipo de encuesta. Solo para los años con al menos 30 países disponibles

## Bibliografía

- [1] United Nations Development Programme. Technical notes: Human development report 2023/24. Technical report, United Nations Development Programme (UNDP), 2023. Accessed: 2024-10-11, [https://hdr.undp.org/sites/default/files/2023-24\\_HDR/hdr2023-24\\_technical\\_notes.pdf](https://hdr.undp.org/sites/default/files/2023-24_HDR/hdr2023-24_technical_notes.pdf).
- [2] R. Gibrat. *Les inégalités économiques*. Sirey, 1931.
- [3] Lautaro Lasorsa. Códigos para la generación de datos simulados, 2024. <https://gist.github.com/LautaroLasorsa/739dae8ec5f9041f243150183070815a>.
- [4] World Bank. Poverty and inequality platform (version 20240627\_2017\_01\_02\_prod) [data set], 2024. World Bank Group. Available at: <https://pip.worldbank.org/>.