Tesis de Licenciatura en Ciencia de Datos (parcial)

Mejora en la medición del Bienestar Económico

Lautaro Lasorsa

Director: Rodrigo Castro
Co-Director: Walter Sosa Escudero
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Ingresos y Bienestar Económico

Definición (Bienestar Económico)

Definimos como **Bienestar Económico** al bienestar o utilidad que un individuo obtiene de sus ingresos.

Ingresos y Bienestar Económico

Definición (Bienestar Económico)

Definimos como **Bienestar Económico** al bienestar o utilidad que un individuo obtiene de sus ingresos.

- $I = Ingresos$
- $BE = Bienestar\ Econmico$
- $BE = log(I)$

Medición del Bienestar Econónomico

Sean $X_1, X_2, \ldots X_N$ los ingresos de los individuos (o familias) en una población:

- **Medición Actual**:

$$log(\frac{1}{N} * \sum_{i=0}^{N} X_i)$$

Problema: El logaritmo del promedio no es el promedio de los logaritmos.

Medición del Bienestar Econónomico

Sean $X_1, X_2, \ldots X_N$ los ingresos de los individuos (o familias) en una población:

- **Medición Actual**:

$$log(\frac{1}{N} * \sum_{i=0}^{N} X_i)$$

  Problema: El logaritmo del promedio no es el promedio de los logaritmos.

- **Medición Ideal (propuesta)**:

$$\frac{1}{N} * \sum_{i=0}^{N} log(X_i)$$

  Problema: Es muy díficil saber los ingresos de absolutamente cada individuo.

Medición del Bienestar Econónomico

Sean $X_1, X_2, \ldots X_N$ los ingresos de los individuos (o familias) en una población:

- **Medición Actual**:

$$log(\frac{1}{N} * \sum_{i=0}^{N} X_i)$$

  Problema: El logaritmo del promedio no es el promedio de los logaritmos.

- **Medición Ideal (propuesta)**:

$$\frac{1}{N} * \sum_{i=0}^{N} log(X_i)$$

  Problema: Es muy díficil saber los ingresos de absolutamente cada individuo.

- **Medición Aproximada (propuesta)**:

$$\frac{1}{G} * \sum_{i=0}^{G} log(\frac{1}{T} \sum_{j=0}^{T} X_{i*T+j})$$

  Teniendo $N = G * T,\ G, T \in Z$

Expectativas

Definición (Granularidad)

La **Granularidad** de una medición aproximada es la cantidad de grupos que utiliza, G. Cada grupo contendrá $N/G = T$ individuos.

Espero que:

Expectativas

Definición (Granularidad)

La **Granularidad** de una medición aproximada es la cantidad de grupos que utiliza, G. Cada grupo contendrá $N/G = T$ individuos.

Espero que:

1. Una medición con mayor granularidad se parezca más a una medición ideal.

Expectativas

Definición (Granularidad)

La **Granularidad** de una medición aproximada es la cantidad de grupos que utiliza, G. Cada grupo contendrá $N/G = T$ individuos.

Espero que:

1. Una medición con mayor granularidad se parezca más a una medición ideal.
2. El parecido entre mediciones con distinta granularidad dependa de la distribución del ingreso.

Expectativas

Definición (Granularidad)

La **Granularidad** de una medición aproximada es la cantidad de grupos que utiliza, G. Cada grupo contendrá $N/G = T$ individuos.

Espero que:

1. Una medición con mayor granularidad se parezca más a una medición ideal.
2. El parecido entre mediciones con distinta granularidad dependa de la distribución del ingreso.
3. Una medición con mayor granularidad sea mejor que una medición con menor granularidad para predecir otras variables.

Partes de la tésis

1. **Datos sinteticos**: Generar datos sinteticos y ver cómo se comportan los primeros dos supuestos.
2. **Datos reales de ingreso**: Evaluar los primeros dos supuestos sobre datos reales y comparar los resultados con lo obtenido en los datos sinteticos.
3. **Predicción empirica**: Evaluar si la medición propuesta mejora la capacidad de predecir otras variables empiricas que **presumo** están relacionadas al bienestar ecónomico.

Generación de datos sinteticos

Lema

Sean $X_1 \dots X_N$ los ingresos de los individuos (o familias):

$$X_i \sim LN(\mu, \sigma^2)$$

$$log(X_i) \sim N(\mu, \sigma^2)$$

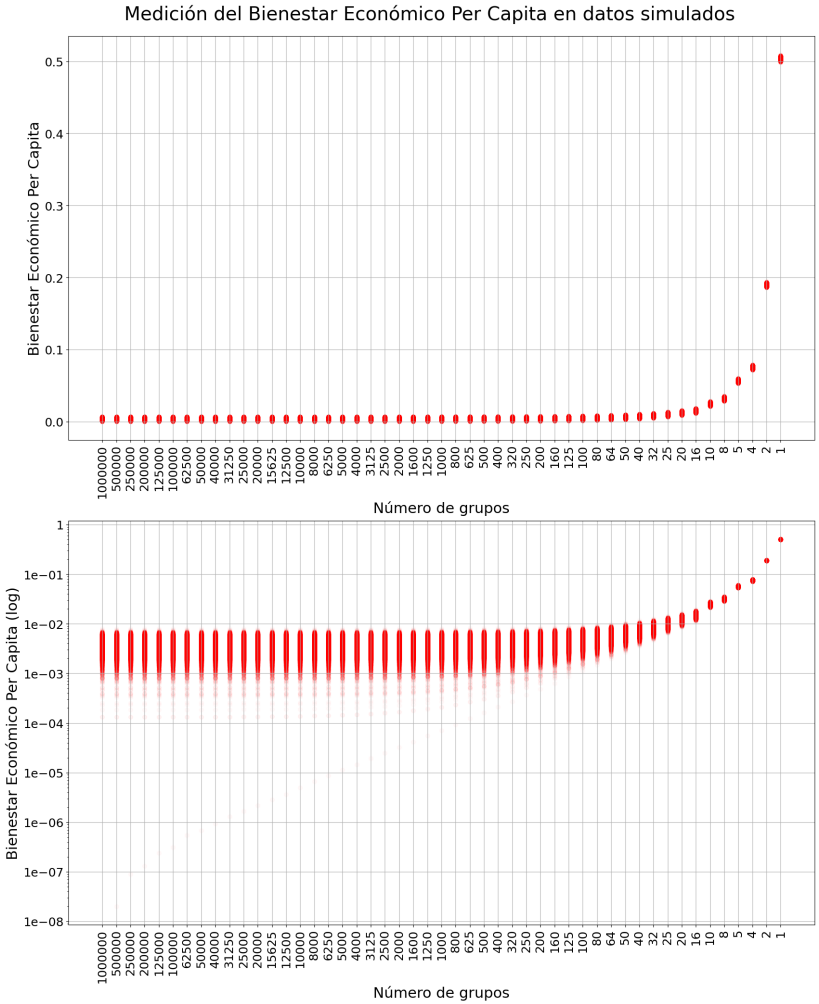
Es decir, los ingresos siguen una distribucion log-normal (LN) de parametros μ y σ^2

Se generarón 2 datasets sinteticos utilizando CUDA para paralelizar los computos.
En ambos se simularon poblaciones con $N = 1\ 000\ 000$ y se hicieron las mediciones parciales para todas las granularidades posibles (incluyen $G = N$ y $G = 1$)

- Un dataset contiene $20\ 000$ poblaciones con distribución $LN(0, 1)$
- El otro dataset contiene 1000 poblaciones con distribución $LN(0, \sigma^2)$ con σ^2 tomando los valores $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$. En total contiene $10\ 000$ observaciones.

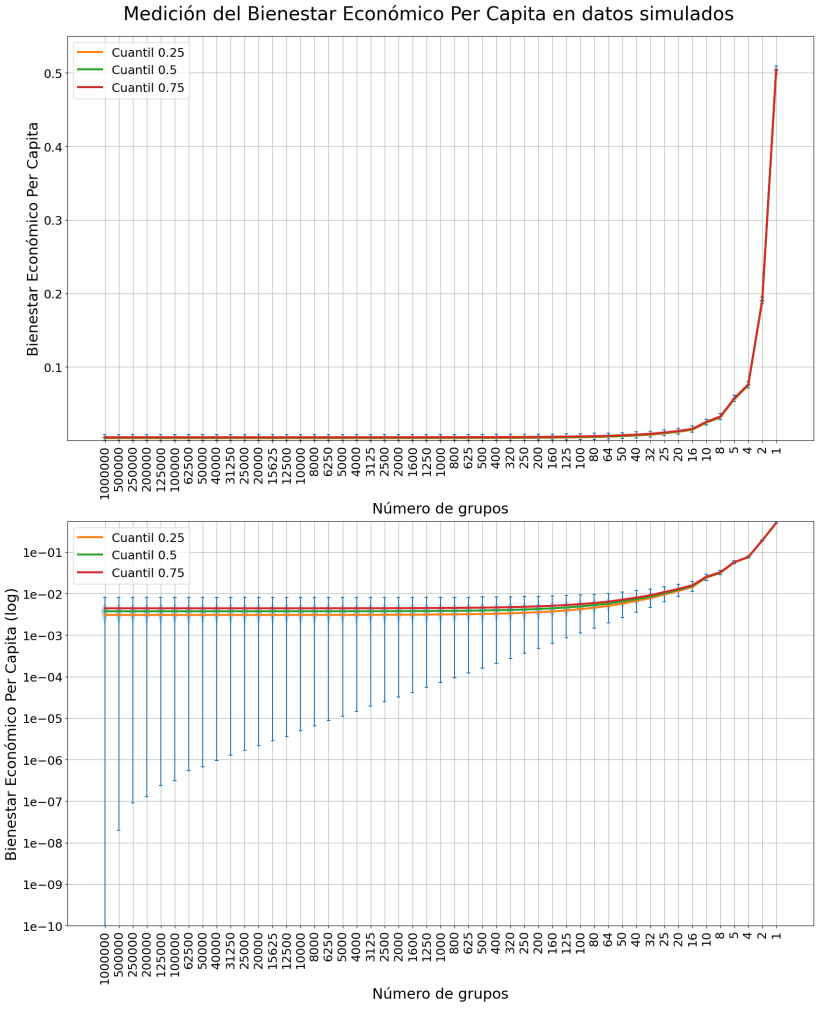
Datos $LN(0,1)$

En este gráfico podemos ver que cada población es un punto en cada valor del eje X, que son las distintas granularidades. En efecto, podemos pensar que el tomar todas las mediciones parciales sobre una población genera un punto en $R^{+|G|}$
Como algunos hay valores negativos (recordar que la medición ideal es $\hat{\mu}$ y $\mu = 0$), para mejorar la visualización se le restó a todos los puntos el menor valor en eje Y entre todos los puntos.
Se puede ver que entre 100 y 200 grupos se alcanzán una distribución dificilmente distinguible de la ideal, mientras que para granularidades menores se nota una diferencia de ordenes de magnitud



Datos $LN(0, 1)$

Podemos ver los mismos datos pero utilizando un violiplot y teniendo marcados los cuartiles de la distribución.



Datos $LN(0,1)$: Correlación

En el siguiente gráfico podemos ver que, consistentemente con lo observado en los gráficos anteriores, aumentar la granularidad aumenta la correlación con la medición ideal

Definición (BE_G)

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

$$T = N/G$$

Datos $LN(0,1)$: Correlación

Conjetura

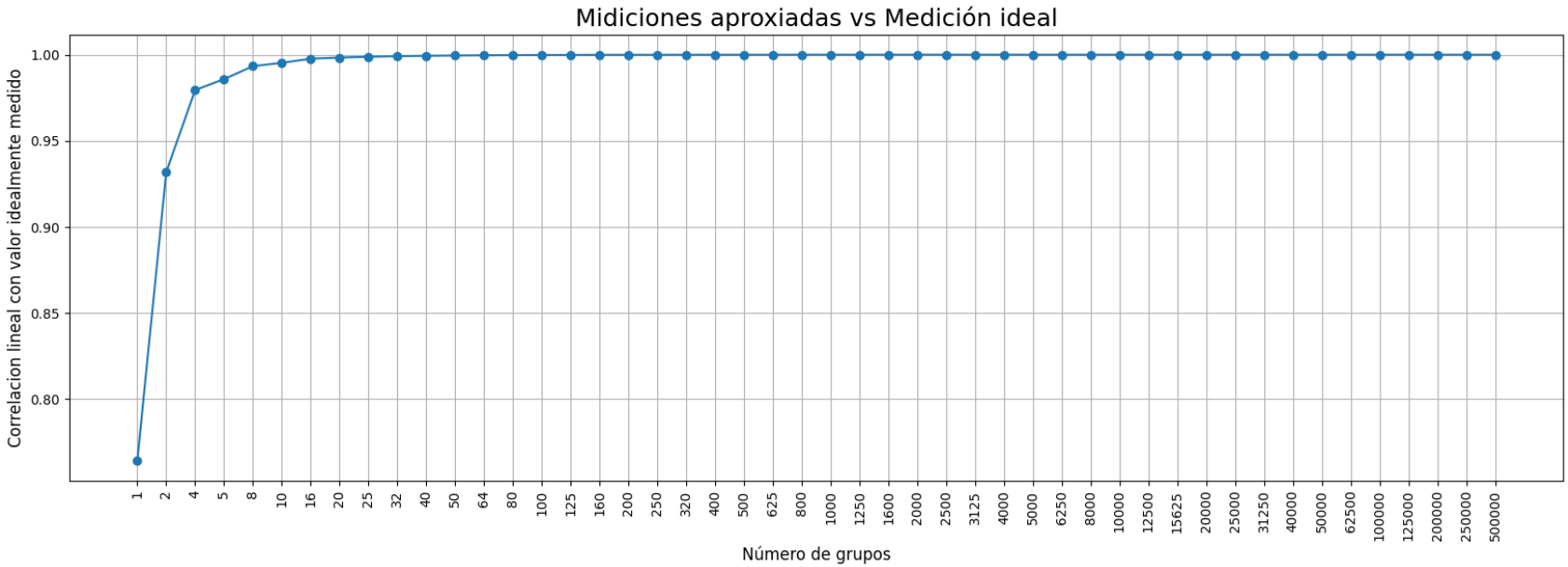
$$i > j \implies Corr(BE_i, BE_N) > Corr(BE_j, BE_N)$$

Datos $LN(0,1)$: Correlación

Conjetura

$$i > j \implies Corr(BE_i, BE_N) > Corr(BE_j, BE_N)$$

Verosimil



Datos $LN(0, \sigma^2)$: Correlación condicional

Podemos ver como el aumentar σ^2 deteriora la capacidad predicitva y nos obliga a utilizar una mayor granularidad.
También podemos ver que llegando a 100 grupos (**percentiles**) la correlación es razonablemente alta para todos los σ^2 evaluados.
Seber esto es importante para la parte empirica porque nos índica que tan bien aproximamos la medición ideal cuando utilizamos datos con menor granularidad.

Datos $LN(0, \sigma^2)$: Correlación condicional

Dados: $X \sim LN(0, \sigma_X^2) \times N$, $Y \sim LN(0, \sigma_Y^2) \times N$ y $\sigma_X^2 < \sigma_Y^2$

Entonces:

$$Corr_X(BE_i, BE_j) > Corr_Y(BE_i, BE_j)$$

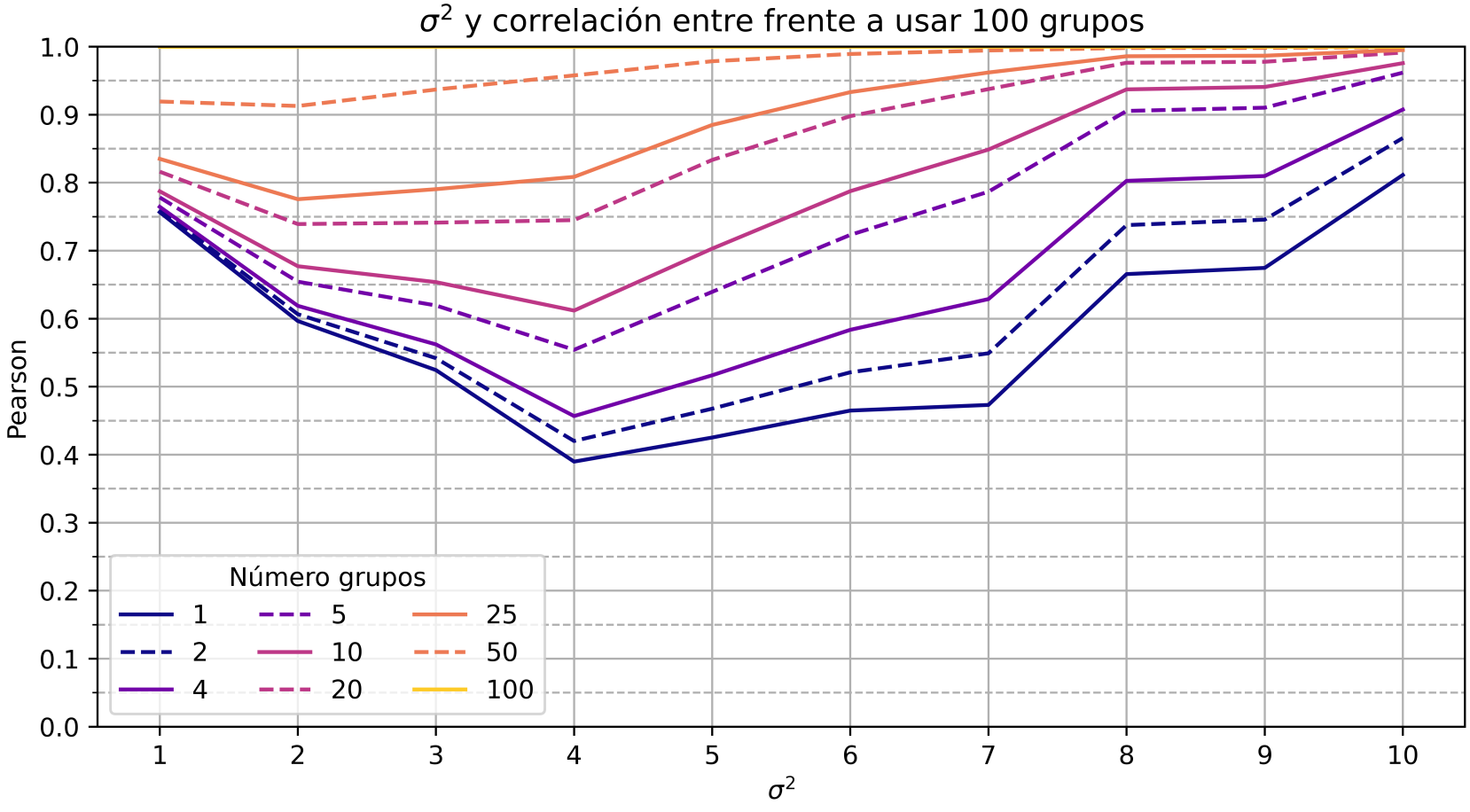
Datos $LN(0, \sigma^2)$: Correlación condicional

Dados: $X \sim LN(0, \sigma_X^2) \times N$, $Y \sim LN(0, \sigma_Y^2) \times N$ y $\sigma_X^2 < \sigma_Y^2$

Entonces:

$$Corr_X(BE_i, BE_j) > Corr_Y(BE_i, BE_j)$$

Inverosimil

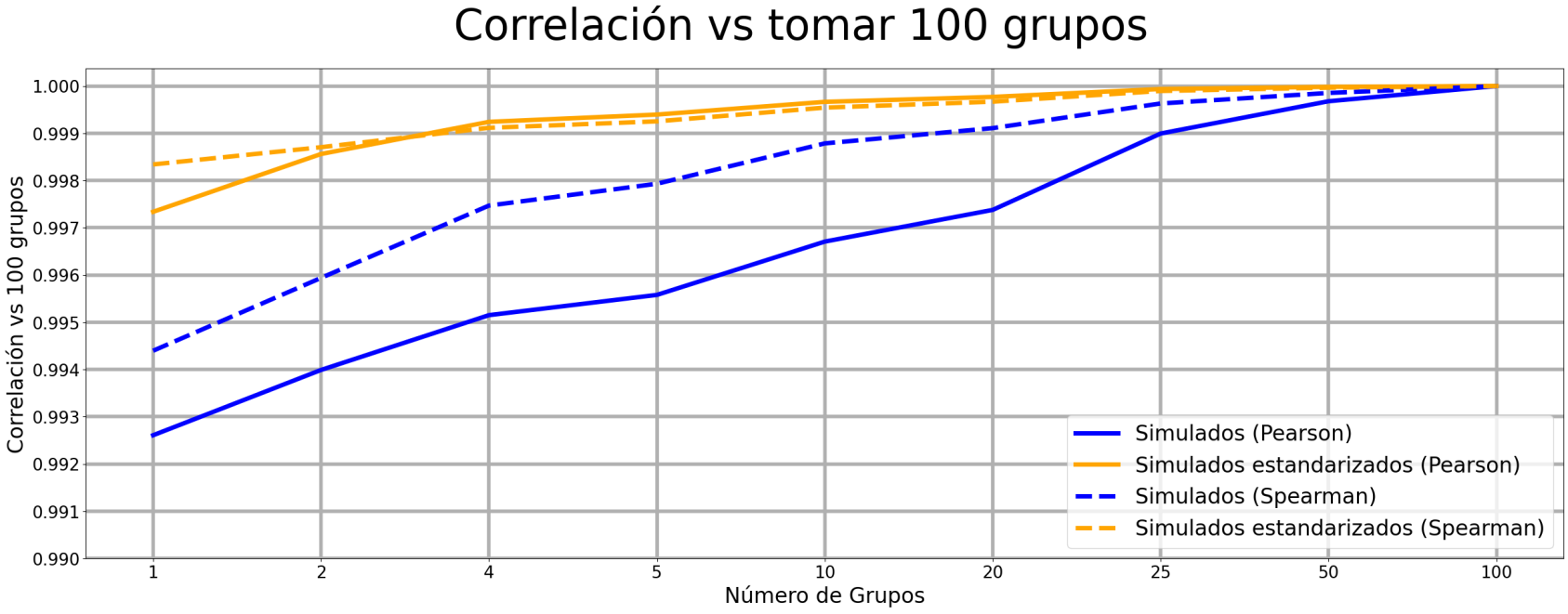


Datos $LN(0, \sigma^2)$: Correlación no condicional

Hemos observado como se comporta la correlación para distintos valores de σ^2, pero hasta ahora siempre condicionando justamente al valor σ^2
Sin embargo, podemos preguntarnos, ya que creamos datos con distintos valores de σ^2. **¿Qué pasa si utilizamos todas las observaciones como una sola población?**

Datos $LN(0, \sigma^2)$: Correlación no condicional

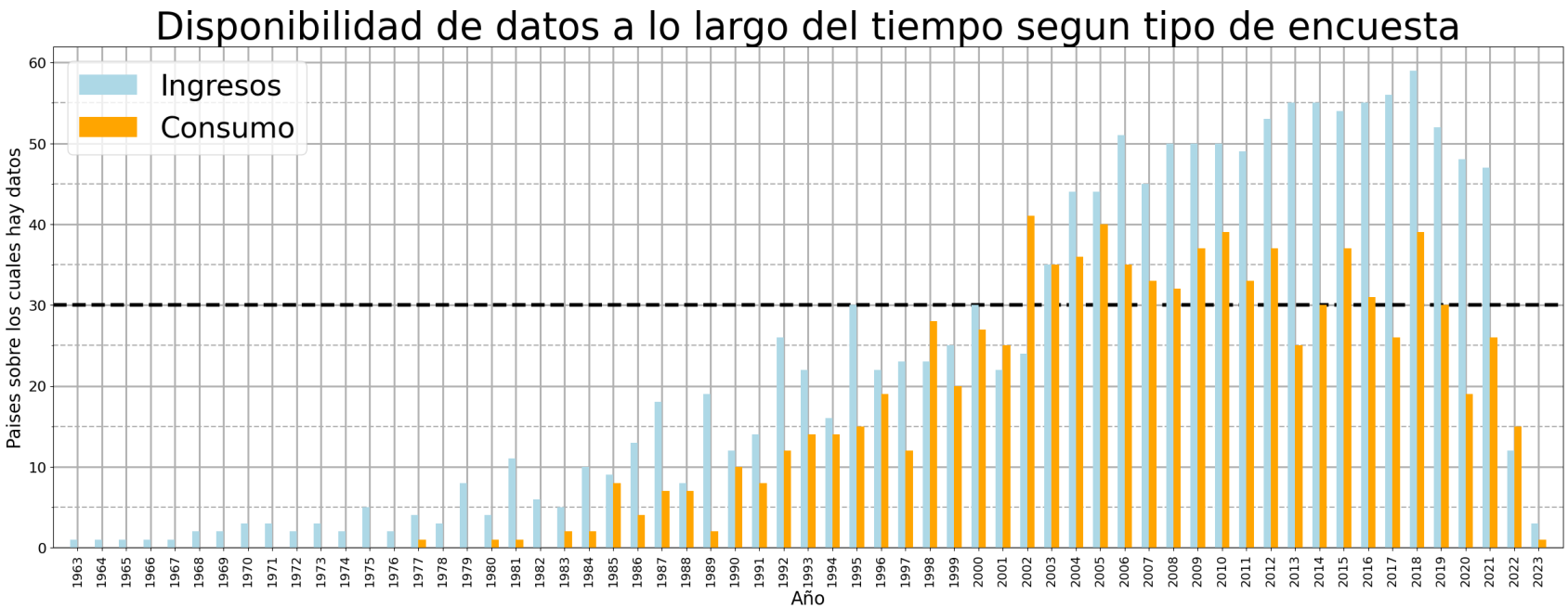
Hemos observado como se comporta la correlación para distintos valores de σ^2, pero hasta ahora siempre condicionando justamente al valor σ^2
Sin embargo, podemos preguntarnos, ya que creamos datos con distintos valores de σ^2. **¿Qué pasa si utilizamos todas las observaciones como una sola población?**
(Prestar mucha atención al eje Y)



Datos reales

Los datos fueron obtenidos del Banco Mundial, y tienen una precisión a nivel de percentil.

- No hay disponibles para todos los países todos los años.
- Se obtienen de los institutos de estadisticas de cada país.
- Hay distintos niveles de alcance: **Nacional**, **Urbano** y **Rural**. Solo usaremos datos de nivel **Nacional**
- Hay dos tipos de encuestas: **Consumo** e **Ingresos**. Estas son metodologicámente incompatibles y deben utilizarse de forma separada.
- Hare todos los análisis en paralelo para ambos tipos de encuestas.



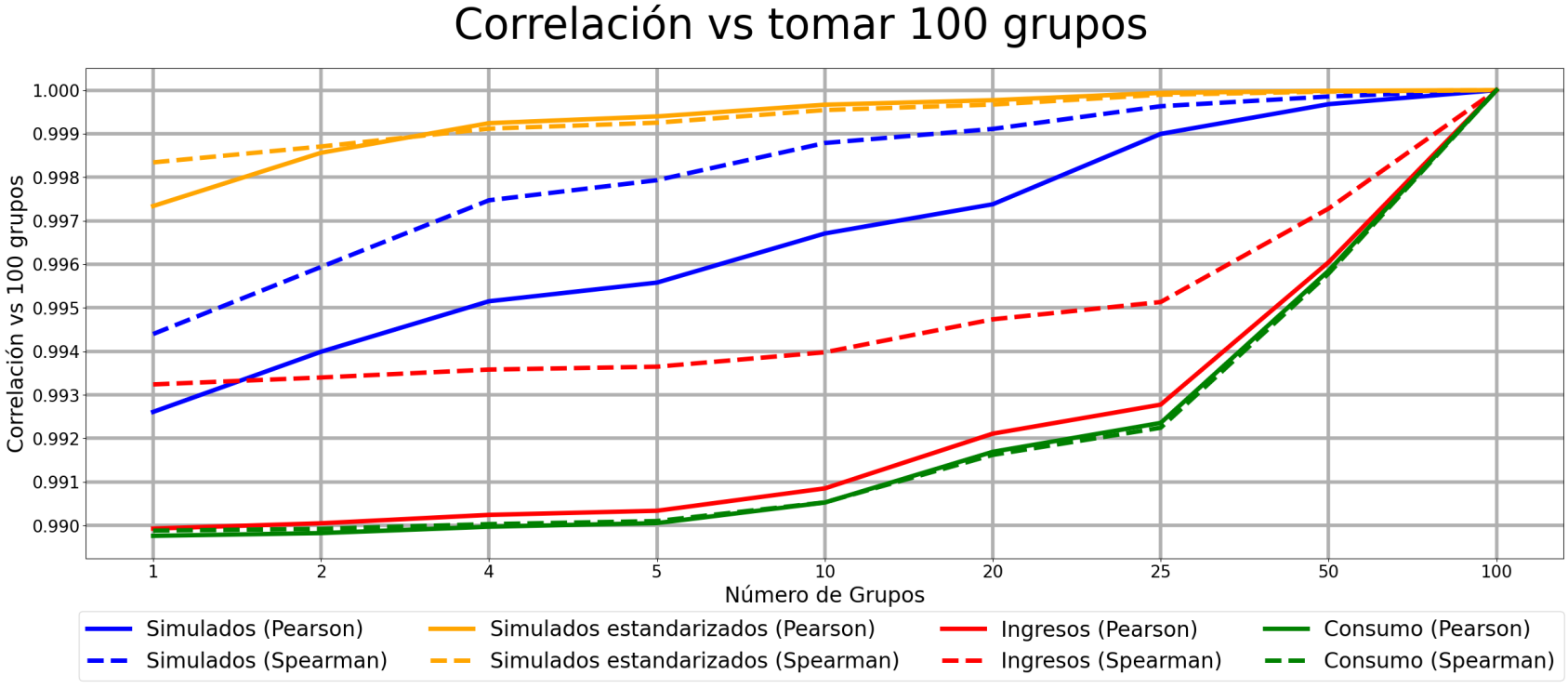
Datos reales: Correlación (no condicional)

Si proponemos reemplazar una medición ya existente (BE_1) con una nueva (BE_{100}), es importante preguntarnos qué tanto se parecen esas 2 mediciones.

Datos reales: Correlación (no condicional)

Si proponemos reemplazar una medición ya existente (BE_1) con una nueva (BE_{100}), es importante preguntarnos qué tanto se parecen esas 2 mediciones.

Se parecen mucho



Datos reales: Esperanza de Vida al Nacer

Definición (Esperanza de Vida al Nacer (EVN))

Es la longitud esperada de la vida al nacer. Se calcula como la edad promedio de muerte en ese año.

La EVN es una variable socioeconómica importante que es razonable suponer que está vinculada con el bienestar ecónomico de la sociedad.

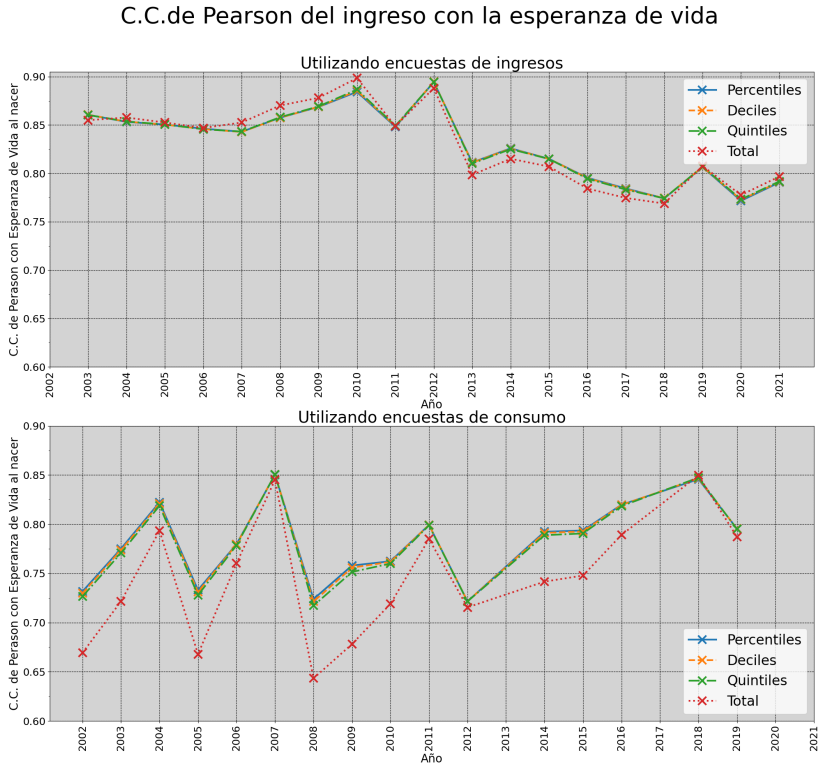
Esperanza de Vida al Nacer: Mejora de la capacidad predictiva

Metodlogía

- Se controlará por año (dado que correlaciona tanto con los ingresos como con la esperanza de vida)
- Dentro de cada año se utilizarán las mediciones con distinta granularidad como variables predictoras de la esperanza de vida.
- Se utilizarán correlaciones de Pearson y Spearman para capturar efectos lineales y no lineales
- Se evaluarán por separado las muestras de **ingresos** y **consumo**
- Solo se tendrán en cuenta en cada caso los años donde haya datos de al menos 30 países.

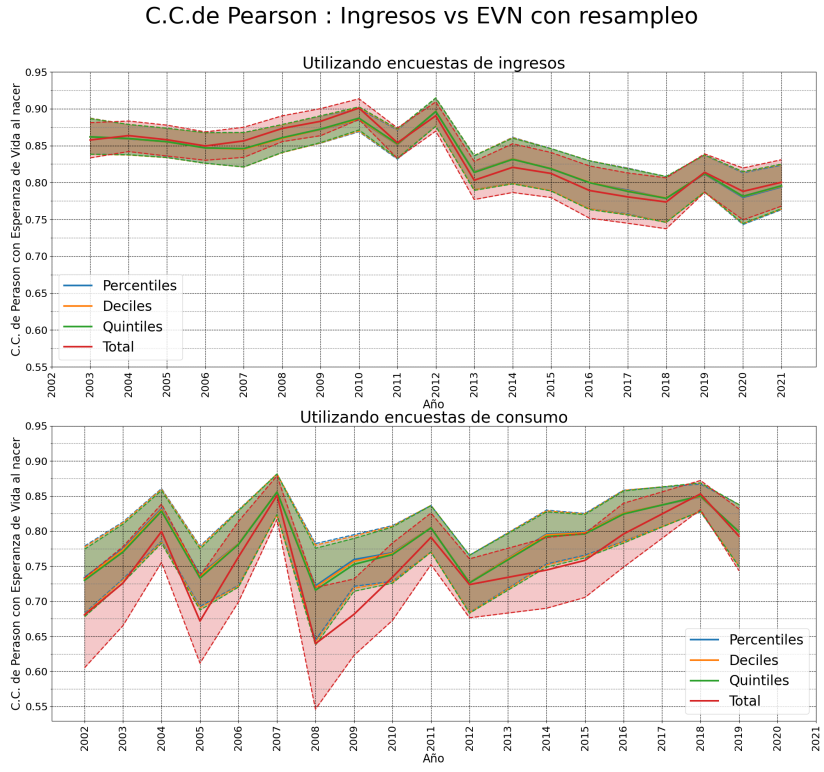
EVN: Resultados

- Los ingresos son mejores predictores que el consumo.
- Utilizando ingresos no hay ganancias significativas (y a veces perdidas) aumentando la granularidad.
- Utilizando consumo hay una mejora significativa al usar quintiles, pero luego no hay más mejora.
- La capacidad predictiva varía mucho entre los distintos años.



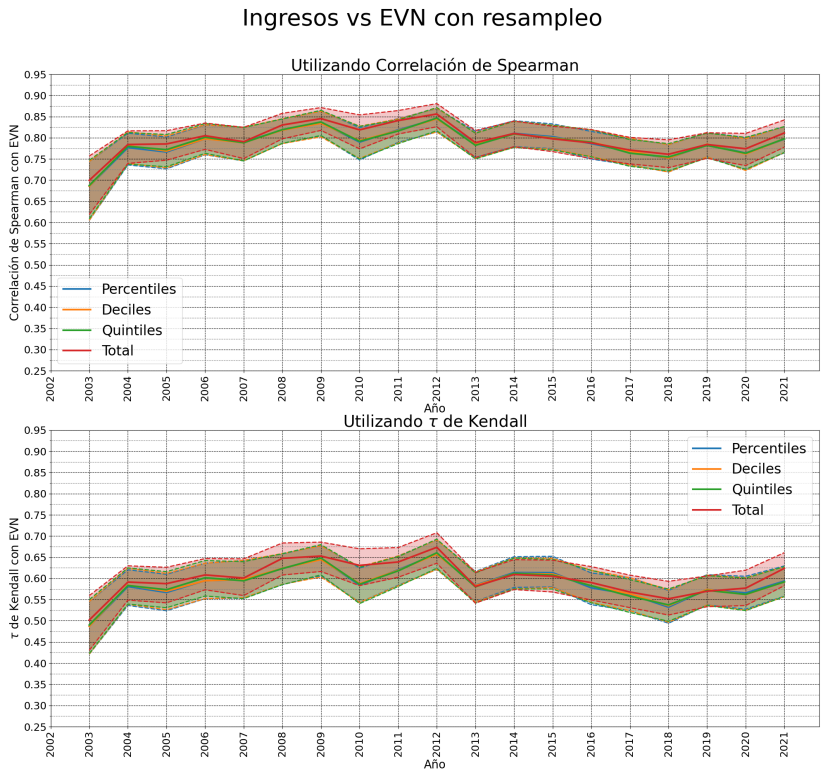
EVN: Remuestreo

- Para cada año, se utilizó bootstrap no parametrico (remuestreo) para estimar la distribución de la correlación las mediciones del bienestar ecónomico y la EVN.
- El sombreado refleja el rango intercuartilico.
- Las conclusiones que podemos alcanzar son similares que con el gráfico anterior.



EVN: Correlación no lineal

- Se repitió el experimento anterior con metricas de correlación no líneal.
- Se mantienen a grandes razgos las conclusiones anteriores.



¡Fin!

¿Preguntas?