Trabajo Práctico Nº1

Un primer encuentro con la EPH

Alumnas

Gete Muriel

Nº de legajo: 33482

Eluney Enria

Nº de legajo: 32503

Milagros Rojas Sanchez

Nº de legajo: 33323

Profesora

Romero, María Noelia

Tutor

Anchorena, Ignacio

Asignatura

Ciencia de Datos

Link al repositorio

https://github.com/murigete/Grupo-ciencia-de-datos.git

<u>Parte I:</u> Famliarizandonos con la base EPH y limpieza

El presente informe tiene como objetivo analizar y comparar las características sociodemográficas y económicas de la población de la región del Gran Buenos Aires, utilizando para ello los microdatos de la Encuesta Permanente de Hogares (EPH) correspondientes al primer trimestre de los años 2005 y 2025. A través del procesamiento y la visualización de los datos, se buscará identificar tanto las estructuras persistentes como los posibles cambios en variables clave como el ingreso, la educación y las condiciones de vida.

Para contextualizar parte del análisis, es fundamental comprender cómo el Instituto Nacional de Estadística y Censos (INDEC) identifica a las personas en situación de pobreza. El método utilizado es la Línea de Pobreza, que consiste en comparar los ingresos de los hogares con el valor de una canasta de bienes y servicios básicos.

El INDEC define dos umbrales:

> **Canasta Básica Alimentaria (CBA):** Contiene un conjunto de alimentos básicos necesarios para satisfacer los requerimientos calóricos mínimos de una persona. Los hogares cuyos ingresos no alcanzan para cubrir la CBA se consideran en situación de **indigencia**.
>
> **Canasta Básica Total (CBT):** A la CBA se le suman una serie de bienes y servicios no alimentarios considerados esenciales, como vestimenta, transporte, educación y salud. Los hogares cuyos ingresos son inferiores al valor de la CBT son considerados pobres.

Este cálculo se ajusta según la composición del hogar mediante el concepto de "adulto equivalente", que asigna a cada miembro un valor en función de sus necesidades nutricionales por edad y sexo. De esta forma, el ingreso total del hogar se compara con el valor de la canasta que le corresponde según su estructura particular.

Procesamiento y Análisis de Datos

1. Selección de Datos y Muestra

Se utilizaron las bases de microdatos y el diccionario correspondiente a los años 2005 y 2025, conformadas por un total de cuatro bases de datos: individual 2005, individual 2025, hogar 2005 y hogar 2025. A partir de esta información se adoptaron decisiones relacionadas con la selección de variables y la depuración de las bases, con el objetivo de elaborar un análisis comparativo entre ambos trimestres.

Para el análisis se seleccionaron 15 variables de interés, de las cuales 8 fueron definidas como obligatorias para los fines del trabajo, mientras que las 7 restantes se incorporaron en función de su relevancia analítica y del interés particular del grupo de investigación. En la Tabla 1 se puede visualizar un resumen de esto.

Tabla 1. Resumen de las variables seleccionadas

Variable	Obligatoriedad	Descripción
CH04	Sí	Sexo
CH06	Sí	Edad en años cumplidos
CH07	Sí	Estado civil
CH08	Sí	Cobertura médica
NIVEL_ED	Sí	Nivel educativo alcanzado
ESTADO	Sí	Condición de actividad (Ocupado, Desocupado, Inactivo)
CAT_OCUP	No	Categoría ocupacional (Patrón, Empleado, etc.)
CAT_INAC	Sí	Categoría de inactividad (Jubilado, Estudiante, etc.)
IPCF	Sí	Ingreso Per Cápita Familiar
IV2	No	Cantidad de ambientes en la vivienda
IV12_3	No	La vivienda está ubicada en villa de emergencia
II7	No	Régimen de tenencia de la vivienda (Propietario, Inquilino, etc.)
IX_TOT	No	Cantidad de miembros en el hogar
IX_MEN10	No	Cantidad de miembros del hogar menores de 10 años
V19_B	No	Percepción de planes sociales (AUH, A. por Embarazo, etc.)

Una vez seleccionadas las variables a manipular, se procedió a realizar la limpieza y unión de las bases de datos para el posterior análisis.

2. Limpieza y Tratamiento de Datos

En primer lugar, antes de proceder a la manipulación de las bases de datos (a partir de este punto, todos los cambios se aplican a las cuatro bases de microdatos), se elaboró una función destinada a generar una tabla exploratoria de las variables, que incluye: nombre de

las columnas, tipo de variable, cantidad de valores nulos, porcentaje de nulos y valores únicos, considerando únicamente las variables de interés previamente filtradas.

Dada la magnitud de las bases, en esta etapa se decidió seleccionar únicamente las variables que serían manipuladas, con el propósito de reducir el costo computacional. Una vez visualizada la base, se procedió a la estandarización de los nombres de las variables. El principal inconveniente identificado consistía en la coexistencia de columnas con nombres en minúsculas y otras en mayúsculas. Asimismo, se eliminaron los espacios finales en los nombres de las columnas. Estas transformaciones facilitaron la posterior unificación de las bases.

Dado que ambas bases contienen información de distintas regiones, se seleccionó el Gran Buenos Aires (GBA) como ámbito de análisis para la presente investigación. Para ello, se filtraron ambas bases mediante la variable "REGION" = 01.

Posteriormente, considerando que algunas variables se encontraban en distintos formatos, se elaboró una función destinada a asegurar la conversión de tipos de datos sin pérdida de información. En particular, todas las variables de tipo *float64* fueron transformadas a *int64*.

Una vez tratadas y estandarizadas las variables de ambas bases, se procedió a su unificación. Para ello, se incorporó una columna adicional denominada "ANIO" en cada una de las cuatro bases, con el fin de identificar la temporalidad de las observaciones una vez integradas. A continuación, se realizaron dos uniones de tipo left join: la primera consistió en vincular la base individual 2025 con la base hogar 2025; la segunda, en vincular la base individual 2005 con la base hogar 2005. De este modo, se obtuvieron dos bases de datos, una para cada año. Finalmente, se llevó a cabo la unión de ambas, conformando la base final denominada "base".

A partir de esta integración, se construyó un heatmap de valores faltantes. En el año 2025 no se registraron valores faltantes en ninguna de las variables (véase Figura 1 en Anexo). En cambio, en el año 2005 se identificaron valores faltantes exclusivamente en la variable IX_MEN10 (cantidad de miembros menores de 10 años en el hogar), con un total de 5.068 casos sin respuesta (véase Figura 2 en Anexo).

Tras la selección y unificación de las bases, se implementó un proceso de depuración orientado a garantizar la calidad y consistencia de la información. La exploración inicial de las 15 variables de interés reveló la existencia de valores que, en una primera revisión, podían interpretarse como inconsistentes. Entre ellos se encontraron códigos numéricos como -1, 0, 9 o 99 en variables categóricas, así como valores atípicos extremos, por ejemplo, una vivienda que declaraba disponer de 31 ambientes.

Al contrastar estos hallazgos con el diccionario de la EPH, se constató que la mayoría de dichos valores no correspondían a errores, sino a códigos específicos utilizados por el INDEC para registrar respuestas particulares. Por ejemplo, el código 0 en la variable

ESTADO representa a “menores de 10 años”, a quienes no se les aplica el cuestionario individual; mientras que el código 9 en CH07 (Estado Civil) refiere a “No sabe / No contesta”. En el caso del código -1, correspondiente a la variable CH06 (Edad), si bien el diccionario no especificaba nada al respecto, se considera imposible poseer una edad negativa.

La decisión metodológica fue tratar estos códigos como datos faltantes para no distorsionar los resultados de los análisis estadísticos posteriores. En lugar de eliminar las filas completas, se optó por reemplazar dichos códigos por el indicador (NaN). Este procedimiento se aplicó a todas las variables afectadas, incluyendo el tratamiento del valor atípico de 31 ambientes, que fue considerado un probable error de carga. De esta manera, se garantiza que los cálculos de promedios, correlaciones y otras métricas se realicen únicamente sobre datos válidos y representativos.

<u>Parte II:</u> Primer Análisis Exploratorio

Una vez procesada y limpia la base de datos, se procedió a realizar nuevamente un gráfico con valores faltantes (véase Figura 3 y 4 en Anexo). Además de un análisis exploratorio inicial para identificar patrones y relaciones en las variables seleccionadas.

3. Composición por Sexo (2005 vs. 2025)

Para analizar la estructura demográfica de la muestra, se realizó un gráfico de barras (vease Figura 5 en Anexo) comparando la composición porcentual por sexo para los años 2005 and 2025 en la región del Gran Buenos Aires.

A partir del gráfico de barras se evidencia una notable estabilidad en la composición por sexo a lo largo de las dos décadas. En el año 2005, la población encuestada se componía de un 52.5% de mujeres y un 47.5% de varones. Para 2025, estas cifras se mantuvieron prácticamente sin cambios, registrando un 52.1% de mujeres y un 47.9% de varones. La diferencia, menor a medio punto porcentual, sugiere que no han ocurrido cambios demográficos significativos en la distribución por sexo dentro de la región estudiada.

4. Análisis de Correlación

Con el objetivo de explorar las relaciones lineales entre las variables sociodemográficas y económicas, se generaron matrices de correlación para los años 2005 y 2025. Previamente, las variables categóricas (CH04, CH07, NIVEL_ED, ESTADO) fueron convertidas a variables *dummy* para permitir el cálculo.

Las matrices de correlación (véase Figura 6 en Anexo) revelan que las relaciones estructurales entre las variables se han mantenido muy estables. Se destaca la educación como un factor clave asociado al ingreso: tener Estudios Superiores Completos (NivelEd_SupComp) presenta una de las correlaciones positivas más fuertes con el Ingreso

Per Cápita Familiar (IngresoPC) en ambos períodos (0.31 en 2005 y 0.23 en 2025). Inversamente, un Primario Incompleto se asocia negativamente con el ingreso.

Asimismo, se observaron patrones lógicos del ciclo de vida, como la fuerte correlación negativa entre la Edad y ser Soltero (-0.71 en 2005 y -0.66 en 2025), la más intensa de la matriz. Esto indica que, a medida que la edad aumenta, la probabilidad de ser soltero disminuye considerablemente. De forma complementaria, la Edad tiene una correlación positiva moderada con ser Casado (EstadoCivil_Casado) en ambos períodos. Un cambio notable es que la relación entre tener Estudios Superiores Completos y estar Ocupado parece haberse fortalecido, pasando de una correlación de 0.16 en 2005 a una más marcada de 0.25 en 2025, lo que podría sugerir una mayor valoración de las credenciales universitarias en el mercado laboral más reciente.

<u>Parte III:</u> Conociendo a los pobres y no pobres

5. No respuesta del Ingreso Total Familiar

Como se menciona en la consigna del TP, uno de los problemas conocidos en la EPH es la falta de respuesta en los ingresos familiares. Para analizar esto, se dividió esta variable en dos bases nuevas:

> "respondieron": compuesta por los hogares donde el ingreso total familiar fue reportado (ITF distinto de 0).
>
> "norespondieron": hogares donde no se informó el ingreso (ITF = 0)

Los resultados generales muestran que 13.680 observaciones respondieron el ingreso total familiar, mientras que 2.985 no lo hicieron. Luego, se analizó por año y en el 2005, casi todos los hogares respondieron (9.371), y solo 113 no lo hicieron. Pero en el 2025 la falta de respuesta creció significativamente, 2.872 hogares no respondieron frente a 4.309 que sí lo hicieron.

Respecto a la condición de actividad (ESTADO) de las personas en los hogares que no respondieron el ingreso total familiar encontramos que en 2005, los pocos casos sin respuesta estaban relativamente distribuidos entre ocupados (17), desocupados (29) e inactivos (36). En 2025, la mayoría de los casos sin respuesta se concentraban en ocupados (1.475) e inactivos (1.003), aunque también se observa un aumento en desocupados (143) y menores de 10 años (230). Esto refleja que el problema de la falta de respuesta sobre ingresos afecta a todos los grupos, pero es más marcado entre los ocupados e inactivos.

6. Nuevas variables: "adulto_equiv" y "ad_equiv_hogar"

Para evaluar la pobreza en los hogares, primero se incorporó la tabla de equivalencia de adultos que permite transformar a cada miembro del hogar en una unidad comparable según edad (CH06) y sexo (CH04). En este sentido, cuanto más pequeño es el individuo y a

partir de la vejez, menor es su peso en términos de necesidad de consumo, además de que hombres y mujeres tienen ponderaciones diferentes en cada rango etario.

Con esta escala se asigna un valor distinto a cada persona. Por ejemplo, un varón recién nacido equivale a 0.35 adultos, a los diez años equivale a 0.79, y entre los 15-17 años llega a 1.04. Luego, en la vejez, su valor relativo desciende a 0.83 entre los 61 y 75 años. En el caso de las mujeres, a los 10 años equivalen a 0.70 adultos, entre los 15 y 17 años a 0.77, y a partir de los 61 años su ponderación disminuye a 0.67, llegando a 0.63 después de los 75 años.

Luego, se sumaron los valores de todos los integrantes de cada hogar (identificados por CODUSU), obteniendo la variable ad_equiv_hogar, que refleja el total de adultos equivalentes en cada hogar. Así logramos estandarizar la comparación entre hogares de distinta composición, lo que resulta clave para evaluar si el ingreso familiar es suficiente para cubrir la canasta básica familiar.

7. Nueva variable: “ingreso_necesario”

Para evaluar el ingreso necesario de cada hogar se incorporó una nueva columna denominada “ingreso_necesario”, calculada como el producto entre la Canasta Básica Total y el valor de “ad_equiv_hogar”. El monto de la Canasta Básica Total varía según el año de la observación: para 2005 corresponde a $205,07, mientras que para 2025 asciende a $365.177. De este modo, a la base “respondieron_equiv” se le adicionó esta nueva variable.

8. Nueva variable: “pobre”

Además, se definió la variable pobre, que toma el valor 1 cuando el ingreso total familiar (ITF) es menor que el ingreso mínimo necesario (calculado como la canasta básica total por el número de adultos equivalentes) y 0 en caso contrario.

Lo que se encontró fue que en 2005, de un total de 9.371 observaciones, 2.485 eran pobres, lo que representa un 26,5% de hogares. En 2025, de 4.309 observaciones, 1.334 eran pobres, lo que equivale a un 30,9% de los hogares. Esto refleja que la incidencia de la pobreza aumentó entre la comparación de ambos años, ya que una mayor proporción de hogares no logra alcanzar el ingreso mínimo necesario para cubrir la canasta básica.

9. Estadísticas descriptivas y gráficos

La Tabla 2 presenta estadísticas descriptivas sobre la pobreza en los años 2005 y 2025. Se observa que, en 2005, de un total de 9.371 observaciones, 2.485 correspondieron a hogares pobres, lo que representa el 26,52% de la muestra. En 2025, sobre 4.309 observaciones, 1.334 hogares fueron clasificados como pobres, alcanzando un 30,96%. En términos relativos, se evidencia un incremento de más de cuatro puntos porcentuales en la proporción de hogares pobres entre ambos períodos.

Tabla 2. Estadísticas descriptivas de pobreza, 2005 y 2025

AÑO	N_TOTAL	N_POBRES	PORCENTAJE	MEDIA	MEDIANA	DESVIACIÓN ESTÁNDAR
2005	9371	2485	26.52	0.265180	0.0	0.441452
2025	4309	1334	30.96	0.309585	0.0	0.462376

Los histogramas de las Figuras 7 y 8 (véase el Anexo) muestran la distribución del Ingreso Total Familiar (ITF), en escala logarítmica (debido a la diferencia de escala se optó por esta forma de normalización), diferenciando hogares pobres y no pobres. En 2005 (Figura 5), se observa una clara separación entre ambas distribuciones: los hogares pobres presentan un ITF promedio de 6,10, mientras que los no pobres alcanzan un promedio de 7,30. En 2025 (Figura 6), aunque la distancia entre los promedios se mantiene (13,35 para pobres y 14,40 para no pobres), las distribuciones presentan una mayor superposición.

Estos resultados sugieren que, si bien la brecha relativa entre los hogares pobres y no pobres persiste en ambos años, la proporción de pobreza aumentó en 2025. Además, la mayor superposición observada en las distribuciones del ITF en ese año podría indicar una mayor heterogeneidad dentro de los grupos, lo que complejiza la caracterización de la pobreza únicamente a partir de los ingresos.

Anexo

Figura 1. Heatmap de valores 2005 previo a la limpieza

Figura 2. Heatmap de valores 2025 previo a la limpieza

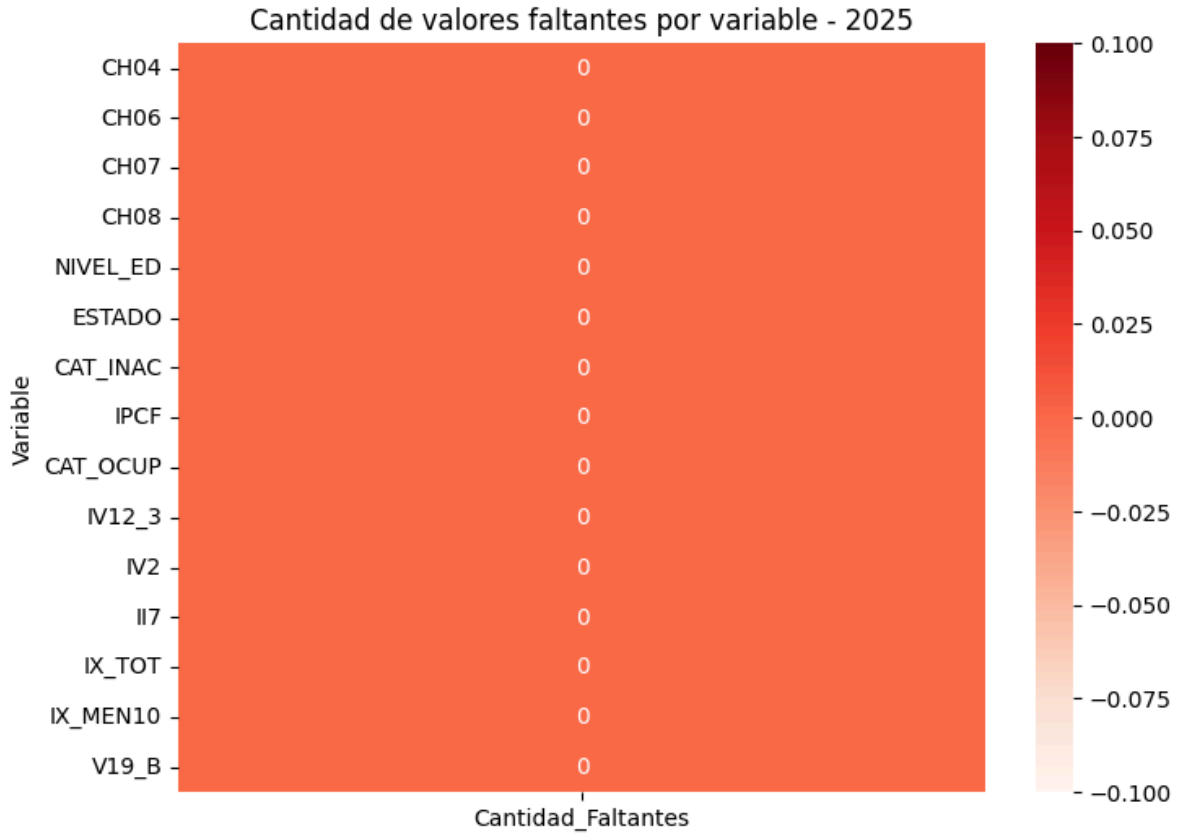

Figura 3. Heatmap de valores 2005 posterior a la limpieza

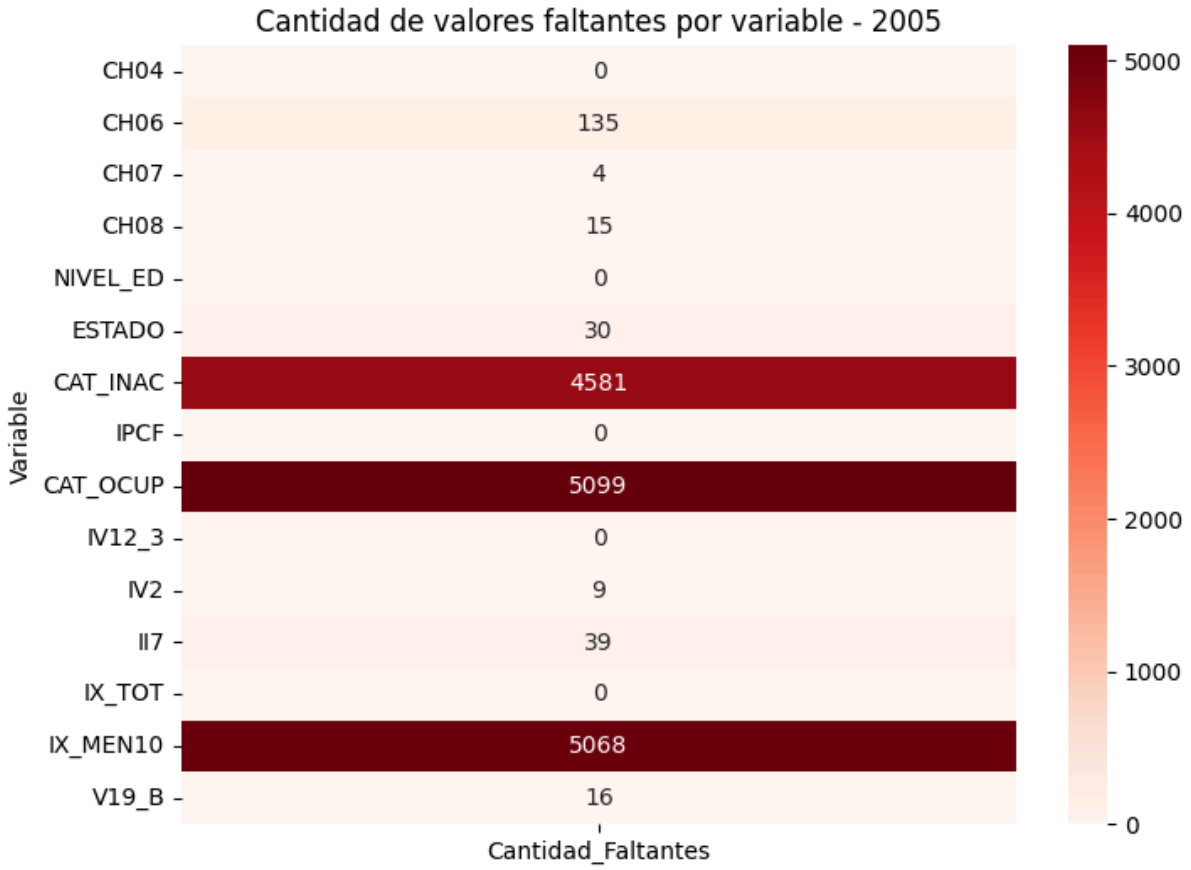

Figura 4. Heatmap de valores 2025 posterior a la limpieza

Figura 5. Gráfico de barras según sexo y año

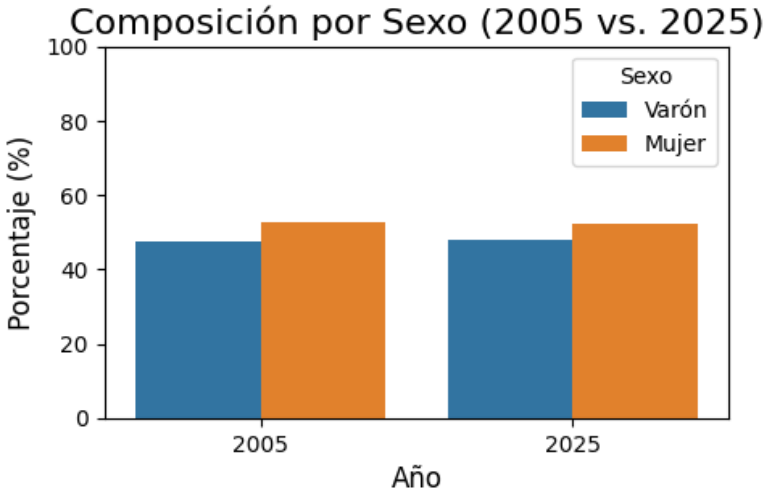

Figura 6. Matriz de correlación

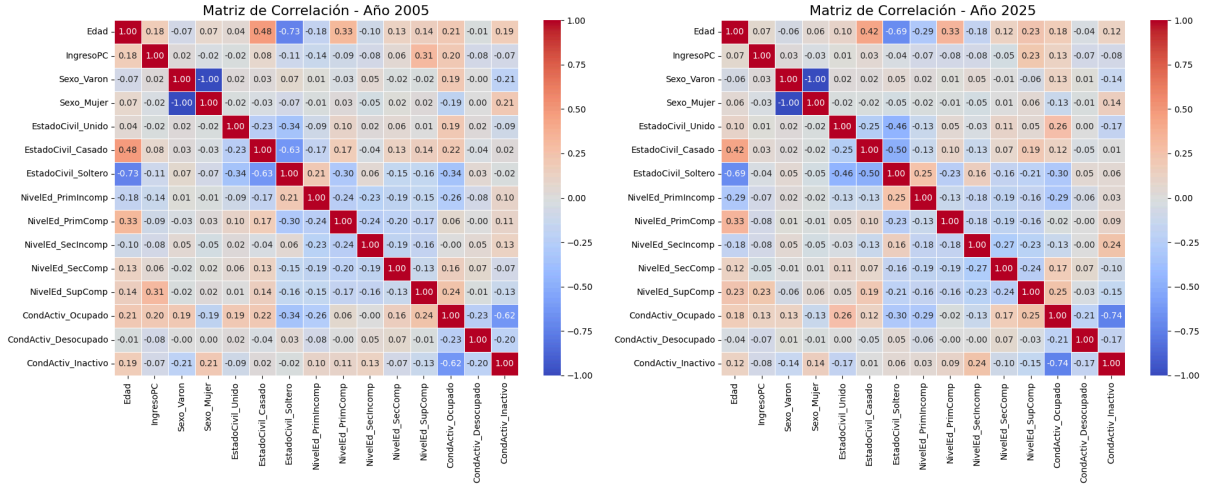

Figura 7. Histograma de ITF, según pobreza del año 2005

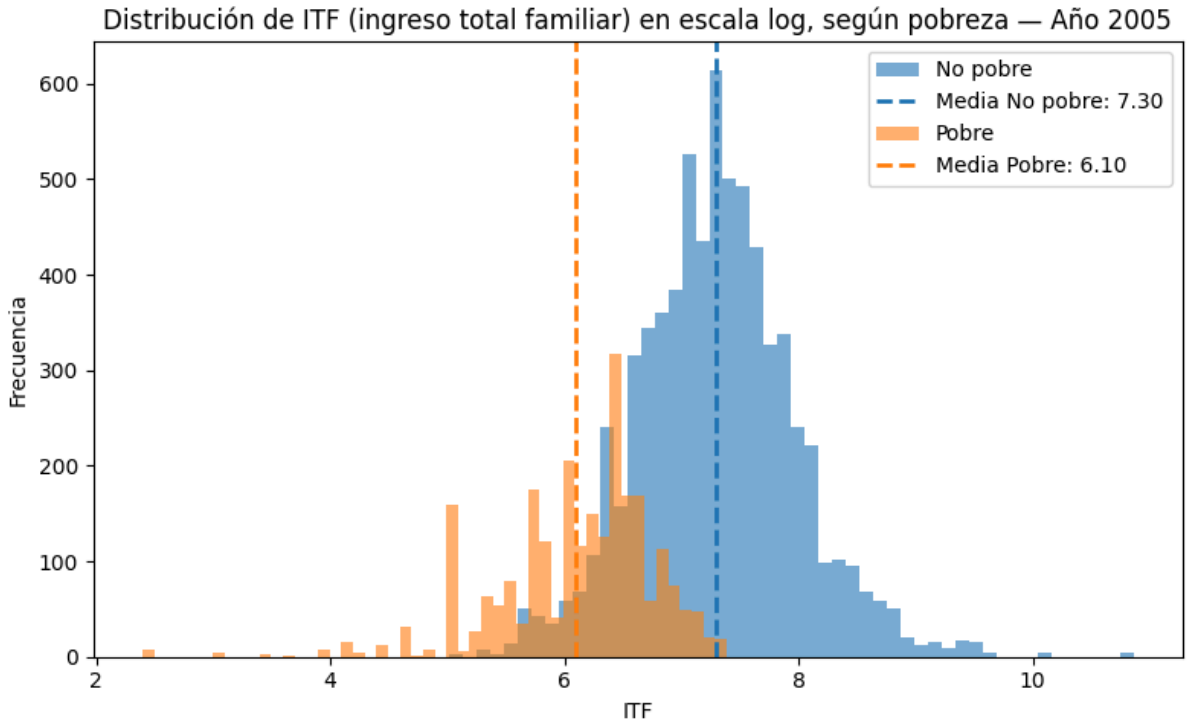

Figura 8. Histograma de ITF, según pobreza del año 2025