

Parcial 2

David Fernández y Juan Karawacki

Primera Parte: Simulación de Eventos Demográficos

Esta primera parte del trabajo final se enfoca en la aplicación de simulación para generar trayectorias de vida individuales. El objetivo central es replicar el comportamiento de una cohorte real mediante la generación de tiempos de espera aleatorios a partir de datos agregados.

Para ello, se utiliza el Método de la Transformada Inversa aplicado sobre tasas específicas por edad, entendidas bajo el supuesto de un Modelo de Riesgo Constante a Intervalos (Piecewise-Constant Hazard).

Selección del Caso de Estudio

Para este ejercicio se seleccionó a Estados Unidos como población de estudio y el año 2015 como período de referencia. Esta elección se justifica por la disponibilidad de datos y el interés en analizar una estructura demográfica moderna, teniendo una alta esperanza de vida y una fecundidad controlada con postergación de la maternidad.

Se basará en analizar dos eventos de naturaleza distinta para poner a prueba el modelo: 1. Fallecimiento (Mortalidad): Un evento universal e inevitable. 2. Primer Hijo (Fecundidad): Un evento no universal (no todas las mujeres lo experimentan).

Los datos provienen de dos fuentes de referencia internacional: - Human Mortality Database (HMD): Para las tasas de mortalidad (M_x). - Human Fertility Database (HFD): Para las tasas de fecundidad condicionales (m_{1x}).

1. Carga de Datos y Análisis

El primer paso consiste en preparar el entorno de trabajo asegurando reproducibilidad. Cargamos las librerías necesarias, la función de simulación personalizada(`ste`).

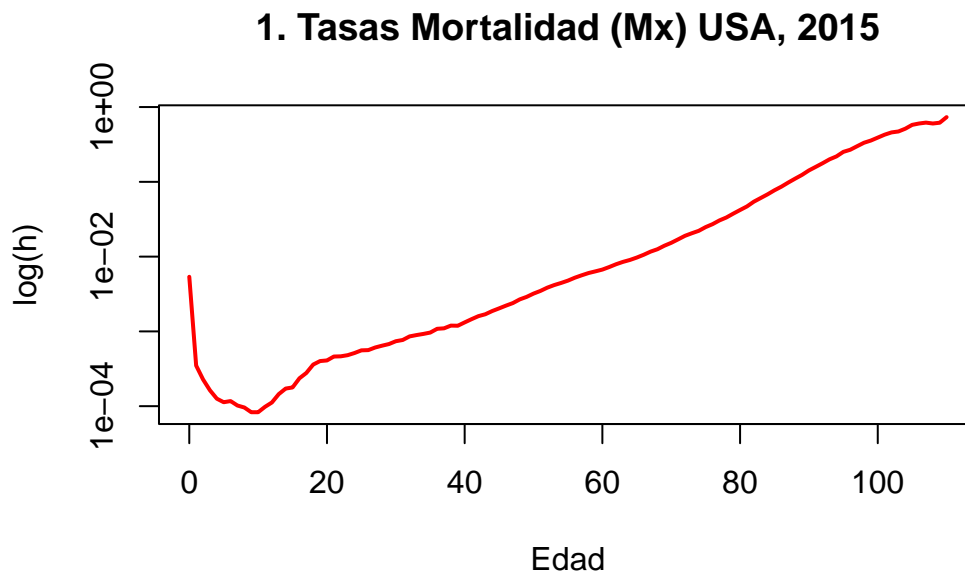
1.1. Procesamiento de las Tablas de Tasas

Se realiza la lectura de los archivos brutos y se limpian para poder trabajar con ellos. Esto es fundamental porque los archivos demográficos pueden tener notaciones de intervalos abiertos (como 110+) o valores perdidos que no permiten el cálculo matemático de las integrales de riesgo, por ejemplo.

Para la mortalidad, se convierte el intervalo abierto en una edad numérica cerrada para permitir la simulación hasta la extinción de la cohorte. Para la fecundidad, imputamos un riesgo cero (0) en aquellas edades donde no se registran nacimientos o existen datos faltantes.

1.2. Análisis Visual de los Insumos (Tasas)

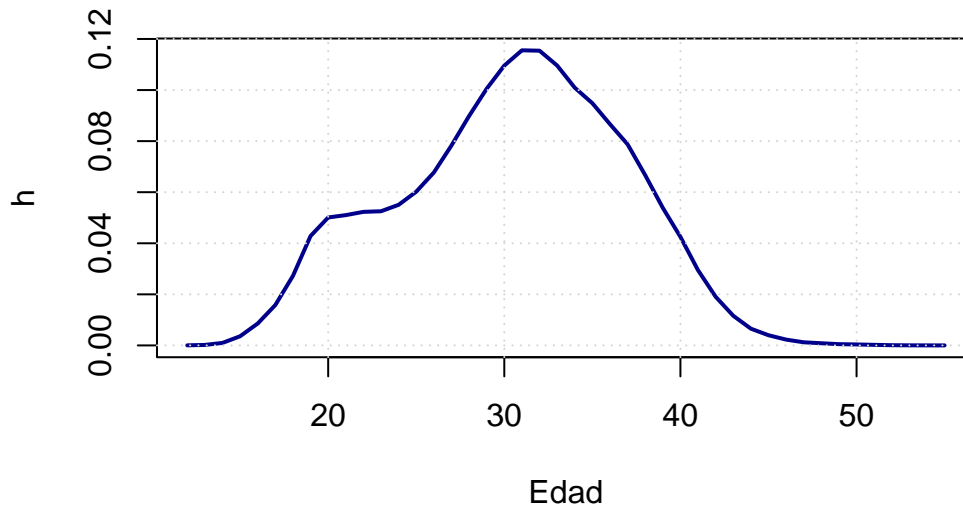
Antes de simular, es importante entender la estructura de los datos de entrada. Graficar las tasas específicas por edad ayuda a verificar que sigan los patrones demográficos esperados.



Descripción: Este gráfico muestra la estructura del riesgo de muerte por edad para las mujeres de Estados Unidos en 2015, se usa una escala logarítmica en el eje vertical para ver adecuadamente las variaciones de magnitud especialmente a edades tempranas. La curva sigue el patrón en forma de “J”, siendo mas característica en las poblaciones modernas. Se observa un nivel de riesgo relativamente elevado en el año 0, que corresponde a la mortalidad infantil, luego se ve un descenso rápido hasta alcanzar el mínimo durante la niñez (alrededor de los 10 años). A partir de la adolescencia y la adultez temprana, la curva toma un ascenso lineal en la escala

logarítmica, lo que significa un crecimiento exponencial del riesgo de muerte que se asocia al proceso biológico.

2. Tasas Condicionales 1er Hijo (m1x) USA, 2015



Descripción: En este gráfico se representa la distribución de las tasas condicionales de tener el primer hijo. Para la cohorte implícita en las tasas de 2015, se observa que la intensidad del inicio de la maternidad es prácticamente nula antes de los 15 años y asciende con 2 modas hasta alcanzar su cúspide pasado los 30 años. Este pico tardío muestra un patrón de postergación de la maternidad visto en sociedades mas desarrolladas. Posteriormente, las tasas descienden de forma acelerada a partir de los 35 años, volviéndose nulas al finalizar el período fértil.

2. Estrategia Metodológica

Para la generación de los tiempos de espera individuales, se usa el Método de la Transformada Inversa. Que permite transformar las tasas demográficas observadas en una distribución de duraciones, bajo el supuesto de que el riesgo es constante dentro de cada intervalo de edad.

Fundamentos del Algoritmo

El procedimiento se basa en tres componentes teóricos que permiten el paso de tasas a individuos:

1. **Modelización del Riesgo Instantáneo:** Se asume que las tasas específicas observadas en las tablas (M_x o m_{1x}) equivalen a la función de riesgo $h(t)$. Para simplificar la integración matemática, se considera que este riesgo se mantiene constante dentro de cada tramo de edad.
2. **Construcción del Riesgo Acumulado:** A partir de las tasas, se llega a la función de riesgo acumulado $H(t)$, que es el resultado de la integración del riesgo hasta el tiempo t . Esta función es creciente y representa la cantidad de riesgo acumulado por un individuo desde el inicio de la exposición.
3. **Inversión de Probabilidad:** Debido a que la función de distribución de supervivencia mapea tiempos a probabilidades en el rango $[0,1]$, se puede hacer el proceso inverso: se genera un número aleatorio U proveniente de una distribución Uniforme $U(0,1)$ y se despeja el tiempo t asociado mediante la formula:

$$T = H^{-1}(-\ln(U))$$

Esto significa buscar la edad exacta T en la cual el riesgo acumulado alcanza el valor determinado por el componente aleatorio $-\ln(U)$.

Implementación Numérica (`ste.r`)

La resolución analítica de la ecuación anterior para miles de casos requiere computación numérica. Para ello, utilizamos la función personalizada `ste()`.

Manejo de Eventos No Universales

Finalmente, el modelo `ste.r` tiene en cuenta que no todos los individuos experimentan todos los eventos (ejemplo, el nacimiento del primer hijo).

Entonces, el algoritmo calcula primero la probabilidad teórica de finalizar el periodo reproductivo sin hijos. Si el número aleatorio U del individuo cae dentro de ese rango de “no ocurrencia”, el sistema le asigna automáticamente un tiempo infinito. Entonces, la simulación distingue de manera correcta entre quienes postergan el evento y quienes nunca lo experimentan.

3. Generación de Datos y Validación del Modelo

Una vez definido la metodología, se hace la simulación de 10.000 trayectorias individuales para cada evento. El objetivo de esta etapa es verificar si el mecanismo propuesto es capaz de reproducir correctamente la estructura de las tasas originales.

3.1. Ejecución de la Simulación

Se usa la función `ste()` descrita anteriormente, tomando como insumo las tasas limpias de mortalidad y fecundidad. El procedimiento no solo devuelve los tiempos simulados, sino también el vector de riesgo acumulado teórico, que es necesario para la validación del modelo.

3.2. Comparación y Validación Gráfica

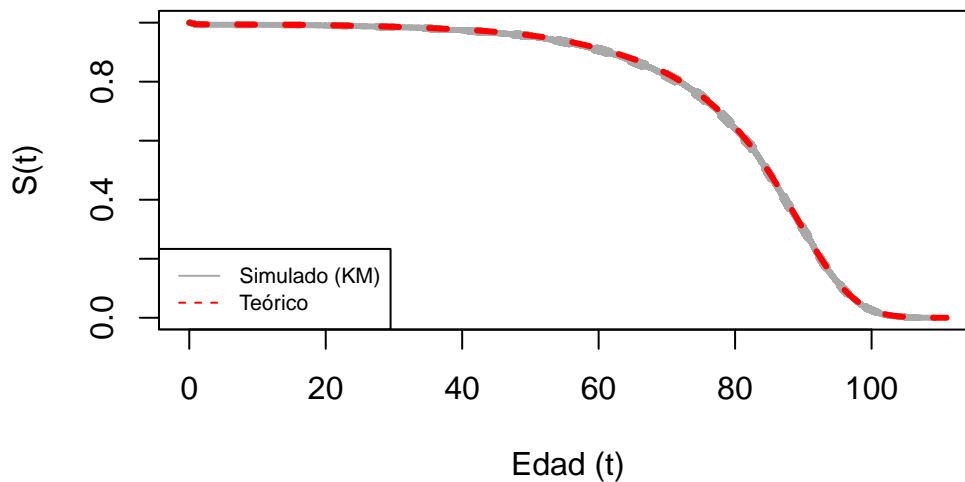
La validación del modelo se realiza con la contraposición de dos curvas de supervivencia:

1. **Curva Esperada:** Calculada matemáticamente a partir de las tasas de entrada ($S(t) = e^{-H(t)}$). Representa el comportamiento ideal del fenómeno.
2. **Curva Simulada:** Estimada mediante el método de Kaplan-Meier sobre los 10.000 tiempos generados por el algoritmo.

Si el procedimiento es correcto, la estimación Kaplan-Meier (línea gris) deberá superponerse a la función teórica (línea de color), confirmando que los datos simulados respetan la distribución de probabilidad original.

Validación de Mortalidad

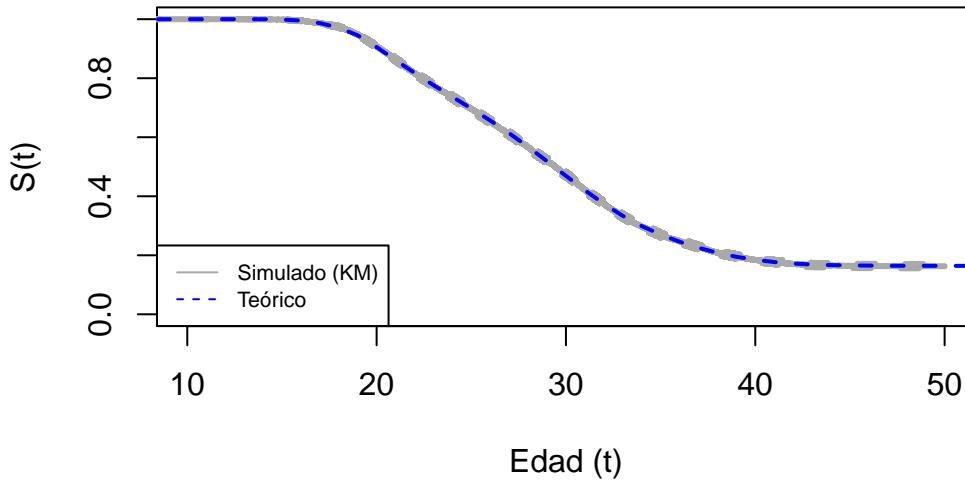
3. S(t) Mortalidad (KM vs Teórico)



Resultados Obtenidos: Este gráfico tiene como objetivo validar la simulación del evento muerte comparando la función de supervivencia teórica ($e^{-H(t)}$) con el estimador Kaplan-Meier derivado de los 10.000 casos simulados. La curva manteniéndose cercana a 1 durante la mayor parte del ciclo de vida indica una alta probabilidad de sobrevivir hasta la vejez. La caída pronunciada de la función se da en las edades avanzadas (mayores de 75 años). La superposición casi exacta entre la curva simulada (gris) y la teórica (roja) confirma la eficacia del método de la transformada inversa para replicar la tabla de mortalidad original.

Validación de Fecundidad (Primer Hijo)

4. S(t) Primer Hijo (KM vs Teórico)



Resultados Obtenidos: Este último gráfico valida la simulación del evento primer nacimiento, comparando la probabilidad de permanecer sin hijos a lo largo del tiempo. La curva muestra un descenso mayor entre los 20 y 35 años, que coincide con las edades de mayor fecundidad. El hecho fundamental de este gráfico es que la función de supervivencia no desciende hasta cero, sino que se estabiliza en una meseta a partir de los 40 años. El nivel que queda al final representa la proporción de mujeres que terminan su vida fértil sin haber tenido hijos. La coincidencia entre ambas curvas en la meseta final valida más el modelo: indica que se asignó correctamente tiempos censurados (infinitos) a las mujeres que no experimentaron la maternidad.

El objetivo de esta segunda parte es trabajar las proyecciones de población. Para completar el ejercicio se deberán seguir los siguientes pasos:

1. Seleccionar un país y dirigirse a <http://www.humanfertility.org> y <http://www.mortality.org> para descargar datos por edades simples de:
 - Tamaño de población por edad y sexo: Human Mortality Database > Period Data > Population Size > 1 Year
 - Tasas específicas de fecundidad por edad: Human Fertility Database > Age-Specific Data > Period > Age-specific fertility rates > Year, Age
 - Tablas de mortalidad (mujeres y hombres por separado): Human Mortality Database > Period Data > Life Tables > 1x1
2. Proyectar la población por sexo y edad desde el primer año posible hasta el último año con datos disponibles.

Importante: Para obtener la proyección es necesario modificar las funciones desarrolladas en clase teniendo en cuenta las diferencias en la estructura de los datos.

3. Comparar la estructura por edades obtenida con el método de los componentes con la estructura observada en el último año.

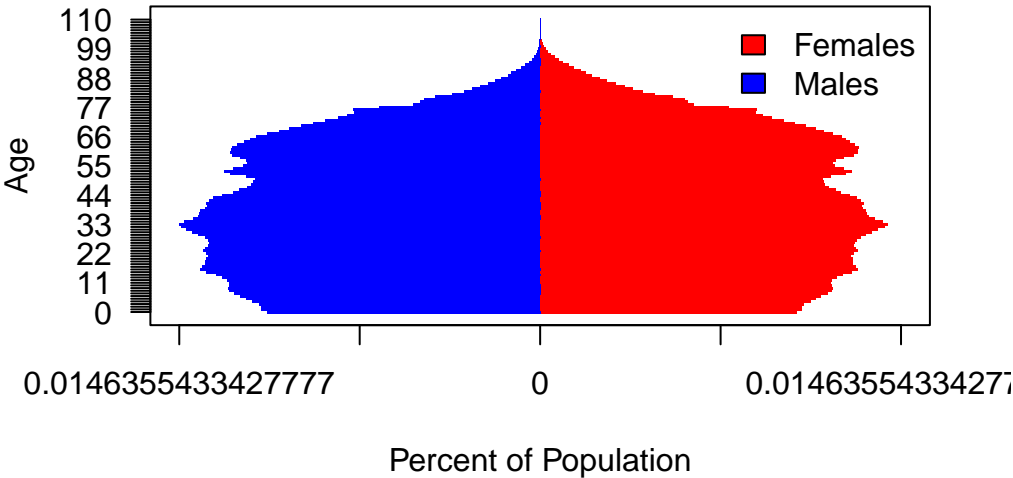
Comparación entre la estructura observada y la estructura proyectada

En esta sección se compara la estructura por edades observada en el último año disponible en las bases de HMD/HFD con la estructura obtenida mediante el método de los componentes ajustado con las tasas específicas de fecundidad y las tablas de mortalidad de período.

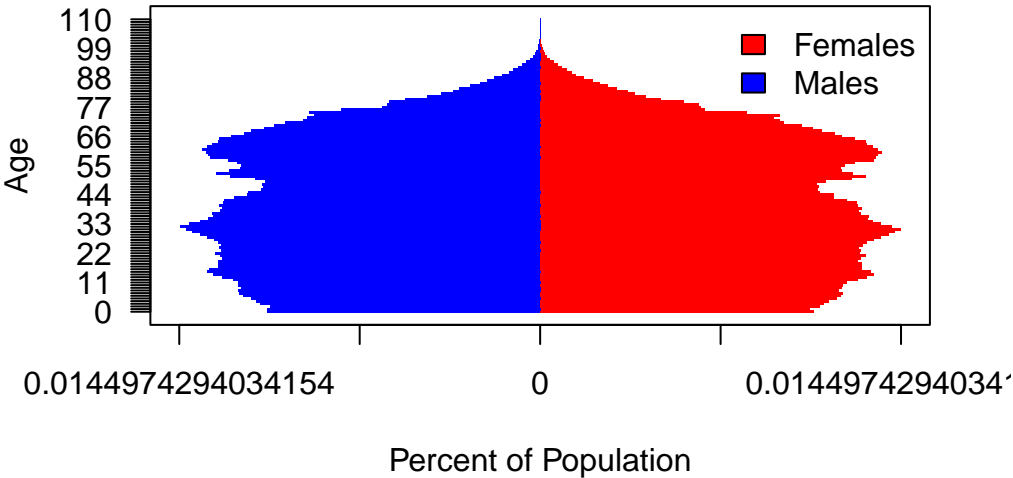
Para ello se construyeron dos pirámides de población:

- La pirámide **observada**, correspondiente al último año con datos disponibles en HMD, que representa la estructura real de la población por sexo y edad.
- La pirámide **proyectada**, obtenida mediante la aplicación del método de los componentes desde el año inicial hasta el último año observado, utilizando como insumos la población base, las tasas específicas de fecundidad por edad y las tablas de mortalidad de mujeres y hombres.

Estructura por Edad de la Población de USA (observada) :



Estructura por Edad de la Población de USA (proyectada) :



Resultados de la comparación

La comparación entre la pirámide **observada (2024)** y la **proyectada (2024)** muestra un grado elevado de coherencia entre ambas estructuras. Las principales conclusiones son las siguientes:

Estructura general

Ambas pirámides presentan una forma muy similar a lo largo de todo el perfil etario, lo que indica que el modelo basado en:

- tablas de mortalidad por período (L_x para mujeres y hombres),
- tasas específicas de fecundidad por edad,
- razón de masculinidad al nacer (SRB),

reproduce adecuadamente la dinámica demográfica central de la población de Estados Unidos.

Edades jóvenes (0–15 años)

Las principales diferencias se observan en la base de la pirámide.

- La población **proyectada** muestra una base ligeramente más regular, mientras que la población **observada** presenta ondulaciones y variaciones más marcadas entre cohortes recientes.
- Estas diferencias son esperables, ya que el modelo **no incluye migración**, un componente clave en la estructura real de la población de EE.UU., especialmente en edades jóvenes y adultas jóvenes.

La ausencia de migración tiende a **suavizar** las cohortes proyectadas y a subestimar o sobrestimar algunos grupos quinquenales recientes.

Edades centrales (20–60 años)

En estas edades, ambas pirámides coinciden notablemente.

- La forma general, los tamaños relativos y el balance por sexo se reproducen con alta precisión.
- Esto refleja que las **tasas de sobrevivencia del período** y la estructura original del año inicial permiten proyectar adecuadamente el “cuerpo” de la pirámide.

Edades avanzadas (65+)

Las diferencias vuelven a ampliarse ligeramente hacia la cúspide de la pirámide.

- La población observada muestra una mayor supervivencia en edades muy avanzadas (especialmente mujeres), mientras que la población proyectada presenta una reducción más marcada en los grupos más longevos.
- Esto es coherente: las **tablas de mortalidad por período** no incorporan las **mejoras futuras** en supervivencia, por lo que la proyección tiende a **subestimar** la longevidad real.

Conclusión

La estructura proyectada para 2024 reproduce de manera consistente la forma observada, especialmente en las edades centrales. Las diferencias en los extremos del perfil etario (jóvenes y longevos) responden a las limitaciones propias del **método de los componentes**:

ausencia de migración

Aunque el modelo no incorpora migración como componente explícito, **sí hereda indirectamente los efectos históricos de la inmigración**, ya que:

- la población inicial contiene migrantes,
- las tasas de fecundidad se calculan sobre población residente,
- las tasas de mortalidad reflejan la composición real del país.

Sin embargo, al no proyectar migración a futuro, se observan diferencias en:

- edades jóvenes (suavización del extremo inferior),
- edades adultas jóvenes (subestimación en 18–40),
- edades avanzadas (subestimación por ausencia de mejoras futuras en mortalidad).

Uso de tasas de fecundidad y mortalidad del período sin mejoras futuras

Cuando el modelo utiliza tasas de fecundidad y mortalidad **del período**, dichas tasas representan únicamente el comportamiento observado en un año específico. Esto implica que:

- no se incorporan **tendencias futuras**, como el aumento sostenido en la esperanza de vida,
- no se consideran cambios en la fecundidad asociados a transformaciones sociales, económicas o migratorias,

- la proyección asume que las condiciones actuales **se mantienen constantes** a lo largo del tiempo.

En consecuencia, la población proyectada tiende a **subestimar la longevidad en edades avanzadas**, ya que no incluye las mejoras que efectivamente ocurren en mortalidad año tras año.

Suavización automática de las cohortes por efecto del modelo Leslie

El modelo Leslie aplica:

1. una matriz de **supervivencia** que desplaza cada cohorte un año hacia adelante, y
2. una matriz de **fecundidad** que genera nacimientos de manera consistente entre años.

Este mecanismo produce una proyección donde:

- las cohortes se vuelven **más suaves y regulares**,
- se eliminan las oscilaciones abruptas presentes en los datos reales,
- las fluctuaciones que en la realidad surgen por migración, crisis económicas o shocks de fecundidad **no aparecen** en el modelo.

Por ello, las pirámides proyectadas suelen mostrar curvas más **limpias y homogéneas**, en contraste con las pirámides observadas, donde la historia demográfica introduce irregularidades visibles en segmentos específicos de edad.

A pesar de estas limitaciones, el método ofrece una aproximación sólida y coherente de la estructura poblacional real de Estados Unidos en el último año disponible.

4. Describir en detalle los procedimientos utilizados y los resultados obtenidos.

Preparación y descarga de datos

Para el ejercicio se seleccionó **Estados Unidos** como país de análisis. Se descargaron datos de las fuentes indicadas en la consigna:

- **Tamaño de población por edad y sexo** (Human Mortality Database, *Period Data* → *Population Size* → *1-year age groups*).
- **Tasas específicas de fecundidad por edad (ASFR)** (Human Fertility Database, *Age-Specific Data* → *Period* → *Age-Specific Fertility Rates*).
- **Tablas de mortalidad por período** para mujeres y hombres (Human Mortality Database, *Period Life Tables* → *1x1*).

Los archivos fueron leídos en R y procesados mediante la función `data_prep()`, la cual:

- filtra los datos desde el primer año común disponible,
- normaliza categorías de edad (por ejemplo, “110+” a “110”),
- convierte edades en valores numéricos,
- reorganiza los datos en formato ancho según año,
- genera la población inicial del año utilizado para comenzar la proyección,
- guarda un archivo `last_pop.txt` con la estructura observada más reciente, utilizada luego para la comparación final.

El año inicial obtenido automáticamente en el caso de Estados Unidos fue de 1933, con lo cual la proyección comienza a partir de la estructura poblacional observada en ese año.

Construcción del modelo de proyección

Las tablas de mortalidad procesadas (L_f , L_m) contienen valores de L_x para edades 0 a 110 por año.

Las tasas específicas de fecundidad contienen valores para edades fértiles 12 a 55.

Se utilizó una variante del modelo Leslie, adaptada para que:

- las **tasas de supervivencia** provengan directamente de L_x observados por año,
- las **tasas específicas de fecundidad** provengan de las ASFR reales de la HFD,
- la **razón de masculinidad al nacer** se mantenga fija en:

$$SRB = 1.05$$

La función `p_pop_t()` realiza la proyección año a año mediante:

1. Construcción de matrices Leslie anuales basadas únicamente en mortalidad.
2. Aplicación de las tasas de fecundidad correctas solo a las edades fértiles (filas 13–56 del vector poblacional).
3. Distribución de nacimientos entre mujeres y varones según el SRB .
4. Almacenamiento de la población femenina y masculina proyectada para cada año.

La proyección total se extendió por:

$\text{iter} = 130 \text{ años}$

lo cual permite llegar hasta el año 2024, último año disponible en los datos descargados.

Generación de pirámides poblacionales

Se generaron dos pirámides para el año final (2024):

- una **observada**, a partir del archivo `last_pop.txt`,
- una **proyectada**, a partir de las últimas dos columnas de `sw_pop`, correspondientes a F_{2024} y M_{2024} .

Ambas estructuras se transformaron al formato largo mediante `reshape_long()`, y se graficaron con `pd_plot()`.

Resultados obtenidos

Estructura observada vs proyectada

La comparación entre las pirámides muestra lo siguiente:

Edades jóvenes (0–15)

- La base de la pirámide **proyectada** es más regular y menos ondulada que la observada.
- La estructura observada exhibe cohortes más irregulares, asociadas a fluctuaciones recientes en migración y fecundidad.
- Estas diferencias surgen porque el modelo **no incluye migración futura**, y por lo tanto los nacimientos se generan únicamente a partir de las mujeres presentes en la población inicial.

Edades centrales (20–60)

- Aquí la coincidencia entre ambas pirámides es notable.
- La forma, la distribución por sexo y los tamaños relativos son muy similares.
- Esto sugiere que las **tasas específicas de fecundidad y mortalidad del período** permiten reproducir de forma adecuada el “cuerpo” de la pirámide estadounidense.

Edades avanzadas (65+)

- La pirámide observada presenta mayor supervivencia a edades muy altas, particularmente entre mujeres.
- La pirámide proyectada muestra una reducción más pronunciada en los grupos longevos.
- Esto se explica porque las **tablas de mortalidad por período no incluyen mejoras futuras en la supervivencia**, y por ello la proyección tiende a subestimar la longevidad real.

Conclusión general

El modelo reproduce adecuadamente la forma general de la estructura poblacional de Estados Unidos en 2024, especialmente en edades centrales.

Las discrepancias en edades jóvenes y avanzadas se explican por:

- ausencia de migración en la proyección,
- ausencia de mejoras en mortalidad,
- suavización de cohortes debido a las propiedades del modelo Leslie.