**Instituto Tecnológico Autónomo de México**

Estadística Aplicada I

Proyecto: “Análisis exploratorio de datos”

Integrantes:

Andrés Albores 131749

Daniela Ordoñez 136977

Karina Ibarra 136608

**Reporte Técnico**

Introducción

En este documento se presenta un análisis de tendencias de casos de violencia intrafamiliar, durante el periodo de 2003 a 2014 en toda la República Mexicana, basándonos en los datos obtenidos por la Secretaría de Salud.

Se presenta en una primera parte el problema que se planteó para después ser resuelto mediante el análisis; posteriormente se describe cómo es que se han obtenido los datos y con qué herramientas fueron usadas. Asimismo, se realiza una breve descripción de los datos y con base a esta se podrá definir el alcance y límite de los datos. Después se realizan pruebas de calidad sobre los datos transformados para que se logre realizar un análisis exploratorio de los datos para que de tal forma se pueda resolver el problema que se planteó en un inicio y, finalmente, se presentan las conclusiones y comentarios del trabajo realizado con la finalidad de que el Gobierno pueda tomar medidas de mejora.

Definición del problema

La violencia intrafamiliar es un tipo de abuso que puede incluir violencia física, violencia sexual, amenazas de violencia o abuso emocional. Su origen es debatido por muchos aunque se cree que las principales causas son falta de recursos económicos, el abuso del alcohol y la drogadicción, asociados comúnmente con la pobreza.

El problema en el cual nos centraremos es obtener las épocas en que se debe prestar mayor atención al fenómeno, cuál es la tendencia del fenómeno, cuáles son las instituciones que atienden un mayor número de casos, qué entidades, qué grupos de edad y qué sexo son más vulnerables al fenómeno para mejorar el apoyo que se brinda a las víctimas y las acciones preventivas que se toman en cada estado.

Obtención y descripción de los datos

Los datos en los que nos basamos para nuestra investigación se obtuvieron de la Dirección General de Epidemiología de la Secretaría de Salud durante el periodo de 2003 a 2014.[[1]](#footnote-1)

Los archivos se descargaron manualmente y se encontraba en formato .pdf. Se dicidió transformarlos, con el objetivo de facilitar la manipulación de los mismos, a formato .csv mediante el programa Tabula[[2]](#footnote-2), mientras que el análisis de datos se realizó en “RStudio”.

Después de la obtención y transformación de datos se observó que la información estaba dividida en

* Casos reportados por fuentes: muestra el número total de víctimas por entidad federativa, reportado por las principales instituciones de salud.
* Casos por grupo de Edad: Presenta el número total de víctimas por entidad federativa, reportado para los grupos de edad establecidos.
* Casos de ocurrencia por mes: Presenta el número total de víctimas, por entidad federativa, reportado por mes.

Asimismo, para cada uno de estos tres grandes grupos la información se divide en tres categorías según género: Femenino, Masculino y ambos.

Los archivos se clasificaron en tres carpetas que nombramos como Fuente (Fue), Grupo (Gru) y Mes (Mes). El nombre que se les asignó fue de tal manera que su identificación fuese sencilla, por ejemplo “FueGen2009” representa al archivo en la carpeta fuente, genero general y año 2009.

La información recopilada determina que se realice un análisis de niveles y que se excluya el análisis de tasas. No se podrán eliminar ciertos sesgos ya que no se hay información disponible que ayude en la corrección de estos.

Calidad de los datos

Utilizando la función “str” en RStudio pudimos conocer la estructura de los archivos consolidados y las variables de cada uno, con base a los extractos obtenidos fue fácil ver que la información carecía de calidad para su manejo y análisis. Además, la información podría generar sesgos ya que para no se reporta el número total de instituciones existentes por entidad en cada año, creando un sesgo respecto al número de casos presentados.

A continuación se presentan los principales puntos (errores) que las tablas tenían y que debían ser corregidos para lograr un buen análisis.

* Columnas desfasadas
* Nombres y orden de columnas distinto
* Múltiples valores faltantes
* Valores faltantes diferentes (N.A. y S/R)
* Títulos que ocupaban filas enteras

Para la prueba de cifras de control se convirtieron ciertos datos a número y se introdujo la columna “TOTAL” que suma las columnas de cada fila, notamos que algunos de los valores totales en las tablas de género General no son iguales a la suma de los totales de las tablas de géneros Masculino y Femenino.

Transformación de los datos

Primero se realizó la lectura de los 108 archivos para los años incluidos en el periodo 2003-2014. Existen 3 tipos de bases de datos: el primer tipo es casos de violencia intrafamiliar reportada por institución y por año, el segundo tipo es casos de violencia intrafamiliar reportados por mes y por año, y, por último, los casos de violencia intrafamiliar reportados en diferentes grupos de edad por año. Cada tipo se encuentra subdivida por casos reportados por género (femenino, masculino y general).

Debido al mal estado de los datos en formato CSV tuvimos que hacer una limpieza manual, con ayuda de Excel y “GVIM”[[3]](#footnote-3), este últimos fue de ayuda para encontrar caracteres ocultos que no permitían una lectura correcta de datos.

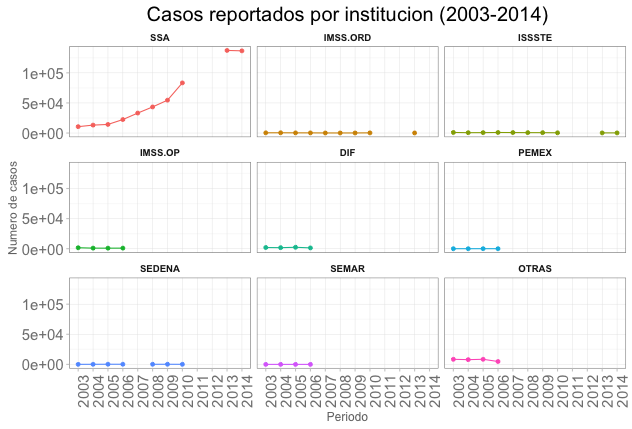
Eliminamos columnas vacías, las columnas de tasa se cambiaron por columnas de incidencia y, en caso de no existir, se crearon con valores faltantes. Además, eliminamos las filas que solo tenían un título.

Después de la limpieza manual utilizamos R para pasar los nombres de las columnas a mayúsculas, reducirlos a una sola palabra y eliminar los espacios sobrantes. Los valores faltantes fueron sustituidos por “NA”.

Siguiendo los pasos para la organización de los datos se juntaron las tres bases de datos en tres data frames los cuales nombramos “violencia.intrafamiliar.fuente”, “violencia.intrafamiliar.edad” y “violencia.intrafamiliar.mes”. Por último, se introdujeron dos nuevas variables “SEXO” y “PERIODO” para poder clasificar los datos y realizar el análisis.

Análisis Exploratorio

Utilizando la tabla, se elabora la siguiente gráfica, mostrando en ella la tendencia del número de incidentes reportados por institución, durante el periodo de observación.

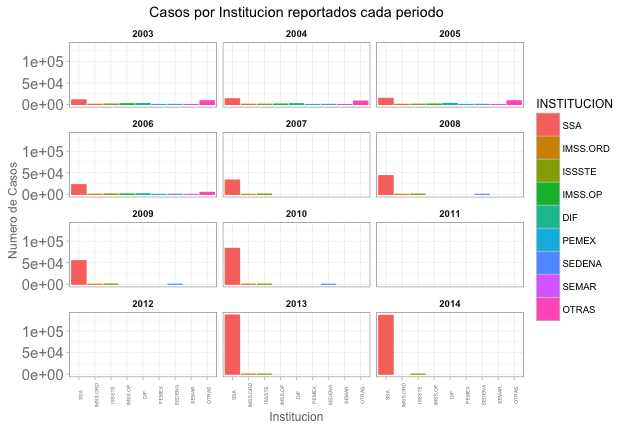


En esta gráfica podemos ver los casos que fueron reportados por institución en cada periodo. Vemos que la Secretaría de Salud empezó en el 2003 con cierto número de casos que con el paso de los años fue aumentando notablemente.

La mayoría del resto de las instituciones mantuvo el número de casos reportados en el 2003, aunque con algunos pequeños cambios.

Utilizando la tabla anterior se logra realizar la siguiente grafica que muestra el número total de reportes por institución que genera un mayor número de reportes en cada uno de 2003 a 2014

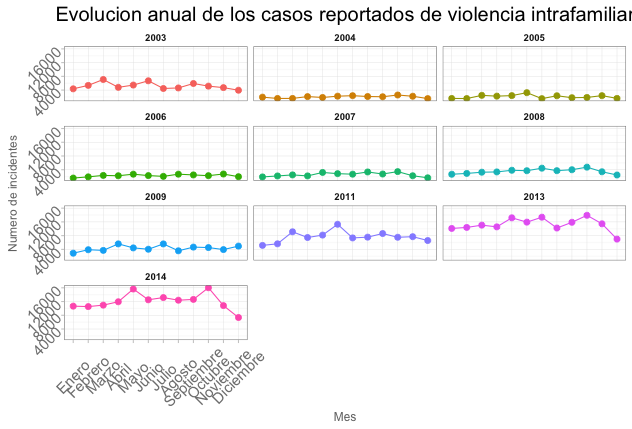
En esta gráfica podemos ver claramente que en cada periodo, la institución que tuvo mayores casos de violencia intrafamiliar reportados fue la Secretaría de Salud. Sin embargo, notamos que en los años 2003, 2004 y 2005 la cantidad de casos reportados a la Secretaría de Salud es similar a la suma de los casos reportados en “otras” instituciones.



Con la finalidad de identificar la época del año que se registra un mayor número de incidentes de violencia doméstica e identificar el comportamiento interanual se creó la siguiente gráfica, que muestra el número total de incidentes de violencia domestica por cada año dentro del periodo de observación.

Encontramos que en los meses de junio, septiembre y octubre se presenta el mayor número de incidentes de violencia doméstica.

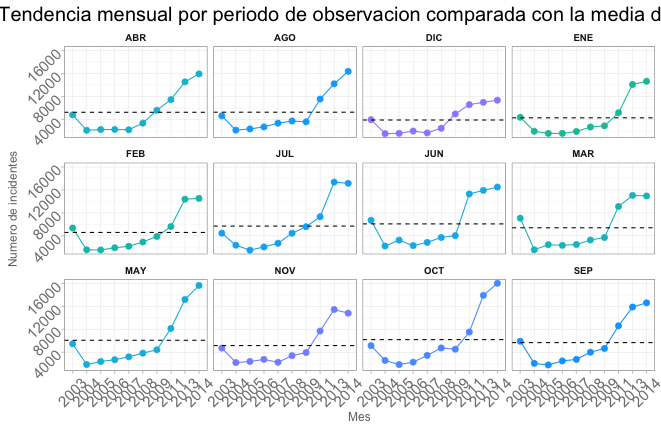
Después, gráficamente se presenta la tendencia inter anual, para cada uno de los años de observación de incidentes de violencia intrafamiliar.



Se observa que el año 2014 tuvo un número elevado de incidentes, en particular durante los meses de mayo y octubre. Otro año que presentó un número notable de incidentes fue el 2013, en particular durante los meses de mayo, julio y octubre. Los años durante los cuales no se presentaron grandes cambios son 2004, 2005, 2006, 2007, 2008 y 2009.

Para ver la tendencia mensual vs su media mensual se rea la siguiente tabla generando el numero promedio de incidentes por mes. Con esta información se obtuvo la siguiente gráfica la cual presenta la tendencia mensual de número total de incidentes a través de los años de observación, comparada con la media para el periodo identificada para cada mes.

Tendencia mensual por periodo de observación comparada con la media

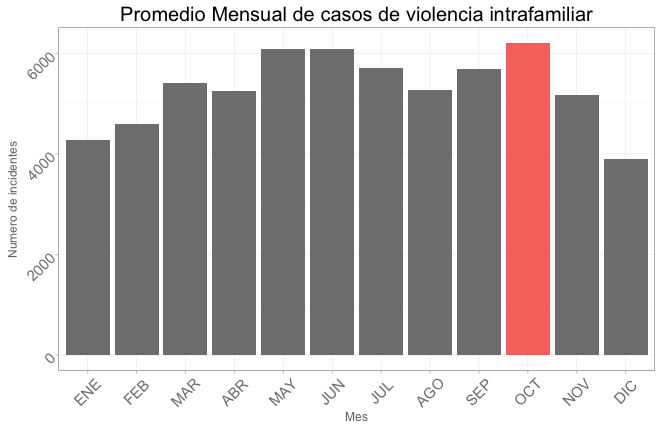


En esta gráfica podemos ver que existe una tendencia a la alza durante el periodo de observación.

A los largo de los años, la diferencia entre el número de casos registrados para hombres y para mujeres es poca, sin embargo, se registra que existe un mayor número de casos de violencia intrafamiliar hacia mujeres. También notamos que el grupo de edad que se encuentra más propenso a sufrir violencia intrafamiliar es el de los 25 a los 44 años.

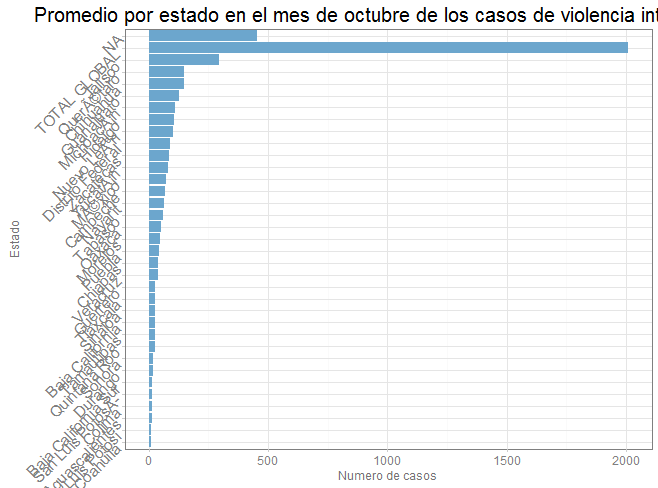
Por un lado, los grupos de edades que presentan un mayor número de casos de violencia intrafamiliar hacia hombres son aquellos menores a los 19 años, pero por otro lado, se registra un mayor número de casos de violencia intrafamiliar hacia mujeres en los grupos de edad que van de los 19 a los 60 años.

Utilizando la tabla generada que contiene el promedio de incidentes mostramos los promedios mensuales de casos registrados.



Por último, para observar el fenómeno de violencia intrafamiliar en el mes de octubre se generó la siguiente tabla, misma que contiene el número promedio de violencia intrafamiliar por entidad durante el periodo de observación.

Promedio por Estado en el mes de octubre de los casos de violencia intrafamiliar



Conclusiones

Desafortunadamente, los resultados de los análisis no son completamente confiables, debido al mal estado de las bases de datos.

Es imperativo que el gobierno mejore sus métodos de recopilación de datos. Con la finalidad de eliminar sesgos en el análisis se sugiere además, incorporar el número de instituciones de salud existentes en cada año, así como la población por entidad para cada uno de los años.

Del análisis realizado se concluye que las gráficas sugieren que el número de reportes de violencia intrafamiliar ha aumentado durante los años estudiados, pero se mantiene constante en cada año. En promedio, el mayor número de reportes ocurre durante el mes de octubre en cada año y se realizan en la Secretaria de Salud.

Como sugerencia para un análisis futuro se sugiera una homogenización de la información contenida en los archivos, pues como hicimos notar anteriormente existen datos que se reportan solo para ciertos grupos o años, así como definir los grupos de edad de manera uniforme, es decir, involucrando cada uno de estos la misma cantidad de años.

Fuentes

* http://www.violenciaenlafamilia.conapo.gob.mx/en/Violencia\_Familiar/Violencia\_en\_la\_Familia
* <http://www.epidemiologia.salud.gob.mx/anuario/html/anuarios.html>
* <http://tabula.technology/>

1. *http://www.epidemiologia.salud.gob.mx/anuario/html/anuarios.html* [↑](#footnote-ref-1)
2. *http://tabula.technology/* [↑](#footnote-ref-2)
3. *http://www.vim.org/download.php* [↑](#footnote-ref-3)