Proyecto Estadistica Aplicada I

Sebastian Morales Peralta 000124317

27 de marzo de 2015

### REPORTE TECNICO

#### I. Introduccion

Este reporte busca presentar, de forma general, el fenomoeno de la Diabetes melitus tipo II (no insulinodependiente) en Mexico. Los datos contenidos abarcan el periodo 2003 a 2013 y son publicados por la Secretaria de Salud en la pagina: <http://www.epidemiologia.salud.gob.mx/anuario/html/anuarios.html>

En esta seccion se encuentra una breve descripcion de los datos, para pasar, mas adelate, a la definicion del problema.

La base de datos esta compuesta por 3 tablas principales, presentadas a continuacion, describen los casos de diabetes reportados de 2003 a 2013. La informacion esta distribuida en dos generos sexuales (masculino y femenino) en las 32 entidades del pais.

FUENTE.  
Informacion dividida por las distintas instituciones que reportaron los casos de diabetes.

## estado genero fecha fuente casos  
## 1 Aguscalientes General 2003 salud 1859  
## 2 Baja California General 2003 salud 2443  
## 3 Baja California Sur General 2003 salud 646  
## 4 Campeche General 2003 salud 1188  
## 5 Coahuila General 2003 salud 1346  
## 6 Colima General 2003 salud 802

## [1] "salud" "imss.ord" "issste" "imss.op" "dif" "pemex"   
## [7] "sedena" "semar" "otras"

MES.  
La informacion se encuantra dividida por cada mes del anio.

## estado genero fecha mes casos  
## 1 Aguscalientes General 2003 ene 268  
## 2 Baja California General 2003 ene 1049  
## 3 Baja California Sur General 2003 ene 171  
## 4 Campeche General 2003 ene 206  
## 5 Coahuila General 2003 ene 1299  
## 6 Colima General 2003 ene 174

EDAD.  
Presenta los casos de diabetes reportados en los distintos grupos de edad, donde ign se refiere a aquellos individuos que no fueron clasifiados en un grupo de edad.

## estado genero fecha edad casos  
## 1 Aguscalientes General 2003 <1 0  
## 2 Baja California General 2003 <1 0  
## 3 Baja California Sur General 2003 <1 0  
## 4 Campeche General 2003 <1 0  
## 5 Coahuila General 2003 <1 2  
## 6 Colima General 2003 <1 0

## [1] "<1" "1a4" "5a9" "10a14" "15a19" "20a24" "25a44" "45a49"  
## [9] "50a59" "60a64" "65+" "ign"

nota: se omitieron la mayoria de los acentos en el presente reporte, debido a que eran cambiados por "??" automaticamente por el sistema.

#### II. Definicion del Problema

El principal objetivo de este reporte es presentar el fenomeno de la diabetes melitus tipo II en Mexico, por lo cual iniciaremos haciendo las preguntas mas generales, hasta llegar a responder las preguntas mas especificas que nuestros datos nos permitan.

1. Es importante conocer como se distribuyen todos nuestros casos reportados de diabetes entre hombres y mujeres de la poblacion.
2. Cual es el estado de la republica que presenta mayores casos de diabetes en Mexico; asi como aquel estado que presenta los menores casos.
3. Presentar la tendencia, a lo largo del tiempo, de los casos reportados en el periodo 2003 - 2013, ??es creciente o decreciente?
4. ??En que a??o se reporto el mayor y menor numero de casos de diabetes en Mexico?
5. ??Cual es el comportamiento, de los casos reportados, en las distintas edades de la poblacion? ??En que grupo de edad se encuentra el mayor numero de casos de diabetes? 6.??Cuales son las instituciones en la republica donde se han reportado el mayor y menor numero de casos?
6. ??Como es el comportamiento mensual de los casos reportados? ??Existe algun tipo de estacionalidad?
7. Cual es el comportamiento de los casos en hombre y mujeres a lo largo del tiempo y en cada uno de los estados de la republica?
8. Cual es el comportamiento de los distintos grupos de edad a lo largo del tiempo y en cada una de las entidades del pais?

Cada una de estas preguntas ser?? respondida mediante un analisis exploratorio, el cual nos permita entender el comportamiento de los casos de diabetes en la poblacion mexicana a lo largo del tiempo, por lo que se incluiran datos de promedios y varianzas para comenzar a "visualizar" que tan solido sera el analisis en la etapa inferencial (no presentada en el actual reporte).

#### III. Obtencion y Descripcion de los Datos

Una base "ideal" que podria responder todas las preguntas de manera puntual seria aquella que contuviera los siguientes campos o variables:

## [1] "estado" "genero" "fecha" "mes" "fuente" "edad" "casos"

En la cual, las variables serian de los siguientes tipos y escalas:  
ESTADO.  
Esta variable es de tipo "caracter" y da cuenta de cada uno de los estados de la republica mexicana.

## [1] "Aguscalientes" "Baja California" "Baja California Sur"  
## [4] "Campeche" "Coahuila" "Colima"   
## [7] "Chiapas" "Chihuahua" "Distrito Federal"   
## [10] "Durango" "Guanajuato" "Guerrero"   
## [13] "Hidalgo" "Jalisco" "Mexico"   
## [16] "Michoacan" "Morelos" "Nayarit"   
## [19] "Nuevo Leon" "Oaxaca" "Puebla"   
## [22] "Queretaro" "Quintana Roo" "San Luis Potosi"   
## [25] "Sinaloa" "Sonora" "Tabasco"   
## [28] "Tamaulipas" "Tlaxcala" "Veracruz"   
## [31] "Yucatan" "Zacatecas"

GENERO.  
El sexo de de los individuos reportados con diabetes melitus tipo II. Es una variable de tipo "caracter".

## [1] "General" "Mujeres" "Hombres"

FECHA.  
Se refiere al anio en el que se presentaron los casos de diabetes, es una variable numerica, de tipo entero que se encuantra entre los rangos 2003 y 2013.

## [1] 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013

MES.  
El mes del anio en el que se reportaron los casos, es una variable de tipo caracter que contiene cada uno de los meses del anio.

## [1] "ene" "feb" "mar" "abr" "may" "jun" "jul" "ago" "sep" "oct" "nov"  
## [12] "dic"

FUENTE.  
Es una variable categorica que permite saber el numero de casos reportados algunas instituciones del pais.

## [1] "salud" "imss.ord" "issste" "imss.op" "dif" "pemex"   
## [7] "sedena" "semar" "otras"

EDAD.  
Variable categorica que agrupa distintos rangos de edades que presentan casos de diabetes. En este caso, la categoria "ign" se refiere a todas aquellas personas que no fueron asignadas a algun grupo de edad.

## [1] "<1" "1a4" "5a9" "10a14" "15a19" "20a24" "25a44" "45a49"  
## [9] "50a59" "60a64" "65+" "ign"

CASOS.  
Es la variable dependiente, es de tipo numerica y se refiere al numero de personas que reflejan un caso de diabetes melitus tipo II (no insulinodependiente). Para este reporte, la variable se encuentra en los siguentes rangos:

## min max  
## casos 0 28024

Este reporte contiene todas estas variables, sin embargo, no se tienen integradas en una sola base de datos, por lo que no es posible dar respuestas tan especificas como por ejemplo:  
Cuantos casos de hombres con diabetes se reportaron en abril de 2008 en el issste en Durango?  
Ya que la informaci??n actual puede dar cuenta de los hombres reportados en abril de 2008 y los hombres reportados en el 2008 por parte del issste pero a lo largo de todo el anio.  
ej.

## estado genero fecha mes casos  
## 3690 Durango Hombres 2008 abr 269

## estado genero fecha fuente casos  
## 2986 Durango Hombres 2008 issste 566

Por lo tanto, utilizaremos las tres bases por separado (fuente, mes, edad) para responder de las preguntas mas generales hasta la mas particular que sea posible.

#### IV Calidad de los Datos

En esta seccion se mostrara la calidad de los datos presentados, es decir, que los datos de cada una de nuestras tablas (fuente, mes, edad) sean los mismos para cada anio, genero, y estado. Esto con el fin de corrobrar que los datos que estamos utilizando coinciden independientemente de la tabla que estemos utilizando, lo cual nos permite tener solidez en el analisis comparativo.

Primero mostraremos que, para cada tabla, los datos son los mismos para cada anio y cada genero.  
Por fuente.

## Fecha General Hombres Mujeres Total Diferencia  
## 1 2003 380332 147171 233161 380332 0  
## 2 2004 404562 156916 247646 404562 0  
## 3 2005 397387 152621 244766 397387 0  
## 4 2006 394360 152836 241524 394360 0  
## 5 2007 404770 160068 244702 404770 0  
## 6 2008 396374 158992 237382 396374 0  
## 7 2009 426802 177825 248977 426802 0  
## 8 2010 420032 173507 246525 420032 0  
## 9 2011 427125 176891 250234 427125 0  
## 10 2012 419493 173725 245768 419493 0  
## 11 2013 397192 166220 230972 397192 0

Por mes.

## Fecha General Hombres Mujeres Total Diferencia  
## 1 2003 380332 147171 233161 380332 0  
## 2 2004 404562 156916 247646 404562 0  
## 3 2005 397387 152621 244766 397387 0  
## 4 2006 394360 152836 241524 394360 0  
## 5 2007 404770 160068 244702 404770 0  
## 6 2008 396374 158992 237382 396374 0  
## 7 2009 426802 177825 248977 426802 0  
## 8 2010 420032 173507 246525 420032 0  
## 9 2011 427125 176891 250234 427125 0  
## 10 2012 419493 173725 245768 419493 0  
## 11 2013 397192 166220 230972 397192 0

Por edad.

## Fecha General Hombres Mujeres Total Diferencia  
## 1 2003 380332 147171 233161 380332 0  
## 2 2004 404562 156916 247646 404562 0  
## 3 2005 397387 152621 244766 397387 0  
## 4 2006 394360 152836 241524 394360 0  
## 5 2007 404770 160068 244702 404770 0  
## 6 2008 396374 158992 237382 396374 0  
## 7 2009 426802 177825 248977 426802 0  
## 8 2010 420032 173507 246525 420032 0  
## 9 2011 427125 176891 250234 427125 0  
## 10 2012 419493 173725 245768 419493 0  
## 11 2013 397192 166220 230972 397192 0

Nota: el Total = Hombres + Mujeres Nota: la Diferencia = Total - General

En total, para cada genero y en cada anio, podemos observar que los datos son los mismos para cada una de las tablas (fuente, mes y edad). Y en un primer analisis nos estaria diciendo el total de hombres y mujeres reportados con un caso de diabetes melitus tipo II, asi como el total de personas reportadas en el periodo.

## Hombres Mujeres Total   
## 1796772 2671657 4468429

## Hombres Mujeres Total   
## 1796772 2671657 4468429

## Hombres Mujeres Total   
## 1796772 2671657 4468429

#### V. Transformacion de los Datos

Los datos que tenemos estan listos para comenzar el analisis exploratorio, simplemente extraeremos las variables de lo general a lo particular, dependiendo de la o las preguntas que se deban contestar.  
ej. Casos de hombres reportados por el DIF.

hombresDif <- subset(fuente, genero == "Hombres" & fuente == "dif")$casos  
sum(hombresDif)

## [1] 27081

Sin embargo, en esta seccion se explica el proceso que se llevo a cabo para obtener los datos listos para utilizarse.

##### V.I. Extraccion de los Datos

Primero, se llevo a cabo la extraccion de los datos desde la pagina de la Secretaria de Salud: <http://www.epidemiologia.salud.gob.mx/anuario/html/anuarios.html> Fueron descargados 99 archivos PDF que contenian la informacion de los casos de diabetes reportados para hombres, mujeres y a nivel general para los a??os 2003 a 2013, dependiendo de la categoria seleccionada: Grupo de Edad, Tipo de Fuente y Mes.

##### V.II. Cambio de Formato (PDF a CSV)

Para poder hacer uso de los datos extraidos, el formato de los archivos debe ser de texto plano, es decir, .txt o .csv. Para hacer el cambio de formato de PDF a CSV utilizo el programa TABULA (<http://www.tabula.tecnology>). De los archivos PDF fueron transformados 99 archivos CSV, a cada archivo le fue asignado un nombre especificando el genero y el a??o de los datos contenidos y despues guardados en carpetas dependiendo de la categoria de los datos(Tipo de Fuente, Grupo de Edad o Mes). ej. "PROYECTO/DATOS/BRUTOS/CSV/MES/G-2009" Se refiere al reporte general de 2009 por mes. "PROYECTO/DATOS/BRUTOS/CSV/FUENTE/M-2003" se refiere al reporte de mujeres en 2003 por tipo de fuente.

Esto con el fin de homologar los nombres y generar una secuencia para leer, procesar y compilar los datos mediante codigo R.

##### V.III. Limpieza y Compilacion de Datos

Mediante codigo R se realiza la lectura, limpieza y compilacion de los datos para, finalmente, generar tres nuevos archivos con la informacion lista para realizar los respectivos analisis.

La lectura consiste en cargar los 33 archivos csv de cada una de las categorias para hombres, mujeres y general durante el periodo 2003 a 2013.

La limpieza de los datos consiste en la homologacion de titulos, orden de las columnas, simplificar los nombres de las variables categoricas, como por ejemplo: "San Luis Potos??" en "San Luis Potosi". Tambien se quitan los espacios entre el numero de casos para que la variable pueda ser tomada en cuenta como numerica, ejemplo: "2 45" a 245.

La compilacion de los datos es simplemente integrar los 33 archivos de cada categoria en una sola tabla.

Para mayor informacion al respecto se puede revisar el codigo en los siguientes archivos:  
PROYECTO/CODIGO/PRELIMINAR/Limpieza.R  
PROYECTO/CODIGO/PRELIMINAR/Limpieza.RData

#### VI. Analisis Exploratorio

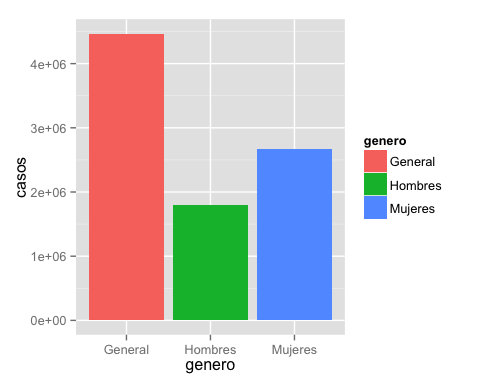
##### Analisis General

En genera, se ha registrado una mayor cantidad de diabetes melitus tipo II en mujeres (periodo 2003-2013)

## Hombres Mujeres Total  
## Casos 1796772 2671657 4468429

El porcentaje en hombres y en mujeres es el siguiente:

## Hombres Mujeres  
## % 40.21037 59.78963



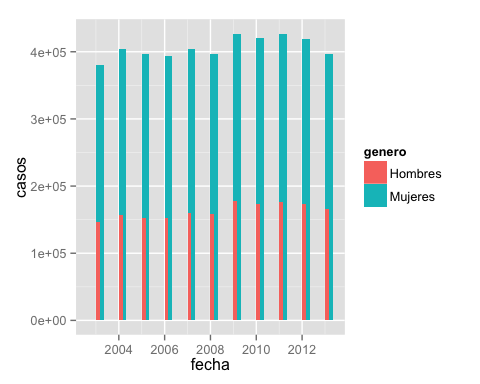
En el anio 2003 se registro el menor numero de casos de diabetes en la poblacion general, mientras que 2001 fue el a??o con mas casos. En hombres, el menor numero se registro en 2003 y el mayor en 2009. En mujeres, el menor numero se registro en 2013 y el mayor en 2011.

## Hombres Mujeres General  
## min 2003 2013 2003  
## max 2009 2011 2011

El promedio de casos registrados por anio en hombres, mujeres y a nivel general es el siguente:

## General Mujeres Hombres  
## 1 406220.8 242877.9 163342.9

La tendencia a lo largo del tiempo en hombres y mujeres:



Los estados que reportan el mayor y menor numero de casos son los siguientes:

## Hombres Mujeres General  
## min Colima Colima Colima  
## max Mexico Mexico Mexico

El porcentaje de casos reportados por estado de la republica en el periodo 2003-2013:

## Estado General  
## 15 Mexico 12.5940235  
## 9 Distrito Federal 8.7529644  
## 14 Jalisco 7.9608739  
## 30 Veracruz 5.3658903  
## 21 Puebla 4.2377757  
## 11 Guanajuato 4.0979951  
## 8 Chihuahua 3.9198340  
## 19 Nuevo Leon 3.8524502  
## 5 Coahuila 3.5855331  
## 25 Sinaloa 3.5653470  
## 2 Baja California 3.4163237  
## 28 Tamaulipas 3.4089162  
## 12 Guerrero 3.3625241  
## 7 Chiapas 3.0118863  
## 16 Michoacan 2.9432492  
## 20 Oaxaca 2.8333000  
## 24 San Luis Potosi 2.7389044  
## 17 Morelos 2.6486266  
## 26 Sonora 2.3040984  
## 10 Durango 2.1687488  
## 31 Yucatan 1.9900730  
## 27 Tabasco 1.9218387  
## 13 Hidalgo 1.7534798  
## 1 Aguscalientes 1.2357811  
## 22 Queretaro 1.1997953  
## 32 Zacatecas 1.0701524  
## 18 Nayarit 0.9651490  
## 29 Tlaxcala 0.8764825  
## 23 Quintana Roo 0.6330189  
## 4 Campeche 0.5819271  
## 3 Baja California Sur 0.5290898  
## 6 Colima 0.4739473

