Reporte de Base de Datos

Primer reporte

Grupo : 4

Integrantes : -Abraham Calderón Canicoba(20203599)
-Jean Pierre Arroyo Gonzales(20211199)
-Jimena Solange López Penadillo(20211552)
-Geraldine Namie Pajuelo Sanchez(20193790)

Facultad : Ciencias Sociales

Profesor : José Mendoza

Año de la encuesta : 2022

1. EJERCICIO 1

Diferencias entre la ENIGH de México y ENAHO de Perú

La Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) de México y la Encuesta Nacional de Hogares (ENAHO) del Perù son encuestas dedicadas a recopilar información acerca de los ingresos, egresos y la condiciòn de vida de los hogares de cada país anteriormente mencionado, por lo que sirven de herramientas importantes para la formulaciòn de políticas basadas en la necesidad de la población, la toma de decisiones económicas y la investigación académica. Así también, ambas encuestas han experimentado cambios y ajustes en su metodología a lo largo de los años. A manera de introducción, la ENAHO fue creada en 1993 por el Instituto Nacional de Estadística e Informática de Perú y la ENIGH, fue creada en 1984 por el Instituto Nacional de Estadìstica y Geografía de México. Asimismo, una de las principales diferencias entre estas dos encuestas es que la ENAHO evalúa muy aparte de los aspectos económicos de los hogares también tratados por la ENIGH, temas basados en el nivel de bienestar y condición de vida de los habitantes, tales como: salud, educación, entre otros. Es por ello que se puede deducir que también existe una distinción de las preguntas y bases de datos a utilizar de acuerdo al enfoque proporcionado por cada una de ellas. De este modo, a partir de estos temas evaluados, existen distintos temas específicos a tratar para la realización de posteriores estudios o trabajos de investigación. Una muestra de ello y dada la diversidad de temas que abarca la ENAHO, se puede especificar que esta encuesta es usada en estudios sobre el bienestar social y el nivel de desempleo, pobreza, salud y educación en el Perú. En el caso de la ENIGH, es relevante para él anàlisis económico, como también la planificación fiscal y la evaluación de las políticas económicas en México. Por otra parte, con respecto a cada cuánto tiempo se recopilan los datos a tratar, este varía en ambas encuestas dado que en la ENIGH se produce cada dos años, mientras que en la ENAHO, se lleva a cabo de manera trimestral y anual. Bajo esta misma línea, es importante aclarar que cada encuesta posee un ente que regula el ingreso a estas bases de datos y además es el encargado de publicar documentos e informes que resumen el contenido de estas dos encuestas. Este ente, para el caso de Perú, es el Instituto Nacional de Estadìstica e Informática (INEI) y para México, se trata sobre el Instituto Nacional de Estadìstica y Geografía (INEGI). Otro detalle a evaluar es el tamaño de la población de ambos países, ya que México tiene una población mucho más grande que Perú, lo que hace que la ENIGH tenga que lidiar con una muestra más grande y diversa. Esto, a su vez, puede afectar la logística y la representatividad de la encuesta en comparación con la ENAHO. Por último y con respecto a la disponibilidad de recursos, se establece que los recursos disponibles para llevar a cabo estas encuestas pueden variar entre México y Perú. Esto puede afectar la calidad de la recopilación de datos y la capacidad de realizar análisis más detallados.En conclusión, cada encuesta anteriormente mencionada abarca metodologías específicas de acuerdo a las polìticas que regulan cada país, así también como las necesidades que posee cada nación, incluyendo las áreas urbanas y zonas rurales, para luego ser publicadas en los respectivos sitios web de cada institución estadística que le corresponde.

Subiendo la Base de Datos para trabajar

El presente trabajo, usa las bases de datos recopiladas por la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) del año 2022. Las bases de datos con las que trabajaremos serán específicamente las de **ingreso, hogares y población**.

```
*ingreso
use
"C:\Users\namie\OneDrive\Escritorio\Laboratorio\stata\entrega1\ingresos2022.dta", clear

*Hogares solo hay tarjeta acá
use
"C:\Users\namie\OneDrive\Escritorio\Laboratorio\stata\entrega1\hogares2022.dta", clear

*Población
use
"C:\Users\namie\OneDrive\Escritorio\Laboratorio\stata\entrega1\poblacion2022.dta", clear
```

2. EJERCICIO 2

Emparejando el módulo de ingresos y caraterísticas del hogar

Para emparejar los módulos usamos el comando **merge**, y los indetificadores de vivienda y hogar, los cuáles son: **folioviv** y **foliohog** respectivamente.
CODIGO:

```
* emparejar el módulo de ingreso y de características del hogar
use
"C:\Users\namie\OneDrive\Escritorio\Laboratorio\stata\entrega1\ingresos2022.dta", clear
merge m:1 folioviv foliohog using
"C:\Users\namie\OneDrive\Escritorio\Laboratorio\stata\entrega1\hogares2022.dta"
```

RESULTADO:

```
Result                      Number of obs
-----------------------------------------
Not matched                            58
    from master                         0  (_merge==1)
    from using                         58  (_merge==2)

Matched                           397,182  (_merge==3)
-----------------------------------------
```

Construcción de la variable (Entidad del Norte o Entidad del Sur)

El territorio mexicano está dividido en 32 entidades federativa. La encuesta las identifica según número como se aprecia en la Figura 1.

Figura 1: Las 32 entidades federativas de mexico

#	Entidad Federativa
1	Aguascalientes
2	Baja California
3	Baja California Sur
4	Campeche
5	Coahuila
6	Colima
7	Chiapas
8	Chihuahua

#	Entidad Federativa
9	Distrito Federal
10	Durango
11	Guanajuato
12	Guerrero
13	Hidalgo
14	Jalisco
15	Mexico
16	Michoacan

#	Entidad Federativa
17	Morelos
18	Nayarit
19	Nuevo León
20	Oaxaca
21	Puebla
22	Querétaro
23	Quintana Roo
24	San Luis

#	Entidad Federativa
25	Sinaloa
26	Sonora
27	Tabasco
28	Tamaulipas
29	Tlaxcala
30	Veracruz
31	Yucatan
32	Zacatecas

Para resolver el ejercicio, el equipo dividió a las entidades federativas mexicanas, en 2 bloques: entidades del norte y entidades del sur. Con el objetivo tener casi la misma cantidad de entidad de entidades en ambos bloques se optó por empezar la división desde Nayarit, Zacatecas, Aguascalientes, San Luis de Potosí y Tamaulipas. De esta forma bajo nuestra división tenemos 15 entidades del norte y 17 del sur como se muestra en la Figura 2.

Figura 2: División de las entidades federativas

Para esta división creamos una variable dicotómica, la cual valdrá 1 si la entidad es del norte y 0 si es del sur. De la siguiente forma:

```
* Para crear norte=1 y sur=0
destring entidad, replace

gen Estado = 0

replace Estado = 1 if inlist(entidad, 1, 2, 3, 5, 6, 8, 10, 14, 18, 19,
24, 25, 26, 28, 32)
replace Estado = 0 if inlist(entidad, 4, 7, 9, 11, 12, 13, 15, 16, 17, 20
, 21, 22, 23, 27, 29, 30, 31)
```

Figura 3: División de las entidades federativas

Recodificando la variable de tenencia de tarjeta y la variable negocio en el hogar

Modificamos los valores que puede tomar la variable "tenencia de tarjeta"(*tarjeta*) que eran 2(no tiene) y 1(si tiene) por 0 y 1 respectivamente. Del mismo modo ser hará para la variable "negocio en el hogar"(*negcua*)

```
* Para recodificar la variable de tendencia de tarjeta de modo que 2(no
tiene)=0 y 1(si tiene), se mantenga
destring tarjeta, replace
recode tarjeta (2=0)

label define tarjeta_1 0 "No tiene" 1 "Sí tiene"
label values tarjeta tarjeta_1

* Para recodificar la varible de negocio en el hogar de modo que 2(no
tiene)=0 y 1 (si tiene)
destring negcua, replace
recode negcua (2=0)

label define negcua_1 0 "No tiene" 1 "Sí tiene"
label values negcua negcua_1
```

Hallando estadísticos descriptivos

Hallaremos los estadísticos del ingreso trimestral(*ing tri*) en función al valor de ciertas variables y lo ponderaremos usando el factor de expansión(*factor*)

Estadísticos descriptivos de los ingresos trimestrales de todos los estados

CODIGO:

```
* Hallamos los estadísticos descriptivos de los ingresos trimestrales de
todos los estados
summarize ing_tri [aw=factor]
```

RESULTADO:

Variable	Obs	Weight	Mean	Std. dev.	Min	Max
ing_tri	397,182	162752624	12781.49	32478.07	0	6854754

INTERPRETACIÓN Son un total de 397,182 observaciones, pero los "datos"de la base de datos se observa que son 397,240. Por tanto, es deducible que 58 personas no poseen registros de sus ingresos trimestrales. La media de ingresos de la población mexicana es de 12781.49 y su desviación estándar es de 32478.07, lo cual es considerablemente grande. Por esa razón podemos deducir que existe una gran diferencia de ingresos entre la población mexicana. Más aún, el máximo ingreso es de 6854754.00 mientras que el mínimo es 0.

Estadisticos descriptivos de los ingresos trimestrales de los estados del norte

CODIGO:

```
* Hallamos los estadisticos descriptivos de los ingresos trimestrales de
los estados del norte
summarize ing_tri if Estado == 1 [aw=factor]
```

RESULTADO:

Variable	Obs	Weight	Mean	Std. dev.	Min	Max
ing_tri	203,756	61536724	14421.11	37558.15	.24	6854754

INTERPRETACIÓN Por la observación anterior sabemos que puede existir una gran diferencia en los ingresos de la población mexicana. Con los nuevos datos es posible analizar como varían estos resultados en el norte y el sur.Hay 203 756 observaciones en el norte, por tanto la mayoría de la población encuestada está en el norte del país. La media de ingresos es 14421.11, lo cual es mayor a la media del total de la población.De igual forma la desviación estandar es 37,558.15, mayor a la desviación estandar del total. Además, el ingreso máximo 6854754 y el mínimo 0.24.

Estadisticos descriptivos de los ingresos trimestrales de los estados del sur

CODIGO:

```
* Hallamos los estadísticos descriptivos de los ingresos trimestrales de
los estado del sur
summarize ing_tri if Estado == 0 [aw=factor]
```

RESULTADO:

Variable	Obs	Weight	Mean	Std. dev.	Min	Max
ing_tri	193,426	101215900	11784.64	28911.77	0	4157609

INTERPRETACIÓN Se observa que el número de observaciones es de 193,426. El promedio de ingresos trimestrales de los estado del sur es de 11784.64, lo cual es mucho menor que el promedio del total de estados y de los estados del norte.La desviación estandar es de 28911.77, lo cual es considereblemente menor respecto a la desviación estantar de los estados del norte y del país en general. El ingreso máximo es de 4157609 y el mínimo de 0. A modo de comentario, podemos decir que la variabilidad de los ingresos de México se encuentra en el norte del país, región en la cúal también hay mas concentración de población.

Estadisticos descriptivos de los ingresos trimestrales de las personas sin tarjeta

CODIGO:

```
*Hallamos los estadísticos descriptivos de los ingresos trimestrales de
los hogares que no tienen tarjeta=0
summarize ing_tri if tarjeta == 0 [aw=factor]
```

RESULTADO:

Variable	Obs	Weight	Mean	Std. dev.	Min	Max
ing_tri	266,302	106475546	10331.28	19469.8	0	4402174

INTERPRETACIÓN A continuación analizaremos los estadísticos descriptivos de los ingresos trimestrales de los hogares que no tienen tarjeta. Notamos que hay 266302 observaciones, es decir, existe un gran porcentaje de la población que no tiene tarjetas. Por otro lado, los ingresos trimestrales son en promedio 10331.28, lo cual es considerablemente bajo en comparación con el promedio nacional. De igual forma, la desviación estándar es mucho menor, siendo de 19469.8. Ello puede significar que si bien los ingresos trimestrales, en promedio de este sector de la población (los que no tienen tarjeta), son menores, tienen menor variabilidad. Es decir, este sector de la población tiene ingresores menores pero son más uniformes. Ello tambien se aprecia en que el minimó (0) el máximo valor (4402174)de ingresos trimestrales no tienen una brecha tan grande como las que hemos visto anteriormente.

Estadisticos descriptivos de los ingresos trimestrales de las personas con tarjeta

CODIGO:

```
*Hallamos los estadísticos descriptivos de los ingresos trimestrales de
los hogares que sí tienen tarjeta=1
summarize ing_tri if tarjeta == 1 [aw=factor]
```

RESULTADO:

Variable	Obs	Weight	Mean	Std. dev.	Min	Max
ing_tri	130,880	56277078	17417.25	47963.57	.48	6854754

INTERPRETACIÓN Analaizando los estadísticos descriptivos de los hogares que sí tienen, podemos ver que hay 130,880, es decir, el grupo de la población que "sí tiene"tarjeta es mucho más reducido que el que "no tiene", pues la cifra es mucho menos de la mitad. Ahorabien, el promedio de los ingresos trimestrales de este grupo de la población es de 17417.25, lo cual es resaltante, pues es considerablemente mayor que el promedio de los que "no tienen"tarjeta, e icluso mayor que el promedio nacional. Por otro lado, la desviación estandar es de 47963.57, lo cual es mucho más grande que la desviación estandar de los ingresos trimestrales de los que no tiene tarjeta. Podemos concluir que este secto de población, si bien tiene ingresos trimestrales mayores, en promedio, dichos ingresos poseen una enorme variabilidad. Ello esta representado en el mínimo (.48) y máximo (6854754) ingreso trimestral, la brecha entre ambos es considerablemente mayor que la brecha que observamos en lon ingresos trimestrales de los que no tienen tarjeta.

Estadisticos descriptivos de los ingresos trimestrales de las personas sin negocio

CODIGO:

```
*Hallamos los estadísticos descriptivos de los hogares que no tienen
negocio=0
summarize ing_tri if negcua == 0 [aw=factor]
```

RESULTADO:

Variable	Obs	Weight	Mean	Std. dev.	Min	Max
ing_tri	347,478	143765353	13157.7	32861.15	0	6854754

INTERPRETACIÓN Resulta interesante analizar ahora los estadísticos descriptivos de la población diviendola entre dos grandes sectores, los hogares que no tienen negocio y los hogares que si tienen negocio. Empezando por los hogares que no tienen negocio, vemos que la cifra es realmente alta, y es que hay un total de 347478 observaciones. Por otro lado, el promedio de ingresos trimestrales es de 13157.7, lo cual es mayor que el promedio nacional. Adional a ello, la desviación estandar es de 32861.15, que también es mayor que la desviación estandar nacional. Además, el mínimo y el máximo valor de los ingresos trimestrales es de 0 y 6854754 respectivamente. Podemos concluir que dentro de este sector de la población existe una gran variabilidad de ingresos trimestrales, y es que la brecha entre el mínimo y el máximo valor es enorme. Es decir, entre los hogares que no tienen negocios hay personas ganando mucho y de igual forma personas ganando muy poco.

Estadisticos descriptivos de los ingresos trimestrales de las personas con negocio

CODIGO:

```
*Hallamos los estadísticos descriptivos de los hogares que sí tienen
negocio=1
summarize ing_tri if negcua == 1 [aw=factor]
```

RESULTADO:

Variable	Obs	Weight	Mean	Std. dev.	Min	Max
ing_tri	49,704	18987271	9932.958	29259.9	0	4157609

INTERPRETACIÓN Bien, ahora analizamos el otro sector de la población que si tienen negocios. Vemos que es un grupo muy pequeño, y es que hay solo 49704 observacione. Ahora, vemos que el promedio de los ingresos trimestrales es de 9932.958, lo cual es considerablemente bajo respecto al sector que no tienen negocios. Por otro lado, tenemos que la desviación estandar es de 29259.9, que es menor que la del sector de los hogares que no tienen negocios. Podemos concluir que hay una menor variabilidad en los ingresos trimestrales de este sector de la población, ello lo podemos ver en la brecha de los valores mínimos y máximos de los ingresos trimestrales, que son 0 y 4157609 respectivamente, lo cual es una brecha mucho menor que el otro sector de la población.
Este resultado es muy interesante, y es que podríamos pensar que el promedio de los ingresos trimestrales del sector de la población que tienen negocios es más grande que el sector que no tiene negocios. Sin embargo, vemos que los resultados demuestran lo contrario.

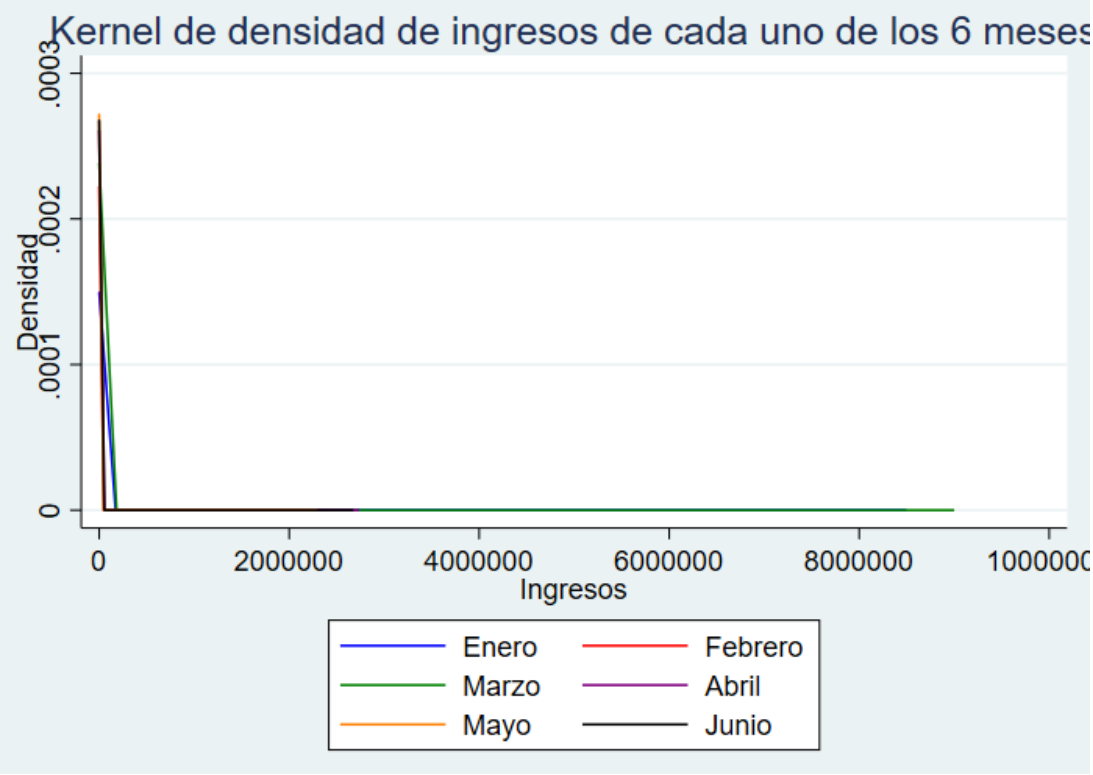
Kernel de densidad

Este código genera un gráfico de densidad de kernel superpuesto para visualizar la distribución de ingresos en seis meses diferentes (Enero a Junio) utilizando líneas de colores distintos para cada mes. Esto permite comparar visualmente las distribuciones de ingresos en diferentes momentos del año.

CODIGO:

```
//KERNEL DE DENSIDAD

kdensity ing_1, generate(x1 y1)
kdensity ing_2, generate(x2 y2)
kdensity ing_3, generate(x3 y3)
kdensity ing_4, generate(x4 y4)
kdensity ing_5, generate(x5 y5)
kdensity ing_6, generate(x6 y6)


twoway (line y1 x1, lcolor(blue)) ///
  (line y2 x2, lcolor(red)) ///
  (line y3 x3, lcolor(green)) ///
  (line y4 x4, lcolor(purple)) ///
  (line y5 x5, lcolor(orange)) ///
  (line y6 x6, lcolor(black)), ///
   legend(order(1 "Enero" 2 "Febrero" 3 "Marzo" 4 "Abril" 5 "Mayo" 6
"Junio")) ///
   title("Kernel de densidad de ingresos de cada uno de los 6 meses")
///
   xtitle("Ingresos") ///
   ytitle("Densidad")
```

RESULTADO:

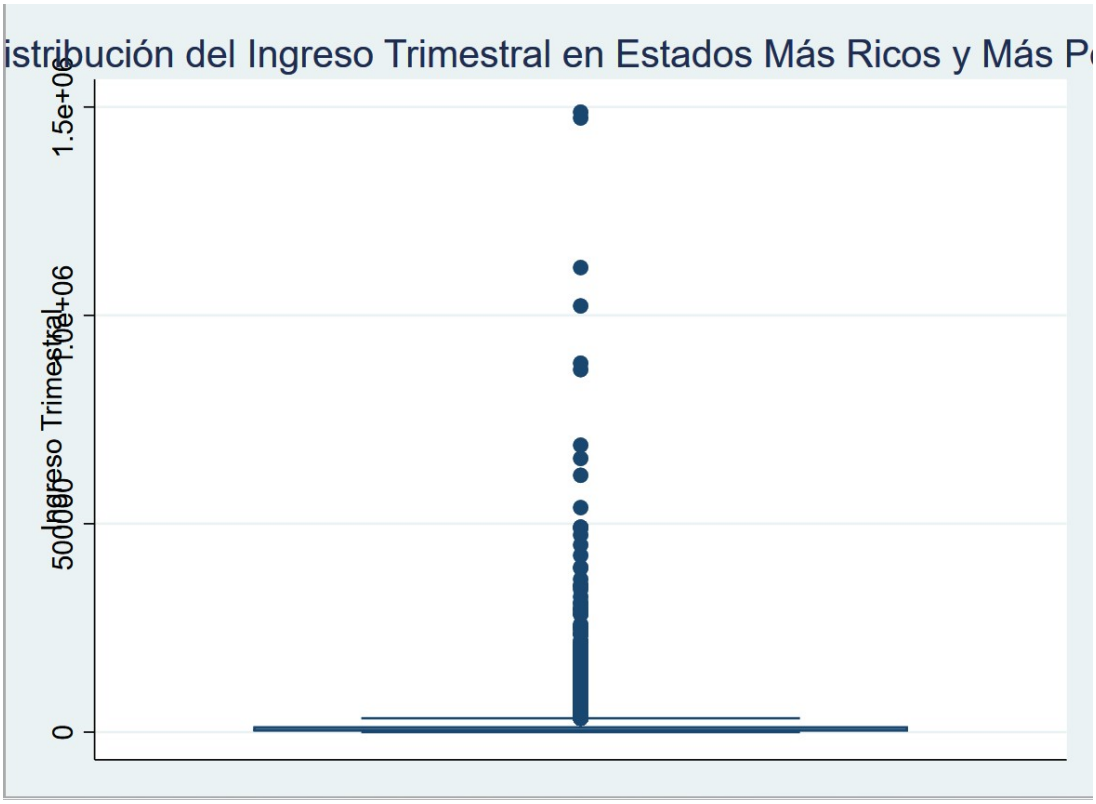


boxplot

Para este parte, realizamos un boxplot comparando los ingresos de los estados más ricos(ciudad de méxico y Nuevo León) y los mas pobres (Oaxaca y Chiapas)
CODIGO:

```
//BOXPLOT DE LOS MAS RICOS
graph box ing_tri if entidad == 15 | entidad == 19 | entidad == 7 |
entidad == 20, ///
    title("Distribución del Ingreso Trimestral en Estados Más Ricos y
Más Pobres") ///
    ytitle("Ingreso Trimestral")
```

RESULTADO:

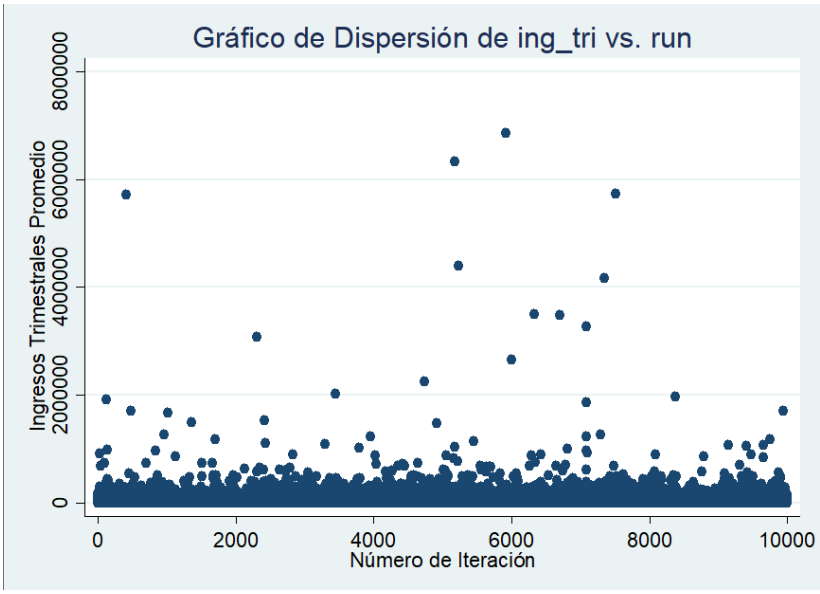


EJERCICIO 3

Para resolver este ejercicio se hizo una simulación, en la cual se creó una variable aleatoria. Se simulará realizar 10 mil observaciones de la variable y se obtendrá el ingreso promedio trimestral considerando su factor de expansión, para posteriormente recuperar el promedio. Se realizará este procedimiento 1000 veces, con ayuda de un *bucle* y se gráficará la variabilidad de los promedios obtenidos, con un boxplot. Para conservar el conjunto de datos usamos el **preserve** y **restore**
CODIGO:

```
* Generar una variable aleatoria
set seed 1
gen variable_aleatoria = runiformint(1, 100)

* Crear una variable para el número de iteración (utilizando egen)
egen run = seq(), from(1) to(10000)

* Realizar el procedimiento 1,000 veces
forval x = 1/1000 {
    preserve
    gen var1 = runiform()
    sort var1
    keep if _n <= 10000  // Mantener las primeras 10,000 observaciones
    collapse (mean) ing_tri = ing_tri
    tempfile b`x'
    save `b`x''.dta, replace
    restore
}

* Limpiar y unir los archivos temporales
clear
forval x = 1/1000 {
    append using `b`x''.dta
}

* Limpiar los archivos temporales
forval x = 1/1000 {
    erase `b`x''.dta
}

* Crear un gráfico de dispersión (scatter plot) de ing_tri vs. run
scatter ing_tri run, title("Gráfico de Dispersión de ing_tri vs. run")
///
xtitle("Número de Iteración") ytitle("Ingresos Trimestrales Promedio")
```

RESULTADO:



3. OPCIONALES

Opcional 1

Opcional 2

Para obtener muestras aleatorias en Stata se usa el comando "sample", el cual extrae un % como muestra. Sin embargo, este comando elimina las observaciones que no estén dentro del porcentaje, y eso es un problema puesto que el ejercicio nos pide extraer muestras consecutivamente empezando con el 5 %, luego el 10 %, así de 5 en 5 hasta llegar al 100 %.
Afortunadamente hay un par de comandos que nos pueden ayudar a preservar la base de datos y de esta forma volver a ejecutar los comandos. Estos comandos son: **preserve** y **restore**; el primero guarda temporalmente el conjunto de datos para posteriormente recuperarlos con el segundo.
Creamos una matriz para almacenar los datos, al mismo tiempo crearemos un bucle para extraer muestras de 5 en 5 por ciento, pero preservaremos y restauraremos los datos con **preserve** y **restore**.
CODIGO:

```
* Creamos una matriz para almacenar los promedios de los ingresos
matrix A = J(20,1,.)

* Creamos una variable local "j" que nos ayude para llenar los
promedios en la matriz, y hacemos un bucle con forvalues
local j = 1
forvalues i = 5(5)100 {
    preserve
    sample `i'

    * Calculamos el promedio
    sum ing_tri, meanonly

    * Almacenamos el promedio en la matrzi
    matrix A[`j',1] = r(mean)
    restore
    local ++j
}

* Mostramos los resultados
matrix list A
```

RESULTADO:

```
  8. }
(377,378 observations deleted)
(357,516 observations deleted)
(337,654 observations deleted)
(317,792 observations deleted)
(297,930 observations deleted)
(278,068 observations deleted)
(258,206 observations deleted)
(238,344 observations deleted)
(218,482 observations deleted)
(198,620 observations deleted)
(178,758 observations deleted)
(158,896 observations deleted)
(139,034 observations deleted)
(119,172 observations deleted)
(99,310 observations deleted)
(79,448 observations deleted)
(59,586 observations deleted)
(39,724 observations deleted)
(19,862 observations deleted)

.
. * Mostramos los resultados
. matrix list A
```

```
A[20,1]
           c1
 r1   12292.78
 r2  12335.874
 r3  12210.964
 r4  12222.824
 r5  12086.784
 r6  12343.415
 r7  12328.362
 r8  12326.919
 r9  12252.139
r10  12337.338
r11  12217.673
r12  12304.752
r13  12253.638
r14  12201.129
r15  12151.624
r16  12222.988
r17  12208.464
r18  12180.817
r19  12231.753
r20  12233.881
```

Para observar mejor las dispersión entre los promedios se usa un boxplot. Pero para ellos convertirmos antes la matriz en conjunto de datos.
CODIGO:

```
* Mostramos los resultados
matrix list A


* Convertimos la matriz en un conjunto de datos
svmat A

* Creamos el box plot
graph box A1
```

RESULTADO: