Introducción a la estadística descriptiva y la gestión de datos en Stata

Iván Andrés Trujillo
Universidad Surcolombiana *

June 29, 2022

Contenido

*

1 Generalidades

El siguiente texto está dirigido a aquellas personas que tienen necesidades laborales o académicas en el manejo de bases de datos,como para aquellos que se introducen en la estadística descriptiva. En general, el esfuerzo en organizar y depurar una base de datos puede verse reducido considerablemente cuando se trabaja con un software como Stata. El contenido de este documento se ha basado en las necesidades mas comunes a los que se podría enfrentar laboralmente, muchos de las rutinas aquí contenidas han sido extraídas o modificadas de trabajos de investigación donde ha participado el autor así como de los requerimiento usuales en su labor, dichas rutinas en gran parte han sido modificadas o extraídas del statlist a quienes el autor agradece y reconoce como autores. cabe señalar que este documento será útil solo para aquellas personas que se introducen en el tema, y que irá perdiendo su valor en el grado que se vaya interactuando con el software, y que el objetivo principal es que el individuo se base en el soporte que da stata, que es más que suficiente.

Existen muchos libros de introducción a Stata, algunos podrán superar al presente en contenido e incluso en la didáctica, No obstante, con este documento pretendemos dar un manejo práctico a algunas bases de datos que son útiles para estudiantes de epidemiología, salud pública y de economía. Se recomienda, para un entendimiento más adecuado de los conceptos estadísticos el libro de(**?**).

En cuanto al contenido implícito el texto esta dividido en cuatro componentes o secciones que pueden considerarse independientes si el lector tiene cierto grado de destreza:

1.0.1 Sintaxis del lenguaje de STATA

En esta seccion se introduce al lenguje del software, dado que Stata "entiende" un lenguaje, respeta una sintaxis.

1.0.2 Gestión de base de datos

La depuración, estandarización y unión de base de datos es importante para la obtención de resultados debido a que menudo la información no se ecuentra consolidada o no está en la forma adecuada para realizar las descripciones necesarias. Por lo que en esta seccion se desarrollan elementos utiles y basicos para la gestion de la base de datos.

1.0.3 Análisis empírico

dentro de esta sección se manejarán los estadísticos básicos que utilizará el investigador para hacer descripciones acerca de la información que posee y se plantee las hipótesis a probar.

2 introducción

Lo primero que debemos preguntarnos es ¿por qué manejar un software como Stata?, la respuesta quizá sea vea justificada cuando se trata obtener resultados de investigación o prácticos sobre cientos, miles o millones de individuos, podríamos estar interesados en conocer el salario promedio de las mujeres en Colombia por grupos etarios en un año particular, cuyo tiempo cálculo en un paquete estadístico como Stata, sería ínfimo, incluso podríamos obtener los resultados para un periodo de tiempo determinado (si se cuenta con la información) sin mayor esfuerzo alguno.

Aceptada la eficiencia en de los paquetes estadísticos, para describir las características de la población objeto de estudio, e identificar patrones en las variables de manera mas sencilla y confiable, nos queda decir que Stata es un programa deseable pues proporciona un lenguaje sencillo pero potente que lo posiciona como uno de los mejores programas en el mercado.

Stata, los paquetes, la comunidad y los desarrolladores.

2.1 Tipos de Variables

En el estudio de los fenómenos existen dos tipos de variables; las variables cualitativas y las cuantitativas, las primeras corresponden a las categorías como el sexo, la ocupación, la religión, la presencia o ausencia de una morbilidad, y dentro de las segundas están las discretas como el número de hijos, comorbilidades, es decir todas aquellas de caracter discreto, número de personas que viven el hogar. Son variables continuas; el peso, la edad , el salario, etc. Stata identifica la naturaleza de la variable para poder realizar operaciones sobre ellas, así las variables podrán ser de texto **string** o númericas **numeric**, por ejemplo la talla y el peso serán de carácter numeric, mientras que la ocupación o profesión será identificada como string. No obstante, existen errores de registro y por lo tanto podríamos encontrar que en la variable peso se ha registrado el peso de un individuo como 1oo en vez de 100, por lo que Stata considerara que esa variable es de carácter string, debido a que algún registro contienen caracteres no numéricos así si queremos calcular el índice de masa corporal de los individuos nos aparecería un error de sintaxis. pues la suma de variables está definida para variables numéricas. más adelante veremos como podemos tratar estos errores. lo que nos interesa en síntesis es saber que Stata identificara como variables numéricas aquellas que en cada una de sus celdas solo se hayan registrado caracteres numéricos y podrá definir como string aquellas que tengan caracteres numéricos, no numéricos o ambas.

Realizaremos un ejercicio para demostrar por que es necesario identificar cuales son los tipo de variables para Stata que contiene nuestra base de datos. No se preocupe por la falta de descripción de los pasos, cada uno de los comando será descrito posteriormente.Este solo es un ejemplo ilustrativo.

Al abrir Stata, en la barra de comandos tecleamos la palabra

```
edit
```

nos aparecerá una ventana similar a una hoja de cálculo allí podremos registrar los datos de interés para el siguiente ejercicio queremos conocer cual fue el promedio de la inflación en Colombia para el año 2018, información disponible en Banco De la Républica

Table 1: Fuente: Banco de la républica

3,18
3,27
3,33
3,23
3,10
3,12
3,20
3,16
3,13
3,14
3,37
3,68

se observará que al introducirse los datos en el programa se genera una variable denominada *var1* dicha variable contiene el registro de los datos sobre inflación, que acabamos de registrar para obtener el promedio de la inflación para el periodo de estudio podemos teclear en la barra de comando

```
sum var1
```

observamos que en la ventana de Stata encontramos una tabla pero no registra ningún resultado y en la columna *Obs* nos indica que hay cero observaciones. ¿por que?

cuando tecleamos

```
describe
```

aparece un recuadro proporcionándonos información sobre la base de datos que tenemos en memoria en este momento, observe que *Obs* nos está indicando por el contrario que tenemos 12 registros,y *vars* nos indica que tenemos solo una variable. No obstante, en la descripción de *var1* nos indica que dicha variable se identificó como *str(string)*, y eso ocurrió debido a que Stata reconoce al "." como separador decimal y no la ",". por lo tanto la variable se almacenó en formato *string* y no podemos realizar el cálculo de la media.

por ahora solucionemos esto tecleando secuencialmente:

```
destring var1, dpcomma replace

describe

sum var1
```

lo que hicimos en la primera línea fue modificar "." por "," en los registros, y reestructurar la naturaleza de la variable a numérica, lo que nos permitirá

efectivamente obtener un promedio de inflación de **3.245** para el periodo de estudio.

Cuando se trabajará en la recolección de datos para una investigación académica o laboral, es importante identificar la naturaleza de las variables en el registro de la base, debido a que algunas investigaciones requieren aplicar algún instrumento, o revisar una historia clínica, se necesita consolidar una base de datos entre un conjunto de personas y puede no haber uniformidad en como los individuos registran las observaciones, por lo tanto es recomendable para este ejercicio que se realice la colaboración en una plataforma como **google Formulario**, pues esta le permite a diferentes usuarios trabajar de manera simultánea en el registro de los datos, consolidar descripciones y parámetros de registro de los mismos. adicional, los puede consolidar en un archivo de extensión .xls o también una hoja de calculo en Excel, que puede ser facilmente importado a Stata.

En algunas ocasiones cuando un individuo no registra información sobre alguna variable, se suelen utilizar acrónimos o colocar palabras como "No registra" o "N.A". Para Stata los datos faltantes o missing son identificados de dos maneras; para las variables cuantitativas con el caracter "." Y para las variables categóricas o string ningún carácter "" (en otras palabras se deja vacío). Si recolectaremos nuestra información en una formato .xls, entonces bastará con no registrar caracteres cuando la variable no cuenta con la observación ya que cuando sea importada a Stata según la naturaleza de cada variable el software por defecto le asignará "." a los valores numéricas y "" a las categorías. Cuando no se participa en el diseño de la recolección de datos y se han utilizado ciertos caracteres para identificar la ausencia del dato tendremos que eliminar o reemplazar dichas palabras, pues para las variables cuantitativas como se había mencionado serán consideradas en formato string por lo que no podríamos aplicar operaciones sobre la misma, como por ejemplo calcular su media o desviación estándar. No obstante, veremos que dicho proceso de reemplazo o eliminación no representa mayor dificultad.

2.2 Generalidades de Stata

Hay aspectos básicos y elementales que se deben conocer acerca del paquete Stata;

los archivos con extensión **.dta** hace referencia a los archivos de bases de datos, la extensión **.do** indica los archivos de texto que contienen una sesión de trabajo; Comandos, directorios o programas. **.log** indica el formato de impresión de resultados en Stata.

si deseamos saber u obtener información acerca de un comando [1] en particular, debemos escribir en la barra de comandos **help comando**. Por ejemplo deseamos tener información del comando **list** que es utilizado para mostrar las observaciones o el valor de las variables;

```
help list
```

Para conocer nuestro escritorio de trabajo utilizamos el comando

```
cd
```

[1] Cabe mencionar que Satata es sensible a la composición de la escritura de los caracteres pues reconoce diferencias entre mayusculas y minisculas

que es muy util por que es el lugar donde el programa reconoce donde stata aplicara las sentencias o comandos por ejemplo donde buscara una base de datos, o donde guardará los archivos producto de la sesión como las gráficas. Stata por determinación trabajará en su carpeta de instalación, para cambiar esta carpeta basta con indicar la ruta de trabajo.

```
cd "C:\Users\MICROSOFT\Desktop\Taller Stata y descriptiva"
```

allí el usuario específica la ruta donde sea mas comodo trabajar.

igualmente Stata trabajará con una sola base de datos en memoria esto es muy importante debido a que en el momento de abrir una base de datos determinada se debe trabajar con la memoria limpia de lo contrario se tornará en error (las versiones mas recientes no tienen problema en limpiar la memoria automáticamente).

el comando **browse** nos permite abrir la base de datos para observar su contenido e igualmense si deseamos modificarla debemos introducir el comando **edit ??**

3 Sintaxis del lenguaje de STATA

Por ahora solo es necesario conocer los aspectos básicos del lenguaje de stata; podemos interactuar con el programa de dos manera uno vía el lenguaje MATA y el otro informalmente como ADO, en este último nos centraremos en este taller.

```
\textbf{Comando Variable(s) condicional (if) rango (in) , opciones}
```

Stata reconoce como primera instancia los comandos, por ejemplo el comando de regresión, el comando de estadística descriptiva etc... posteriormente stata reconoce la lista de variables a las cuales se les va a aplicar tales comandos, posteriormente acepta el condicional para hacer subestimaciones en categorías o en ciertos intervaos, y posteriormente acepta el rango de las observaciones por ejemplo las diez primeras y por ultimo las opciones de cada comando que son complementarias.

```
sysuse dir
```

observamos cada una de las bases de datos que tiene incoporado STATA.

Utilizaremos la base de datos cancer por ejemplo.

```
sysuse cancer, clear
```

se agrega la opción debido a que STATA Debe permanecer con su memoria de trabajo limpia.

```
help list

sysuse auto, clear

list price if price>1000


sysuse cancer, clear
list studytime in 1/19


list studytime age if age>20

list studytime age if age>20, table nocompress sep(0)




\textit{

  bysort Variables: Comando Variable(s) condicional (if) rango (in) , opciones


  }
```

Stata y algunos comandos permiten que se les agregue un comando que repite las operaciones por submuestras o categorías.

```
bysort died: list died, sep(0)
```

4 Operadores

4.1 Operados lógicos

Los operadores lógicos son importantes por que nos permiten obtener resultados para submuestas que cumplen una condicón o sentencia lógica como vimos en la sección anterior para obtener la edad de los mayores a 20 años.

1. & es "y"
2. | es "o"
3. $\sim$ es la negación
4. $\leq$ menor o igual
5. $\geq$ mayor o igual

6. == igual
7. ! = diferente a

4.2 Operadores aritméticos

1. + adición
2. − sustracción
3. ∗ multiplicación
4. / division
5. ∧ potenciación

4.3 Operados de letras (string)

aquí se utilizan dos operados + y ∗

Ejercicio

1. Compruebe la formula $(a + b)^2 = a^2 + 2ab + b^2$ para cualquier par de números.
2. realice la siguiente operación $\frac{a+b}{c} + \frac{a+c}{b}$

5 Programa y Hojas de ruta

cuando hacemos un análisis o la gestión de una base de datos podemos utilizar nuestra propia hoja de ruta que utilizará en STATA como extensión **.do**

5.1 Macros

Los macros los podemos dividir en dos categorías en local y global, su función es un item que almacena varios carácteres.

```
local a = 2
display ‘a’

global a = 2
display $a


**
sysuse auto, clear
tabulate rep78
local x " 1 2 3 4 5"
foreach k of local x {
sum price if rep78==‘k’
}
```

```
** Resumen

levelsof rep78, local(l)
foreach  k of local l{
quietly sum price if rep78==‘k’
display as text " La media del precio de los carros con ‘k’ reparaciones es " = ‘r(mean)’
}



doedit

*Mi primer dofile

 cd "C:\Users\acer\Desktop"

use data.dta, clear

exit


do miprimerdo
```

Ahora los programas son especiales para generar nuestra ruta de comandos[2]

```
capture program drop salude
display as text " ¿Quiubo? ‘0’"
end




caputure program drop mylist

list ‘0’ , sep(0) nocompress

end

sysuse cancer, clear

mylist age
```

5.2 Ejercicio

1. Genere un programa que establezca un directorio de trabajo para una sesión de Stata.

[2]Los programas pueden ser cargados directamente si los guardamos como extensión **.ado** y los almacenamos en la carpeta de directorio personal para ello escribimos en la barra de comandos el comando **sysdir**

2. Genere un texto como mínimo de dos parrafos que tenga coherencia utilizando los macros.

Codigo base

```
capture program drop texto1
program define texto1

display as text "mi nombre es ‘1’ soy de la ciudad de ‘2’ y estoy estudiando ‘3’."

end
```

SSC install

En algunas ocasiones, existen programas desarrollados por usuarios para diseños especificos, asi si conocemos el nombre del programa bastara con escribir en la barra de comando, dentro de estos paquetes estara ssc install estout,

6 Gestión de Base de datos

En la investigación empírica el insumo primordial son los datos aveces contamos con ellos pero no poseen las información relevante para nuestra investigación puesto que por ejemplo necesitamos constuir indicadores. Una buena descripción de una base de datos también permite colaborar en los grupos de investigación lo que agiliza el proceso o permite a tercero valorar o validar nuestros resultados.

En el proceso de investigación muchas veces las investigaciones cuenta con k procesos consecutivos lo que usualemente genera la necesidad de agregar observaciones o variables, esto es útil para consolidar la base de datos y obtener los resultados esperados.

```
sysuse lifeexp, clear
describe
```

Obtendremos una descripción de la base de datos de manera, que podamos conocer sobre que datos estamos trabajando o de de que entidad o proyecto de investigación se extrayerón.

6.1 los datos en Stata

La creacíón de variables, la transformación de su escala, la creación de indicadores, es importante para obtener relaciones empiricas y hacer mas robusto el proceso de investigación en cuesitón.

Averiguemos que información tenemos disponible en el Banco Mundial

6.1.1 Creación de Variables

A veces se necesita constuir una nueva variable, que puede ser por ejemplo; un identificador, una transformación o un indicador.

```
help generate
```

Utilizando la función logarítmica obtenemos ;

```
sysuse census,clear
generate lnpop= ln(pop)
```

6.2 Rename & Label

En ocasiones es importante renombrar las variables, o incluso dejarles etiquetas para saber cual es su descripción.

```
sysuse cancer,clear
```

Note que la variable died toma dos valores

```
codebook died
```

pero nos indica que tiene una etiqueta descriptiva que las observaciones con el número uno significan que murieron tales pacientes.

6.3 Keep & drop

Estos comandos son importantes por que nos permiten mantener o elimnar observaciones o variables.

6.4 Reshape

El comando reshape es de gran utilidad, puesto que permite llevar a formato de panel algunas bases de datos, por eso es importante organizar las variables de modo que se tenga en cuenta el efecto temporal es decir; si tenemos la población para diferentes periodos es bueno tenerla de la manera

Table 2: Base en formato Ancho

id	pob2000	pob2001	pob2002	pob2003	.	.	.	pob200k
Municipio 1								
Municipio 2								
Municipio 2								
.								
.								
.								
Municipio N								

```
reshape wide(long), i(id) j(year)
```

Table 3: Base en Formato Largo

id	año	Pob
Municipio 1	**2001**	
Municipio 1	**2002**	
Municipio 1	**2003**	
Municipio 1	**2004**	
.	.	
.	.	
Municipio 1k	**200k**	
Municipio 2	**2001**	
Municipio 2	**2002**	
Municipio 2	**2003**	
.	.	
.	.	
.	.	
Municipio 2k	**200k**	
Municipio N1	**2001**	
Municipio N2	**2002**	
.		
.		
Municipio Nk	**200k**	

j es una una variable o una existente, tiene la ultima condición cuando vamos de formato largo a corto, y es una una variable cuando vamos de formato ancho a corto.

```
*Uso del comando reshape
```

```
wbopendata, language(en - English) country() topics(8 - Health) indicator() long
```

6.5 Ejercicio

observar como evoluciona la tasa de fertilidad através de de los años de manera global, y comparar si este resultado ha sido equivalente al

6.6 if & cond()

Los condicionales son importantes para la generación de variables, por ejemplo para la creación de categorías dada una condición particular.

```
help cond
```

la funcionalidad del comando **cond** es importante debido a que sigue la estructura lógica si x entonces p.

```
sysuse cancer,clear
gen live= cond(died==0, "muerto", "vivo")
```

Lo que específica que al generar la variable live tomará la categoría de muerto si el número de la variable **died** es cero de lo contrario se le asignara la categoría vivo.

6.7 Ejercicio

6.8 Insheet

Hay que tener en cuenta que las bases de datos de algunas entidades oficiales vienen en formato plano.

1. Escriba un programa para obtener los logarimos de un varlist.

6.9 import excel (vs 12 ¿)

el Comando *import excel* nos facilita la importacion de la base de datos en formato .xls a Stata. Sin embargo, para que la base sea importada de manera adecuada, sin mayores ajustes, es recomendable que la estructura de la Matriz o base de datos, en el foramto .xls se considere como nombre de la variable la primera fila, es decir; no pueden existir celdas por encima del nombre de la variable dado que esta es la primera fila y debajo de ella deben apracer los registros de inmediato y segundo que no existan celdas compartidas.

```
import excel using archivo.xlsx, sheet("hoja") first
```

Se le debe especificar al comando que hoja de trabajo es la que se va incorporar en la base de datos de stata así mismo la opción first resalta la condición de que la primera columna de la hoja corresponde al nombre de las variables. recuerde que Stata solo trabaja con una base de datos en memoria por lo que es necesario anteceder una linea de comando con *clear all* o utilizar la opcion clear, permitida por la mayor parte de comandos de importacion.

6.10 Append & Merge

Para esto vamos a ver la intuición;

podemos abrir y dejar solo 30 países despues, y despues pegarle el resto con append, con merge si los dos modulos y hagale..

6.11 La homogenización de la base

Antes de iniciar está sesión es mas que obvio que hay diferentes formas de conseguir el mismo resultado, es por esto que el programa es tan versátil.No obstante, hemos tratado de realizar rutinas generalizables con el objetivo de que sirvan para cualquier base de datos y que la rutina sea intuitiva y legible fácilmente. Sin embargo, el lector puede modificar las rutinas de forma que le sean más optimas en su trabajo, hemos dicho previamente que Stata es sensible a las Mayúsculas y Minúsculas, así por ejemplo en el registro de los departamentos de residencia habitual de los individuos para la tabulación de frecuencias con el comando

```
tab
```

se obtendrán diferentes resultados si en los registros no hay uniformidad "Bogota", "bogota" aunque en conjunto sean el mismo departamento para Stata son dos diferentes.

```
replace dpto=ustrupper(dpto)
```

lo que se hace es transformar todas las letras correspondientes minúsculas a mayúsculas. No obstante, el problema se seguirá presentandose para el caso siguiente "BOGOTÁ" "BOGOTA" una forma de corregir este proceso es homogeneizar el registro sin los caracteres de acentuación por medio de la función *subinstr()* que tiene como dominio cuatro parámetros *subinstr(1,2,3,4)* en el 1 se indica la variable, en 2 se indica que caracter se quiere reemplazar y en el 3 por cual, en el 4 la posición en la cadena de caracteres que se quiere modificar, puede ser el primero, segundo, tercero, o hasta la n-ésima posición, si se quiere modificar o reemplazar en toda la cadena de caracteres se utiliza el "."

```
replace dpto=subinstr(dpto,"Á","A",.)
```

En algunas ocasiones, se trabajan con bases de datos en la cual desde Excel parecen homogéneas Empero, Stata reconoce el " " (espacio) como un caracter por lo que los registros "BOGOTA " Y "BOGOTA" son diferentes, note que el primero tiene un espacio después de la A, para corregir esto solo debemos reemplazar dicho caracter por el vacío;

```
replace dpto=subinstr(dpto," ","",.)
```

Notará el lector que la estructura de la línea anterior sirve para eliminar cualquier carácter indeseable dentro de una variable, por ejemplo en algunas series númericas se presenta con facilidad el singo "$" como también el de "%".

Para corregir esto de manera general (En cualquier base de datos), se puede generar un programa denominado *work1* que preserva el nombre del programa personal del autor que en estructura general son iguales excepto por unas mínimas variaciones.

```
ds, has(type numeric)
 foreach  x  in  `r(varlist)'{
replace `x' =upper(`x')
replace `x' = subinstr(`x'," ","",.)
replace `x' = subinstr(`x',"Á","A",.)
replace `x' = subinstr(`x',"É","",.)
replace `x' = subinstr(`x',"Í","I",.)
replace `x' = subinstr(`x',"Ó","O",.)
replace `x' = subinstr(`x',"Ú","U",.)
}
```

Note que a este programa le puede agregar las siguientes lineas para hacer el ejercicio más dinámico y detectar errores

```
tab `x', missing

tab `x'
```

Table 4: Variable con múltiples categorías

varn1	varn2
conyugue,madre	conyugue y madre
hijos, conyugue, otra persona	hijos , conyugue y otra persona
padre,madre,conyugue	padre, madre y conyugue
solo	solo

En algunas ocasiones, cuando se trabaja con alguna base de datos modificada e importada desde Excel directamente a Stata, se generan variables que no contienen ningún dato, lo mismo ocurre cuando se desea analizar un subconjunto de la población para la cual múltiples variables pueden no contener información, para solucionar el problema de manera automática podemos utilizar el siguiente programa:

```
capture program eliminate
ds, has(type numeric)
foreach x in ‘r(varlist)’{
sum ‘x’
if ‘r(N)’==0{
drop ‘x’
end
}
}
```

Notará el lector que este programa puede guardarse como un ADO y posteriormente ser llamado directamente desde el comando *work1*

Otro problema recurrente es la aparición de columnas con múltiples categorías en algunas ocasiones separados por algún caracter como ”-” o ”,” lo que puede limitar la capacidad de obtener información más detallada a la hora de obtener resultados:

La anterior tabla presenta las variables *varn1* y *varn2* que en esencia contienen la misma información, pero fueron registradas de manera diferente en la base de datos, en la primera columna de la tabla anterior las categorías se encuentran separadas por ”,” mientras que en la segunda no. Podríamos separar la primera columna en dos debido a que contiene el separador ”,” no obstante para la segunda columna identificaremos los términos (como ejercicio podría realizar un programa que genere variables que contengan las categorías en variables).

las siguientes funciones son importantes: agregar a esta sección strpos, int, substr.

En algunas ocasiones nos interesa dejar o eliminar solo caracteres o cifras de una variable de igual longitud, el caso particular de la Base de Datos de Defunciones dado que está codificada con el código CIE-10 y necesitamos agrupar

en algunas ocasiones, las variables vienen etiquetadas pero sus nombres son muy largos

```
   local i = 0
ds, has(type string)
 foreach d in ‘r(varlist)’{
local n‘d’ = substr("‘d’",1,15)
rename ‘d’ v‘i’‘n‘d’’
local i= ‘i’ +1
}
```

Para algunos fines prácticos nos interesa dejar de una cadena de caracteres numéricos o no-numéricos, tan solo los primeros dos o tres dígitos;

La función *cond(1,2,3)* está compuesta de 3 argumentos, el 1 indica la variable en la cual se va a evaluar la sentencia lógica por ejemplo Price¿#, o Price==# donde # es cualquier número real, el 2 indica que valor se determinara si la sentencia lógica se cumpla y 3 el valor en caso en la cual dicha sentencia no se cumpla, podemos utilizarla para generar variables:

```
Gen var=cond(Price>10,1,2,)
```

No obstante hay que tener cuidado si la variable presenta datos faltantes o missings debido a que les asignara un valor dado que no cumple la condición, para corregir esto podemos hacer:

```
Gen var=cond(missing(var1),.,cond(Price>10,1,2))
```

Asi la función º*missing* indentificará que si es un dalto faltante , le proporcione un punto dado que me interesa que la variable resultante tenga el carácter nùmerico.

6.12 Hallar un resultado

El siguiente ejercicio se basa en la siguiente base de datos,

```
clear all

input str6 var1 str6 var2 str6 var3
A  H F  A
B  K N  P
C  A T  C
D  T A  D
E  J R  D
end
```

y consistirá en identificar si en cualquiera de las variables *var1-var3* está contenida letra "A".

```
forvalue i=1/2{
display as text "var‘i’"
local f    var‘i’,  ‘f’
 display  as text "‘f’"
}

gen id2=1 if inlist("A",  ‘f’  var3)
```

6.13 Expresiones regulares

Las expresiones regulares nos serán de utilidad en la medida que deseemos depurar nuestra información textual de la manera mas general y eficiente posible.

Para una revisión mas detallada de este método de depuración, ver el apéndice de generalidades de informática por ahora, basta con dar los siguientes ejemplos prácticos, cuando queremos sustituir y o eliminar caracteres ya sea alfanuméricos o no alfanuméricos que contiene una variable, en especial de tipo string.

para esto utilizaremos la función **regex** que se basa en dos parámetros o argumentos:

```
regex(varname,"regexpression")
```

varname indica la variable de tipo string y el segundo argumento "regexpression" es nuestra expresión regular, que detallaremos enseguida como funcionan, regex retorna el valor de 1 cuando varname cumple la expresión regular y cero en el caso contrario.

```
clear all
input str8 sumas
"1+1"
"1+2"
"1+3"
"1+4"
end
gen suma=1 if regex(sumas,"1+1")

gen suma2=1 if regex(sumas,"1\+1")
```

la variable sumas, esta compuesta del operador binario de adición, y dos números naturales, queremos identificar donde se encuentra la expresión "1+1", al ejecutar la primera linea de comandos, nos dará como resultado elementos vacíos, para esta base de datos y esta expresión regular solamente se obtienen vacíos, no obstante, si agregamos el backslash (\) al signo de adición (+) obtenemos lo que deseamos, esto es así debido a que ese signo es un metacarater y por lo tanto el computador lo reconoce como un símbolo especial, con el (\) se anula su significado dando énfasis a su expresión textual.

Notemos hasta el momento el parecido de regex con subinstr, cabría de preguntarnos cual es la diferencia, y esta radica en la capacidad que poseen los metacaracteres como motor de búsqueda y que son soportados por regex.

En algunas ocasiones, las variables contienen datos de carácter alfanumérico(letras y números), pero para algún fin práctico necesitaremos remover ya sean los números o las letras, hasta el momento regex nos permite identificar cuando se cumple la sentencia, para el reemplazo directo utilizamos la función **regexr** que consta de 3 argumentos:

```
regexr(varname, "regex", "newstring")
```

varname corresponde a la variable que queremos modificar, regex la definición de la expresión regular y por ultimo esos carácteres por cuales queremos modificar.

```
clear all
input str8 iscod
is-123
io-124
il-123
io-124
end

replace iscod=regexr(iscod,"[a-z]*\-","")
```

como podrá observar en su pantalla se han removido de la variable las letras y el guión, si no nos hubiese interesado los caracteres numéricos si no las letras sería necesario el siguiente código:

```
gen iscod2=regexr(iscod,"\-[0-9]*","")
```

Debemos tener en cuenta que hemos hecho uso de una expresión regular para los caracteres alfabéticos "[a-z]" en minúscula, y estos se deben distinguir en mayúscula ["A-Z"] cuando sea el caso. No obstante podemos también utilizar una expresión para ambos caracteres ["A-Za-z"], pero se espera que en este momento el lector ya haya hecho una depuración previa de su base de datos, y esta se encuentre homogenizada.

6.14 While

Esta es importante por que mantiene ciertas condiciones

```
local i = 0

while ‘i’<10{
display as text " el número ‘i’
local i = ‘i’ + 1
}
```

Ejercicio

6.15 forvalue

En algunas ocasiones necesitamos realizar ciertas operaciones o ejecutar sobre diferentes archivor por ejemplo necesitamos empalmar las n-esimas bases data1 , data2, data3, ..., datan. para eso *forvalue i* nos permitira realizar el empalme de manera eficiente.

```
forvalue  i = 1/10{
display ‘i’
}

los nùmeros pares e impares.
```

Asi para empalmar las bases de datos de manera consecutiva bastara con especificar el bucle a la variable del texto, asi por ejemplo si se tienen 10 bases con la denominacion data:

```
use data1,
forvalue i = 1/10{
append data ‘i’
}
```

obser que la base maestra esta en formato .dta, por lo que es necesario que las demas bases tambien se encuentren en el mismo formato.

6.16 datos atípicos

Es importante entender si existen datos atípicos (extremadamente heterogéneos) o si hay errores de registro(por ejemplo registrar la talla de un individuo en centímetros cuando la de los demás esta en metros), este apartado utiliza los conceptos de posición relativa no obstante se presentan aquí con el objetivo de que el lector vaya formando su

programa cada vez mas óptimo, los detalles con mayor exactitud en el apartado de las medidas de posición relativa. No obstante el teorema de shebysheb y la regla empírica nos pueden ayudar a sospechar de algunas observaciones, hago énfasis en sospechar debido a que en algunos casos por ejemplo para el producto interno bruto (PIB) existe una brecha de tamaña importancia entre los países y no se debe a observaciones con error en el registro.

6.16.1 Teorema de shebysheb y la regla empirica

La regla empirica es un resultado derivado de la normalidad de una variable, no obstante el teorema de shebysev es para cualquier función y por lo tanto para nosotros particularmente más adecuado. El teorema expresa que *dado un número* $k \geq 1$ *y un número* n *de mediciones entonces,* ***al menos*** $1 - \frac{1}{k^2}$ *de las observaciones estarán a* k *desviaciones estándar de la media*

```
display as text "{center:{ title: El teorema de shebyshev}}"
forvalue i  = 1/4{
display as text "{it: a ‘i’ desviaciones estandar de un conjunto de n mediciones, al menos
 " (1- 1/(‘i’)^2) " de las mediciones estára dentro de ‘i’ desvaiaciones estadnar de la media}"
}
 end
```

6.16.2 Ejercicio

Realice un programa que identifique los valores atipicos a 3,4 y 5 desviaciones estandar de la media de la distribucion, de las variables que desee especificar en un varslit (lista de variables) y que nos proporcione informacion con respecto al porcentaje de datos atipicos en la muestra por variable.

El Banco Mundial organizacion provee una util herramienta para los usuaros de Stata, es un comando que permite

```
clear all
** por determinación estamos trabajando en la carpeta de stata
ssc install wbopendata

wbopendata, language(en - English) country() topics(3 - Economy & Growth) indicator() long
save economy
clear
wbopendata, language(en - English) country() topics(8 - Health) indicator() long
save health
clear all

wbopendata, language(en - English) country() topics(6 - Environment) indicator() long
save enviorement

clear all


use economy


 merge 1:m  countrycode year using health.dta

 drop _merge
```

```
merge 1:m countrycode year using enviorement.dta, nogenerate

save panel
```

7 Análisis empírico

En esta sección se tratará en orden los siguientes temas;

1. Medidas de tendencia central; Media, moda, mediana.
2. Medidas de variabilidad; varianza y rangos
3. teorema de shebyshev
4. Medidas de posición relativa
5. Graficos.

7.1 Gráficas descriptivas

7.1.1 Pastel

Se usa para gráficar la frecuencia relativa o o la propoción de observaciones en una cateogría dada del total, por ejemplo el porcentaje de hombres y mujeres en la población de Neiva.

```
help graph
```

seleccionamos

```
graph pie
```

para ver la descripción del comando.

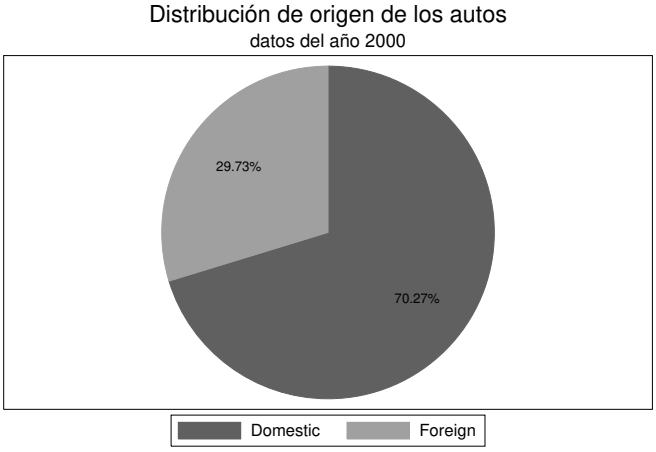
piense como se grafícará la torta si tuviera que hacerla manualmente.

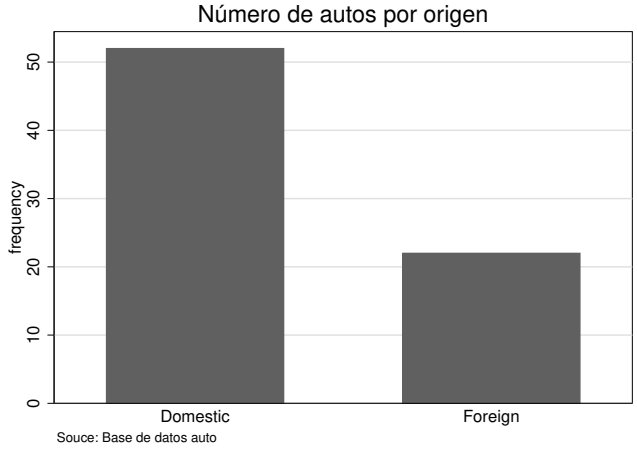
Para la base de datos **cancer** grafiquemos el porcentaje de vivos y muertos de la muestra.

```
sysuse cancer, clear
graph pie
```

7.1.2 gráfica de barras

se utiliza especialmente para mostrar las frecuencias absolutas de las categorías. por ejemplo se le pregunta a 150 pacientes de un hospital que califiquen en excelente, bueno, regular y malo la antención en la entidad. Otro gráfico de barras sería la asignación de presupuesto anual a los ministerios. Piense también en el ingreso promedio de diferentes grupos poblacionales, es decir por rango de edad, por sexo, por etnia etc.





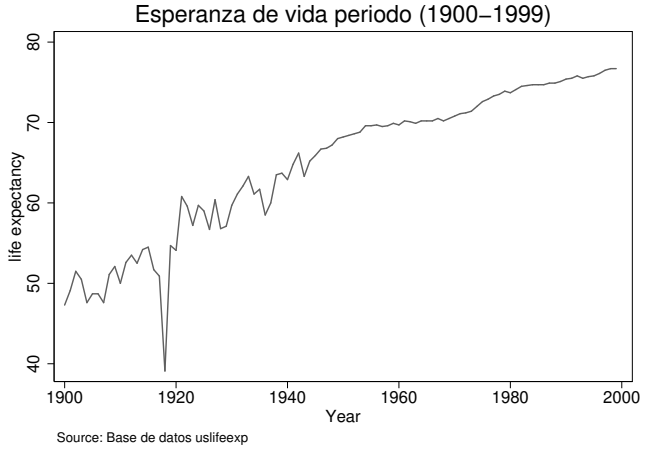
7.1.3 Gráfica de lineas

Se utiliza para observar la **tendencia** de una serie de tiempo, es decir una variable que ha sido registrado a través del tiempo. Un ejemplo es el caso de afiliados al régimen contributivo través del tiempo, o el gasto del ministerio de salud en sus diferentes rubros a través del tiempo, el número de Aquí hay que tener cuidado con las escalas de los ejes.

7.1.4 histograma

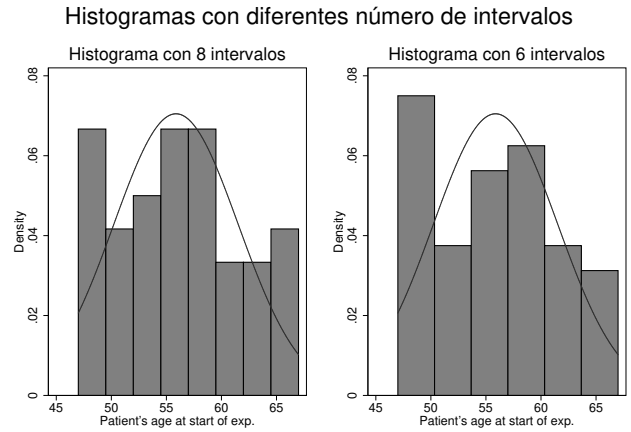
El histograma es utilizado para observar la distribución de la variable misma en intervalos, el histograma es sensible a la cantidad de clases que se construyan. por ejemplo deseamos observar la distribución de una variable por ejemplo el peso de recién nacidos en un hospital, o la presión sanguínea.

hay que distinguir en histogramas con clases de igual longitud o no.



hay que tener en cuenta el método de inclusión a la izquierda. el área del histograma?

Una distribución puede estar sesgada a la derecha o a la izquierda; es decir cuando está sesgada a la derecha existen pocos datos anormalmente grandes y el caso contrario para el lado izquierdo.



3

7.2 Ejercicio

utilizando la información que puede conseguir en Dane, realice los siguientes ejercicios:

[3]Es decir podemos concluir que se pueden utilizar tanto las barras como las frecuencias relativas para algunos casos

1. Observar la tendencia de la población de Huila, Tolima, Bogotá y Chocó.

   Mirar si se pueden programar los ejercicios que tengo en excel de demografia en STATA.

2. Observar la tasa de crecimento de la población promedio para un periodo de tiempo 1980-2010.

Codigo base

```
encode departamento, gen(dpto)

label list dpto

reshape long a , i(dpto) j(year)

xtline a if dpto==6 | dpto==19
```

7.3 Medidas de tendencia central

Las medidas de tendencia central nos sirven para tener información acerca del comportamiento de la variable a estudiar, como su nombre lo indica las medidas de tendencia central dan un una estimativo del centro de una variable estudiada.

7.3.1 La media

La media es un estadístico usual por definición la media es;

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

que posee unas propiedades por ejemplo de la ecuación anterior se deduce que;

$$\bar{x}n = \sum_{i=1}^{n} x_i$$

uno de los problemas que tiene la media es que es muy sensible a los datos atípicos por ejemplo.

En la ventana de STATA escribimos

```
sysuse cancer, clear

help tabstat

tabstat age
tabstat age studytime
tabstat age studytime , by(died)
tabstat studytime age, by(died) s(N)
```

```
tabstat studytime age, by(died) s(mean N)
```

7.3.2 La mediana

La mediana es la posición relativa en la cual el 50% de los datos son más grandes que ella, y el otro 50% son menores a ella, esta se consigue ordenando los datos de menor a mayor:

$$1 - 2 - 3 - 4 - 5 - 6 - 7$$

la mediana es 3, cuando el número de observaciones son par entonces escogemos el promedio de las dos centrales.

7.3.3 la Moda

La moda es una medida de tendencia central que nos permite obtener el dato que encuentra con mayor frecuencia en los datos.

7.4 Ejercico

1. Utilizando la base de datos del banco mundial que hemos guardado como panel encuentre el promedio de la esperanza de vida de los paises agrupados por al menos dos regiones incluyendo en cada región como minimo 3 países , y comparale con el agragado de la base.

2. Genere un comando que obtenga el promedio de una variable específica.

Codigo base

```
use panel, clear
gen region#=1 if country==# | country==# | country==3
gen region#=1 if country==# |.......
```

Para la creación del programa que obtenga la media de una variable podemos utilizar varios métodos(incluso unos muy sencillos) u otros más rudimentarios.

Codigo base

```
** La media aritmetica  como programa:

capture program drop mean
program define mean

egen suma = total(‘1’)
scalar sum = suma[_n]
drop suma
```

```
quietly describe ‘1’
scalar media= sum/‘r(N)’

display as text " la media observada de la variable ‘1’ es ="  media

end
```

7.5 Medidas de variabilidad

Las medidas de variabilidad son importantes por que nos dan un estimativo de la dispersión de los datos, unas se basan en el centro como la varianza y otras en el ancho de la medidición como el rango.

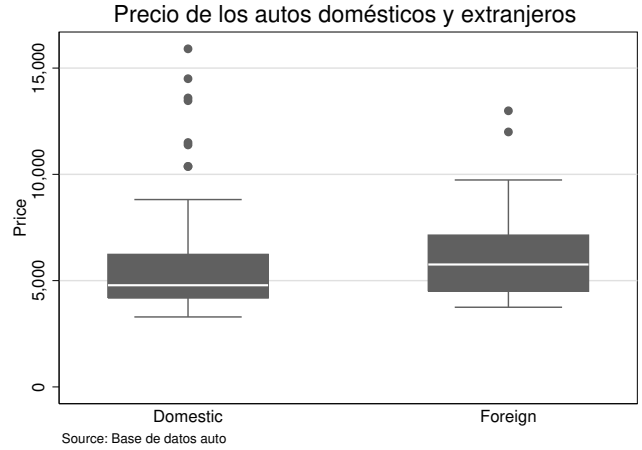
7.6 El rango

El rango es una medida de variabilidad que considera la distancía o diferencia entre el mayor valor y el menor valor de la distribución. Sin embargo, no es tan sensitivo a la distribución de los datos en general.

```
sum x
‘r(max)’ - ‘r(min)’
```

7.6.1 Gráfico de caja

el gráfico de caja está basado en las medidas de posición relativa, también se utiliza para observar la distribución de las variables entre dos grupos.



7.7 La varianza

Como podemos medir la variabilidad de los datos;

$$x_i - \bar{x}$$

Sin embargo, para las n observaciones no podemos obtener la suma;

$$\sum_{i=1}^{n}(x_i - \bar{x})$$

Los conceptos de muestra y población se hacen necesarios tenerlos en mente en el concepto de varianza. Es una medida de la dispersión de los datos a su centro. o en otras palabras el promedio de las distancias al cuadrado.

¿por qué es necesario elevarlo al cuadrado?

La varianza poblacional es

$$\frac{\sum(x_i - \bar{x})^2}{n}$$

la varianza muestral ususalmente consume un grado de libertad por que empiricamente los resultados son mejores.

la desviación estandar es simplemente la raíz cuadrada de la varianza.

$$\frac{\sum(x_i - \bar{x})^2}{n-1}$$

que pasa con la varianza si?

$$\sum_{i=0}^{n} x_i = na$$

La variabilidad de los datos es importante por ejemplo piensese en un análisis discriminatorio o una variable que tiene gran variabilidad para diferentes categorías.

podemos obtener;

$$\frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

para ver la igualdad en **??**

7.8 Ejercicio

1. Escriba un programa que solamente cálcule la varianza
2. utilizando la base de datos **census.dta** análice la variable *pop* por regiones.

Código base

```
tabstat varlist, s(statname) by(varname)
```

7.9 Medidas de posición relativa

Las medidas de posición relativa son importantes;

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

```
** Variables estandarizadas
capture program drop standar
program define standar
foreach x of local 0{
quietly sum ‘x’
gen std‘x’= (‘x’-‘r(mean)’)/(‘r(sd)’)
}
end
```

Vamos a comprobar su consistencia

```
sysuse cancer, clear
```

```
capture program drop rate
program define rate
foreach x of local 0{
gen rate‘x’ = (‘x’-‘x’[_n-1])/‘x’[_n-1]
}
end
```

8 Ejercicio práctico

En esta sección trataremos de integrar lo aprendido para analizar la evidencia empírica del caso Colombiano en las defunciones: La siguiente sección está basada en el DO-file producto de un ejercicio académico de investigación por lo que sus resultados aquí obtenidos tienen mucha utilidad practica en los estudios de mortalidad. Para un entendimiento más adecuado se sugiere la lectura del documento capitulo 1 y capitulo 3 del libro de (**?**)

La información (microdatos) aquí utilizada se encuentra disponible en: Defunciones

la tasa especifica de mortalidad:

$$TBM = \sum_{i=1}^{n} = TEM_i \frac{P_i}{P}$$

8.1 Exportación de resultados

Para ver detalladamente como podemos exportar resultados a formato o .rtf ver Estimación

Ejercicio

1. realice un análisis de las tasas de mortalidad especificas
2. ¿Dada la base de datos cual es el departamento que más presenta homicidios?
3. ¿Como se relaciona con el nivel de riqueza de las unidades geográficas?

9 Apéndice Generalidades de la Informática

Esta sección la hemos incluido como apéndice por dos razones; primero por no aburrir al novel con demasiada información, que un principio le puede ser irrelevante y segundo, algunos conceptos serán necesarios a medida que el lector incremente su destreza y sus ambiciones con la práctica.

En este apéndice trataremos cosas que son importantes a saber, de manera muy general en la informática para tener mayor claridad con el uso de paquetes estadísticos, un **bit** es la mínima unidad de información.

los datos de tipo string o cadenas de caracteres son frecuentes en las bases de datos, por eso debemos tener claras que operaciones podemos realizar sobre ellas; podemos concatenar, localizar o extraer una subcadena de otra, entras operaciones, pero las principales aquí abordadas son las mencionadas anteriormente.

Un alfabeto es un conjunto de símbolos(caracteress o grafemas)utilizados en una lengua para la comunicación. Sin embargo, dichas letras, deben representarse de manera que el ordenador las entienda (un mapeo de bits a símbolos), para ello se han usado algunos códigos que han tratado de homogeneizar dicha representación el código ASCII, el unicode, el UTF-8, y demás.

ASCII (American Standard Code for Information Interchange) es un código basado en el alfabeto latino, que representa de caracteres.

9.0.1 Regular expressions

las expresiones regulares también denominadas regexs nos ayudan a definir patrones de busqueda de caracteres,

9.0.2 Metacaracteres

Hay una lista de caracteres que debemos tener en cuenta, dado que tienen un significado especial para los ordenadores.

1. .
2. +
3. ^
4. *
5. -
6. ()

7. { }
8. [ ]
9. |

^ (caret) este caracter es útil para indicar que se debe buscar al inicio del string, recuerde que regex retorna 1 si se cumple la sentencia y 0 en el caso contrario.

```
di regex("America","^A")
```

nos dará como resultado en la pantalla de Stata un 1, indicado que se cumple la sentencia.

10 Apéndice A

Algunas tasas como tasa bruta

Tasa especifica de fecundidad (TEF):

$$TEF_i = \frac{N_i}{P_{fi}}$$

número de hijos nacidos vivos dividio por mujeres en el rango de edad i.

11 Apéndice B

Propiedades de la sumatoria:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + ... + x_n$$

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + ... + x_n^2$$

$$\sum_{i=1}^{n} x_i^2 \neq (\sum_{i}^{n} x_i)^2$$

$$\sum_{i=1}^{n} ax_i = ax_1 + ax_2 + ax_3 + ... + a_x n = a(x_1 + x_2 + x_3 + ... + x_n) = a\sum_{i=1}^{n} x_i$$

la formula computacional de la varianza:

$$\begin{aligned}
\sum (x_1 - \bar{x})^2 &= \sum (x_1^2 - 2x_i\bar{x} + (\bar{x})^2) \\
&= \sum_{i=0}^{n} x_i^2 - 2\bar{x}\sum_{i=0}^{n} x_i + \sum \bar{x}^2 \\
&= \sum_{i=0}^{n} x_i^2 - 2n\bar{x}^2 + nx_i^2 \\
&= \sum_{i=0}^{n} x_i^2 - n\left(\frac{\sum_{i=0}^{n} x_i}{n}\right)^2
\end{aligned}$$

12 Modelos de probabilidad

En este sentido trataremos de examinar los modelos de probabilidad es decir; Cuales son los efectos que tienen sobre la probabilidad de ocurrencia. Ciertos factores asociados.

Nuestra pregunta de investigación es cual es la probabilidad de estar afiliado a salud:

13 Introducción a la probabilidad

Ley de la regularidad estadística; Cuando aumenta el número de intents en un experimentos las frecuencias relativas convergen o se estabilizan.

El termino probabilidad se refiere al límite cuando el numero de sucesos se hace infinito:

$$\lim_{N \to \infty} = \frac{n(A)}{N} = P(A)$$

incluir la definicón de laplace..

13.1 La probabilidad condicionada

Cuando tenemos un especaio Muestral Ω la probabilidad de un evento A_i queda reducido al conteo de dentro de espacio muestral. Sin embargo, la ocurrencia de tal evento a veces queda reducia un subconjunto del espacio muestral, por ejemplo cuando otro suceso ya ha ocurrido $P(A \mid B)$.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Que tambien se puede reescribir como;

$$P(A \cap B) = \frac{P(A \mid B)}{P(B))}$$

Las reglas de probabilidad, son especialmente importantes como se ha visto. Sin embargo, existe un teorema particular que nos permitirá hacer diagnostico.

13.2 El teorema de Bayes

Por ejemplo la probabilidad de inferencia del evento A dado que ha ocurrido B, sin embargo la pregunta ahora es como podemos inferir la ocurrencia de B dado que ha ocurrido A.

Lo primero que debemos observar aquí es la relación que existe específicamente con la regularidad estadística y los procesos de la vida díaria o aplicado en términos concretos.

Cual es la probabilidad de obtener en una muestra exactamente 4 pacientes de una poblacion total de 20 donde hay 6 que padecen una enfermedad Z.

```
prevalenciae 20 6 4
```

```
*Este dato no lo podemos obtener por que no es programa propio de Stata

* lo puedo ejecutar por que está en el directorio "Ado" personal
```

¿Como lo obtuve? la serie de pasos que se siguen para crear un programa personal son concretos y sencillos, el lenguaje ADO, de Stata nos permite formular de manera general nuestro problema;

Eso con el fín de formular la pregunta de manera que podamos obtener información mas acertada y ligera:

¿ Cual es la probabilidad de que en una población de n personas con k infectados obtengamos una muestra de exactamente j de ellos ?

```
/* la sintaxis del comando es:


prevalenciae n k j


*/
```

Podemos estudiar algunos comportamientos de nuestra muestra supongamos por ejemplo que tenemos los mismos n pacientes pero ahora se incrementa el número de infectados.

¿que pasa con la probabilidad cuando aumenta el número de tamaño de la población?

```
forvalue i = 1/40 {
  prevalenciae 80 ‘i’ 4
  }


forvalue i = 1/20 {
  prevalenciae 80 20 ‘i’
  }


forvalue i= 40/80 {
prevalenciae ‘i’ 20 3
}
```

¿que sustenta el comando? es decir cual es la formula matemática que utilicé? véamonos algunos términos:

$$\binom{n}{k} = C(n,k) = \frac{n!}{k!(n-k)!}$$

Que se puede expresar como un teorema si se quiere y que vimos que puede resultar de gran ayuda o interés para un investigador.

$$\frac{\binom{k}{j}\binom{(n-k)}{(k-j)}}{\binom{n}{k}}$$

13.3 3 distribuciones

por ser la más sencilla y discreta empezaremos con la binomial:

13.4 La distribución binomial

se trata de los ensayos bernoulli debido a que hay dos posibles resultados asociados con exito y fracaso.

Pensemos en un momento en el lanzamiento de una moneda al aire, los dos posibles resultados son cara o sello. Por lo tanto, podemos estudiar el comportamiento de la variable Aleatoria "Número de caras obteneidas" al obtener una cara decimos que el resultado del experimento ha sido un exito para este caso la probabilidad asociada del evento exito es $1/2$ por lo que le fracaso tedrá la misma probabilidad de ocurrir debido a que;

$$P(e) + p(e)^c = 1$$

alguna de las dos debe ocurrir, entonces

$$p(e)^c = 1 - p(e)$$

Ademas la probabilidad de obtener un evento k veces;

$$p(e)^k$$

Ahora como estamos interesados en experimentos repetidos nos preguntamos la probabilidad de obtener en n ensayos k exitos, por lo que habrían $n - k$ fracasos;

$$p(e)^k \left(1 - p(e)\right)^{n-k}$$

pero hay también $\binom{n}{k}$ formas o maneras de obtener los k exitos entre n por lo que se deduce la formula de la distribución binomial;

$$\binom{n}{k} p(e)^k \left(1 - p(e)\right)^{n-k}$$

Para aclarar más esto observemos que de 4 monedas 2 éxitos se pudieron obtener de la manera:

$$\begin{array}{l} \mathrm{E}_1 E_2 F_3 F_4 \\ \mathrm{E}_1 E_3 F_2 F_4 \\ \mathrm{E}_1 E_4 F_2 F_3 \end{array}$$

$$\mathrm{E}_2 E_3 F_1 F_4$$
$$\mathrm{E}_2 E_4 F_1 F_3$$
$$\mathrm{E}_3 E_4 F_1 F_2$$

Por lo que harían esas 4 posibles combinaciones de multiplicar las probabilidades por eso tenemos ese resultado.

ejemplo; En el HUHMP se tiene una prevalencia de la enfermedad Z de $x\%$ en los pacientes, cuanto pacientes obtendremos en una muestra de j.

En la construcción del modelo binomial hemos supuesto por ejemplo que la probabilidad de exito es constante a través de los ensayos, es decir podemos obtener un exito con la misma probabilidad en cada una de las pruebas, y sobra decir que las categorías son excluyentes es decir; se presenta la enfermedad o no. Otra de las caracterisitcas importantes en la construcción del modelo es que los eventos son independientes entre sí,l igualmente el tamaño poblacional es constante, como consecuencia de la definición de probabilidad, y por supuesto el supuesto de aletoriedad de los datos.

para el ejemplo anterior, de 150 pacientes cual es la probabilidad de que exactamente 30 tengan la enfermedad Z si esta tiene una prevalencia entre los pacientes del 30%.

```
display comb(150,30)* 0.3^(30) * (1-0.3)^(150-30)

display comb(100, 60) * 0.6^(60) * (1-0.6)^(100-60)
display comb(100, 60) * 0.6^(60)* (0.4)^(40)

help binomialp
binomialp(n,k,p)
/*

Donde n significa el número de casos, k es el número de exitos
con una tasa de probabiliad p

*/
```

Ahora observemos una función que tiene stata para hacer ese calculo de manera más abreviada.

```
display binomialp(150,30,0.3)

help binomial(n,k,p)
help binomialtail(n,k,p)
```

ejemplo de mendenhall:

en un examen hay 100 preguntas, y hay 5 posibles respuestas con cada una.

¿cual es la calificación esperada para un estudiante que está adivinando cada pregunta?

Entonces el valor esperado y la desviación estándar de la variable aleatoria binomial es:

$$\mu = np$$
$$\sigma = \sqrt{npq}$$

entonces x es el número de respuestas correctas en un examen de 100 preguntas, probabilidad de éxito de obtener una respuesta correcta es de 1/5.

y su valor esperado es por lo tanto $20 = \mu$.

```
range x 0 10 11

gen pexito= binomialp(10, x, 0.5)

list x p

list x p , separator(0)
```

Como sería una variable binomial simétrica??

13.5 Ejercicio

1. para el ejemplo de la sección calcule la probabilidad de que 30 o menos pacientes padezcan la enfermedad Z
2. para el mismo ejemplo calcule la probabilidad de que 30 o mas pacientes padezcan la enfermedad Z

la distribución de poisson es utilizada en health economics,
ver ese para acabar binomial

13.6 Distribución de poisson

According to the former binomial distribution $X \sim b(p, n)$ the two parameter are the shape a form of the distribution. the poisson distribution is the case when the variable follow a binomial distribution with a $n \to \infty$

In the limit case, the occurrence of a only event is only guaranteed in the measure that the space is very small, for instance if the ocurrence of the events is simultaneous, you should not consider a Poisson distribution. the FD we can dervied of a binomial distribution in the following way $E(x) = np = \lambda$, thus:

$$\frac{\lambda}{n} = p$$

according to FD of a $x \sim b(n, p)$

$$\frac{n!}{(n-k)!k!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k}$$

$$\frac{(n-k+1)!}{n^k k!}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-k}$$

$e = \lim_{x\to\infty}(1+\frac{1}{n})^n$we must use $t = \frac{n}{k}$, and thus $\frac{n+k}{n} = 1 + \frac{k}{n}$

$$\lim_{n\to\infty} = \frac{e^{-k}\lambda^k}{k!}$$

thus a ramdon variable follow a poisson distrbution with a paramter λ $X \sim p(\lambda)$ and its FD is rewritten as:

$$p(X = x) = \frac{e^{-x}\lambda^x}{x!}$$

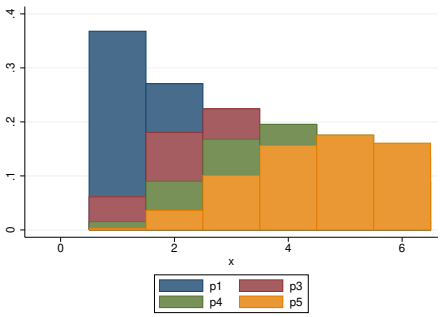


Figure 1: poisson distribution distribution (changes in mean)

13.7 Normal distribution

The normal distribution

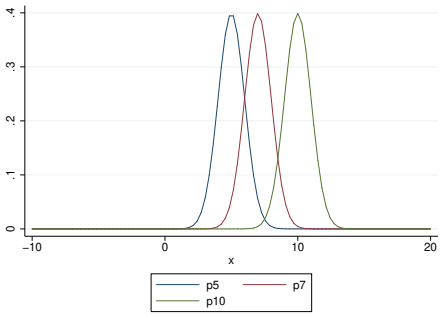


Figure 2: Normal distribution (changes in mean)

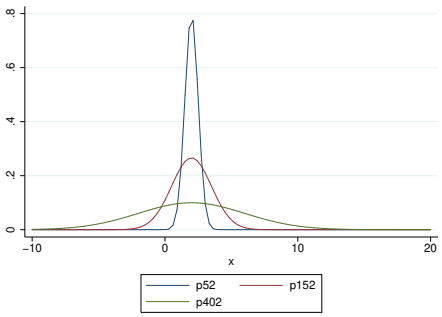


Figure 3: Normal distribution (changes in sd)

13.8 Simulations in Stata

```
* binomialp(n,k,p))
* Observing k successes, in n trials with probability p.

clear all
range x 0 6  7
local bn 1 3 4 5
foreach t of local bn{
gen b‘t’=binomialp(6,x,(‘t’/10))
}

graph twoway (bar  b1 x) (bar b3 x) (bar b4 x) (bar b5 x) , graphregion(color(white))

clear all
range x -10 20 100
* normalden(x, mean, sd)
local normal  5 7 10 // mean
gen a=normalden(x, 20, 1)

foreach t of local normal{
gen p‘t’=normalden(x, ‘t’, 1)
}

graph twoway (line p5  x) (line p7 x) (line p10 x) , graphregion(color(white))



local normal2 5 15 40 //sd
foreach t of local normal2{
gen p‘t’2=normalden(x, 2, (‘t’/10))
}

graph twoway (line p52  x) (line p152 x) (line p402 x) , graphregion(color(white))
```

proyecto poisson Ejercicio.

```
gen p= poissonnp(2.4,x)

list x p
```

Como podemos utilizar la distribución de poisson en la literatura empírica para una idea de esto véase: possion

13.9 Distribución ji-cuadrada

supongamos que tenemos un variable aletoria tal que $X \sim N(0,1)$

El concepto de grado de libertad es importante, entendemos como la probabilidad de obtener un número dada la distribución.

por ejemplo para el caso de la distribución normal estandarizada, piense que obtener un resulado cercano a uno es más probable que un 3 debido a que las mediciones a 3 desviaciones son muy improbables.

```
** Rate este  programa toca condicionarlo...
capture program drop rate
program define rate
foreach x of local 0{
gen rate‘x’ = (‘x’-‘x’[_n-1])/‘x’[_n-1]
}
end
```

Ejercicio con base en la base de datos del Banco Mundial trate de comprobar la hipótesis de convergencia: que enuncia que los países de mayor ingreso tienen menores tasas de crecimiento en comparación con los que tienen menor ingreso.