COMANDOS DE STATA REVISADOS EN ESTA CLASE

Examinar los datos:

<describe *varnames*> entrega información sobre las variables y etiquetas
<browse> permite explorar los datos
<count [if *exp*] [in *exp*]> cuenta el número de observaciones en la base de datos o que satisfagan una condición específica
<sum *varnames* [, detail]> entrega estadísticas descriptivas
<list *varnames* [in *exp*]> listado de los valores de las variables y observaciones especificadas
<codebook *varnames*> resume la codificación de las variables
<tab *varname*, [no label missing]> entrega distribuciones de frecuencia (las opciones permiten no desplegar las etiquetas y mostrar los datos perdidos)
<tab *varname1 varname2* [, row col cell chi2]> genera tablas de doble entrada (las opciones permiten obtener porcentajes de fila, columna, celda y test de independencia chi2)
<tab1 *varnames*> genera distribuciones de frecuencia separadas para cada variable

Describir los datos perdidos:

<mdesc *varlist* [if *exp*] [in *exp*]> entrega el número y proporción de casos perdidos

Examinar gráficamente los datos:

<histogram *varname*[, norm disc freq> muestra la distribución de frecuencia de una variable con barras verticales (opción norm despliega la curva de distribución normal, disc especifica que la variable es discreta y freq despliega las frecuencias)
<dotplot *varname*> muestra la distribución de frecuencia de una variable con filas de puntos horizontales
<graph box *varname*>diagrama de caja para una variable
<scatter *varname1 varname2*> gráfico de puntos para dos variables
<lowess *varname1 varname2*> línea de regresión suavizada entre dos variables
<graphmatrix *varnames*> gráfico de puntos para todas las combinaciones entre variables
<graphsave *filename* [, replace]> guarda el gráfico en un archive .gph
<graph use *filename*> abre un gráfico previamente guardado
<kdensity varname, norm> grafica la función de kernel y con la opción norm permite evaluar normalidad en la distribución de una variable

Análisis bivariados:

<correlate *varlist*> calcula correlaciones entre variables botando todas las observaciones con casos perdidos
<pwcorr *varlist* [,star]> calcula correlaciones botando solamente las observaciones con casos perdidos en una pareja de variables
<ttest *contvar*, by(*dicvar*)> test de diferencia entre las medias de contvar entre los grupos de dicvar

Administrar los datos:
<sort> ordena las observaciones en una base de datos
<order *varnames* [, opciones]> reordena las variables en una base de datos
<drop *varnames*> bota variables
<keep *varnames*> guarda variables
<drop [in *range*] [if *exp*]>bota observaciones
<keep [in *range*] [if *exp*]>guarda observaciones

DEMOSTRACIÓN PRÁCTICA

```
. use "C:\vs_chile_2005_v9.dta", clear

. *contar el número de observaciones en la base de datos

. count
 1000

.  *obtener estadísticas descriptivas

. sum lsat fsat hap

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        lsat |       992    7.243952    2.035746          1         10
        fsat |       993    5.797583    2.348282          1         10
         hap |       998    3.134269     .727734          1          4

. sum lsat-hap

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        lsat |       992    7.243952    2.035746          1         10
        fsat |       993    5.797583    2.348282          1         10
         hap |       998    3.134269     .727734          1          4

. sum age, detail

                            age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           18             18
 5%           19             18
10%           22             18       Obs                1000
25%           29             18       Sum of Wgt.        1000

50%           41                      Mean             42.931
                        Largest       Std. Dev.      16.97694
75%           55             85
90%           68             85       Variance       288.2165
95%           74             85       Skewness       .4331384
```

```
99%         82.5             85       Kurtosis       2.270827

. *listado de los valores de variables específicas para observaciones
      específicas

. list female age mstat in 1/12, sep(3)

     +--------------------------+
     | female   age       mstat |
     |--------------------------|
  1. |   Male    50      MarPar |
  2. | Female    61   DivSepWid |
  3. |   Male    30      MarPar |
     |--------------------------|
  4. | Female    30      MarPar |
  5. |   Male    72      MarPar |
  6. |   Male    71   DivSepWid |
     |--------------------------|
  7. | Female    35      MarPar |
  8. |   Male    25      MarPar |
  9. |   Male    53      MarPar |
     |--------------------------|
 10. |   Male    49   DivSepWid |
 11. |   Male    18      Single |
 12. | Female    44      MarPar |
     +--------------------------+

. *lo mismo pero con la línea default de separación (cada 5
      observaciones) y sin etiquetas para los valores

. list female age mstat in 1/12, nol

     +----------------------+
     | female   age   mstat |
     |----------------------|
  1. |      0    50       1 |
  2. |      1    61       2 |
  3. |      0    30       1 |
  4. |      1    30       1 |
  5. |      0    72       1 |
     |----------------------|
  6. |      0    71       2 |
  7. |      1    35       1 |
  8. |      0    25       1 |
  9. |      0    53       1 |
 10. |      0    49       2 |
     |----------------------|
 11. |      0    18       3 |
 12. |      1    44       1 |
     +----------------------+

. codebook educ

-------------------------------------------------------------------------------
      --------
```

```
educ        RECODE of educf (RECODE of x025 (highest educational level
      attained))
-------------------------------------------------------------------------
      --------

                  type:  numeric (byte)
                 label:  educ_l

                 range:  [1,4]                        units:  1
          unique values:  4                       missing .:  1/1000

             tabulation:  Freq.   Numeric  Label
                            149         1  No formal education or
                                           Incomplete primary school
                            282         2  Less than high school
                            388         3  High school
                            180         4  More than high school
                              1         .

. *tablas de frecuencia

. tab hap

              hap |      Freq.     Percent        Cum.
------------------+-----------------------------------
not at all happy |         14        1.40        1.40
  not very happy |        164       16.43       17.84
     quite happy |        494       49.50       67.33
      very happy |        326       32.67      100.00
------------------+-----------------------------------
           Total |        998      100.00

. *lo mismo pero incluyendo los valores perdidos

. tab hap, m

              hap |      Freq.     Percent        Cum.
------------------+-----------------------------------
not at all happy |         14        1.40        1.40
  not very happy |        164       16.40       17.80
     quite happy |        494       49.40       67.20
      very happy |        326       32.60       99.80
               . |          2        0.20      100.00
------------------+-----------------------------------
           Total |      1,000      100.00

. *tablas de doble entrada

. tab dtrust female

                      |        female
               dtrust |      Male     Female |     Total
----------------------+----------------------+----------
Most people can be tr |        65         57 |       122
Can't be. Need to be  |       378        484 |       862
```

```
---------------------+----------------------+----------
               Total |       443        541 |       984


. *tablas de doble entrada con porcentaje de fila

. tab dtrust female, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

                     |        female
              dtrust |      Male     Female |     Total
---------------------+----------------------+----------
Most people can be tr |        65         57 |       122
                     |     53.28      46.72 |    100.00
---------------------+----------------------+----------
Can't be. Need to be  |       378        484 |       862
                     |     43.85      56.15 |    100.00
---------------------+----------------------+----------
               Total |       443        541 |       984
                     |     45.02      54.98 |    100.00


. *tablas de doble entrada con porcentaje de columna

. tab dtrust female, col

+------------------+
| Key              |
|------------------|
|     frequency    |
| column percentage |
+------------------+

                     |        female
              dtrust |      Male     Female |     Total
---------------------+----------------------+----------
Most people can be tr |        65         57 |       122
                     |     14.67      10.54 |     12.40
---------------------+----------------------+----------
Can't be. Need to be  |       378        484 |       862
                     |     85.33      89.46 |     87.60
---------------------+----------------------+----------
               Total |       443        541 |       984
                     |    100.00     100.00 |    100.00


. *tablas de doble entrada con porcentaje de celda sobre el total

. tab dtrust female, cel
```

```
+-----------------+
| Key             |
|-----------------|
|    frequency    |
| cell percentage |
+-----------------+
```

dtrust	female Male	female Female	Total
Most people can be tr	65	57	122
	6.61	5.79	12.40
Can't be. Need to be	378	484	862
	38.41	49.19	87.60
Total	443	541	984
	45.02	54.98	100.00

```
. *tablas de doble entrada con todos los porcentajes y un test de
      independencia chi2, sin la leyenda (key)

. tab dtrust female, row col cel chi nokey
```

dtrust	female Male	female Female	Total
Most people can be tr	65	57	122
	53.28	46.72	100.00
	14.67	10.54	12.40
	6.61	5.79	12.40
Can't be. Need to be	378	484	862
	43.85	56.15	100.00
	85.33	89.46	87.60
	38.41	49.19	87.60
Total	443	541	984
	45.02	54.98	100.00
	100.00	100.00	100.00
	45.02	54.98	100.00

```
          Pearson chi2(1) =   3.8373   Pr = 0.050

. *tabulaciones de distribuciones de frecuencia separadas para cada
      variable, desplegando los casos perdidos

. tab1 dtrust female, m

-> tabulation of dtrust

                           dtrust |      Freq.     Percent        Cum.
----------------------------------+-----------------------------------
```

```
       Most people can be trusted |        122       12.20       12.20
Can't be. Need to be very careful |        862       86.20       98.40
                                . |         16        1.60      100.00
----------------------------------+-----------------------------------
                            Total |      1,000      100.00

-> tabulation of female

     female |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |        449       44.90       44.90
     Female |        551       55.10      100.00
------------+-----------------------------------
      Total |      1,000      100.00


. *ocupando condiciones lógicas: < es menor que, > es mayor que, == es
      igual a, <= es menor o igual a, >= es mayor o igual a, != no es
      igual a

. *varias condiciones lógicas se pueden conectar ocupando los siguientes
      símbolos:

. * &significa "y"

. * | significa "o"

. * (paréntesis agrupan las condiciones)

. codebook shlth

-------------------------------------------------------------------------------
      --------
shlth
      (unlabeled)
-------------------------------------------------------------------------------
      --------

                  type:  numeric (byte)
                 label:  shlth_l

                 range:  [2,5]                        units:  1
         unique values:  4                        missing .:  0/1000

            tabulation:  Freq.   Numeric  Label
                            52         2  poor
                           277         3  fair
                           475         4  good
                           196         5  very good

. sum shlth

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       shlth |      1000       3.815    .8046256          2          5
```

```
. sum shlth if female==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       shlth |       551    3.735027    .8149891          2          5

. sum shlth if female==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       shlth |       449     3.91314    .7814999          2          5

. sum shlth if female==0 &inc>=9 &inc<=10

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       shlth |         9    4.333333    .7071068          3          5

. *obtener ayuda en Stata, por ejemplo, para aprender a contar
      observaciones perdidas y válida

. help count

. count if inc<11
  946

. count if inc>10
   54

. count if inc==.
   53

. *ojo que Stata trata a los valores perdidos como las magnitudes más
      grandes posibles, partiendo por .y seguido por .a hasta .z

. *buscar keywords

. search missing

Keyword search

        Keywords:  missing
          Search:  (1) Official help files, FAQs, Examples, SJs, and STBs

Search of official help files, FAQs, Examples, SJs, and STBs


[I]     missing values  . . . . . . . . . . Quick reference for missing
      values
        (help missing values)

[U]     Chapter 12  . . . . . . . . . . . . . . . . . . . . . . . . . .
      Data
        (help datatypes, strings, label, missing, notes, format)
```

```
 --Break--
r(1);

. list inc if mi(inc) in 1/10, sep(0)

     +-----+
     | inc |
     |-----|
  1. |  .e |
  5. |   . |
     +-----+

. count if !mi(inc)
  946

. count if mi(inc)
   54

. help mdesc

. net search mdesc
(contacting http://www.stata.com)

1 package found (Stata Journal and STB listed first)
---------------------------------------------------

mdesc from http://fmwww.bc.edu/RePEc/bocode/m
    'MDESC': module to tabulate prevalence of missing values / Produces a
    table with the number of missing values, total number / of cases, and
    percent missing for each variable in varlist. mdesc / works with both
    numeric and character variables.  / KW: missing values / KW: data

. mdesc inc

    Variable    |     Missing          Total     Percent Missing
----------------+-----------------------------------------------
           inc |          54          1,000            5.40
----------------+-----------------------------------------------


. *Examinar gráficamente los datos. Siempre hay dos ejes: vertical (y) y
      horizontal (x). Para examinar un gráfico primero hay que entender
      qué representa cada eje.

. *Eje vertical (y) de un histograma automáticamente muestra la densidad
      de las respuestas.

. histogram mstat
(bin=29, start=1, width=.06896552)

. histogram mstat, disc
(start=1, width=1)

. histogram mstat, disc freq
(start=1, width=1)
```

```
. tab mstat

  RECODE of |
       x007 |
   (marital |
    status) |      Freq.     Percent        Cum.
------------+-----------------------------------
     MarPar |        568       56.97       56.97
  DivSepWid |        141       14.14       71.11
     Single |        288       28.89      100.00
------------+-----------------------------------
      Total |        997      100.00

. histogram age, norm
(bin=29, start=18, width=2.3103448)

. *Dotplots muestran información similar, pero con filas de puntos en el
      eje horizontal(x).

. dotplot age

. *Los gráficos de kernel permiten evaluar la normalidad en la
      distribución de una variable (skewness and kurtosis).

. kdensity age, norm

. *El diagrama de caja indica la mediana con la línea horizontal del
      medio, el percentil 75 y 25 con las cajas, los valores adyacentes
      máximos y mínimos con loscorchetes y los outliers con puntos.

. graph box age

. preserve

. replace age =100 in 1
(1 real change made)

. graph box age

. restore

. *Los scatterplots ilustran la relación entre dos variables con puntos.

. scatter ttins age

. *lowess es muy útil para ilustrar la dirección  y linealidad de la
      relación entre dos variables.

. lowess ttins age

. *ejemplo ficticio de una relación positiva y lineal

. gen age2 = age+inc
(54 missing values generated)
```

```
. scatter age age2

. lowess age age2

. *graph matrix permite visualizar múltiples relaciones bivariadas al
      mismo tiempo

. graph matrix ttins age swb inc

. *Guardar y desplegar un gráfico. Para copiar y pegar en Word ocupar el
      botón de la derecha del mouse.

. graph save agegraph.gph, replace
(file agegraph.gph saved)

. *Calcular matriz de correlaciones entre variables. Cualquier
      observación con casos perdidos es eliminada (listwisedeletion).

. corr ttins age swb inc
(obs=946)

             |    ttins      age      swb      inc
-------------+------------------------------------
       ttins |   1.0000
         age |   0.1677   1.0000
         swb |   0.0167  -0.1644   1.0000
         inc |   0.0443  -0.1793   0.4237   1.0000


. *Calcular matriz de correlaciones entre variables. Cualquier
      observación con casos perdidos en una pareja de variables es
      eliminada (pairwisedeletion).
. pwcorr ttins age swb inc, obs

             |    ttins      age      swb      inc
-------------+------------------------------------
       ttins |   1.0000
             |     1000
             |
         age |   0.1658   1.0000
             |     1000     1000
             |
         swb |   0.0234  -0.1538   1.0000
             |     1000     1000     1000
             |
         inc |   0.0443  -0.1793   0.4237   1.0000
             |      946      946      946      946
             |

. *lo mismo, pero incluyendo la significancia estadística.

. pwcorr ttins age swb inc, sig star(.05)

             |    ttins      age      swb      inc
```

```
-------------+------------------------------------
       ttins |   1.0000
             |
             |
         age |   0.1658*  1.0000
             |   0.0000
             |
         swb |   0.0234  -0.1538*  1.0000
             |   0.4597   0.0000
             |
         inc |   0.0443  -0.1793*  0.4237*  1.0000
             |   0.1734   0.0000   0.0000
             |

. *la relación bivariada entre una variable continua y otra dicotómica se
      puede explorar ocupando t-tests para diferencia de medias entre
      grupos

. ttest age, by(female)

Two-sample t test with equal variances
------------------------------------------------------------------------------
      -----
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf.
      Interval]
---------+--------------------------------------------------------------------
      -----
    Male |     449    42.94655    .8289185    17.56447     41.3175
      44.5756
  Female |     551    42.91833    .7028716    16.49878    41.53769
      44.29897
---------+--------------------------------------------------------------------
      -----
combined |    1000      42.931    .5368579    16.97694     41.8775
      43.9845
---------+--------------------------------------------------------------------
      -----
    diff |            .0282176    1.079886               -2.090889
      2.147324
------------------------------------------------------------------------------
      -----
    diff = mean(Male) - mean(Female)                                 t =
      0.0261
Ho: diff = 0                                     degrees of freedom =
      998

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff >
      0
 Pr(T < t) = 0.5104         Pr(|T| > |t|) = 0.9792          Pr(T > t) =
      0.4896

. ttest chur, by(female)

Two-sample t test with equal variances
```

```
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf.
      Interval]
---------+--------------------------------------------------------------------
    Male |     425    20.36706    2.300835    47.43293     15.8446
      24.88952
  Female |     542    32.32472    2.508347    58.39656    27.39743
      37.25202
---------+--------------------------------------------------------------------
combined |     967    27.06929    1.741463    54.15363     23.6518
      30.48677
---------+--------------------------------------------------------------------
    diff |           -11.95766    3.489353               -18.80526   -
      5.110069
------------------------------------------------------------------------------
    diff = mean(Male) - mean(Female)                                 t =  -
      3.4269
Ho: diff = 0                                     degrees of freedom =
      965

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff >
      0
 Pr(T < t) = 0.0003         Pr(|T| > |t|) = 0.0006          Pr(T > t) =
      0.9997

. *para reordenar las observaciones en la base de datos ocupamos el
      comando sort. Por ejemplo la variable swb va de 1 a 8 y podemos
      reordenar las observaciones de acuerdo a ese puntaje.

. browse

. sort swb

. browse

. *también podemos ordenar las variables de la base de datos

. order swb age age2

. browse

. *el comando drop permite botar variables

. drop age2 year

. *el comando keep permite elegir qué variables conservar en la base de
      datos

. keep age-female
```

```
. *los mismos comandos se pueden utilizar para botar o conservar
      observaciones específicas

. drop in 20/1000
(981 observations deleted)

. browse

. keep if age>=50
(12 observations deleted)

. browse

. keep if namea!="Chile"
(7 observations deleted)

. count
    0
```

TAREA PARA LA SIGUIENTE CLASE

Ver los tres últimos videos de IDRA-UCLA (managing data, analyzing data, general information):
http://www.ats.ucla.edu/stat/stata/notes/default.htm

Como alternativa se pueden revisar los 11 videos de UNAV (STATA 2.1 a STATA 4.3). Son los tópicos similares a los anteriores descritos en español:
http://www.unav.edu/departamento/preventiva/recursos_bioestadistica

Imprimir las notas de las clases antes de ver el video e intentar replicar los procedimientos realizados por la profesora. Guardar el .log file y enviarlo al ayudante al menos una hora antes de la siguiente clase.