Curs bàsic d'Anàlisi de dades amb Stata

Sessió 3

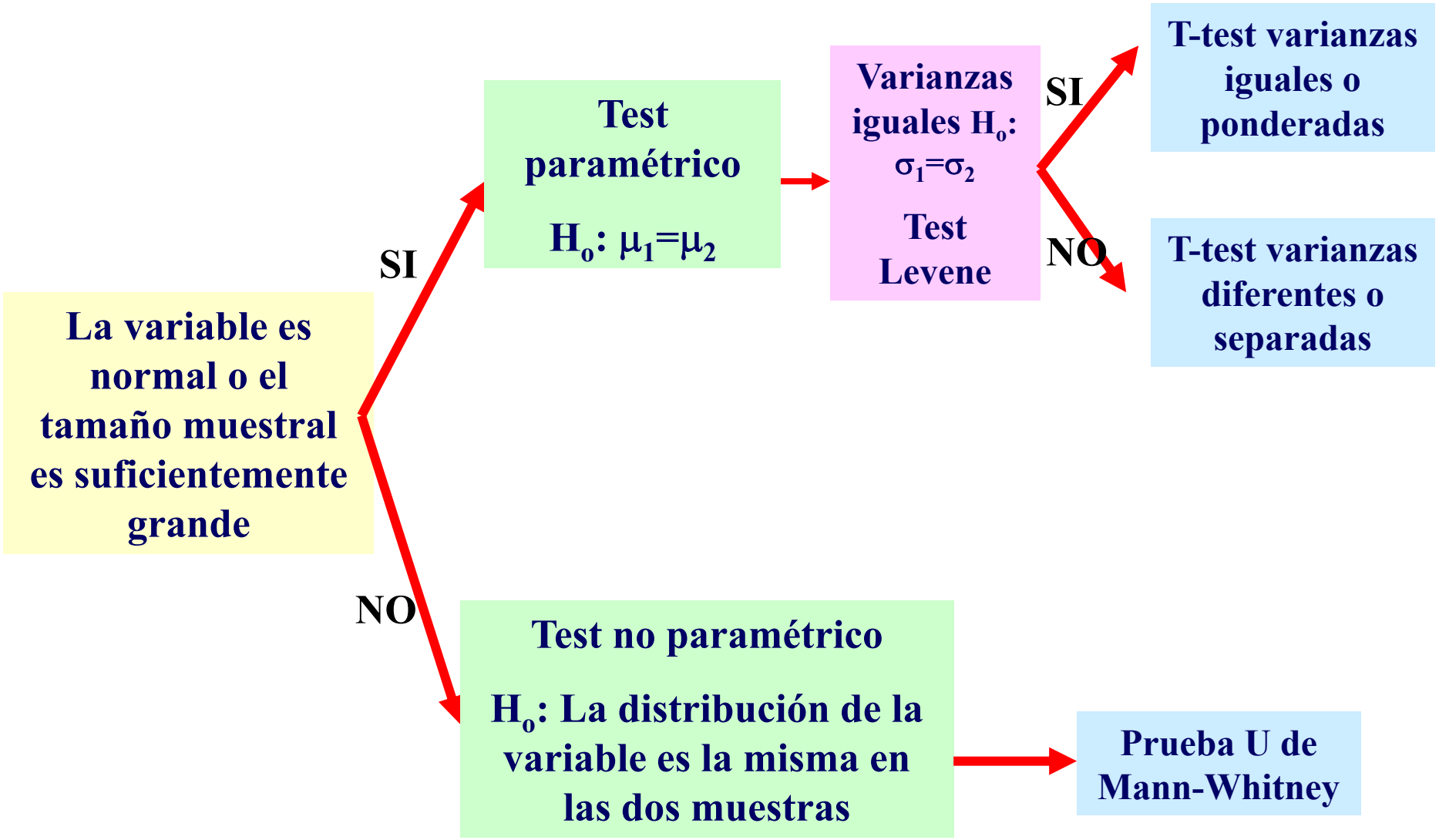
- Estimació i contrast d' hipòtesi
  - **Test per una mostra t-test**
  - **Test per 2 mostres. T-test- Mann-Whitney**
  - **Test per 3 o mes mostres: Anova, Kruskal Wallis**
  - **Probes de Normalitat**
  - **Test per variables qualitatives: Ji-cuadrat**
- Correlació i Regressió
  - **Gràfics de dispersió**
  - **Introducció a la Correlació**
  - **Introducció a la regressió lineal simple**
  - **Exercici Pràctic**

Base dades MUNS mortalitat en UCI

- Quants morts hi ha, quin percentatge
- Quina es la mitjana d'edat dels vius i dels morts
- Quina es la mediana de la gravetat TISS i SAPS en els morts
- Es el % de universitaris en la mostra mes gran en els morts que en els vius

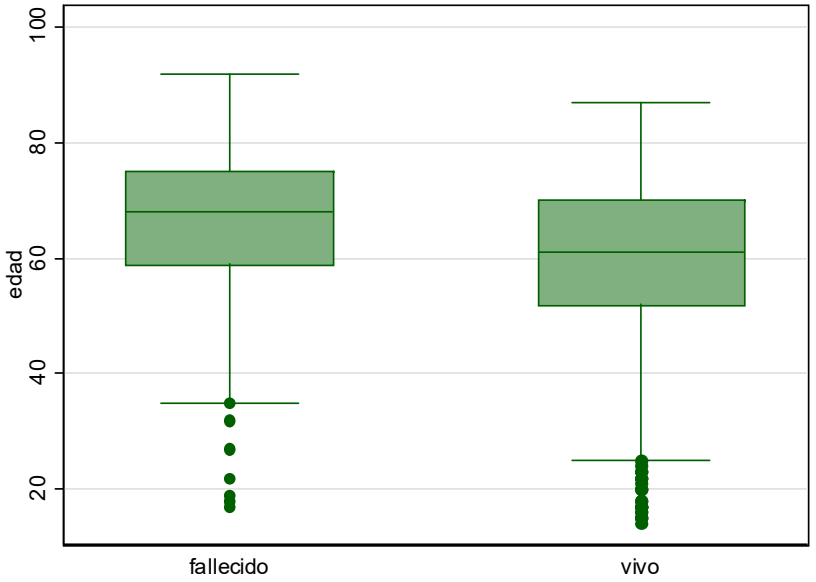
Relación entre variable cuantitativasegún los niveles de una variable cualitativa

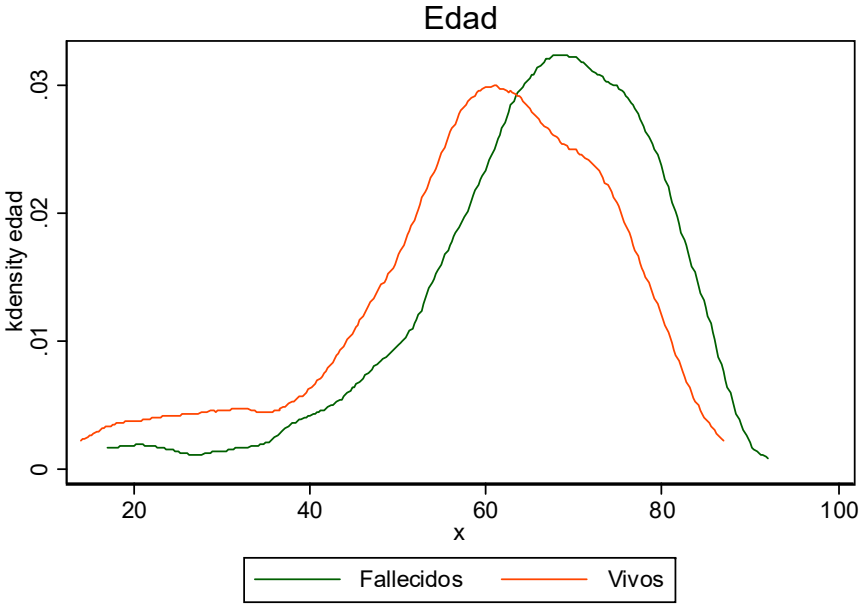
2 Muestras independientes



Ejemplo: Comparación de edat en mortalidad a las dos semanas de salir en UCI

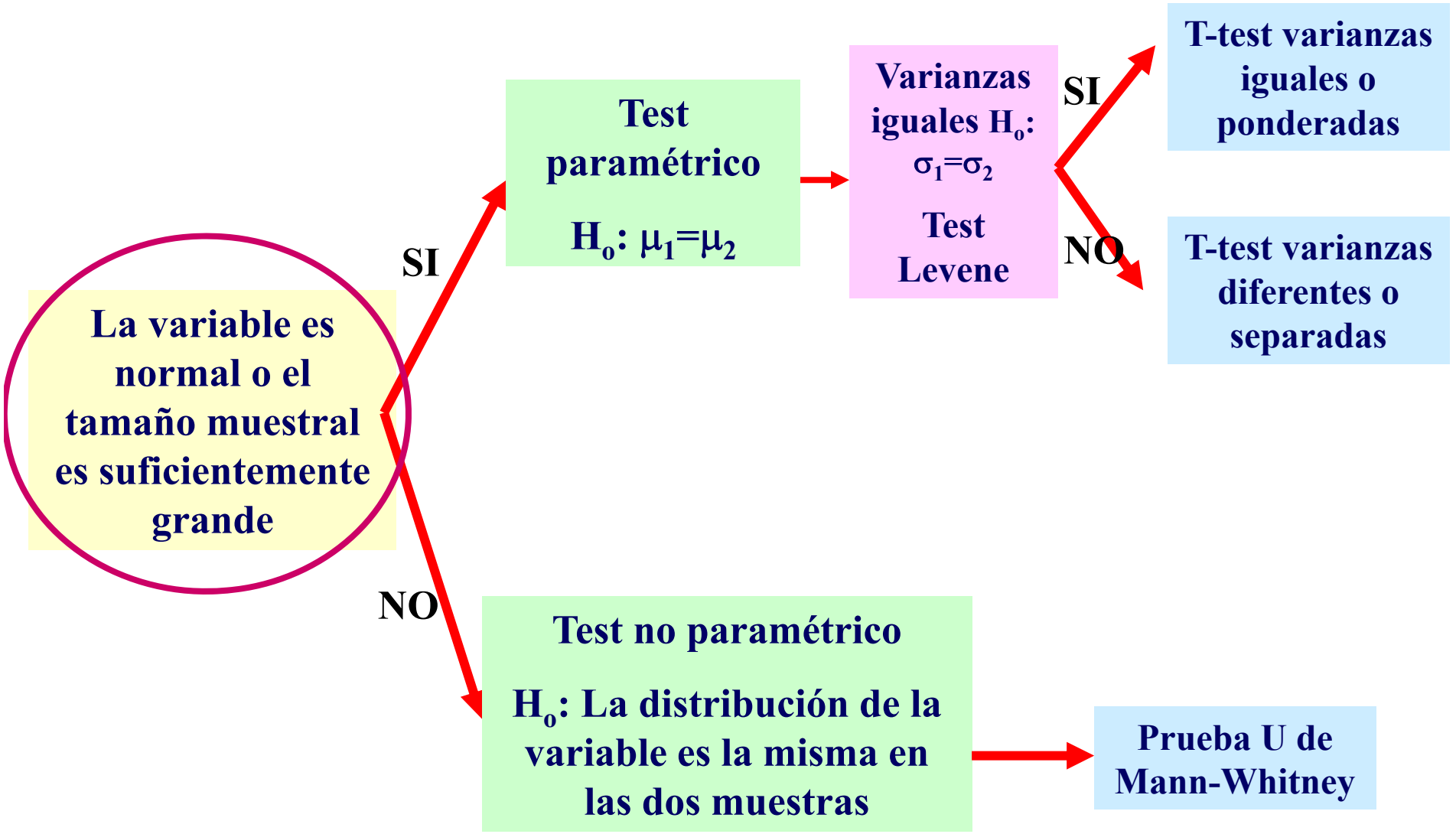
```
graph box edad, over(mortdos1)
```






```
twoway kdensity edad if mortdos1==1 ///
    || kdensity edad if mortdos1==2 ///
    ||,  legend( label( 1 Fallecidos ) ///
          label(2 Vivos) ) title(Edad)
```

2 Muestras independientes



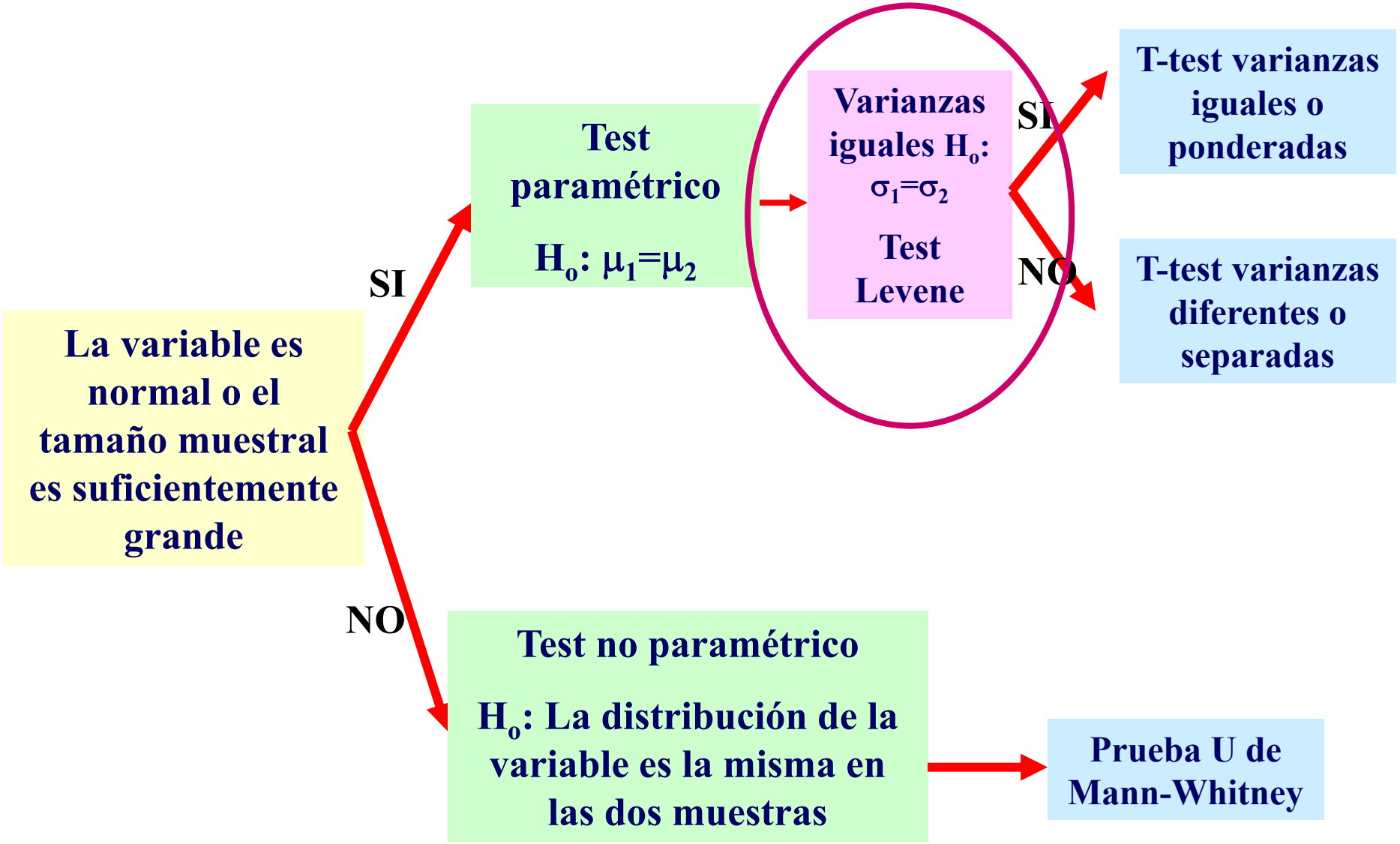
Comparación de normalidad

- Test de normalidad (Saphiro-Wilk, Shaphiro-Francia,Skeness/Kurtosis)

```
swilk var
sfrancia var
sktest var
```

```
. swilk edad
                    Shapiro-Wilk W test for normal data

    Variable |    Obs       W           V         z       Prob>z
-------------+-------------------------------------------------
        edad |    844    0.94565     29.377     8.312    0.00000

. sfrancia edad
                   Shapiro-Francia W' test for normal data

    Variable |    Obs       W'          V'         z       Prob>z
-------------+-------------------------------------------------
        edad |    844    0.94628     30.985     7.080    0.00001

. sktest edad
                     Skewness/Kurtosis tests for Normality
                                                      ------- joint ------
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------
        edad |    844      0.0000         0.0020            .          0.0000
```

2 Muestras independientes



Comparación de varianzas

- Test de Barlett

sdtest *var1*,**by(***vargrupo***)**

```
. sdtest edad,by(mortdos1)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
fallecid |     189    65.66138    .9846041    13.53607    63.71908    67.60367
    vivo |     554    58.87365    .6462106    15.20999    57.60432    60.14297
---------+--------------------------------------------------------------------
combined |     743    60.60027    .5534556    15.08611    59.51374    61.68679
------------------------------------------------------------------------------
    ratio = sd(fallecid) / sd(vivo)                               f =   0.7920
Ho: ratio = 1                                  degrees of freedom = 188, 553

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.0291         2*Pr(F < f) = 0.0583          Pr(F > f) = 0.9709
```

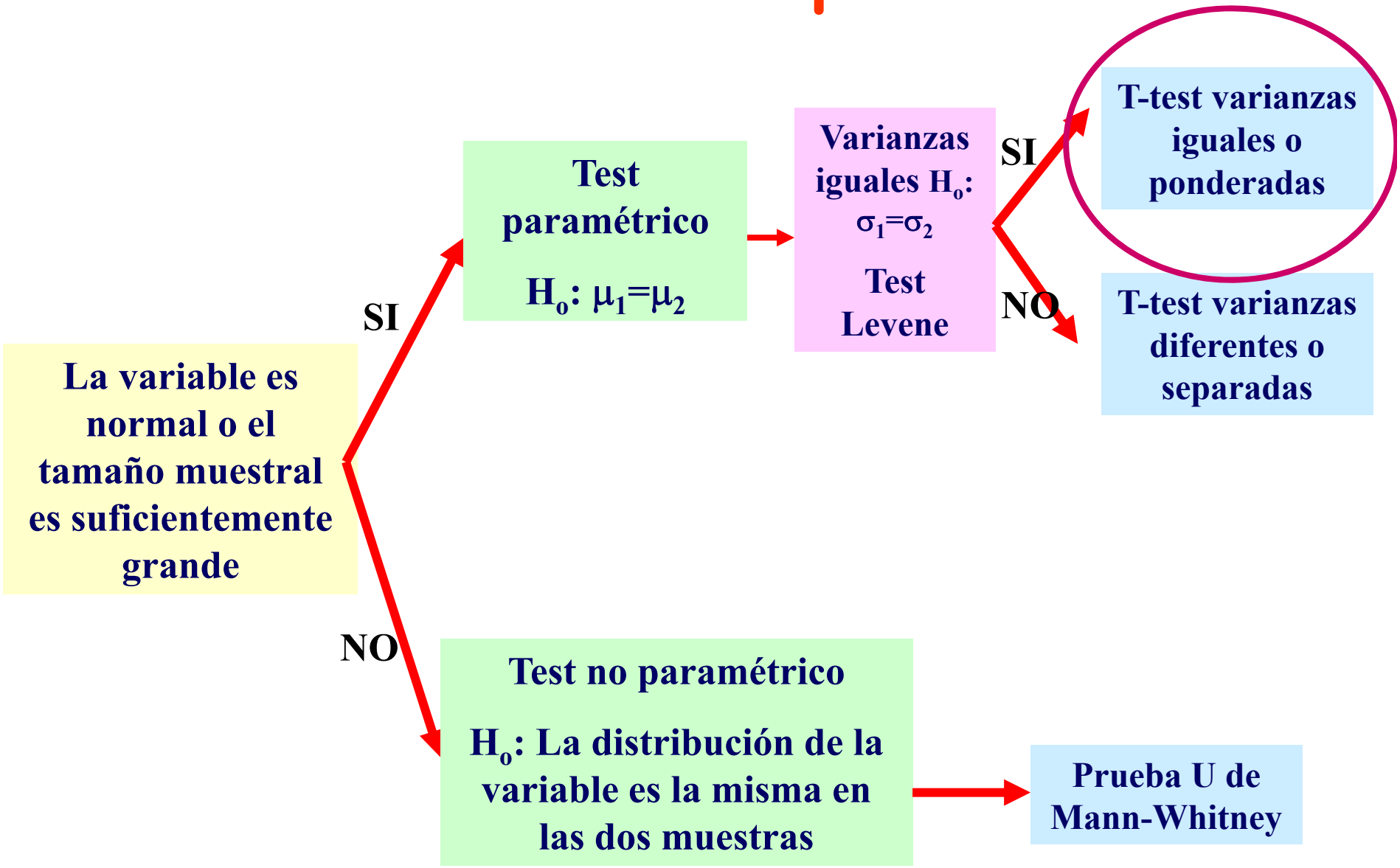
Comparación de varianzas

- Test de Levene y variaciones (+ 2 grupos)

  robvar *var1*,**by(***vargrupo***)**

```
. robvar edad,by(mortdos1)

 mortalidad |
   tras dos |           Summary of edad
      meses |        Mean   Std. Dev.       Freq.
------------+------------------------------------
  fallecido |   65.661376   13.536068         189
       vivo |   58.873646   15.209992         554
------------+------------------------------------
      Total |   60.600269   15.086109         743

W0  = 1.6295302   df(2, 841)     Pr > F = .19663994

W50 = 1.5424108   df(2, 841)     Pr > F = .21446926

W10 = 1.5807144   df(2, 841)     Pr > F = .20643891
```

2 Muestras independientes



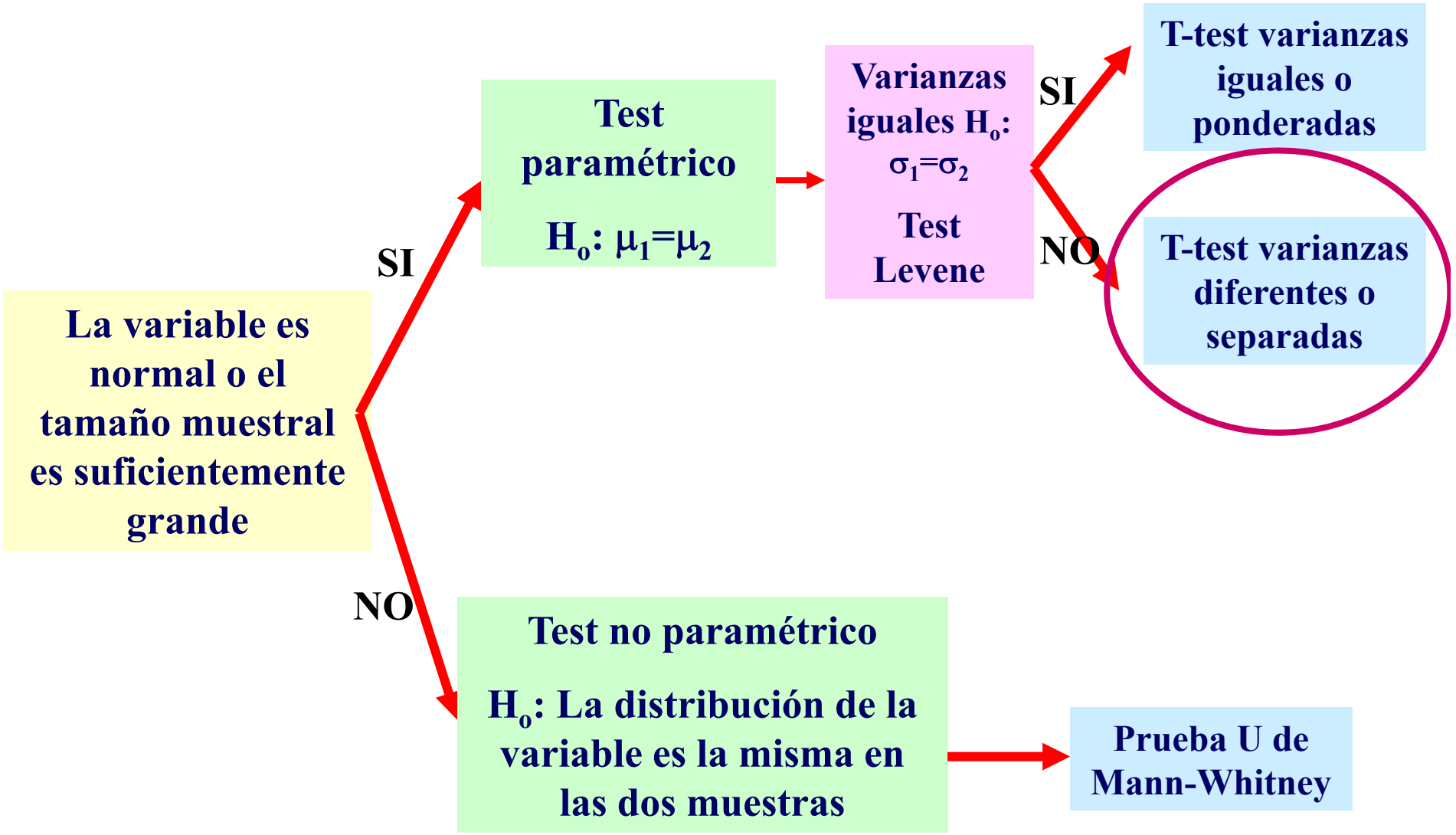
Comparación de medias

- T-test para varianzas iguales

  ttest *var1*,**by(***vargrupo***)**

```
ttest edad,by(mortdos1)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
fallecid |     189    65.66138    .9846041    13.53607    63.71908    67.60367
    vivo |     554    58.87365    .6462106    15.20999    57.60432    60.14297
---------+--------------------------------------------------------------------
combined |     743    60.60027    .5534556    15.08611    59.51374    61.68679
---------+--------------------------------------------------------------------
    diff |            6.787729    1.246996                4.339663    9.235796
------------------------------------------------------------------------------
    diff = mean(fallecid) - mean(vivo)                            t =   5.4433
Ho: diff = 0                                     degrees of freedom =      741

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

2 Muestras independientes



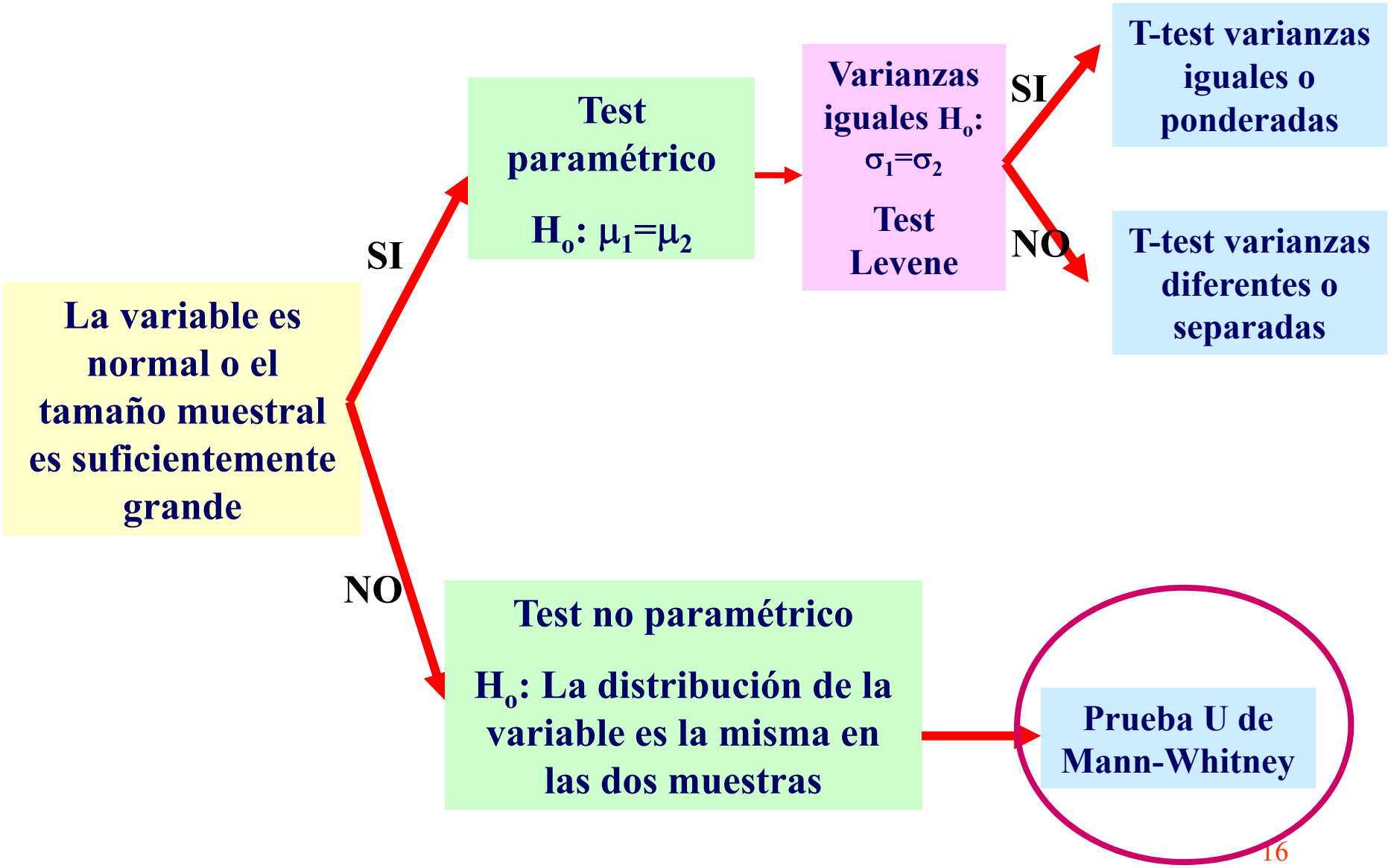
Comparación de medias

- T-test para varianzas diferentes o separadas

  ttest *var1*,**by(***vargrupo***) unequal**

```
test edad,by(mortdos1) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
fallecid |     189    65.66138    .9846041    13.53607    63.71908    67.60367
    vivo |     554    58.87365    .6462106    15.20999    57.60432    60.14297
---------+--------------------------------------------------------------------
combined |     743    60.60027    .5534556    15.08611    59.51374    61.68679
---------+--------------------------------------------------------------------
    diff |            6.787729    1.177724                4.47169    9.103769
------------------------------------------------------------------------------
    diff = mean(fallecid) - mean(vivo)                            t =   5.7634
Ho: diff = 0                     Satterthwaite's degrees of freedom =  362.009

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

2 Muestras independientes



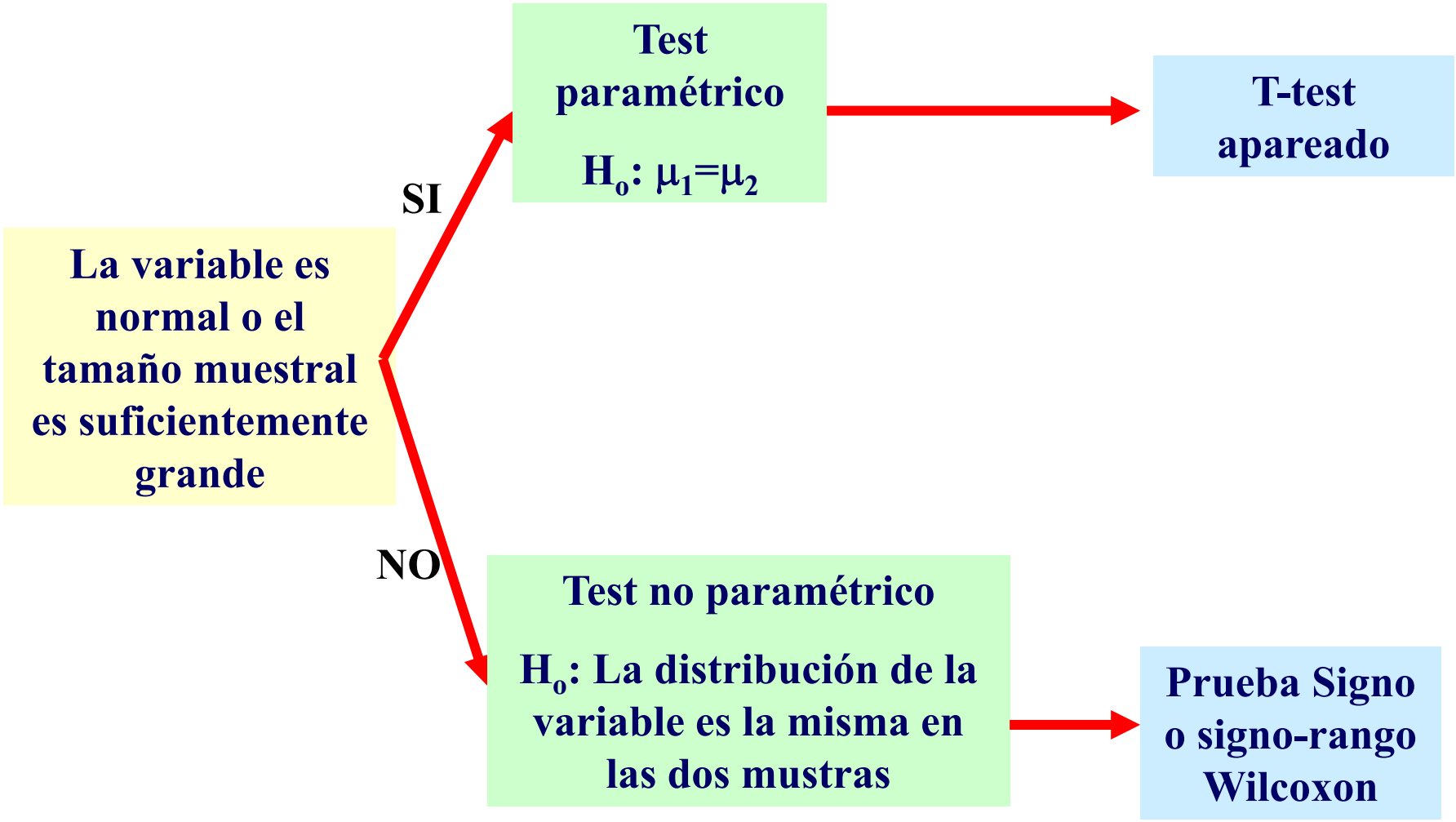
Comparación no paramétrica

- Prueba Suma-rango de Wilcoxon o U de Mann-Witney

  ranksum *var1*,**by(***vargrupo***)**

```
ranksum edad,by(mortdos1)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

    mortdos1 |      obs    rank sum    expected
-------------+---------------------------------
   fallecido |      189       85060       70308
        vivo |      554      191336      206088
-------------+---------------------------------
    combined |      743      276396      276396

unadjusted variance  6491772.00
adjustment for ties    -3990.77
                     ----------
adjusted variance    6487781.23

Ho: edad(mortdos1==fallecido) = edad(mortdos1==vivo)
              z =   5.792
    Prob > |z| =   0.0000
```

2 Muestras dependientes



Comparación de medias

- T-test apareado

  ttest *var1=var2*

```
. ttest tiss_20= saps_10

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 tiss_20 |     828   -.1763285    .3729507    10.73165   -.9083699    .5557129
 saps_10 |     828   -.6400966    .1660687    4.778624   -.9660623    -.314131
---------+--------------------------------------------------------------------
    diff |     828    .4637681    .3000386    8.633608   -.1251587    1.052695
------------------------------------------------------------------------------
     mean(diff) = mean(tiss_20 - saps_10)                          t =   1.5457
 Ho: mean(diff) = 0                           degrees of freedom =      827

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9387         Pr(|T| > |t|) = 0.1226          Pr(T > t) = 0.0613

.
```

Comparación no paramétrica

- Prueba signo-rango de Wilcoxon

signrank *var1*,**by(***vargrupo***)**

```
signrank tiss_20= saps_10

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+--------------------------------
    positive |      310      155982      171015
    negative |      470      186048      171015
        zero |       48        1176        1176
-------------+--------------------------------
         all |      828      343206      343206

unadjusted variance    47391029
adjustment for ties  -66284.125
adjustment for zeros      -9506
                     ----------
adjusted variance      47315238

Ho: tiss_20 = saps_10
              z =  -2.185
    Prob > |z| =   0.0289
```

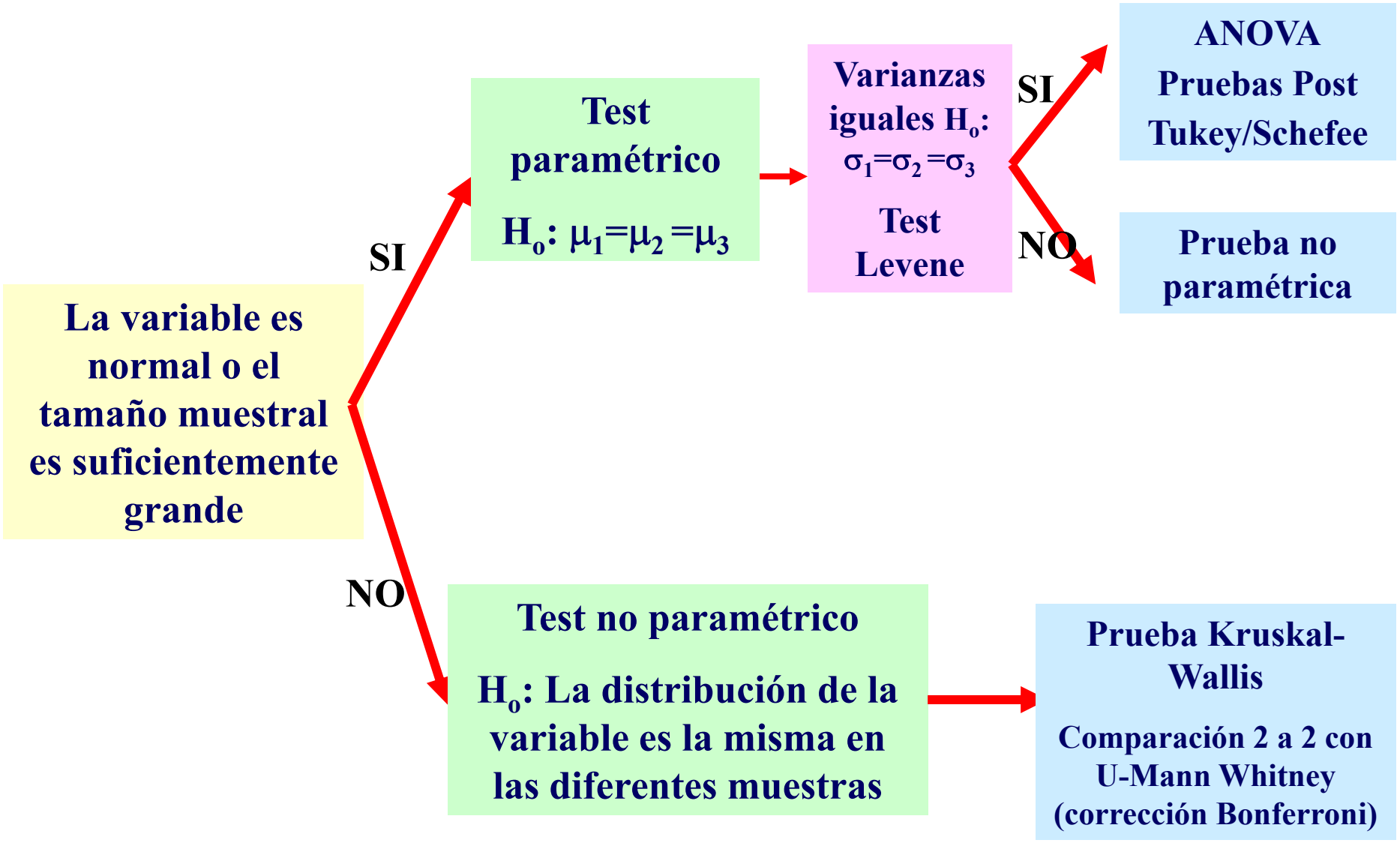
Comparación no paramétrica

- Prueba signo

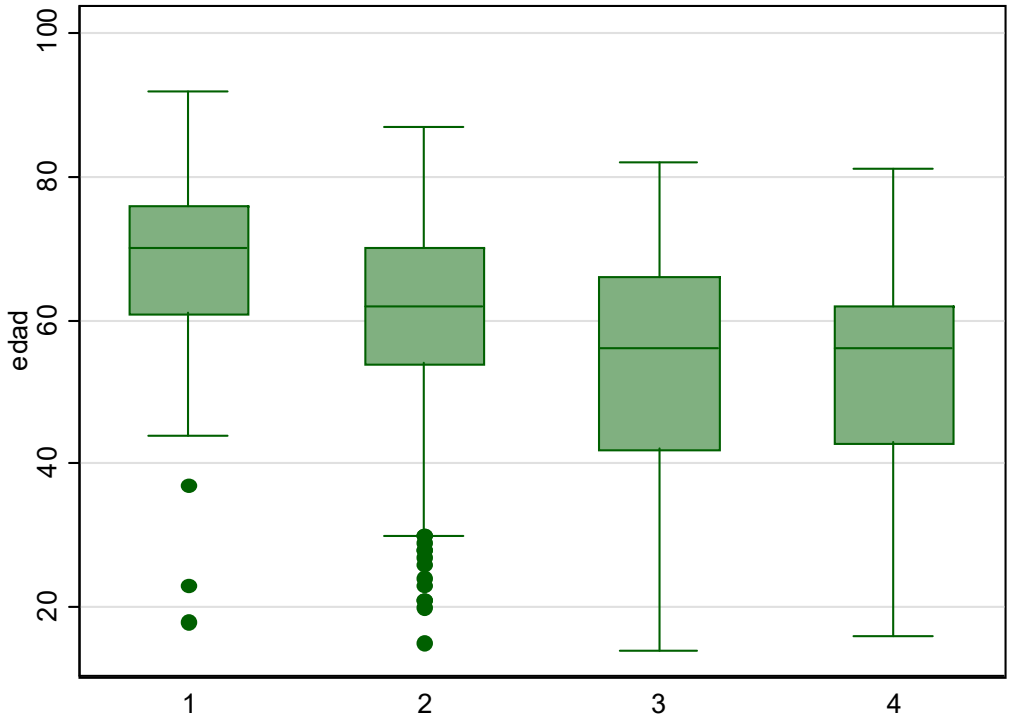
signtest *var1 = var2*

```
 signtest tiss_20= saps_10

Sign test
        sign |    observed    expected
-------------+------------------------
    positive |         310         390
    negative |         470         390
        zero |          48          48
-------------+------------------------
         all |         828         828

One-sided tests:
  Ho: median of tiss_20 - saps_10 = 0 vs.
  Ha: median of tiss_20 - saps_10 > 0
      Pr(#positive >= 310) =
         Binomial(n = 780, x >= 310, p = 0.5) =  1.0000
  Ho: median of tiss_20 - saps_10 = 0 vs.
  Ha: median of tiss_20 - saps_10 < 0
      Pr(#negative >= 470) =
         Binomial(n = 780, x >= 470, p = 0.5) =  0.0000
Two-sided test:
  Ho: median of tiss_20 - saps_10 = 0 vs.
  Ha: median of tiss_20 - saps_10 != 0
      Pr(#positive >= 470 or #negative >= 470) =
         min(1, 2*Binomial(n = 780, x >= 470, p = 0.5)) =  0.0000
```

>2 Muestras independientes



Ejemplo: Comparación de la edad en función del nivel educativo de enfermos en UCI



Comparación de medias

- ANOVA

oneway *var1 vargrupo* **,tabulate means standard bonferroni scheffe**

```
. oneway edad educacio, tabulate means standard bonferroni scheffe

   nivel de |     Summary of edad
   estudios |        Mean   Std. Dev.
------------+------------------------
          1 |    68.18617   11.489238
          2 |   60.620098   12.924095
          3 |   52.067485   18.161528
          4 |   52.470588   15.314507
------------+------------------------
      Total |   60.141975   15.069459

                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups        25887.11       3   8629.03668     44.07     0.0000
 Within groups      157827.563     806   195.815835
------------------------------------------------------------------------
    Total           183714.673     809   227.088594

Bartlett's test for equal variances:  chi2(3) =  44.7951  Prob>chi2 = 0.000
```

Comparación de medias

- ANOVA

oneway *var1 vargrupo* **,tabulate means estándar bonferroni scheffe**

```
. oneway edad educacio, tabulate means standard bonferroni scheffe


     Comparison of edad by nivel de estudios
                                (Bonferroni)
Row Mean-|
Col Mean |          1          2          3
---------+------------------------------------
       2 |   -7.56607
         |      0.000
         |
       3 |   -16.1187   -8.55261
         |      0.000      0.000
         |
       4 |   -15.7156   -8.14951    .403104
         |      0.000      0.001      1.000

                  Comparison of edad by nivel de estudios
                                  (Scheffe)
Row Mean-|
Col Mean |          1          2          3
---------+------------------------------------
       2 |   -7.56607
         |      0.000
         |
       3 |   -16.1187   -8.55261
         |      0.000      0.000
         |
       4 |   -15.7156   -8.14951    .403104
         |      0.000      0.002      0.998
```

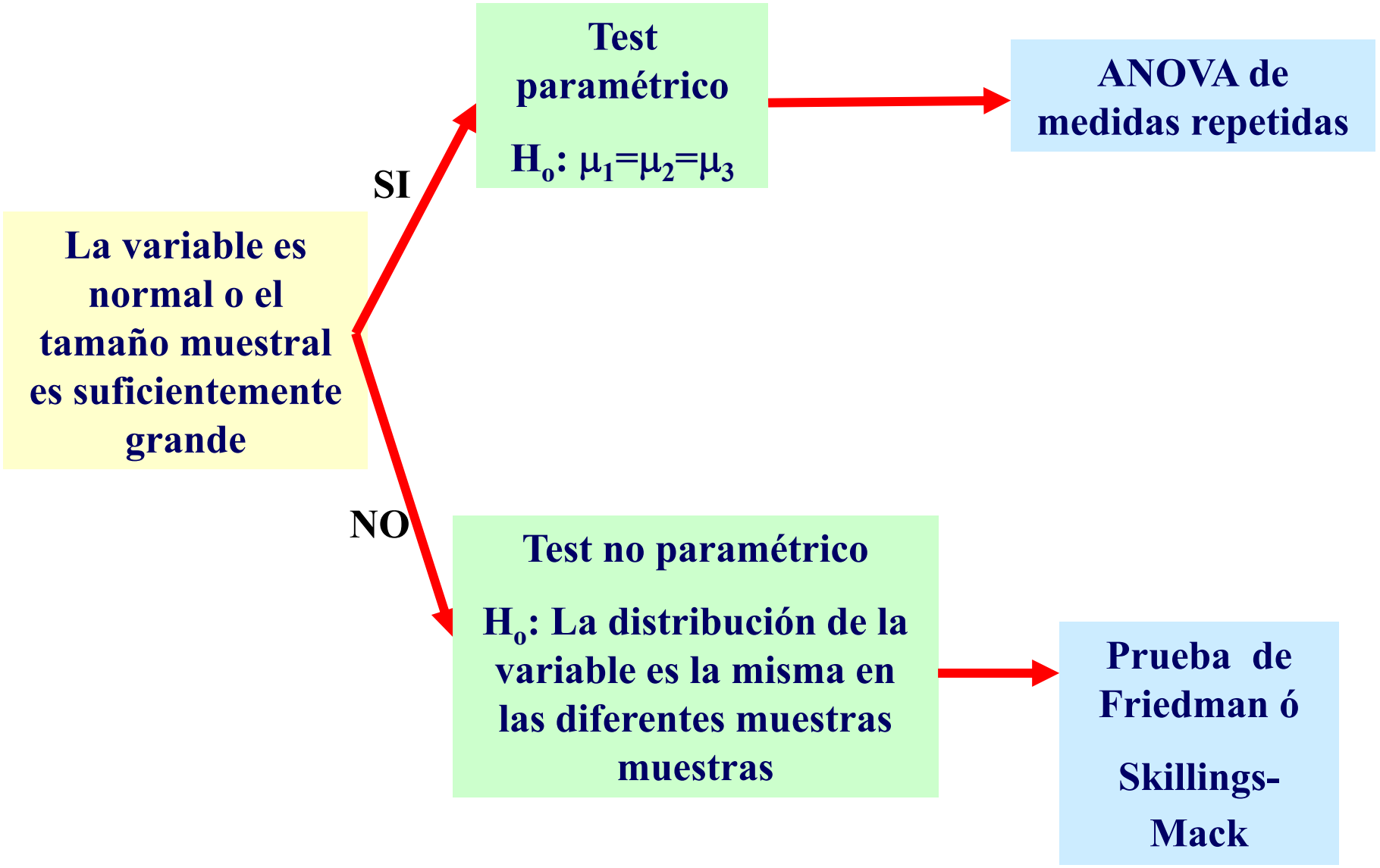
Comparación no paramétrica

- Prueba Kruskal-Wallis

kwallis *var1*,**by(***vargrupo***)**

```
kwallis edad,by(educacio)

Test: Equality of populations (Kruskal-Wallis test)

  +---------------------------+
  | educacio | Obs |  Rank Sum |
  |----------+-----+-----------|
  |        1 | 188 | 101002.50 |
  |        2 | 408 | 163864.00 |
  |        3 | 163 |  49147.50 |
  |        4 |  51 |  14441.00 |
  +---------------------------+

chi-squared =   105.860 with 3 d.f.
probability =     0.0001

chi-squared with ties =   105.923 with 3 d.f.
probability =     0.0001
```

>2 Muestras dependientes



Comparación de medias

- ANOVA de medidas repetidas. Datos “long2

  anova *var1 varmedicion id*, **repeated(***medicion***)**

```
. anova tas medicion numero, repeated(medicion)

                           Number of obs =       639     R-squared     =  0.6067
                           Root MSE      =   10.4854     Adj R-squared =  0.5582

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  96330.6179    70  1376.15168       12.52     0.0000
                         |
                medicion |  11716.8386    11  1065.16714        9.69     0.0000
                  numero |   82921.013    59   1405.4409       12.78     0.0000
                         |
                Residual |  62448.3054   568    109.9442
              -----------+----------------------------------------------------
                   Total |  158778.923   638  248.869786
```

Comparación de medias

- ANOVA de medidas repetidas. Datos "long"

anova *var1 varmedicion id*, **repeated(***medicion***)**

```
. anova tas medicion numero, repeated(medicion)
.....

Between-subjects error term:  numero
                     Levels:  60          (59 df)
     Lowest b.s.e. variable:  numero

Repeated variable: medicion
                                    Huynh-Feldt epsilon        =  0.8579
                                    Greenhouse-Geisser epsilon =  0.7316
                                    Box's conservative epsilon =  0.0909

                                      ------------ Prob > F ------------
              Source |     df      F    Regular    H-F      G-G       Box
           ----------+-----------------------------------------------------
            medicion |     11     9.69   0.0000   0.0000   0.0000   0.0030
            Residual |    568
           -----------------------------------------------------------------
```

Comparación no paramètrica

- Test de Skillings-Mack. Datos “long”

skilmack *var1* **,i(***ident***) repeated(***medicion***)**

```
skilmack tas ,id(numero) repeated(medicion)

Weighted Sum of Centered Ranks

 medicion |      N  WSumCRank         SE    WSum/SE
----------+-------------------------------------------
 1        |     57     136.81      24.06       5.69
 2        |     57     102.14      24.06       4.24
 3        |     55      40.97      24.02       1.71
 4        |     53     -22.63      23.77      -0.95
 5        |     52     -11.31      23.85      -0.47
 6        |     52     -88.39      23.85      -3.71
 7        |     51     -69.85      23.64      -2.95
 8        |     52     -25.41      23.85      -1.07
 9        |     52     -37.92      23.85      -1.59
 10       |     52      -1.16      23.85      -0.05
 11       |     51      10.73      23.64       0.45
 12       |     52     -33.97      23.85      -1.42
-------------------------------------------------
 Total                    0
```

Comparación no paramètrica

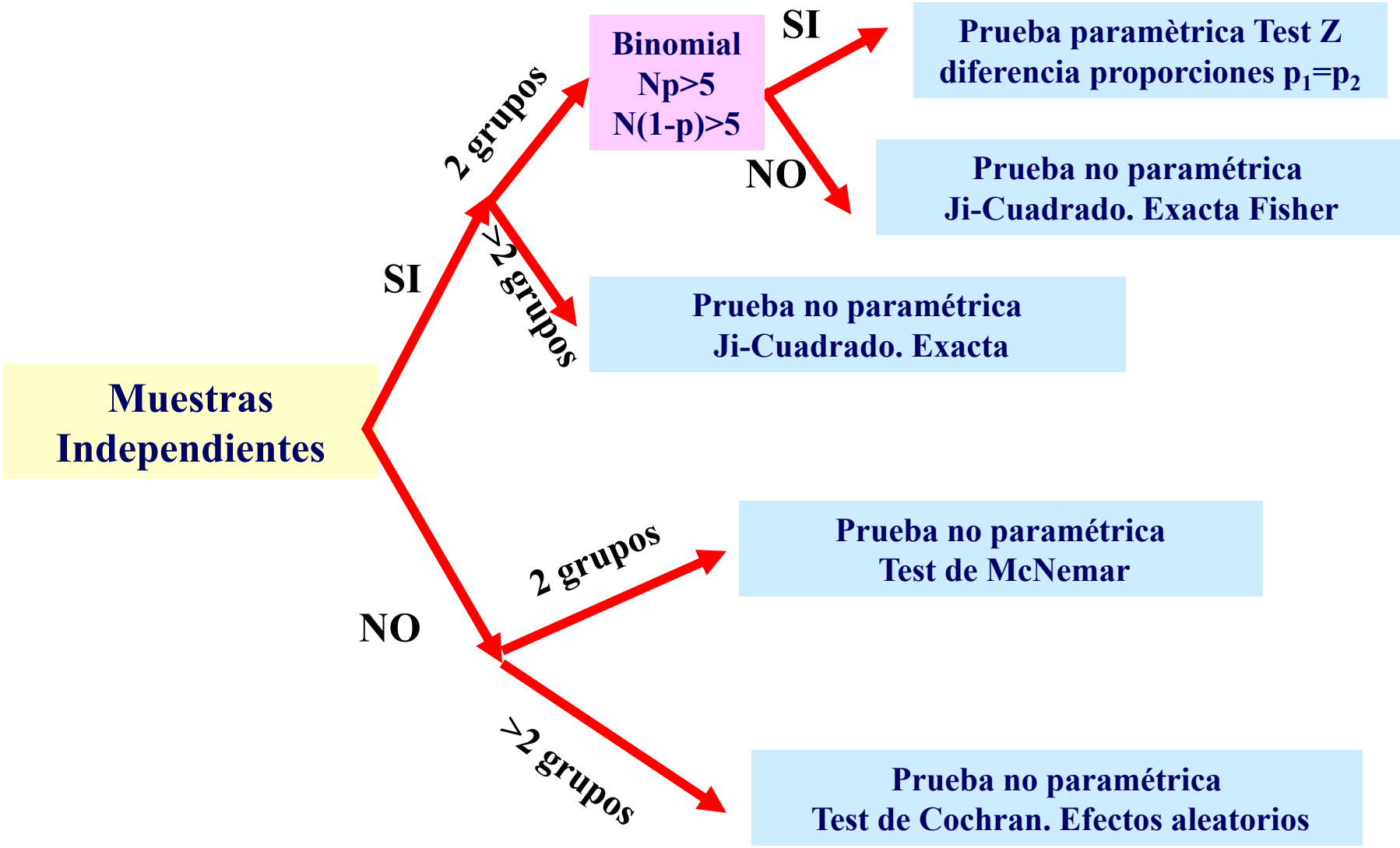
- Test de Skillings-Mack. Datos "long"

skilmack *var1* **,i(***ident***) repeated(***medicion***)**

```
skilmack tas ,id(numero) repeated(medicion)

.....

Note N= 3 not included as only had one observation

Skillings Mack    =  76.340
P-value (No ties) =   0.0000
     N.B. As P-value <0.02, it is likely to be conservative (unless n large).
Consider obtaining a p-value from a simulated null
     distribution of SM - see options.

  Ties exist. Above SEs and P-value approximate, if not too many ties;
  639 rows of [numero, tas]; 308 different combinations;  n(numero) = 60

  Consider using the p-value below, (which is found from a simulated
        conditional null distribution of SM   - see options -
  simulating ..........)

Empirical P-value (Ties)    ~   0.0000
```

Relación entre variable cualitativa (2 niveles) según los niveles de una variable cualitativa

Relación variables categòricas



Muestras Independientes

- Prueba Ji cuadrado, Exacta de Fisher

tabulate *variable1 variable2* **, chi exact exp**

```
. tabulate expcateg aids, chi exact exp

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
+--------------------+

  exposure |    aids diagnosis
  category |       yes         no |     Total
-----------+----------------------+----------
       HSH |        20         63 |        83
           |      18.7       64.3 |      83.0
-----------+----------------------+----------
       UDI |        74        260 |       334
           |      75.3      258.7 |     334.0
-----------+----------------------+----------
     Total |        94        323 |       417
           |      94.0      323.0 |     417.0

          Pearson chi2(1) =   0.1434   Pr = 0.705
           Fisher's exact =                 0.769
   1-sided Fisher's exact =                 0.402
```

2 Muestras Independientes

- Test z paramètrico para diferencia proporciones (variable codificada como 0 y 1)

prtest *variable1 ,by(variable2)*

```
. prtest aids, by(expcateg)


Two-sample test of proportion                     HSH: Number of obs =       83
                                                  UDI: Number of obs =      334
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         HSH |   .2409639   .0469427                      .1489578    .3329699
         UDI |   .2215569   .0227239                      .1770189    .2660949
-------------+----------------------------------------------------------------
        diff |    .019407   .0521536                     -.0828121    .1216261
             |  under Ho:   .0512489     0.38   0.705
------------------------------------------------------------------------------
        diff = prop(HSH) - prop(UDI)                              z =   0.3787
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.6475         Pr(|Z| < |z|) = 0.7049          Pr(Z > z) = 0.35
```

2 Muestras Dependientes

- Test Mcnemar (2 mediciones variables (0,1) a través de un estudio casos control apareado

mcc *var1 var2*

```
. mcc rtas1 rtas12

                 | Controls                |
Cases            |   Exposed   Unexposed   |      Total
-----------------+-------------------------+-----------
         Exposed |          12          19 |         31
       Unexposed |           3          18 |         21
-----------------+-------------------------+-----------
           Total |          15          37 |         52

McNemar's chi2(1) =     11.64    Prob > chi2 = 0.0006
Exact McNemar significance probability      = 0.0009

Proportion with factor
        Cases          .5961538
        Controls       .2884615     [95% Conf. Interval]
                -------- -        --------------------
        difference  .3076923        .1327038    .4826808
        ratio       2.066667        1.349346     3.16532
        rel. diff.  .4324324        .2452494    .6196155

        odds ratio  6.333333        1.864327    33.41648    (exact)
```

2 Muestras Dependientes

- Test de simetria (útil para 2 niveles de la variable)

  symmetry *var1 var2*

```
. symmetry rtas1 rtas12
-------------------------------------
TAS         |
medicion    |       TAS medicion 12
1           | <140mmHg  >140mmHg    Total
------------+------------------------------
 <140mmHg   |    18          3         21
 >140mmHg   |    19         12         31
            |
      Total |    37         15         52
-------------------------------------

                                                chi2     df
Prob>chi2
-------------------------------------------------------------------------------
Symmetry (asymptotic)                         |     11.64       1        0.0006
Marginal homogeneity (Stuart-Maxwell)         |     11.64       1        0.0006
-------------------------------------------------------------------------------
symmi  18 3 \ 19 12
```

>2 Muestras Dependientes

- Test de Cochran (variable respuesta de 2 niveles de la variable)

  cochran *var1 var2 …varN*

```
. cochran rtas1 rtas2 rtas12, detail

Test for equality of proportions of nonzero
outcomes in matched samples (Cochran's Q):

    Variable | Proportion      Count
-------------+----------------------
       rtas1 |   .5961538         31
       rtas2 |   .4615385         24
      rtas12 |   .2884615         15
-------------------------------------

Number of obs       =        52
Cochran's chi2(2)   =  13.78571
Prob > chi2         =    0.0010
```

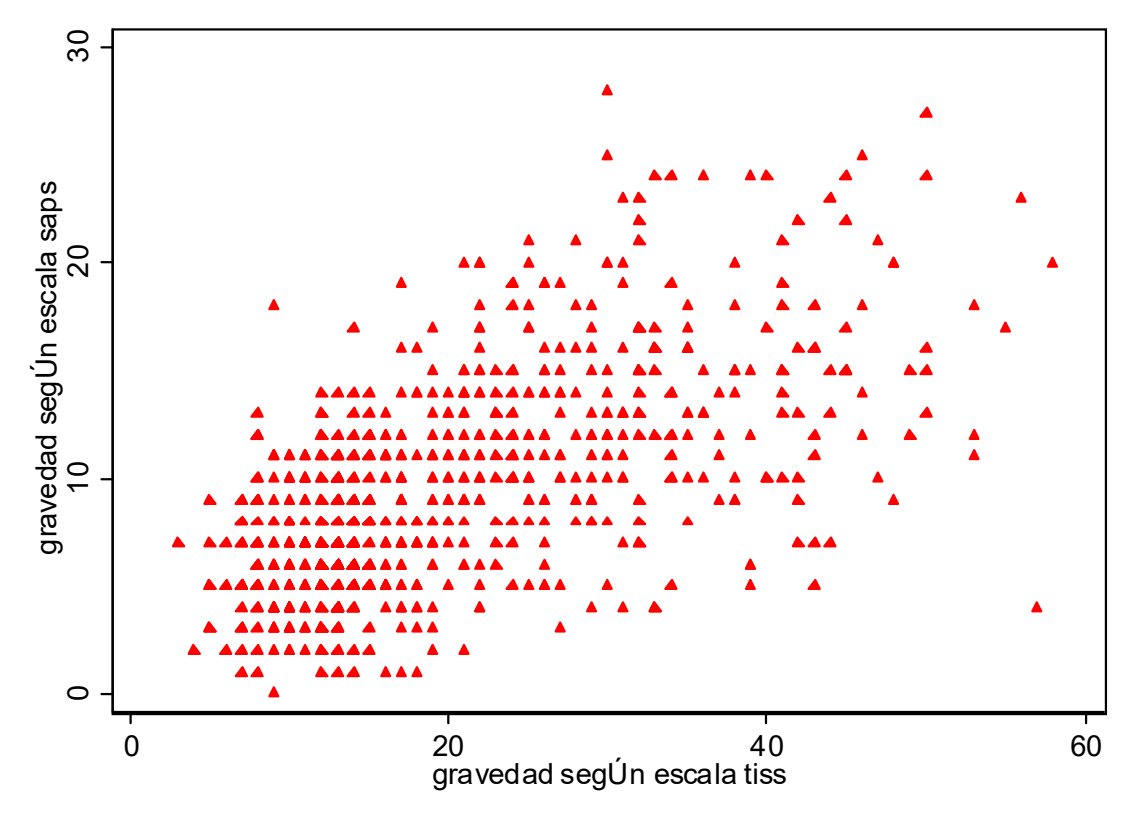
Relación entre 2 variables cuantitativas (correlación y regresión)

Gràfico de Dispersion

- Scatterplot

sc *var1 var2, msymbol(sym) mcolor(color) msize(size)*

```
sc saps tiss, mcolor(red) msymbol(triangle) msize(small)
```

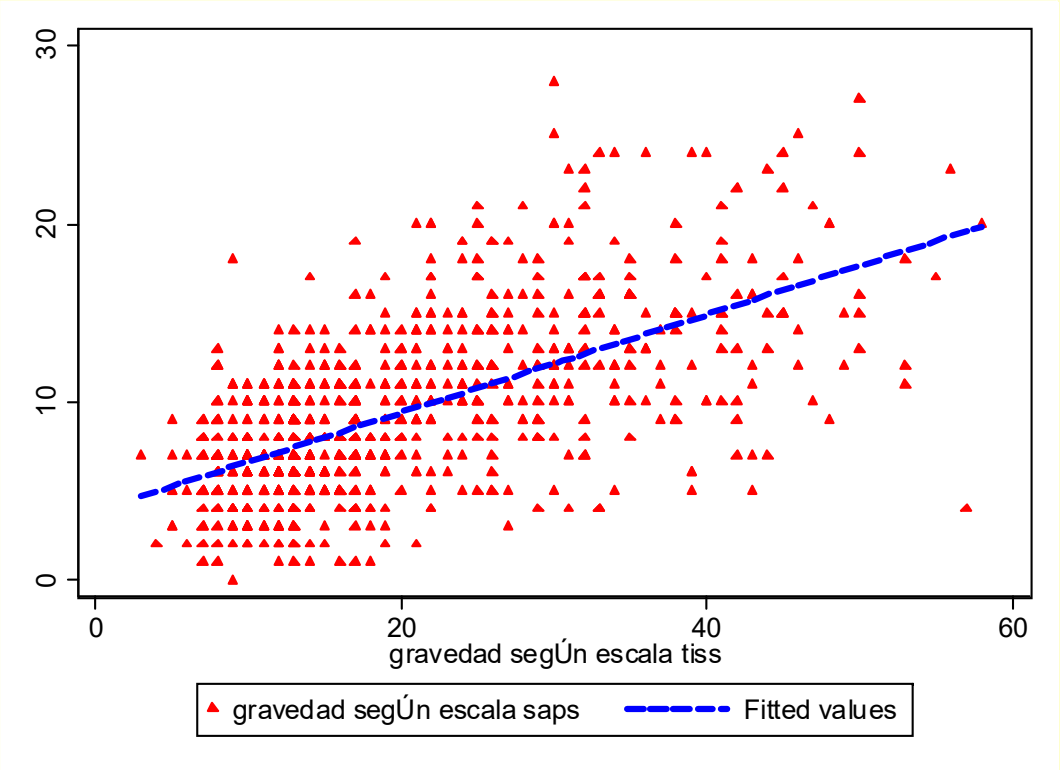


Añadir gràficos (Twoway)

- Es poden combinar gràficos amb el comando twoway para mirar la forma de la relación entre las dos variables

```
twoway sc var1 var2, msymbol(sym) mcolor(color) msize(size)
    || lfit var1 var2, lpattern(lp) lwidth(lw) lcolor(lc)
```

```
twoway sc saps tiss, mcolor(red) msymbol(triangle) msize(small)
|| lfit saps tiss, lpattern(dash) lwidth(thick) lcolor(blue)
```

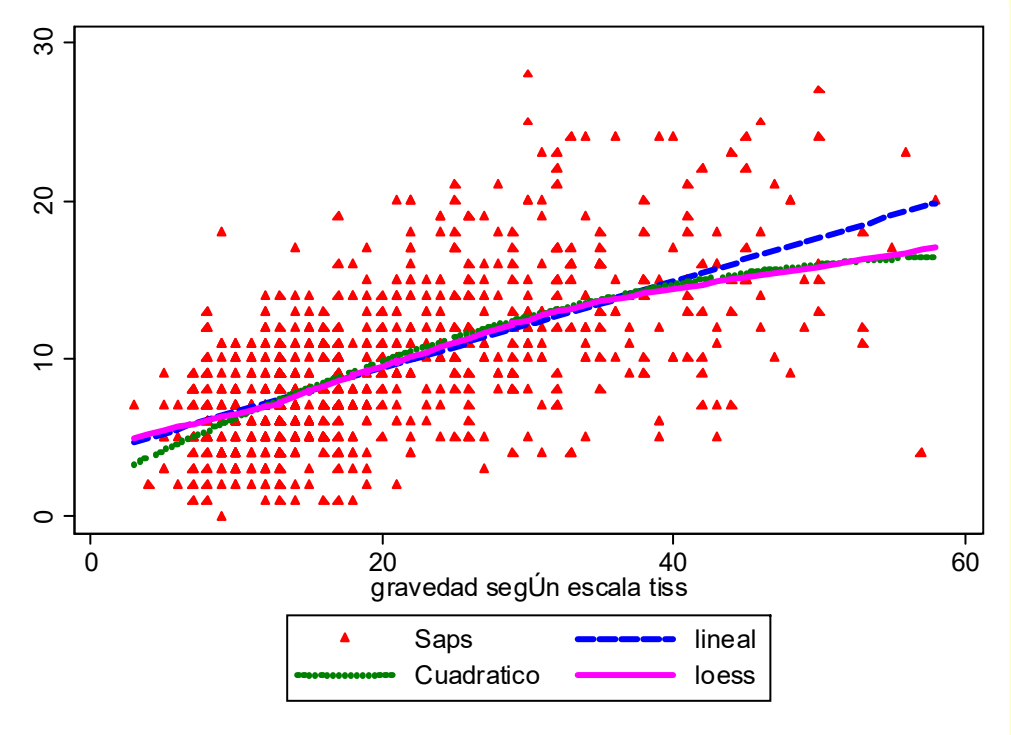


Añadir gràficos (Twoway)

- Otras opciones para el twoway son

```
||lfit   ||mspline   ||lowess   ||lfit ||qfit
||fpfit  ||lfitci    ||qfitci   ||fpfitci
||, title() subtitle() note() legend( label(# eti#) )
```

```
twoway sc saps tiss, mcolor(red) msymbol(triangle) msize(small)
|| lfit saps tiss, lpattern(dash) lwidth(thick) lcolor(blue)
|| qfit saps tiss, lpattern(dot) lwidth(thick) lcolor(green)
|| lowess saps tiss, lpattern(solid) lwidth(thick) lcolor(magenta)
||, legend( label(1 Saps) label(2 lineal) label(3 Cuadratico) label(4 loess))
```



Correlación

- Existen dos comandos para obtener la correlación

corr var1 var2 var3

pwcorr var1 var2 var3, sig obs

```
. corr tiss saps edad
(obs=827)
             |     tiss     saps     edad
-------------+---------------------------
        tiss |   1.0000
        saps |   0.6183   1.0000
        edad |  -0.0579   0.2145   1.0000
. pwcorr tiss saps edad, sig obs

             |     tiss     saps     edad
-------------+---------------------------
        tiss |   1.0000
             |
             |      829
             |
        saps |   0.6188   1.0000
             |   0.0000
             |      828      837
             |
        edad |  -0.0592   0.2118   1.0000
             |   0.0889   0.0000
             |      828      836      844
             |
```

Regressión lineal con Stata

- Regresión líneal entre dos variables cuantitativas

  regress depvar **[**indepvars**]** **,** **opciones**

```
regress saps tiss

      Source |       SS       df       MS              Number of obs =     828
-------------+------------------------------           F(  1,   826) =  512.47
       Model |  7230.58002     1  7230.58002            Prob > F      =  0.0000
    Residual |  11654.1688   826  14.1091632           R-squared     =  0.3829
-------------+------------------------------           Adj R-squared =  0.3821
       Total |  18884.7488   827  22.8352464           Root MSE      =  3.7562

------------------------------------------------------------------------------
        saps |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tiss |   .2755289   .0121711    22.64   0.000     .2516389    .2994189
       _cons |   3.897909   .2743255    14.21   0.000     3.359452    4.436366
------------------------------------------------------------------------------
```