Mode	Range	27.10348
	Interquartile Range	6.05068

Tabla 3:

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	1018.146	Pr > |t|	<.0001
Sign	M	50	Pr >= |M|	<.0001
Signed Rank	S	2525	Pr >= |S|	<.0001

Tabla 4

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.987233	Pr < W	0.4535
Kolmogorov-Smirnov	D	0.057461	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.068799	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.444214	Pr > A-Sq	>0.2500

Tabla 5

Quantiles (Definition 5)

Quantile	Estimate
100% Max	513.830
99%	512.736
95%	508.014
90%	506.315
75% Q3	503.434
50% Median	500.870
25% Q1	497.383
10%	494.459
5%	490.826
1%	487.514
0% Min	486.727

Tabla 6:

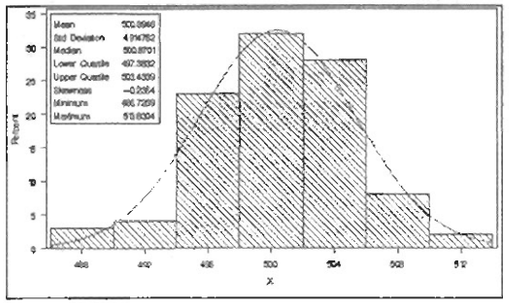
Extreme Observations

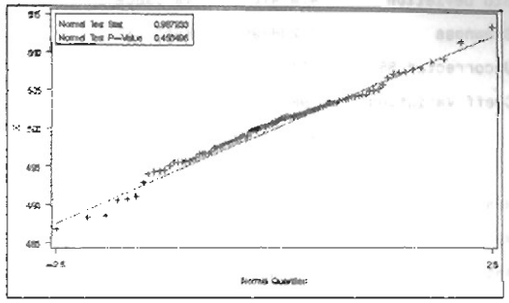
------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
486.727	80	508.138	98
488.302	97	508.944	22
488.468	53	509.380	95
490.574	94	511.641	49
490.699	37	513.830	45

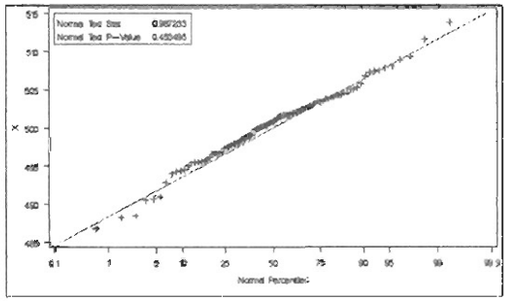
Datos guardado en el data univout

Obs	MEDIA	DESV_STD	MEDIANA	PERCENTIL_ 50	PERCENTIL_ 95	PERCENTIL_ 97_5	PERCENTIL_ 100
1	500.395	4.91476	500.870	500.870	508.014	509.380	513.830

Los gráficos solicitados en el anterior ejemplo, lucen como se muestra a continuación:







```
/*solicitando intervalos de confianza bilaterales para la media, la desviación estándar y la varianza
(CIBASIC(TYPE=twosided) )de X, e intervalos de confianza no paramétricos (CIPCTLOF(TYPE=SYMMETRIC)) y
paramétricos (CIPCTLNORMAL(TYPE=twosided)) para percentiles, adicionalmente, se especifica el valor de
significancia para pruebas e intervalos que se construyan y el valor para las hipótesis de igualdad
(paramétrica basada en la distribución normal, y no paramétricas) para la media (MU0=500)*/
proc univariate data=DATOS alpha=0.05 CIBASIC(TYPE=twosided) CIPCTLOF(TYPE=SYMMETRIC)
CIPCTLNORMAL(TYPE=twosided) MU0=500;
var X;
run;
```

Las salidas de la anterior programación son las siguientes, estas se diferencian del ejemplo anterior en las salidas referentes a: "Basic Confidence Limits Assuming Normality", "Tests for Location Mu0=500", cuantiles con límites de confianza asumiendo normalidad y límites de confianza libres de distribución para cuantiles.

```
The UNIVARIATE Procedure
Variable:  X
                          Moments
N                     100    Sum Weights              100
Mean           500.394595    Sum Observations   50039.4595
Std Deviation    4.9147617    Variance           24.1548826
Skewness        -0.2363992    Kurtosis           0.55617509
Uncorrected SS  25041866.4    Corrected SS       2391.33338
Coeff Variation 0.98217722    Std Error Mean     0.49147617
              Basic Statistical Measures
    Location                    Variability
Mean     500.3946     Std Deviation            4.91476
Median   500.8701     Variance                24.15488
Mode        .         Range                   27.10348
                      Interquartile Range      6.05068

     Basic Confidence Limits Assuming Normality
Parameter          Estimate     95% Confidence Limits
Mean              500.39459     499.41940   501.36979
Std Deviation       4.91476       4.31519     5.70936
Variance           24.15488      18.62090    32.59676

          Tests for Location: Mu0=500
Test           -Statistic-     -----p Value------
Student's t    t  0.802877     Pr > |t|    0.4240
Sign           M         7     Pr >= |M|   0.1933
Signed Rank    S       309     Pr >= |S|   0.2903

          Quantiles (Definition 5)
                          95% Confidence Limits
Quantile       Estimate      Assuming Normality
100% Max       513.830
99%            512.736    510.1707     513.9540
95%            508.014    507.1428     510.1519
90%            506.315    505.5045     508.1484
75% Q3         503.434    502.6996     504.8642
50% Median     500.870    499.4194     501.3698
25% Q1         497.383    495.9250     498.0896
10%            494.459    492.6408     495.2847
5%             490.826    490.6373     493.6464
```



```
1%             487.514    486.8352     490.6185
0% Min         486.727

          Quantiles (Definition 5)
                 95% Confidence Limits     -------Order Statistics-------
Quantile          Distribution Free       LCL Rank  UCL Rank    Coverage
100% Max
99%            509.380    513.830          98        100        55.46
95%            506.814    513.830          91        100        96.59
90%            504.579    508.944          85         97        95.23
75% Q3         502.459    504.383          67         84        95.13
50% Median     499.837    501.888          41         61        95.40
25% Q1         495.675    498.523          17         34        95.13
10%            490.574    495.613           4         16        95.23
5%             486.727    494.409           1         10        96.59
1%             486.727    488.468           1          3        55.46
0% Min

          Extreme Observations
------Lowest-----          -----Highest-----
   Value      Obs            Value      Obs
 486.727       80          508.138       98
 488.302       97          508.944       22
 488.468       53          509.380       95
 490.574       94          511.641       49
 490.699       37          513.830       45

/*solicitando una media recortada 10%*/
proc univariate data=DATOS TRIMMED=0.1(TYPE=TWOSIDED);
var X;
run;
```

Las salidas de la anterior programación presentan como información adicional las medias recordadas y los respectivos límites de confianza:

```
The UNIVARIATE Procedure
Variable:  X
                          Moments
N                     100    Sum Weights              100
Mean           500.394595    Sum Observations   50039.4595
```

```
Std Deviation        4.9147617    Variance          24.1548826
Skewness            -0.2363992    Kurtosis          0.55617509
Uncorrected SS      25041866.4    Corrected SS      2391.33338
Coeff Variation     0.98217722    Std Error Mean    0.49147617

              Basic Statistical Measures

     Location                    Variability

Mean     500.3946     Std Deviation            4.91476
Median   500.8701     Variance                24.15488
Mode        .         Range                   27.10348
                      Interquartile Range      6.05068

           Tests for Location: Mu0=0

Test           -Statistic-     -----p Value------
Student's t    t  1018.146     Pr > |t|    <.0001
Sign           M        50     Pr >= |M|   <.0001
Signed Rank    S      2525     Pr >= |S|   <.0001


                              Trimmed Means

Percent    Number                  Std Error
Trimmed    Trimmed      Trimmed      Trimmed     95% Confidence
in Tail    in Tail        Mean         Mean          Limits              DF
  11.00        11      500.4976     0.449585   499.6023  501.3928        77
          Trimmed Means

Percent
Trimmed      t for H0:
in Tail       Mu0=0.00    Pr > |t|
  11.00       1113.244     <.0001


Quantiles (Definition 5)

Quantile       Estimate
100% Max        513.830
99%             512.736
95%             508.014
90%             506.315
75% Q3          503.434
50% Median      500.870
25% Q1          497.383
10%             494.459
5%              490.826
1%              487.514
```



```
0% Min          486.727

              Extreme Observations
------Lowest-----          -----Highest-----
  Value        Obs           Value        Obs
 486.727        80           508.138       98
 488.302        97           508.944       22
 488.468        53           509.380       95
 490.574        94           511.641       49
 490.699        37           513.830       45
```

2.5.2.3 PROC TTEST

Este procedimiento SAS permite obtener intervalos de confianza para: medias, diferencias de medias (muestras independientes o pareadas) y para la desviación estándar, de poblaciones normales. También permite realizar la prueba t para la media y la diferencia de medias (con muestras independientes o pareadas) de poblaciones normales. Para la comparación de medias con muestras independientes, también ofrece resultados para la prueba de igualdad de varianzas, así como los estadísticos t para los casos de varianzas iguales (método POOLED) y diferentes (aproximación Satterthwaite y Cochran). La sintaxis general es como sigue:

```
PROC TTEST DATA=base de datos SAS< opciones > ;
CLASS variable ;
PAIRED variables ;
BY variable ;
VAR variables ;
FREQ variable ;
WEIGHT variable ;
RUN;
```

Las opciones posibles en la línea del PROC TTEST son:

ALPHA=p con 0<p<1, especifica el nivel de significancia para las pruebas y el nivel de confianza para los intervalos es (1-p)x100%. Por defecto, si no se especifica esta opción, el nivel usado es 0.05.

CI=EQUAL ó UMPU ó NONE , especifica si los intervalos de confianza para la desviación estándar han de ser presentados y de qué clase. Cuando **CI=EQUAL**, los intervalos bilaterales son de colas

iguales, y por defecto esta es la especificación que considera el procedimiento. **CI=UMPU** especifica que los intervalos deben basarse en el estimador insesgado uniformemente más potente. El **CI=NONE** solicita que no se imprima por pantalla ningún intervalo de confianza. Se pueden especificar simultáneamente en la misma invocación del procedimiento **EQUAL** y **UMPU**.

COCHRAN solicita se use la aproximación de Cochran y Cox para el nivel de probabilidad del estadístico t cuando las varianzas son diferentes en una prueba sobre diferencia de medias de poblaciones normales con muestras independientes.

H0=m especifica el valor del parámetro (media o diferencia de medias independientes o pareadas) bajo la hipótesis nula, para las pruebas t. Por defecto m es cero (si no se especifica).

CLASS se usa en el caso de comparación de medias de dos poblaciones normales usando muestras independientes, y especifica la variable de clase que identifica a cuál población pertenece cada observación en el conjunto de datos. En este caso la estructura de datos debe incluir una sola variable de análisis especificada en **VAR** y una variable de clasificación con únicamente dos niveles o categorías. Cuando se usa esta declarativa, no debe usarse la declaración **PAIRED**. Si sólo se tiene una población de análisis, entonces se omite **CLASS**.

PAIRED Cuando se quieren hacer comparaciones pareadas de medias (es decir, con dos muestras no independientes), se invoca esta declarativa seguida por las variables a comparar separadas por asterisco (*) o por dos puntos (:). El asterisco requiere comparaciones entre cada variable ubicado a su izquierda con cada variable indicada a su derecha. Los dos puntos solicitan la comparación entre la primera variable a su izquierda con la primera variable a su derecha, de la segunda variable a su izquierda con la segunda a su derecha, etc. en este caso, el número de variables a izquierda y derecha de los dos puntos debe ser igual. Tanto en **CLASS** como en **PAIRED**, las diferencias se calculan en el orden de variable a la izquierda menos variable a la derecha. Las declaraciones **CLASS** y **VAR** se omiten cuando se usa **PAIRED**.

Algunos ejemplos con el uso del asterisco y los dos puntos en PAIRED:

```
PAIRED X*Y            Compara X vs. Y
PAIRED X*Y  Z*W             Compara X vs. Y y Z vs. W
PAIRED (X Y)*(Z W)    Compara X vs. Z, X vs. W, Y vs. Z y  Yvs. W
PAIRED (X Y):(Z W)          Compara X vs. Z y Y vs. W
```

BY variable para obtener análisis separados de las observaciones según grupos o categorías definidos por la variable indicada en **BY**. Por defecto, el procedimiento espera que los datos se hayan ordenado

ascendentemente por los niveles de la variable en **BY**. Si los datos no han sido ordenados previamente con **PROC SORT** o el ordenamiento no es con base en la variable indicada, entonces puede especificarse una de las siguientes opciones: **NOTSORTED** o **DESCENDING**. La primera opción indica que los datos no necesariamente están ordenados en orden alfabético o creciente según las categorías o niveles de la variable especificada en **BY**. **DESCENDING** especifica que el conjunto de datos está ordenado en forma descendente según los niveles de la variable especificada en **BY**.

VAR nombra las variables a ser usadas en los análisis. El análisis de una sola muestra es ejecutado cuando la declaración **VAR** es usada sin invocar la declaración **CLASS**, en tanto que se realizan comparaciones de grupos (dos muestras independientes cada una de una población normal) cuando se especifica **CLASS**. Si no se especifica la declaración **VAR**, el procedimiento realiza análisis univariado de cada una de las variables numéricas que existan en el correspondiente conjunto de datos, excepto las especificadas en **CLASS**, **BY**, **FREQ** y **WEIGHT**. **VAR** no se invoca cuando se usa **PAIRED**.

FREQ variable identifica en el conjunto de datos una variable que contiene la frecuencia de ocurrencia de cada observación, es decir, **PROC TTEST** trata cada observación como si ésta apareciese n veces, siendo n el valor de la variable especificada en **FREQ** en la respectiva observación. La variable de frecuencias sólo puede tomar valores enteros, en caso contrario sólo la parte entera es considera para la frecuencia y si este valor es menor de uno o falta, entonces la correspondiente observación es excluida del análisis. Si la declaración **FREQ** no es usada, entonces por defecto el procedimiento asume una frecuencia unitaria para cada observación en el conjunto de datos.

WEIGHT variable pondera cada observación en el conjunto de datos con el valor correspondiente de la variable especificada en **WEIGHT**. Los valores de ésta variable de ponderación deben ser positivo y menores o iguales a 1. Si algún valor de esta variable es cero, negativo o faltante, entonces la correspondiente observación no es usada en los cálculos. Por defecto, cuando no se invoca la declaración **WEIGHT**, el procedimiento asigna una ponderación de 1 a todas las observaciones en el conjunto de datos.

Ejemplo 1:

Considere que los datos de la tabla que aparece a continuación, corresponden a una muestra aleatoria de los puntajes en cierta prueba, de 7 hombres y 7 mujeres. Se quiere determinar si existe diferencia entre los puntajes promedios de hombres y mujeres en dicha prueba. Asumiendo que el puntaje promedio en cada género se distribuye normal, entonces procedemos a su análisis.

La tabla de datos es como sigue:

GENDER(género)	SCORE(puntaje)	GENDER(género)	SCORE(puntaje)
f	75	m	82
f	76	m	80
f	80	m	85
f	77	m	85
f	80	m	78
f	77	m	87
f	73	m	82

Entrada de datos para una prueba de diferencia de medias: Se deben ingresar dos columnas una columna debe indicar el grupo (Ej. Variable Gender: f= femenino, m= masculino) y otra con los valores de cada grupo (Ej. Variable Score: indica el puntaje de las mujeres y de los hombres) que se quieren comparar. La entrada de esta tabla en SAS, se puede hacer de la siguiente forma:

```
options nodate nocenter nonumber PS=60 LS=80;
data scores;
input Gender $ Score @@;
cards;
f 75  f 76  f 80  f 77  f 80  f 77  f 73
m 82  m 80  m 85  m 85  m 78  m 87  m 82
;
run;
proc print data=scores;run;

The SAS System
Obs    Gender    Score
  1      f        75
  2      f        76
  3      f        80
  4      f        77
  5      f        80
  6      f        77
  7      f        73
  8      m        82
  9      m        80
 10      m        85
 11      m        85
 12      m        78
 13      m        87
 14      m        82
```


La comparación de las dos muestras a un nivel de significancia del 5% y solicitando el método Cochran para la prueba t en el caso de varianzas diferentes:

```
proc ttest data=scores cochran ci=equal alpha=0.05;
class Gender;
var Score;
run;
```

Salidas sas del proc ttest: Asumiendo normalidad y varianzas desconocidas, usando la distribución t en la construcción de intervalos de medias y las respectivas pruebas de hipótesis, la tabla 1 nos dá lo siguiente: límites de los intervalos de confianza del 95% para las medias de cada población, para las diferencias de las medias de las poblaciones 1 y 2, valores de las medias muestrales de cada muestra y de la diferencia de estas medias. También nos dá los límites de intervalos de confianza para las desviaciones estándar de cada población, para las diferencias de las observaciones que se consideraría en el caso de muestras pareadas, las desviaciones estándar estimadas de cada muestra y para las diferencias y finalmente, el error estándar estimado de las medias muestrales de las dos muestras y de las diferencias muestrales (también aparecen los valores mínimos y máximos en cada muestra):

La tabla 2 da los grados de libertad (DF) los valores observados (t Value) de los estadísticos para la prueba t sobre la hipótesis H0: mu1=mu2 vs. H1: mu1≠mu2 para el caso sigma1=sigma2 pero desconocidas (método Pooled) y para el caso sigma1≠sigma2 (método Satterthwaite y método Cochran), también podemos ver los valores P de estas pruebas bilaterales en la columna Pr > |t| con los cuales debemos decidir si se rechaza o no H0.

La tabla 3 presenta los valores del estadístico F (F Value) para la prueba de igualdad de varianzas H0: sigma1=sigma2 vs. H1: sigma1≠sigma2 , que utiliza el método denominado FOLDED, el cual construye el estadístico de prueba como la razón entre la varianza muestral más grande sobre la varianza muestral más pequeña de las dos muestras, los grados de libertad del numerador y del denominador de la distribución F correspondiente (Num DF y Den DF respectivamente), y el valor P de la prueba (Pr > F) también son suministrados.

NOTA: Si la prueba de igualdad de varianzas no rechaza H0, entonces la prueba de medias evalua los resultados para la prueba de igualdad de medias arrojados por el método Pooled; pero si la prueba sobre

las varianzas rechaza la igualdad de éstas, entonces para la prueba t sobre la igualdad de medias se evaluan los resultados para la prueba de igualdad de medias arrojados por el método Satterthwaite.

Tabla 1:

The TTEST Procedure

Statistics

Variable	Gender	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev
Score	f	7	74.504	76.857	79.211	1.6399	2.5448
Score	m	7	79.804	82.714	85.625	2.028	3.1472
Score	Diff (1-2)		-9.19	-5.857	-2.524	2.0522	2.8619

Statistics

Variable	Gender	Upper CL Std Dev	Std Err	Minimum	Maximum
Score	f	5.6039	0.9619	73	80
Score	m	6.9303	1.1895	78	87
Score	Diff (1-2)	4.7242	1.5298		

Tabla 2:

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
Score	Pooled	Equal	12	-3.83	0.0024
Score	Satterthwaite	Unequal	11.5	-3.83	0.0026
Score	Cochran	Unequal	6	-3.83	0.0087

Tabla 3:

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
Score	Folded F	6	6	1.53	0.6189

Uso de la opción CI: Observe en los resultados que no aparecen los límites de confianza para las desviaciones estándar ni los valores máximos y mínimos de las muestras:

```
proc ttest data=scores alpha=0.01 CI=none;
class Gender;
var Score;
run;
```


The TTEST Procedure

Statistics

Variable	Gender	N	Lower CL Mean	Mean	Upper CL Mean	Std Dev	Std Err
Score	f	7	73.291	76.857	80.423	2.5448	0.9619
Score	m	7	78.304	82.714	87.124	3.1472	1.1895
Score	Diff (1-2)		-10.53	-5.857	-1.184	2.8619	1.5298

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
Score	Pooled	Equal	12	-3.83	0.0024
Score	Satterthwaite	Unequal	11.5	-3.83	0.0026

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
Score	Folded F	6	6	1.53	0.6189

Realice el siguiente ejemplo:

```
/*simulación de dos muestras normales independientes
de medias mu1=540 mu2=600 sigma1=10 y sigma2=5, con n1=n2=40*/
data datos;
do i=1 to 40;
pob=1;          /*crea variable de clasificación para identificar las muestras de la población 1*/
x=540+10*normal(0); /*genera una observación normal de media 540 y desviación estándar de 10*/
output;
end;
do i=41 to 80; /*solicita repetir 40 veces el procedimiento pero i comienza desde 41*/
pob=2; /*crea variable de clasificación para identificar las muestras de la población 2*/
x=600+5*normal(0); /*genera una observación normal de media 600 y desviación estándar de 5*/
output;
end;*concluye ciclo DO */
run;

/*para visualizar los datos simulados*/
proc print data=datos;run;

/*para análisis descriptivo y evaluación del supuesto de normalidad*/
proc univariate data=datos normaltest noprint;
by pob;
var x;
```

```
histogram x;
inset mean std;
qqplot x;
inset normaltest probn;
run;
quit;

/*En los gráficos de probabilidad normal solicitados deben aparecer los valores del estadístico de
prueba (Normal Test Stat) y el valor P (Normal Test P-Value) para la prueba de hipótesis: H0: X se
distribuye normal vs. H1: X no se distribuye normal. Si las hipótesis nulas en cada prueba, no son
rechazadas, entonces se ha probado la normalidad y podemos usar el proc ttest para las inferencias
sobre las diferencias de medias de dos poblaciones normales con muestras independientes.*/

/*procedimiento para obtener estadísticos para prueba de igualdad de medias e igualdad de varianzas,
de dos poblaciones normales asumiendo que desconocemos las varianzas y que las muestras son
independientes*/

proc ttest data=datos;
class pob;
var x;
run;
```

Ejemplo 2: Datos pareados

La entrada de datos de los pares debe organizarse en dos columnas diferentes (las dos variables pareadas). Considere la siguiente tabla de observaciones pareadas por individuo, de la presión antes y después de aplicar un tratamiento:

Individuo	PSSantes	PSSdespues
1	120	128
2	124	131
3	130	131
4	118	127
5	140	132
6	128	125
7	140	141
8	135	137
9	126	118
10	130	132
11	126	129
12	127	135

```
data presion;
```



```
input PSSantes PSSdespues @@;
cards;
120 128   124 131   130 131   118 127
140 132   128 125   140 141   135 137
126 118   130 132   126 129   127 135
;
run;
```

Realización de la prueba utilizando el PROC TTEST (opción paired var1*var2;) presenta todos los valores de los estadísticos, intervalos de confianza y el respectivo valor p.

```
proc ttest data=presion;
paired PSSantes*PSSdespues;
run;
```

Los resultados se muestran a continuación:

```
The SAS System
The TTEST Procedure
                                  Statistics
                                 Lower CL           Upper CL  Lower CL
Difference                   N      Mean    Mean      Mean   Std Dev  Std Dev
PSSantes - PSSdespues       12    -5.536  -1.833    1.8598    4.1288   5.8284
                              Statistics
                            Upper CL
Difference                 Std Dev    Std Err    Minimum    Maximum
PSSantes - PSSdespues       9.8958     1.6825         -9          8
                           T-Tests
Difference                     DF    t Value    Pr > |t|
PSSantes - PSSdespues          11      -1.09      0.2992
```

Finalmente considere el siguiente ejemplo en el cual se pide análisis univariado de las medias y desviaciones estándar de cada una de las variables en el data presión. Tenga en cuenta que hacer esto supone que las variables se distribuyen normales:

```
proc ttest data=presion;
run;
```

Con lo anterior se obtiene:

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev
PSSantes	12	124.26	128.67	133.07	4.911	6.9326	11.771
PSSdespues	12	126.74	130.5	134.26	4.1909	5.9161	10.045

Statistics

Variable	Std Err	Minimum	Maximum
PSSantes	2.0013	118	140
PSSdespues	1.7078	118	141

T-Tests

Variable	DF	t Value	Pr > |t|
PSSantes	11	64.29	<.0001
PSSdespues	11	76.41	<.0001

2.5.3 PROC FREQ: TABLAS DE FRECUENCIAS

Muestra totales y proporciones por celda, fila y columna, y puede también realizar algunas pruebas de hipótesis sobre los datos. Para tablas de una sola vía, PROC FREQ puede calcular pruebas Chi-cuadrado de igualdad o comparación de proporciones especificas. Para tablas de 2 vías, PROC FREQ calcula pruebas y medidas de asociación. Para tablas de n vías, PROC FREQ hace análisis estratificado, calculando estadísticos dentro y para cada estrato.

Los estadísticos incluidos para las tablas de contingencia son:

- Medidas y pruebas chi-cuadrado.
- Medidas de asociación.
- Riesgos (proporciones binomiales) y diferencias de riesgo para tablas 2x2.
- Razones de odds y riesgos relativos para tablas 2x2.
- Pruebas para tendencia.
- Pruebas y medidas de concordancia.
- Estadísticos de Cochran-Mantel-Haenszel.

Sintaxis

```
PROC FREQ <opción(es)>;
BY <DESCENDING> variable-1 <...<DESCENDING> variable-n> <NOTSORTED>;
EXACT statistic-palabra(s)_clave(s)</ opción(es)>;
OUTPUT statistic-keyword(s) <OUT=archivo-SAS>;
```



```
TABLES request(s) </ opción(es)>;
TEST statistic-palabra(s)_clave(s);
WEIGHT variable;
```

A continuación se definirán las declaraciones más usadas.

DECLARACIÓN PROC FREQ

```
PROC FREQ DATA= <option(s)>;
```

Controles para la salida de impresión

COMPRESS Comienza la próxima tabla de una vía en la página actual, aun si la tabla entera no cabe en la misma.

FORMCHAR= Especifica caracteres a ser usados como divisores de celda y bordes en las tablas de contingencia.

NOPRINT Suprime toda la salida.

ORDER= Especifica el orden para los valores de las variables.

PAGE Muestra una tabla por página.

Detalles

ORDER=DATA | FORMATTED | FREQ | INTERNAL

Ordena los valores de la tabla de frecuencia o de contingencia de acuerdo al orden especificado, donde:

DATA Ordena los valores de acuerdo a su orden en el archivo de entrada.

FORMATTED Ordena los valores por sus valores formateados. Por defecto, el orden es ascendente.

FREQ Ordena los valores por frecuencia de conteo descendente.

INTERNAL Ordena los valores por sus valores no formateados, el cual coincide con el mismo orden del PROC SORT.

DECLARACIÓN BY

```
BY <DESCENDING> variable-1 <...<DESCENDING> variable-n> <NOTSORTED>;
```

Argumentos Requeridos

variable

Especifica la variable que el procedimiento usa para formar los grupos BY. Usted puede especificar más de una variable. Si no se usa la opción NOTSORTED en la declaración BY, las observaciones en

la base de datos deben estar ordenadas por todas las variables que usted especificó, o se deben indexar apropiadamente.

Opciones

DESCENDING Especifica que las observaciones están ordenadas en forma descendiente por la variable que sigue inmediatamente de la palabra DESCENDING en la declaración BY.

NOTSORTED Especifica que las observaciones no están necesariamente dadas en orden alfabético o numérico. Las observaciones son agrupadas en otra forma, por ejemplo, en orden cronológico. El requerimiento de ordenar o indexar las observaciones según los valores de las variables es suspendido cuando usted usa la opción NOTSORTED. De hecho, el procedimiento no usa un indice si usted especifica NOTSORTED. El procedimiento define un grupo como un conjunto de observaciones consecutivas que tienen los mismos valores, esto para todas las variables. Si las observaciones con los mismos valores no son consecutivas, el procedimiento toma cada conjunto consecutivo como un grupo BY diferente.

DECLARACIÓN TABLES

```
TABLES requests < / options > ;
```

El comando requests en la declaración TABLES sirve para construir tablas de frecuencia (1 vía) o de contingencia (de 2 a n vías) y estadísticos para esas tablas. Se pueden incluir cualquier número de declaraciones TABLES. Si la declaración TABLES no es dada, se produce una tabla de frecuencias de una vía para todas las variables en el archivo de SAS que no han sido listadas en otras declaraciones.
El siguiente argumento es requerido por la declaración TABLES.

requests

Especifica las tablas de frecuencia o de contingencia que serán producidas. Un request está compuesto por un nombre de variable o un grupo de nombres de variable separado por asteriscos. Para producir una tabla de frecuencias de una vía, use una sola variable. Para producir una tabla de contingencia de dos vías, use un asterisco entre las dos variables. Para producir una tabla de contingencia de n vías (donde n>2), separe las variables deseadas con asteriscos. Los valores de estas variables forman las filas, columnas, y estratos de la tabla.

Para tablas de dos a n vías, los valores de la última variable forman las columnas de la tabla de contingencia, mientras que aquellos en la penúltima variable forman las filas. Cada nivel (o

combinación de niveles) de las demás forman un estrato. PROC FREQ produce una tabla de contingencia separada para cada estrato. Por ejemplo, la especificación de A*B*C*D en la declaración TABLES produce k tablas, donde k es el número de diferentes combinaciones de valores para A y B. Cada tabla tiene los valores de C como filas y los valores de D como columnas.

Se pueden tener múltiples declaraciones TABLES en el PROC FREQ. Para especificar múltiples tablas rápidamente, use la sintaxis abreviada siguiente:

```
Sintaxis Abreviada
Request                    Equivalente a
tables A*(B C);            tables A*B    A*C;
tables (A B)*(C D);        tables A*C    B*C    A*D    B*D;
tables (A B C)*D;          tables A*D    B*D    C*D;
tables A-C;                tables A    B    C;
tables (A-C)*D;            tables A*D    B*D    C*D;
```

Sin opciones

Si no se especifican opciones al crear una tabla de frecuencias para una variable, PROC FREQ produce frecuencias, frecuencias acumuladas, porcentajes, y porcentajes acumulados para cada valor de la variable. Si no se especifican opciones en tablas de contingencia de 2 a n vías, PROC FREQ produce tablas de contingencia que incluyen en cada celda frecuencias, porcentajes respecto al total global, porcentajes por fila y por columna. El procedimiento excluye observaciones con valores faltantes de la tabla pero muestra la frecuencia total de valores faltantes en la parte inferior de cada tabla.

Opciones

La siguiente tabla lista las opciones disponibles con la declaración TABLES.

Controles para Análisis Estadístico

AGREE calcula pruebas y medidas de concordancia de clasificación.

ALL calcula pruebas y medidas de asociación producidas por CHISQ, MEASURES, y CMH.

ALPHA= especifica el nivel de confianza para los límites de confianza.

BINOMIAL calcula la proporción binomial, los límites de confianza y la prueba para tablas de una vía.

CHISQ calcula las pruebas chi-cuadrado y medidas de asociación basadas en la chi-cuadrado.

CL calcula límites de confianza para los estadísticos en la declaración MEASURES.

CMH calcula todos los estadísticos de Cochran-Mantel-Haenszel.

CMH1 calcula el estadístico de correlación CMH, riesgos relativos ajustados y razones de odds.

CMH2 calcula estadísticos de correlación CMH y scores medios por fila (ANOVA), riesgos relativos ajustados y razones de odds.

CONVERGE= especifica el criterio de convergencia para calcular la correlación polychoric.

FISHER calcula la prueba exacta de Fisher para las tablas de más de 2×2.

JT calcula la prueba de Jonckheere-Terpstra.

MAXITER= especifica el número máximo de iteraciones para calcular la correlación polychoric.

MEASURES calcula medidas de asociación y sus errores estándar asintóticos.

MISSING trata los valores faltantes como si no lo fueran.

PLCORR calcula la correlación polychoric.

RELRISK calcula medidas de riesgo relativas para tablas 2×2.

RISKDIFF calcula riesgos y diferencias de riesgo para tablas 2×2.

SCORES= especifica el tipo de scores por fila y columna.

TESTF= especifica frecuencias esperadas para la prueba chi-cuadrado en una tabla de una vía.

TESTP= especifica proporciones esperadas para la prueba chi-cuadrado en una tabla de una vía.

TREND calcula la prueba Cochran-Armitage para la tendencia

Controles para Información Adicional de las Tablas

CELLCHI2 muestra la contribución de cada celda al estadístico chi-cuadrado de Pearson.

CUMCOL muestra el porcentaje acumulado en la columna en cada celda.

DEVIATION muestra la desviación de la frecuencia observada del valor esperado en cada celda.

EXPECTED muestra la frecuencia esperada para cada celda.

MISSPRINT muestra las frecuencias de valores faltantes.

SPARSE lista todas las posibles combinaciones de niveles de las variables aun cuando una combinación no ocurre.

TOTPCT muestra el porcentaje de la frecuencia total en tablas de n vias cuando n>2.

Controles para la Salida Mostrada

NOCOL suprime el porcentaje de la columna para cada celda.

NOCUM suprime las frecuencias y porcentajes acumulados en tablas de frecuencia de una vía y en formato de lista.

NOFREQ suprime la frecuencia en cada celda.

NOPERCENT suprime los porcentajes en cada celda, por fila y por columna en tablas de contingencia, o los porcentajes, y porcentajes acumulados en tablas de frecuencia de una vía y en formato de lista.

NOPRINT suprime las tablas pero muestra los estadísticos.

NOROW suprime el porcentaje de la fila para cada celda.

LIST muestra las tablas de 2 a n vías en formato de lista.

PRINTKWT muestra los coeficientes ponderados de kappa.

SCOROUT muestra los scores por fila y por columna.

Crear un Archivo SAS de Salida

OUT= especifica un archivo de salida que contiene los valores de variable y conteos de frecuencia.

OUTEXPECT incluye la frecuencia esperada de cada celda en el archivo de salida.

OUTPCT incluye los porcentajes respecto a frecuencias por columna, por fila, y en tablas de dos vías en el archivo de salida.

DECLARACIÓN WEIGHT

```
WEIGHT variable ;
```

La declaración WEIGHT especifica una variable numérica cuyos valores representan la frecuencia de las observaciones. La declaración WEIGHT es comúnmente usada para entrar los conteos por celda. Al usar la declaración WEIGHT, PROC FREQ asume que una observación representa n observaciones, donde n es el valor de la variable. El valor de la variable weight no es necesariamente un entero. La suma de los valores de la variable weight representa el número total de observaciones.

Ejemplo 1. Crear un archivo de salida con las frecuencia por celda de una tabla:

El color de los ojos y del cabellos de niños de dos diferentes regiones de Europa son entrados en el archivo SAS Color. Una vez obtenida una observación por niño, los datos son llevados como conteos por celda, donde la variable CONTEO corresponde al número de niños encontrados que coinciden en cada una de las 15 combinaciones de colores de ojos y cabello.

```
data Color;
input Region Ojos $ Cabello $ Conteo @@;
label Ojos ='Color de los Ojos' Hair ='Color del Cabellos' Region='Región Geográfica';
cards;
1 azul  claro   23  1 azul  rojo     7  1 azul  castano 24
```

```
1 azul   oscuro   11  1 verde claro   19  1 verde rojo    7
1 verde castano 18  1 verde oscuro   14  1 cafe claro   34
1 cafe rojo      5  1 cafe castano 41  1 cafe oscuro   40
1 cafe negro     3  2 azul  claro   46  2 azul  rojo    21
2 azul  castano 44  2 azul  oscuro   40  2 azul  negro   6
2 verde claro   50  2 verde rojo     31  2 verde castano 37
2 verde oscuro  23  2 cafe claro    56  2 cafe rojo     42
2 cafe castano 53  2 cafe oscuro    54  2 cafe negro    13
;
run;
```

El siguiente programa lee el archivo COLOR y crea un archivo de salida que contiene las frecuencias, porcentajes, y frecuencias esperadas por celda de la tabla de dos vías entre OJOS y CABELLO. La declaración TABLES construye tres tablas: Tablas de frecuencia para OJOS y CABELLO, y una tabla de contingencia de OJOS por CABELLO. La opción OUT= crea el archivo FreqCnt, el cual contiene las frecuencias de la tabla de contingencia. La opción OUTEXPECT adiciona las frecuencias esperadas por celda al archivo FreqCnt, y la opción SPARSE incluye las celdas con conteos de cero. La declaración WEIGHT especifica que CONTEO contiene los pesos de las observaciones.

```
proc freq data Color;
weight Conteo;
tables Ojos Cabello Ojos*Cabello/out=FreqCnt outexpect;
title 'Color de Ojos y Cabello de niños Europeos';
run;
```

```
Color de Ojos y Cabello de ninos Europeos
The FREQ Procedure
```

Color de los Ojos

Ojos	Frequency	Percent	Cumulative Frequency	Cumulative Percent
azul	222	29.13	222	29.13
cafe	341	44.75	563	73.88
verde	199	26.12	762	100.00

Color del Cabello

Caballo	Frequency	Percent	Cumulative Frequency	Cumulative Percent
castano	217	28.48	217	28.48
claro	228	29.92	445	58.40
negro	22	2.89	467	61.29
oscuro	182	23.88	649	85.17
rojo	113	14.83	762	100.00

Table of Ojos by Cabello

Ojos(Color de los Ojos) Cabello(Color del Cabello)

Frequency|
Percent |
Row Pct |
Col Pct |

	castano	claro	negro	oscuro	rojo	Total
azul	68	69	6	51	28	222
	8.92	9.06	0.79	6.69	3.67	29.13
	30.63	31.08	2.70	22.97	12.61	
	31.34	30.26	27.27	28.02	24.78	
cafe	94	90	16	94	47	341
	12.34	11.81	2.10	12.34	6.17	44.75
	27.57	26.39	4.69	27.57	13.78	
	43.32	39.47	72.73	51.65	41.59	
verde	55	69	0	37	38	199
	7.22	9.06	0.00	4.86	4.99	26.12
	27.64	34.67	0.00	18.59	19.10	
	25.35	30.26	0.00	20.33	33.63	
Total	217	228	22	182	113	762
	28.48	29.92	2.89	23.88	14.83	100.00

Por defecto, PROC FREQ muestra los valores de las variables en orden alfabético. La especificación 'Ojos*Cabello' produce una tabla de contingencia con Color de Ojos como las filas y Color del Cabello como las columnas en la tabla. Una celda con conteo cero para Ojos verdes y Cabello negro indica que esta combinación de color de Ojos y Cabello no ocurre en los datos.

```
proc print data=FreqCnt noobs;
title2 'Archivo de salida del PROC FREQ';
run;
```

Color de Ojos y Cabello de ninos Europeos

Archivo de salida del PROC FREQ

Ojos	Cabello	COUNT	EXPECTED	PERCENT
azul	castano	68	63.220	8.9239
azul	claro	69	66.425	9.0551
azul	negro	6	6.409	0.7874
azul	oscuro	51	53.024	6.6929
azul	rojo	28	32.921	3.6745
cafe	castano	94	97.109	12.3360
cafe	claro	90	102.031	11.8110
cafe	negro	16	9.845	2.0997
cafe	oscuro	94	81.446	12.3360
cafe	rojo	47	50.568	6.1680
verde	castano	55	56.671	7.2178
verde	claro	69	59.543	9.0551
verde	negro	0	5.745	0.0000
verde	oscuro	37	47.530	4.8556
verde	rojo	38	29.510	4.9869

El archivo de salida contiene conteos de frecuencia y porcentajes para la última tabla. El archivo incluye también una observación para la celda de conteo cero (SPARSE) y una variable con la frecuencia esperada por celda (OUTEXPECT).

Ejemplo 2. Calcular pruebas Chi-cuadrado para tablas de frecuencias de una vía.

Se examinará si el color de Cabello de los niños (del ejemplo anterior) tiene una distribución multinomial para las dos regiones. La distribución hipotética para el color de Cabello es 30% claro, 12% rojo, 30% castaño, 25% oscuro, y 3% negro. Para que se realice una prueba de hipótesis para cada región, los datos son primero ordenados por la variable REGION. Entonces el procedimiento FREQ usa una declaración BY para producir una tabla separada para cada grupo BY (Region).

La opción ORDER=DATA ordena los valores de la tabla de frecuencias (color de cabello) por su orden en el archivo original. La declaración TABLES calcula una tabla de frecuencia para el color de cabello, y la opción NOCUM suprime las frecuencias acumuladas y porcentajes.

La opción TESTP= especifica los porcentajes hipotéticos para la prueba chi-cuadrado; el número de porcentajes especificados es igual al número de niveles de la tabla, y los porcentajes suman 100.


```
proc sort data=Color;
by Region;run;
proc freq data=Color order=data;
weight Count;
tables Cabello/nocum testp=(30 12 30 25 3);
by Region;
title 'Color de Cabello de Niños Europeos';
run;
```

Color de Cabello de Niños Europeos

Región Geográfica=1

The FREQ Procedure

Color del Cabello

Cabello	Frequency	Percent	Test Percent
claro	76	30.89	30.00
rojo	19	7.72	12.00
castano	83	33.74	30.00
oscuro	65	26.42	25.00
negro	3	1.22	3.00

Chi-Square Test for Specifiad Proportions

Chi-Square	7.7602
DF	4
Pr > ChiSq	0.1008

Sample Size = 246

Color de Cabello de Niños Europeos

Región Geográfica=2

The FREQ Procedure

Color del Cabello

Cabello	Frequency	Percent	Test Percent
claro	152	29.46	30.00
rojo	94	18.22	12.00

castano	134	25.97	30.00
oscuro	117	22.67	25.00
negro	19	3.68	3.00

Chi-Square Test
for Specified Proportions

Chi-Square	21.3824
DF	4
Pr > ChiSq	0.0003
Sample Size = 516	

La columna "Test Percent" lista los porcentajes hipotéticos para la prueba chi-cuadrado. Siempre se debe chequear que el orden de los porcentajes en la opción TESTP= coinciden correctamente con el orden de los niveles de la variable.

PROC FREQ calcula un estadístico chi-cuadrado para cada región. Dicho estadístico es significativo al nivel 0.05 para la Región 2 (p=0.0003) pero no lo es para la Region 1. Esto indica que los porcentajes hipotéticos son significativos en la Region 2.

Ejemplo 3. Calcular proporciones binomiales para tablas de frecuencia de una vía

La proporción binomial es calculada como la proporción de observaciones para el primer nivel de la variable que se está estudiando. El programa siguiente calcula la proporción de niños con ojos cafés y prueba este valor contra la hipótesis que la proporción es 50%. También, se prueba si la proporción de niños con cabello claro es 28%.

```
proc freq data=Color order=freq;
weight Count;
tables Eyes / binomial alpha=.1;
tables Hair / binomial(p=.28);
title 'Hair and Eye Color of European Children';
run;
```

La primera declaración TABLES produce una tabla de frecuencias para el color de los ojos. La opción BINOMIAL calcula la proporción binomial, los límites de confianza, y las pruebas de hipótesis de que esa proporción para el primer nivel de color de ojos (café) es 0.5. La opción ALPHA=.1 especifica que los límites del 90% serán calculados. La segunda declaración TABLES crea una tabla de frecuencias

para el color de cabello y calcula la proporción binomial, los límites de confianza, y realiza una prueba de que la proporción para el primer color de cabello (claro) es 0.28.

Color de Cabello y Ojos de ninos Europeos

The FREQ Procedure

Color de los Ojos

Ojos	Frequency	Percent	Cumulative Frequency	Cumulative Percent
cafe	341	44.75	341	44.75
azul	222	29.13	563	73.88
verde	199	26.12	762	100.00

Binomial Proportion
for Ojos = cafe

Proportion	0.4475
ASE	0.0180
90% Lower Conf Limit	0.4179
90% Upper Conf Limit	0.4771

Exact Conf Limits

90% Lower Conf Limit	0.4174
90% Upper Conf Limit	0.4779

Test of H0: Proportion = 0.5

ASE under H0	0.0181
Z	-2.8981
One-sided Pr < Z	0.0019
Two-sided Pr > |Z|	0.0038
Sample Size = 762	

La tabla de frecuencias muestra los valores de la variable en orden descendente de frecuencias. Así el primer nivel de variable es 'café', PROC FREQ calcula la proporción binomial de niños con ojos cafés. PROC FREQ también calcula sus errores estándar asintóticos (ASE), y límites de confianza asintóticos y exactos del 90%. Si no se especifica la opción ALPHA=, entonces PROC FREQ calcula por defecto los límites de confianza del 95%. Debido a que el valor de Z es menor que cero, PROC FREQ calcula

un valor-p de cola izquierda (0.0019). Este valor-p pequeño soporta la hipótesis alternativa de que el valor verdadero de la proporción de niños con ojos cafés es menor del 50%.

Color del Cabello

Cabello	Frequency	Percent	Cumulative Frequency	Cumulative Percent
claro	228	29.92	228	29.92
castano	217	28.48	445	58.40
oscuro	182	23.88	627	82.28
rojo	113	14.83	740	97.11
negro	22	2.89	762	100.00

Binomial Proportion
for Cabello = claro

Proportion	0.2992
ASE	0.0166
95% Lower Conf Limit	0.2667
95% Upper Conf Limit	0.3317
Exact Conf Limits	
95% Lower Conf Limit	0.2669
95% Upper Conf Limit	0.3331
Test of H0: Proportion = 0.28	
ASE under H0	0.0163
Z	1.1812
One-sided Pr > Z	0.1188
Two-sided Pr > |Z|	0.2375
Sample Size = 762	

Los anteriores resultados corresponden a la segunda declaración TABLES. PROC FREQ calcula los límites de confianza del 95%, ya que la opción ALPHA= no es especificada. El valor de Z es mayor que cero, de manera que PROC FREQ calcula el valor-p de cola derecha (0.1188). Este valor-p proporciona evidencia insuficiente para rechazar la hipótesis nula de que la proporción de niños con cabello claro es 28%.

Ejemplo 4. Analizando una tabla de contingencia 2x2

Este ejemplo calcula pruebas chi-cuadrado y la prueba exacta de Fisher para comparar la probabilidad de enfermedades coronarias para dos tipos de dieta. Esto también estima los riesgos relativos y calcula límites de confianza exactos para las razones de odds. El archivo FatComp contiene datos hipotéticos para un estudio de caso-control de dietas altas en grasas y el riesgo de enfermedades coronarias. Los datos son tomados como conteos por celda, donde la variable CONTEO corresponde a las frecuencias para cada combinación de exposición y respuesta. Los datos están ordenados descendentemente por las variables Exposición y Respuesta, así que la primera celda de la tabla 2×2 contiene la frecuencia de una exposición positiva y una respuesta positiva. El procedimiento FORMAT crea formatos para identificar el tipo de exposición y respuesta con valores tipo caracter.

```
proc format;
value ExpFmt 1='High Cholesterol Diet' 0='Low Cholesterol Diet';
value RspFmt 1='Yes' 0='No';
run;

data FatComp;
input Exposure Response Count;
label Response='Heart Disease';
datalines;
0 0  6
0 1  2
1 0  4
1 1 11
;
run;

proc sort data=FatComp;
by descending Exposure descending Response;
run;
```

En el siguiente programa, la declaración TABLES crea una tabla de dos vías, y la opción ORDER=DATA ordena los valores de la tabla de contingencia values por su orden en los datos originales. La opción CHISQ produce varias pruebas chi-square, mientras que la opción RELRISK produce medidas de riesgo relativo. La declaración EXACT crea la prueba exacta chi-cuadrado de Pearson y límites de confianza exactos para la razón de odds.

```
proc freq data=FatComp order=data;
weight Count;
tables Exposure*Response / chisq relrisk;
exact pchi or;
format Exposure ExpFmt. Response RspFmt.;
title Estudio de Caso-Control de Dietas Altas en Grasa/Colesterol';
run;
```

```
Estudio de Caso-Control de Dietas Altas en Grasa/Colesterol
The FREQ Procedure
Table of Exposicion by Respuesta
Exposicion        Respuesta(Enfermedades del corazón)
Frequency         |
Percent           |
Row Pct           |
Col Pct           |Si      |No      |  Total
------------------+--------+--------+
Dieta Alta en Co  |     11 |      4 |     15
lesterol          |  47.83 |  17.39 |  65.22
                  |  73.33 |  26.67 |
                  |  84.62 |  40.00 |
------------------+--------+--------+
Dieta Baja en Co  |      2 |      6 |      8
lesterol          |   8.70 |  26.09 |  34.78
                  |  25.00 |  75.00 |
                  |  15.38 |  60.00 |
------------------+--------+--------+
Total                   13       10        23
                     56.52    43.48    100.00
```

La tabla de contingencia muestra los valores de variables de modo que la primera celda contiene la frecuencia en la primera celda en el archivo de datos, esto es, la frecuencia de exposición y respuesta positiva (cuando ambos toman el valor de 1).

```
Statistics for Table of Exposicion by Respuesta
Statistic                      DF       Value      Prob
-------------------------------------------------------
Chi-Square                      1      4.9597    0.0259
Likelihood Ratio Chi-Square     1      5.0975    0.0240
Continuity Adj. Chi-Square      1      3.1879    0.0742
Mantel-Haenszel Chi-Square      1      4.7441    0.0294
Phi Coefficient                        0.4644
Contingency Coefficient                0.4212
Cramer's V                             0.4644
WARNING: 50% of the cells have expected counts less than 5.
         (Asymptotic) Chi-Square may not be a valid test.
```



```
Pearson Chi-Square Test
----------------------------------
Chi-Square                   4.9597
DF                                1
Asymptotic Pr >  ChiSq       0.0259
Exact      Pr >= ChiSq       0.0393
Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)         11
Left-sided Pr <= F           0.9967
Right-sided Pr >= F          0.0367
Table Probability (P)        0.0334
Two-sided Pr <= P            0.0393
```

Debido a que los conteos esperados en algunas celdas son pequeños, PROC FREQ muestra un mensaje de advertencia de que las pruebas chi-cuadrado asintóticas podrían no ser apropiadas. En este caso, las pruebas exactas serán las adecuadas. La hipótesis alternativa para este análisis establece que las enfermedades coronarias están más asociadas con una dieta alta en grasas, así que una prueba unilateral es requerida. La prueba exacta de cola derecha de Fisher analiza si la probabilidad de enfermedades coronarias en el grupo alto en grasa excede la probabilidad de enfermedades coronarias en el grupo bajo en grasa; así que si el valor-p es pequeño, la hipótesis alternativa es soportada.

```
Estimates of the Relative Risk (Row1/Row2)
Type of Study                  Value      95% Confidence Limits
-------------------------------------------------------------
Case-Control (Odds Ratio)     8.2500      1.1535       59.0029
Cohort (Col1 Risk)            2.9333      0.8502       10.1204
Cohort (Col2 Risk)            0.3556      0.1403        0.9009

  Odds Ratio (Case-Control Study)
-----------------------------
```

```
Odds Ratio                    8.2500
Asymptotic Conf Limits
95% Lower Conf Limit          1.1535
95% Upper Conf Limit         59.0029


Exact Conf Limits
95% Lower Conf Limit          0.8677
95% Upper Conf Limit        105.5488
Sample Size = 23
```

2.5.4 PROCEDIMIENTOS DE REGRESIÓN

2.5.4.1 PROC REG

Es el procedimiento SAS para hacer regresión lineal simple y múltiple. La sintaxis más usada es como sigue:

```
PROC REG  DATA=archivo_sas< opciones > ;
< etiqueta: > MODEL variable_dependientes=<variable_explicatorias> < / opciones > ;
BY variables ;
FREQ variable ;
VAR variables ;
WEIGHT variable ;
<etiqueta: > MTEST <ecuación, ... ,ecuación> < / opciones > ;
PLOT <yvariable*xvariable> <=símbolo1>
      < ...yvariable*xvariable> <=símbolo> < / opciones > ;
< etiqueta: > TEST ecuación,<, ...,ecuación>  ;
OUTPUT < OUT=Archivo_SAS > palabras_claves=nombres;
run;
```

En la lista anterior, los < > denota una especificación opcional,. En todos los casos la etiqueta es opcional.

Para ajustar un modelo, éste debe especificarse con la declaración **MODEL**. Si sólo se requiere usar las opciones disponibles en la declaración **PROC REG**, no se necesita usar **MODEL**, pero se debe usar **VAR**. Varias declaraciones **MODEL** pueden usarse. También varias declaraciones **MTEST**, **OUTPUT**, **PAINT**, **PLOT**, **PRINT**, **RESTRICT**, y **TEST** pueden ser usadas pero a continuación de cada declaración **MODEL**. Las declaraciones **BY**, **FREQ**, **VAR**, y **WEIGHT** se especifican opcionalmente y una sola vez para todo el paso **PROC REG**.

Veamos algunos detalles: Las variables dependiente y predictoras deben ser de tipo numéricas.

Algunas de las opciones posibles en la línea del **PROC REG**, son:

ALL solicita la presentación de muchas tablas y es equivalente a usar la opción **ALL** en cada declaración Model. **ALL** también implica las opciones **CORR, SIMPLE, y USSCP.**

ALPHA=número (entre 0 y 1) Ajusta el nivel de significancia usado en la construcción de intervalos de confianza; por defecto es 0.05. Esta opción afecta a las opciones **TABLEOUT** de la línea del **PROC REG**; a las opciones **CLB, CLI, y CLM** de la declaración **MODEL**; a los valores de **LCL, LCLM, UCL, y UCLM** de la declaración **OUTPUT**; a los valores de **LCL., LCLM., UCL., UCLM, CONF y PRED** de la declaración **PLOT**.

CORR muestra la matriz de correlación entre todas las variables listadas en la declaración **MODEL** o **VAR.**

COVOUT produce la matriz de covarianzas de los parámetros estimados para la base de datos SAS especificada con **OUTEST=archivo_SAS, COVOUT** es válida si se usa **OUTEST=archivo_SAS**.

EDF ó RSQUARE produce el número de regresores en el modelo incluyendo y excluyendo el intercepto, los grados de libertad del error y el R^2, para la base de datos SAS especificada con **OUTEST=archivo_SAS.**

NOPRINT suprime la salida por pantalla de los resultados del procedimiento.

OUTEST=Archivo_SAS solicita que los parámetros y estadísticos opcionales de resumen del ajuste del modelo sean direccionados a dicho conjunto de datos SAS.

OUTSEB produce los errores estándar de los parámetros estimados para **OUTEST=archivo_SAS**, en este conjunto de datos los valores SEB de la variable **_TYPE_** identifican los errores estándar. Esta opción está disponible para todos los métodos de selección excepto para **RSQUARE, ADJRSQ, y CP**

OUTSTB esta opción produce estimaciones de los parámetros estandarizados como también las usuales para **OUTEST=archivo_SAS** cuando la opción **RIDGE= o PCOMIT=** es especificada.

OUTVIF produce los factores de inflación de varianza VIF para **OUTEST=archivo_SAS** cuando la opción **RIDGE= o PCOMIT=** es especificada. Los factores son los elementos de la diagonal de la inversa de la matriz de correlación de los regresores ajustados por la regresión ridge o análisis IPC.

PRESS produce el estadístico PRESS para la base de datos SAS creada con **OUTEST=archivo_SAS.** Los valores de este estadístico son guardados en la variable **_PRESS_**. Esta opción está disponible para todos los métodos de selección de modelos excepto **RSQUARE, ADJRSQ, y CP.**

RIDGE=lista requiere un análisis de regresión Ridge y especifica los valores de la constante ridge k. Cada valor de k produce un conjunto de estimaciones de regresión ridge que son colocados en

OUTEST=archivo_SAS. Los valores de k son guardados en la variable **_RIDGE_** y el valor de la variable **_TYPE_** es RIDGE para identificar las estimaciones.

TABLEOUT produce para el conjunto de datos SAS creado con **OUTEST=archivo_SAS**, los errores estándar y los límites de confianza (1-alpha)100% para las estimaciones de los parámetros, los estadísticos t para la prueba de siginficancia de los parámetros, y los valores p asociados a dichas pruebas, La variable **_TYPE_** toma respectivamente los valores **STDERR, LNB, UNB, T, y PVALUE**, para identificar los anteriores valores en el **OUTEST=** . Por tanto es necesario que se use la opción **OUTEST=archivo_SAS** para que TABLEOUT tenga efecto.

Las opciones en la declaración **MODEL**, son entre otras:

ADJRSQ calcula el R^2 ajustado por los grados de libertad para cada modelo seleccionado.

AIC calcula el criterio de información de Akaike para cada modelo seleccionado.

ALL solicita todas las siguientes opciones: **ACOV, CLB, CLI, CLM, CORRB, COVB, I, P, PCORR1, PCORR2, R, SCORR1, SCORR2, SEQB, SPEC, SS1, SS2, STB, TOL, VIF, y XPX.**

ALPHA=número ajusta el nivel de significancia para los intervalos de confianza y de predicción para el modelo.

B se usa con los métodos se selección de modelo basados en **RSQUARE, ADJRSQ, y CP**, y calcula los coeficientes de regresión estimados para cada modelo seleccionado.

BEST=n se usa con los métodos de selección de modelos **RSQUARE, ADJRSQ, y CP**. Si se especifica **SELECTION=CP o SELECTION=ADJRSQ**, la opción **BEST=n** especifica el número máximo de submodelos a presentar por pantalla o guardar en un data con la opción **OUTEST=archivo_SAS**. Para **SELECTION=RSQUARE**, la opción **BEST=n** especifica el número máximo de submodelos para cada tamaño. Si la opción **BEST=n** es usada sin la opción **B**, las variables en cada **MODEL** son listadas en el orden de inclusión en vez del orden en el cual ellas aparecen en la declaración **MODEL**. Si la opción **BEST** es omitida y el número de regresores es menor que 11, todos los posibles submodelos son evaluados.

BIC calcula el criterio de información bayesiana para cada modelo seleccionado.

CLB calcula los intervalos de confianza de (1-alpha)100% para los parámetros.

CLI calcula los intervalos de predicción de (1-alpha)100% para las respuestas individuales en todos los valores de las variables regresoras especificadas en el DATA.

CLM calcula los intervalos de confianza de (1-alpha)100% para la respuesta media en cada observación.

COLLIN solicita un análisis detallado de colinealidad entre regresores. Incluye valores propios, índices de condición, y descomposición de las varianzas de las estimaciones con respecto a cada valor propio.

COLLINOINT lo mismo que **COLLIN** pero excluyendo al intercepto de los diagnósticos.

CORRB produce la matriz de correlación de los estimadores, es decir la matriz $(X'X)^{-1}$ escalada a diagonal unitaria.

COVB produce la matriz de varianzas covarianzas de los parámetros estimados.

CP calcula el estadístico Cp de Mallows para cada modelo seleccionado.

DETAILS=nombre especifica el nivel de detalle cuando se usan los métodos de selección secuencial **BACKWARD, FORWARD o STEPWISE**. Los nombres pueden ser **ALL, STEPS o SUMMARY**. Por defecto es **STEP**, si se omite esta opción en un proceso de selección secuencia.

DW Produce el estadístico Durbin - Watson para probar autocorrelación de primer orden entre los errores.

I presenta la matriz $(X'X)^{-1}$.

INFLUENCE requiere un análisis detallado de la influencia de cada observación sobre las estimaciones y los valores predichos.

MSE calcula el MSE de cada modelo seleccionado.

NOINT corre una regresión por el origen (sin intercepto).

P produce en la salida los valores predichos del modelo ajustado, incluyendo el número de la observación, el valor actual y los residuales. Si se invoca **CLI, CLM, o R**, la opción **P** es innecesaria.

PARTIAL solicita gráficos leverage de regresión parcial para cada regresor.

R produce en la salida análisis de los residuales del modelo ajustado. Los resultados incluyen los mismos de la opción **P** más los errores estándar de la media predicha y de los residuales, los residuales estudentizados, el estadístico *Cook's D* para medir la influencia de cada observación sobre los parámetros estimados.

RMSE produce la raíz cuadrada del MSE para cada modelo seleccionado.

SBC calcula el estadístico SBC para cada modelo seleccionado.

SELECTION=nombre especifica el método de selección de modelos usado, donde nombre puede ser **FORWARD (ó F), BACKWARD (ó B), STEPWISE, MAXR, MINR, RSQUARE, ADJRSQ, CP, o NONE**, este último es el valor por defecto de esta opción y produce el modelo completo.

SEQB produce una secuencia de estimaciones de parámetros a medida que cada variable es ingresada al modelo. Esto es presentado como una matriz donde cada fila es un conjunto de parámetros estimados.

SLENTRY=valor ó SLE=valor especifica el nivel de significancia para entrar una variable al modelo cuando se usan los métodos **FORWARD y STEPWISE.**

SLSTAY=valor especifica el nivel de significancia para que una variable permanezca en el modelo cuando se usan los **métodos BACKWARD y STEPWISE.**

SS1 despliega la sumas de cuadrados secuenciales (sumas de cuadrados tipo I) junto con las estimaciones de los parámetros para cada término en el modelo.

SS2 despliega las sumas de cuadrados parciales (sumas de cuadrados tipo II) junto con las estimaciones para cada término en el modelo.

SSE calcula el SSE para cada modelo seleccionado.

STB produce los coeficientes de regresión estandarizados. Un coeficiente de regresión estandarizado es obtenido dividiendo el parámetro estimado por la razón de la desviación estándar muestral de la variable dependiente a la desviación estándar muestral de la variable explicatoria correspondiente.

TOL produce los valores de tolerancia para las estimaciones. La tolerancia para una variable es definida como $1-R^2$, donde R^2 es obtenido de la regresión de la variable sobre todas las demás variables explicatorias en el modelo.

VIF produce los factores de inflación de varianza con los parámetros estimados. El VIF es el recíproco de la tolerancia.

XPX produce la matriz de productos cruzados X'X, la cual es rodeada por las matrices X'Y y Y'Y.

BY variable realiza análisis separados de las observaciones según los niveles de la variable de clasificación especificada en **BY**. Por defecto esta declaración asume que las observaciones están ordenadas en forma creciente o alfabética según los niveles de la variable de clasificación. Si los datos no han sido ordenados previamente, se puede usar la opción **NOTSORTED** o si los datos están ordenados en forma descendente se usa la opción **DESCENDING**.

La declaración **FREQ variable** se usa si en los datos existe una variable que exprese las frecuencias de cada una de las observaciones. Por defecto se asume que la frecuencia de cada observación es 1.

La declaración **WEIGHT variable** esta declaración nombra una variable en los datos que contiene los valores que son los pesos relativos para un ajuste de mínimos cuadrados ponderados. Las observaciones con peso cero o negativo, son omitidas del ajuste.

La declaracióu **OUTPUT < OUT=Archivo_SAS > Palabras_claves=nombres** crea un nuevo conjunto de datos SAS que guarda los estadísticos y medidas de diagnósticos de cada observación,

indicados con las palabras claves, del último modelo ajustado (la declaración **MODEL** más reciente). Se incluyen los datos originales junto con las variables creadas con el **OUTPUT**. Las palabras claves que pueden invocarse son (en todos los casos, nombre es el nombre que ud. da al respectivo estadístico o medida, con el cual será identificado dentro del archivo SAS creado en el output):

COOKD=nombre Distancia de Cook.

COVRATIO=nombre medida de influencia de las observaciones sobre la covarianza de los betas.

DFFITS=nombre medida de influencia de las observaciones sobre los valores predichos.

H=nombre leverage, $x_i(X'X)^{-1}x_i'$.

LCL=nombre límite inferior del intervalo de predicción individual.

LCLM=nombre límite inferior del intervalo de confianza para la respuesta media.

UCL=nombre límite superior del intervalo de predicción individual.

UCLM=nombre límite superior del intervalo de confianza para la respuesta media.

PREDICTED=nombre ó P=nombre valores predichos.

PRESS=nombre i-ésimo residual dividido por (1-h), h es el leverage, y donde el modelo ha sido ajustado sin la i-ésima observación.

RESIDUAL=nombre ó R=nombre residuales calculado como el ACTUAL menos PREDICTED.

RSTUDENT=nombre residuales estudentizados con la observación actual borrada.

STDI=nombre error estándar de los valores individuales predichos.

STDP=nombre error estándar de los valores medios ajustados.

STDR=nombre error estándar de los residuales.

STUDENT=nombre residuales estudentizados, los cuales son los residuales divididos por sus errores estándar.

La declaración **PLOT < yvariable*xvariable >< =símbolo > < ... yvariable*xvariable > < =símbolo > < / opciones >** crea gráficos de dispersión con yvariable la variable en el eje vertical y xvariable la variable en el eje horizontal. Los gráficos son en alta resolución a menos que se especifique opciones como **HPLOTS=n VPLOTS=m** (n es el número de gráficos por renglón o fila y m el número de gráficos por columna). Más de un gráfico puede ser invocado en un mismo plot, y si se quieren superponer se usa la opción **OVERLAY**. Los ajustes gráficos especificados con **GOPTIONS** y **SYMBOL** más recientes, son usados para los gráficos del PROC REG.

Cuando en los gráficos una de las variables es un estadístico o medida de la regresión, deben usarse las palabras claves mencionadas antes en la declaración **OUTPUT**, pero seguidas por un punto, por ejemplo: **COOKD. , COVRATIO. , DFFITS. , H. , LCL. , LCLM., PREDICTED. | PRED. | P.,**

PRESS. , **RESIDUAL.** | **R.** , etc. las opciones **CONF** y **PRED** agregan las bandas de confianza y de predicción al gráfico de Y vs. X (si es una regresión simple).

Otras palabras claves usadas con estadísticos de diagnóstico, son:

NPP. para crear un gráfico de probabilidad normal, **NQQ.** para crear un gráfico cuantil-cuantil normal. También se puede graficar estadísticos resumen de ajuste de modelos, contra el número de parámetros, **NP.** o contra el número de variables predictoras **IN.** , en el modelo incluyendo el intercepto: **ADJRSQ.** el R cuadrado ajustado, **AIC.** el criterio de Akaike, **BIC.** El criterio bayesiano, **CP.** el estadístico Cp de Mallows, **MSE.** el error cuadrático medio, **RMSE.** La raíz cuadrada del MSE , **SSE.** La suma de cuadrados del error, o **RSQ.** el **R2.** , entre otros.

La declaración < **etiqueta:** > **TEST ecuación** < , ... , **ecuación** > < / **opciones** > permite probar hipótesis acerca de los parámetros estimados en la declaración **MODEL** precedente. Cada ecuación especifica una hipótesis lineal a ser probada. Las filas de las hipótesis son separadas por comas. Se usan los nombres de las variables regresoras y cada nombre de variable invocada representa el correspondiente coeficiente de la variable en el modelo. Se puede usar etiquetas para identificar cada prueba. La palabra clave **INTERCEPT** puede usarse para referirse al intercepto del modelo. El PROC REG realiza pruebas F para las hipótesis conjuntas especificadas en una sola declaración **TEST**. Más de una declaración TEST puede invocarse para un mismo modelo. Por ejemplo

```
model y=x1 x2 x3 x4;
test1: test x1+x2=1;
test2:    test x3=0, x4=0;
test3:    test x3, x4;
```

las dos últimas declaraciones son equivalentes.

Veamos algunos ejemplos:

```
PROC REG DATA=paciente ALL;
  MODEL presart=edad;
RUN;
```

Al ejecutar este programa (desde luego después de haber entrado lo datos con el paso DATA), produce una salida con todos los resultados de la regresión.

Para guardar los resultados y graficarlos:

Si se utiliza la instrucción PLOT se especifica la variables que se quieran graficar. Esta instrucción se ilustra en el siguiente ejemplo:


```
PROC REG DATA=paciente ALL;
  MODEL presart=edad;
  PLOT presart*edad='*';
RUN;
```

Para colocar una gráficas sobre otras podemos realizar lo siguiente:

```
PROC REG DATA=paciente ALL;
  MODEL presart=edad;
  PLOT presart*edad= L95M.*edad U95M.*edad / overlay;
RUN;
```

Mostrará una gráfica con las bandas de confianza para la media del 95% de confianza. Si no se coloca el overlay se generarán tres gráficas diferentes.

Si queremos obtener varias gráficas (en baja resolución) en una sola página realizamos lo siguiente:

```
PROC REG DATA=paciente ALL;
  MODEL presart=edad;
  PLOT presart*edad= L95M.*edad U95M.*edad / VPLOT=3;
RUN;
```

Generará tres gráficas diferentes en una página. Si se quieren horizontales se utiliza la opción HPLOT=número, donde número indica la cantidad de gráficas a lo ancho de la hoja. Se pueden realizar combinaciones de tal forma que quede una matriz de gráficas.

En realidad este PROC tiene varias opciones que pueden ser consultadas en el manual. Además un poco de curiosidad por mirar "qué sucede si ..." sirve de mucha ayuda para entender muchas de las opciones.

Ejemplo 1:

Veamos un ejemplo que puede aclarar el uso de este PROC. El ejemplo a continuación presentado corresponde a un ejercicio en el libro PROBABILIDAD Y ESTADISTICA de Walpole y Myers, cuarta edición, pág. 382 ejercicio 7.

Un comerciante al menudeo llevó a cabo un estudio para determinar la relación entre los gastos de publicidad semanal y las ventas.

a) Dibuje un diagrama de dispersión.

b) Encuentre la ecuación de la línea de regresión para pronosticar las ventas semanales resultantes de los gastos de publicidad.

c) Estime las ventas mensuales cuando los gastos de publicidad ascienden a $35.

A continuación se muestra el listado del programa con los datos del problema:

```
/*****************************************************/
/* Ejemplo del uso del PROC REG.  Datos tomados del libro de*/
```

```
/* Walpole 4. ed., pag. 381 Ejercicio 7        */
/*************************************************************/
OPTIONS NODATE NOCENTER NONUMBER;
goptions reset=global gunit=pct border cback=white
colors=(black blue green red)
ftitle=swissb ftext=swiss htitle=6 htext=3;
DATA COMERC;
 INPUT PUBLIC VENTAS;
 CARDS;
  40  385
  20  400
  25  395
  20  365
  30  475
  50  440
  40  490
  20  420
  50  560
  40  525
  25  480
  50  510
  35   .
 ;
RUN;
PROC REG DATA=COMERC USSCP;
 MODEL VENTAS=PUBLIC/P;
 PLOT VENTAS*PUBLIC/CONF PRED;
 PLOT R.*(P. PUBLIC NQQ.) NPP *R.;
 OUTPUT OUT=RESID R=RES;
 RUN;QUIT;
```

Uncorrected Sums of Squares and Crossproducts

Variable	Intercept	PUBLIC	VENTAS
Intercept	12	410	5445
PUBLIC	410	15650	191325
VENTAS	5445	191325	2512925

Model: MODEL1

Dependent Variable: VENTAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	17030	17030	6.75	0.0266
Error	10	25226	2522.62056		
Corrected Total	11	42256			

Root MSE	50.22570	R-Square	0.4030
Dependent Mean	453.75000	Adj R-Sq	0.3433
Coeff Var	11.06902		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	343.70558	44.76618	7.68	<.0001
PUBLIC	1	3.22081	1.23961	2.60	0.0266

Output Statistics

Obs	Dep Var VENTAS	Predicted Value	Residual
1	385.0000	472.5381	-87.5381
2	400.0000	408.1218	-8.1218
3	395.0000	424.2259	-29.2259
4	365.0000	408.1218	-43.1218
5	475.0000	440.3299	34.6701
6	440.0000	504.7462	-64.7462
7	490.0000	472.5381	17.4619
8	420.0000	408.1218	11.8782
9	560.0000	504.7462	55.2538
10	525.0000	472.5381	52.4619
11	480.0000	424.2259	55.7741
12	510.0000	504.7462	5.2538
13	.	456.4340	.

Sum of Residuals	0
Sum of Squered Residuals	25226
Predicted Residual SS (PRESS)	35848