# Computación Estadística

# EPG3308

Profesora: María Inés Godoy,

Ayudante: María De Los Angeles Villena

Primer Semestre 2017: 18 abril

# PROC FREQ

El procedimiento FREQ, produce tablas de contingencia y frecuencias. Para las tablas de 2 entradas este procedimiento calcula test y medidas de asociación. La sintaxis corresponde a,

**PROC FREQ** < options > ;

**BY** variables;

**EXACT** statistic-options < / computation-options> ;

**OUTPUT** <OUT=SAS-data-set > output-options;

**TABLES** requests < /options > ;

**TEST** options;

**WEIGHT** variable < / option >;

# PROCEDIMIENTO FREQ

- **PROC FREQ** Invoca el procedimiento.
- **BY** Proporciona análisis separados para cada por grupo.
- **EXACT** Las solicitudes exactas para los test.
- **OUTPUT** Requiere un conjunto de datos de Salida.
- **TABLES** Especifica tablas y requerimiento de análisis.
- **TEST** Requisitos del test para las medidas de asociación.
- **WEIGHT** Identifica la variable de peso.

# PROCEDIMIENTO FREQ

```
data Color;
input Region Eyes $ Hair $ Count @@;
label Eyes  ='Eye Color'
Hair  ='Hair Color'
Region='Geographic Region';
datalines;
1 blue  fair   23  1 blue  red     7  1 blue  medium 24
1 blue  dark   11  1 green fair   19  1 green red     7
1 green medium 18  1 green dark   14  1 brown fair   34
1 brown red     5  1 brown medium 41  1 brown dark   40
1 brown black   3  2 blue  fair   46  2 blue  red    21
2 blue  medium 44  2 blue  dark   40  2 blue  black   6
2 green fair   50  2 green red    31  2 green medium 37
2 green dark   23  2 brown fair   56  2 brown red    42
2 brown medium 53  2 brown dark   54  2 brown black  13
;
```

# PROCEDIMIENTO FREQ

```
proc freq data=Color;
tables Eyes Hair Eyes*Hair / out=FreqCount outexpect ;
weight Count;
title 'Eye and Hair Color of European Children';
run;
proc print data=FreqCount noobs;
title2 'Output Data Set from PROC FREQ';
run;
```

# PROCEDIMIENTO FREQ

```
proc freq data=Color order=freq;
tables Hair Hair*Eyes / plots=freqplot(type=dotplot);
tables Hair*Region / plots=freqplot(type=dotplot scale=percent);
weight Count;
title 'Eye and Hair Color of European Children';
run;
```

# PROCEDIMIENTO FREQ

```
proc freq data=Color order=data;
tables Hair / nocum chisq testp=(30 12 30 25 3)
plots(only)=deviationplot(type=dotplot);
weight Count;
by Region;
title 'Hair Color of European Children';
run;
```

# PROC CORR

El procedimiento CORR calcula los coeficientes de correlación de Pearson, tres medidas de asociación no paramétricas, y las probabilidades asociadas a estas estadísticas. Las estadísticas de correlación incluyen los siguientes:

- Correlación momento-producto de Pearson (es una medida paramétrica de una relación lineal entre dos variables).
- Correlación por rangos de Sperman (utiliza las filas de los valores de los datos).
- Coeficiente de Tau-b de Kendall (utiliza el número de concordancias y discordancias en las observaciones pareadas).
- Medida de dependencia D de Hoeffding (es otra medida no paramétrica de asociación que detecta salidas más generales de la independencia).
- Correlaciones Parcial de Pearson, Spearman y Kendall (Una correlación parcial proporciona una medida de la correlación entre dos variables después de controlar los efectos de otras variables).
- entre otros.

# PROCEDIMIENTO CORR

La sintaxis corresponde a,

**PROC CORR** <options>;

**BY** variables;

**FREQ** variable;

**ID** variables;

**PARTIAL** variables;

**VAR** variables;

**WEIGHT** variable;

**WITH** variables;

# PROCEDIMIENTO CORR

- **PROC CORR** Invoca el procedimiento.
- **BY** Especifica grupos en los que se realizan los análisis de correlación separadas.
- **FREQ** Especifica la variable que representa la frecuencia de ocurrencia de los demás valores en la observación.
- **ID** Especifica una o más variables adicionales de punta para identificar observaciones en gráficos de dispersión y las matrices de dispersión de la trama.
- **PARTIAL** La declaración PARTIAL identifica el control de las variables para calcular Pearson, Spearman, o Kendall coeficientes de correlación parcial.

# PROCEDIMIENTO CORR

- **VAR** Enumera las variables numéricas para ser analizados y su orden en la matriz de correlación. Si se omite la declaración VAR, se utilizan todas las variables numéricas no mencionados en otras declaraciones.
- **WEIGHT** Identifica la variable cuyos valores de peso cada observación para calcular Pearson producto-momento de correlación.
- **WITH** Enumera las variables numéricas con las que las correlaciones deben ser computados.

# Detalles: Procedimiento CORR

http://support.sas.com/documentation/cdl/en/procstat/67528/HTML/default/viewer.htm#procstat_corr_details.htm

- Pearson Product-Moment Correlation
- Kendall's Tau-b Correlation Coefficient
- Partial Correlation
- Polychoric Correlation
- Cronbach's Coefficient Alpha
- Missing Values
- Output Tables
- ODS Table Names
- Spearman Rank-Order Correlation
- Hoeffding Dependence Coefficient
- Fisher's z Transformation
- Polyserial Correlation
- Confidence and Prediction Ellipses
- In-Database Computation
- Output Data Sets
- ODS Graphics

# PROCEDIMIENTO CORR

En este ejemplo se calcula pruebas de chi-cuadrado y la prueba exacta de Fisher para comparar la probabilidad de enfermedad coronaria para dos tipos de dieta. También estima los riesgos relativos y calcula los límites de confianza exactos para la odds ratio.

Los datos FatComp contiene datos hipotéticos para un estudio de casos y controles de la dieta rica en grasas y el riesgo de enfermedad coronaria. Los datos se registran como recuentos de células, donde el conde variable contiene las frecuencias para cada exposición y la combinación de la respuesta.

# PROCEDIMIENTO CORR

```
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
;
```

# PROCEDIMIENTO CORR

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

proc freq data=FatComp order=data;
format Exposure ExpFmt. Response RspFmt.;
tables Exposure*Response / chisq relrisk;
exact pchi or;
weight Count;
title 'Case-Control Study of High Fat/Cholesterol Diet';
run;
```

# PROCEDIMIENTO CORR

Estas mediciones se realizaron en hombres que participan en un curso de aptitud física en la Universidad de Carolina del Norte Estado. Las variables son la edad (años), peso (kg), Tiempo de ejecución (tiempo de correr 1,5 millas en minutos), y Oxígeno (consumo de oxígeno, ml por kg de peso corporal por minuto).

# PROCEDIMIENTO CORR

```
data Fitness;
input Age Weight Oxygen RunTime @@;
datalines;
44 89.47 44.609 11.37    40 75.07 45.313 10.07
44 85.84 54.297  8.65    42 68.15 59.571  8.17
38 89.02 49.874   .      47 77.45 44.811 11.63
40 75.98 45.681 11.95    43 81.19 49.091 10.85
44 81.42 39.442 13.08    38 81.87 60.055  8.63
44 73.03 50.541 10.13    45 87.66 37.388 14.03
45 66.45 44.754 11.12    47 79.15 47.273 10.60
54 83.12 51.855 10.33    49 81.42 49.156  8.95
51 69.63 40.836 10.95    51 77.91 46.672 10.00
48 91.63 46.774 10.25    49 73.37   .    10.08
57 73.37 39.407 12.63    54 79.38 46.080 11.17
52 76.32 45.441  9.63    50 70.87 54.625  8.92
51 67.25 45.118 11.08    54 91.63 39.203 12.88
51 73.71 45.790 10.47    57 59.08 50.545  9.93
49 76.32   .      .      48 61.24 47.920 11.50
52 82.78 47.467 10.50
;
```

# PROCEDIMIENTO CORR

```
proc corr data=Fitness pearson spearman kendall hoeffding
plots=matrix(histogram);
var Weight Oxygen RunTime;
run;
```

# Ejercicio:

Realicemos un ejercicio completo, usando todos los procedimientos de estadísticos básicos aprendidos en la ultimas clases.

```
data Sim (drop=i);
do i=1 to 400;
X = rannor(135791);
Batch = 1 + (i>150) + (i>300);
if Batch = 1 then Y = 0.3*X + 0.9*rannor(246791);
if Batch = 2 then Y = 0.25*X + sqrt(.8375)*rannor(246791);
if Batch = 3 then Y = 0.3*X + 0.9*rannor(246791);
output;
end;
run;
```

- Con la data anterior realice lo siguiente

1. Importe las Datas a SAS a una librería creada por usted.
2. Calcule el promedio de las variables X e Y por Batch.
3. Crea una DATA con dos nuevas variables X2 e Y2 tal que X2 será la diferencia entre X y el promedio de X por batch. e Y2 será la diferencia entre Y y el promedio de Y por batch.
4. Realice un proc Summary de las variables creadas, identificando los grupos con mayor diferencias.
5. Realice una tabla de promedios de las variables X e Y por Batch usando PROC TABULATE.
6. Calcule la covarianza entre X e Y por BATCH.
7. Realice una tabla de frecuencia con las variables X e Y pero categorizadas en 4 categorías cada una de ellas. Tal que los puntos de cortes sean los percentiles 0.25,0.50 y 0.75.