

**SOFTWARE ESADÍSTICO I**

**FACULTAD DE ESTUDIOS ESTADÍSTICOS**

**PRÁCTICA 2- “CONTAMINA”**

**Nº24**

Pablo Galarón Mateo

**ÍNDICE**

1. APARTADO A……………3
2. APARTADO B……………6
3. APARTADO C……………11
4. APARTADO D……………14
5. APARTADO E……………18
6. APARTADO F……………24
7. APARTADO G…………...29
8. APARTADO H……………33

**Apartado A.**

Crear un formato horario: en el que se distinga la madrugada (0-7 horas), la mañana (8-12), el medio día (13-16), la tarde (17-20) y la noche (21-24). Crear otro formato para los meses (1-3 invierno, 4-6 primavera, 7-9 verano, 10-12 otoño). Ambos sólo se utilizarán en algunos apartados, por tanto, no se debe incluir en ningún paso data, sólo en los procedimientos que así lo requieran.

**Programación.**

|  |
| --- |
| /\* PRÁCTICA-2 Nº 24 PABLO GALARÓN MATEO \*/  /\* APARTADO A \*/  **proc** **format** library = sasuser. formatos\_practica\_24\_;  value horario **0** - **7** = 'Madrugada'  **8** - **12** = 'Mañana'  **13** - **16** = 'Mediodía'  **17** - **20** = 'Tarde'  **21** - **24** = 'Noche';  value estaciones **1** - **3** = 'Invierno'  **4** - **6** = 'Primavera'  **7** - **9** = 'Verano'  **10** - **12** = 'Otoño';  **run**;  /\* para recordar que lo hemos introducido en sas user por nuestra propia cuenta, no haría falta hacer este paso data pero lo hago para que observemos que es lo que necesitamos, cual es nuestro conjunto de datos. \*/  **data** sasuser.contamina;  set sasuser.contamina;  **run**;  **proc** **print** data = sasuser.contamina (obs= **10**); **run**; /\* imprimo 10 observaciones para apreciar un poco el conjunto de datos \*/ |

**Resultados.**

Formato estaciones:

Tabla

Descripción generada automáticamente

Formato horario:

Interfaz de usuario gráfica, Texto, Tabla

Descripción generada automáticamente

Conjunto de datos:

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

**Comentarios.**

En primer lugar, el archivo proporcionado para realizar la práctica es un conjunto de datos, no un archivo sas y para poder trabajar con él o se le llama mediante la sentencia libname o se introduce manualmente en la carpeta que queramos.

En mi caso, he preferido descargar el archivo, dirigirme a “My sas files”, después a “9.4” y pegar el conjunto de datos. Automáticamente, se acaba de introducir a la librería permanente sasuser.

En el primer apartado, hay que crear dos formatos, los cuales también los he guardado en la librería permanente sasuser. Nótese que los formatos están creados, pero no los he introducido en el conjunto de datos, ya que solo harán falta cuando lo pida el ejercicio, por lo que dentro de cada apartado los introduciré según convenga.

Más adelante, se imprimen solo las 10 primeras observaciones del conjunto de datos para observar que está guardado correctamente y ver que los datos guardan una cierta coherencia.

**Apartado B.**

Crear una gráfica con la evolución media del Ozono a través de los años (eje x) para cada estación (primavera, verano,) y cada bloque horario. (son 4\*5=20 figuras). En cada gráfica representar la función en la que se unan los valores de unos años con otros. así mismo debería también aparecer una recta de regresión asociada a cada grafica (se piden por tanto 2 graficas en cada una de las 20 figuras) Comentarlas.

**Programación.**

|  |
| --- |
| /\* APARTADO B \*/  **proc** **means** data = sasuser.contamina nway noprint;  var O3;  class ANNO hora mes;  output out = ej\_b mean = medias\_ozono;  format hora horario.  mes estaciones.; /\* introduzco los formatos ahora que si que me hacen falta \*/  **run**;  **Proc** **sort** data = ej\_b;  by hora mes;  **run**;  title h = **2** color = black "Evolución media del ozono";  footnote h = **1.7** color = black "Para cada estación y horario";  symbol1 v = dot i = j color = blue; /\* para el plot1 \*/  symbol2 w = **2** i = r color = red; /\* para el plot2 \*/ /\*w me mide el grosor de la linea \*/  **proc** **gplot** data = ej\_b;  by hora mes;  plot1 medias\_ozono \* anno /LEGEND; /\* /legend es para que aparezca abajo la leyenda \*/  plot2 medias\_ozono \* anno /LEGEND;  label anno='Año' medias\_ozono='Media de Ozono';/\* con estos nombres queda mejor \*/  **run**; |

**Resultados.**

Proc means para cada año, grupo horario y estación:

Tabla

Descripción generada automáticamente

Gráficos:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

**Comentarios.**

En primer lugar, debemos calcular las medias para cada año, estación y grupo horario. Para ello, realizamos un proc means con la sentencia class y le añadimos nway(importante), ya que solo queremos que nos aparezcan las observaciones con el \_type\_ de todas con todas, Creamos el conjunto de datos para luego realizar las gráficas y ahora introducimos los formatos para el grupo horario y las estaciones.

Una vez tenemos las medias, realizamos un proc sort para ordenar los datos por grupo horario y estación para luego realizar la sencencia by en el proc gplot.

Más adelante, realizamos el proc gplot llamando al conjunto de datos de las medias, incluimos la sentencia by y realizamos varios plots y symbols. Realizamos varios ya que queremos que es una misma gráfica aparezca la evolución del ozono y una recta de regresión para cada grupo horario y estación. En el primer symbol, decimos que se unan todos los puntos y el el segundo ponemos i = r, lo que nos hará que nos parezca una recta de regresión.

Por último, realizamos los últimos retoques en el symbol como el color, la longitud y la anchura de las líneas y ponemos un título y un subtitulo para que las gráficas queden más claras y ordenadas.

Al haber obtenido 20 figuras, adjuntaré cuatro de ellas para observar un poco los resultados.

-Invierno madrugada: muestra una tendencia general de aumento en el ozono con un pico notable en 2012 antes de una caída.

-Invierno Mañana: también muestra un aumento continuo, sugiriendo un incremento más constante en la concentración del ozono durante estas horas.

-Verano madrugada: presenta fluctuaciones con una tendencia a la baja hacia el final del período, lo cual podría estar relacionado con cambios en las condiciones atmosféricas estacionales.

-Otoño madrugada: se aprecia cierta tendencia cambiante considerable con picos y caídas, pero con una tendencia creciente en la regresión, lo que indica un aumento general a pesar de las variaciones anuales.

**Apartado C.**

Determinar si existe diferencia en los valores medios del ozono por grupos horarios. ¿y por tipo\_area? (podéis utilizar los procedimientos npar1way y/o ttest).

**Programación.**

|  |
| --- |
| /\* APARTADO C \*/  /\* Lo haré con npar1way porque el Ttest no puede procesar en la sentencia class más de dos niveles en esa variable \*/  **proc** **npar1way** data = sasuser.contamina wilcoxon;  var O3;  class hora;  format hora horario.;  **run**;  **proc** **npar1way** data = sasuser.contamina wilcoxon;  var O3;  class TIPO\_AREA;  **run**; |

**Resultados.**

Por grupos horarios

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Por tipo de área

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

**Comentarios.**

Para la realización de este apartado, se debe hacer con el procedimiento npar1way ya que el ttest la sentencia class solo puede tomar dos valores y en este caso toma más de tres.

A la hora de hacer el npar1way, es muy importante incluir la opción wilcoxon (test de diferencia de medias) ya que vamos a hacer “una prueba de rangos”, la cual es una prueba no paramétrica y nos sirve para evaluar en este caso si existen diferencias entre las medias por grupo horario y por tipo de área.

Realizamos ambos procedimientos cambiando únicamente la sentencia class y formateando los grupos horarios y observamos los resultados.

Para los grupos horarios, observamos que hay diferencias significativas ya que el p-valor nos sale significativamente bajo, <0.0001. La probabilidad de que todos los grupos horarios tengan la misma media de ozono es muy baja.

A su vez, observamos que los valores más altos se producen en el mediodía y a medida que llega la noche, disminuyen de manera considerable, lo cual se puede deber a los factores ambientales y atmosféricos del ozono en la tierra.

En cuanto al tipo de área, observamos que también hay diferencias significativas en las medias del ozono para cada tipo de área. El p-valor nos sale muy bajo, también <0.0001. En este caso, el valor más alto lo alcanzan las zonas rurales, luego las suburbanas y por último las urbanas. Estos resultados, guardan cierta coherencia ya que las zonas rurales suelen estar más elevadas y más expuestas a la luz solar, lo que puede dar lugar a una mayor concentración de ozono.

**Apartado D.**

Intentar estimar una función de distribución que sirva para explicar cómo se producen los valores del ozono según el bloque horario y la estación del año en las zonas rurales. (utilizar el procedimiento Univariate o NPAR1WAY).

**Programación.**

|  |
| --- |
| /\* APARTADO D \*/  **data** rural;  set sasuser.contamina;  if TIPO\_AREA = 'RURAL';  format hora horario.  mes estaciones.;  **run**;  **proc** **sort** data = rural;  by hora mes;  **run**;  **proc** **univariate** data = rural;  var O3;  by hora mes;  histogram / normal lognormal exponential beta  gamma weibull;  **run**; |

**Resultados.**

Conjunto de datos con solo RURAL

Interfaz de usuario gráfica, Tabla, Excel

Descripción generada automáticamente

Resultados del proc Univariate

-HORA= Noches, MES = Otoño:

Gráfico, Histograma

Descripción generada automáticamente

NORMAL:

Tabla

Descripción generada automáticamente

GAMMA:

Tabla

Descripción generada automáticamente

**Comentarios.**

Para poder comprobar si un conjunto de datos se parece a cierta distribución he decidido utilizar el procedimiento univariate. En este apartado me piden solo para el tipo de área rural, entonces debo filtrar el conjunto de datos original y que solo contenga las observaciones cuyo tipo de área sea rural.

Después realizo el proc univariate y pongo muchas distribuciones continuas para ver si se aproxima a alguna distribución.

Al poner tantas distribuciones, he querido mostrar solo si se parece a una normal o a una GAMMA, para la noche en el mes de otoño.

Como se puede observar, el p-valor es muy bajo en ambos casos, por lo tanto, no se puede estimar ni a una normal ni a una gamma. No obstante, si observamos la gráfica parece que se parece a la normal, no está tan alejado. Pero al tener un conjunto de datos tan grande, con tantas observaciones el p-valor nos sale muy bajo.

**Apartado E.**

Hacer lo mismo que en d para las zonas suburbana y urbana. Comentar los resultados.

**Programación.**

|  |
| --- |
| /\* APARTADO E \*/  /\* Para tipo de área urbana \*/  **data** urbana;  set sasuser.contamina;  if TIPO\_AREA = 'URBANA';  format hora horario.  mes estaciones.;  **run**;  **proc** **sort** data = urbana;  by hora mes;  **run**;  **proc** **univariate** data = urbana;  var O3;  by hora mes;  histogram / normal lognormal exponential beta  gamma weibull;  **run**;  /\*Para tipo de área suburbana \*/  **data** suburbana;  set sasuser.contamina;  if TIPO\_AREA = 'SUBURBANA';  format hora horario.  mes estaciones.;  **run**;  **proc** **sort** data = suburbana;  by hora mes;  **run**;  **proc** **univariate** data = suburbana;  var O3;  by hora mes;  histogram / normal lognormal exponential beta  gamma weibull;  **run**; |

**Resultados.**

Conjunto de datos con solo URBANA

Tabla

Descripción generada automáticamente

Resultados del proc Univariate

-HORA= Tarde, MES = Verano:

Gráfico, Histograma

Descripción generada automáticamente

NORMAL:

Tabla

Descripción generada automáticamente

WEIBULL:

Tabla

Descripción generada automáticamente

Conjunto de datos con solo SUBURBANA

Tabla

Descripción generada automáticamente

Resultados del proc Univariate

-HORA= Mediodía, MES = Otoño:

Gráfico, Histograma

Descripción generada automáticamente

NORMAL:

Tabla

Descripción generada automáticamente

LOGNORMAL:

Tabla

Descripción generada automáticamente

**Comentarios.**

En este apartado se procede igual que en el anterior, filtramos para urbana y suburbana y hacemos el univariate. También he decidido mostrar 2 para cada tipo de área ya que no puedo mostrar todas.

En este apartado, todos los p-valores salen también muy bajos, pero se ve más claramente como algunas gráficas si parecen muy parecidas a una normal. Como comenté antes se debe a el tamaño del conjunto de datos. Si quisiéramos ser más exactos en los calculo y discrepar el tamaño del conjunto de datos, podemos hacer un pplot y ver si los punto se acercan a la recta.

**Apartado F.**

Identificar el grupo horario con mayor media de ozono, independientemente de la temporada. Crear un formato que separe en grupos, las provincias cantábricas, las mediterráneas y las de interior. Identificar si hay diferencias en los valores del ozono obtenidos en cada uno de esos grupos para cada estación, utilizando sólo la información relativa al grupo horario con mayor índice de ozono.

**Programación.**

|  |
| --- |
| /\* APARTADO F \*/  /\* PRUEBA para luego saber si lo tengo bien \*/  **proc** **means** data = sasuser.contamina noprint nway;  var O3;  class hora;  output out = prueba  mean = ;  format hora horario.;  **run**;  /\* Observamos que el grupo horario es MEDIODÍA  , nos servirá para comprobar nuestro resultado  más adelante \*/  **proc** **means** data = sasuser.contamina noprint nway;  var O3;  class hora;  output out = medias  mean = medias\_O3;  format hora horario.;  **run**;  **proc** **means** data = medias noprint;  var medias\_O3;  output out = max\_medias (drop = \_type\_ \_freq\_)  max =  idgroup (max(medias\_O3) out[**1**] (hora) = max\_hora);  **run**;  /\* Creamos el formato con las regiones \*/  **proc** **freq** data = sasuser.contamina;  tables comunidad;  **run**;  **data** sasuser.contamina;  set sasuser.contamina;    if comunidad in ('ASTURIAS(PRINCIPADO DE)', 'CANTABRIA', 'PAÍS VASCO', 'GALICIA', 'NAVARRA(COMUNIDAD FORAL)') then region = 'Cantábrica';  else if comunidad in ('CATALUÑA', 'COMUNIDAD VALENCIANA', 'MURCIA (REGIÓN DE)', 'BALEARES (ISLAS)') then region = 'Mediterránea';  else if comunidad in ('MADRID', 'CASTILLA Y LEÓN', 'CASTILLA-LA MANCHA', 'ARAGÓN', 'EXTREMADURA', 'RIOJA (LA)') then region = 'Interior';  else region = 'Otra';  **run**;  /\* Filtrar datos para el grupo horario con la mayor media de ozono---MEDIODÍA(hemos comprobado en proc means)\*/  **data** filtrados;  set sasuser.contamina;  if hora in (**13**,**14**,**15**,**16**); /\*LAS HORAS PERTENECIENTES AL MEDIODÍA \*/  format hora horario.;  **run**;  /\* PASO FINAL: hacemos el proc npar1way y observamos \*/  **proc** **sort** data = filtrados;  by mes;  format mes estaciones.;  **run**;  **proc** **npar1way** data = filtrados wilcoxon;  var O3;  class region;  by mes;  format mes estaciones.;  **run**; |

**Resultados.**

Prueba para luego comprobar:

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Grupo horario con mayor media de Ozono con IDGROUP en proc means:

Calendario

Descripción generada automáticamente con confianza media

Creación de nueva variable región en el conjunto de datos:

Tabla

Descripción generada automáticamente

Conjunto de datos filtrado con solo el MEDIODÍA:

Tabla

Descripción generada automáticamente

Paso Final, Proc Npar1way:

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

**Comentarios.**

Para que este apartado se realice correctamente hay que ir por orden y poco a poco. Primero calculamos las medias y vemos que grupo horario es el que más nivel de ozono alberga. Nos sale mediodía. No obstante, para hacerlo mas preciso y claro hay que hacer máximo de las medias por lo que utilizaré el idgroup.

Más delante hago un proc freq para saber que tipo de comunidades alberga el conjunto y con que nombre para poder realizar un paso data, sobrescribiendo contamina y creando la nueva variable región.

Para hacer el proc npar1way hay que filtrar los datos para que tenga solo las observaciones pertenecientes al grupo horario mediodía (grupo con mayor media de ozono). Filtro lo datos, importante poner como si fuese variable numérica porque está guardada como numérica por mucho que le haya puesto un formato.

Una vez tengo lo datos filtrados realizo el npar1way y observo los resultados.

Observamos que hay diferencias significativas en los distintos tipos de región al tener un p-valor significativamente bajo, <0.0001. Se puede apreciar como el valor más alto lo toma la región “otra”, luego la mediterránea, y la que menos alberga es la región cantábrica. Quizá se deba a que el sol (más presente en mediterránea y otra(Andalucía, Canarias)) favorezca la aparición y concentración del ozono.

**Apartado G.**

Obtener las medias del ozono por grupo horario, grupo regional (cantábricas, mediterráneas,), mes y año. Crear una variable que contenga la información de la fecha sobre la que calculamos esas medias suponiendo que los datos de cada mes se dieran en el día 15. Representar gráficamente para cada grupo horario (una figura diferente) la evolución en el tiempo (eje x la fecha) de las medias de ozono con las gráficas diferenciadas según grupo regional (tantas curvas como grupos regionales haya). Comentarlas.

**Programación.**

|  |
| --- |
| /\* APARTADO G \*/  **proc** **means** data = sasuser.contamina noprint nway;  var O3;  class hora region mes ANNO;  output out = medias\_ozono mean = medias\_O3;  format hora horario.  mes estaciones.;  **run**;  **data** fecha\_medias;  set medias\_ozono;  fecha = mdy(mes,**15**,ANNO);  format fecha yymmdd10.; /\* Aunque se pueden utilizar otros formatos para las fechas \*/  **run**;  title h = **2** color = black "Evolución media del ozono por grupo horario";  footnote h = **1.7** color = black "Para cada tipo de región en España";  symbol1 v = dot i = j color = red l = **1** w = **2**;  symbol2 v = dot i = j color = green l = **2** w = **2**;  symbol3 v = dot i = j color = blue l = **3** w = **2**; /\* DISTINTOS SYMBOLS PARA CADA REGIÓN \*/  symbol4 v = dot i = j color = yellow l = **4** w = **2**; /\* Pongo distitos tipos de líneas para que se observe mejor\*/  **proc** **gplot** data = fecha\_medias;  by hora;  plot medias\_O3 \* fecha = region / overlay legend ; /\* Para que me saque todas las regiones en un mismo gráfico \*/  label fecha='Fecha';  label medias\_O3='Media de Ozono';  **run**; |

**Resultados.**

Proc means por grupo horario, región, estación y año con la variable FECHA**:**

**Tabla

Descripción generada automáticamente**

Gráficos:

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

**Comentarios.**

Para este apartado primero debemos hacer el proc means creando un conjunto de datos con la media por grupo horario, región, año y estación. Después hacemos un paso data para crear las fechas, hacemos set para llamar al conjunto de datos de las medias. Para la fecha, utilizo la sentencia mdy y pongo las variables año mes y el día 15 que me pide el enunciado. Después, le pongo un formato para que me coja la fecha.

Por último, realizo el proc gplot llamando al conjunto de datos de las fechas y realizo los gráficos. Salen muy juntos y es por ello que he introducido muchos symbols para que se vea más claro. He intentado ordenar el conjunto de datos para la variable x, pero me sale igual. Además, abajo en la variable x se puede observar como está ordenado por la fecha.

Las conclusiones de los gráficos son las siguientes:

-Madrugada:

Cantábrica: muestra fluctuaciones significativas en el nivel de ozono, con algunos picos y caídas a lo largo del tiempo

Interior: Parece tener la menor consistencia en los niveles de ozono, tendencia menos clara

Mediterráneo: Parece tener mejor tendencia, pero también tiene fluctuaciones.

Otra: La que mayor valor toma, tendencia más clara y picos menos extremos.

-Mañana:

Las curvas varían bastante, pero en comparación con la mañana los niveles de ozono son más altos. Las tendencias por región son similares a las de la madrugada, pero en este caso el valor más alto(pico) del ozono lo toma la región mediterránea.

**Apartado H.**

Crear un conjunto de datos que contenga los máximos y los mínimos para cada grupo horario, grupo regional y tipo de zona para cada mes y año como en el apartado G. Incluir también la misma variable fecha, y representar esos máximos y mínimos comprobando su evolución en el tiempo de la misma forma a como se hizo con las medias en el apartado g.

**Programación.**

|  |
| --- |
| /\* APARTADO H \*/  **proc** **means** data = sasuser.contamina noprint nway;  var O3;  class hora region mes ANNO;  output out = max\_min\_ozono max = max\_O3 min = min\_O3;  format hora horario.  mes estaciones.;  **run**;  **data** fecha\_max\_min;  set max\_min\_ozono;  fecha = mdy(mes,**15**,ANNO);  format fecha yymmdd10.; /\* Aunque se pueden utilizar otros formatos para las fechas \*/  **run**;  /\* REPRESENTACIÓN PARA LOS MÁXIMOS \*/  title h = **2** color = black "Evolución de los máximos niveles del ozono por grupo horario";  footnote h = **1.7** color = black "Para cada tipo de región en España";  symbol1 v = dot i = j color = red l = **1** w = **2**;  symbol2 v = dot i = j color = green l = **2** w = **2**;  symbol3 v = dot i = j color = blue l = **3** w = **2**; /\* DISTINTOS SYMBOLS PARA CADA REGIÓN \*/  symbol4 v = dot i = j color = yellow l = **4** w = **2**;  **proc** **gplot** data = fecha\_max\_min;  by hora;  plot max\_O3 \* fecha = region / overlay legend ; /\* Para que me saque todas las regiones en un mismo gráfico \*/  label fecha='Fecha';  label medias\_O3='Max del Ozono';  **run**;  /\* REPRESENTACIÓN PARA LOS MÍNIMOS \*/  title h = **2** color = black "Evolución de los mínimos niveles del ozono por grupo horario";  footnote h = **1.7** color = black "Para cada tipo de región en España";  symbol1 v = dot i = j color = red l = **1** w = **2**;  symbol2 v = dot i = j color = green l = **2** w = **2**;  symbol3 v = dot i = j color = blue l = **3** w = **2**; /\* DISTINTOS SYMBOLS PARA CADA REGIÓN \*/  symbol4 v = dot i = j color = yellow l = **4** w = **2**;  **proc** **gplot** data = fecha\_max\_min;  by hora;  plot min\_O3 \* fecha = region / overlay legend ; /\* Para que me saque todas las regiones en un mismo gráfico \*/  label fecha='Fecha';  label medias\_O3='Min del Ozono';  **run**; |

**Resultados.**

Cómo es el mismo proceso mostraré solo alguno de los gráficos:

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

**Comentarios.**

La manera de proceder es exactamente igual que el apartado anterior pero simplemente cambiando las medias por máximos y mínimos y realizando dos gplots.

Las conclusiones obtenidas de los gráficos son las siguientes:

-Mínimos mediodía:

Cantábrica: muestra fluctuaciones considerables, aumentando los niveles de O3 con el tiempo.

Interior: Tendencia estable y con valores medio-bajos de ozono.

Mediterráneo: Tiene un incremento gradual en los mínimos, especialmente después de 2020(pandemia).

Otra: Variabilidad muy significativa con bastantes picos y caídas.

-Máximos mañana:

Cantábrica: cierta variabilidad, pero con tendencia ascendente.

Interior: muchas fluctuaciones y no tenemos un patrón claro.

Mediterráneo: Picos notables hacia el final del gráfico que quizá se deban a cambios ambientales o urbanización.

Otra: Aunque con ciertas fluctuaciones, los niveles parecen bastante estables.

Por último, podemos observar como la gráfica de los mínimo y máximo más o menos se complementan en la mayoría de los casos, lo que nos da la certeza de que nuestros cálculos se han realizado correctamente.