Universidad de Valladolid

Estadística Descriptiva

PAC 1

García Prado, Sergio
Taboada Rodero, Ismael José

9 de diciembre de 2016

Índice

1. Introducción

1.1. Enunciado

"Con el fin de analizar el consumo energético de una empresa productores de acero se inspeccionaron durante cinco días cada una de las tres líneas de producción. En cada una de ellas se anotaron las variables más relevantes para distintas horas del turno, salvo en la última hora donde sólo se inspeccionó durante cuatro días."

1.2. Variables

- **Consumo**: Consumo energético de la empresa (Megavatios - hora).
- **pr.tbc**: Producción del tren de bandas calientes (Tm de acero).
- **pr.cc**: Producción de colada continua (Tm de acero).
- **pr.ca**: Producción del convertidor de acero (Tm de acero).
- **pr.galv1**: Producción de galvanizado de tipo I (Tm de acero).
- **pr.galv2**: Producción de galvanizado de tipo II (Tm de acero).
- **m pr.pint** Producción de chapa pintada (Tm. de acero).
- **Línea**: Línea de producción empleada (A, B o C).
- **Hora**: Hora en la que se recogieron los datos
- **Temperatura**: Temperatura del sistema: alta (Alta), media (Media) y baja (Baja).
- **Averias**: Presencia de averías (Sí, No).
- **Naverias**: número de averías detectadas.
- **Sistema**: Activación de un sistema de detección de sobrecalientamiento (encendido (ON),apagado (OFF)).

1.2.1. Variables artificiales

A continuación se muestran las variables artificiales utilizadas para el desarrollo de la entrega:

- **pr.total**: Suma de las variables relacionadas con la producción, con el fin de información sobre la producción general de la empresa.

1.3. Notación

La notación que se ha decidido seguir en este documento es la siguiente:

- X: Variable a estudiar, siendo X_i cada una de las muestras con $i \in (0, n(X))$.
- $n(X)$: Cardinalidad de la variable X, es decir, el número de muestras que contiene.
- $\overline{X} = \sum \frac{X_i}{n(X)}$: Media muestral de la variable X.
- $me(X)$: Mediana muestral de la variable X.
- $S_x^2 = \sum \frac{X_i^2}{n(X)} - \overline{X}^2$: Varianza muestral de la variable X.
- $\sigma_x = S_x = \sqrt{S_x^2}$: Desviación típica muestral de la variable X.

- $S_{xy} = \sum \frac{X_i Y_i}{n(X,Y)} - \overline{XY}$: Covarianza muestral de la variable X e Y.
- $r_{xy} = \frac{S_{xy}}{S_x S_y}$: Coeficiente de correlación muestral entre las variable X e Y.
- $max(X)$: Valor máximo muestral de la variable X.
- $min(X)$: Valor mínimo muestral de la variable X.
- $range(X)$: Rango muestral de la variable X.
- $percentil(X, P)$: Percentil $P * 100$ de la distribución X.

2. Ejercicios

2.1. Ejercicio 1

Consideremos la variable *consumo*.

- Calcular media, mediana, desviación típica y rango de esta variable. Realizar un polígono de frecuencias relativas a partir de un histograma con 18 clases.
- Realizar un diagrama de cajas múltiple para la variable *consumo* en los grupos marcados por la variable *hora*. En caso de existir datos atípicos, identificarlos indicado la línea de producción y el estado del sistema de detección de sobrecalentamiento. Escribir una frase indicando lo que se observa en este gráfico.
- ¿Se obtiene los mismos resultados si hacemos todos los análisis anteriores para cada línea por separado?. Realiza los análisis para justificar tu respuesta.

El número de muestras que contiene la variable consumo es: $n(consumo) = 117$

El valor de la media muestral de la variable consumo es:

$$\overline{consumo} = \frac{\sum consumo_i}{n(consumo)} = 139{,}456$$

La mediana muestral de la variable consumo es:

$$me(consumo) = 140{,}07$$

Cabe destacar la cercanía que existe entre el valor de la *media* y la *mediana*, lo cuál nos da indicios de que la distribución tiene una forma simétrica, es decir, no contiene colas hacia ninguna de las direcciones. Esto se puede confirmar al visualizar la figura 1

En cuanto a la desviación típica podemos obtener su valor a partir de la siguiente fórmula, la cuál representa la variación global de la distribución:

$$\sigma_{consumo} = S_x = \sqrt{S_x^2} = \sqrt{\sum \frac{X_i^2}{n(X)} - \overline{X}^2} = 55{,}18353$$

Para calcular el rango primero necesitamos obtener los valores tanto máximo como mínimo dentro de la muestra, para calcular la diferencia en valor absoluto que hay entre ellos:

$$range(consumo) = \left|max(consumo) - min(consumo)\right| = |290{,}72 - 17{,}5| = 273{,}22$$

El polígono de frecuencias relativas generado a partir del particionamiento de la distribución en 18 clases se puede visualizar en la figura 1, la cual contiene tanto el Histograma de frecuencias relativas como su correspondiente polígono de frecuencias.

El diagrama de cajas múltiple obtenido a partir del particionamiento de la variable *consumo* en las categorías indicadas por la variable *hora* se muestra en la figura 2. Este permite analizar la distribución de los valores que toma dicha variable repartidos por horas.

Existe un outlier en la muestra $consumo_{29}$. Tal y como se puede apreciar en el gráfico de la figura 2. Este se encuentra dentro de la categoría de la 5^a *hora* y es un punto atípico inferior.

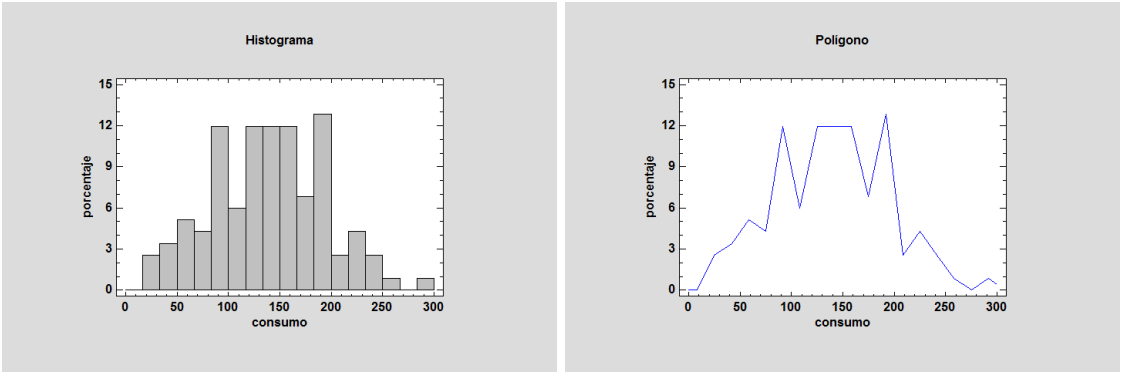

Figura 1: Histograma y Polígono de Frecuencias de *consumo*

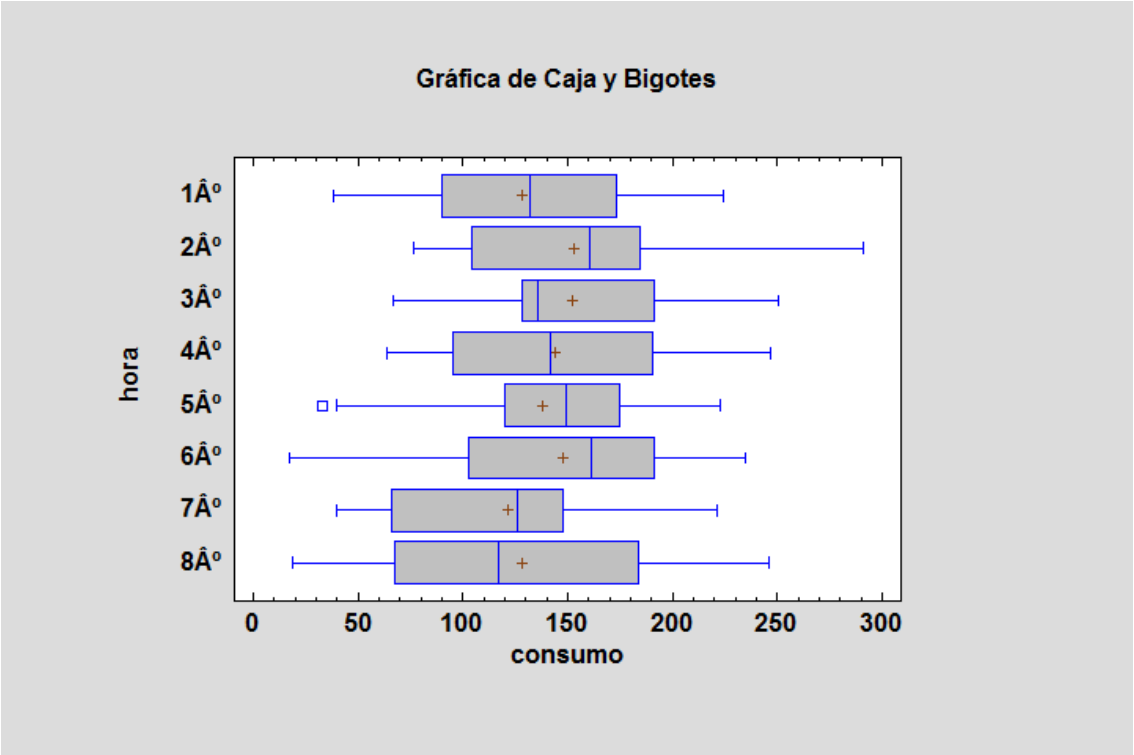

Figura 2: Diagrama de Cajas de *consumo* particionada sobre *hora*

Una interpretación posible para este suceso podría ser la siguiente: La jornada laboral está particionada en dos intervalos de 4 horas, por lo que la *5ª hora* se corresponde con el inicio del segundo intervalo. Por tanto, una posible causa de ello sería (debido a que estamos analizando la variable *consumo*) que durante esa muestra se retrasara el inicio por lo que el consumo producido sería inferior al tener la maquinaria en reposo.

Dividiendo por líneas de producción el consumo, los indicadores estadísticos resultantes se recogen en la tabla 1.

Parámetro	Linea A	Línea B	Línea C
Tamaño	39	39	39
Media	109,841	137,548	170,98
Mediana	99,37	143,82	173,1
Desviación Típica	39,6747	59,6624	47,31871
Máximo	226,38	245,74	290,72
Mínimo	33,18	17,5	68,3
Rango	193,2	228,24	222,42

Cuadro 1: Estadísticos de la variable *consumo* particionada sobre *linea*

La relación entre estos estadísticos y los de la variable sin particionamiento son las siguientes: En el caso de la *media*, el valor se reparte proporcionalmente por lo que a partir de cada línea se puede obtener el valor original. Por contra, la *mediana* no se puede recalcular a partir de los indicadores de cada línea. Con la desviación típica ocurre el mismo caso, depende de cómo esté distribuida internamente cada subvariable y no se puede obtener el global sin conocer cada una de ellas internamente. En el caso de los valores *máximo*, *mínimo* y *rango*, aplicando las operaciones pertinentes para obtener los extremos según cada caso, es posible recuperar el rango total.

2.2. Ejercicio 2

Se desean estudiar las variables *sistema* de detección de sobrecalentamiento y *temperatura*

- Construir una tabla cruzada donde aparecezcan únicamente las frecuencias absolutas y las frecuencias relativas de la variable *sistema* condicionadas por temperatura.
- Realizar un gráfico apropiado. Calcular en estadístico adecuado e interpretar el resultado.

	Alta	Media	Baja
OFF	24 52,17 %	16 48,48 %	19 50,00 %
ON	22 47,83 %	17 51,52 %	19 50,00 %

Cuadro 2: Tabla de Frecuencias para temperatura por sistema

La tabla 2 representa la relación existente entre las variables de *sistema* y *temperatura*. Cada casilla contiene dos valores numéricos, el primero representa la frecuencia absoluta de la relación entre el valor de sistema y el valor de temperatura; la segunda indica la frecuencia relativa del valor de *sistema* condicionada por el valor de *temperatura*.

Debido a que el estadístico *Chi-Cuadrada* utilizado para el análisis de dependencia entre filas y columnas de la tabla 2 tiene un valor de $0,9471$ no podemos rechazar la hipótesis de que las variables *sistema* y *temperatura* sean independientes con un nivel de confianza del $95,0\,\%$.

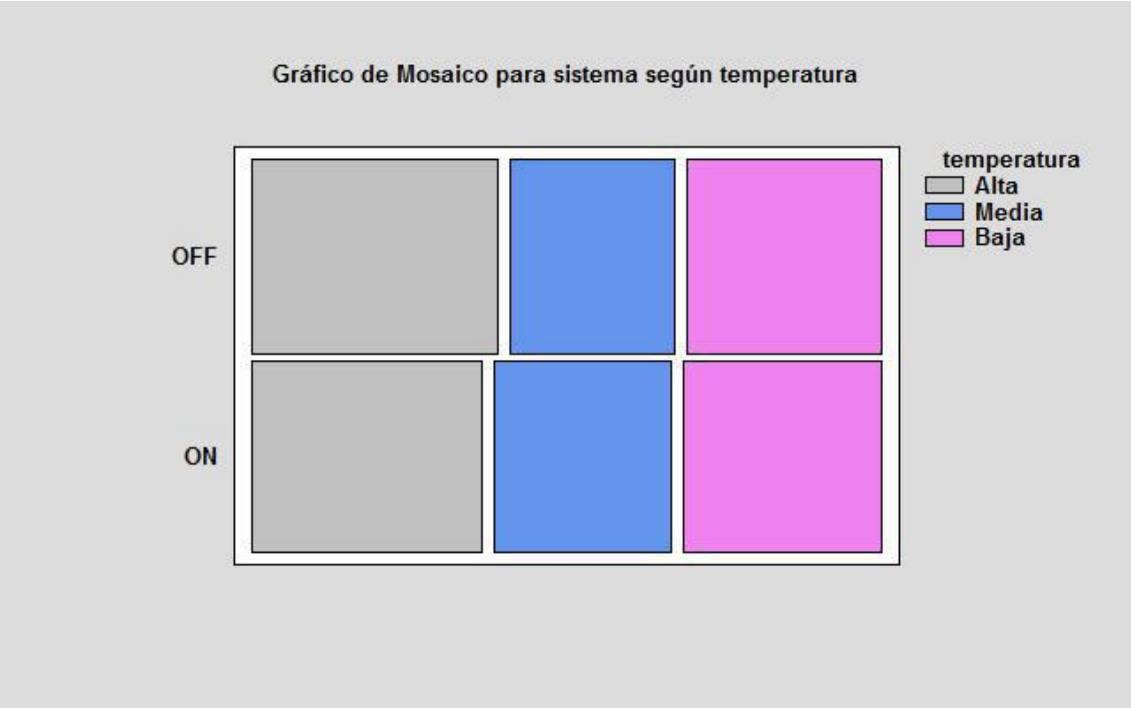

Figura 3: Diagrama de mosaico sobre las variables *sistema* y *temperatura*

Es decir, que como vemos en la gráfica de la figura 3 y como indicábamos en el párrafo anterior. Las variables *sistema* y *temperatura* son independientes entre sí, por lo tanto el valor del sistema en un caso concreto puede no estar condicionado por el valor de *temperatura*.

2.3. Ejercicio 3

Completar las frases siguientes:

- El número de datos recogidos en momentos en que no hubo averías fue ....., esto es .....% del total de datos disponibles. Para esos momentos, el *consumo* promedio fue de ..... Megavatios-hora.
- El percentil 85 de la variable *consumo* es .....

El número total de muestras de la variable *averías* que se incluye en el fichero de datos se recogen a continuación:

- $n(averias) = 117$ número total de muestras de la variable avería que contiene el fichero de datos.
- $n(averias_{si}) = 28$ número de muestras de la variable avería cuyo valor es *si*.
- $n(averias_{no}) = 89$ número de muestras de la variable avería cuyo valor es *no*.

Puesto que la variable *averias* es de carácter binario tan solo puede tomar uno de los dos valores indicados previamente, podemos demostrar este hecho comprobando que la siguiente operación se cumple:

$$n(averias) = n(averias_{si}) + n(averias_{no}) = 28 + 89 = 117$$

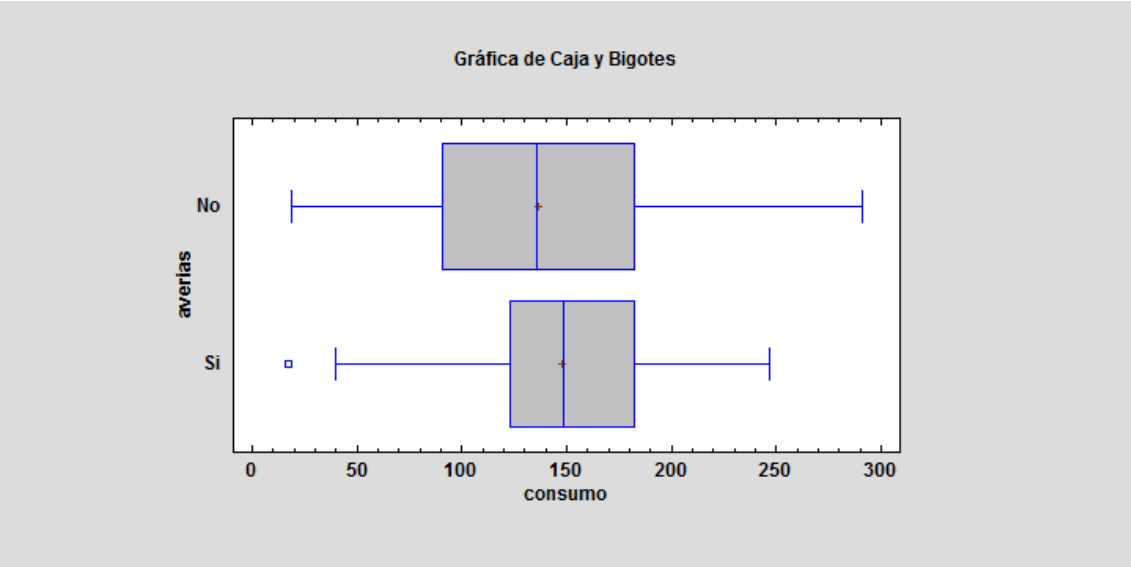

Figura 4: Diagrama de Cajas de *consumo* particionada sobre *averias*

A partir de estos valores es fácil obtener el porcentaje de veces que no ocurre ninguna avería:

$$\frac{n(averias_{no})}{n(averias)} * 100 = \frac{89}{117} * 100 = 76{,}068\,\%$$

El consumo promedio de energía cuando no se producen averías es el siguiente:

$$\overline{consumo/averias_{no}} = 136{,}759$$

En la figura 4 se muestra el diagrama múltiple de caja y bigotes de la variable *consumo* particionada sobre *averías*. Cabe destacar que este es menor que el consumo promedio global, lo que nos permite inferir que cuando surgen averías el consumo aumenta, probablemente debido a las reparaciones pertinentes u otras causas similares.

Por tanto, el texto completado es el siguiente:

- "El número de datos recogidos en momentos en que no hubo averías fue **89**, esto es **76.968 %** del total de datos disponibles. Para esos momentos, el *consumo* promedio fue **136.759 Megavatios/Hora**"

Para rellenar la segunda oración es necesario calcular el siguiente valor, que tal y como se puede apreciar en la figura 5, el resultado es correcto:

$$percentil(consumo, 0{,}85) = 194{,}92$$

Por tanto, el texto completado es el siguiente:

- "El percentil 85 de la variable *consumo* es **194.92 Megavatios/Hora.**"

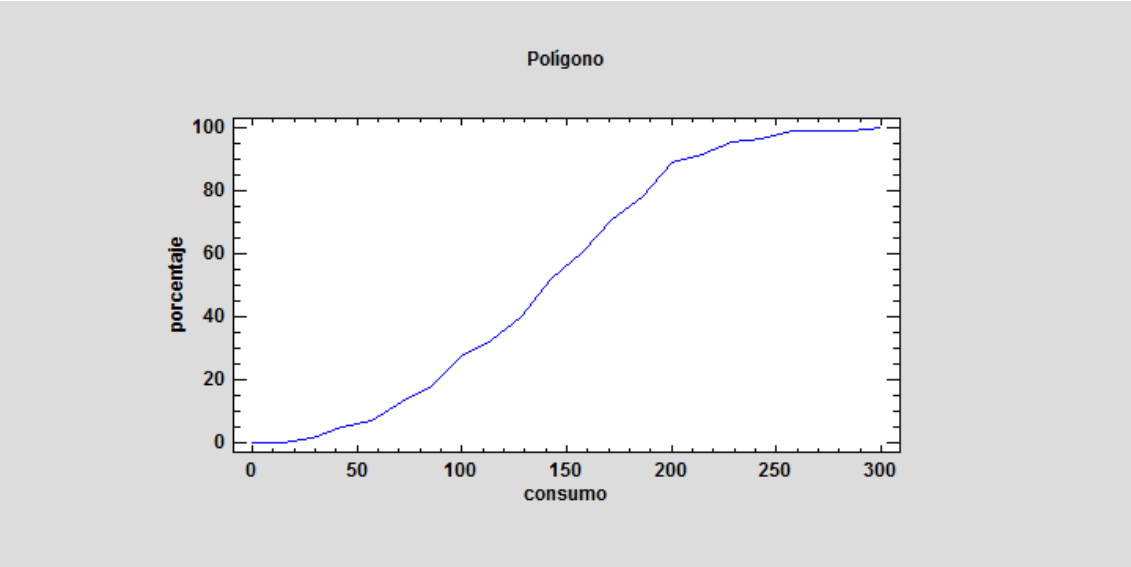

Figura 5: Polígono de frecuencias relativas acumuladas de *consumo*

2.4. Ejercicio 4

Realizar una matriz de gráficos planos para las variables *consumo*, *pr.tbc*, *pr.cc*, *pr.galv1*, *pr.galv2*. En este gráfico no deben representarse ningún diagrama de cajas en la diagonal y los puntos deben de tener color indicativo de la variable *línea*. Calcular la matriz de correlaciones. ¿Qué se puede decir de la relación entre todas estas variables?. ¿Se obtendría el mismo resultado si se realiza este análisis sólo para los datos que corresponden a momentos sin averías?

A continuación en la figura 6 se muestra la matriz de gráficos planos para las variables *consumo*, *pr.tbc*, *pr.cc*, *pr.galv1* y *pr.galv2*.

	consumo	pr.tbc	pr.cc	pr.galv1	pr.galv2
consumo		0,7433 0,0000	0,3853 0,0000	0,4013 0,0000	0,2407 0,0089
pr.tbc	0,7433 0,0000		0,1540 0,0974	0,0661 0,4786	0,1022 0,2727
pr.cc	0,3853 0,0000	0,1540 0,0974		0,3001 0,0010	0,0711 0,4463
pr.galv1	0,4013 0,0000	0,0661 0,4786	0,3001 0,0010		0,0496 0,5950
pr.galv2	0,2407 0,0089	0,1022 0,2727	0,0711 0,4463	0,0496 0,5950	

Cuadro 3: Matriz de correlaciones

La tabla 3 muestra la correlaciones entre las variables *consumo*, *pr.tbc*, *pr.cc*, *pr.galv1* y *pr.galv2*. En esta tabla, por cada celda se muestra el valor de correlación y el *Valor-P*. Los siguientes pares tienen un *Valor-P* menor que $0,05$:

- consumo y pr.tbc
- consumo y pr.cc

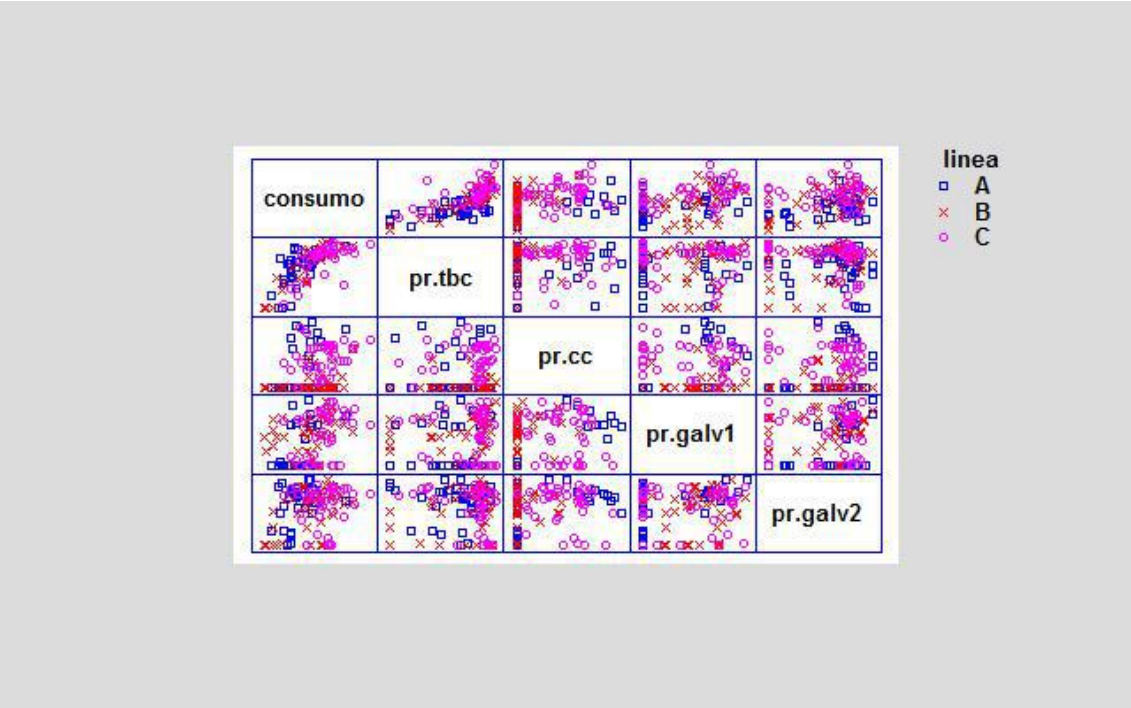

Figura 6: Matriz de gráficos planos

- consumo y pr.galv1
- consumo y pr.galv2
- pr.cc y pr.galv1

Lo que significa que existe una correlación entre la variable *consumo* y las demás variables, y entre las variables *pr.cc* y *pr.galv1*, con un nivel de confianza del $95,0\,\%$.

Tomando solo aquellos casos sin averías se obtendría un comportamiento similar como podemos observar en la tabla.

	consumo	pr.tbc	pr.cc	pr.galv1	pr.galv2
consumo		0,7167 0,0000	0,4028 0,0001	0,4343 0,0000	0,2794 0,0080
pr.tbc	0,7167 0,0000		0,1617 0,1300	0,0973 0,3644	0,1208 0,2596
pr.cc	0,4028 0,0001	0,1617 0,1300		0,2855 0,0067	0,1262 0,2385
pr.galv1	0,4343 0,0000	0,0973 0,3644	0,2855 0,0067		0,0837 0,4354
pr.galv2	0,2794 0,0080	0,1208 0,2596	0,1262 0,2385	0,0837 0,4354	

Cuadro 4: Matriz de correlaciones para momentos sin averías

Esta tabla nos muestra que el *Valor-P* en los pares anteriores sigue estando por debajo de $0,5$.

2.5. Ejercicio 5

Un elemento a valorar es la relación entre consumo de energía y producción.

- Ajusta una recta de regresión de la variable *consumo* en función de la producción del tren de bandas calientes *pr.galv1*.
- Realizar un gráfico de residuales studentizados frente a la variable independiente. ¿Existe algún punto influyente o residual atípico?
- Valora la bondad de este ajuste. ¿Es apropiado utilizar este modelo?
- Completar la siguiente tabla.

Intercep	
Pendiente	
Coeficiente de correlación entre las dos variables	
Residual correspondiente al caso 1	

Nota 1: Si en la respuesta incluyéseis un gráfico de dispersión junto con la recta de regesión, no ha de aparecer ninguna línea adicional.

La recta lineal de ajuste obtenida por Statgraphics a partir de las variables *consumo* y *pr.galv*1 es:

$$Y = 112{,}825 + 0{,}0661151X$$

Esta función se ilustra gráficamente a en la figura 7:

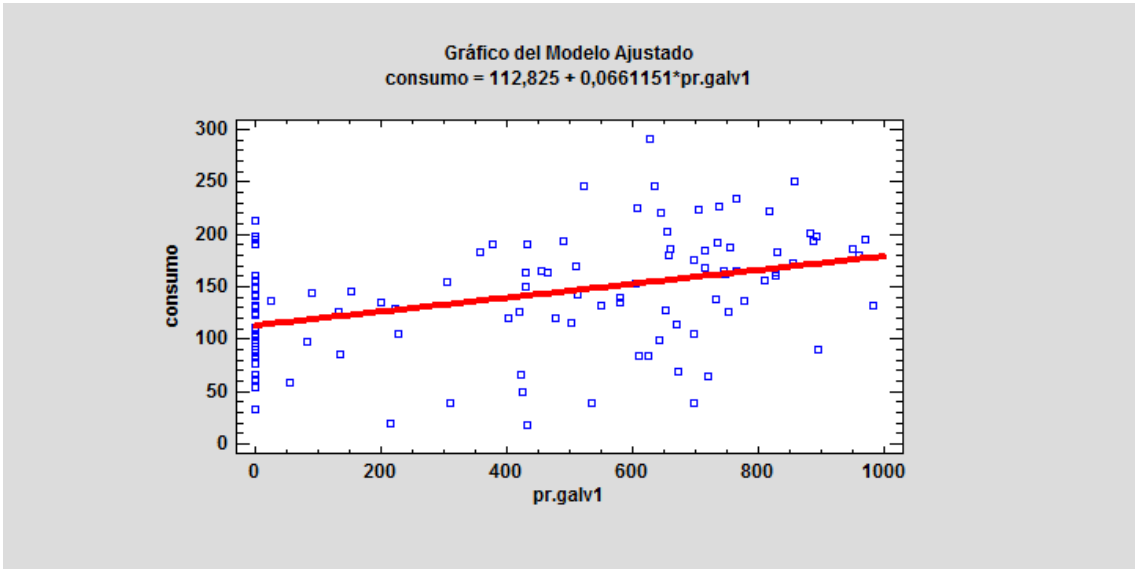

Figura 7: Diagrama de regresión lineal de *consumo* sobre *pr.galv1*

El gráfico de residuales studentizados frente a la variable independiente *pr.galv*1 se muestra en la figura 8:

No existen puntos que destaquen como influyentes, sin embargo, la influencia media de estos es del: 0,017094. En cuanto a los valores residuales atípicos, se han recogido en la tabla 5.

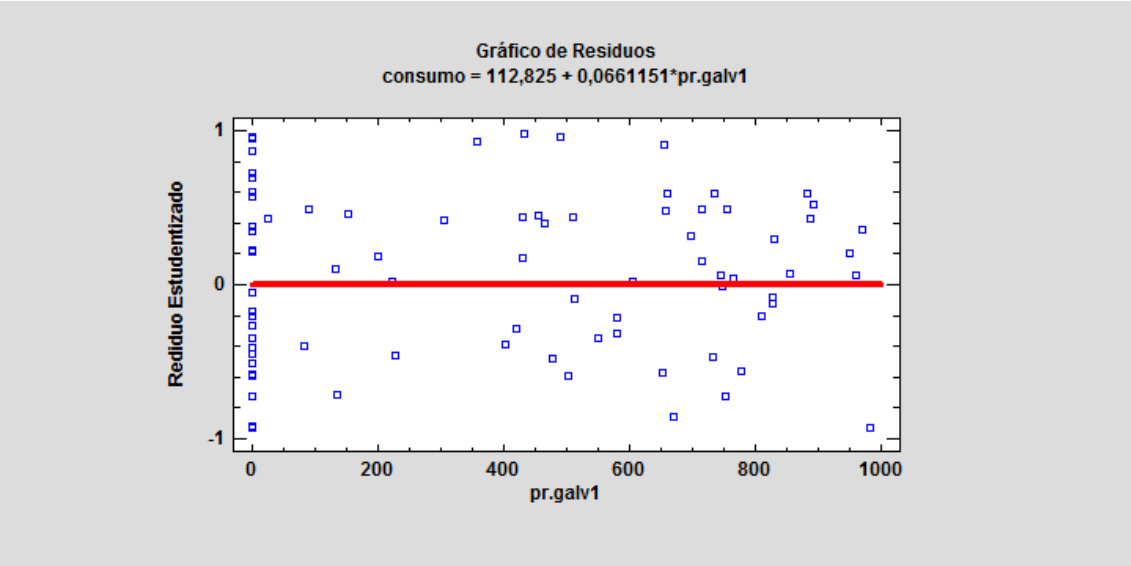

Figura 8: Diagrama de residuos studentizados correspondientes a la regresión lineal de *consumo* sobre *pr.galv1*

Fila	X	Y	Residuos Studentizados.
55	216	19.07	-2.17
56	536	38.39	-2.21
60	698	39.72	-2.42
61	432	17.50	-2.51
88	627	290.72	2.08
117	0	213.23	2.03

Cuadro 5: Residuos studentizados correspondientes a la regresión de la variable *consumo* sobre *pr.galv1*

La bondad del ajuste se analiza a partir del coeficiente de correlación r_{xy}:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{7416{,}625}{55{,}1853 * 334{,}929} = \frac{7416{,}625}{18483{,}157} = 0{,}401264$$

El coeficiente de correlación lineal de pearson toma valores en el intervalo $[-1, 1]$, reflejando una correlación fuerte cuando este valor se aproxima a los extremos del intervalo y la inexistencia de la misma en caso de que tienda a 0. Puesto que el valor obtenido es cercano a 0,5 podemos afirmar que entre las dos variables existe una correlación positiva. A pesar de ello está correlación no es muy fuerte, tal y como se puede visualizar gráficamente en el diagrama de dispersión lineal de la figura 7. Puesto que el valor es positivo significa que las dos variables están relacionadas positivamente, es decir, al aumentar una de ellas también aumenta la otra.

Los resultados de la tabla que se pide completar en el enunciado del ejercicio se recogen en la tabla 6.

Parámetro	Valor
Itercepto	112,825
Pendiente	0661151
Coeficiente de Correlación	0,401264
Residual del caso 1	−0,311627

Cuadro 6: Tabla requerida en el enunciado del ejercicio

El *intercepto* representa el valor constante en la función lineal de regresión, mientras que la *pendiente* se corresponde con el valor multiplicador que acompaña a la variable de orden 1. Tal y como se ha explicado previamente, el *coeficiente de correlación* representa el grado de similitud entre las distribuciones de las 2 variables, tomando valores en el intervalo $[-1, 1]$ y tendiendo a 0 en caso de que la similitud sea alta. Por último, el *residual correspondiente al caso 1* se refiere al valor residual que surge en la primera muestra del conjunto de datos.

2.6. Ejercicio 6

Proponer alguna cuestión sobre estos datos que consideres interesante y darle respuesta. (Plantea una pregunta que requiera realizar al menos un gráfico y/o un estadístico)

Enunciado propuesto En el ejercicio 1 se ha realizado un análisis para la variable *consumo* en los grupos marcados por la variable *hora*. Analice los estadísticos sobre las variables relacionadas con la producción (*pr.tbc*, *pr.cc*, *pr.galv1*, *pr.galv2* y *pr.pint*), y realice la gráfica de caja y bigotes sobre la producción total agrupada por la variable *hora*. Realice otras gráficas que considere convenientes y saque conclusiones sobre las relaciones entre estos datos.

	Promedio	Desviación Estándar	Coeficiente de Variación	Máximo	Rango
pr.tbc	7567,82	3002,7	39,6773 %	10979,0	10979,0
pr.cc	295,53	358,619	121,348 %	1204,0	1204,0
pr.ca	124,479	161,991	130,135 %	677,0	677,0
pr.galv1	402,803	334,929	83,1495 %	982,0	982,0
pr.galv2	1159,72	577,792	49,8217 %	1963,0	1963,0
pr.pint	188,556	289,446	153,507 %	898,0	898,0
Total	1623,15	2966,56	182,766 %	10979,0	10979,0

Cuadro 7: Tabla de comparación de estadísticos sobre variables de producción

La tabla 7 muestra una comparación de los estadísticos de las distintas variables de producción dentro del estudio del ejercicio. Caben destacar las grandes diferencias que hay entre las distintas producciones. Teniendo mayor relevancia aquellas como la producción del tren de bandas calientes(*pr.tbc*) o la producción de galvanizado de tipo II(*pr.galv2*) que soportan la mayor parte de la producción en Tm. de acero de la empresa. Estas diferencias se muestran mejor en la figura 9.

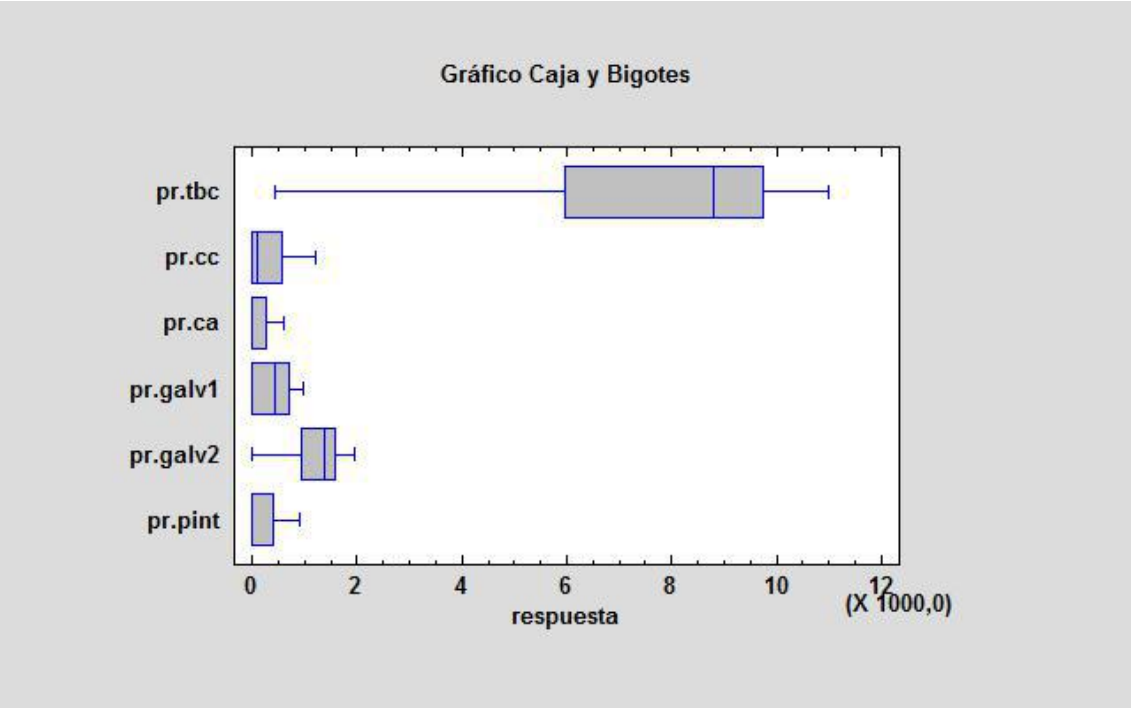

Figura 9: Diagramas de cajas sobre variables de producción

Este diagrama muestra mejor las grandes diferencias entre los estadísticos de la tabla 7. Como podemos comprobar de forma gráfica la producción correspondiente *pr.tbc* no es equiparable con las otras líneas de producción de la fábrica.

Respecto a la comparación de la producción total sobre las agrupaciones surgidas de la variable *hora*, en las figuras 10 y 11 se muestran las diferencias entre las medias y las medianas de la variable artificial de producción total[1]

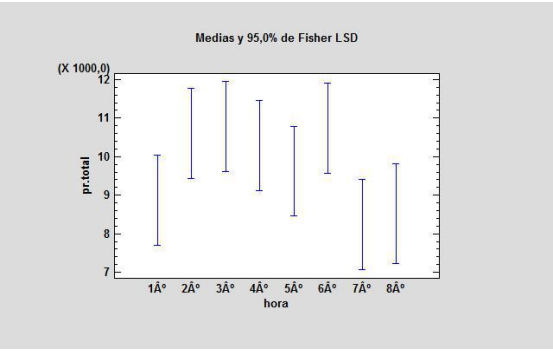

Figura 10: Diagramas de comparación de medias sobre 95 % de confianza

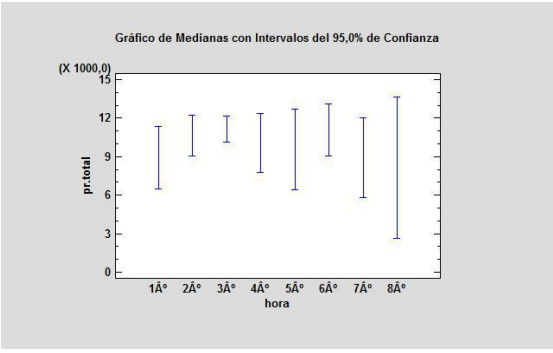

Figura 11: Diagramas de comparación de medianas sobre 95 % de confianza

[1] Veáse la sección 1.2.1, donde se expone la naturaleza de esta variable.

Entre ellas hay diferencias respecto a la interpretación que se le pueda dar. Sobre la primera figura podemos indicar que las medias sobre las diferentes horas de la jornada laboral son dispares entre sí, teniendo las medias más altas entre la segunda, tercera y sexta hora; en cambio teniendo las más bajas entre la primera, séptima y la última hora de la jornada.

La figura 11 muestra la mediana de producción total(*pr.total*) sobre cada hora, y el rango de esta. En esta figura observamos que la variación sobre la producción total en la tercera hora es muy ajustada, en cambio, en la octava hora este rango es mucho mayor, lo cual puede indicar, junto con la conclusión de una media de producción total baja, la posibilidad de ser una hora bastante menos productiva que las demás.

En la figura 12 nos ofrece una comparación mediante diagramas de cajas de la naturaleza de la variable de producción total (*pr.total*) sobre las agrupaciones según *hora*. Al igual que en los párrafos anteriores, distinguimos entre los rangos de valores que puede tomar la producción general de toneladas de acero.

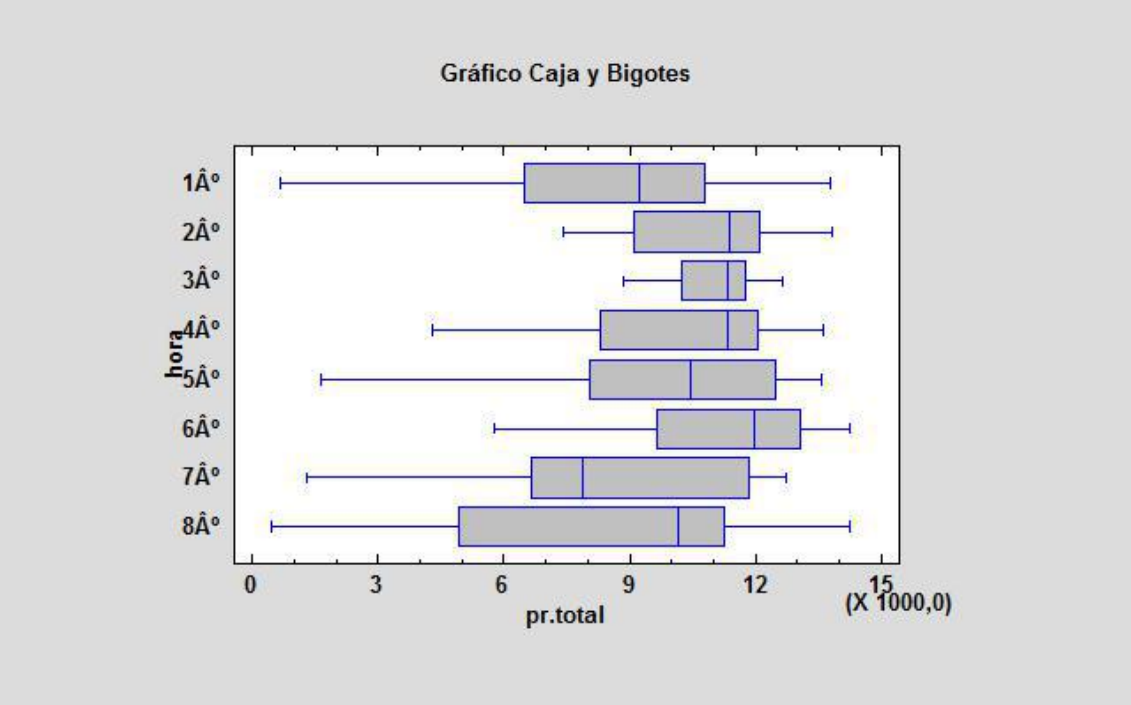

Figura 12: Diagramas de cajas sobre variables de producción agrupadas por la variable *hora*

Por último, la figura 13 muestra mediante un diagrama de sector el peso de producción según la hora. Al igual que en las conclusiones que podíamos sacar mediante los diagramas anteriores, aunque hay diferencias pequeñas en cuanto al reparto de la producción general respecto a *hora*, los mayores porcentajes se encuentran entre la tercera y la sexta hora. Teniendo mayor peso la octava hora, aunque en la figura 12 pudiésemos comprobar que era la que mayor rango de valores tenía.

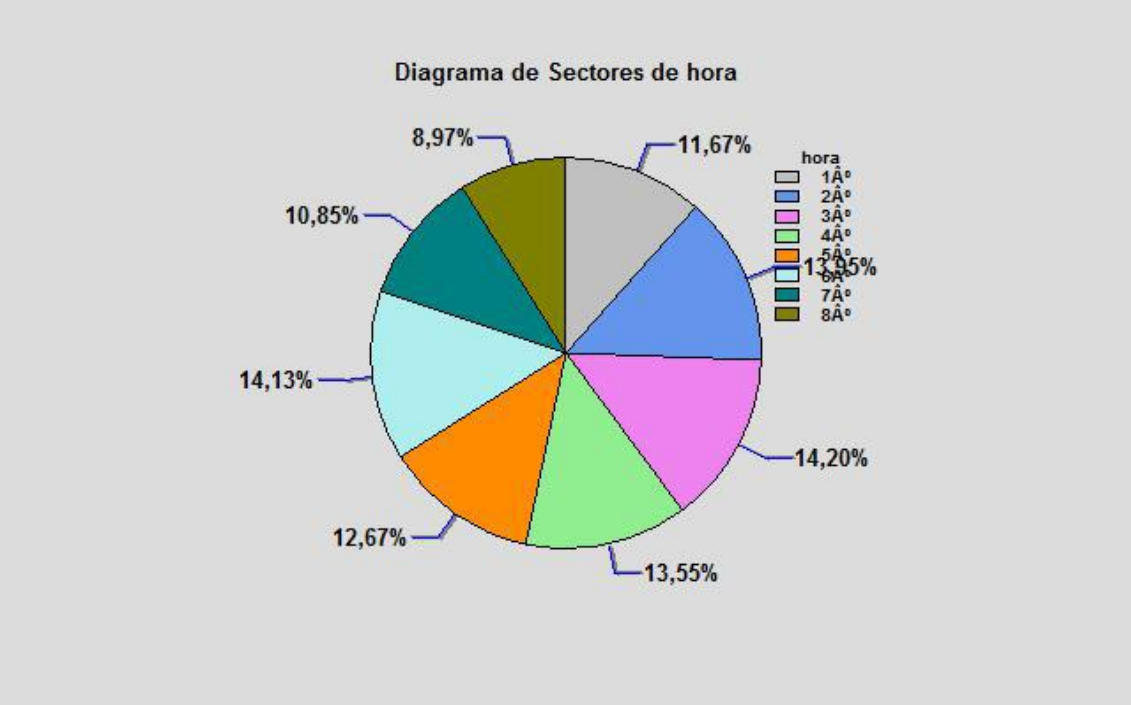

Figura 13: Diagrama de sector sobre porcentaje de producción agrupadas por la variable *hora*

3. Conclusiones

A través de los ejercicios propuestos por la práctica se han analizado los datos correspondientes a una explotación industrial destinada a la producción de distintos tipos de acero.

Se ha prestado especial atención a la variable **consumo**, tratando de descubrir información relevante acerca de la misma y su relación con otras como su división por **líneas**, su relación con la existencia de **averías** o la relación con la **producción** que se ha realizado en los ejercicios 4 y 5.

También ha sido de especial interés el estudio acerca del funcionamiento del sistema de **sobrecalentamiento** con respecto a la **temperatura** del mismo, revelando la independencia de estas.

Por último, mediante análisis de valores estadísticos hemos podido conocer detalladamente la taxonomía de producción de la explotación industrial estudiada mediante el ejercicio de libre enunciado, obteniendo como conclusión que el componente principal en que se basa es la producción de **bandas calientes de tren**.

Con la realización de esta práctica se ha comprendido mediante un caso práctica cómo los sucesos reales están fuertemente relacionados por los valores de un amplio número de variables que nos permiten conocer detalles significativos que ha simple vista no seríamos capaces de observar ni analizar. Por tanto, la estadística descriptiva es una herramienta imprescindible en el análisis de dichos sucesos.