TRABAJO PRÁCTICO 2: INFERENCIA ESTADÍSTICA

CURSO 2021-2022

"ANÁLISIS DEL DATASET G02_DATOS_DEM_G1"

Autores:

Enrique Campos Alonso

Patricia Renart Carnicero

Sergio Rodríguez Vidal

Álvaro Pereira Chagoyen

Madrid, abril de 2022

Índice de Contenidos

1. INTRODUCCIÓN

Este documento expone el trabajo de Enrique Andrés Campos Alonso, Patricia Renart Carnicero, Sergio Rodríguez Vidal y Álvaro Pereira Chagoyen, alumnos de la Escuela Técnica Superior de Ingeniería (ICAI) de la Universidad Pontificia Comillas. Dicho trabajo lleva como título "ANÁLISIS DEL DATASET G02_DATOS_DEM_G1" y se ha realizado dentro de la asignatura "Probabilidad y Estadística" de 1º de iMAT.

En el trabajo se plantea una serie de cuestiones acerca de la demanda convencional del gas y la temperatura máxima diaria, y su relación mediante un enfoque cuantitativo-cualitativo.

El documento está estructurado en 5 apartados. En el apartado segundo se plantea una introducción general al problema que permite entender la motivación del mismo. En el apartado tercero se detallan modelados probabilísticos de la demanda

2. MODELADO PROBABILÍSTICO DE LA TEMPERATURA

En este apartado se analizan algunos aspectos de las temperaturas disponibles.

2.1. Estimación de la temperatura TMAX esperada en julio

Estima un intervalo de confianza al 95% para la media de TMAX en el mes de julio. Para decidir el método de estimación de dicho intervalo, realiza un contraste no paramétrico para determinar si se puede considerar que la variable TMAX sigue una distribución normal, obrando en consecuencia.

En este apartado se estima la temperatura máxima que se puede esperar en el mes de julio a partir de la muestra de datos disponible (ver Figura 1).

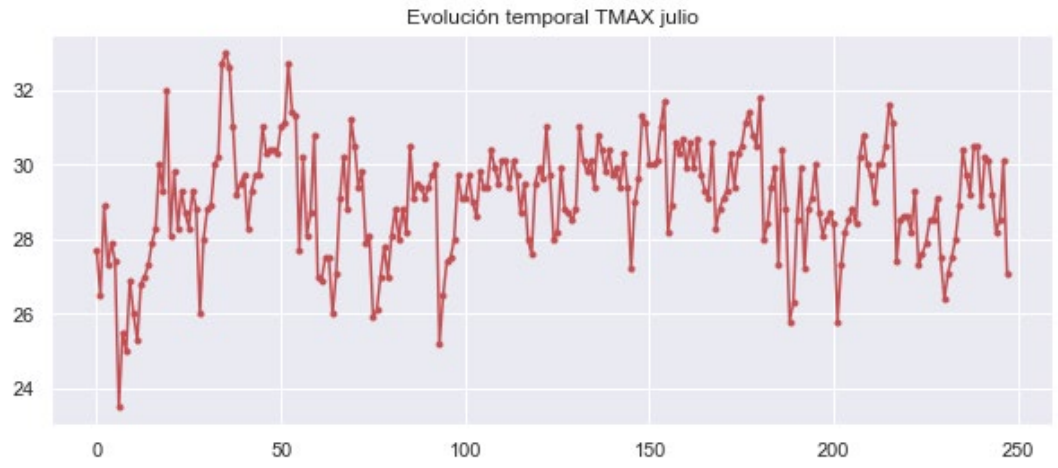

Figura 1. Evolución temporal TMAX de julio

Para realizar dicha estimación se han utilizado todos los datos de TMAX disponibles para el mes de julio. En la Figura 2 se muestra el histograma de dichos datos, junto con el diagrama de cajas correspondiente.

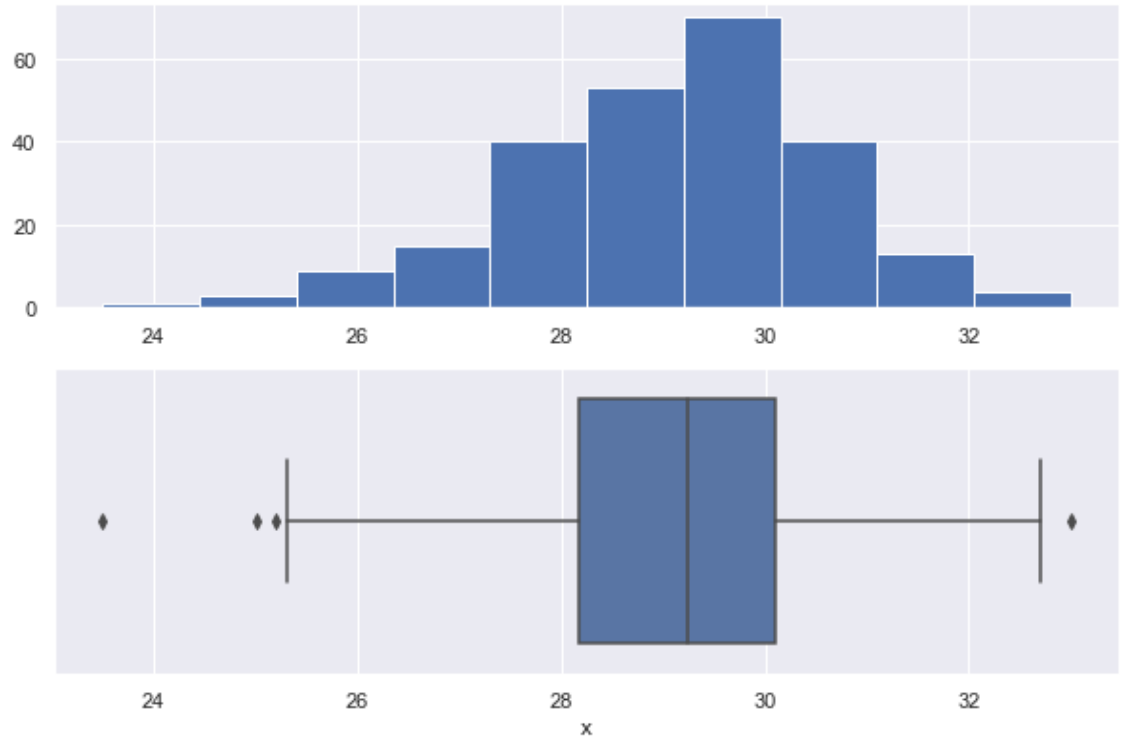

Figura 2. Histograma y diagrama de cajas de la variable TMAX en el mes de julio.

Ahora se quiere determinar si se puede considerar que la TMAX sigue una normal, para lo que se observa el qqplot (ver Figura 9) o el test KS (ver Figura 8). Para ambos métodos, la muestra está al límite entre poder ajustarse a una normal y no, por lo que se optará por no asumir nada. Por ello se optará por el uso de Bootstrap para calcular el intervalo de confianza.

Empleando este método, se estima que el intervalo de confianza al 95% para la media de la TMAX de julio es [28.8830, 29.2593] (ver Figura).

2.2. Distribución de TMAX según los meses del año

En este apartado se estimará la función de densidad de probabilidad de la TMAX para cada mes. El procedimiento para conseguir dicha función es igual para cada mes. Primero se mostrará el gráfico qqplot respecto de la normal y se hará el contraste Kolmogorov – Smirnov (KS), donde se sacarán la distancia máxima y el p_valor. En caso vamos a utilizar un nivel de significación del 5%. A partir de los datos obtenidos se decidirá si se estimará con la normal o habrá que utilizar una estimación de densidad del kernel (KDE). Esto último se comparará con un histograma (20 bins) de la distribución de cada mes para ver si la estimación realizada es buena.

Una vez se han seleccionado los datos y divididos en variables individuales las TMAX de cada mes, se puede trabajar con los datos. Cómo ya se tiene una ligera idea de cómo

se distribuyen gracias al último trabajo realizado, podemos discurrir directamente a los qqplots y su comparación con la normal. Los resultados obtenidos son los siguientes:

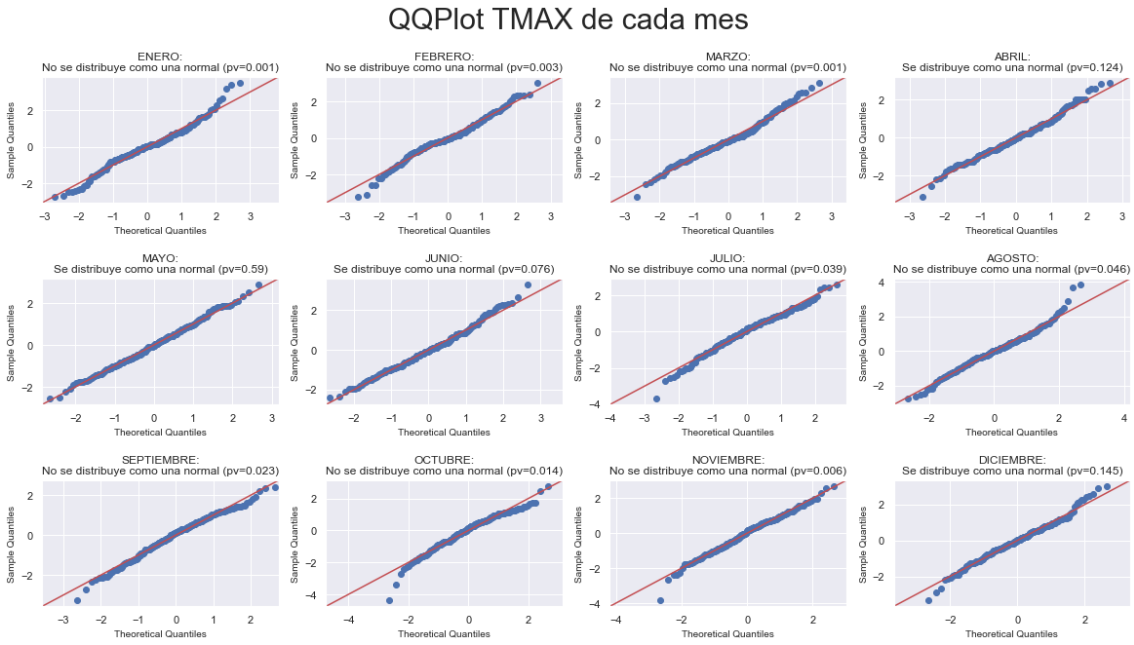

Figura 3: Qqplots TMAX por cada mes

- **ENERO** - KS en normal: máx. dist. = 0.077895 -> p_valor = 0.001
- **FEBRERO** - KS en normal: máx. dist. = 0.077659 -> p_valor = 0.002931
- **MARZO** - KS en normal: máx. dist. = 0.087519 -> p_valor = 0.001
- **ABRIL** - KS en normal: máx. dist. = 0.053631 -> p_valor = 0.124477
- **MAYO** - KS en normal: máx. dist. = 0.037814 -> p_valor = 0.589656
- **JUNIO** - KS en normal: máx. dist. = 0.056861 -> p_valor = 0.075595
- **JULIO** - KS en normal: máx. dist. = 0.060883 -> p_valor = 0.039044
- **AGOSTO** - KS en normal: máx. dist. = 0.059703 -> p_valor = 0.045753
- **SEPTIEMBRE** - KS en normal: máx. dist. = 0.064822 -> p_valor = 0.023116
- **OCTUBRE** - KS en normal: máx. dist. = 0.067401 -> p_valor = 0.013639
- **NOVIEMBRE** - KS en normal: máx. dist. = 0.072458 -> p_valor = 0.006197
- **DICIEMBRE** - KS en normal: máx. dist. = 0.051986 -> p_valor = 0.144756

De los resultados anteriores podemos sacar varias conclusiones. Si comparamos los p_valores de cada mes con el nivel de significación del 5% declarado anteriormente, podemos afirmar que los que superan el 0.05 (en verde) se pueden aproximar con una normal. Estos son los meses de ABRIL, MAYO, JUNIO y DICIEMBRE. El resto de los meses: ENERO, FEBRERO, MARZO, JULIO, AGOSTO, SEPTIEMBRE, OCTUBRE y NOVIEMBRE tendremos que realizar KDEs. Este análisis también se puede observar con los qqplots. En los meses que se puede aproximar por una normal, los datos están mucho más próximos a la línea roja, mientras que en el resto suele haber mucha más variación en los extremos. Los mejores casos para ver estas diferencias son con los meses de mayo, muy estable, con enero, mucho más variable.

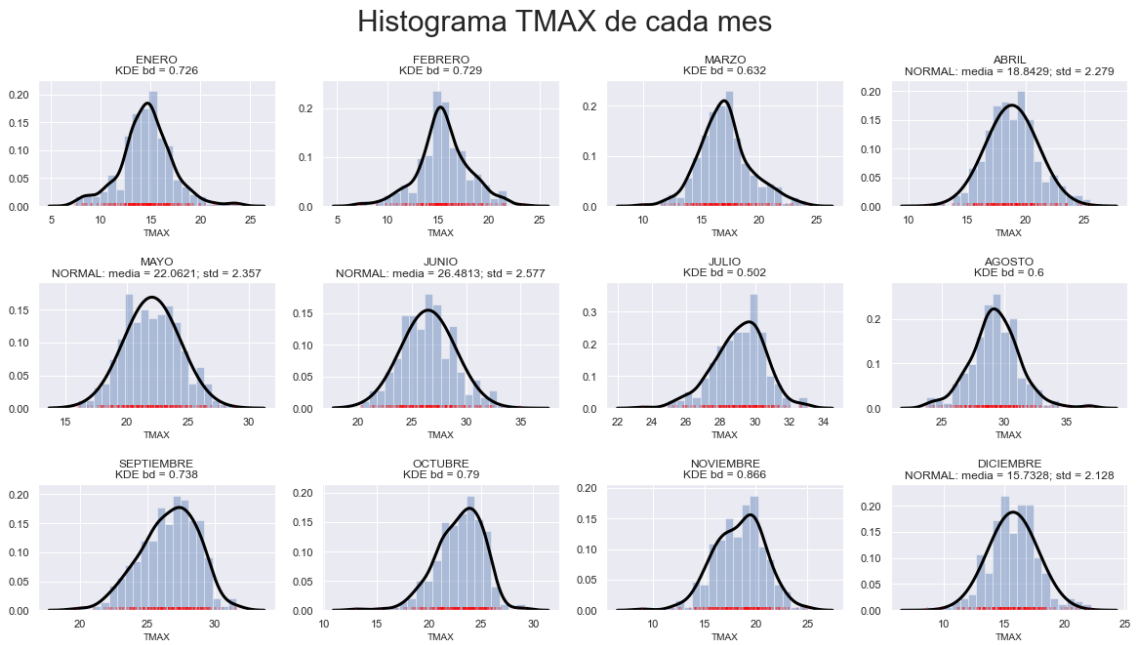

Figura 4: Histogramas TMAX por cada mes

A continuación, se procederá a estimar las funciones de densidad de probabilidad y compararlos con los histogramas de sus respectivas distribusiones. En el caso de lo histogramas se han utilizado 20 bins. Para estimar las KDE, el bandwith se ha calculado automáticamente en referencia a una normal. Las normales se han calculado con las respectivas medias y desviaciones típicas de cada distribución. Los resultados son los siguientes:

3. MODELADO PROBABILÍSTICO DE LA DEMANDA

En este apartado se trabaja con las demandas de los días laborables (es decir, ignorando fines de semana y festivos), de todos los años salvo los datos de los años 2020 y 2021.

3.1. Distribución de la demanda de los laborables según los meses del año

En este apartado se trata de identificar y ajustar un modelo de densidad de probabilidad para la demanda diaria de los días laborables, condicionado al mes del año. Es decir, el objetivo es estimar la función de densidad de probabilidad de la demanda de los laborables para cada mes.

En primer lugar, hay que determinar para qué meses es razonable ajustar una distribución normal. La decisión final se tomará utilizando un contraste no paramétrico (gráfico qqplot y el contraste KS). Para el resto de meses para los que no es razonable utilizar una distribución normal, estima la función de densidad de probabilidad usando KDE. Indica gráficamente para cada mes la distribución ajustada junto al histograma de los datos en un subplot de 3 x 4. También muestra los parámetros que define la distribución en cada mes (para KDE el bandwidth utilizado). Comenta los resultados.

En este apartado se va a analizar la demanda energética por cada mes desde 2014 hasta 2019 (ambos inclusive). El objetivo de ello será determinar cuáles son los meses cuya distribución de demanda puede ser aproximada a una distribución normal y cuáles no.

Previamente al desarrollo de la solución, los datos deberán ser modificados para garantizar que la información con la que se trabaja no sea errónea y pertenezca al ámbito de estudio pedido (días laborables desde 2014 a 2019 por meses). Tras crear un dataframe que contenga la información proporcionada, este será variado para únicamente contener los días laborables, es decir, que entren en días del 2 al 6 en la clasificación semanal del archivo de datos dado. Posteriormente se descartarán los pertenecientes a los años del 2020 al 2022 y las columnas de información innecesaria serán también extraídas del dataframe.

Para ello se utilizará la representación del gráfico qqplot y el posterior cálculo del p-valor, estableciendo este último como principal condicionante de la decisión.

Aquellos meses cuya demanda cuente con un p-valor menos al 0.05 han sido considerados como no aptos para ser aproximados por medio de una distribución normal. En estos casos se utilizará el ajuste por medio de KDE con un bandwidth calculado de manera automática para cada una de las distribuciones.

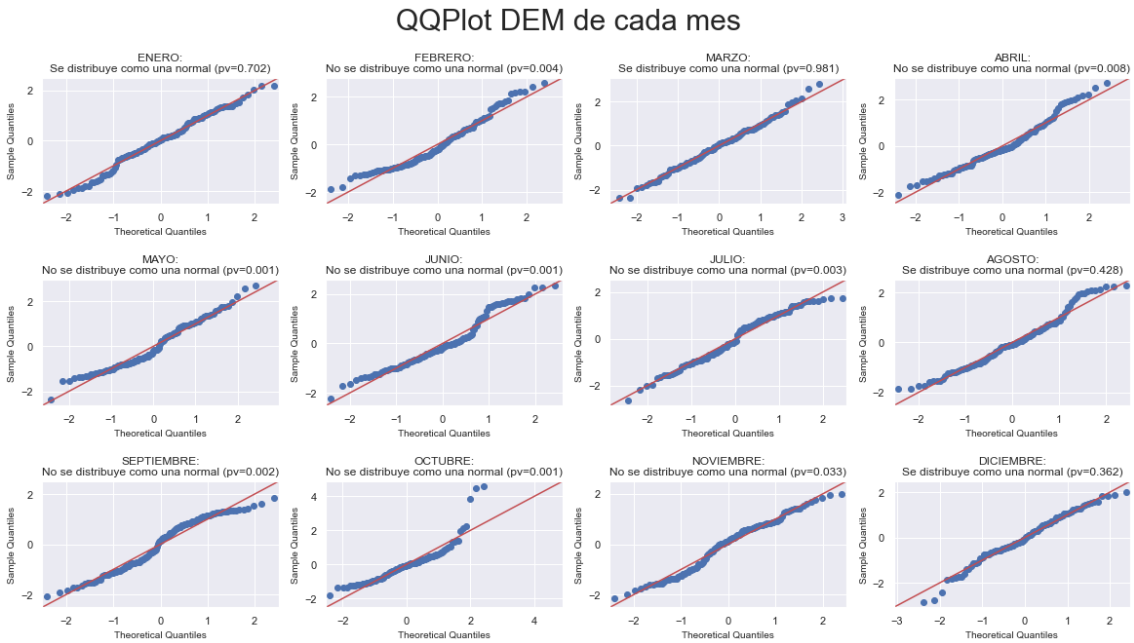

Figura 5: QQplots demanda por cada mes (laborales – 2014 a 2019)

- **ENERO** - KS en normal: máx. dist. = 0.0492 -> p_valor = 0.7019
- **FEBRERO** - KS en normal: máx. dist. = 0.1054 -> p_valor = 0.003557
- **MARZO** - KS en normal: máx. dist. = 0.03429 -> p_valor = 0.9805
- **ABRIL** - KS en normal: máx. dist. = 0.09917 -> p_valor = 0.00838
- **MAYO** - KS en normal: máx. dist. = 0.1233 -> p_valor = 0.000999
- **JUNIO** - KS en normal: máx. dist. = 0.1207 -> p_valor = 0.000999
- **JULIO** - KS en normal: máx. dist. = 0.1036-> p_valor = 0.002717
- **AGOSTO** - KS en normal: máx. dist. = 0.05807-> p_valor = 0.4280
- **SEPTIEMBRE** - KS en normal: máx. dist. = 0.1071 -> p_valor = 0.00223
- **OCTUBRE** - KS en normal: máx. dist. = 0.1170 -> p_valor = 0.000999
- **NOVIEMBRE** - KS en normal: máx. dist. = 0.08698 -> p_valor = 0.03328
- **DICIEMBRE** - KS en normal: máx. dist. = 0.06293 -> p_valor = 0.3616

Tras realizar el análisis previamente expuesto, se llega a la conclusión de la existencia de únicamente 4 meses del año cuya distribución puede ser aproximada por una normal, como podemos observar claramente en la Figura 5. Dichos meses serán **enero, marzo, agosto y diciembre.** Por lo tanto, el resto de los meses tendrán distribuciones cuyo p-valor sea menor que 0.05 y serán rechazadas como normales, siendo posteriormente aproximadas por medio del KDE, también observable en la Figura 5, ya que ninguno de dichos meses tiene una aproximación que pueda asemejarse a una normal, comúnmente debido a la existencia de más de un máximo en la distribución, como es el caso de **julio**.

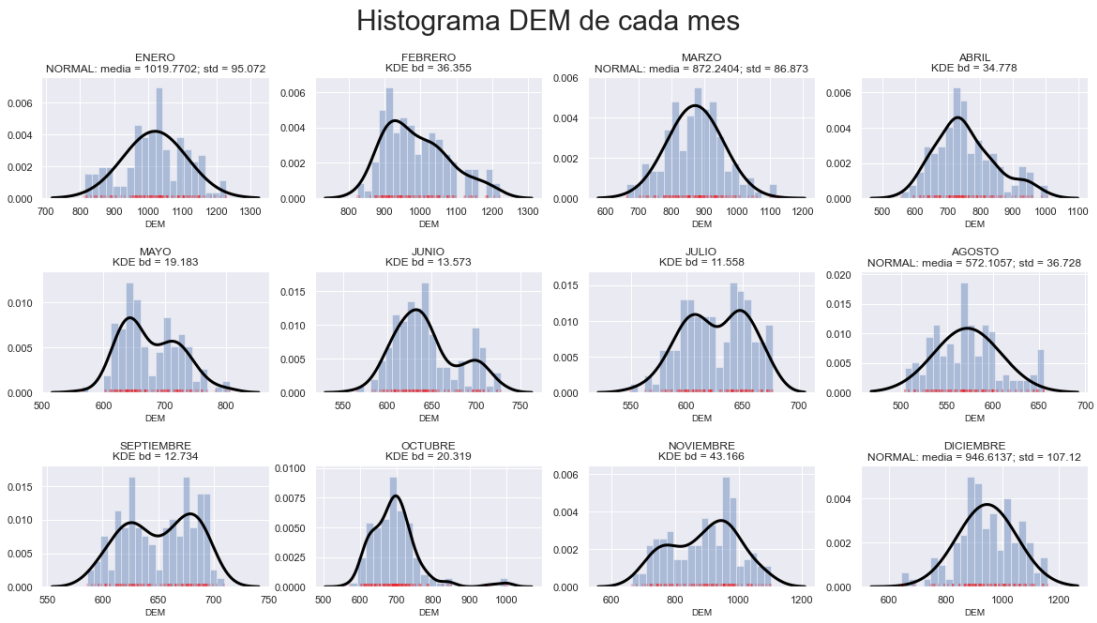

Figura 6: Histogramas demanda por cada mes (laborales – 2014 a 2019)

3.2. ANOVA de la demanda de los laborables con los meses

Responde claramente a la pregunta ¿Cambia la demanda de los laborables con los meses? Para ello realiza un ANOVA, incluyendo la tabla ANOVA resultante, los gráficos de cajas del factor y la conclusión final en vista de los resultados obtenidos.

Como se indica en el nombre del apartado, se ha realizado un ANOVA para determinar si la demanda de los días laborables varía según el mes del año (ver Figura 11. Demanda de laborables por mes).

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	33924380.50085	11	3084034.590986	557.51304	0.0	2.001333
Within Groups	8292125.061108	1499	5531.771222			
Total	42216505.561958	1510	27957.950703			

Figura 7: Tabla ANOVA

Observando la tabla, se puede afirmar que la demanda varía con los meses, como también se puede ver en la Figura .

4. CONCLUSIONES

La demanda convencional del gas en España en días laborales (de 2014 a 2019) **depende fuertemente del mes en el que se mida**. Esto probablemente se deba a que **mucha gente demande una mayor cantidad de gas en meses más fríos** y, por consecuente, menor en meses más cálidos (para calentarse). Otro motivo, puede ser a que **en meses de verano, invierno o pascua** (abril), **muchos trabajadores libran para pasar tiempo con sus familias o de vacaciones** (sin ser necesariamente días festivos).

El **próximo julio** (2022), podríamos encontrarnos en torno a **temperaturas máximas de entre 28.8830 y 29.2593 grados con un nivel de confianza del 95%**. Estas **temperaturas son bastante "típicas" o "medias"** a comparación del resto de temperaturas máximas medidas en julio. La temperatura máxima media en julio es del 29.0713 ºC. **Por tanto, deberíamos poder esperar una demanda diaria igual de "típica" en julio** (alrededor de los **603.007 GWH**), aunque con la bajada de la demanda en los últimos meses de 2022 podríamos esperar una demanda menor.

Los **meses** que siguen una **distribución de la TMAX más irregular** son **enero, febrero y marzo**. Casualmente, estos son los meses de invierno. En cuanto a la **demanda en días laborales**: **mayo, junio y octubre** son los meses más variables.

5. ANEXO A

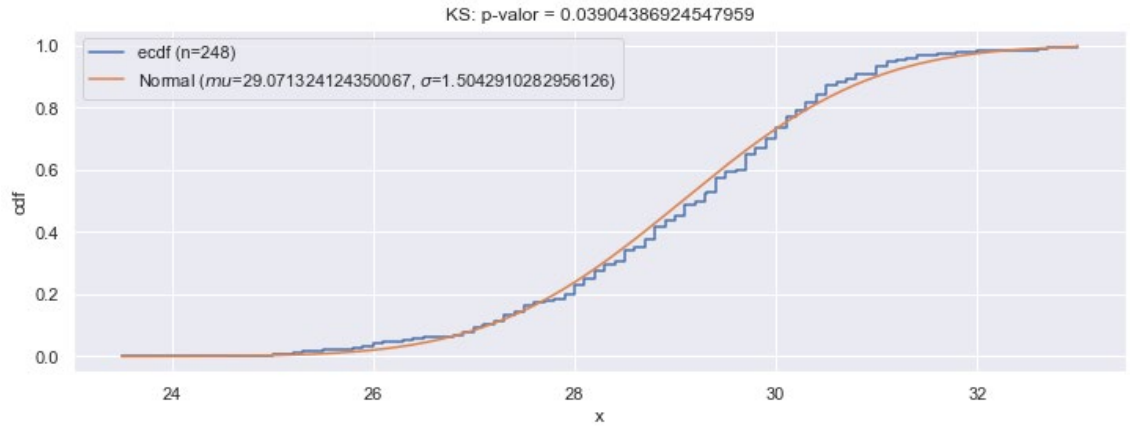

Figura 8. Gráfica de las distribuciones empírica y teórica

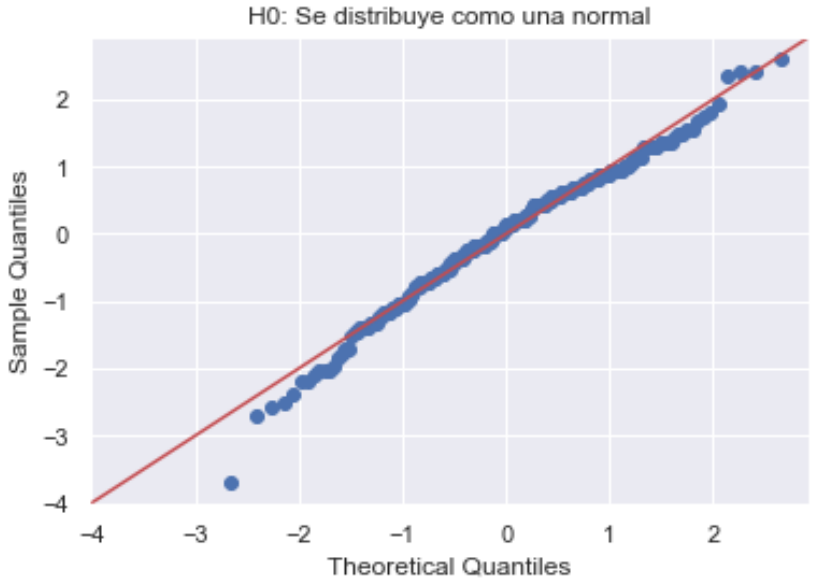

Figura 9. Qqplot respecto a una normal

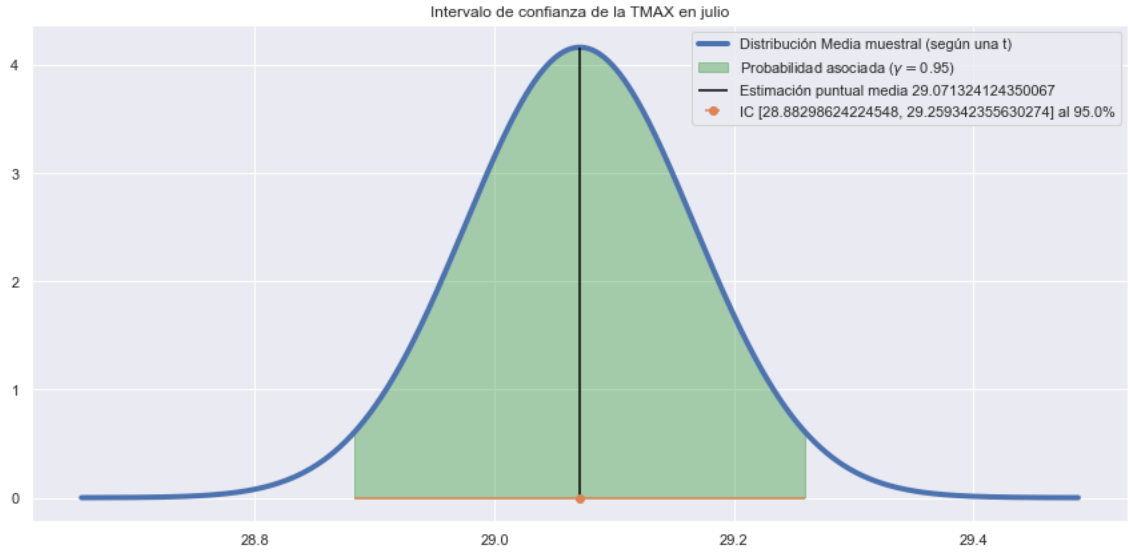

Figura 10. Intervalo de confianza de la media, comparado con una t de Student

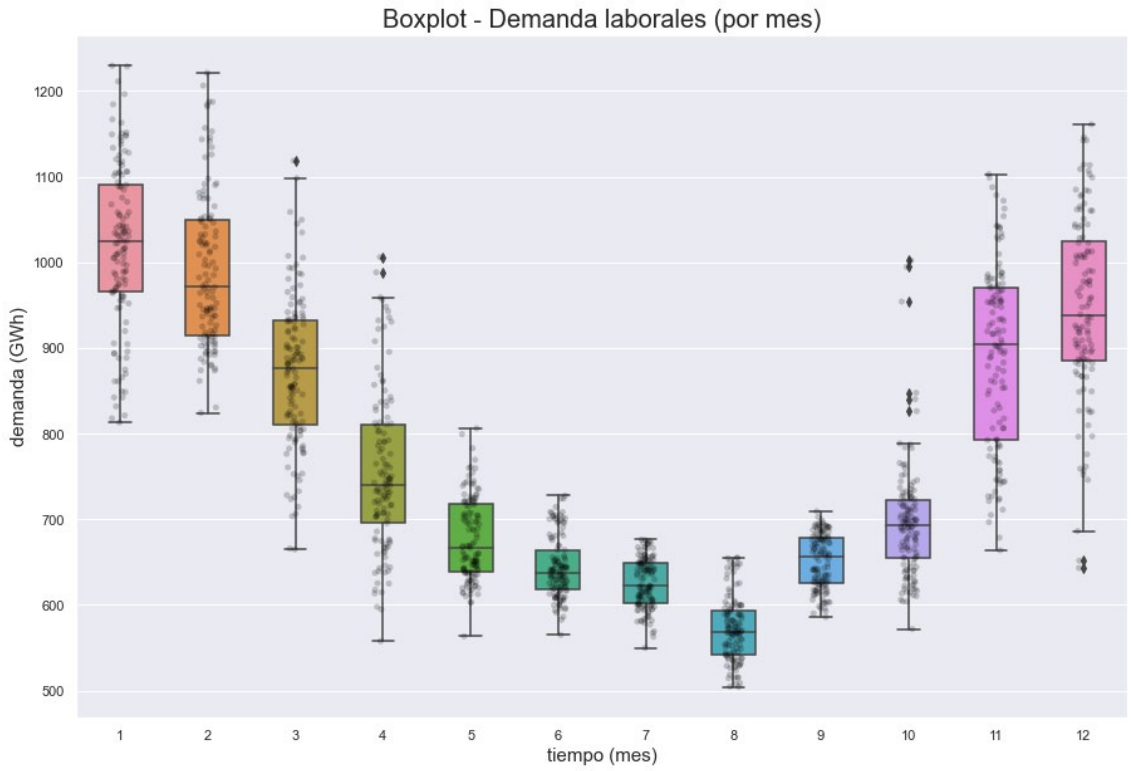

Figura 11. Demanda de laborables por mes