Trabajo práctico 4

Fundamentos de análisis de datos

Luciana Gattuso

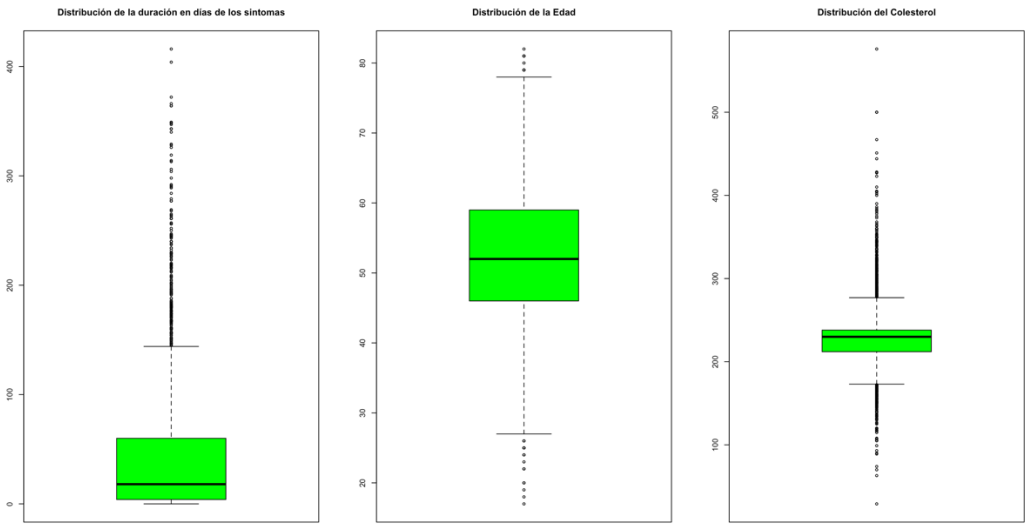
Ejercicio 1

Se analizaron los datos *acath*.sav que corresponden a una muestra de 3504 pacientes que acudieron al centro con dolor en el pecho, para los que se recogieron diversas variables cuyo nombre en la base de datos y descripci´on es la siguiente:

- sigdz : variable binaria que toma valores 1 y 0, indicando si el paciente presenta estrechamiento de alguna de las arterias coronarias de al menos un 75 % (sigdz = 1) o no (sigdz = 0).
- tvdlm : lo mismo que la anterior pero corresponde a tres arterias con estrechamiento.
- sex: variable categórica que indica el género del paciente: 0 corresponde al género masculino, 1 al género femenino.
- age: variable continua que representa la edad en años del individuo.
- choleste: variable continua que expresa los Mg/dl de colesterol
- duración: variable continua que recoge la duración, en días, de los síntomas de la enfermedad coronaria.

Para un mejor análisis del conjunto, la variable tvdlm quedó excluida del dataframe por no ser relevante. En un primer, análisis descriptivo es posible visualizar que, de 3504 registros, 1246 corresponden a NAs de la variable Colesterol (choleste). Estos registros fueron reemplazados por la media de la columna e imputados al dataframe arrojando nuevos resultados.

Figura 1



Del conjunto de datos sólo fueron graficadas las variables continuas, es decir la edad del paciente, duración de los síntomas y el nivel colesterol. La variable duración no muestra una distribución simétrica, es decir que la mediana no está situada al centro de la caja. Por otro lado, edad y colesterol, a pesar de ser complicado observar un sesgo en la figura 1, presentan valores similares de medias y medianas respectivamente indicando una distribución simétrica. La edad media y mediana de los pacientes es de 52 y el nivel de colesterol de 229. Además, todas las variables presentan valores atípicos. El conjunto de datos contempla pacientes desde los 17 a 59 años. El máximo nivel de colesterol registrado fue 576 mientras que el 3er cuartil de 238.

En segundo lugar, se aplicó un modelo de regresión logística simple utilizando sigdz como variable explicativa, la misma indica si un paciente presenta estrechamiento de alguna arteria coronaria. El modelo utilizó como única variable explicativa la variable colesterol. El resumen del modelo cómo el colesterol es una variable predictora relevante, ya que el p valor (4.81e-11) es muy bajo y un valor z de 6.57.

Asimismo, se calculó que la probabilidad de que una persona con colesterol 199 tenga estrechamiento arterial es de 0.62. Teniendo en cuenta que para una persona con un nivel de colesterol medio 229 tiene una probabilidad de 0.66, una persona con colesterol 576 -el valor máximo registrado en el dataframe – tiene una probabilidad de 0.94 y, en comparación, de 29 -el mínimo registrado- de colesterol cuenta con 0.36 de probabilidad. Esto indica que la variable colesterol es significativa para el análisis del estrechamiento arterial.

En tercer lugar, se utilizó un modelo de regresión logística considerando todas las variables no categóricas. En este caso, el colesterol y la edad de los pacientes obtuvieron valores de p debajo del 0.05, a diferencia de la variable duración de los síntomas, que con 0.73 supera el umbral. Además, edad y colesterol contaron con valores de z más alejados del cero (13.56 y 6.45) mientras que el valor z para Duración se acercó con 0.34. Lo que evidencia este segundo modelo es que los niveles de colesterol y la edad de un paciente son significativos para explicar la probabilidad de sufrir un estrechamiento arterial. Sin embargo, la duración de los síntomas de la enfermedad coronaria no explica la probabilidad de ocurrencia de un estrechamiento arterial. Se realizaron predicciones para utilizando los valores de las medianas de cada variable y se obtuvo como resultado 0.67.

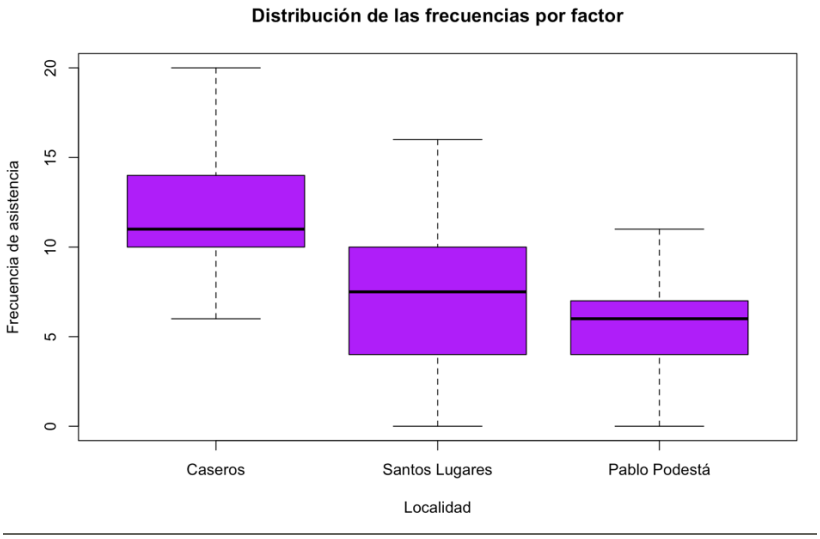
Asimismo, se incorporó la variable sexo como factor al modelo de regresión logística. Los resultados del comando summary del modelo es un valor p para sexo menor a 0.05 (2e-16) y un valor z de -22.34 indicando una influencia significativa sobre el estrechamiento arterial. Así la predicción del modelo, para las medianas de las variables y distinguiendo por sexo, da como resultado que para personas del género masculino (0) cuentan con un 0.81 de probabilidad de estrechamiento arterial. A diferencia de las personas de género femenino que, con los mismos valores de colesterol, edad y duración de los síntomas, presentan un 0.36 de probabilidad.

Ejercicio 2

El departamento de psicología de la Universidad de Tres de Febrero ha realizado un estudio sobre la frecuencia con que los alumnos asisten a clases teóricas no obligatorias, pertenecientes a tres localidades del partido, tomadas durante un cuatrimestre. Se han utilizado los datos que tomaron los profesores a algunos estudiantes y se busca conocer si existen diferencias estadísticamente significativas entre la cantidad estudiantes universitarios que asisten a clase dependiendo de la localidad a la que pertenecen.

Los datos están conformados por la frecuencia de asistencia a clase y la información sobre la localidad a la que pertenecen los estudiantes. Se graficó por Localidad la frecuencia de asistencia en la figura 2.

Figura 2



Se observa que los estudiantes de Caseros cuentan con la mediana más alta del conjunto (11.00). En segundo lugar, se encuentra Santos Lugares (7.50) y, por último, Pablo Podestá con la mediana de asistencia más baja (6.00). Respecto de las medias, Caseros tiene una asistencia media 11.60, Santos Lugares 6.90 y Pablo Podestá 5.45. Las tres variables analizadas presentan algún tipo de asimetría en la distribución de los datos, en otras palabras, ninguna es de distribución simétrica y, no presentan valores atípicos.

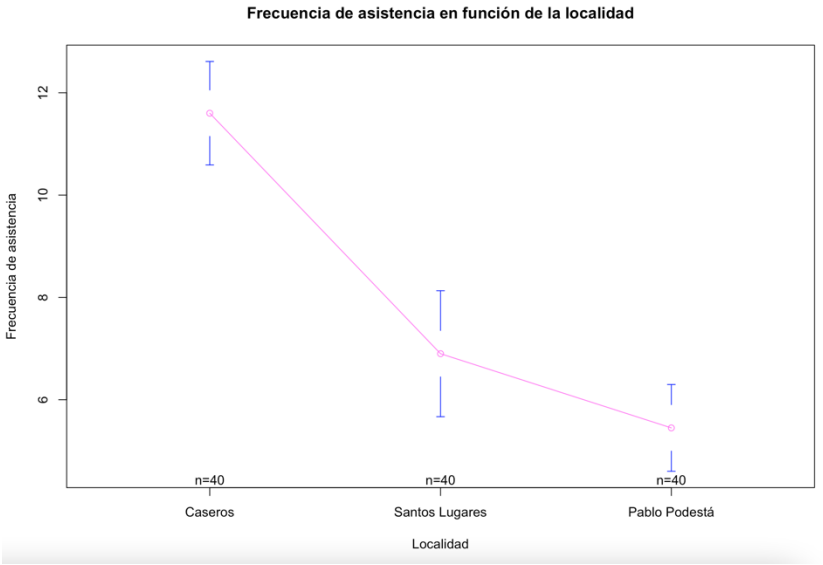
Con el objetivo de explicar la relación entre la frecuencia de asistencia y la localidad se realizó un test de ANOVA de factor fijo que compara las medias de las tres poblaciones. La hipótesis nula indica que la localidad no incrementa de manera significativa la frecuencia de asistencia y, por eso, todas las medias de las localidades analizadas deberían ser iguales. La hipótesis alternativa, por su

parte, establece que existe al menos una media de las localidades que es significativamente diferente a las demás, por lo tanto, la localidad a la que pertenece el alumno influenciaría sobre su asistencia a clase.

El resultado del modelo ANOVA indica que el estadístico estudiado que figura como valor F es de 38.98 y, que el p valor para las localidades se encuentra debajo del umbral del 0.05 (1.07e-13) reflejando una diferencia entre las medias. También devuelve la varianza entre clases para la localidad (413.4) e intraclases para los residuos (10.6).

El p valor permite no rechazar la hipótesis alternativa. Se podría concluir que la media de la localidad de residencia es estadísticamente diferente. Por lo tanto, está sí afecta sobre la asistencia a clases a la Universidad.

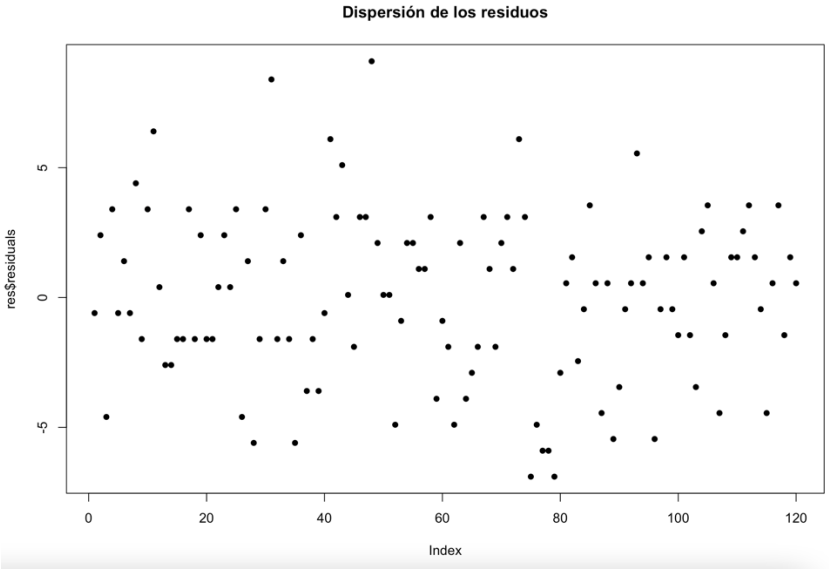
Figura 3



La figura 3 grafica las frecuencias de asistencia en función de la localidad, utilizada como factor, se observa el promedio y la desviación estándar para cada factor. Caseros presenta una diferencia evidente respecto de Santos Lugares y Pablo Podestá que se ubican en frecuencias similares.

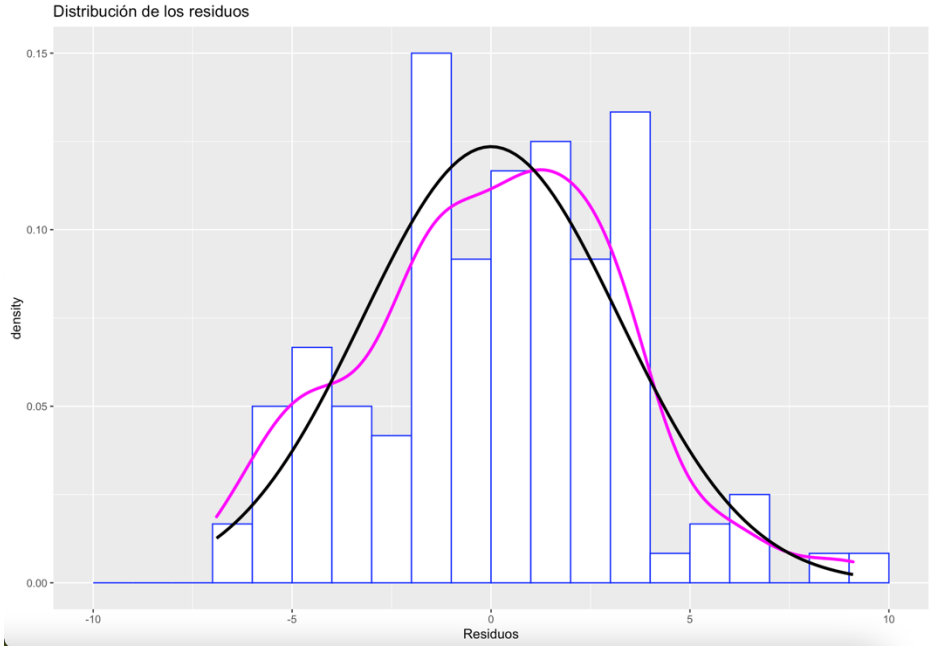
Se realizó una validación del modelo, en primer lugar, se evaluó la independencia de los residuos utilizando un gráfico de dispersión.

Figura 4



En la figura 4 es posible visualizar que no existe ningún tipo patrón entre las variables ni se detecta tendencias a un tipo de correlación sino, por el contrario, los registros se distribuyen de manera independiente sobre el gráfico. Asimismo, una hipótesis de normalidad fue utilizada para evaluar la distribución de los residuos que fue graficado en un histograma.

Figura 5



Se observa que la distribución aproximada graficada en magenta se acerca a la distrución teórica de normalidad en negro sin embargo no termina de ser una distribución normal.

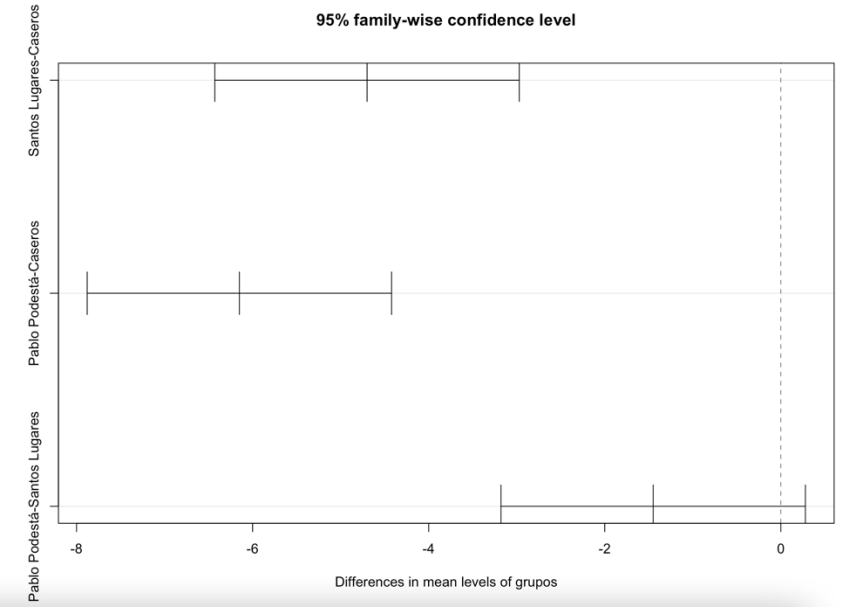
Por último, también se realizó un análisis de las varianzas para validar el modelo. La desviación estándar respecto de la localidad calculada fue 3.16 para Caseros, 3.84 Lantos Lugares y 2.65 Pablo Podestá. La división entre el máximo y el mínimo resultó 1.44, es decir que existe una desviación significativa.

A modo de conclusión, la validación del modelo no fue exitosa porque si bien los residuos corroboraron ser independientes, no se pudo validar una distribución normal ni una varianza similar. Es necesario evaluar que otro método que permita reevaluar la hipótesis del modelo.

Resulta relevante evaluar el contexto del problema, la Universidad cuenta con una sede central ubicada en Caseros. La localidad más cercana a Caseros es Santos Lugares y, la más lejana, Pablo Podestá. La universidad también cuenta con otra sede en Ciudad Jardín Lomas del Palomas, más cercana a Pablo Podestá, es posible que los estudiantes de esa localidad elijan esa sede.

La prueba de ANOVA devolvió una diferencia significativa en las medias delos grupos, para determinar cuáles son esas diferencias se evaluaron los resultados con el test HSD de Tukey.

Figura 6

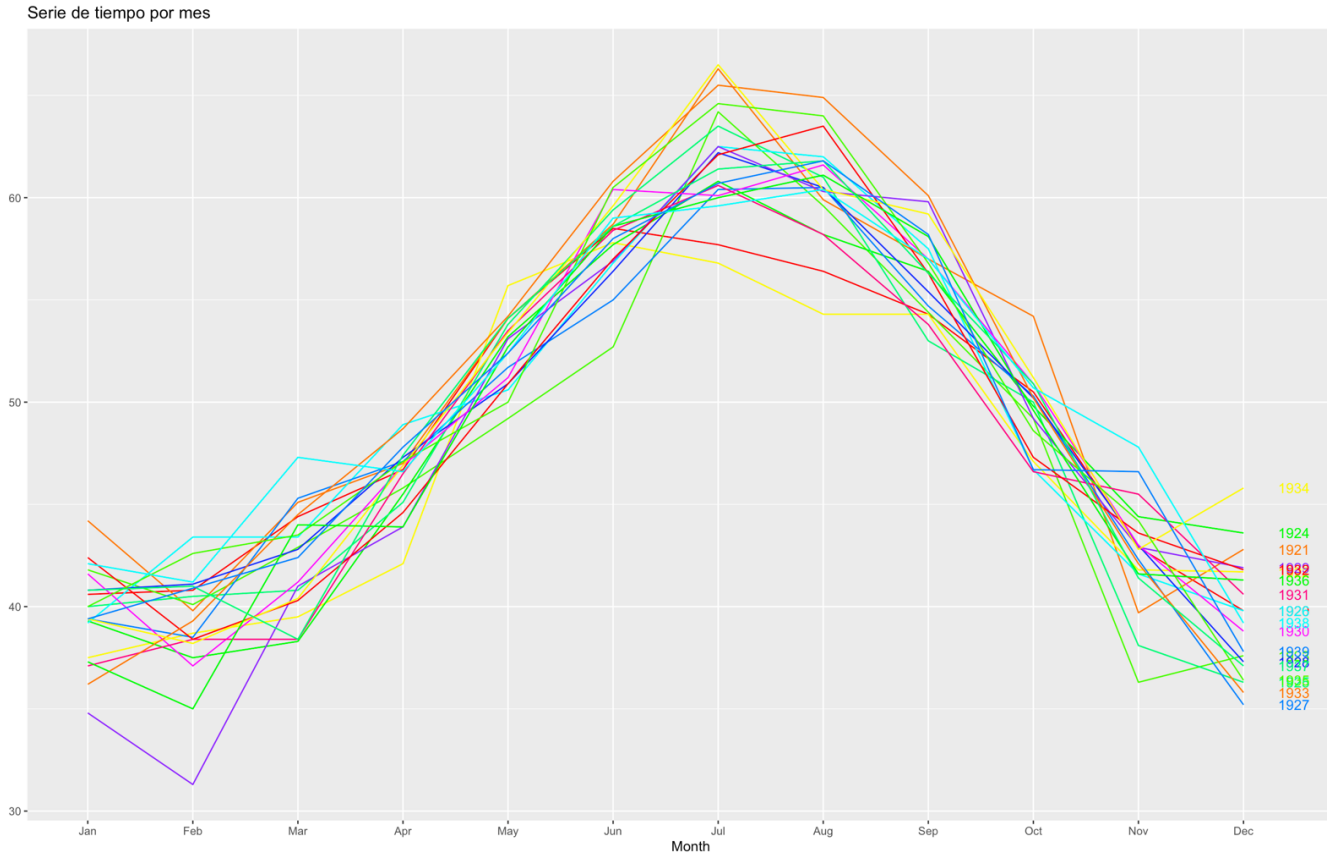


El mismo demostró que sólo hay diferencia significativa de las medias dentro del intervalo que corresponde a Pablo Podestá y Santos Lugares.

Ejercicio 3

Se evaluó las series de tiempo del dataset *nottem* que contiene las temperaturas promedio por mes de la ciudad de Nottinham desde 1920 a 1939. Se corroboró que el formato del dataset fuese una serie de tiempo con el comando *class* descartando la necesidad de definir la serie de tiempo. Los datos tienen como fecha inicial enero de 1920 con 40.6 °F y final diciembre de 1939 con 37.8 °F.

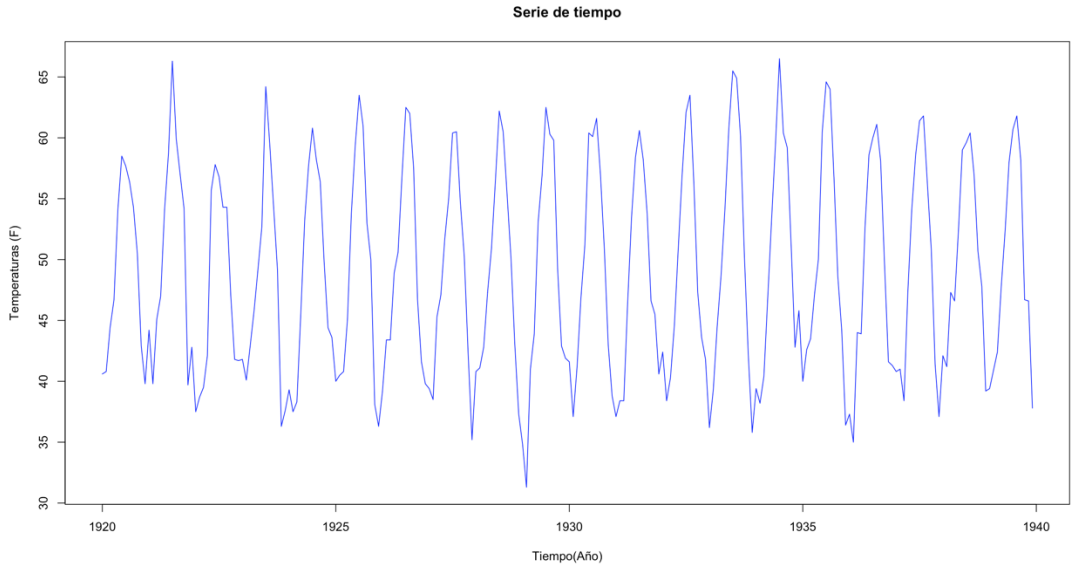
Figura 7



El gráfico de series refleja el verano en Junio-Julio con temperaturas altas y el invierno en Enero-Febrero alcanzando temperaturas mínimas. Es posible observar la variabilidad por año de las temperaturas para cada mes. Por ejemplo, la temperatura más baja registrada en el conjunto de datos fue registrada en el mes de Febrero.

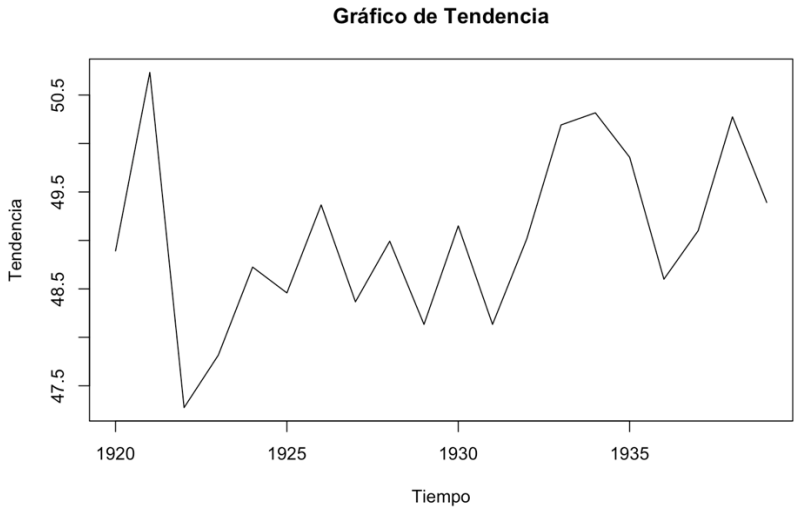
Se gráfica con el comando decompose el análisis de cada una de las componentes.

Figura 8



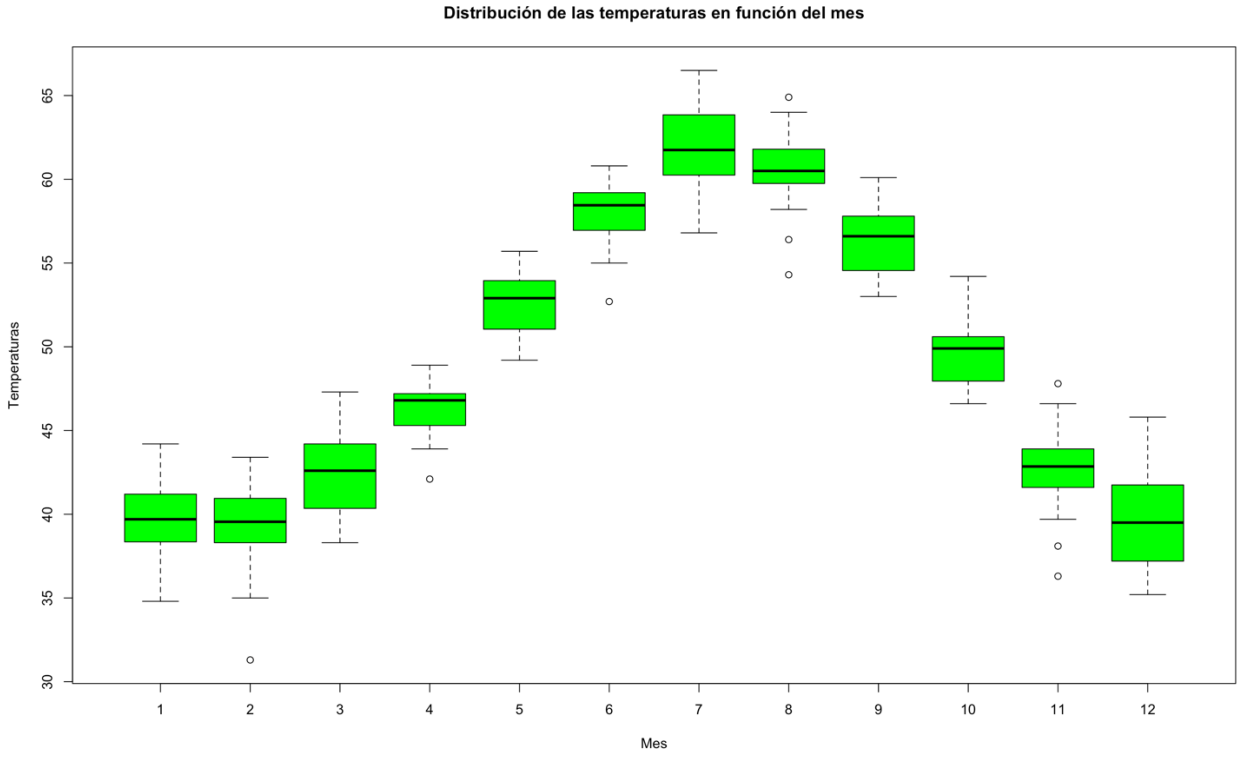
La figura 8 demuestra algunos valores atípicos como picos de máximos y mínimos, algunos cambios repentinos como una amplitud de la temperatura de 25 grados un año y el año siguiente sólo de 5 aproximadamente. No se observa una tendencia a aumentar o disminuir a largo plazo. Se visualizan movimientos ciclicos, no son intervalos regulares pero tampoco son aleatorios.

Figura 9



El gráfico de tendencia refleja la presencia de valores atípicos, una varianza no constante por la diferencia entre los mínimos y máximos y, una covarianza tampoco constante, dada por la variabilidad de amplitud entre los intervalos.

Figura 10



El gráfico de cajas anual permite visualizar outliers en los meses de abril, junio, agosto y noviembre. También meses con tendencia a una distribución normal como enero, febrero, mayo y noviembre.

En base a las figuras 8, 9 y 10 se puede concluir que los datos de la serie de tiempo no poseen estacionalidad debido a que la variabilidad de los registros cambia en el tiempo.

Por último, se analizó un subconjunto de entrenamiento desde enero 1920 a diciembre 1938. El último año de la serie se reservó para la prueba. El modelo ARIMA devolvió valores como el error medio absoluto, en este caso, 1.69 y la raíz cuadrada del error 2.19.

Figura 11

```
> summary(modelo1)
Series: AP
ARIMA(1,0,2)(1,1,2)[12] with drift

Coefficients:
         ar1     ma1     ma2     sar1     sma1     sma2   drift
      0.1526  0.1225  0.1100  -0.5181  -0.5447  -0.2066  0.0041
s.e.  0.3759  0.3716  0.1229   0.1742   0.1882   0.1744  0.0041

sigma^2 estimated as 5.236:  log likelihood=-490.45
AIC=996.91   AICc=997.6   BIC=1023.91

Training set error measures:
                     ME     RMSE      MAE        MPE     MAPE      MASE
Training set -0.06289015 2.190789 1.696956 -0.4251381 3.650275 0.6174907
                    ACF1
Training set 0.002508707
>
```

El comando forecast se realizó una predicción sobre los 12 posteriores y se compararon con el subconjunto de testeo. Se calculó el error cuadrático medio 3.75, un error no significativo entre para valores que oscilan entre los 30 y 60 grados F.