



11.67 Estadística Aplicada

Comisión Especial

Informe Final

Azul de los Ángeles Makk

Paula Ariana González

Cosatto Ammann, Pedro Camilo

Índice

1. Introducción	2
2. Análisis descriptivo	3
2.1. Selección de datos y limpieza	3
2.2. Análisis Descriptivo de datos numéricos	3
2.2.1. fixed.acidity	4
2.2.2. density	7
2.2.3. alcohol	11
2.2.4. pH	15
2.3. Relaciones entre variables	19
2.4. Estudio de una variable a elección	24
3. Estimación Puntual	28
3.1. Elección de variables	28
3.2. Cálculo de estimadores	29
3.2.1. $q_1 = x \cdot 0.9$	29
3.2.2. $q_2 = P(x > 3.40)$	30
4. Bondad de Ajuste	31
3.1. Log-verosimilitudes máximas	31
3.2. Construcción e interpretación de QQ-Plots	31
3.3. Distribución empírica ajustada a cada modelo	32
3.4. Histograma con la densidad ajustada a cada modelo	33
5. Intervalos de confianza	34
6. Regresión	35
6.1. Análisis Exploratorio	35
6.2. Diagnóstico	38
6.2.1. Supuesto de linealidad de la regresión	39
6.2.2. Supuesto de normalidad de los errores	40
6.2.3. Supuesto de homocedasticidad de los errores	41
6.2.4. Supuesto de independencia de los errores	42
6.2.5. Outliers y puntos influyentes	43
6.3. Validación del modelo	45
7. Conclusiones	48
8. Bibliografía	49

1. Introducción

El siguiente informe tiene como objetivo realizar un análisis descriptivo y de regresión de un conjunto de datos relacionados con propiedades físico-químicas de vinos de la cepa Vinho Verde, cosechada en Portugal. Se abordarán variables numéricas como la acidez fija, densidad, contenido de alcohol y pH, categorizadas en función a su calidad.

En el análisis descriptivo, se seleccionarán los datos relevantes y se realizará una limpieza inicial para garantizar su calidad. Se llevará a cabo un análisis de las variables numéricas, incluyendo medidas de tendencia central, dispersión y distribución.

Además, se explorarán las relaciones entre las variables para identificar posibles patrones o correlaciones. Se dará especial atención a una variable seleccionada, estudiándola en detalle para comprender su comportamiento y su relación con las demás variables.

Se realizará la estimación puntual de dos valores de interés: el percentil 90 de una variable y la probabilidad de que otra variable supere un valor determinado. Se elegirán las variables adecuadas y se calcularán los estimadores correspondientes.

En cuanto a la bondad de ajuste, se utilizará el enfoque de log-verosimilitudes máximas para comparar diferentes modelos y se construirán e interpretarán QQ-Plots. Además, se ajustará una distribución empírica a cada modelo y se construirán histogramas con la densidad ajustada.

En la sección de regresión, se llevará a cabo un análisis exploratorio de la relación entre una variable de respuesta y una variable explicativa. Se realizará un diagnóstico para evaluar supuestos como linealidad, normalidad de los errores, homocedasticidad e independencia de los errores. También se identificarán posibles outliers y puntos influyentes.

Se realizará una validación del modelo para evaluar su desempeño y se presentarán las conclusiones obtenidas a partir del análisis realizado.

2. Análisis descriptivo

2.1. Selección de datos y limpieza

Para el presente trabajo práctico se decidió estudiar la base de datos que muestra la calidad de distintos tipos de vinos tintos llamada "winequality-red.csv". La misma contiene observaciones de vinos de la variedad Vinho Verde, elaborada en Portugal. La base de datos seleccionada posee 12 columnas: acidez fija, acidez volátil, acidez cítrica, azúcar residual, cloruro, sulfuro de dióxido neto, sulfuro de dióxido total, densidad, ph, sulfatos, alcohol y calidad.

Para el alcance de este trabajo se seleccionaron únicamente las columnas de acidez fija, densidad, alcohol y ph, siendo las cuatro variables numéricas. Adicionalmente, a modo de realizar un análisis categórico se tomó la variable 'quality', el cual le asigna un valor numérico a la calidad del vino. Para poder profundizar en el comportamiento de las observaciones de cada categoría, se han seleccionado las que se le asignan los números 4, 6 y 8 a modo de reflejar una calidad baja, media y alta respectivamente. La muestra aleatoria tomada de un total de 709 observaciones, es de un total de 500 muestras -sin reposición-.

Se ha identificado que la base de datos se trata de una muestra transversal ya que se basa en la observación de los vinos al mismo tiempo. Su obtención fue mediante el sitio web Kaggle (véase bibliografía).

2.2. Análisis Descriptivo de datos numéricos

En esta sección se realiza un análisis de cada una de las 4 variables seleccionadas para poder entender mejor cómo se comportan. Para llevar a cabo el análisis se utilizó R en RStudio y para producir los gráficos se instaló el paquete "tidyverse" que viene con la librería de *GGPLOT2*. Dicha librería fue posteriormente empleada para generar todas las visualizaciones disponibles en las figuras adjuntas, de manera en la que se dispongan de manera prolífica y personalizada.

2.2.1. fixed.acidity

$$F = \text{Acidez fija de un vino Vinho Verde.}$$

La variable F mide la suma de todos aquellos ácidos que, al someter el vino al calor, no se evaporan. Primeramente, se observa que la mayor acidez que se hallada en la muestra fue de 14.3 gramos, mientras que la menor fue de 4.7 gramos, otorgando un rango de 9.6 gramos. A fines de poder observar mejor a la variable, se grafica la función de distribución empírica que se puede observar en la *Figura I*.

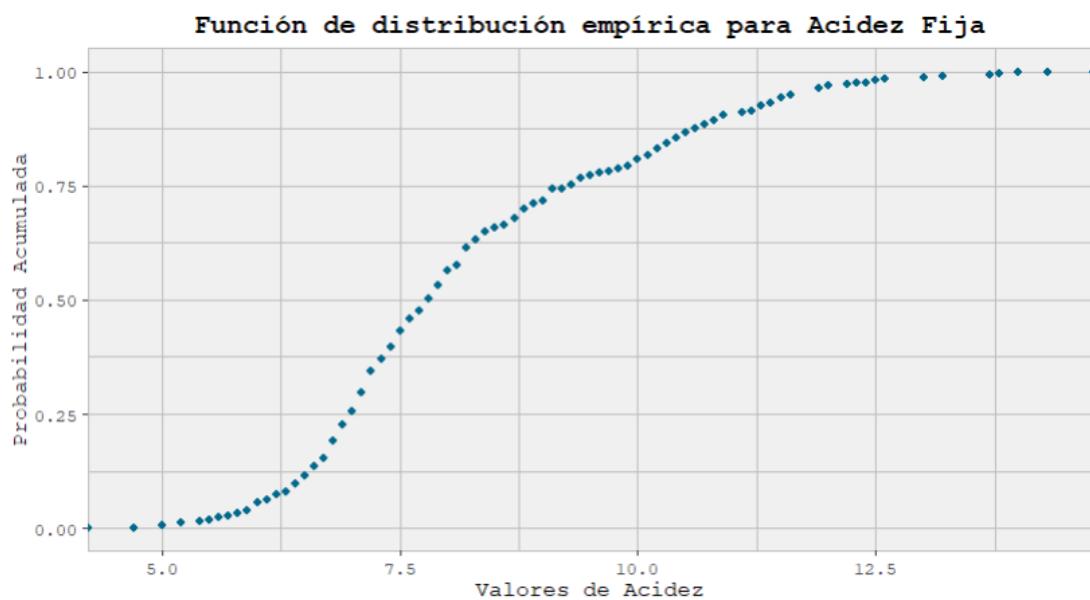


Figura I

Es posible observar en el gráfico que como los puntos son muy próximos unos de otros, esta variable es continua. Por esta razón, se grafica un histograma (véase *Figura II*) para poder distinguir la distribución de la variable F.

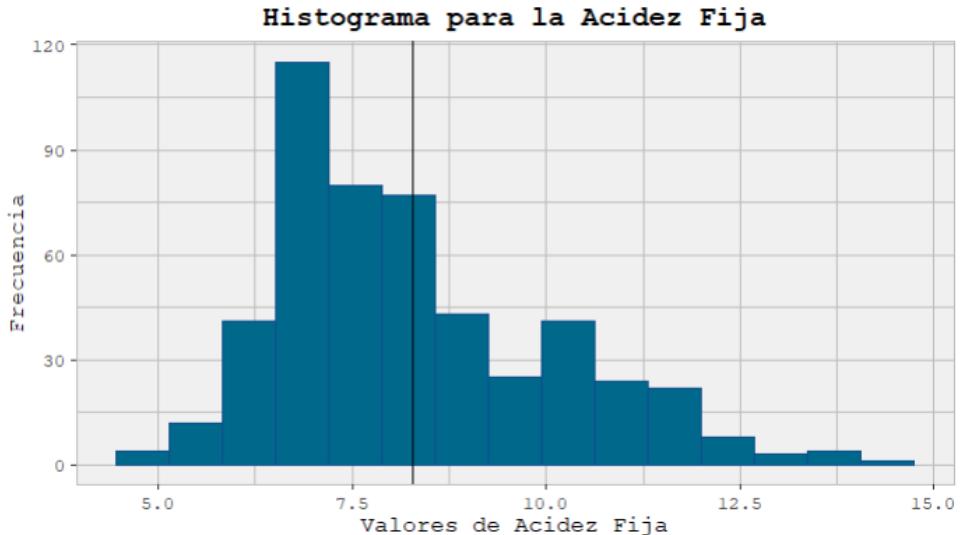


Figura II

A su vez, se grafica la función de densidad (véase *Figura III*), y es posible observar mejor que en el histograma que hay un pico de densidad cercano a $x = 7$ y luego va disminuyendo progresivamente hasta aproximadamente $x = 10.5$, donde hay un nuevo pico y vuelve a bajar.

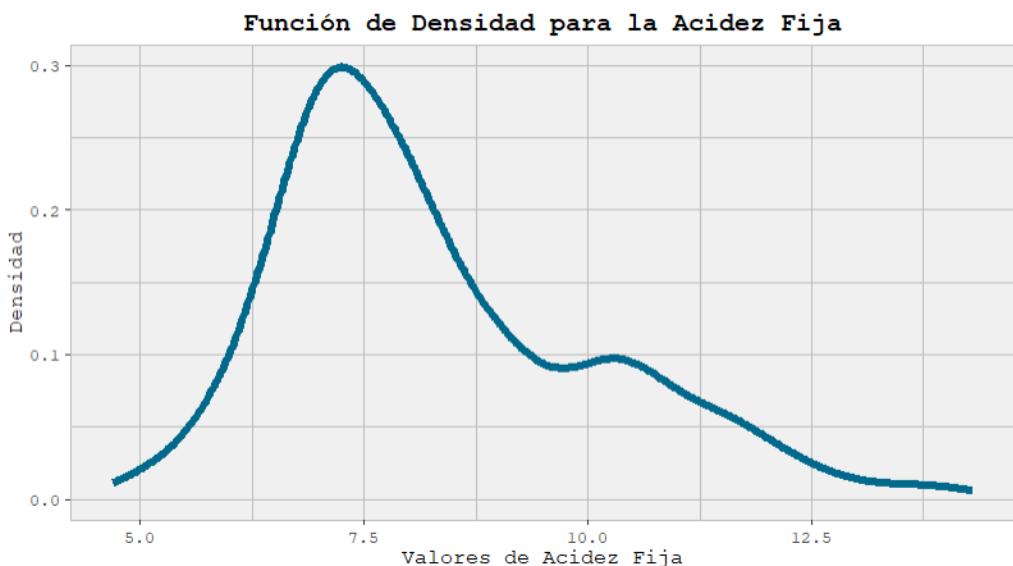


Figura III

Luego, se estima la media muestral de $\bar{f} = 8.27$ y la mediana de $f_{0.5} = 7.9$, de manera que al ser mayor la mediana que la media es posible afirmar que la distribución tiene asimetría positiva -luego respaldado en *Tabla II*. Esto se puede confirmar en el boxplot de la *Figura IV*. En

este gráfico también se pueden observar un par de outliers -marcados en negro en la parte superior-, y los cuantiles son:

Cuantil	Expresión matemática	Resultado numérico
0.25	$f_{0.25} = F^{-1}(0.25) = \min f_i \text{ tq}\{f_i: \hat{F}(f_i) \geq 0.25\}$	7.0
0.5	$f_{0.5} = F^{-1}(0.5) = \min f_i \text{ tq}\{f_i: \hat{F}(f_i) \geq 0.5\}$	7.9
0.75	$f_{0.75} = F^{-1}(0.75) = \min f_i \text{ tq}\{f_i: \hat{F}(f_i) \geq 0.75\}$	9.3

Tabla I

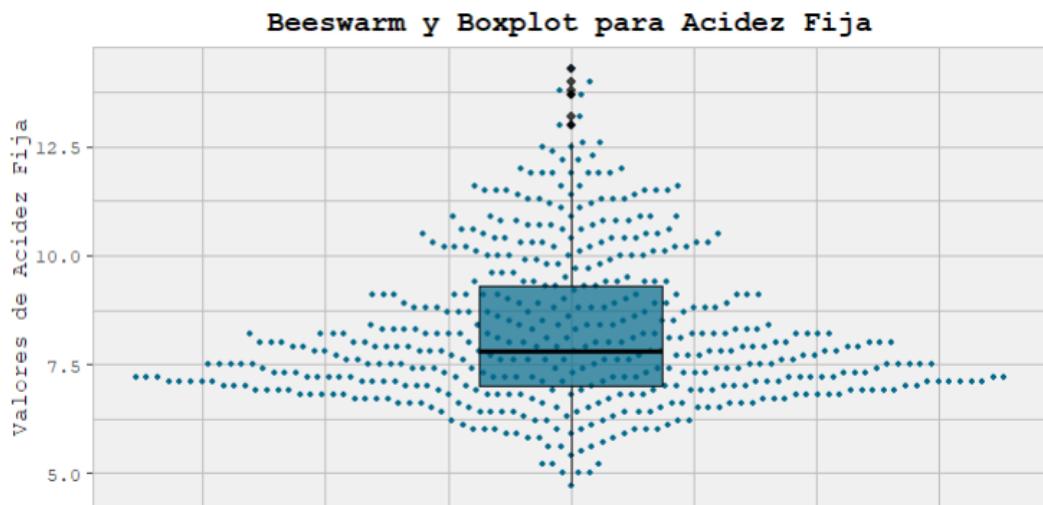


Figura IV

Cómo se puede ver en la *Figura IV*, hay una presencia de valores atípicos (los puntos de color negro, que muestran la cantidad de outliers) en la cantidad de acidez del vino pero se optó no excluirlos ya que no son significativos ya que no alteran significativamente a las medidas no robustas con respecto a las robustas. Asimismo, se ha realizado un gráfico para visualizar si hay cambios en la distribución si se eliminan los outliers y el resultado fue que las curvas de las distribuciones se superponen, por lo que consideramos que no es un gráfico que aporte tanto valor al informe y nos confirma que no es necesario eliminar los valores atípicos.

Para finalizar el análisis de esta variable, se ha confeccionado la siguiente *Tabla II* para visualizar un breve resumen estadístico de la variable de acidez del vino. De todos los valores, no resulta de principal interés destacar que el coeficiente de curtosis es mayor a 3, eso quiere decir que la distribución de la variable es de colas pesadas.

Resumen estadístico para la <i>fixed.acidity</i>		
Media muestral	$\bar{f} = \frac{\sum_{i=1}^n F_i}{N}$	8.27
Desvío estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (F_i - \bar{F})^2}{n}$	1.79
Coeficiente de variación	$r = \frac{\bar{V}(f)}{ f }$	0.21
Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\sum (\frac{f_j - \bar{f}}{\sqrt{V(f)}})^3}{n}$	0.81
Coeficiente de curtosis muestral	$\hat{k} = \frac{\sum (\frac{f_j - \bar{f}}{\sqrt{V(f)}})^4}{n}$	3.20

Tabla II

2.2.2. *density*

D = Densidad de un vino *Vinho Verde* (gr/ml)

La variable *D* mide la densidad del vino tinto en gramos por mililitro. El vino que registró menor densidad fue de 0.99 y el de mayor valor registrado de 1.003, presentando un rango total de 0.013. A continuación, en la *Figura V*, se observa su gráfico de la función de distribución empírica.

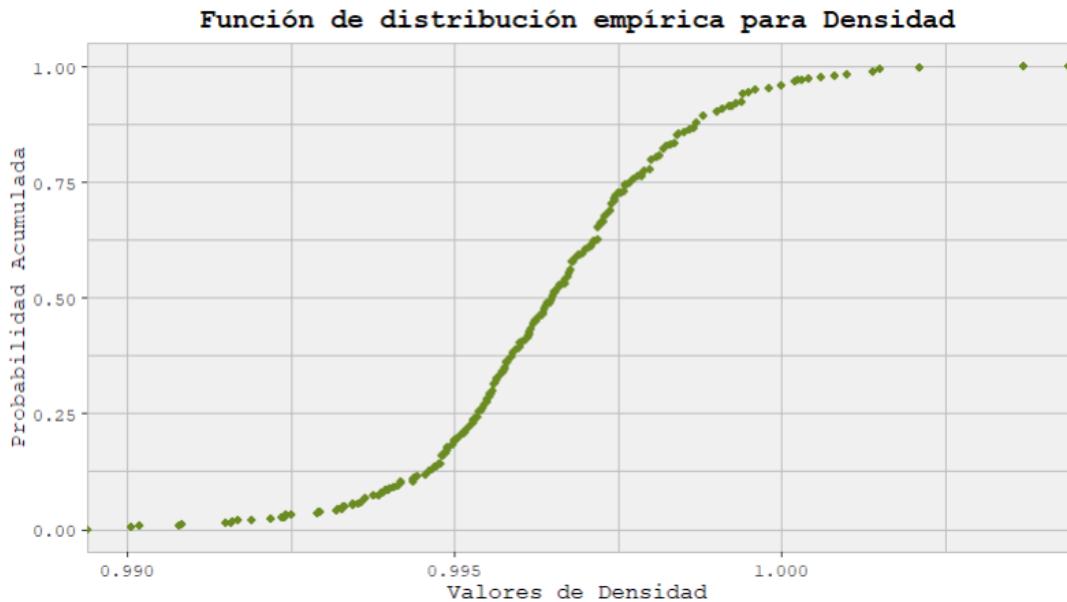


Figura V

Se podría decir que la variable D se comporta de manera aleatoria continua. También es posible visualizar que la gran mayoría de los puntos se encuentran concentrados en el centro de la distribución, cercano a la media. Procedemos a graficar su histograma y la función de densidad correspondiente (véase Figura VI y Figura VII) para ver mejor cuál es la forma de la distribución.

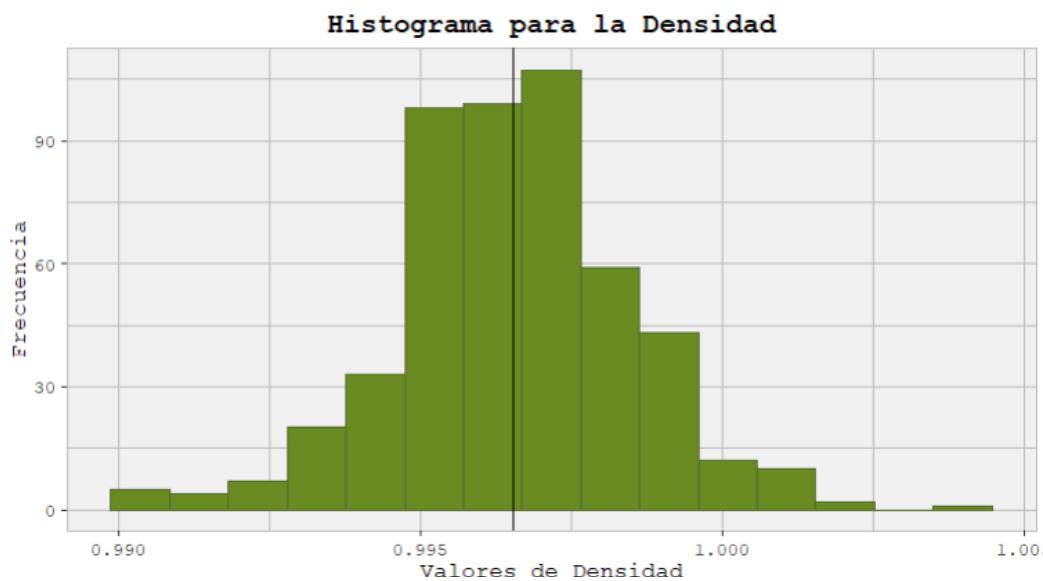


Figura VI

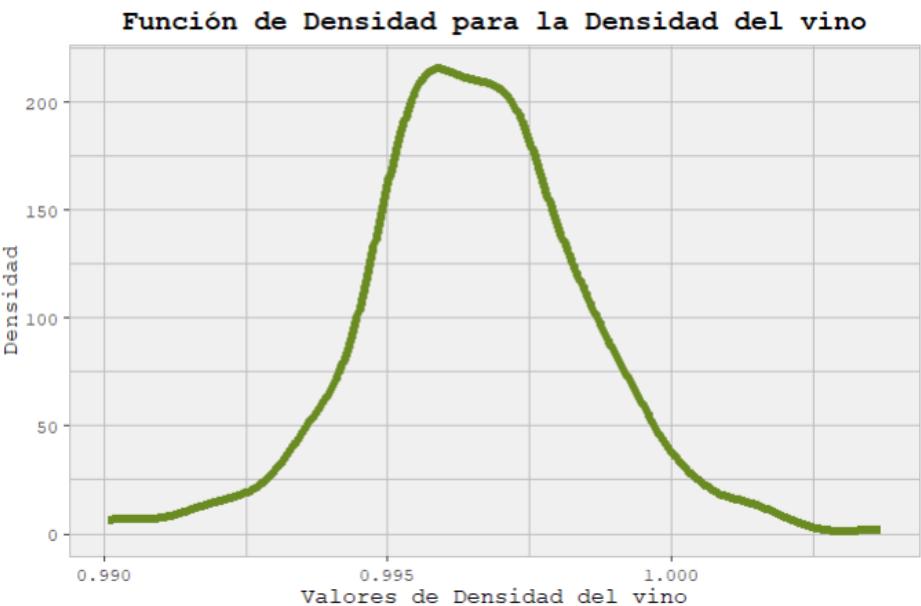


Figura VII

La *Figura VII* muestra que la curva se parece a una campana de Gauss (con algunas irregularidades). Luego, con su posterior cálculo identificamos que la media y la mediana están representadas por el mismo valor ($\bar{d} = d_{0.5} = 0.9$), lo que nos muestra que la distribución es simétrica y posee normalidad. Esto se puede confirmar en el boxplot de la *Figura VIII*. En este gráfico también se pueden observar algunos outliers de ambos extremos y la delimitación de los cuantiles (representados también en *Tabla III*).

Cuantil	Expresión matemática	Resultado numérico
0.25	$d_{0.25} = F^{-1}(0.25) = \min d_i \text{ tq}\{d_i; \hat{F}(d_i) \geq 0.25\}$	0.99
0.5	$d_{0.5} = F^{-1}(0.5) = \min d_i \text{ tq}\{d_i; \hat{F}(d_i) \geq 0.5\}$	0.99
0.75	$d_{0.75} = F^{-1}(0.75) = \min d_i \text{ tq}\{d_i; \hat{F}(d_i) \geq 0.75\}$	0.99

Tabla III

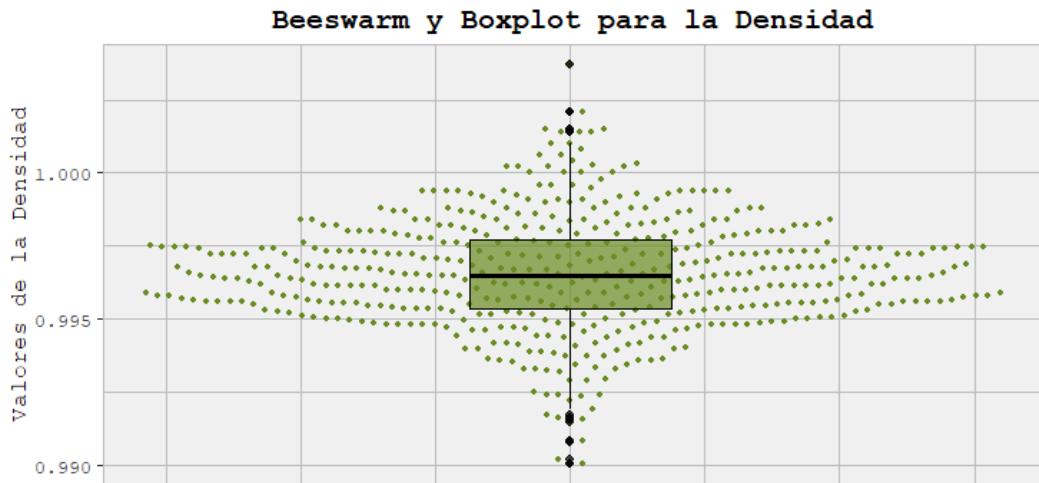


Figura VIII

Como se ha mencionado previamente, en la parte superior e inferior del gráfico es observable la presencia de outliers -marcados en color negro-. Hemos tomado la decisión de no separar tales valores de la muestra al no resguardar una significativa distancia con los límites del rango intercuartílico. De todas formas, a modo de evaluar el impacto de la inclusión de estos datos y su posible eliminación, en la *Figura IX*, se puede observar cómo se altera su distribución en ambos escenarios. De esta forma, es posible afirmar que no es sustancial su diferenciación y no altera significativamente la manera en la que esta se distribuye.

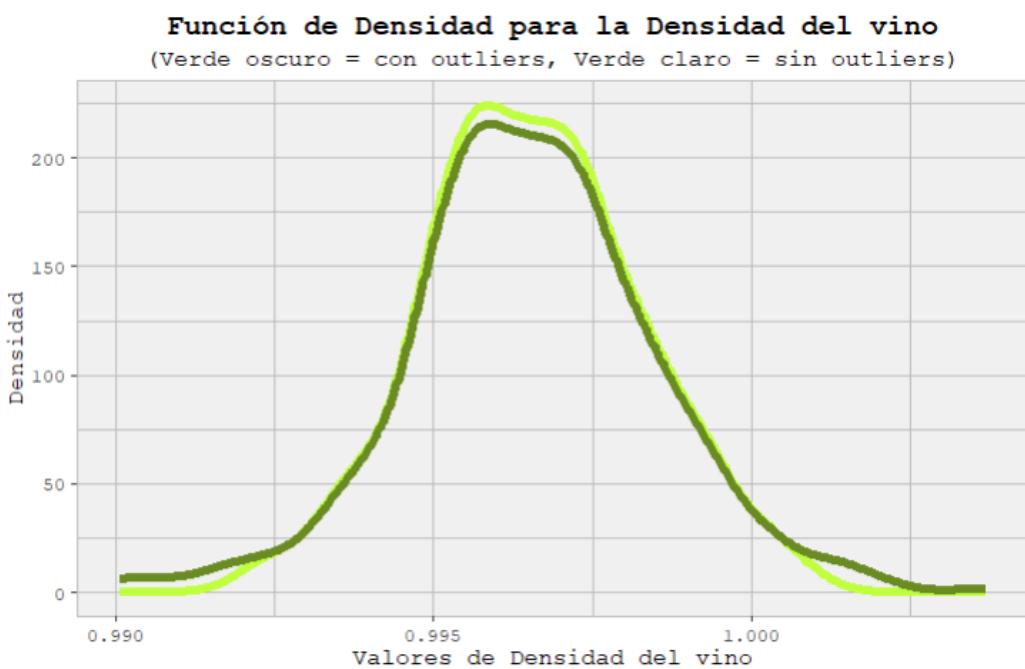


Figura IX

Para finalizar el análisis de esta variable, confeccionamos la *Tabla IV* a fines de visualizar un breve resumen estadístico de la variable de la densidad del vino. De esta tabla se puede destacar, nuevamente, que el coeficiente de curtosis nos dió mayor a cero, lo cual nos confirma que esta variable aleatoria tiene una distribución de colas pesadas como habíamos supuesto anteriormente y presenta una asimetría levemente negativa.

Resumen estadístico para density		
Media muestral	$\bar{d} = \frac{\sum_{i=1}^n D_i}{N}$	0.99
Desvió estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n}$	0.001
Coeficiente de variación	$r = \frac{\bar{V}(d)}{ \bar{d} }$	0.001
Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\sum (\frac{d_j - \bar{d}}{\sqrt{\bar{V}(d)}})^3}{n}$	-0.03
Coeficiente de curtosis muestral	$\hat{k} = \frac{\sum (\frac{d_j - \bar{d}}{\sqrt{\bar{V}(d)}})^4}{n}$	5.85

Tabla IV

2.2.3. alcohol

A = Porcentaje de alcohol de un vino Vinho Verde.

La variable alcohol mide la graduación alcohólica del vino. La misma se expresa en grados y mide el contenido de alcohol absoluto en, es decir, el porcentaje de alcohol que esta posee. A modo de representación, obsérvese en la *Figura X* su función empírica. Al ver cómo se comportan los puntos, se podría decir que A es una variable aleatoria continua.

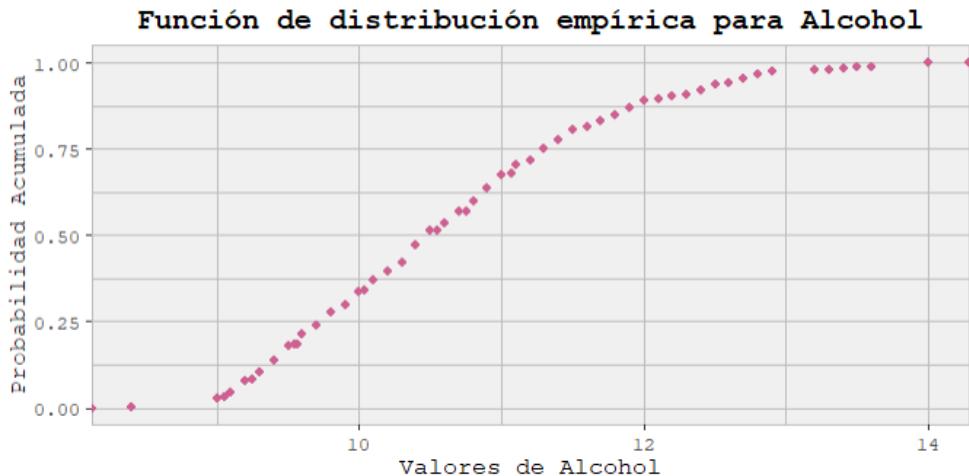


Figura X

Los valores registrados de tal variable son, como mínimo 8.4 y como máximo 14, presentando un rango de 5.6. Véase *Figura X* para observar en un histograma la manera en la que los datos se distribuyen.

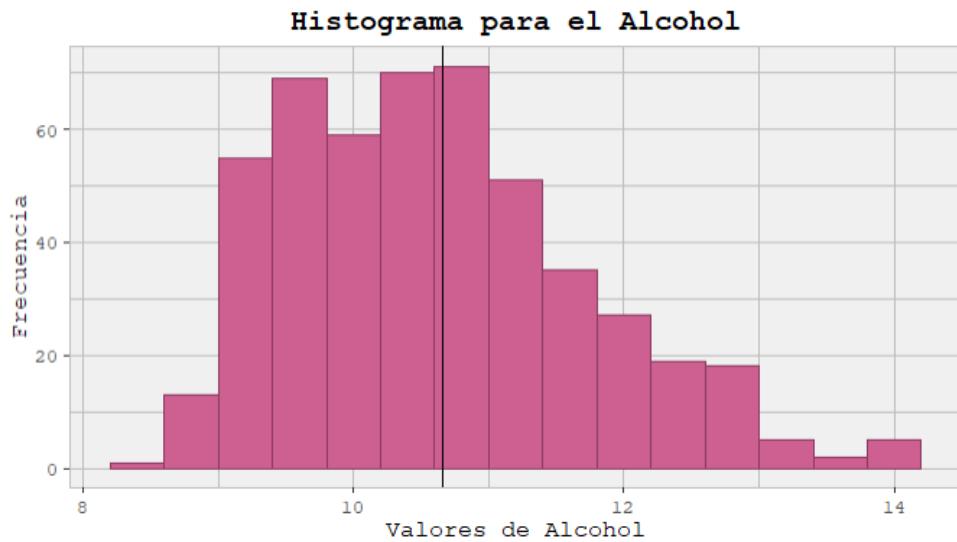


Figura XI

Es posible observar una amplia concentración de los datos cerca de la media (marcada con una línea vertical negra), prácticamente no teniendo agrupación o un significativo peso en sus colas, especialmente a la izquierda. Esto quiere decir que la distribución tiene asimetría positiva. A su vez, obsérvese en la *Figura XII* el diagrama de densidad de dicha variable. Nótese en aquel, un predominante aumento en la densidad,

a medida que se acerca a $x=9.5$ y una progresiva baja a partir de $x=10.5$ hasta $x=13$.

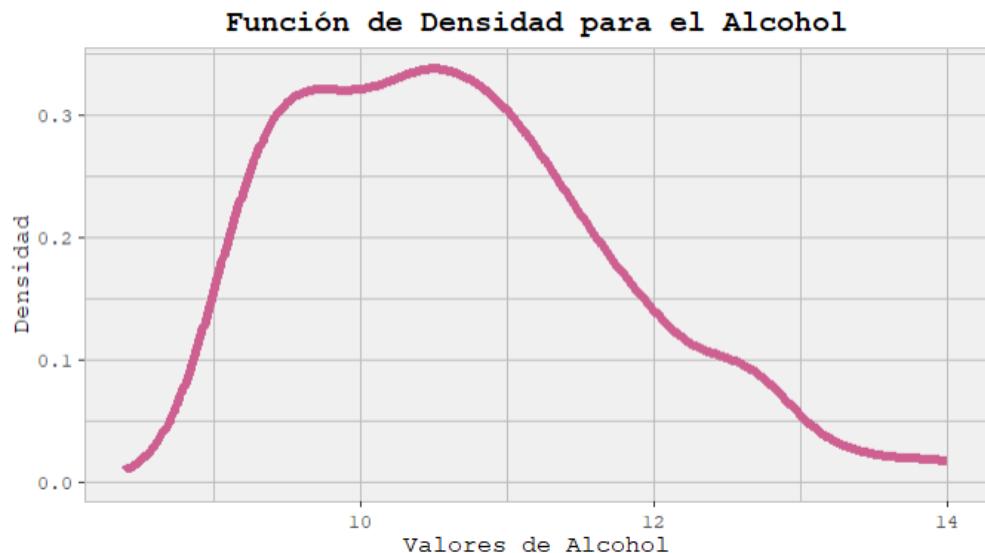


Figura XII

Continuamos luego con el cálculo de la media y la mediana. Ambas otorgan un valor cercano entre sí, siendo que la media es de 10.65 mientras que la mediana es de 10.5. Para mayor información de cuartiles, obsérvese la *Tabla V*, la cual muestra los valores numéricos para cuantiles 0.25, 0.5 y 0.75.

Cuantil	Expresión matemática	Resultado numérico
0.25	$a_{0.25} = F^{-1}(0.25) = \min a_i \text{ tq} \{a_i: \hat{F}(a_i) \geq 0.25\}$	9.80
0.5	$a_{0.5} = F^{-1}(0.5) = \min a_i \text{ tq} \{a_i: \hat{F}(a_i) \geq 0.5\}$	10.50
0.75	$a_{0.75} = F^{-1}(0.75) = \min a_i \text{ tq} \{a_i: \hat{F}(a_i) \geq 0.75\}$	11.32

Tabla V

A continuación se puede visualizar el diagrama de caja y bigotes (véase *Figura XIII*) con un gráfico Bee Swarm. Es observable en tal caso que no existe una predominancia de datos fuera del rango intercuartílico, ubicándose sólo 4 valores atípicos que tienen el mismo valor fuera del rango superior estipulado. Al no ser una distancia muy predominante, no nos resultó de crucial importancia estudiarlo al no alterar

significativamente las medidas no robustas como la media. Por esa razón es que también hemos decidido no excluirlos de la muestra. Asimismo, hemos realizado un gráfico para visualizar si hay cambios en la distribución con la eliminación de outliers y el resultado fue que las curvas se superponen, por lo que consideramos que no es un gráfico que aporte tanto valor al informe y nos confirma que no es necesario eliminar los valores atípicos.

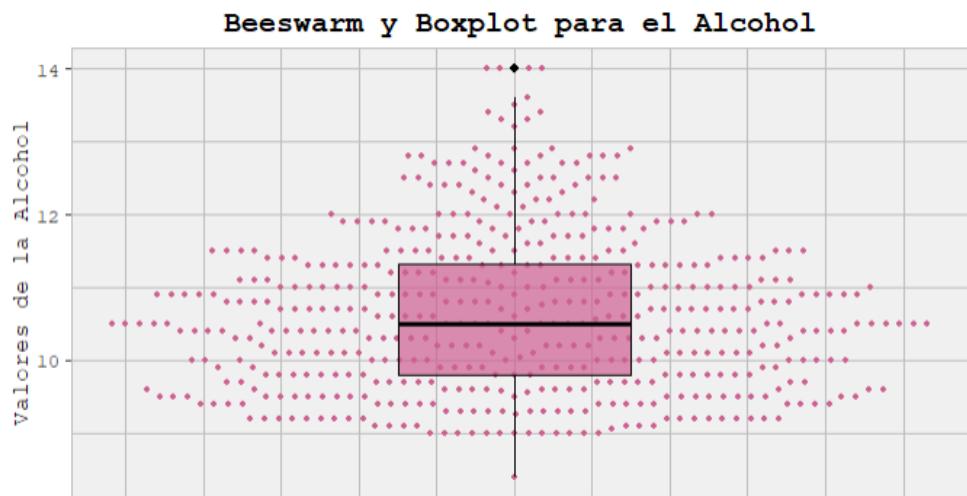


Figura XIII

Finalmente, a continuación en la *Tabla VI* se pueden observar junto a sus expresiones, los valores numéricos para la media muestral, el desvío estándar muestral, el coeficiente de variación, el coeficiente de asimetría muestral y el coeficiente de curtosis muestral. Podemos verificar que la muestra posee una leve asimetría positiva y una curtosis cercana a 3, lo cual sugiere la normalidad de la muestra.

Resumen estadístico para alcohol		
Media muestral	$\bar{a} = \frac{\sum_{i=1}^n A_i}{N}$	10.65
Desvío estándar muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (A_i - \bar{A})^2}{n}$	1.09
Coeficiente de variación	$r = \frac{\bar{V}(a)}{ \bar{a} }$	0.10

Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\sum(\frac{a_j - \bar{a}}{\sqrt{V(a)}})^3}{n}$	0.62
Coeficiente de curtosis muestral	$\hat{k} = \frac{\sum(\frac{a_j - \bar{a}}{\sqrt{V(a)}})^4}{n}$	2.94

Tabla VI

2.2.4. pH

P = Medida de pH de un vino Vinho Verde.

La variable P describe qué tan *ácido* o *básico* es un vino en una escala del 0 -muy ácido- al 14 -muy básico-, la mayoría de los vinos suelen estar en un rango de 3-4. A modo de representación, obsérvese en la Figura XIV la función empírica. Al ver que la mayoría de los datos se encuentran concentrados en la mitad, y simultáneamente adquieran una forma de “s” suave, suponemos que la forma de la distribución adquirirá la forma de una campana de Gauss. Asimismo, es posible afirmar que la variable P se comporta como una aleatoriedad continua.

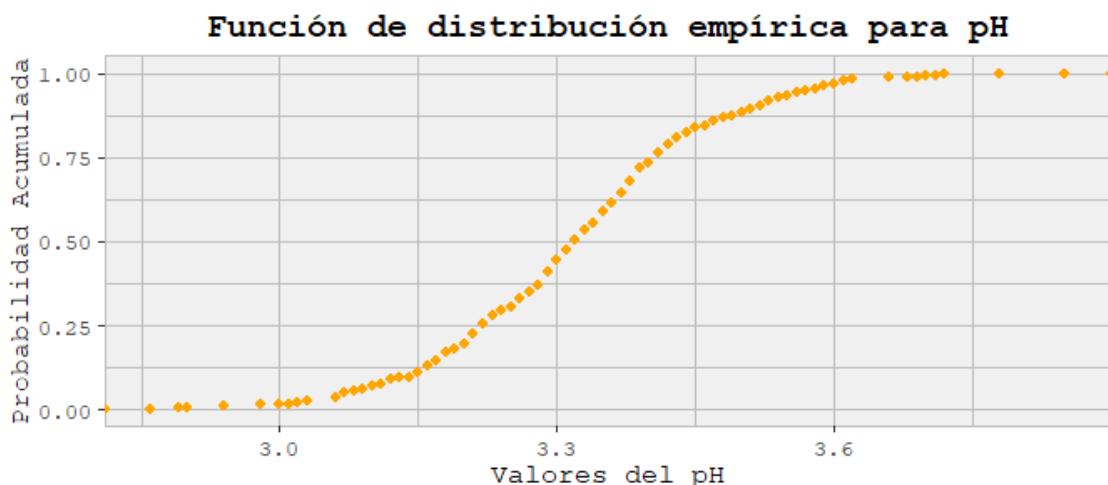


Figura XIV

En la muestra tomada es posible observar un valor mínimo de 2.86 y un máximo de 3.85, presentando un rango total de 0.99. Véase la Figura XV para el análisis de su distribución. En el mismo, se puede confirmar nuestra suposición de que la variable P morfológicamente se asemeja a

grandes rasgos a una campana de Gauss, presentando una gran concentración cerca de la media (marcada con una línea vertical negra) y con las colas similares a ambos lados, equitativamente presentando bajo peso.

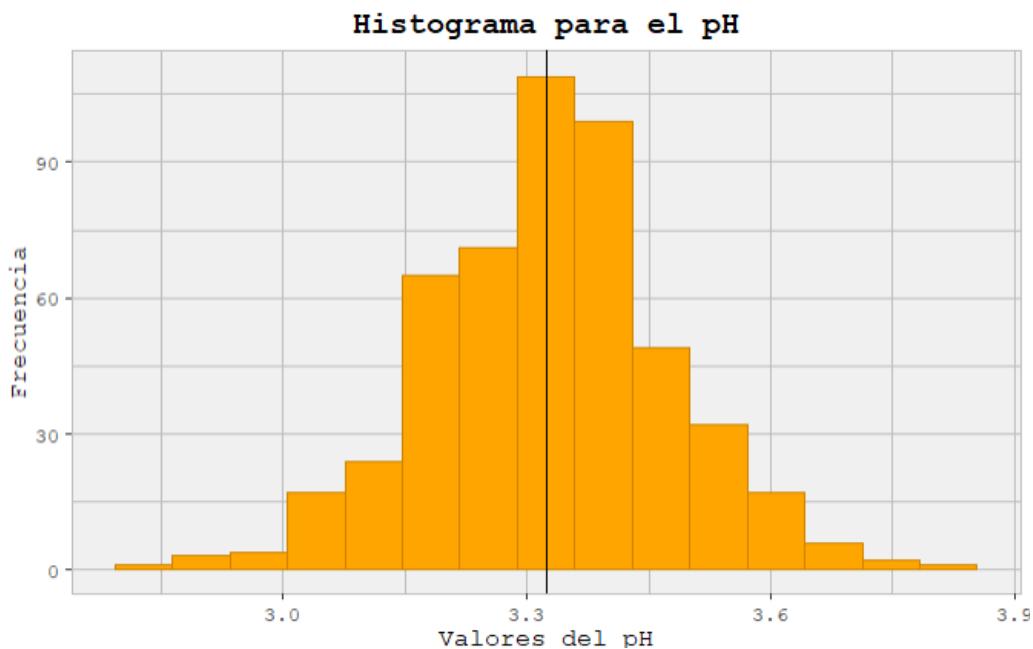


Figura XV

Adicionalmente, al confeccionar la curva de densidad es observable (*Figura XVI*) que la misma coincide con la descripción de normalidad previamente mencionada. Alrededor de la media, existe una notable predominancia de datos.

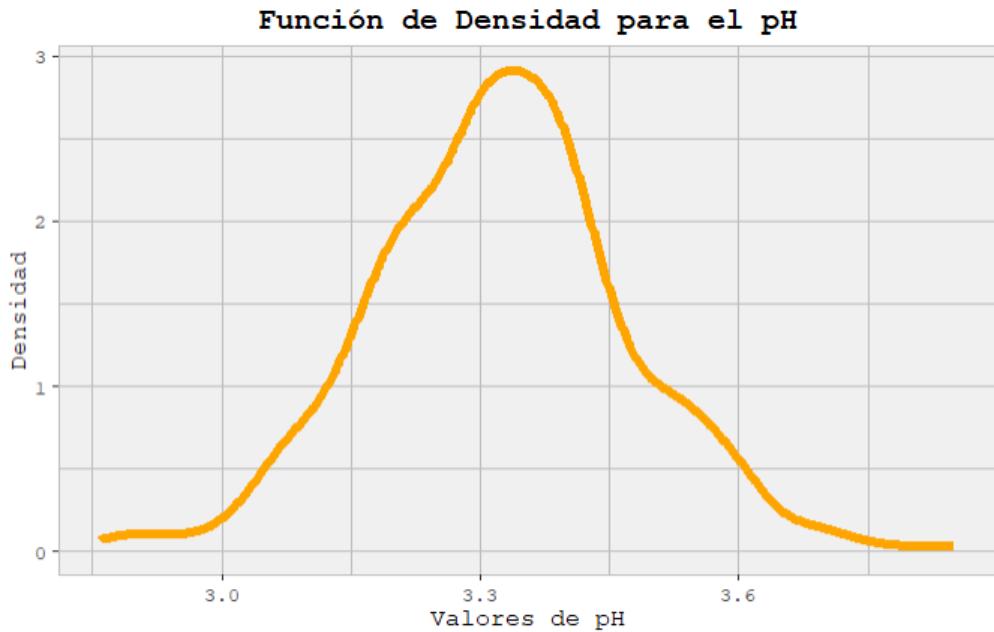


Figura XVI

Asimismo, se puede ver una notable similitud entre la mediana 3.32 y la media 3.3237. Al ser una medida robusta como la mediana tan similar a aquella no robusta, es posible afirmar que hay simetría en la distribución. A continuación en la *Tabla VII* se enlistan los cuantiles 0.25 0.5 y 0.75. También, obsérvese el diagrama de caja y bigotes con un gráfico Bee Swarm (*Figura XVII*).

Cuantil	Expresión matemática	Resultado numérico
0.25	$p_{0.25} = F^{-1}(0.25) = \min p_i \text{ tq} \{p_i: \hat{F}(p_i) \geq 0.25\}$	3.22
0.5	$p_{0.5} = F^{-1}(0.5) = \min p_i \text{ tq} \{p_i: \hat{F}(p_i) \geq 0.5\}$	3.32
0.75	$p_{0.75} = F^{-1}(0.75) = \min p_i \text{ tq} \{p_i: \hat{F}(p_i) \geq 0.75\}$	3.41

Tabla VII

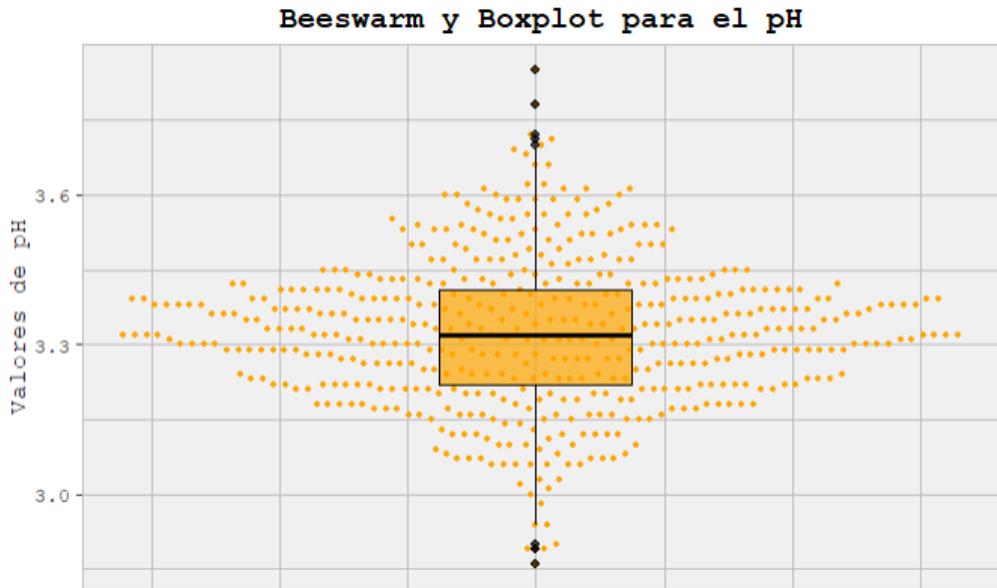


Figura XVII

Es observable en el boxplot una mayor presencia de outliers en relación con otras variables. Tanto en valores mínimos y máximos, se identifican valores atípicos. Sin embargo, en su gran mayoría ninguno de ellos se aleja alarmantemente del rango intercuartílico, por esa razón se ha decidido no excluirlos de la muestra. Cuando se realizó el gráfico de las distribuciones con y sin los valores atípicos las curvas se superponen, confirmando nos que no es necesario eliminar los outliers ya que no cambia en nada la distribución.

Para finalizar, en la Tabla XVIII se puede apreciar enlistados la media muestral, el desvío estándar muestral, el coeficiente de variación, el coeficiente de asimetría muestral y el coeficiente de curtosis muestral. En ella, se ha identificado un coeficiente de asimetría positiva muy cercano al cero, lo que hace la muestra casi simétrica. A su vez, su curtosis es mayor a 3, lo cual nos confirma nuestra hipótesis de que la distribución es de colas pesadas.

Resumen estadístico para pH		
Media muestral	$\bar{p} = \frac{\sum_{i=1}^n p_i}{N}$	3.32

Desvió estández muestral	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (P_i - \bar{P})^2}{n}$	0.14
Coeficiente de variación	$r = \frac{\bar{V}(p)}{ \bar{p} }$	0.04
Coeficiente de asimetría muestral	$\hat{\gamma} = \frac{\sum (\frac{p_j - \bar{p}}{\sqrt{V(p)}})^3}{n}$	0.05
Coeficiente de curtosis muestral	$\hat{k} = \frac{\sum (\frac{p_j - \bar{p}}{\sqrt{V(p)}})^4}{n}$	3.44

Tabla VIII

2.3. Relaciones entre variables

En esta sección se analizarán las relaciones entre las variables presentadas previamente. A fines de llevar adelante un análisis integral, se ha ejecutado la matriz de correlación (*Tabla IX*), la cual detalla numéricamente la dependencia lineal entre dos variables. El determinante de la matriz es de 0.12888 por lo que, sabiendo que a mayor semejanza con el 0 mayor correlación, las variables guardan cierta dependencia lineal.

	fixed.acidity	density	alcohol	pH
fixed.acidity	1.00	0.68	-0.95	-0.70
density	0.68	1.00	-0.51	-0.34
alcohol	-0.09	-0.51	1.00	0.17
pH	-0.70	-0.34	0.17	1.00

Tabla IX

Para poder facilitar el análisis de las correlaciones, se decidió ilustrar el siguiente gráfico (véase Figura XVIII), que visualizan los valores dados en la *Tabla IX*. El color rojo indica una fuerte correlación positiva entre dos variables mientras que un color azulado demuestra una correlación negativa.

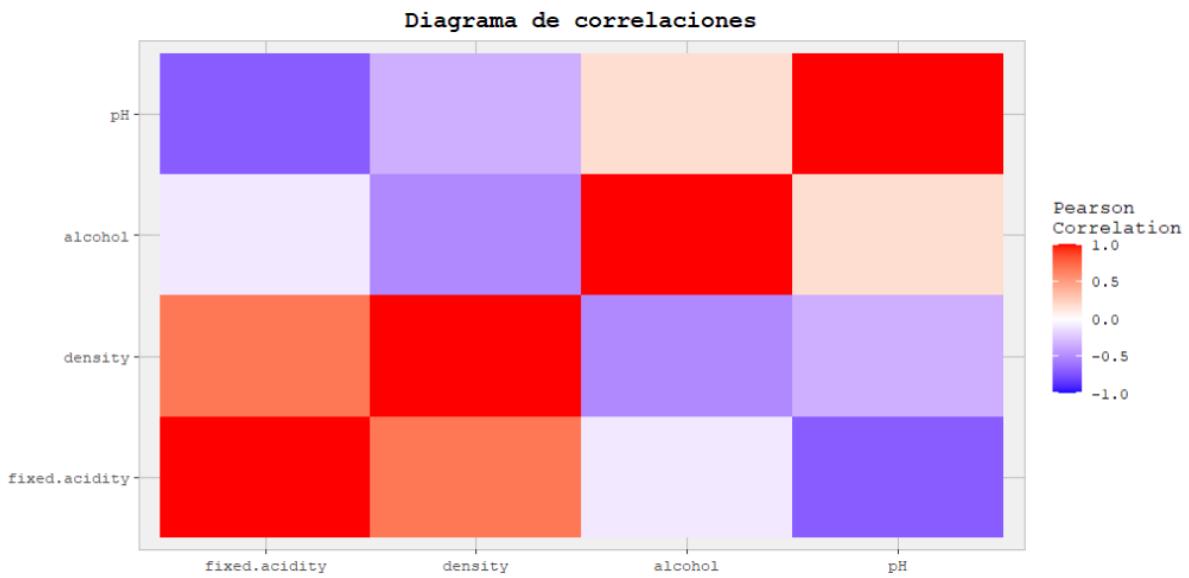


Figura XVIII

Luego, se eligió hacer 4 relaciones entre las variables para ver cómo se ven gráficamente.

En primer lugar, se decidió relacionar la acidez del vino con el pH, lo cual había arrojado una correlación de -0.70, que señala que una estrecha relación entre ambas al estar cerca de -1. Tal información es verificada visualmente mediante un scatter plot (*Figura XIX*), donde se puede ver una tendencia negativa posicionando a la acidez fija en el eje horizontal y el ph en el eje vertical. En consecuencia, se concluyó que a mayor acidez fija menor es el pH o, alternativamente, a mayor pH tiende a ser menor la acidez fija.

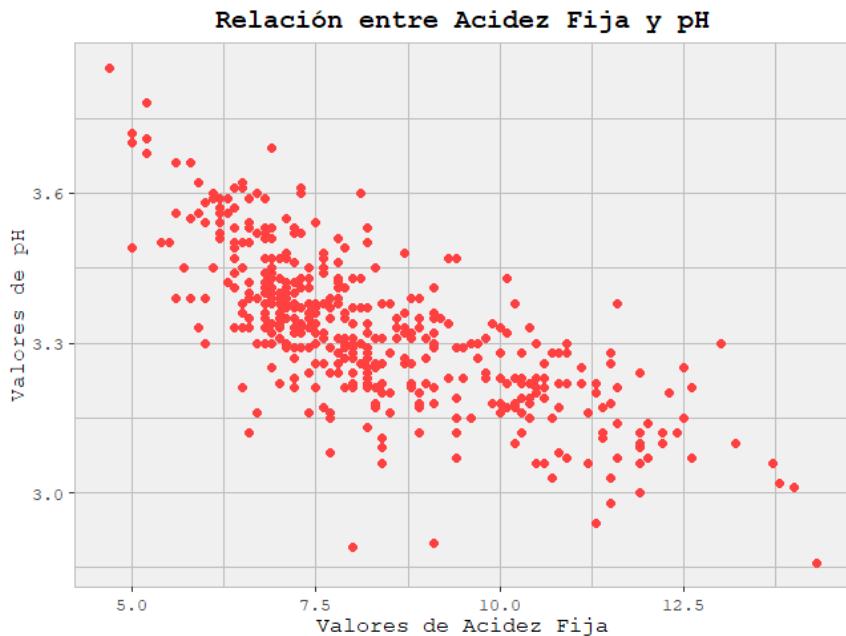


Figura XIX

Luego, se continuó con el mismo procedimiento para analizar la relación entre las variables de densidad de alcohol. Al igual que en la *Figura XIX*, en este nuevo gráfico (*Figura XX*) se puede ver una clara tendencia decreciente -ubicando la densidad en el eje horizontal y el alcohol en el eje vertical-. Por lo tanto, se infiere que a mayor densidad, menor es el nivel de alcohol y viceversa. En este caso, se puede observar una menor concordancia o menos delimitada esta tendencia, lo cual es correspondido al presentar una concordancia más cercana al 0 que la variable relación anterior, siendo de -0.51.

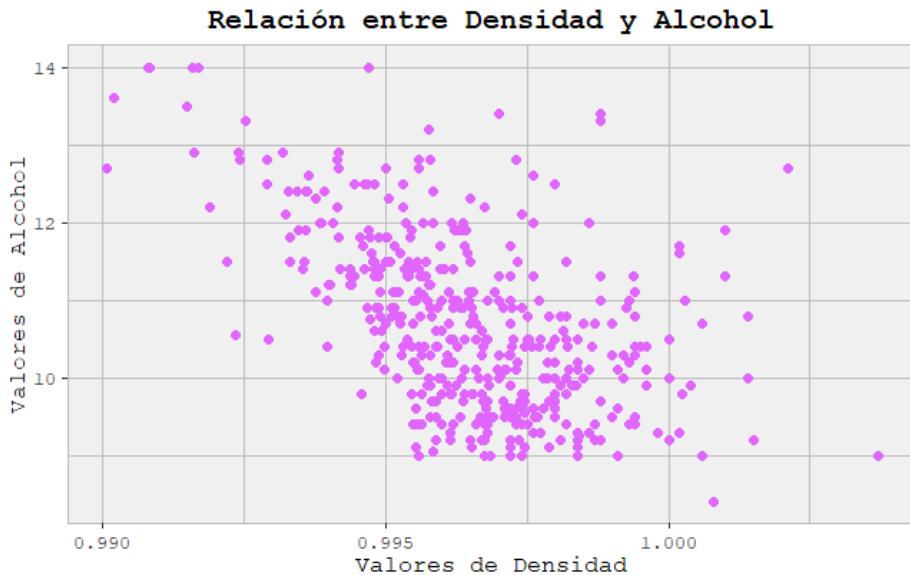


Figura XX

Para el tercer análisis sobre relaciones de variables se decidió observar la conexión entre el alcohol y el pH (*Figura XXI*). Previamente en la *Tabla IX* se arroja como resultado numérico una correlación de 0.17, notablemente cercano al 0. Si bien podría marcarse una tendencia ascendente (a mayor alcohol, mayor pH y viceversa), esto no es lo suficientemente diferenciable como para tomarlo como una afirmación válida.

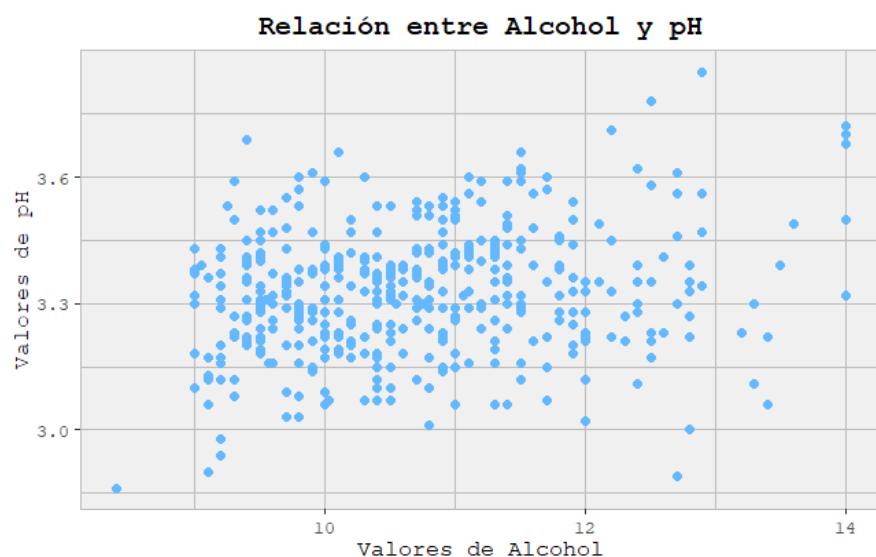


Figura XXI

Para finalizar, se eligió evaluar la relación entre la densidad y el pH (*Figura XXII*). Previamente en la *Tabla IV*, el resultado numérico nos muestra que su correlación es de -0.34, notablemente cercano al 0. Es sensato entonces, que al observar su visualización, las mismas presentan una leve tendencia negativa (a mayor densidad, menor pH y viceversa), pero es la predominancia de dispersión lo que imposibilita determinarlo.

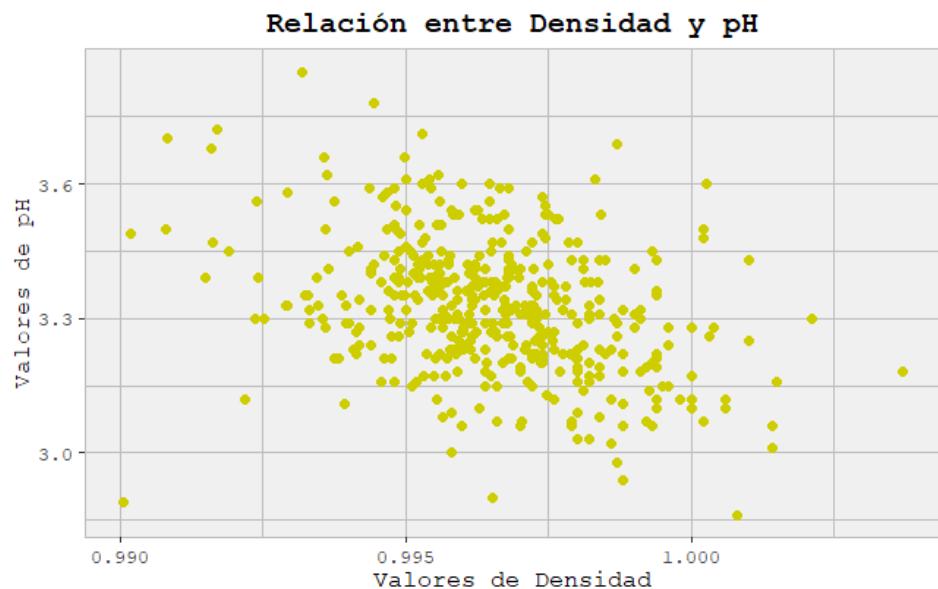


Figura XXII

Por último, a fines de poder visualizar las diferencias entre los cuatro diagramas se ha diseñado la Figura XXIII, la cual respeta los esquemas de colores a fin de poder ser correctamente identificados.

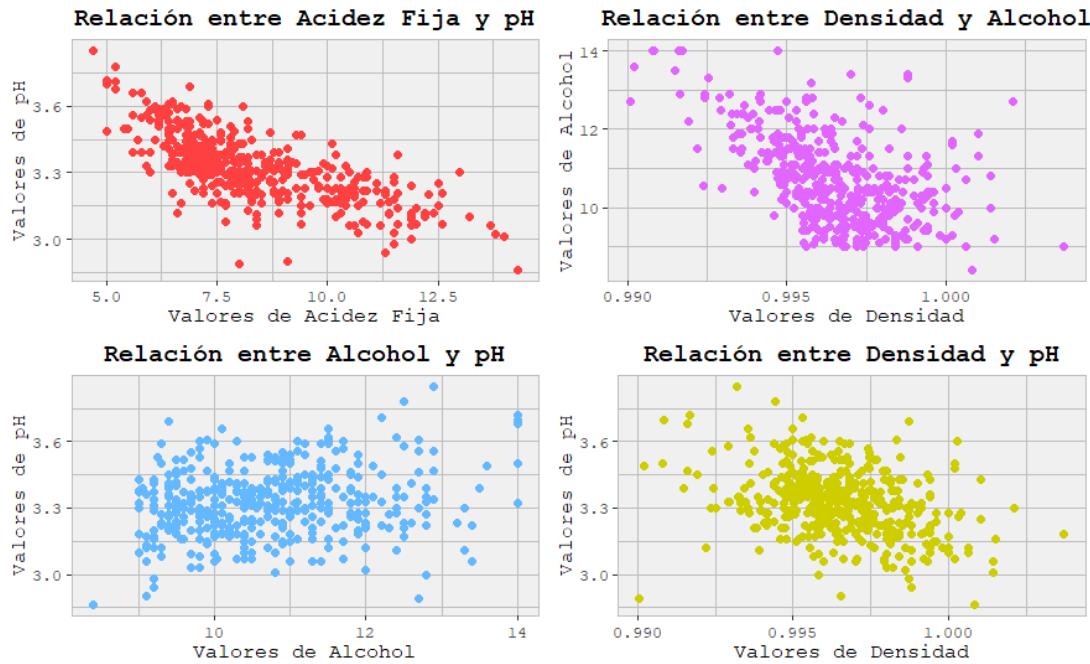


Figura XXIII

2.4. Estudio de una variable a elección

Para esta sección se eligió la variable numérica del pH y se hizo un análisis gráfico de cómo cambia la distribución de los datos en los distintos niveles de la variable categórica *quality*. Como se ha mencionado anteriormente, la calidad en el dataset seleccionado está representada por un valor numérico. Dentro de ellos se decidió seleccionar aquellos de calidad 4 representando una calidad *baja*, los de calidad 6 para un nivel *medio* y los de calidad 8 para una categorización *alta*.

Primeramente, en la *Figura XXIV*, se puede observar el histograma de la variable P filtrado únicamente para la calidad baja. Se puede apreciar que la misma se distribuye con ciertas irregularidades, principalmente debido a que está orientado a la izquierda toma una forma que se asemeja a una campana, presenta una gran baja para luego volver a subir.

Tal irregularidad probablemente esté relacionado a que la muestra tomada solo presenta 40 registros para la calidad baja, por lo que no son los suficientes para asegurar innegablemente que la distribución se asemeje a la de una campana.

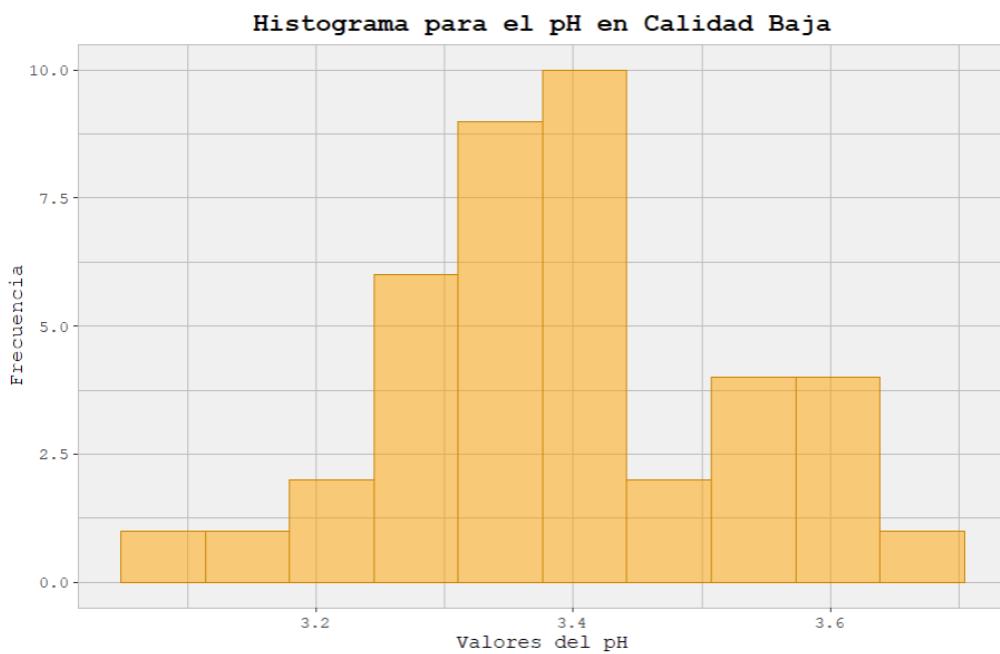


Figura XXIV

Continuando con el análisis descriptivo de la variable P, pasamos a estudiar su calidad media (véase *Figura XXV*). La misma se distribuye ciertamente con cierto rango de normalidad. Posee colas tanto a izquierda como a derecha, siendo la derecha levemente más significativa. De los tres análisis realizados, se infiere que la calidad media es la que adquiere una mayor normalidad, pero a su vez es la única que se alimenta de una numerosa cantidad de registros.

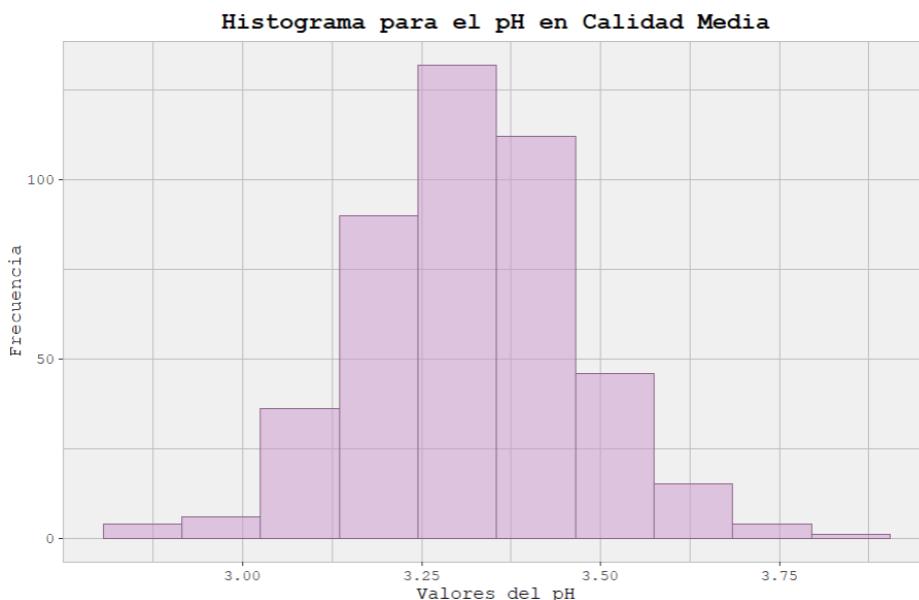


Figura XXV

Se procede luego para el análisis de la variable P para la calidad alta (véase *Figura XXVI*). La misma se distribuye de manera totalmente irregular, con repentinos picos y ausencia de registros en determinados rangos. Dicha inestabilidad, imposibilita el intento de señalar una posible distribución, lo cual no es novedad al tan solo tener un total de 14 observaciones.

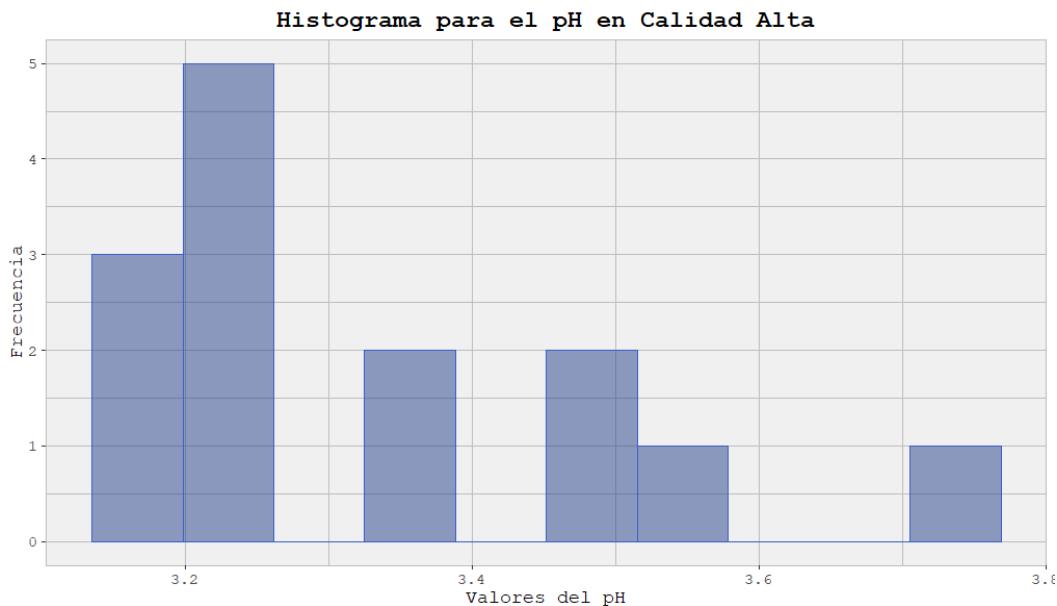


Figura XXVI

Finalmente, decidimos superponer los 3 histogramas mostrados previamente (véase *Figura XXVII*) y se puede observar que existe una gran predominancia de observaciones de calidad media en relación a los de calidad baja y alta. A su vez, es posible afirmar que los menores valores de pH que tienen los vinos de calidad alta son mayores que los valores de las otras calidades.

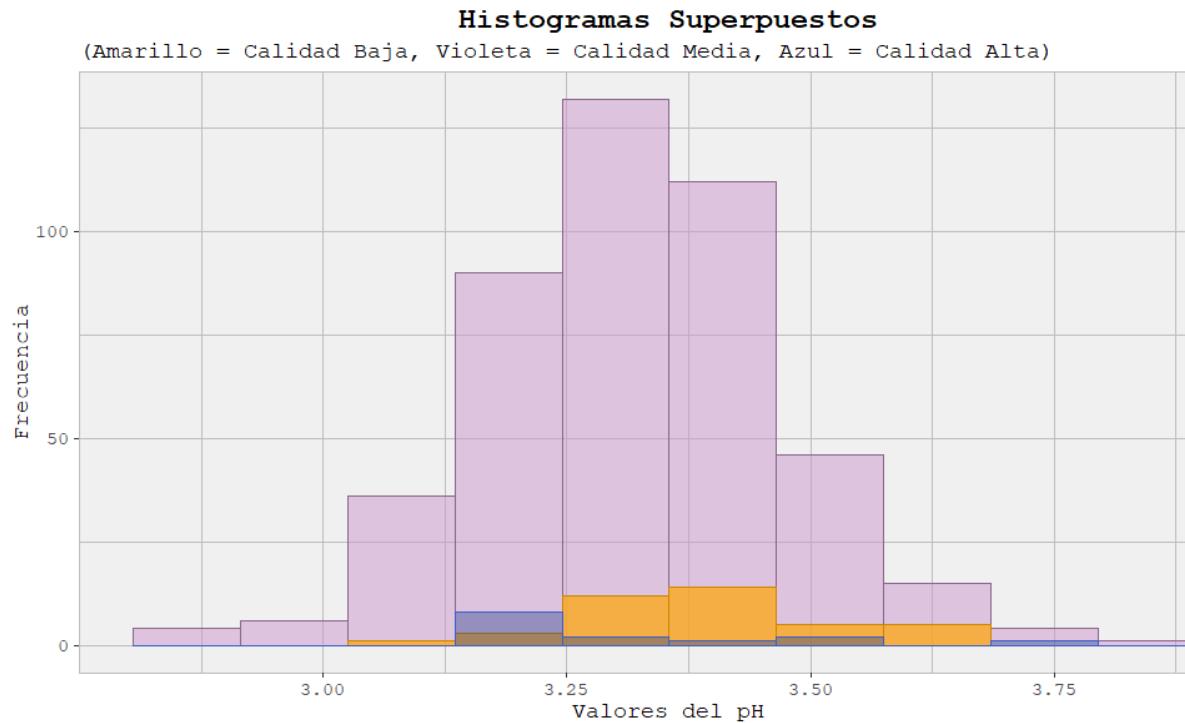


Figura XXVII

Similar a la disposición vista en la *Figura XXVII*, decidimos hacer un gráfico de densidades con el objetivo de nivelar y poder observar en mayor detalle cómo se distribuyen, independientemente de sus frecuencias. De esta manera, obtenemos los resultados dispuestos en la *Figura XXVIII* en donde graficamos la variable ph en su calidad alta (línea azul), media (línea rosa) y baja (línea naranja). Es observable que sus distribuciones no se asemejan necesariamente entre sí, principalmente en cuanto a su asimetría y curtosis.

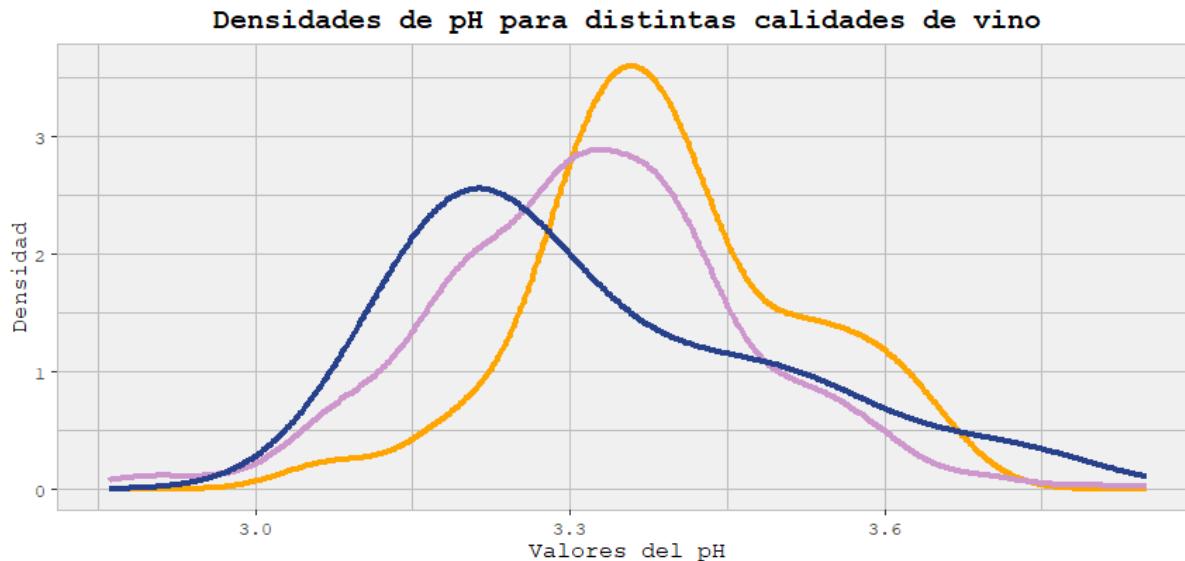


Figura XXVIII

3. Estimación Puntual

Para la siguiente sección se tomaron como valor de referencia únicamente a los valores de la variable P (que indica el pH del vino) para la categoría de calidad media. Esta decisión se debe principalmente a que esta categoría concentra la amplia mayoría de observaciones, con un total de 446 entre un total de 500 -decisión acordada previamente con la cátedra-.

3.1. Elección de variables

Se eligió analizar 2 cantidades de interés q_1 y q_2 ; q_1 siendo el valor del pH que acumula el 0.1 de probabilidad hacia la derecha y q_2 , siendo la proporción poblacional cuyo valor del pH es mayor a 3.4. Se ha elegido este valor ya que es un valor cercano a la media del pH en la calidad media (3.31).

$$q_1 = x_{0.9}$$

$$q_2 = P(X > 3.40)$$

En cuanto a la sensibilidad ante los datos atípicos, se identificó que por un lado, q_1 no es sensible a los datos atípicos. Es decir, si es que se ordenan los datos de la muestra de menor a mayor, la presencia de

datos extremadamente altos o bajos en sus extremos no afectan el cuantil, al tratarse de una estimador robusto.

Por otro lado, en el caso de q_1 , se identificó que la probabilidad es un estimador que depende de manera directa de los parámetros de la distribución. La presencia de datos atípicos tiene un efecto sobre tales parámetros, lo que infiere que la probabilidad es afectada ante la presencia de anomalías, correspondiendo a un estimador no robusto.

3.2. Cálculo de estimadores

3.2.1. $q_1 = x_{0.9}$

Para la cantidad q_1 se calculará \hat{q}_{11} y \hat{q}_{12} , lo que implica el cálculo de en cuantil $x_{0.9}$ de forma paramétrica -tomando distribución normal- y de forma no paramétrica. Véase la *Tabla X* para analizar las expresiones de tales estimaciones. Luego, en la *Tabla XI* se deja su resultado matemático, junto con el resultado de su error.

Forma paramétrica	Forma no paramétrica
$P(Z > r) = 1 - \phi(r) = 0.9$	$X_{0.9} = F^{-1}(0.9) = \min \widehat{X}_i \text{ tq } \{X_i: \widehat{F}(X_i) \geq 0.9\}$

Tabla X

Estimador	Distribución	Resultado	Error Estándar
\hat{q}_{11}	No Paramétrico	3.51	0.0163
\hat{q}_{12}	Paramétrico Normal	3.50	0.0167

Tabla XI

Analizando cercanamente el resultado numérico del error estándar para ambas estimaciones -véase *Tabla XI*- podemos observar que para \hat{q}_{11} el resultado del error es menor, denotando que la estimación paramétrica normal posee un menor desvío, sobre lo cual se puede inferir que se acerca mayormente al valor poblacional que buscamos estimar.

3.2.2. $q_z = P(x > 3.40)$

Para la cantidad q_z , se calcularon dos estimadores \hat{q}_{21} y \hat{q}_{22} para luego compararlos. En la siguiente *Tabla XII* se observan las fórmulas paramétricas que usaremos para calcular esta segunda cantidad de interés.

Forma paramétrica Normal	Forma paramétrica Log-Normal
$P(Z > z) = 1 - \phi(\sigma, \mu)(z) = 0.2903$	$P(Z > z) = 1 - \phi(m, D)(z) = 0.2903$

Tabla XII

Se optó por calcular los estimadores asumiendo la distribución Normal por métodos de los momentos y suponiendo la distribución Log-Normal usando el método de máxima verosimilitud. La razón por la cual elegimos estas dos distribuciones es porque viendo la *Figura XXVI*, se pudo observar que la curva color lila se puede llegar a asemejar a una campana de Gauss con unas leves distorsiones en la curva. Por ese motivo, probamos el primer estimador asumiendo una distribución normal. Asimismo, se observa una ligera asimetría positiva en la curva, donde la mayoría de los valores se concentran en el lado izquierdo de la curva, y por ello planteamos obtener el segundo estimador utilizando la distribución Log-Normal.

Los resultados estimadores puntuales con su error estándar se pueden observar a continuación en la *Tabla XXXI*.

Estimador	Distribución	Método de obtención	Resultado	Error Estándar
\hat{q}_{21}	Paramétrico Normal	Momentos	0.28	0.01638
\hat{q}_{22}	Paramétrico Log-Normal	Máxima verosimilitud	0.58	0.016813

Tabla XIII

Al analizar los resultados de la tabla, resulta sorprendente observar que el estimador obtenido mediante la técnica de máxima verosimilitud con distribución log-normal es casi el doble del estimador obtenido con distribución normal con el método de los momentos. A pesar de esta diferencia, es notable que ambos estimadores presentan un error estándar casi equivalente. A pesar de lo mencionado anteriormente, se

podría afirmar que el estimador obtenido bajo asumir una distribución normal es apenas más preciso que aquel obtenido bajo la suposición de una distribución log-normal.

4. Bondad de Ajuste

En esta sección, se estará comparando el modelo de distribución normal, que se eligió como la más apropiada para describir el conjunto de datos numéricos del pH, con un segundo modelo de distribución Gamma, para observar cual es la que mejor se ajusta a nuestro conjunto de datos.

3.1. Log-verosimilitudes máximas

En este apartado, se obtendrán los valores de las log-verosimilitudes máximas pertenecientes a cada modelo de distribución mencionado anteriormente para evaluar cuál se ajusta mejor al conjunto de datos de la variable P .

	Distribución Normal	Distribución Gamma
Log-verosimilitud máxima	144.6616	245.2669

Tabla XIV

Como se puede observar en la *Tabla XIV*, la distribución gamma tiene una mayor log-verosimilitud máxima que la distribución normal. Es por esto que se puede concluir que el modelo de la distribución gamma es más adecuado para describir la variabilidad de los datos, ya que a mayor valor, mejor es el ajuste.

3.2. Construcción e interpretación de QQ-Plots

En este apartado se empleará el uso de QQ-Plots que permiten observar cuán cerca están las distribuciones normales y gamma del conjunto de datos.

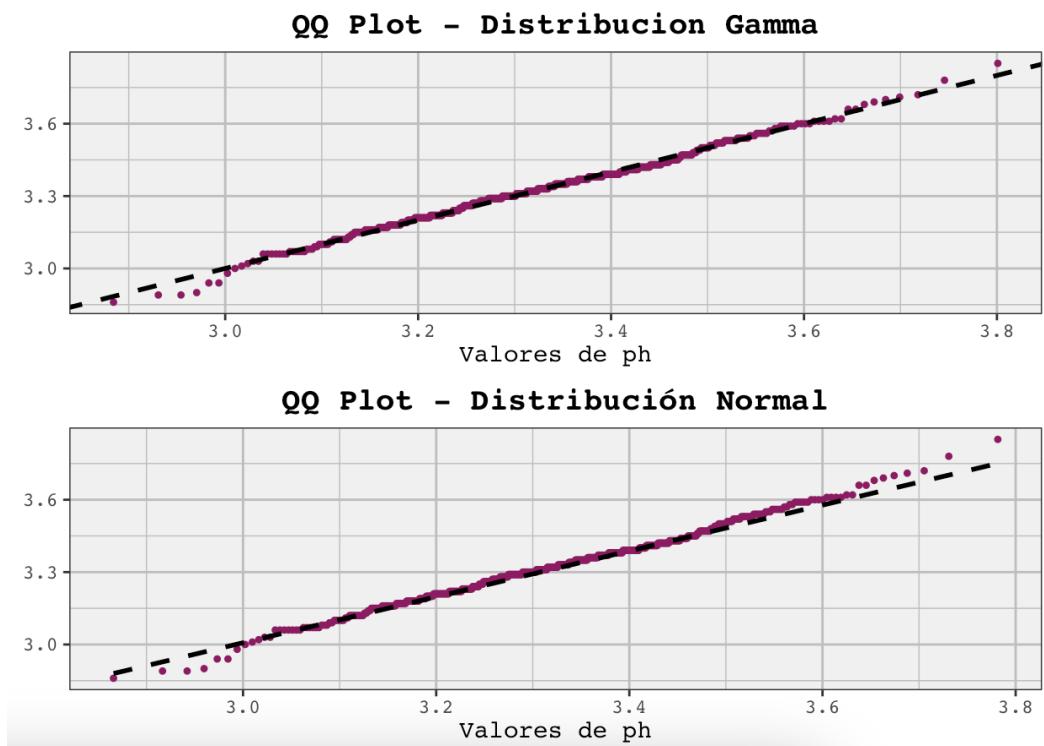


Figura XXIV

Como se puede apreciar en la *Figura XXIV*, se presentan dos gráficos de QQ-plot, en el cual se puede observar que ambas gráficas se ajustan correctamente a la línea punteada. Sin embargo, si bien su ajuste es notablemente acorde a lo óptimo para la debida distribución, podemos observar principalmente para los valores mayores a 3.4 en adelante una leve desviación que luego se hace más notoria. Consecuentemente, esta apreciación puede ser validada con el valor de la log-verosimilitud, siendo que se ha obtenido una cifra mayor para la distribución Gamma para el caso de la variable pH estudiada.

3.3. Distribución empírica ajustada a cada modelo

En este apartado se analizarán las distribuciones empíricas ajustadas a cada modelo de distribución. Como se puede observar en la *Figura XXX*, la linea negra indica la función empírica de los datos originales de los pH, la línea punteada violeta representa la función empírica de la distribución gamma y la línea punteada amarilla representa la de la distribución normal.

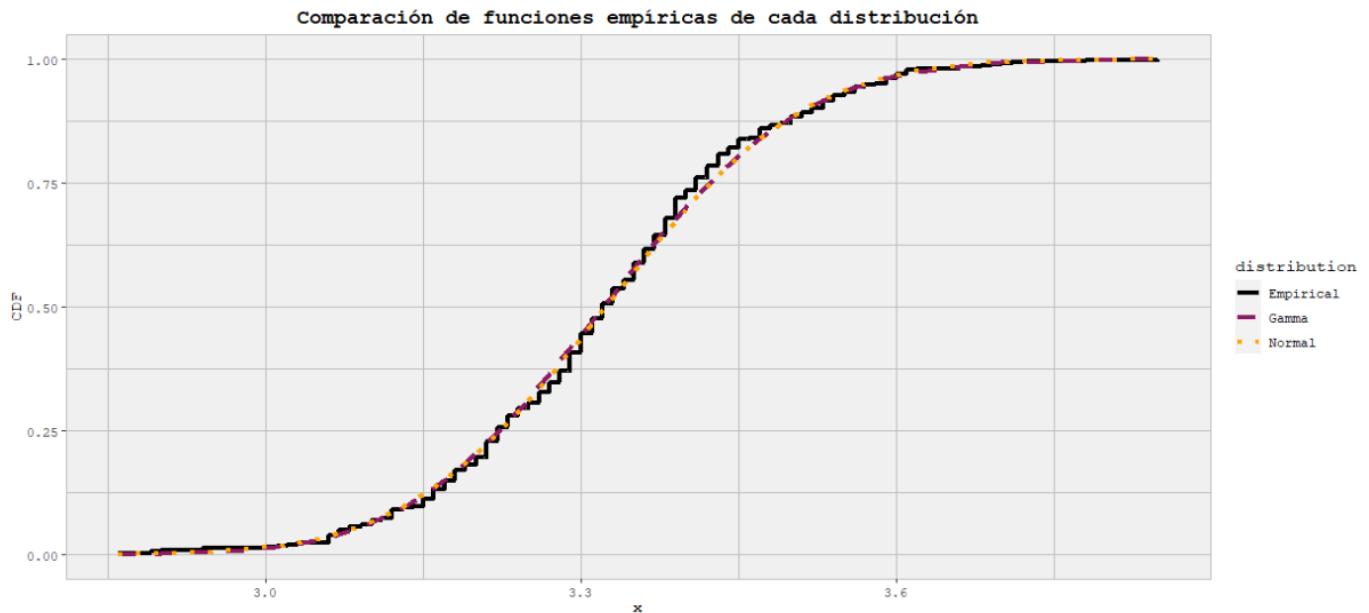


Figura XXX

Como se puede apreciar, las funciones de probabilidad acumulada correspondientes a los datos originales, en comparación con las funciones empíricas de las dos distribuciones seleccionadas, muestran una notable similitud, llegando a superponerse en diversos puntos. Esto sugiere que ambas distribuciones elegidas se ajustan de manera satisfactoria a los datos, presentando una adecuada adecuación en un grado similar para ambas.

3.4. Histograma con la densidad ajustada a cada modelo

En este apartado, se van a comparar las densidades de las distintas distribuciones elegidas. Para esto, se realizó el siguiente gráfico (véase Figura XXXI) en donde se puede observar que las distribuciones gamma y normal no difieren tanto entre sí, lo cual nos sorprende al haber obtenido valores heterogéneos para las log-verosimilitudes máximas. Asimismo, se puede observar cómo, al tener un alpha elevado ($\alpha = 502,29$) se tiene una distribución gamma que se asemeja a una normal.

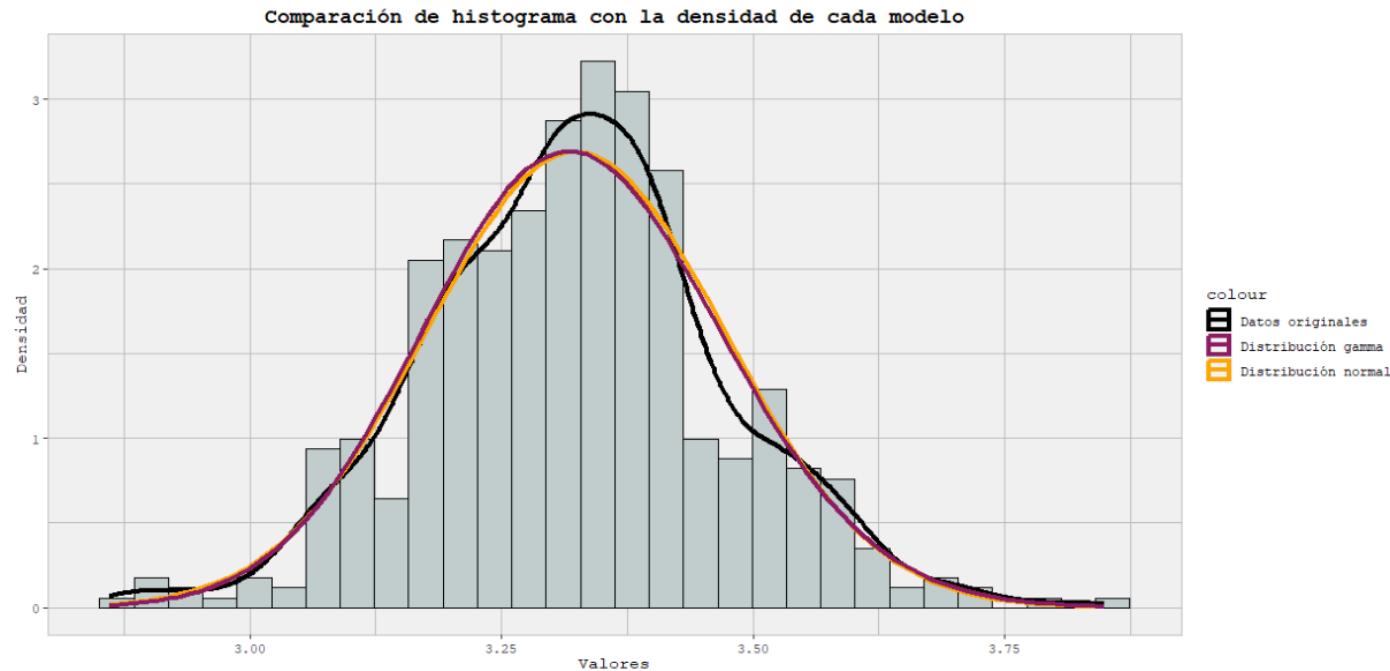


Figura XXXI

5. Intervalos de confianza

En esta sección se realizará una inferencia estadística para los estimadores enunciados en la sección 2.2. Para el desarrollo de ambos ensayos se ha tomado un nivel de significancia del 90%.

$$5.1. q_1 = x_{0.9}$$

$$\widehat{q}_1 = \widehat{F}^{-1}(0.9) = \min\{x: \widehat{F}(x_i) \geq 0.9\}$$

$$(A \leq x_{0.9} \leq B) = 1 - \alpha$$

$$(-z_{0.95} \leq \frac{x_{0.9} - x_{0.9}^*}{sd(x_{0.9}^*)} \leq z_{0.95})$$

$$(-z_{0.95} * sd(x_{0.9}^*) \leq x_{0.9} - x_{0.9}^* \leq z_{0.95} * sd(x_{0.9}^*))$$

$$(x_{0.9}^* - z_{1-\alpha} * sd(x_{0.9}^*) \leq x_{0.9} \leq x_{0.9}^* + z_{1-\alpha} * sd(x_{0.9}^*))$$

$$(3,519 - 1,644 * 0,0129 \leq x_{0.9} \leq 3,519 + 1,644 * 0,0129)$$

$$(3,498 \leq x_{0.9} \leq 3,542)$$

Con un nivel de 90% de confianza podemos afirmar que el valor para la variable pH del cuantil 0.9 se encuentra comprendido entre los valores establecidos, siendo el mínimo 3,498 y 3,542 el máximo.

5.2. $q_{\alpha} = P(x > 3.40)$

$$(A \leq P(x > 3.40) \leq B) = 1 - \alpha$$

$$(-z_{0.95} \leq \frac{P(x > 3.40) - P(x > 3.40)^*}{sd(P(x > 3.40))} \leq z_{0.95})$$

$$(-z_{0.95} * sd(P(x > 3.40))^*) \leq P(x > 3.40) - P(x > 3.40)^* \leq z_{0.95} * sd(P(x > 3.40))^*)$$

$$(P(x > 3.40)^* - z_{1-\alpha} * sd(P(x > 3.40))^*) \leq P(x > 3.40)^* \leq P(x > 3.40)^* + z_{1-\alpha} * sd(P(x > 3.40)^*))$$

$$(0,3037 - 1,644 * 0,0167 \leq x_{0.9} \leq 0,3037 + 1,644 * 0,0167)$$

$$(0,2761 \leq x_{0.9} \leq 0,331)$$

Con un nivel de 90% de confianza podemos afirmar que la probabilidad de que la variable pH tome un valor mayor a 3,4 se encuentra comprendido entre los valores establecidos, siendo el mínimo 0,2761 y 0,331 el máximo.

6. Regresión

En esta sección, se estará haciendo un análisis de distintos modelos de regresión con las variables de la base de datos de los vinos.

6.1. Análisis Exploratorio

En esta sección se estará tomando la variable numérica del pH como variable explicada y el resto de las variables del dataset como variables explicativas para analizar las distintas combinaciones de variables para ajustar modelos de regresión. En cada caso se calcularán distintas métricas como el coeficiente de determinación R ajustado, la varianza residual, el determinante de la matriz de correlaciones, el coeficiente Cp de Mallows y la suma de cuadrados de la predicción (PRESS) obtenida por validación cruzada. Todos estos valores se pueden observar en la Tabla XVI.

A fines de entender las variables de la tabla de abajo se recordarán las variables trabajadas anteriormente.

F = Acidez fija de un vino Vinho Verde

P = Medida de pH de un vino Vinho Verde

D = Densidad de un vino Vinho Verde (gr/ml)

A = Porcentaje de alcohol de un vino Vinho Verde

Adicionalmente, consideramos una variable previamente mencionada como categórica, siendo tal la calidad. Se estipula de la siguiente manera:

C = Calidad del vino Vinho Verde

Previamente, se ha mencionado que esta variable C se divide en 3 categorías 4,6 y 8, y con estos valores, se crearon las clasificaciones "Calidad baja" para los números 4, "Calidad Media" en el caso de los números 6, y "Calidad alta" para los número 8.

A continuación se desarrolla la *Tabla XV*, la cual propone una totalidad de diez modelos, los cuales parten de la base de tomar la variable elegida -PH- y evaluar diferentes combinaciones junto con las otras tres variables numéricas -acidez fija, alcohol y densidad- así como la variable categórica descrita previamente. Los criterios utilizados son los siguientes:

- **Coeficiente de determinación R^2 ajustado:** Medida corregida de bondad de ajuste (precisión de modelo) para los modelos lineales. Identifica el porcentaje de varianza en el campo de destino que se explica por la entrada o las entradas. R^2 tiende a estimar de forma optimista el ajuste de la regresión lineal.

$$R^2 = \frac{\sum_{t=1}^T (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^T (Y_t - \bar{Y})^2}$$

$$R_{ajust}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

- **Varianza residual:** Proporciona información sobre qué tan bien se ajustan los valores predichos por el modelo a los valores observados. Una varianza residual baja indica que los residuos están cercanos a cero y que el modelo puede explicar

la mayor parte de la variabilidad en los datos. Por otro lado, una varianza residual alta indica que los residuos están dispersos y que el modelo no puede explicar adecuadamente la variabilidad en los datos.

$$S^2 = \frac{\sum_i [Y_i - \hat{\varphi}(x_i)]^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

- **Determinante de la matriz de correlaciones:** El determinante de esta matriz proporciona una medida cuantitativa de la multicolinealidad. Un determinante cercano a cero indica que las variables predictoras están altamente correlacionadas entre sí, lo que sugiere la presencia de multicolinealidad.
- **Coeficiente C_p de Mallows:** Compara la precisión y el sesgo del modelo completo con modelos que incluyen un subconjunto de los predictores. Generalmente, debe buscar modelos en los que el valor de C_p de Mallows sea pequeño y esté cercano al número de predictores en el modelo más la constante (p). Así como demostrado a continuación:

$$\begin{aligned} E(C_p) &\simeq E\left[\frac{SC_{Error}(p)}{s^2} - (n - 2p)\right] \\ E(C_p) &\simeq E\left[\frac{SC_{Error}(p)}{s^2}\right] - (n - 2p) \\ E(C_p) &\simeq E(n - p) - (n - 2p) \\ E(C_p) &\simeq (n - p) - (n - 2p) \\ E(C_p) &\simeq p \end{aligned}$$

- **Suma de cuadrados de la predicción (PRESS) obtenida por validación cruzada:** Medida de error calculada en el proceso de validación cruzada. Para cada repetición de la validación cruzada, se ajusta el modelo a los datos de entrenamiento y se realiza una predicción en los datos de prueba. Luego, se calcula la diferencia entre los valores reales y las predicciones al cuadrado y se suman estos valores. Tiene en cuenta tanto el sesgo (diferencia entre los valores reales y las predicciones) como la varianza (variabilidad de las predicciones).

Modelo	Combinacione s	Coeficiente R Ajustado	Varianza Residual	Determinante Matriz	Coeficiente Cp de Correlaciones Mallows / p	PRESS
				-	1	8670,41
I	P, C	0,014	0,021	-	1	8670,41
II	P, A	0,028	0,021	-	1	537,86
III	P, F	0,499	0,011	-	1	276,21
IV	P, D	0,119	0,019	-	1	490,26
V	P, C, D	0,135	0,018	0,646	1	482,20
VI	P, A, F	0,510	0,011	0,483	1	11.293.656,00
VII	P, D, F	0,532	0,009	0,249	1	179.584,40
VIII	P, F, A, D	0,609	0,008	0,128	1	219,38
IX	P, D, F, C	0,534	0,009	0,181	1	260,93
X	P, A, D, C	0,134	0,018	0,448	1	484,48

Tabla XV

Analizando los resultados de la *Tabla XV*, desde el punto de vista del coeficiente R ajustado, el modelo que más efectividad tiene para explicar la variable del pH es el el modelo de regresión múltiple que relaciona pH con acidez fija, alcohol y densidad -modelo VIII- (casualmente las 4 variables numéricas de nuestra base de datos). Asimismo, este modelo es el que presenta menor varianza residual de todos. Sin embargo, decidimos descartar debido a que su determinante de la matriz de correlaciones es cercano a 0,1 -siendo este un criterio para descartar el modelo- (situación similar para el modelo IX). A su vez, los modelos de regresión de las variables pH, combinado con acidez fija y densidad y pH combinado con acidez fija y alcohol presentan un PRESS llamativamente alto. Por esta razón, tomamos el modelo III como definitivo para trabajar más adelante, siendo que su varianza residual es baja, al igual que el PRESS y su R^2 ajustado es alto.

6.2. Diagnóstico

Se eligió el modelo de regresión lineal simple del pH como variable como variable explicada y la acidez fija como variable explicativa para realizar un diagnóstico de los supuestos de linealidad de la regresión,

de normalidad de los errores, de homocedasticidad de los errores, de independencia de los errores y analizar los outliers y puntos influyentes. Se eligió este modelo de regresión lineal de dos variables ya que es que mayor valor de R-ajustado tiene (véase Tabla XV). Todo este análisis se llevará a cabo para evaluar si el modelo es apto o no.

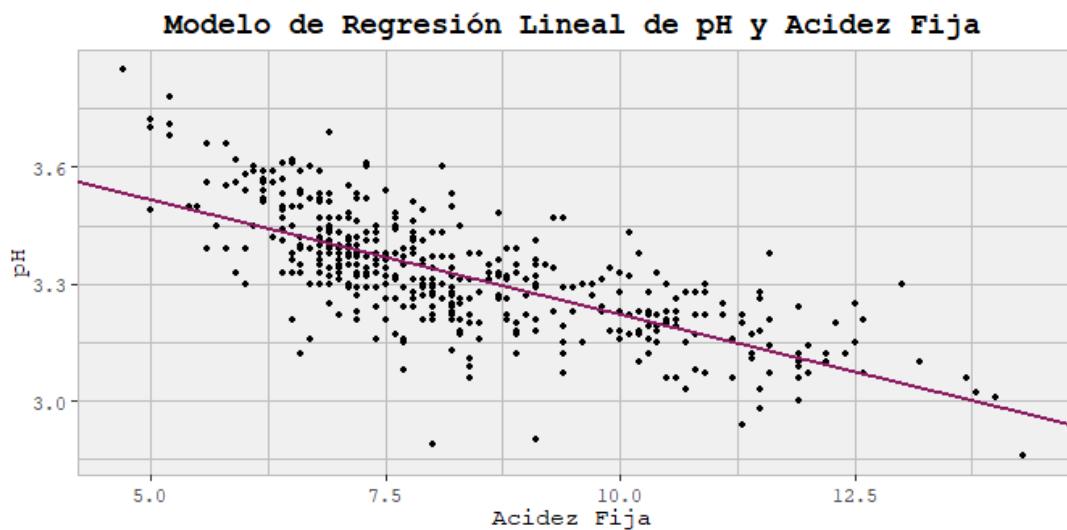


Figura XXXII

En la *Figura XXXII* se puede observar el modelo de regresión lineal ilustrado para poder tener un mejor entendimiento del modelo.

6.2.1. Supuesto de linealidad de la regresión

Para este punto, se va a trabajar con los residuos estandarizados del modelo comparándolos con los valores de la acidez fija. Como se puede observar en el gráfico de dispersión de la *Figura XXXIII*, los puntos no forman una linealidad, lo que indica que no se cumple el primer supuesto necesario para el modelo de regresión lineal. Esto afecta ya que el modelo podría generar estimaciones sesgadas.

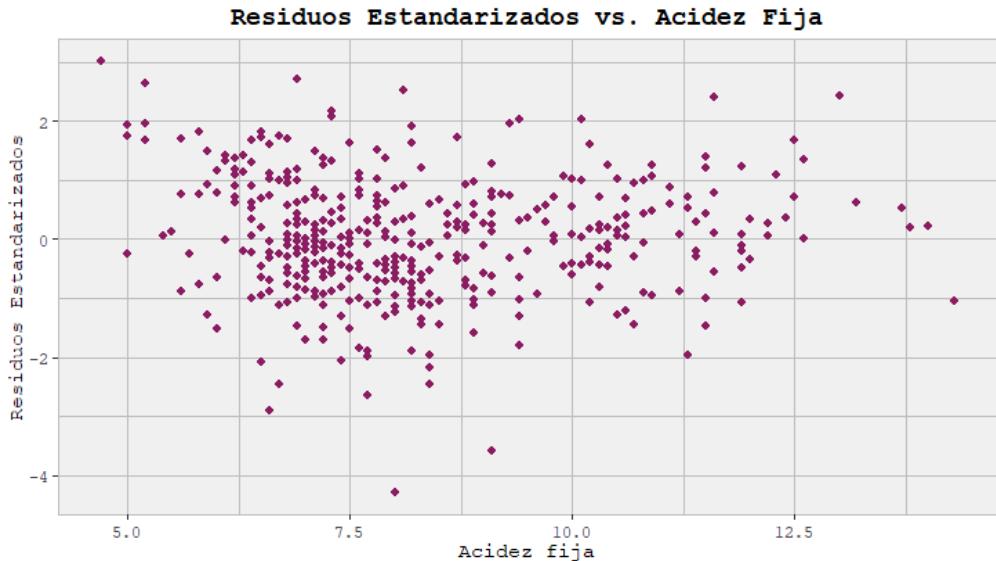


Figura XXXIII

6.2.2. Supuesto de normalidad de los errores

Para este punto, se va a evaluar el supuesto de normalidad trabajando, como en el punto anterior, con los errores estandarizados. Para observar la forma que toman los puntos, se graficó un histograma (véase Figura XXXIV), en donde se puede ver que hay una forma acampanada que podría llegar a ser aproximarse a una normal.

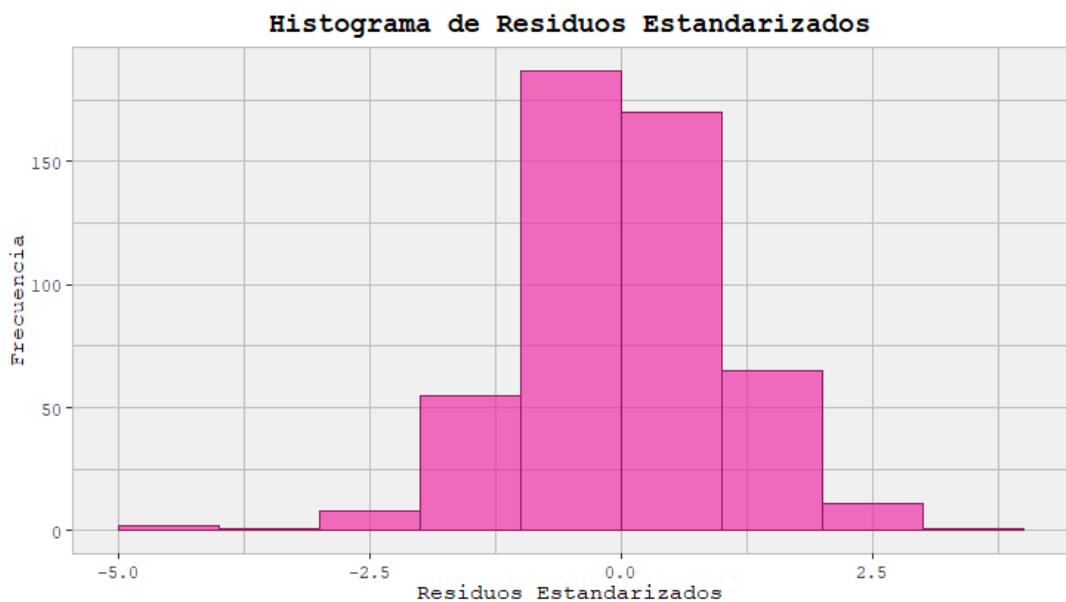


Figura XXXIV

De todos modos, se decidió comprobar si realmente la distribución se asemejaba a una normal empleando un QQPlot que se puede observar en la *Figura XXXV*. Como se puede ver, el conjunto de residuos no parece seguir una distribución normal, especialmente cerca de las colas. A pesar de que gráficamente se puede identificar que la distribución tiene colas pesadas, la misma asemejarse a una distribución normal debido a que su curtosis es leve y su asimetría no es pronunciada.

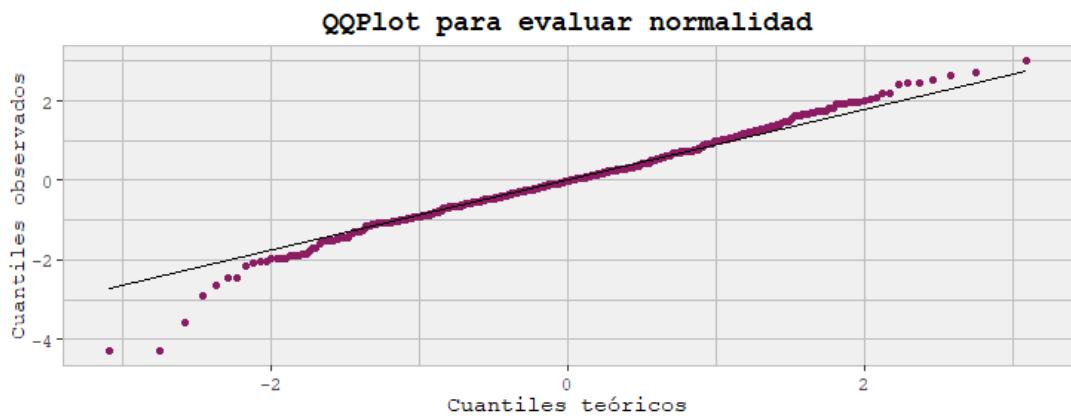


Figura XXXV

6.2.3. Supuesto de homocedasticidad de los errores

En este apartado se estará analizando si los residuos son homocedásticos, es decir, que sus varianzas son constantes. Para eso, se graficó la siguiente *Figura XXXVI* en donde se puede observar que los residuos tienen forma de embudo. Por este motivo no se puede afirmar homocedasticidad.

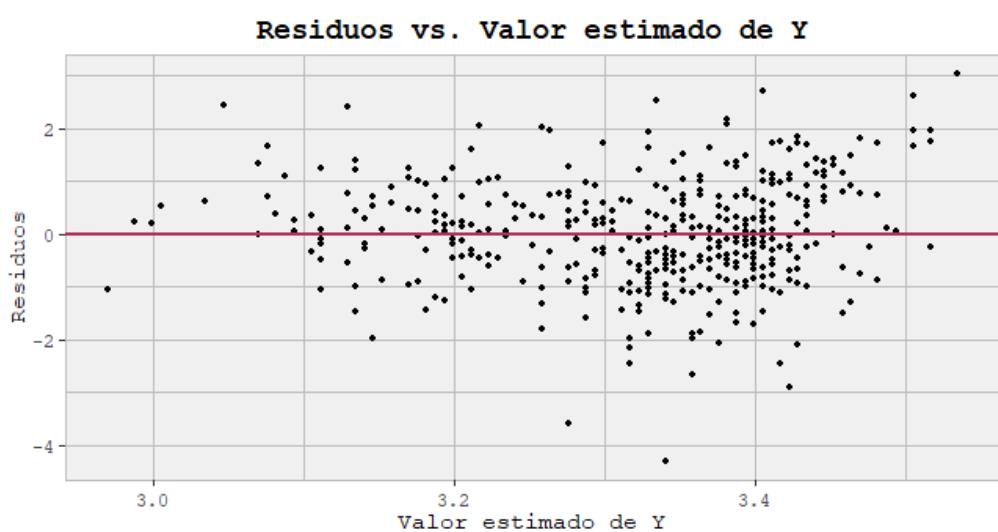


Figura XXXVI

Debido a que no se puede afirmar con exactitud de que haya homocedasticidad en los residuos a lo largo de los valores ajustados se realizará una transformación de Box-Cox para ayudar a estabilizar la varianza de los residuos y cumplir con el supuesto de homocedasticidad.

La transformación de Box-Cox es una técnica utilizada en estadística para estabilizar la varianza de los datos y mejorar la linealidad de la relación entre variables. Es aplicada a una variable dependiente o respuesta y busca encontrar una transformación óptima que maximice la aproximación a una distribución normal y homocedasticidad (varianzas constantes) en el modelo de regresión.

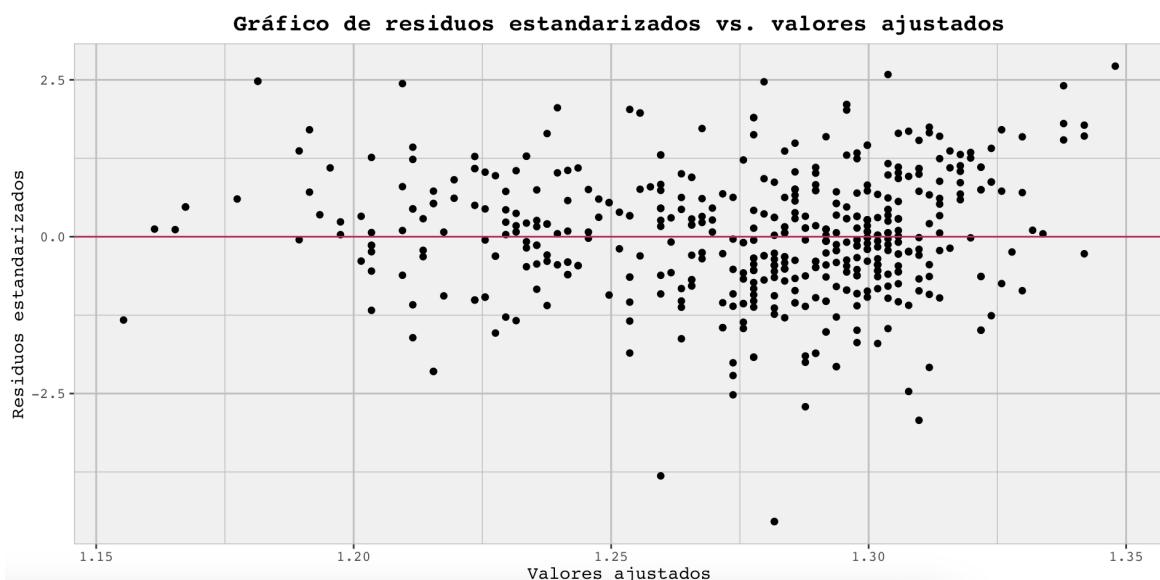


Figura XXXVII

En la *Figura XXXVII* se puede observar cómo queda el modelo de regresión lineal simple aplicando la transformación de Box-Cox.

6.2.4. Supuesto de independencia de los errores

En esta sección se quiso analizar la independencia de los errores dentro del modelo de regresión lineal elegido. Para eso se graficó la siguiente *Figura XXXVIII* en donde se observa que los residuos no tienen ninguna tendencia. Como se puede observar, los puntos no muestran ningún patrón entre ellos, lo que confirma el supuesto de independencia de los errores.

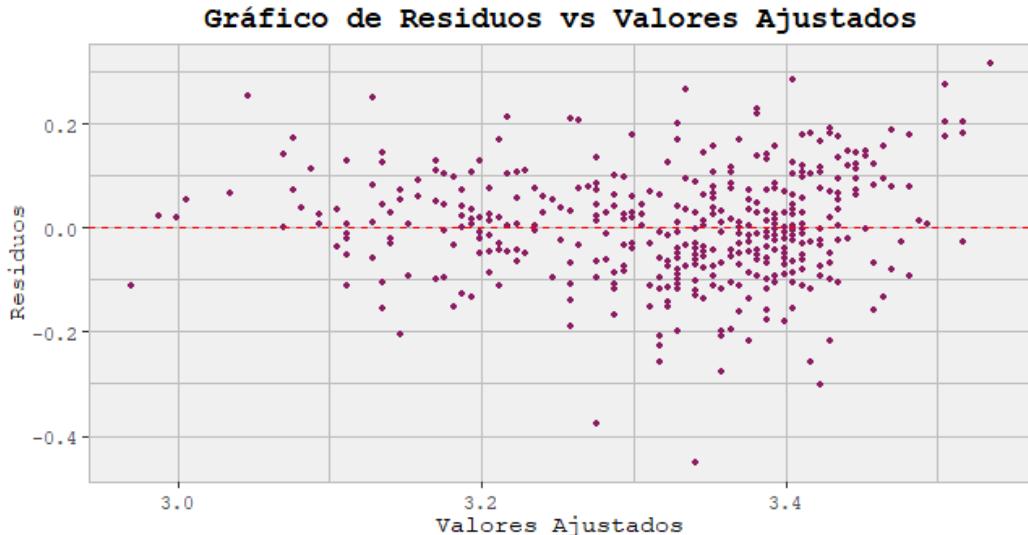


Figura XXXVIII

6.2.5. Outliers y puntos influentes

En este apartado se analizarán los valores atípicos dentro del modelo de regresión lineal elegido, y, si es que los hay, si éstos influyen en el modelo. Para eso, se va a emplear el cálculo de la Distancia de Cook cuyo objetivo es calcular cuán influentes son cada uno de los puntos para la recta de regresión, y si los mismos son “palanca”.

$$D_i = \frac{\sum_{j=1}^N (\hat{Y}_j - \hat{Y}_{j(-i)})^2}{p \cdot S^2}$$

Para el cálculo de las Distancias de Cook de cada punto de la muestra, se usa la ecuación de arriba, que lo que hace es ver si hay diferencia entre el valor predecido al valor estimado de la variable explicada (\hat{Y}) con el punto analizado dentro del modelo y el valor predecido de \hat{Y} sin el punto analizado dentro del modelo. Una vez calculados los valores de D_i , se generó la Figura XXXIX para visualizarlos. Consideraremos que aquellos que son outliers son aquellos que $D_i > 1$ por lo que no hay razón aparente para apartarlos.

Como se puede observar, si bien existen tres observaciones de mayor magnitud que sugieren una influencia significativa en la variación de la línea de regresión, estas no exceden el 0,5. Finalmente, para un segundo punto de vista sobre su permanencia en el modelo, se realizó un gráfico

(véase *Figura XL*) comparando las rectas de regresión con y sin esos valores.

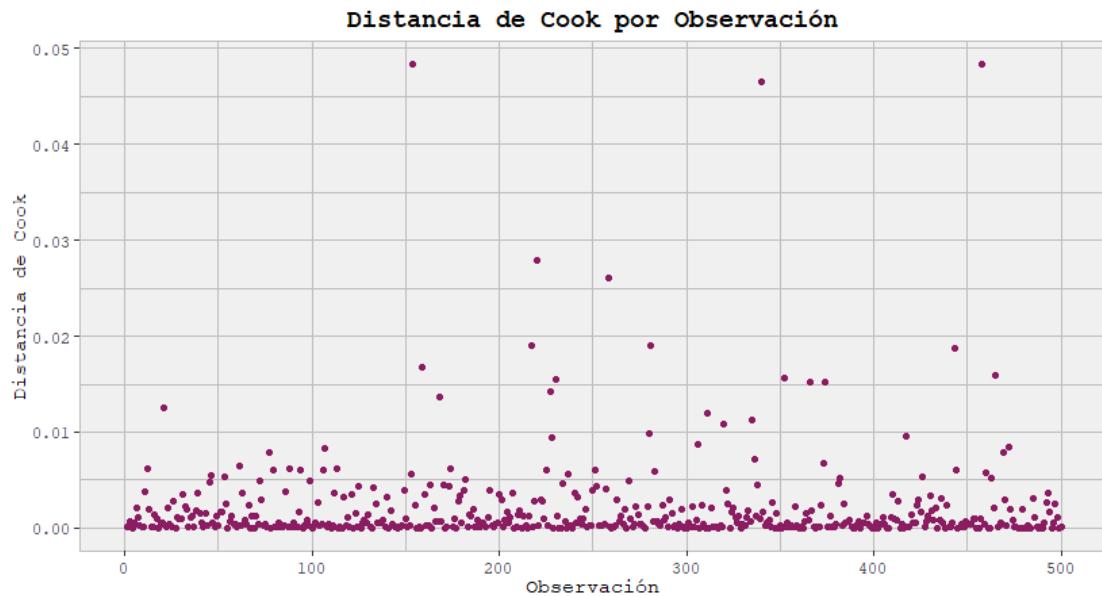


Figura XXXIX

En la *Figura XL* se puede ver las dos rectas de regresión mostrando cómo se modifican las rectas si se sacan los 3 puntos palancas que se observaron en la *Figura XXXIX*. Como se puede apreciar en el gráfico, las rectas están muy superpuestas, al punto de que no se puede distinguir claramente la diferenciación de sus colores. Por esta razón, consideramos que estos puntos de palanca no tienen un impacto significativo en la línea de regresión lineal y hemos decidido mantenerlos en el conjunto de datos.

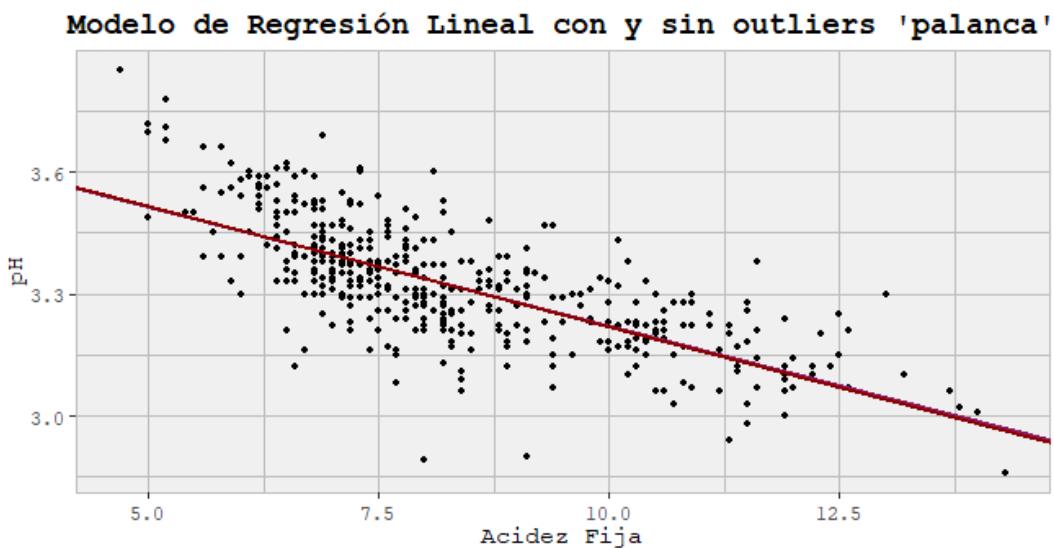


Figura XL

6.3. Validación del modelo

En esta sección, se plantea el modelo definitivo de regresión lineal simple del pH como variable dependiente y la acidez fija como variable explicativa. Es importante aclarar que, como se mencionó en el inciso anterior, se cumplen los supuestos de normalidad, linealidad, homocedasticidad e independencia de los residuos del modelo.

Para encontrar la ecuación al modelo se encontraron los valores de los parámetros utilizando el método de cuadrados mínimos que se expresan a continuación. Los resultados obtenidos para los parámetros de este modelo (cuyos datos fueron entrometidos en una transformación) se encuentran en la *Tabla XVI*.

$$\hat{\beta}_1 = \frac{s_{ap}}{S_a^2} = \frac{\sum(a_i - \bar{a}) \cdot (p_i - \bar{p})}{\sum(a_i - \bar{a})^2}$$

$$\hat{\beta}_0 = \bar{p} - \hat{\beta}_1 \cdot \bar{a}$$

$$S^2 = \frac{SCR}{n-k} = \frac{\sum(p_i - \hat{p}_i)^2}{n-p}$$

$$\widehat{Se}(\hat{\beta}_1) = \sqrt{\frac{S^2}{\sum(p_i - \hat{p}_i)^2}}$$

k = cantidad de parámetros

Parámetro	$\hat{\beta}_1$	$\hat{\beta}_0$	S^2	$\widehat{Se}(\hat{\beta}_1)$
Valor	-0,058	3,811	0,011	0,031

Tabla XVI

Luego, se decidió hacer intervalos de confianza del 90% para $\hat{\beta}_1$ y $\hat{\beta}_0$. Con el intervalo de confianza de $\hat{\beta}_1$, se desea evaluar la siguiente hipótesis:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

A modo de interpretación, cuando el coeficiente beta (β) de una variable explicativa en un modelo de regresión lineal es igual a cero -hay evidencia para afirmar que H_0 es verdadera-, significa que no hay una relación lineal significativa entre esa variable y la variable de respuesta (variable dependiente). En otras palabras, la variable explicativa no tiene un impacto estadísticamente significativo en la variable de respuesta en el contexto del modelo de regresión lineal. A continuación, se presentan las expresiones para calcular los intervalos de confianza para los coeficientes del modelo de regresión.

$$(A \leq \hat{\beta}_1 \leq B) = 1 - \alpha$$

$$(t_{n-2, 0.05} \leq \frac{\hat{\beta}_1 - \hat{\beta}_1}{\widehat{Se}(\hat{\beta}_1)} \leq t_{n-2, 0.95})$$

$$(t_{n-2, 0.05} * \widehat{Se}(\hat{\beta}_1)) \leq \hat{\beta}_1 - \widehat{Se}(\hat{\beta}_1) \leq t_{n-2, 0.95} * \widehat{Se}(\hat{\beta}_1))$$

$$(\widehat{\beta}_1 + t_{n-2, \frac{\alpha}{2}} * \widehat{Se}(\hat{\beta}_1)) \leq \hat{\beta}_1 \leq (\widehat{\beta}_1 + t_{n-2, 1-\frac{\alpha}{2}} * \widehat{Se}(\hat{\beta}_1))$$

$$(A \leq \hat{\beta}_0 \leq B) = 1 - \alpha$$

$$(t_{n-2, 0.05} \leq \frac{\hat{\beta}_0 - \hat{\beta}_0}{\widehat{Se}(\hat{\beta}_1)} \leq t_{n-2, 0.95})$$

$$(t_{n-2, 0.05} * \widehat{Se}(\hat{\beta}_1)) \leq \hat{\beta}_0 - \widehat{Se}(\hat{\beta}_1) \leq t_{n-2, 0.95} * \widehat{Se}(\hat{\beta}_1))$$

$$(\widehat{\beta}_0 + t_{n-2, \frac{\alpha}{2}} * \widehat{Se}(\hat{\beta}_1)) \leq \hat{\beta}_0 \leq (\widehat{\beta}_0 + t_{n-2, 1-\frac{\alpha}{2}} * \widehat{Se}(\hat{\beta}_1))$$

Los resultados a los intervalos de confianza para los coeficientes del modelo de regresión lineal se encuentran a continuación en la *Tabla XVII*.

Parámetro	$\hat{\beta}_1$		$\hat{\beta}_0$	
Límite del intervalo	Inferior	Superior	Inferior	Superior
Valores	-0,111	-0,006	3,759	3,863

Tabla XVII

Así como se ha indicado previamente, el intervalo de confianza del $\widehat{\beta}_1$ no puede contener al 0, ya que de esa manera podría suceder que la pendiente del modelo de regresión lineal sea nula y el modelo lineal no se ajuste correctamente. Dado que $0 \notin IC$ rechazamos H_0 por lo que indica que el modelo tiene una relación lineal.

Parámetro	$\widehat{\beta}_1$	$\widehat{\beta}_0$	S^2	$\widehat{Se}(\widehat{\beta}_1)$
Valor	-0,058	3,811	0,011	0,031

Tabla XVIII

$$\varphi(x) = E(Y|x) = 3,811 - 0,058x$$

6.4. Aplicación

Para esta última sección, se hará una aplicación del modelo definitivo de la regresión lineal simple que definimos en el apartado anterior. Para ello se realizará un intervalo de predicción para la variable respuesta, dado un valor fijo para las variables explicativas -en este caso acidez-. Un intervalo de predicción proporciona un rango dentro del cual se espera que se encuentre el valor real de una nueva observación con cierto nivel de confianza. Este intervalo tiene en cuenta tanto el error de estimación del modelo como la variabilidad inherente de los datos.

Puntualmente, se evaluará con un nivel de significancia de 90%, un intervalo de predicción para la variable explicada cuando la acidez fija es igual a 8. Para ello, se calculan los siguientes estimadores:

$$\widehat{\sigma}_{a_0} = \sqrt{S^2 \left(1 + \frac{1}{n} - \frac{(a_0 - \bar{a})^2}{\sum (a_i - \bar{a})^2} \right)}$$

$$(A \leq p \leq B) = 1 - \alpha$$

$$(t_{n-2, 0.05} \leq \frac{p - \widehat{p}_{a=8}}{\widehat{\sigma}_{a=8}} \leq t_{n-2, 0.95})$$

$$(t_{n-2, 0.05} * \widehat{\sigma}_{a=8} \leq p - \widehat{p}_{a=8} \leq t_{n-2, 0.95} * \widehat{\sigma}_{a=8})$$

$$\left(\widehat{p}_{a=8} + t_{n-2, \frac{\alpha}{2}} * \widehat{\sigma}_{a=8} \leq p \leq \widehat{p}_{a=8} + t_{n-2, 1-\frac{\alpha}{2}} * \widehat{\sigma}_{a=8} \right)$$

Parámetro	p	
Límite de intervalo	Inferior	Superior
Valores	3,167	3,513

Tabla XX

7. Conclusiones

En conclusión, en este informe se llevó a cabo un análisis exhaustivo de una base de datos de vinos “winequality-red.csv”, aplicando diversas técnicas estadísticas.

Se realizó un análisis descriptivo detallado de las variables numéricas, lo que permitió comprender mejor las características y distribuciones de los datos. Además, se realizaron estimaciones puntuales de algunas variables clave, lo que proporcionó una idea precisa de los valores esperados. Posteriormente, se realizaron pruebas de bondad de ajuste para evaluar qué tan bien se ajustaban los datos a ciertas distribuciones teóricas. Los intervalos de confianza se calcularon para estimar rangos de valores probables, lo que ayudó a tomar decisiones más informadas. Por último, se llevaron a cabo análisis de regresión para explorar las relaciones entre variables y predecir valores. Estos análisis revelaron patrones y tendencias significativas en los datos.

Este trabajo nos ha brindado una comprensión profunda de la importancia de la estadística en el estudio de cualquier fenómeno o conjunto de datos. Nos ha permitido apreciar cómo la estadística es una herramienta fundamental para la toma de decisiones basada en evidencia, tanto en el ámbito académico como en el mundo real.

La aplicación de la estadística trasciende la materia en sí, ya que se puede utilizar en diversos campos y situaciones de la vida cotidiana. Desde la toma de decisiones empresariales hasta el análisis de políticas públicas, la estadística nos ayuda a comprender los datos, identificar patrones, realizar predicciones y tomar decisiones fundamentadas.

8. Bibliografía

- Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009. [en linea]. Obtenido en:
<https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>

Recursos adicionales utilizados:

- Documentacion de libreria GGPLOT [en linea]
<https://www.rdocumentation.org/packages/ggplot2/versions/3.4.1>