TP 1 - AID

Florencia Altschuler

# Base de datos y muestreo

Se utilizó la base de datos estructurada y abierta “*Conjunto de datos de calidad del vino*” (Cortez et al., 2009.) de 13 atributos y 6497 instancias. De todos los atributos, 11 son variables basadas en características fisicoquímicas (variables cuantitativas continuas), una es una variable cardinal que indica la calidad del vino (1 al 10) y la última es una variable cualitativa nominal que indica el tipo de vino (tinto o blanco).

Se tomó como supuesto la aleatoriedad e independencia de las muestras. Se realizó un muestreo balanceado estratificado según la variedad de vino.

# Medidas descriptivas

Para todas las variables, según la variedad del vino, se obtuvieron medidas de tendencia central (media, mediana y moda), de dispersión (varianza, rango muestral, desviación estándar, coeficiente de variación, rango intercuartílico), de posición (cuartil 1, cuartil 3) y de la forma de la distribución (asimetría y curtosis). Para el cálculo de la moda para las 11 variables continuas se utilizó el metodo *meanshift*. Para el cálculo del coeficiente de asimetría se utilizó el método *default* (Fisher) ya que no tiene restricciones.

# Representación gráfica

Las variables continuas se graficaron con un *boxplot* mientras que para la variable *Calidad* se realizó un histograma. La variable variedad de vino se excluyó de la gráfica ya que en el muestreo se fijó un N=1000 para cada variedad, por lo que la representacion grafica no aportaria mayor informacion de la que ya se disponía.

Se realizaron dos diagramas circulares para representar las proporciones de calidad de vino para las dos variedades y un gráfico de barras adyacentes.

# Tests de hipótesis

Se eligieron las variables anhidrido sulfuroso total y anhidrido sulfuroso libre con la idea de que podrían llegar a estar directamente relacionadas ya que al tratarse del mismo compuesto no sería difícil imaginarse un vino que contenga altos niveles de una pueda llegar a contener altos niveles de la otra variable.

***H0****: Los niveles de anhidrido sulfuroso libre son independientes de los niveles de anhidrido sulfuroso total*

***H1****: Las variables están asociadas*

Para poner a prueba tal hipótesis se definieron nuevas variables categóricas dividiendo las variables por la mediana aproximadamente. Se aplicó el test Chi-cuadrado y, viendo que se cumplían los supuestos, se rechazó la H0 y se analizaron las subtablas de contingencia concluyéndose que, con un nivel de confianza del 95%, niveles bajos de anhidrido sulfuroso libre están asociados con niveles bajos de anhidrido sulfuroso total y lo mismo para niveles altos de ambas variables.

Otra pregunta que se quiso responder es si existían diferencias de pH según la variedad del vino. Para eso se estimó un intervalo confianza (α=0.05, suponiendo una distribución cercana a la normal por el teorema central del límite) de pH para tintos (3.18 - 3.20) y blancos (3.31 - 3.33). Como las dos muestras no seguían una distribución normal (aún intentando remover hasta 5 outliers de cada variedad), se optó por el test no paramétrico para dos poblaciones de Mann-Whitney-Wilcoxon a una cola, resultó en un p-valor < 2.2e-16, por lo que se puede rechazar la hipótesis nula; el pH del vino tinto es menos del de vino blanco con una confianza del 95%.

Luego se deseaba saber si existen diferencias significativas en la calidad del vino tinto respecto al vino blanco. Para eso se realizó un test de homogeneidad y se plantearon las hipótesis:

***H0****: la calidad es homogénea en las variedades*

***H1****: la calidad no es homogénea en las variedades*

Se aplicó el test Chi-cuadrado y habiéndose cumplido los supuestos se obtuvo p < 2.2e-16, realizó un test post hoc de bonferroni y se encontró que hay más probabilidad que un vino sea blanco dado que tiene un puntaje de 5 y hay mas probabilidad que un vino sea tinto dado que tiene un puntaje de 7 y 8-9 . (el puntaje 9 se agrupo con el 8 porque si no no se cumplia el supuesto que las frecuencias esperadas sean mayores a 5).

Por último quiso evaluar si existían diferencias significativas en las proporciones de alcohol entre los vinos de calidad baja (3 a 5) , media (6 a 7) y alta (8 a 9). Para ello se realizó el test no parametrico de kruscal wallis (dado que no se cumplían los supuestos de normlidad y homocedasticidad de los errores)

***Ho*** *Las tres calidades de vino tienen la misma posición para la proporcion de alcohol dada por el puntaje*

***Ha*** *al menos una calidad de vino tiene diferente posición para proporcion de alcohol dada por el puntaje.*

Se encontraron diferencias significativas entre la proporcion de alcohol y la calidad (para las tres comparaciones) siendo los de menor calidad los vinos de menor proporción de alcohol y aquellos de mayor calidad los de mayor volúmen de alcohol.