**Ejercicio Propuesto 2  (Resuelto)**

**Se pretende estudiar la acción que ejercen tres tipos de levadura distintos en el proceso de fermentación del vino, para ello se desea conocer cómo es la cantidad de levadura en 1 mm3 de vino contadas según los tres tipos distintos de levadura, que son Kloeckera, Pichia y Mycoderma al finalizar todo el proceso. Se toman 10 muestras aleatorias de 1 mm3 a cada una de las tres especies, y se toma el recuento de la cantidad de levadura:**

****

Tabla 1: Datos del Ejercicio Propuesto 2

1. **¿Influye el tipo de levadura en el proceso de fermentación del vino? Utilice un nivel de significación del 5%.**

**2. En caso afirmativo en el apartado anterior, ¿entre qué tipos? Utilice un nivel de significación del 5%.**

**3. Estudiar las hipótesis del modelo: Homocedasticidad, independencia y normalidad.**

**Solución**

1. **¿Influye el tipo de levadura en el proceso de fermentación del vino? Utilice un nivel de significación del 5%.**

Nos interesa saber si el tipo de levadura influye significativamente en la cantidad de levadura en 1 mm3 de vino

* La cantidad de levadura en 1 mm3 de vino es una **variable numérica**
* El tipo de levadura es una **variable categórica** que divide a los individuos en grupos

Nuestro contraste de hipótesis es:



Es decir,

$$\begin{matrix}H\_{0}:la cantidad de levadura en 1 mm3 es la misma para los tres tipos\\H\_{1}:la cantidad de levadura en 1 mm3 es diferente según el tipo \end{matrix}$$

En primer lugar, cargamos el paquete BrailleR

> library(“BrailleR”)

almacenamos los datos de las dos variables en dos vectores:

> tipo = factor(c(rep("Kloeckera",10), rep("Pichia",10), rep("Mycoderma",10)),levels = c("Kloeckera", "Pichia", "Mycoderma"))

> cantidad=c(279, 301, 337, 287, 331, 297, 326, 297, 304, 294, 204, 177, 212, 224, 197, 220, 198, 211, 187, 229, 190, 200, 220, 180, 160, 169, 203, 230, 186, 198)

Agrupamos las 2 variables en un data frame, al que llamamos *levadura*:

> levadura = data.frame (tipo, cantidad)

Realizamos el ANOVA pedido mediante la función **OneFactor** sin que nos muestre el Test post-hoc.

> OneFactor("cantidad", "tipo", levadura, HSD = FALSE)

[Muestra una salida en html:](file:///C%3A%5CUsers%5CUsuario%5CDesktop%5Cdatos%5CCantidad.Tipo-OneFactor.html)

# **Analysis of the Levadura data, using Cantidad as the response variable and Tipo as the single grouping factor.**

#### **Prepared by BrailleR**

A continuación, se muestra y se detalla cada resultado:

1. **Una tabla resumen que incluye para cada grupo la media, la desviación típica, el número de datos y el error estándar**

****

Tabla 2: Resumen para cada tipo de levadura

1. **Diagramas de caja y bigotes comparativos para cada grupo**

****

Figura 1: Diagrama de caja y bigotes para cada tipo de levadura

* + Los grupos Pichia y Mycoderma parecen seguir una distribución simétrica, mientras que en el grupo Kloeckera no se ve claramente tal simetría.
	+ No hay valores atípicos.
	+ El tamaño de las cajas es similar para los distintos niveles, por lo que no hay indicios de falta de homocedasticidad.
	+ Se puede intuir que existirá diferencia de medias entre los grupos Pichia y Mycoderma con respecto a Kloeckera.
1. **Gráfico de puntos comparativos para cada grupo**

****

Figura 2: Gráfico de puntos para cada tipo de levadura

Al igual que en el diagrama de caja y bigotes, se puede tener una idea intuitiva de si las medias difieren unas de otras o no. Obteniendo las mismas conclusiones que con el gráfico anterior.

1. **Análisis de la varianza de un factor**

 Df Sum Sq Mean Sq F value Pr(>F)

tipo 2 75028 37514 101 2.92e-13 \*\*\*

Residuals 27 10027 371

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Esta es la tabla ANOVA que nos proporciona un p-valor=2.92e-13 por lo que se rechaza la hipótesis nula, esto quiere decir que existen diferencias significativas en la cantidad de levadura en 1 mm3 de vino según el tipo de la misma (al menos en dos de ellos). Esto es, el tipo de levadura influye en el proceso de fermentación del vino.

1. **En caso afirmativo en el apartado anterior, ¿entre qué tipos? Utilice un nivel de significación del 5%.**

Se realiza un análisis post-hoc, mediante el método Tukey HSD. Añadimos a la función el test post-hoc de la siguiente forma:

> OneFactor("cantidad", "tipo", levadura)

Esta nos devuelve todo lo anterior junto con lo que se especifica a continuación:

1. **Test post-hoc mediante el método de Tukey Honestly Significant Difference**

 Tukey multiple comparisons of means

 95% family-wise confidence level

 factor levels have been ordered

Fit: aov(formula = cantidad ~ tipo, data = levadura)

$tipo

 diff lwr upr p adj

Pichia-Mycoderma 12.3 -9.068645 33.66864 0.3415057

Kloeckera-Mycoderma 111.7 90.331355 133.06864 0.0000000

Kloeckera-Pichia 99.4 78.031355 120.76864 0.0000000



Figura 3: Intervalos de confianza de la diferencia de medias para cada par de tipos

Tanto numéricamente mediante el p-valor como gráficamente se tiene que:

* Pichia y Mycoderma: Puesto que p-valor=0.3415057, no se puede rechazar H0, es decir, no hay diferencias significativas en la cantidad de levadura en 1 mm3 de vino según si el tipo es Pichia o Mycoderma.
* Kloeckera y Mycoderma: Puesto que p-valor=0, se rechaza H0, es decir, existen diferencias significativas en la cantidad de levadura en 1 mm3 de vino entre los tipos Kloeckera y Mycoderma.
* Kloeckera y Pichia: Puesto que p-valor=0, se rechaza H0, es decir, existen diferencias significativas en la cantidad de levadura en 1 mm3 de vino entre los tipos Kloeckera y Pichia.

En definitiva, el tipo de levadura Kloeckera influye en el proceso de fermentación del vino de manera diferente a Pichia y Mycoderma.

Se pueden deducir las mismas conclusiones, observando que los intervalos de confianza, tanto analítica como gráficamente no contienen al cero.

1. **Estudiar las hipótesis del modelo: Homocedasticidad, independencia y normalidad.**

Para comprobar la idoneidad del modelo propuesto, la salida de la función **OneFactor** muestra los gráficos:

1. **Valores ajustados frente a residuos**
2. **Gráfico Q-Q de normalidad**
3. **Valores ajustados frente a raíz cuadrada de los residuos estandarizados (en valor absoluto)**
4. **Residuos estandarizados frente a leverages**

****

Figura 4: Gráficos de comprobación de las hipótesis

Los gráficos 1 y 3 se utilizan para contrastar gráficamente si la media de los residuos es cero y la homocedasticidad de los residuos. La idea intuitiva para que se cumplan estas hipótesis es que no se observe ningún patrón ni forma de embudo. Además, en el gráfico 3 se usa como regla general que cualquier punto por encima de 2 en el eje Y sugeriría heterogeneidad de varianza. Por tanto, se confirma que los residuos tienen media cero y son homocedásticos.

Observando el gráfico Q-Q, se observa que los residuos estandarizados están próximos a la diagonal punteada que aparece en el gráfico, por tanto, los residuos siguen una distribución normal.

El gráfico de Leverage para cada grupo frente a los residuos estandarizados se utiliza para detectar valores atípicos. En caso de detectarse alguna observación fuera del rango [-2,2] debe estudiarse este punto de forma aislada para detectar, por ejemplo, si la elevada importancia de esa observación se debe a un error. No se detectan valores atípicos.

La hipótesis de independencia es intrínseca al diseño del experimento, por tanto, se supone que las observaciones son independientes ya que se están considerando distintos tipos de levadura.

1. **Dos tests de homogeneidad de varianzas**

Vamos a confirmar la homocedasticidad mediante dos métodos analíticos.

* **El test de homogeneidad de varianzas de Bartlett**

 Bartlett test of homogeneity of variances

data: cantidad by tipo

Bartlett's K-squared = 0.58681, df = 2, p-value = 0.7457

Considerando como nivel de significación 0.05, y observando el p-valor=0.7457 no tenemos evidencia muestral para rechazar la hipótesis nula, por lo que la varianza permanece constante entre grupos.

* **El test de homogeneidad de varianzas de Fligner-Killeen**

 Fligner-Killeen test of homogeneity of variances

data: cantidad by tipo

Fligner-Killeen:med chi-squared = 0.48253, df = 2, p-value = 0.7856

Considerando como nivel de significación 0.05, y observando el p-valor=0.7856 no tenemos evidencia muestral para rechazar la hipótesis nula, por lo que la varianza permanece constante entre grupos.