Laboratorio 9. Comprobación de las hipótesis del contraste ANOVA

Para realizar un contraste de igualdad de múltiples medias mediante los tests de ANOVA, tenemos que comprobar dos hipótesis:

- Normalidad de los datos dentro de cada nivel.
- Igualdad de varianzas de los datos dentro de todos los niveles (homocedasticidad).

En este laboratorio aprenderemos cómo comprobar las hipótesis anteriores.

1. Comprobación de la normalidad de los datos: técnica `qqplot` o `qqnorm`

Una de las técnicas que se utilizan para contrastar la bondad de ajuste de unos datos a una distribución teórica son las técnicas gráficas. Ya hemos visto, en laboratorios anteriores, cómo se puede comprobar si el perfil de una distribución de una muestra se acerca al perfil de la campana de Gauss asociada (la que tiene por media y por desviación típica las muestrales) utilizando los histogramas de densidades. La comparación por curvas de densidad estimada y teórica en los histogramas es difícil de evaluar de forma visual.

Otros métodos gráficos hacen más sencilla la interpretación la comparación visual de una distribución de una muestra con una teórica. Los más básicos son los conocidos como los gráficos de "*probabilidad contra probabilidad*"más conocidos por su acrónimo en inglés *ppplot* y también los gráficos de *cuatil contra cuantil*, en inglés *qqplot*.

1.1. Función de distribución empírica

Dado un conjunto de datos $x_1, \ldots, x_n$, llamaremos función de distribución empírica F_E a la función definida a continuación.

Sean $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ los datos anteriores ordenados. Entonces la función F_E se define de la forma siguiente: dado un valor x cualquiera, $F_E(x)$ vale:

$$F_E(x) = \begin{cases} 0, & \text{si } x < x_{(1)}, \\ \frac{1}{n}, & \text{si } x_{(1)} \leq x < x_{(2)}, \\ \cdots & \cdots \\ \frac{k}{n}, & \text{si } x_{(k)} \leq x < x_{(k+1)}, \\ \cdots & \cdots \\ 1, & \text{si } x \geq x_{(n)}. \end{cases}$$

Por ejemplo, para los 15 datos siguientes

```
> n <- 15
> x <- rnorm(n)
> round(x, 2)
```

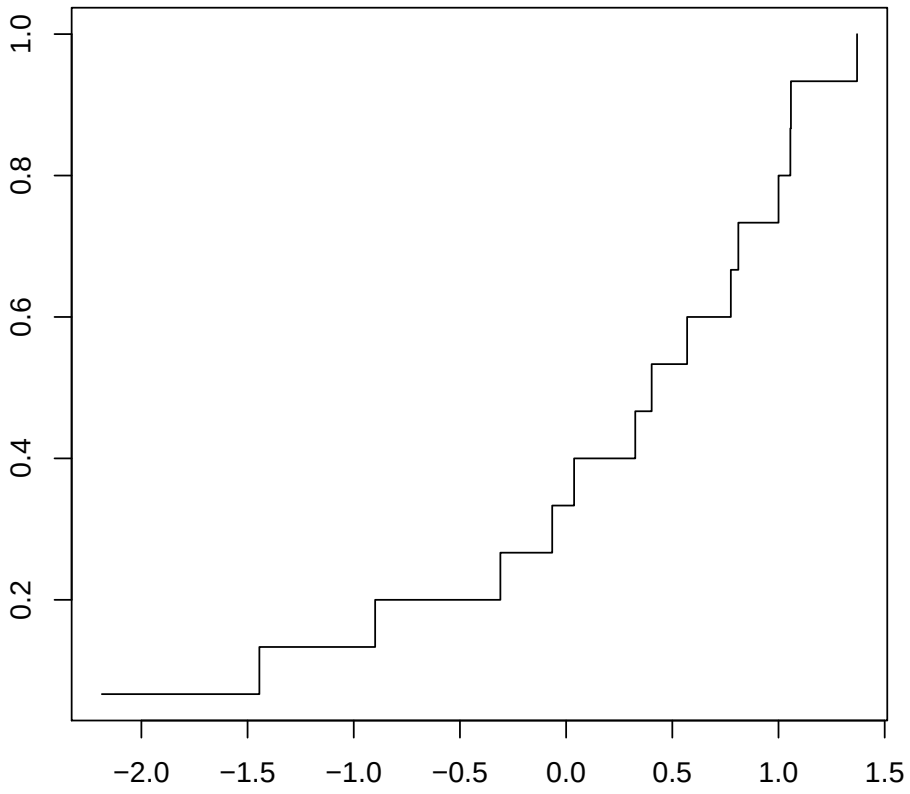
```
 [1]  0.81  0.57  0.78  1.06  1.06 -1.44  0.04 -0.90 -2.19 -0.31  1.00  0.33
[13] -0.07  0.40  1.37
```

cuyos datos ordenados son

```
> round(sort(x), 2)
```

```
 [1] -2.19 -1.44 -0.90 -0.31 -0.07  0.04  0.33  0.40  0.57  0.78  0.81  1.00
[13]  1.06  1.06  1.37
```

la función de distribución empírica aparece en la figura siguiente:



Sin embargo, la función de distribución empírica anterior tiene problemas prácticos que no entraremos a justificar por falta de espacio y tiempo.

Por dicho motivo, los paquetes estadísticos y entre ellos R "refinan" la definición de la función de distribución empírica de la forma siguiente:

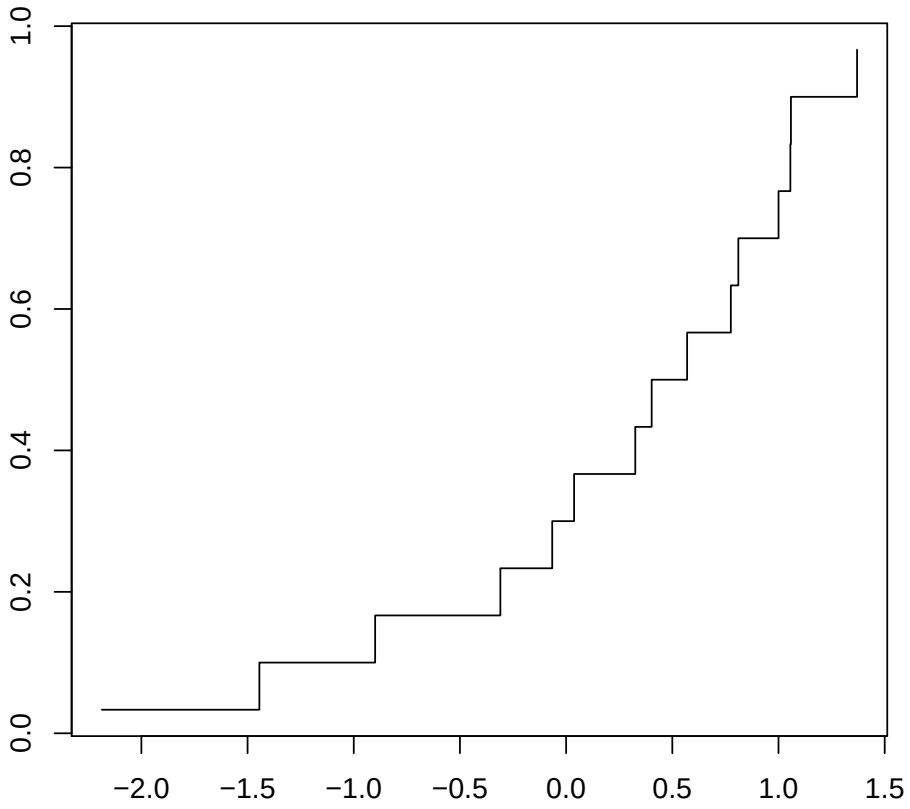
a) Si $n \leq 10$, dicha función dado un valor cualquiera x se define como:

$$F_E(x) = \begin{cases} 0, & \text{si } x < x_{(1)}, \\ \frac{1-3/8}{n+1/4}, & \text{si } x_{(1)} \leq x < x_{(2)}, \\ \cdots & \cdots \\ \frac{k-3/8}{n+1/4}, & \text{si } x_{(k)} \leq x < x_{(k+1)}, \\ \cdots & \cdots \\ \frac{n-3/8}{n+1/4}, & \text{si } x \geq x_{(n)}. \end{cases}$$

b) Si $n > 10$, dicha función dado un valor cualquiera x se define como:

$$F_E(x) = \begin{cases} 0, & \text{si } x < x_{(1)}, \\ \frac{1-1/2}{n}, & \text{si } x_{(1)} \leq x < x_{(2)}, \\ \cdots & \cdots \\ \frac{k-1/2}{n}, & \text{si } x_{(k)} \leq x < x_{(k+1)}, \\ \cdots & \cdots \\ \frac{n-1/2}{n}, & \text{si } x \geq x_{(n)}. \end{cases}$$

El gráfico de la función de distribución empírica refinada por `R` aparece en el gráfico siguiente para los valores anteriores:



La función de distribución empírica se interpreta como una aproximación de la función de distribución de la muestra de datos $x_1, \ldots, x_n$.

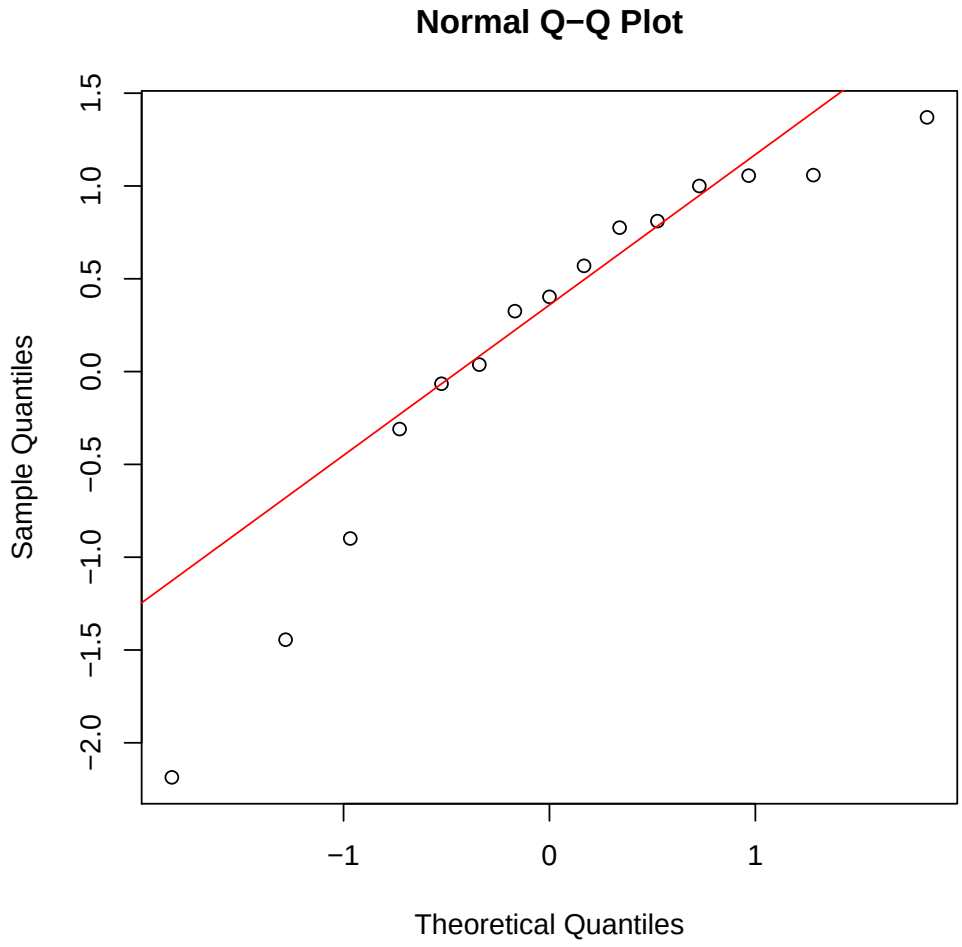
Dicha función se puede usar para "comprobar" si la muestra de datos anterior proviene de una variable aleatoria continua X con función de distribución F_X. ¿Cómo se lleva a cabo dicha comprobación?

Si nos fijamos en la definición de la función de distribución "refinada" por `R` (para fijar ideas, suponemos $n > 10$), tenemos que $F_E(x_{(k)}) = \frac{k-1/2}{n}$, para $k = 1, \ldots, n$. Si la muestra de datos proviene de una variable X cuya función de distribución es F_X, tendremos que $F_E \approx F_X$ y, por tanto, $F_X^{-1}\left(\frac{k-1/2}{n}\right) \approx x_{(k)}$. De aquí que si realizamos un gráfico de los puntos $\left(F_X^{-1}\left(\frac{k-1/2}{n}\right), x_{(k)}\right)$, dichos puntos al ser la coordenada x aproximadamente igual a la coordenada y deberían estar en una recta. Las coordenadas x de los puntos anteriores, $F_X^{-1}\left(\frac{k-1/2}{n}\right)$ se denominan cuantiles teóricos y las coordenadas y, $x_{(k)}$, cuantiles empíricos. Si en el gráfico anterior, los puntos no están aproximadamente alineados, significará que la muestra de de datos no proviene de la variable continua X.

La técnica anterior se denomina hacer un `qq-plot` de los datos.

En el caso en que X sea normal, se realiza lo que se denomina un `qqnorm`.

En los puntos considerados si usamos la instrucción `qqnorm(x)`, `R` nos da el gráfico siguiente:



Comprobamos que efectivamente los cuantiles teóricos y empíricos están alineados y por tanto, podemos deducir que los valores anteriores provienen de una distribución normal.

Si usamos las instrucciones `qqnorm(x)[[1]]` y `qqnorm(x)[[2]]`, `R` nos da los cuantiles teóricos y empíricos respectivamente.

Usando la instrucción `lm` que vistéis en Matemáticas I, podemos ver lo bueno que es el ajuste:

```
> summary(lm(qqnorm(x)[[2]] ~ qqnorm(x)[[1]]))

Call:
lm(formula = qqnorm(x)[[2]] ~ qqnorm(x)[[1]])

Residuals:
    Min      1Q  Median      3Q     Max
-0.5796 -0.2390  0.1370  0.2383  0.3211

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)     0.16681    0.08584   1.943    0.074 .
qqnorm(x)[[1]]  0.96694    0.08956  10.797 7.34e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3325 on 13 degrees of freedom
Multiple R-squared: 0.8997,        Adjusted R-squared: 0.8919
F-statistic: 116.6 on 1 and 13 DF,  p-value: 7.341e-08
```

Tenemos un ajuste del 89.97 %.

2. Comprobación de igualdad de varianzas

Otra de las hipótesis que hemos de comprobar antes de aplicar el test ANOVA es si todas las muestras de los k niveles tienen la misma varianza.

Para ello, usaremos el denominado **test de Bartlett**.

Para fijar ideas, supongamos que tenemos k muestras correspondientes a los k niveles de nuestro contraste ANOVA y queremos ver si las varianzas de las k muestras son la misma o no. Nuestra tabla de datos será:

Nivel del factor			
1	2	$\cdots$	k
X_{11}	X_{21}	$\cdots$	X_{k1}
X_{12}	X_{22}	$\cdots$	X_{k2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_{1n_1}	X_{2n_2}	$\cdots$	X_{kn_k}

El contraste que queremos realizar es:

$$\left.\begin{array}{ll} H_0: & \sigma_1^2 = \cdots = \sigma_k^2, \\ H_1: & \exists i,j \mid \sigma_i^2 \neq \sigma_k^2, \end{array}\right\}$$

donde σ_i^2 representaría la varianza de la muestra correspondiente al nivel k.

Para realizar el contraste anterior usaremos el llamado estadístico de Barlett:

$$X^2 = \frac{(N-k)\ln(s_p^2) - \sum_{i=1}^k (n_i-1)\ln(s_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^k \left(\frac{1}{n_i-1}\right) - \frac{1}{N-k}\right)},$$

donde $N = \sum_{i=1}^k n_i$, $s_p^2 = \frac{\sum_i (n_i-1)s_i^2}{N-k}$ y $s_i^2 = \frac{\sum_j (X_{ij} - \overline{X}_{i\bullet})^2}{n_i-1}$.

Si la hipótesis nula es cierta, el estadístico anterior tiene aproximadamente la distribución χ^2_{k-1}. Por tanto, rechazaremos la hipótesis nula o la igualdad de varianzas a un nivel de significación α si $X^2 > \chi^2_{k-1,1-\alpha}$ y la aceptaremos en caso contrario. El p-valor del contraste es: $p = p(\chi^2_{k-1} > X^2)$.

Si hacemos el contraste del test de Bartlett al ejemplo de los apuntes del crecimiento del microorganismo, tendremos lo siguiente:

```
> ej_anova <- matrix(c(62.6, 50.9, 45.5, 29.5, 24.9, 59.6, 44.3,
+     41.1, 22.8, 17.2, 64.5, 47.5, 29.8, 19.2, 7.8, 59.3, 49.5,
+     38.3, 20.6, 10.5, 58.6, 48.5, 40.2, 29.2, 17.8, 64.6, 50.4,
+     38.5, 24.1, 22.1, 50.9, 35.2, 30.2, 22.6, 22.6, 56.2, 49.9,
+     27, 32.7, 16.8, 52.3, 42.6, 40, 24.4, 15.9, 62.8, 41.6, 33.9,
+     29.6, 8.8), 50, 1)
> ej_anova <- cbind(ej_anova, rep(seq(1:5), 10))
> ej_anova <- as.data.frame(ej_anova)
> niveles <- as.factor(ej_anova[, 2])
> (bartlett.test(ej_anova[, 1] ~ niveles))

        Bartlett test of homogeneity of variances

data:  ej_anova[, 1] by niveles
Bartlett's K-squared = 1.0701, df = 4, p-value = 0.899
```

Al tener un p valor de 0.899 aceptamos la hipótesis nula y concluimos que las varianzas de los 5 niveles son iguales.