Análisis de la Varianza (ANOVA).

1. Planteamiento del problema.

Se desea contrastar si las medias de p poblaciones independientes son todas iguales o si existen diferencias entre al menos dos de ellas:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_p \\ H_1 : \exists \ \ i, j : \mu_i \neq \mu_j \end{cases}$$

Datos disponibles:

Se observan p poblaciones en las que se mide una variable respuesta Y. De cada población i se extrae una muestra aleatoria de tamaño n_i, $i = 1, 2, \ldots, p$ (los n_i pueden ser distintos entre sí). Los datos disponibles son, por tanto, de la forma:

Muestra 1	Muestra 2	...	Muestra i	...	Muestra p
Y_{11}	Y_{21}	...	Y_{i1}	...	Y_{p1}
Y_{12}	Y_{22}	...	Y_{i2}	...	Y_{p2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{1n_1}	Y_{2n_2}	...	Y_{in_i}	...	Y_{pn_p}

Hipótesis del modelo:

1. El valor observado de la variable Y en el objeto j de la población i sigue un modelo de la forma:
$$Y_{ij} = \mu + \beta_i + \varepsilon_{ij}, \ \ i = 1, \ldots, p$$

2. Las variables ε_{ij} (residuos) son independientes y con la misma distribución $N(0, \sigma_\varepsilon)$ $\forall i, j$. Ello significa que, en la población i:
$$Y_i \approx N(\mu + \beta_i, \sigma_\varepsilon)$$

Interpretación del modelo:

- μ representa la media global de Y sobre todas las poblaciones, esto es, $\mu = E[Y]$.
- $\mu_i = \mu + \beta_i$ representa la media de Y en la población i. Por tanto $\beta_i = \mu_i - \mu$ representa la diferencia media de los valores de Y en la población i con respecto a la media global.
- $\varepsilon_{ij} = Y_{ij} - \mu - \beta_i = Y_{ij} - \mu_i$ representa lo que se aparta cada objeto de la media de su población.
- $Y_i \approx N(\mu + \beta_i, \sigma_\varepsilon)$ indica que, en cada población i, la variable Y tiene una distribución normal con media $\mu + \beta_i$ y desviación típica σ_ε^2 común para todas las poblaciones (esta condición de varianza común recibe el nombre de *homoscedasticidad*).

Estimación del modelo:

En cada muestra estimamos su media y varianza:

	Muestra 1	Muestra 2	...	Muestra i	...	Muestra p
Media	$\bar{Y}_1$	$\bar{Y}_2$	...	$\bar{Y}_i$	...	$\bar{Y}_p$
Varianza	s_1^2	s_2^2	...	s_i^2	...	s_p^2

Asimismo, estimamos:

$\hat{\mu} = \bar{Y} = \dfrac{\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}}{N} = \dfrac{\sum_{i=1}^{p} n_i \bar{Y}_i}{N}$	$\bar{Y}$ es la media global de todos los datos, $N = n_1 + n_2 + \cdots + n_p$
$\hat{\beta}_i = \hat{\mu}_i - \hat{\mu} = \bar{Y}_i - \bar{Y}$	$\hat{\beta}_i$ es el estimador del *efecto* de pertenecer a la población $i - ésima$.
$S_R^2 = \dfrac{\sum_{i=1}^{p}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{N-p} = \dfrac{\sum_{i=1}^{p} (n_i - 1) S_i^2}{N-p}$	S_R^2 es una estimación conjunta de σ_ε^2, la *varianza residual* dentro de cada población. Esta varianza se supone que es la misma para todas las poblaciones.
$S_E^2 = \dfrac{\sum_{i=1}^{p} n_i (\bar{Y}_i - \bar{Y})^2}{p-1}$	S_E^2 es una estimación de la variabilidad entre las poblaciones. Se suele conocer como *varianza explicada* por el factor que define las poblaciones.

Estadístico de contraste:

$$F_{\text{exp}} = \frac{S_E^2}{S_R^2} = \frac{\frac{1}{p-1}\sum_{i=1}^{p} n_i(\bar{Y}_i - \bar{Y})^2}{\frac{1}{N-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}$$

Regla de decisión:

Si H_0 es cierta, el estadístico F_{exp} sigue una distribución F de Fisher-Snedecor con $p-1$ y $N-p$ grados de libertad.

- Si $F_{exp} \leq F_{p-1,N-p,\alpha} \Longrightarrow$Aceptar $H_0 : \mu_1 = \mu_2 = ... = \mu_p$
- Si $F_{exp} > F_{p-1,N-p,\alpha} \Longrightarrow$Rechazar H_0. Por tanto $\exists\ i, j : \mu_i \neq \mu_j$

Tabla del análisis de la varianza.

Los valores necesarios para la realización del contraste anterior suelen disponerse en una tabla como la siguiente:

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Medias Cuadradas	F_{exp}
Factor (variabilidad entre grupos)	$\sum_{i=1}^{p} n_i(\bar{Y}_i - \bar{Y})^2$	$p-1$	$S_E^2 = \frac{1}{p-1}\sum_{i=1}^{p} n_i(\bar{Y}_i - \bar{Y})^2$	S_E^2/S_R^2
Residuos (variabilidad dentro de los grupos)	$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$	$N-p$	$S_R^2 = \frac{1}{N-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$	

2. Validación de los supuestos del Análisis de la Varianza.

Para validar el resultado del contraste deben comprobarse los supuestos del modelo ANOVA:

- *Homoscedasticidad:* debe comprobarse que las varianzas de las p poblaciones o grupos son iguales (o al menos que no existen diferencias significativas entre ellas). Para ello,

llamando σ_k^2 a la varianza de Y en la población k, debe realizarse el contraste:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2 \\ H_1 : \exists\, i, j : \sigma_i^2 \neq \sigma_j^2 \end{cases}$$

Este contraste puede llevarse a cabo mediante el test de Bartlett o el test de Levene, siendo el segundo más robusto que el primero frente a la ausencia de normalidad.

- *Normalidad de los residuos:* Los residuos del modelo deben seguir una distribución normal (en caso contrario, la distribución F no sería adecuada para el contraste). El test de Shapiro-Wilk permite contrastar si los residuos son normales.

3. Contrastes a posteriori: Método de Scheffé.

Si en el contraste de la F del análisis de la varianza se acepta la existencia de diferencias significativas entre algunas de las medias, resulta de interés determinar qué poblaciones tienen medias diferentes y cuál es la magnitud de dichas diferencias. Para contrastar si las medias de las poblaciones r y s son iguales ó distintas:

$$\begin{cases} H_0 : & \mu_r - \mu_s = 0 \\ H_1 : & \mu_r - \mu_s \neq 0 \end{cases}$$

puede utilizarse el siguiente test debido a Scheffe:

$$\text{Si } \left| \frac{\bar{Y}_r - \bar{Y}_s}{S_R \sqrt{(p-1) \cdot \left(\frac{1}{n_r} + \frac{1}{n_s}\right)}} \right| \leq \sqrt{F_{p-1,N-p,\alpha}} \Longrightarrow \text{Aceptar } H_0$$

$$\text{En caso contrario} \Longrightarrow \text{Rechazar } H_0$$

Además un intervalo de confianza para $\mu_r - \mu_s$ es:

$$\mu_r - \mu_s \in \left[\bar{Y}_r - \bar{Y}_s \pm S_R \sqrt{(p-1) \cdot F_{p-1,N-p,\alpha} \cdot \left(\frac{1}{n_r} + \frac{1}{n_s}\right)} \right]$$

4. Análisis de la varianza con R

Datos.

Los datos deben disponerse en dos variables:

- La variable que define los grupos o poblaciones, que llamaremos grupo. Esta variable habrá de ser de tipo factor.
- La variable respuesta Y, que llamaremos y.

Así pues, el *data.frame* que contiene los datos será usualmente de la forma:

grupo	y
1	y_{11}
⋮	⋮
1	y_{1n_1}
2	y_{21}
⋮	⋮
2	y_{2n_2}
⋮	⋮
p	y_{pn_p}

Tabla del análisis de la varianza.

La tabla del análisis de la varianza, con la variabilidad explicada, la variabilidad residual y el test F con su correspondiente p-valor se obtiene mediante:

```
> adeva = aov(y ~ grupo)
> summary(adeva)
```

Si $p - valor \geq \alpha$ se acepta $H_0 : \mu_1 = \mu_2 = ... = \mu_p$; en caso contrario concluimos H_1, esto es $\exists\, i, j : \mu_i \neq \mu_j$.

Estimación de las medias y efectos por grupo.

Podemos estimar la media μ_i de cada grupo mediante:

```
> model.tables(adeva, "means")
```

Asimismo, si estamos interesados en la estimación de los efectos $\hat{\beta}_i = \hat{\mu}_i - \hat{\mu}$, podemos obtenerla directamente mediante:

```
> model.tables(adeva, "effects")
```

Contrastes de homoscedasticidad.

El método más conveniente para contrastar la homoscedasticidad es el test de Levene. Para utilizarlo hay que cargar previamente la librería *car*:

```
> require(car)
> levene.test(y ~ grupo)
```

(**Nota:** en versiones antiguas del paquete *car* el test de Levene se ejecuta mediante una sintaxis ligeramente distinta, *leveneTest()*).

Cuando los residuos tienen distribución normal, se puede usar el test de Bartlett, que no requiere ninguna librería:

```
> bartlett.test(y ~ grupo)
```

En ambos casos:

Si $p - valor \geq \alpha$ se acepta $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2$; en caso contrario concluimos H_1, esto es $\exists\, i, j : \sigma_i^2 \neq \sigma_j^2$

Contraste de normalidad.

La normalidad de los residuos se contrasta mediante:

```
> shapiro.test(residuals(adeva))
```

Si $p - valor \geq \alpha$ se acepta que los residuos siguen una distribución normal.

Tests de Scheffé.

Para realizar el test de Scheffé hay que cargar la librería *agricolae*:

```
> require(agricolae)
> scheffe.test(adeva, "grupo", alpha = 0.05)
```

Esta función contrasta todos los pares de variables e indica entre cuáles existen diferencias significativas.

Nótese que la variable *grupo* debe especificarse entre comillas. Asimismo, si se va a utilizar 0.05 como nivel de significación, no es preciso especificarlo, ya que se toma por defecto.

Alternativas al ANOVA: Transformación de los datos o Test de Kruskal-Wallis.

En caso de que *no se verifiquen* las condiciones para la aplicación del análisis de la varianza –bien sea porque falle la normalidad, bien porque falle la heterocedasticidad– se pueden probar las siguientes alternativas:

1. Transformar la variable respuesta a logaritmo, raiz cuadrada, inversa o cuadrado y comprobar si la variable transformada cumple las condiciones de aplicación del ANOVA. En tal caso proceder con el ANOVA sobre los datos transformados. La sintaxis para estas transformaciones es:

```
> ytransf = log(y)
> ytransf = sqrt(y)
> ytransf = 1/y
> ytransf = y^2
```

2. En caso de que ninguna transformación produzca el resultado buscado, utilizar el *test de Kruskal-Wallis:*

```
> kruskal.test(y ~ grupo)
```