Tema 8: Análisis de la varianza

Estadística. Grado en Ciencias del Mar

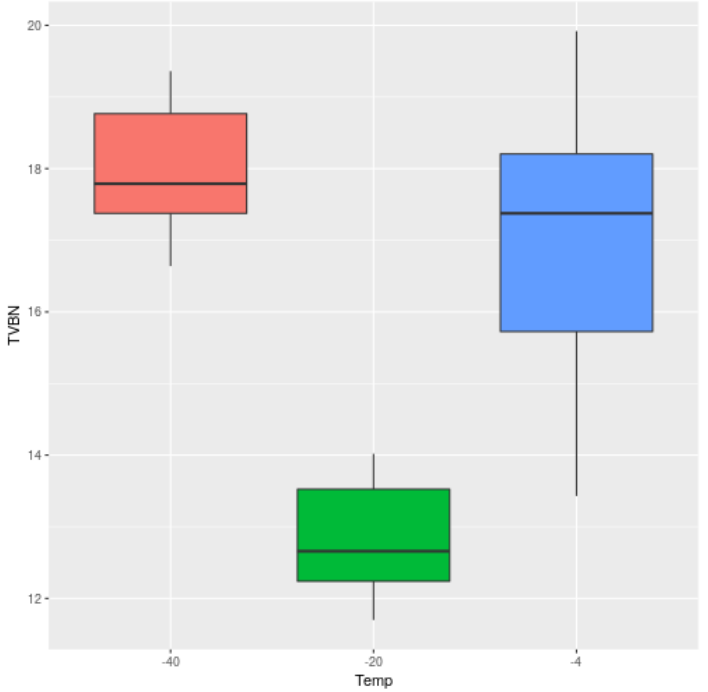
Planteamiento del problema: Ejemplo

Las condiciones de conservación del pescado se evalúan a través de la concentración de TVBN (*Total Volatile Base Nitrogen*). A mayor concentración de este elemento, peor es el estado de conservación de la pieza. Con objeto de determinar la temperatura que produce la menor concentración de TVBN, se eligen al azar 30 atunes recién pescados, todos de idéntico peso y características generales. Se separan en tres grupos de 10 piezas cada uno. El primer grupo se congela a -4ºC, el segundo a -20ºC y el tercero a -40ºC. La tabla adjunta muestra la concentración de TVBN en cada pieza después de 2 semanas de congelación.

Temperatura	-4ºC	-20ºC	-30ºC
	18.3	11.7	16.64
	15.92	12.87	17.83
	18.71	11.77	19.01
	17.92	12.23	17.33
	15.66	13.62	17.06
	17.14	13.24	18.04
	15.21	14.02	17.51
	19.92	13.66	19.11
	17.61	12.27	17.75
	13.43	12.45	19.36
Media (Sd)	**17.96 (0.92)**	**12.78 (0.82)**	**16.98 (1.92)**

Planteamiento del problema: Ejemplo

Podemos representar estos datos mediante un boxplot:



Obviamente las medias muestrales son distintas:

	-4ºC	-20ºC	-30ºC
Media (sd)	17.96 (0.92)	12.78 (0.82)	16.98 (1.92)

Pero: **¿Podemos deducir de estos datos que existe evidencia suficiente de que las medias poblacionales difieren entre sí?**

Planteamiento del problema: Ejemplo

En definitiva, nos planteamos el contraste de hipótesis:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \mu_3 \\ H_{1:} & \exists i, j: \mu_i \neq \mu_j \end{cases}$$

siendo μ_1, μ_2 y μ_3 las respectivas medias poblacionales de TVBN a cada una de las tres temperaturas.

Análisis de la varianza.

En el caso más elemental, se denomina **análisis de la varianza** *con un factor de variación* al contraste:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_p \\ H_{1:} & \exists i, j: \mu_i \neq \mu_j \end{cases}$$

- El **factor de variación** es la variable que define las poblaciones que se comparan. El factor de variación es siempre una variable categórica, y sus valores se denominan **niveles**.
- En nuestro ejemplo el factor de variación es la temperatura, que toma tres niveles (-4ºC, -20ºC y -40ºC)
- Nótese que aunque el contraste se denomina *análisis de la varianza* es realmente un contraste de comparación de medias. Pronto veremos la razón de este nombre.

Análisis de la varianza.

- ¿Por qué no realizar varias comparaciones mediante el test de la t de Student que ya hemos visto?
- Imaginemos que p (el número de medias a comparar) es $p = 7$. Si queremos comparar todas las medias dos a dos tendríamos que hacer $\binom{7}{2} = 21$ contrastes.
- Si cada contraste se realiza con un nivel de significación $\alpha = 5\%$, ello significa que de cada 100 contrastes en los que H_0 fuese cierta, dicha hipótesis se rechazaría en 5; o lo que es lo mismo, se rechazaría incorrectamente en uno de cada 20 contrastes.
- Por tanto, si $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$ es cierta, al hacer los 21 tests para comparar estas 7 medias dos a dos, **¡podemos estar casi seguros de que en alguno de los contrastes decidiríamos que las dos medias implicadas son distintas!**, lo que nos obligaría a rechazar H_0.
- Dicho de otra forma, la probabilidad de rechazar H_0 siendo cierta se "infla" cuando se realizan muchos contrastes de la t de Student para comparar las distintas medias dos a dos.

Análisis de la varianza: Procedimiento

Los datos disponibles para el análisis de la varianza normalmente son de la forma:

Grupo 1	Grupo 2		Grupo i		Grupo p
y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{p1}
y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{p1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
y_{1n_1}	y_{2n_2}	$\cdots$	y_{in_i}	$\cdots$	y_{pn_p}

- Tenemos p grupos
- En el grupo i hay n_i observaciones
- Llamaremos $\bar{y}_i$ a la media del grupo i, e $\bar{y}$ a la media de todos los valores observados.

Análisis de la varianza: Procedimiento

Sean:

- $N = \sum_{i=1}^{p} n_i$ el número total de observaciones.
- $S_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_i - 1}$ la varianza de los valores dentro del grupo i
- $S_R^2 = \frac{\sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{N - p} = \frac{\sum_{i=1}^{p} (n_i - 1) S_i^2}{N - p}$ la **Variabilidad dentro de los grupos**; esta medida es un promedio de las varianzas dentro de cada grupo.
- $S_E^2 = \frac{\sum_{i=1}^{p} n_i (\bar{y}_i - \bar{y})^2}{p - 1}$ la **variabilidad entre grupos**; esta cantidad mide la variabilidad entre las medias de los grupos.

Análisis de la varianza: Procedimiento

Sea σ_i^2 la varianza poblacional en el grupo i. Supondremos que se dan las siguientes condiciones:

- **Homoscedasticidad:** Todos los grupos tienen la misma variabilidad: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2 = \sigma^2$
- **Normalidad:** Dentro de cada grupo, la variable respuesta observada sigue una distribución normal: $Y_i \approx N(\mu_i, \sigma)$

Si se dan estas condiciones se puede probar que:

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

Si `$H_0 es cierta la variabilidad entre grupos S_E^2$ debe ser menor que la variabilidad dentro de los grupos $S_R^2$, y por tanto $F_{exp}$` debe ser un valor pequeño.

Análisis de la varianza: Procedimiento

Por tanto la regla de decisión consistirá en determinar cuál es el mayor valor que puede tomar F_{exp} por puro azar y rechazar H_0 si F_{exp} es mayor que dicho valor. Concretamente, si fijamos α como nivel de significación, la regla de decisión es:

- Si $F_{exp} \leq F_{p-1,N-p,\alpha} \Longrightarrow$ Aceptar $H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$
- Si $F_{exp} > F_{p-1,N-p,\alpha} \Longrightarrow$ Rechazar H_0 y concluir que $\exists \;\; i, j : \mu_i \neq \mu_j$

Análisis de la varianza: Ejemplo

En nuestro ejemplo de la conservación de atunes se tiene que:

Temp	-4ºC	-20ºC	-40ºC	GLOBAL
media	17.964	12.783	16.982	15.91
sd	0.919	0.823	1.923	
var	0.845	0.677	3.698	

Entonces:

- $$S_E^2 = \frac{\sum_{i=1}^{p} n_i(\bar{Y}_i - \bar{Y})^2}{p-1} = \frac{10(17.964-15.91)^2+10(12.783-15.91)^2+10(16.982-15.91)^2}{3-1} = 75.73$$
- $$S_R^2 = \frac{\sum_{i=1}^{p} (n_i-1)S_i^2}{N-p} = \frac{9 \cdot 0.845+9 \cdot 0.677+9 \cdot 3.698}{30-3} = 1.74$$

$$\left.\begin{array}{rl} F_{exp} = & \frac{S_E^2}{S_R^2} = \frac{75.73}{1.74} = 43.52 \\ F_{p-1,N-p,\alpha} = & F_{2,27,0.05} = 3.354 \end{array}\right\} \Rightarrow F_{exp} > F_{2,27,0.05} \Rightarrow \text{Se rechaza } H_0$$

Análisis de la varianza con R: Ejemplo

Para realizar el análisis de la varianza con R, los datos deben guardarse en dos columnas, una con la temperatura y otra con el valor de TVBN:

Como el p-valor (3.58e-09) es menor que el nivel de significación 0.05 concluimos que existe evidencia suficiente para rechazar `$H_0$`

Análisis de la varianza: Presentación de resultados

Normalmente, los resultados de un análisis de la varianza se presentan en una tabla análoga a la que ha mostrado R:

Fuente de variación	g.l.	Suma de cuadrados	Cuadrados medios	F	P
Factor	$p-1$	$\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2$	$S_E^2 = \frac{1}{p-1}\sum_{i=1}^{p} n_i(\bar{Y}_i - \bar{Y})^2$	S_E^2/S_R^2	
Residual	$N-p$	$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y - \bar{y}_i)^2$	$S_R^2 = \frac{1}{N-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$		

El p-valor se calcula como:

$$p = P\left(F_{p-1,N-p} > \frac{S_E^2}{S_R^2}\right)$$

Análisis de la varianza: Validación

Para validar el resultado de un análisis de la varianza hay que comprobar que se cumplen las condiciones en que se basa:

- Homoscedasticidad: se comprueba mediante el test de Levene, cuya hipótesis nula es que los datos son homoscedásticos.
- Normalidad: se comprueba mediante el test de Shapiro-Wilk, cuya hipótesis nula es que los datos son normales.

Ambos contrastes son laboriosos de realizar. Para llevarlos a cabo utilizaremos R y tomaremos la decisión basándonos en el p-valor de cada contraste:

- Si $p - valor < \alpha$ se rechaza H_0
- Si $p - valor \geq \alpha$, se acepta H_0

Veamos a continuación la aplicación de estos contrastes en nuestro ejemplo.

Análisis de la varianza: Validación con R

- **Homoscedasticidad:**

```
library(car)
leveneTest(TVBN~factor(Temp), data=atunes)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  3.2265 0.0554 .
##       27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el p-valor es mayor que 0.05 podemos aceptar la hipótesis de homoscedasticidad.

Análisis de la varianza: Validación con R

- **Normalidad:**

- Contraste gráfico

```
modelo=aov(TVBN~factor(Temp),dat
qqPlot(residuals(modelo))
```

- Contraste numérico

```
shapiro.test(residuals(modelo))
```

```
##
##     Shapiro-Wilk normality test
##
## data:  residuals(modelo)
## W = 0.9783, p-value = 0.7788
```

Como el p-valor es mayor que 0.05 podemos aceptar la hipótesis de normalidad

Análisis de la varianza: Contrastes a posteriori

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 **Test de Scheffe**

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

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

Análisis de la varianza: Contrastes a posteriori

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

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

Análisis de la varianza: Test de Scheffe con R

```
library(DescTools)
ScheffeTest(modelo)
```

```
##
##   Posthoc multiple comparisons of means : Scheffe Test
##     95% family-wise confidence level
##
## $`factor(Temp)`
##           diff    lwr.ci     upr.ci    pval
## -20--40 -4.881 -6.434821 -3.3271794 5.5e-08 ***
## -4--40  -0.982 -2.535821  0.5718206  0.2788
## -4--20   3.899  2.345179  5.4528206 3.0e-06 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```