Análisis de la varianza de un factor

El **test t de 2 muestras** se aplica cuando se quieren comparar las medias de dos poblaciones con **distribuciones normales** con **varianzas iguales** y se observan **muestras independientes** para cada población. Ahora consideraremos una generalización para el caso en que se quieren comparar tres o más medias.

Ejemplo: En la tabla siguiente se muestran los resultados obtenidos en una investigación acerca de la estabilidad de un reactivo fluorescente en diferentes condiciones de almacenamiento.

Se conservaron tres muestras en cada una de 4 condiciones. Supongamos (porque a veces puede ocurrir) que para una de las condiciones, la medición no pudo realizarse o se detectó una falla y fue eliminada. Los datos observados son:

Condiciones	Mediciones observadas (señales de fluorescencia)	Media Muestral
Recientemente preparada	102 100 101	101
Una hora en la oscuridad	101 101 104	102
Una hora con luz tenue	97 95 99	97
Una hora con luz brillante	92 94	93

Mirando los promedios muestrales se ven diferencias y nos preguntamos si las condiciones de almacenamiento no influyeron sobre la fluorescencia de las muestras (ésta será nuestra H_0), ¿cuál es la probabilidad de que por simple azar se observen diferencias de esta magnitud entre las medias muestrales?

Para generalizar podemos pensar que observamos k muestras (en el ejemplo k=4). Suponemos el siguiente modelo:

Modelo de k muestras normales independientes con varianzas iguales.

Muestra 1: $X_{11}, X_{12}, \ldots\ldots X_{1n_1}$ v. a. i.i.d $N(\mu_1, \sigma^2)$

.....

Muestra i: $X_{i1}, X_{i2}, \ldots\ldots X_{in_i}$ v. a. i.i.d $N(\mu_i, \sigma^2)$

.....

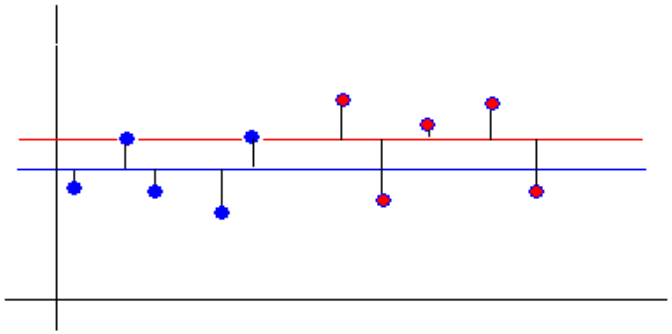
Muestra k: $X_{k1}, X_{k2}, \ldots\ldots X_{kn_k}$ v. a. i.i.d $N(\mu_k, \sigma^2)$

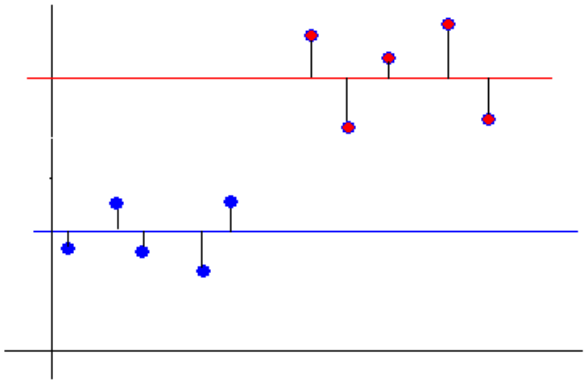
y asumimos que **las v. a. de una muestra son independientes de las v. a. de otra muestra**.

Llamaremos $\overline{X}_i$ y s_i^2 a la media y la varianza muestrales de la muestra i = 1,2, ...,k.

Vamos a testear:

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ vs. H_1: existen i y j para los cuales $\mu_i \neq \mu_j$





Parece natural proponer un estimador de σ^2 basado en un promedio ponderado de las varianzas de cada muestra s_i^2, tal como se hacemos con el s_P^2 cuando comparamos dos muestras. Se puede demostrar que el mejor estimador insesgado de σ^2 bajo el modelo anterior es:

$$s_p^2 = \frac{(n_1 - 1)*s_1^2 + \ldots + (n_k - 1)*s_k^2}{n_1 + \ldots + n_k - k} = \frac{\sum_{i=1}^{k}(n_i - 1)*s_i^2}{n - k} = \frac{SS_W}{n - k} \qquad (1)$$

En la última expresión hemos llamado

$$n = \sum_{i=1}^{k} n_i$$

al número total de observaciones.

Bajo la hipótesis nula:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$

todas las observaciones tienen la misma distribución.

Llamemos

$$\overline{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{n} = \frac{\sum_{i=1}^{k} n_i \overline{X}_i}{n}$$

a la media general de todas las observaciones.

El estadístico para el test óptimo para este problema, tiene al estimador de la varianza (dado por (1)) en el denominador y una medida de las diferencias (similar a la variancia) entre las medias de las distintas muestras en el numerador. Esta medida es:

$$\frac{\sum_{i=1}^{k} n_i (\overline{X}_i - \overline{X})^2}{k-1} = \frac{SS_B}{k-1} \qquad (2)$$

El estadístico del test se obtiene dividiendo (2) sobre (1):

$$F = \frac{\left(\sum_{i=1}^{k} n_i (\overline{X}_i - \overline{X})^2 \right) / (k-1)}{s_p^2} = \frac{SS_B / k-1}{SS_W / n-k} \qquad (3)$$

Si H_0 fuera cierta, el denominador y el numerador serían parecidos, por lo tanto el cociente sería cercano a 1.

Si las medias poblacionales no son todas iguales, como vimos en el gráfico, el numerador tiende a ser mayor que el denominador y por lo tanto, el cociente será mayor a 1.

Test *F*:

1er. paso: Calculo el estadístico $F = \dfrac{\left(\sum_{i=1}^{k} n_i (\overline{X}_i - \overline{X})^2\right)/(k-1)}{s_p^2} = \dfrac{SS_B / k-1}{SS_W / n-k}$

Nota: Si H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ es cierta, este estadístico tiene distribución F con **k-1** grados de libertad en el **numerador** y **n-k** grados de libertad en el **denominador**.

¿De dónde surgen los grados de libertad? Se puede demostrar, que si se satisfacen los supuestos del análisis de varianza que hemos realizado, entonces:

Bajo H_0: i) $\dfrac{SS_W}{\sigma^2} \sim \chi^2_{n-k}$ ii) $\dfrac{SS_B}{\sigma^2} \sim \chi^2_{k-1}$ y además son independientes.

2do. paso: Si $F > F_{k-1,n-k;\alpha}$, rechazamos H_0.

Con frecuencia los resultados del Análisis de Varianza se presentan una tabla como la que sigue:

Análisis de Varianza					
Fuente	**SS**	**gl**	**MS**	**F**	**Prob > F**
Between	SSB	k-1	MSB = SSB/k-1	MSB/MSW	
Within	SSW	n-k	MSW = SSW/n-k		
Total	SST	n-1	MST = SST/n-1		

Veamos como quedaría en nuestro ejemplo:

Fuente	gl	SS	MS	F	P
BETWEEN	3	122.182	40.7273	15.84	0.0017
WITHIN	7	18.000	2.57143		
TOTAL	10	140.182			

Rechazamos la hipótesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ al nivel 0.01, es decir las medias de la fluorescencia difieren significativamente a este nivel. O dicho de otro modo: concluimos que la media de la fluorescencia depende de las condiciones de almacenamiento.

La pregunta ahora es: ¿cuáles son las que difieren?

Comentarios sobre la “tabla del análisis de la varianza”.

Se puede demostrar que vale la siguiente igualdad:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2 \quad = \quad \sum_{i=1}^{k} n_i(\overline{X}_i-\overline{X})^2 \quad + \quad \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2$$

En la expresión anterior aparecen tres “sumas de cuadrados”:

“suma de cuadrados entre grupos” (SS_B: Between)
“suma de cuadrados dentro de grupos” (SS_W: Within)
“suma de cuadrados total” (SS_T: Total)

Suposiciones del modelo. Diagnóstico.

El test *F* ha sido deducido bajo el supuesto de que las k muestras aleatorias tienen distribución normal, con igual varianza y son independientes. Cuando el tamaño de la muestra de cada grupo es grande, el test *F* es válido en forma aproximada (el valor p calculado es aproximado) aunque la variable no tenga distribución exactamente normal.

En la práctica no es esperable que el modelo se cumpla exactamente, pero sí en forma aproximada. Al igual que con el test t, hay que analizar los datos para detectar si el modelo es aproximadamente cierto o si en cambio es falso.

Boxplots Paralelos

Cuando hay una cantidad suficiente de observaciones se pueden realizar boxplots paralelos de las observaciones originales por tratamiento.

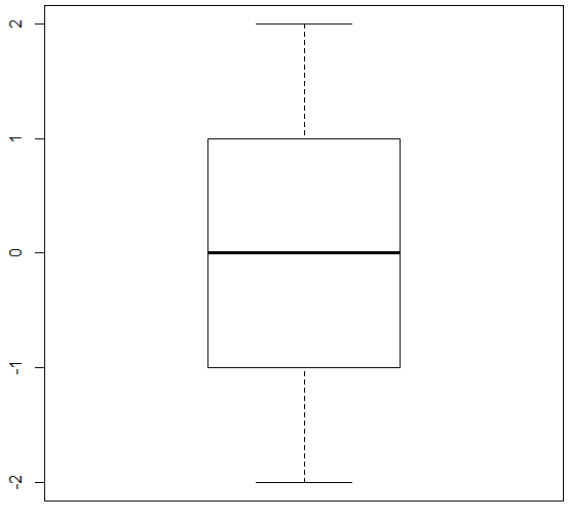
En el presente ejemplo, hay solo 3 y hasta 2 observaciones por casilla, con lo cual no parece muy razonable este gráfico. En su lugar podemos realizar un boxplot de los residuos todos juntos.

Para cada observación, el residuos r_{ij} se calcula como:

$$r_{ij} = X_{ij} - \overline{X}_{i}$$

El siguiente gráfico muestra el boxplot correspondiente a los residuos del ejemplo de fluorescencia:

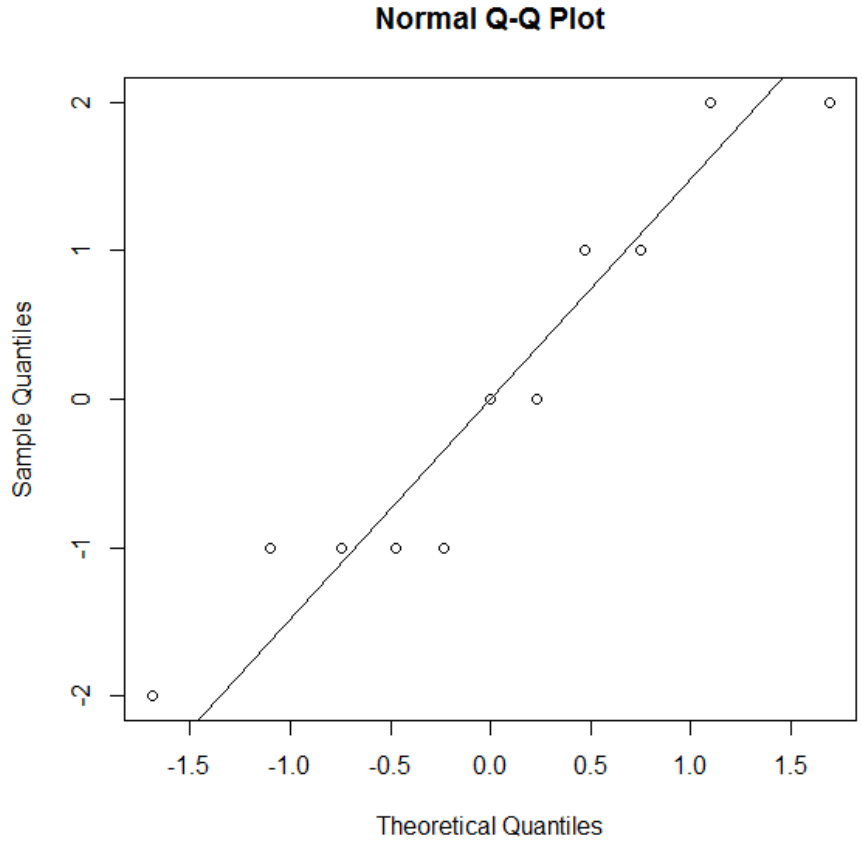
```
boxplot(salida$res)
```



Los residuos parecen tener una distribución simétrica y no se observan datos atípicos, por lo que no parece haber importantes apartamientos de la normalidad.

QQ-plot y Test de Shapiro-Wilk en nuestro ejemplo

```
qqnorm(salida$res)
qqline(salida$res)
```



```
shapiro.test(salida$res)

        Shapiro-Wilk normality test

data:  salida$res
W = 0.9081, p-value = 0.2315
```

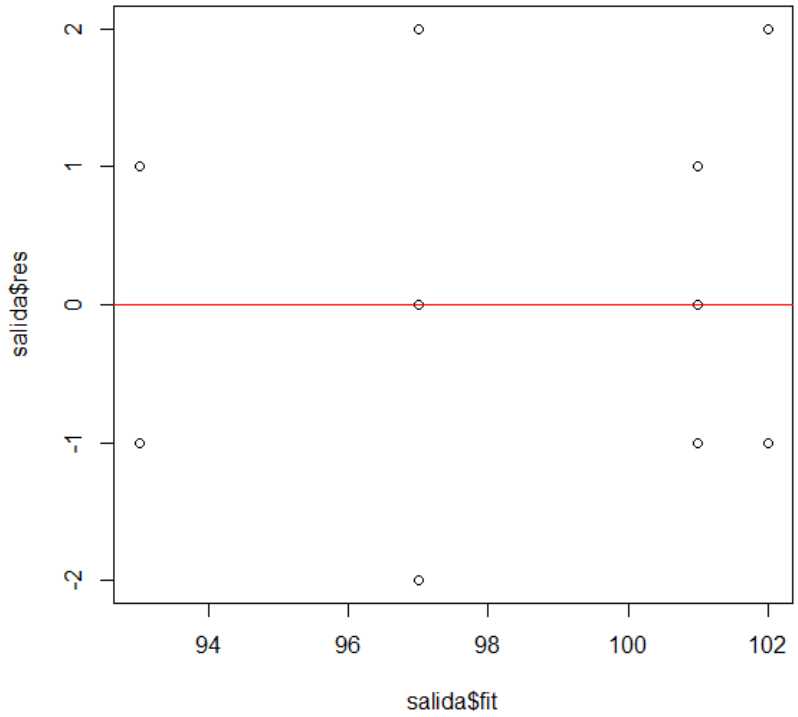
En nuestro ejemplo el estadístico del test de Shapiro-Wilk es 0.9081 y el p-valor correspondiente es de 0.2315, con lo cual no rechazamos el supuesto de normalidad.

Tests para estudiar si las varianzas son iguales

Para estudiar la suposición de igualdad de varianzas podemos graficar y también se pueden realizar algunos tests.

Respecto del gráfico podemos considerar un scatter-plot o diagrama de dispersión de los promedios muestrales versus los residuos.
En el ejemplo de Fluorescencia resultaría:

Se observan algunas diferencias en la dispersión de los residuos, pero no parece haber grandes apartamientos del supuesto de homoscedasticidad en este caso. Sin embargo, deberíamos aplicar un test para chequear este supuesto.



Respecto de tests existen algunas alternativas.

Consideremos el modelo

$$X_{ij} \sim N(\mu_i, \sigma_i^2) \quad (i=1,...,k;\ j=1,...,n_i) \text{ independientes}$$

y la hipótesis a testear será

$$H_0: \ \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

Hay varios tests. El más antiguo es el **test de Bartlett**. Se basa en un estadístico que tiene distribución aproximadamente $\boldsymbol{\chi^2_{k-1}}$ bajo H_0.

Si hay *k* muestras con tamaño n_i y varianzas de las muestras S_i^2 , como en nuestro problema, entonces estadístico de prueba de Bartlett, que se basa en una escala logarítimica, es:

$$X^2 = \frac{(n-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i-1)\ln(S_i^2)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(\frac{1}{n_i-1}) - \frac{1}{n-k}\right)}$$

El numerador tiende a dar valores grandes cuando las varianzas muestrales difieren mucho, por lo tanto se rechaza la hipótesis nula de igualdad de varianzas cuando el estadístico es grande.

La zona de rechazo es

$$X^2 > \chi^2_{k-1,\ \alpha}$$

bartlett.test(FLUOR,luz.f)

```
        Bartlett test of homogeneity of variances

data:  FLUOR and luz.f
Bartlett's K-squared = 0.7515, df = 3, p-value = 0.861
```

En nuestro ejemplo el estadístico del test de Bartlett es 0.7515 con un p-valor de 0.861, por lo tanto no rechazamos el supuesto de homogeneidad de varianzas

Sin embargo, este test tiene una alta sensibilidad a la falta de normalidad. Por esta razón, es necesario disponer de alguna alternativa más resistente a la falta de normalidad.

Un test que es poco sensible a la falta de normalidad es el **test de Modificado de Levene**. Para aplicarlo, primero se calculan

$$d_{ij} = | X_{ij} - \tilde{X}_i |$$

donde $\tilde{X}_i$ denota la mediana del tratamiento i . Luego se calcula el estadístico *F* del análisis de un factor a los d_{ij}.

Si la hipótesis H: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ es cierta y los n_i "no son muy pequeños", el estadístico tiene distribución aproximadamente F con k-1 y n-k grados de libertad. Esto permite aplicar un test aproximado de la hipótesis de igualdad de varianzas. Rechazamos la igualdad de varianzas si el estadístico toma un valor muy grande.

```
medians<-tapply(FLUOR,luz.f,median)
abs.dif<- abs(FLUOR-medians[luz.f])
summary(aov(abs.dif~luz.f))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
luz.f        3  0.667  0.2222   0.167  0.916
Residuals    7  9.333  1.3333
```

Como el p-valor = 0.916, no rechazamos el supuesto de homoscedasticidad.