# ANÁLISIS DE VARIANZA

Cuando es necesario hacer comparaciones entre tres o más medias muéstrales para determinar si provienen de poblaciones iguales utilizamos la técnica de análisis de varianza. Esta técnica se realiza utilizando la distribución de probabilidad F vista anteriormente. Para el uso de esta técnica es necesario seguir los siguientes supuestos:

1) Las poblaciones siguen una Distribución de Probabilidad Normal
2) Las poblaciones tienen desviaciones estándar ($\sigma$) iguales
3) Las muestras se seleccionan de modo independiente

La técnica del análisis de varianza descompone la variación total en dos componentes de variación llamados variación debida a los tratamientos y variación aleatoria.
Cuando estamos frente a un problema de análisis de varianza lo primero que debemos hacer es identificar en términos del problema lo siguiente:

***Variable dependiente o variable respuesta:*** Es la variable que nos interesa medir o respuesta que se va a estudiar para determinar el efecto que tiene sobre ella la variable independiente.

***Variable independiente o factor:*** Es la variable o factor que puede influenciar en la variabilidad de la respuesta o variable dependiente.

***Nivel o tratamiento del factor:*** Es un valor o condición del factor bajo el cual se observa la respuesta medible.

***Unidad experimental:*** Es el objeto (persona, animal o cosa) donde se aplica un determinado tratamiento, para obtener una medición de la variable respuesta.

***Error experimental:*** Es la variación que no se puede atribuir a un cambio de tratamiento; es decir, la que se produce por los factores extraños que pueden influir en la respuesta y que deben ser eliminados o controlados por el investigador.

***Aleatorización:*** Consiste en asignar en forma aleatoria los tratamientos a las unidades experimentales con el propósito de remover los posibles sesgos sistemáticos y neutralizar los efectos de todos aquellos factores externos que no se encuentran bajo el control del investigador, pero pueden estar presentes en el experimento.

Nosotros estudiaremos el diseño Completamente Aleatorizado con un solo factor o unifactorial.

Este modelo es apropiado en aquellas situaciones donde se tiene un solo factor o variable independiente con "c" niveles o tratamientos. En este diseño nos interesa probar las siguientes hipótesis:

$H_0$: Las medias de las ***c*** poblaciones son iguales
$H_1$: No todas las medias de las ***c*** poblaciones son iguales

Otra forma de platear las hipótesis es:

$H_0$: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_c$
$H_1$: Alguna de las medias difiere

También se puede plantear la hipótesis en función de los efectos de los tratamientos asi:

$H_0$: Los tratamientos no producen efecto
$H_1$: Alguno de los tratamientos produce efecto

$H_0$: $\alpha_1 = \alpha_2 = \alpha_3 = \ldots = \alpha_c$
$H_1$: Algún $\alpha$ es diferente

Para probar esta hipótesis se toma una muestra aleatoria de cada una de las ***c*** poblaciones y se examina la cantidad de variación dentro de cada una de estas muestras en relación con la cantidad de variación entre las muestras.

Si no se rechaza $H_0$ entonces las medias de las ***c*** poblaciones son iguales; es decir, no existe ningún efecto de los tratamientos sobre la variable respuesta.

Para realizar un contraste de hipótesis de este tipo debemos seguir los siguientes pasos:

1) Planteamiento de hipótesis

   Se pueden plantear en cualquiera de estas formas

   $H_0$: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_c$
   $H_1$: Alguna de las medias difiere

   $H_0$: Los tratamientos no producen efecto
   $H_1$: Alguno de los tratamientos produce efecto

   $H_0$: $\alpha_1 = \alpha_2 = \alpha_3 = \ldots = \alpha_c$
   $H_1$: Algún $\alpha$ es diferente

2) Se realizan los siguientes cálculos para obtener la tabla ANOVA

| TRATAMIENTO O NIVELES DEL FACTOR | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | … | j | … | C |
| $Y_{11}$ | $Y_{12}$ | … | $Y_{1j}$ | … | $Y_{1c}$ |
| $Y_{21}$ | $Y_{22}$ | … | $Y_{2j}$ | … | $Y_{2c}$ |
| ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| $Y_{i1}$ | $Y_{i2}$ | … | $Y_{ij}$ | … | $Y_{ic}$ |
| ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| $Yn_1 1$ | $Yn_2 2$ | … | $Yn_j j$ | … | $Yn_c c$ |
| T.1 | T.2 | … | T.j | … | T.c |
| n1 | n2 | … | nj | … | Nc |
| $\overline{Y}.1$ | $\overline{Y}.2$ | … | $\overline{Y}.j$ | … | $\overline{Y}.c$ |

Donde:

| | |
|---|---|
| T.j | son los totales de los tratamientos |
| nj | son los tamaños muéstrales |
| $\overline{Y}.j$ | son las medias muéstrales |

$$n = \sum_{j=1}^{c} n_j$$

$$T.. = \sum_{j=1}^{c} T_{.j}$$

$$SCT = \left[ \sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij})^2 \right] - \frac{(T..)^2}{n}$$

$$SCTr = \left[ \sum_{j=1}^{c} \frac{(T_{.j})^2}{n_j} \right] - \frac{(T..)^2}{n}$$

$$SCE = SCT - SCTr$$

3) Se fija la región crítica

Para un análisis de varianza vamos a considerar las pruebas de cola derecha utilizando una distribución F con grados de libertad $\upsilon_1 = (c-1)$ y $\upsilon_2 = (n-c)$

4) Se completa la Tabla ANOVA y se obtiene el estadístico de prueba

<table>
<tr><th colspan="5">TABLA ANOVA</th></tr>
<tr><td>Fuente de Variación</td><td>Grados de Libertad</td><td>Suma de Cuadrados</td><td>Cuadrados Medios</td><td>Fc</td></tr>
<tr><td>Tratamientos</td><td>c-1</td><td>SCTr</td><td>$CMTr = \frac{SCTr}{c-1}$</td><td rowspan="3">$Fc = \frac{CMTr}{CME}$</td></tr>
<tr><td>Error</td><td>n-c</td><td>SCE</td><td>$CME = \frac{SCE}{n-c}$</td></tr>
<tr><td>Total</td><td>n-1</td><td>SCT</td><td></td></tr>
</table>

5) Se toma la decisión contrastando el estadístico de prueba con el valor crítico

Una vez que se ha completado la tabla y se ha calculado el estadístico de prueba, se compara el Fc con el Ft.

Si Fc > que Ft rechazo $H_0$, lo que indica que alguna de las medias difiere o que alguno de los tratamientos está produciendo algún efecto.

Una vez que se ha rechazado la hipótesis nula (si se rechaza), se desea comparar todas las parejas de "a" medias de tratamientos, se puede aplicar la prueba de Tukey o DMS como sigue:

**PRUEBA DE TUKEY**

Se utiliza para diseños balanceados (todos los tratamientos tienen asignado el mismo número de elementos). Se utiliza el estadístico T

$$\boxed{T = q_{\alpha,c,n-c}\sqrt{\frac{CME}{r}}}$$

Se compara T vs la diferencia en valor absoluto de cada par de medias, si esta diferencia Excede a T, las medias son diferentes o iguales en caso contrario.

Otro método más conservador es el la **DIFERENCIA MÍNIMA SIGNIFICATIVA (DMS)**

$$\boxed{DMS = \sqrt{\frac{2(CME)F_{\alpha,1,n-c}}{r}}}$$

Para el caso de diseños no balanceados se utiliza el método **DMS alternativo** para comparar cada par de muestras

$$\boxed{DMS_{j,k} = \sqrt{\left[\frac{1}{r_j}+\frac{1}{r_k}\right](CME)F_{\alpha,c-1,n-c}}}$$

$r_j$ es el número de elementos asignados al tratamiento j
$r_k$ es el número de elementos asignados al tratamiento k

**EJERCICIOS**

1. Supóngase que el gerente de producción de una planta en la cual refabrica y envasa cereal en cajas de 368 gramos, considera sustituir una máquina antigua que afecta directamente la producción. Es más, supóngase que tres proveedores le han permitido utilizar sus equipos para efectuar pruebas, y cuyos precios de compra y contratos de mantenimiento son fundamentalmente iguales. Para tomar la decisión de compra, el gerente de producción decide llevar a cabo un experimento para determinar las diferencias más importantes entre las tres marcas de equipo en el tiempo promedio (en segundos), que necesitan los obreros para su producción. Se asignan en forma aleatoria quince operarios con experiencia, capacidad y edad similares, para recibir adiestramiento en una de las tres máquinas de modo que cada máquina tenga cinco operadores. Después de una capacitación adecuada, suficiente y práctica, el gerente de producción mide el tiempo (en segundos), que necesitan los operadores para trabajar con sus equipos respectivos. En la tabla adjunta se presentan los resultados de este experimento

**Máquina**

| I | II | III |
|---|---|---|
| 25,40 | 23,40 | 20,00 |
| 26,31 | 21,80 | 22,20 |
| 24,10 | 23,50 | 19,75 |
| 23,74 | 22,75 | 20,60 |
| 25,10 | 21,60 | 20,40 |

Con un nivel de significación de 1% ¿a qué conclusión debe llegar el gerente de producción?

2. Cuatro catalizadores que pueden afectar la concentración de un componente en una mezcla líquida de tres componentes están siendo investigados. Se obtienen las siguientes concentraciones:

**Catalizador**

| A | B | C | D |
|---|---|---|---|
| 58.2 | 56.3 | 50.1 | 52.9 |
| 57.2 | 54.5 | 54.2 | 49.9 |
| 58.4 | 57 | 55.4 | 50 |
| 55.8 | 55.3 | | 51.7 |
| 54.9 | | | |

Determinar si existe diferencia significativa con un nivel de significación de 1%

3. Para determinar si existe diferencia significativa en el nivel de Matemáticas de 4 grupos de estudiantes de Administración se realizó un examen aleatorio a 6 individuos por grupo. Determine cuáles son los grupos en los cuales existen diferencias a un 95% de nivel de confianza.

| A | B | C | D |
|---|---|---|---|
| 75 | 78 | 55 | 64 |
| 93 | 91 | 66 | 72 |
| 78 | 97 | 49 | 68 |
| 71 | 82 | 64 | 77 |
| 63 | 85 | 70 | 56 |
| 76 | 77 | 68 | 95 |

4. Las calificaciones en el examen a 18 empleados de tres unidades de negocio Se muestran a continuación:

| A | B | C |
|---|---|---|
| 85 | 71 | 59 |
| 75 | 75 | 64 |
| 82 | 73 | 62 |
| 76 | 74 | 69 |
| 71 | 69 | 75 |
| 85 | 82 | 67 |

Probar si no hay diferencia entre las unidades a un 5% de nivel de significancia.

5. Probar si hay diferencia en los tiempos de servicio de 4 unidades de negocio para el mismo servicio a un nivel de significancia del 5%.

| A | B | C | D |
|---|---|---|---|
| 5.4 | 8.7 | 11.1 | 9.9 |
| 7.8 | 7.4 | 10.3 | 12.8 |
| 5.3 | 9.4 | 9.7 | 12.1 |
| 7.4 | 10.1 | 10.3 | 10.8 |
| 8.4 | 9.2 | 9.2 | 11.3 |
| 7.3 | 9.8 | 8.8 | 11.5 |