# PRESENTACIÓN

Análisis de Varianza ANOVA es el onceavo fascículo, de una serie de guías de estudio en las que se desarrollan los temas de los programas de las asignaturas del área de Probabilidad y Estadística, así como temas selectos que complementan el aprendizaje de de esta disciplina. Tienen la característica de que el estudiante adquiera sólo aquella que trate el tema que necesite reforzar o el que sea de su propio interés.

Estas guías de estudio pretenden reorientar y actualizar el enfoque con el que se debe abordar el estudio de los métodos estadísticos, despertando la inquietud por aprender y resolver los problemas y casos planteados.

Cada guía integra el desarrollo del tema con ejercicios, casos de estudio y con la sección llamada Aprendiendo.com. En esta última sección se le proporciona al estudiante un ambiente interactivo, utilizando los recursos disponibles en Internet, de tal forma que los casos planteados los desarrolle en ambientes de aprendizaje que le permitan encontrarse con el conocimiento, "manipularlo", hacerlo suyo. Con esta filosofía se utilizan applets, sitios de internet con acceso a bases de datos reales, software de uso libre y en general los recursos de la Web 2.0, que se refieren a una segunda generación en la historia de la Web basada en comunidades de usuarios, que fomentan la colaboración y el intercambio ágil de información entre los mismos.

Nuestro reconocimiento a la Dirección General de Asuntos del Personal Académico de nuestra Casa de Estudios, que a través del Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza (PAPIME) ha apoyado nuestro proyecto "Implantación de un Laboratorio Virtual de Estadística y Elaboración de las Guías de Estudio con Soporte Multimedia" clave PE302709.

Los Autores

## Introducción

La observación y la experimentación son la base en que se apoya la investigación para el estudio de fenómenos, presentes en la naturaleza. Mediante la observación se describe el fenómeno con todas las circunstancias que lo rodean, pero no se puede atribuir sus efectos a una causa específica. Con la ayuda de la experimentación se estudian dichos fenómenos en forma controlada, aislando aquellos factores que pudieran enmascarar el efecto que ocasiona la causa de interés sobre dicho fenómeno.

En el estudio experimental de un fenómeno se plantea una hipótesis, para cuya prueba se diseña un procedimiento de ejecución, que se denomina diseño del experimento. Esta hipótesis, al ser probada requiere generalizarla por lo que es necesario asociarle una medida de probabilidad. Los diseños experimentales, cuya metodología es ampliamente usada en la investigación se realizan a fin de comparar los efectos de las diferentes variables experimentales independientes conocidas con el nombre de factores sobre una variable dependiente conocida como variable de respuesta. .

Un diseño experimental debe adecuarse al material experimental con que se cuenta y a las preguntas que desea contestar el investigador. Sus resultados se resumen en un cuadro de Análisis de Varianza y en una Tabla de Comparación de Medias de Tratamientos, que indican las diferencias entre dichas medidas. El análisis de varianza proporciona la variación de la variable de interés, en fuentes explicables por algunos factores y la variación debida a fuentes para las cuales el investigador no tiene control, no puede medir y no le es posible explicar o atribuir a algún factor en particular; variaciones que conforman el llamado error experimental. Por ejemplo: si se realiza un experimento en el cual se estudian 4 tipos de dietas para cerdos de engorda y se medie la ganancia de peso, la variación de dicha ganancia puede descomponerse en la fuete de variación atribuible a las diferentes dietas y alas fuentes desconocidas o error experimental

## Diseño Completamente Al Azar (Dca)

Este diseño consiste en la asignación de los tratamientos en forma completamente aleatoria a las unidades experimentales. Se entiende por unidades experimentales a los objetos sobre los cuales se hacen mediciones. Debido a su aleatorización irrestricta, es conveniente que se utilicen unidades experimentales lo más homogéneas posibles: animales de la misma edad, del mismo peso, similar estado fisiológico; parcelas de igual tamaño, etc., de tal manera que se minimice la magnitud del error experimental, ocasionado por la variación intrínseca entre las unidades experimentales.

## Aleatorización

Para ejemplificar el proceso de aleatorización irrestricta de los tratamientos a las unidades experimentales, supóngase que en la elaboración de las donas se quieren probar cuatro tipos de aceite, ya que parece que la cantidad de aceite absorbida por la masa depende del tipo de aceite. Se tiene un solo factor que es el aceite con cuatro niveles o tratamientos: aceite de cártamo (A), aceite de girasol (B), aceite de semilla de maíz (C) y aceite de soya (D). Para probar si existe efecto del tipo de aceite sobre la variable de respuesta, es decir, sobre la cantidad de aceite absorbido se planea realizar un diseño unifactorial o completamente al azar. Supongamos que se elabora la masa y que cada aceite se prueba en 6 porciones de 100 gramos cada una. Las unidades experimentales son las porciones de masa que son homogéneas, por lo que el diseño completamente al azar es adecuado.

El proceso de aleatorización puede realizarse de la siguiente manera:
*Paso 1.* Se numeran las unidades experimentales, es decir las 24 pociones de 100 gr de masa del 00 al 23
Paso 2. Utilizando una tabla de números aleatorios se seleccionan, números de dos dígitos comprendidos entre 00 y 23. Los primeros 6 seleccionados indican las porciones de masa que serán asignadas al tipo de aceite A, los siguientes 6 indican las porciones que serán asignadas al tipo de aceite B y así sucesivamente. La asignación aleatoria de las porciones a los tipos de aceite se muestra en la siguiente tabla.

| Tipo de Aceite | | | |
|---|---|---|---|
| A | B | C | D |
| 14 | 16 | 24 | 17 |
| 23 | 10 | 07 | 19 |
| 09 | 02 | 15 | 05 |
| 11 | 20 | 04 | 18 |
| 06 | 03 | 13 | 01 |
| 12 | 22 | 08 | 21 |

Tabla 1. Aleatorización de las porciones de las porciones de masa

Estos valores aleatorios aseguran que cada una de las asignaciones de tratamientos posibles tenga la misma probabilidad de ocurrencia

**Análisis De Varianza**

El análisis de varianza (ANOVA), se refiere en general a un conjunto de situaciones experimentales y procedimientos estadísticos para el análisis de respuestas cuantitativas de unidades experimentales. El problema más sencillo de ANOVA se conoce como el análisis de varianza de un solo factor o diseño completamente al azar, éste se utiliza para comparar dos o más tratamientos, dado que sólo consideran dos fuentes de variabilidad, los tratamientos y el error aleatorio.

En este todas las corridas experimentales se deben de realizar en un orden aleatorio. De esta manera, si durante el estudio se hacen $N$ pruebas, éstas se corren al azar, de manera que los posibles efectos ambientales y temporales se vayan repartiendo equitativamente entre los tratamientos.

Vamos a suponer que se tienen $k$ poblaciones o tratamientos, independientes y con medias desconocidas $\mu_1, \mu_2, \ldots, \mu_k$, así como varianzas también desconocidas, pero que se supone que son iguales $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$. Las poblaciones pueden ser $k$ métodos de producción, $k$ tratamientos, $k$ grupos, etc., y sus medias se refieren o son medidas en términos de la variable de respuesta.

Si se decide hacer un experimento completamente al azar para comparar las poblaciones, que cumpla las condiciones antes mencionadas, entonces se tiene que hacer mediante la hipótesis de igualdad de medias:

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

$$H_1: \mu_i \neq \mu_j \text{ para algún } i \neq j$$

Los datos generados para un diseño completamente al azar para comparar dichas poblaciones se pueden escribir tal y como se muestra en la tabla 2

El número de tratamientos $k$ es determinado por el investigador y depende del problema en particular de que se trata. El número de observaciones en cada tratamiento debe escogerse con base a la variabilidad que se espera observar en los datos, así como en la diferencia mínima que el experimentador considera que es importante detectar. Por lo general se recomiendan entre $5$ y $30$ mediciones en cada tratamiento. En caso de que los tratamientos tengan efecto, las observaciones $y_{ij}$ de la tabla 1 se puede escribir como el modelo estadístico lineal dado por:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (1)$$

donde:

$\mu$ – Es el parámetro de escala común a todos los tratamientos, llamado media global
$\tau_i$ – Es un parámetro que mide el efecto del tratamiento $i$
$\varepsilon_{ij}$ – es el error atribuido a la medición $y_{ij}$

| TRATAMIENTOS | | | | | | | |
|---|---|---|---|---|---|---|---|
| $T_1$ | $T_2$ | $T_3$ | $\cdots$ | $T_i$ | $\cdots$ | $T_k$ | |
| $y_{11}$ | $y_{21}$ | $y_{31}$ | $\cdots$ | $y_{i1}$ | $\cdots$ | $y_{k1}$ | |
| $y_{12}$ | $y_{22}$ | $y_{32}$ | $\cdots$ | $y_{i2}$ | $\cdots$ | $y_{k2}$ | |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ | $\vdots$ | |
| $y_{1j}$ | $y_{2j}$ | $y_{3j}$ | $\cdots$ | $y_{ij}$ | $\cdots$ | $y_{kj}$ | |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ | $\cdots$ | $\vdots$ | |
| $y_{1n}$ | $y_{2n}$ | $y_{3n}$ | $\cdots$ | $y_{in}$ | $\cdots$ | $y_{kn}$ | |
| $n_1$ | $n_2$ | $n_3$ | $\cdots$ | $n_i$ | $\cdots$ | $n_k$ | $N$ |
| $y_{1\bullet}$ | $y_{2\bullet}$ | $y_{3\bullet}$ | $\cdots$ | $y_{i\bullet}$ | $\cdots$ | $y_{k\bullet}$ | $Y_{\bullet\bullet}$ |
| $\bar{y}_{1\bullet}$ | $\bar{y}_{2\bullet}$ | $\bar{y}_{3\bullet}$ | $\cdots$ | $\bar{y}_{i\bullet}$ | $\cdots$ | $\bar{y}_{k\bullet}$ | $\bar{Y}_{\bullet\bullet}$ |

Tabla 2. Diseño completamente al azar

Este modelo implica que actuarían a lo más dos fuentes de variabilidad: los tratamientos y el error aleatorio. La media global de la variable de respuesta no se considera una fuente de variabilidad por ser una constante en todos los tratamientos, que es un punto de referencia con el cual se comparan las respuestas medias de los tratamientos, tal como lo muestra la figura 1.

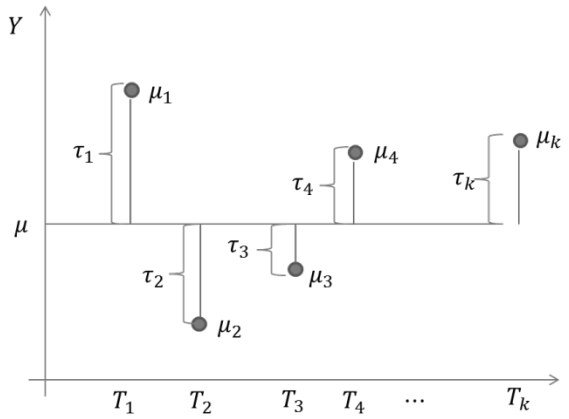


Figura 1. Representación de los efectos de los tratamientos en un diseño completamente al azar.

Entonces las hipótesis también se pueden escribir como:

$$H_0\colon\ \tau_1 = \tau_2 = \ldots = \tau_k = 0$$

$$H_1\colon\ \tau_i \neq 0 \text{ para algún } i$$

Si la respuesta media de un tratamiento particular $\mu_i$, es muy diferente de la media global $\mu$, es un síntoma de que existe un efecto de dicho tratamiento, ya que $\tau_i = \mu_i - \mu$. La diferencia que deben tener las medias entre sí para concluir que hay un efecto, es decir que los tratamientos sean diferentes, no lo dice el análisis de la varianza.

El modelo estadístico de la ecuación (1) describe dos situaciones diferentes con respecto a los efectos de los tratamientos. La primera, los $k$ tratamientos pueden ser elegidos explícitamente por el investigador; en este caso se quieren probar las hipótesis acerca de las medias de los tratamientos y las conclusiones se aplican únicamente a los niveles del factor considerados en el análisis. Las conclusiones no pueden extenderse a tratamientos similares que no fueron considerados expresamente. También si se quisiera estimar los parámetros del modelo $(\mu, \tau_i, \sigma^2)$. A éste se le llama *modelo con efectos fijos*. De manera alternativa, los $k$ tratamientos podrían ser una muestra aleatoria de una población más grande de tratamientos. En esta situación se desearía poder extender las conclusiones (las cuales se basan en la muestra de los tratamientos) a la totalidad de los tratamientos de la población, que se haya considerado en el análisis o no. Aquí las $\tau_i$ son variables aleatorias, y el conocimiento de las $\tau_i$ particulares que se investigaron es relativamente inútil. Más bien, se prueban las hipótesis acerca de la variabilidad de las $\tau_i$ y se intenta estimar su variabilidad. A éste se le llama *modelo con efectos aleatorios* o modelo de los componentes de la varianza.

### Análisis del estadístico del modelo con efectos fijos

El análisis de varianza es la técnica central en el análisis de datos experimentales, la idea es separar la variación total en partes con las que contribuye cada fuente de variación en el experimento, en el diseño completamente al azar se separa la variabilidad debida a los tratamientos y debida al erros.

Cuando la primera predomina claramente sobre la segunda, es cuando se concluye que los tratamientos tienen efecto también se puede decir que las medias son diferentes. Cuando los tratamientos no dominan y contribuyen igual o menor que el error, por los que se concluye que las medias son iguales y que no hay diferencias significativas entre los tratamientos, esto lo podemos ver en la figura 2.

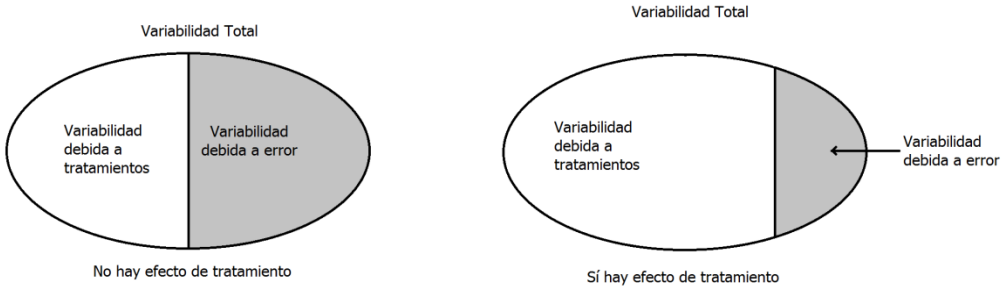


Figura 2. División de la variación total en sus componentes en un Diseño completamente al azar

De la figura 1 se puede observar claramente que:

$$\tau_i = \mu_i - \mu \quad (2)$$

En la figura 3 podemos observar el error residual de cada una de las observaciones y este se puede escribir de la forma:

$$\varepsilon_{ij} = y_{ij} - \mu_i \quad (3)$$

Pero la ecuación (2) y (3) se pueden escribir como

$$\mu_i = \tau_i + \mu \quad (4)$$
$$\mu_i = \varepsilon_{ij} + y_{ij} \quad (5)$$

Igualando la ecuación (4) y (5) se tiene:

$$\tau_i + \mu = \varepsilon_{ij} + y_{ij} \quad (6)$$

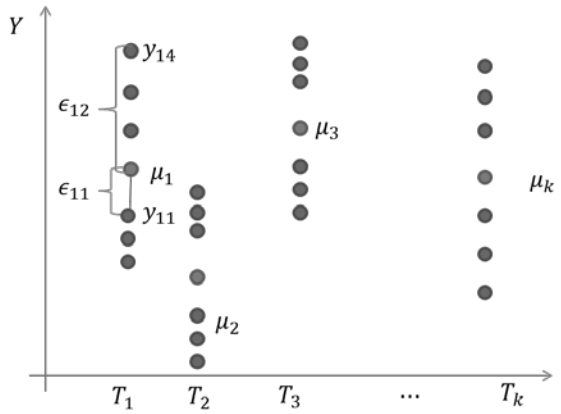


Figura 3. El error residual para el tratamiento uno.

Despejando a $y_{ij}$ tenemos:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (7)$$

Sabe que la media de $y_{ij}$ difiere de la media de la población, entonces la ecuación (7) se transforma en:

$$y_{ij} - \mu = \tau_i + \varepsilon_{ij} \quad (8)$$

Sustituyendo las ecuaciones (2) y (3) en la ecuación (8) obtendremos:

$$(y_{ij} - \mu) = (\mu_i - \mu) + (y_{ij} - \mu_i) \ldots\ldots(9)$$

En la ecuación (9) podemos ver que la desviación de una observación con relación a la gran media (*desviación total*) se descompone en el efecto de los tratamientos (*desviación entre tratamientos*) o desviación de cada tratamiento en relación a la gran media y en el error residual (*desviación dentro de cada tratamiento*) o desviación de cada observación con relación a su propio tratamiento.

Si tomamos los datos de la tabla 1 observamos que la identidad correspondiente a nuestro modelo para muestras es:

$$(y_{ij} - \bar{Y}_{\bullet\bullet}) = (\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet}) \ldots..(10)$$

Si elevamos al cuadrado los términos de ambos miembros de la igualdad de la ecuación (10), obtenemos las distintas sumas de cuadrados:

$$(y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \left[(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet})\right]^2$$

$$(y_{ij} - \bar{Y}_{\bullet\bullet})^2 = (\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + (y_{ij} - \bar{y}_{i\bullet})^2 + 2(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})(y_{ij} - \bar{y}_{i\bullet} \quad (11)$$

obteniendo las sumatorias la ecuación (11) se puede escribir como:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})(y_{ij} - \bar{y}_{i\bullet}) \qquad (12)$$

el último término de la ecuación (12) se puede escribir como:

$$2\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})(y_{ij} - \bar{y}_{i\bullet}) = 2\sum_{i=1}^{k}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})$$

tomando el último término tenemos que:

$$\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet}) = \sum_{j=1}^{n} y_{ij} - \sum_{j=1}^{n}\bar{y}_{i\bullet}$$

se sabe que

$$\bar{y}_{i\bullet} = \frac{y_{i\bullet}}{n}$$

entonces

$$\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet}) = y_{i\bullet} - n\left(\frac{y_{i\bullet}}{n}\right) = y_{i\bullet} - y_{i\bullet} = 0$$

por lo tanto

$$2\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet}-\bar{Y}_{\bullet\bullet})(y_{ij}-\bar{y}_{i\bullet})=0$$

la suma de cuadrados de la ecuación (12) nos queda como

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_{\bullet\bullet})^2=\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2+\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\bullet})^2$$

donde:

$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_{\bullet\bullet})^2 = SC_{Tot} =$ suma de cuadrados total.

$\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2 = SC_{Trt} =$ suma de cuadrados entre tratamientos

$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\bullet})^2 = SC_{error}$

$=$ suma de cuadrado dentro de los tratamientos

Es posible obtener eficientes fórmulas para las sumas de cuadrados, expandiendo y simplificando las expresiones anteriores. Esto produce

$$SC_{Tot}=\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_{\bullet\bullet})^2=\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}^2-\frac{Y_{\bullet\bullet}^2}{N} \qquad (13)$$

Y

$$SC_{Trt}=\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2=\sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n}-\frac{Y_{\bullet\bullet}^2}{N} \qquad (14)$$

La suma de cuadrados del error se obtiene mediante sustracción como

$$SC_{Tot} = SC_{Trt} + SC_{error}$$

$$SC_{error} = SC_{Tot} - SC_{Trt} \qquad (16)$$

En las ecuaciones anteriores se puede ver que las sumas de cuadrado son los numeradores de las varianzas respectivas, que el ANOVA se llama cuadrados medios. A partir de las sumatorias de cuadrados, es posible obtener dos estimadores insesgados de la varianza poblacional $\sigma^2$. Se puede demostrar que cuando las medias de los tratamientos son iguales $(H_0: verdadera)$ tanto la suma de cuadrados de los tratamientos como la suma de cuadrados del error divididas entre sus respectivos grados de libertad proporcionan estimadores insesgados e independientes de $\sigma^2$.
Dentro de los tratamientos, se tiene que:

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2}{n-1}$$

proporciona un estimador insesgado de la varianza de su grupo y bajo el supuesto de que las varianzas de los tratamientos son todas iguales, se pueden ponderar las varianzas de los $k$ tratamientos para obtener:

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2}{\sum_{i=1}^{k} n-1} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2}{N-k} = s^2_{error} = CM_{error}$$

que es la varianza dentro de los tratamientos o varianzas del error o cuadrados medios del error.

El segundo estimador de $\sigma^2$ se obtiene de la varianza de medias conocida (Teorema del Limite Central) $\sigma^2_{\bar{x}} = \sigma^2/n$, que al despejar $\sigma^2$ se tiene $\sigma^2 = n\sigma^2_{\bar{x}}$. Pero, un estimador insesgado de $\sigma^2_{\bar{x}}$ calculado a partir de las $k$ muestras:

$$s^2_{\bar{x}} = \frac{\sum_{i=1}^{k}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{k-1}$$

de donde

$$ns_{\bar{x}}^2 = \frac{n\sum_{i=1}^{k}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{k-1}$$

Se puede ver que el numerador de esta expresión es la suma de cuadrados entre tratamientos. Esta suma de cuadrados dividida entre los correspondientes grados de libertad, es llamada la varianza entre tratamientos o cuadrados medios entre tratamientos.

Si la hipótesis nula es cierta, se esperará que estos dos estimadores de $\sigma^2$ sean aproximadamente iguales y el cociente:

$$\frac{CM_{Trt}}{CM_{error}} = \frac{s_{Trt}^2}{s_{error}^2}$$

Que es una variable de F de Fisher y será la unidad o casi la unidad. Por el contrario si $(H_0: falsa)$; es decir, si los efectos de los tratamientos no son nulos, esto tenderá a ser significativamente menor que la unidad. Así se rechazará $(H_0)$ si

$$\frac{CM_{Trt}}{CM_{error}}$$

es mayor que la F de tablas ( F teórica) $F_{1-\alpha,gl\,Trt,gl\,error}$ donde $gl_{Trt} = k-1$ y $gl_{error} = N-k$. Todo esto se puede sintetizar en una tabla llamada tabla de ANOVA (tabla3).

| Fuente de Variación | Grados de libertad (gl) | Suma de Cuadrados (SC) | Varianza o Cuadrados Medios (CM) | $F_{calc}$ |
|---|---|---|---|---|
| "entre" Tratamientos | $k-1$ | $SC_{Trt}=\sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n}-\frac{Y_{\bullet\bullet}^2}{N}$ | $CM_{Trt}=\frac{SC_{Trt}}{k-1}$ | $F_{calc}=\frac{CM_{Trt}}{CM_{error}}$ |
| "dentro" Tratamientos o error residual | $N-k$ | $SC_{error}\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}^2-\sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n}$ | $CM_{error}=\frac{SC_{error}}{N-K}$ | |
| Total | $N-1$ | $SC_{Tot}=\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}^2-\frac{Y_{\bullet\bullet}^2}{N}$ | | |

Tabla 3. Tabla de ANOVA para un diseño completamente al azar

Vamos a realizar un ejemplo, suponga que la empresa "Sabrosita" se dedica a la elaboración y venta de botanas, un ingeniero en alimentos a observado que en las frituras, las rosquillas absorben grasa en diversas cantidades, y está interesado en saber si la cantidad absorbida depende del tipo de grasa que se utiliza, para esto, desarrolló un experimento en el cual se prepararon 24 bolsas de 250 gramos y se cocieron con cuatro diferentes tipos de aceites (aceite de cártamo (A), aceite de girasol (B), aceite de semilla de maíz (C) y aceite de soya (D)). Los datos obtenidos de grasa absorbida en miligramos se muestran en la tabla 4.

El ingeniero se pregunta si estos datos aportan suficiente evidencia para indicar una diferencia entre los diferentes tipos de aceites.

En análisis de la varianza se desarrolla como sigue:

Primero se deben plantear las hipótesis, para este ejemplo se tiene:

$$H_0\colon\ \mu_A = \mu_B = \mu_C = \mu_D$$

$$H_1\colon\ \mu_i \neq \mu_j \text{ para algún } i \neq j$$

o bien

$$H_0\colon\ \tau_A = \tau_B = \tau_C = \tau_D = 0$$

$$H_1\colon\ \tau_i \neq 0 \text{ para algún } i$$

| Tipo de Aceite | | | |
|---|---|---|---|
| A | B | C | D |
| 64 | 88 | 75 | 55 |
| 72 | 91 | 93 | 66 |
| 68 | 97 | 71 | 49 |
| 77 | 82 | 73 | 64 |
| 65 | 85 | 76 | 60 |
| 85 | 77 | 78 | 58 |

Tabla 4. Gramos de grasa absorbida

Ahora vamos a obtener los totales y las medias de acuerdo a lo planteado en la tabla 1.

| | A | B | C | D | |
|---|---|---|---|---|---|
| $n$ | 6 | 6 | 6 | 6 | $N = 24$ |
| $y_{i\bullet}$ | 431 | 520 | 466 | 352 | $Y_{\bullet\bullet} = 1769$ |
| $\bar{y}_{i\bullet}$ | 71.83 | 86.67 | 77.67 | 58.67 | $\bar{Y}_{\bullet\bullet} = 73.71$ |

Para hacer más sencillos los cálculos, tenemos:

$$\frac{Y_{\bullet\bullet}^2}{N} = \frac{(1769)^2}{24} = \frac{3129361}{24} = 130{,}390.04$$

y

$$\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 = (64)^2 + (72)^2 + (68)^2 + \ldots + (58)^2 = 133{,}941$$

Sustituyendo estos valores en la ecuación (13):

$$SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N} = 133{,}941 - 130{,}390.04 = 3{,}550.96$$

Para calcular la sumatoria de cuadrados de los tratamientos primero realizaremos:

$$\sum_{i=1}^{k} \frac{y_{i\bullet}^2}{n} = \frac{(431)^2 + (520)^2 + (466)^2 + (352)^2}{6} = 132{,}870.17$$

Sustituyendo en la ecuación (14)

$$SC_{Trt} = \sum_{i=1}^{k} \frac{y_{i\bullet}^2}{n} - \frac{Y_{\bullet\bullet}^2}{N} = 132{,}870.17 - 130{,}390.04 = 2{,}480.12$$

Para obtener la sumatoria de cuadrados del error solo sustituimos los dos resultados anteriores en la ecuación (16):

$$SC_{error} = SC_{Tot} - SC_{Trt} = 3{,}550.96 - 2{,}480.12 = 1{,}070.83$$

Ahora determinaremos los grados de libertad para las sumatorias de cuadrados y esto se obtienen de la siguiente manera:

Para los tratamientos $k - 1 = 4 - 1 = 3$
Para el error $N - K = 24 - 4 = 20$
Para el total $N - 1 = 24 - 1 = 23$

En seguida calcularemos los cuadrados medios del erro y de los tratamientos, dividiendo sus respectivas sumas cuadradas entre sus respectivos grados de libertad:

$$CM_{Trt} = \frac{SC_{Trt}}{k-1} = \frac{2{,}480.12}{3} = 826.70$$

$$CM_{error} = \frac{SC_{error}}{N-K} = \frac{1{,}070.83}{20} = 53.54$$

Por ultimo calcularemos el valor de $F_{calc}$

$$F_{calc} = \frac{CM_{Trt}}{CM_{error}} = \frac{826.70}{53.54} = 15.44$$

Con los datos obtenidos elaboraremos la tabla de análisis de la varianza:

Con un nivel de significancia $\alpha = 5\%$ y 3, 20 grados de libertad, se encuentra el valor en la tabla, siendo para este caso $F_{tablas} = 2.87$. Como $F_{calc} = 15.44 > F_{tablas} = 2.87$, se rechaza la hipótesis nula $H_0$, como lo podemos ver en la siguiente figura. Por lo que sí existe suficiente evidencia estadística al nivel del 5% de que hay una diferencia significativa entre los diferentes tipos de grasa que absorben las rosquillas.

| Fuente de Variación | Grados de libertad (gl) | Suma de Cuadrados (SC) | Varianza o Cuadrados Medios (CM) | $F_{calc}$ |
|---|---|---|---|---|
| Tratamientos | 3 | 2,480.12 | 826.70 | 15.44 |
| Error residual | 20 | 1,070.83 | 53.54 | |
| Total | 23 | 3,550.96 | | |

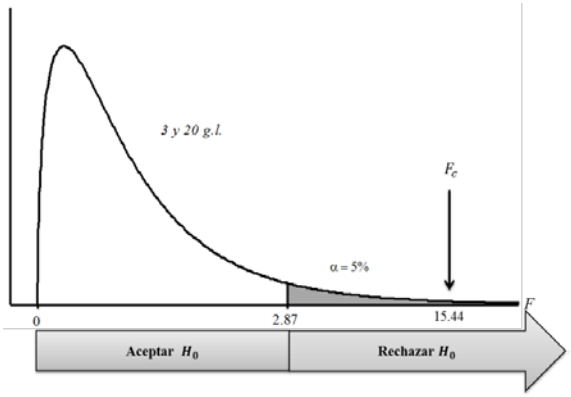


Figura 4. Zona de aceptación y de rechazo de la hipótesis nula.

## Intervalos de Confianza

Como vimos anteriormente la media del tratamiento *i*-ésimo se define como:

$$\mu_i = \tau_i + \mu$$

Un estimador puntual de $\mu_i$ es $\hat{\mu}_i = \bar{Y}_{i\bullet}$. Entonces, si se supone que los errores tienen una distribución normal, el promedio de cada tratamiento tiene una distribución normal con media $\mu_i$ y varianza $\sigma^2/n$. Por lo tanto, si se conociera $\sigma^2$, podría usarse la distribución normal para construir un intervalo de confianza. Usando los cuadrados medios del error como un estimador de $\sigma^2$, el intervalo de confianza se basaría en la distribución *t*, ya que:

$$t = \frac{\bar{Y}_{i\bullet} - \mu_i}{\sqrt{\frac{CM_{error}}{n}}}$$

Tiene una distribución *t* con $N - k$ grados de libertad. En base a lo anterior, se puede calcular un intervalo de confianza del $100(1-\alpha)$ porciento pala media del tratamiento *i*-ésimo $\mu_i$, es:

$$\bar{y}_{i\bullet} - t_{\frac{\alpha}{2},N-K}\sqrt{\frac{CM_{error}}{n}} \leq \mu_i \leq \bar{y}_{i\bullet} + t_{\frac{\alpha}{2},N-K}\sqrt{\frac{CM_{error}}{n}}$$

Vamos a calcular los intervalos de confianza de 95% para cada tipo de grasa. Las estimaciones de la media de tipo de grasa utilizado en el ejemplo anterior son:

Tipo de aceite A: $\hat{\mu}_A = \bar{y}_{A\bullet} = 71.83,$
Tipo de aceite B: $\hat{\mu}_B = \bar{y}_{B\bullet} = 86.67$
Tipo de aceite C: $\hat{\mu}_C = \bar{y}_{C\bullet} = 77.67$
Tipo de aceite D: $\hat{\mu}_D = \bar{y}_{D\bullet} = 58.67$

Los intervalos de confianza del 95% para los tipos de aceite se encuentran usando la fórmula anterior y usando una tabla de la distribución $t$ de student como sigue:

Tipo de aceite A:

$$\bar{y}_{i\bullet} \pm t_{\frac{\alpha}{2},N-K}\sqrt{\frac{CM_{error}}{n}}$$

$$71.83 \pm t_{\frac{0.05}{2},20}\sqrt{\frac{53.54}{6}}$$

$$71.83 \pm 2.08\sqrt{\frac{53.54}{6}}$$

$$60.82 \leq \mu_A \leq 82.84$$

Podemos decir que con un nivel de confianza del 95% que la cantidad de grasa absorbida por el tipo de aceite A se encuentra entre 60.82 7y 82.84 miligramos.

De manera similar para los demás tipos de aceite se tiene:

Tipo de aceite B:

$$75.66 \leq \mu_B \leq 97.68$$

Tipo de aceite C:

$$66.66 \leq \mu_C \leq 88.68$$

Tipo de aceite D:

$$47.66 \leq \mu_A \leq 69.68$$

**Verificación de los supuestos del modelo**

Las suposiciones en el análisis de varianza son fundamentalmente:

- Normalidad de las distribuciones
- Igualdad de Varianzas
- Independencia

Debido a estos supuestos la validez de los resultados obtenidos en cualquier análisis de la varianza quedará sujeta a que se cumplan dichos supuestos. Esto es, la respuesta $(y)$ se debe de distribuir normalmente, con la misma varianza en cada tratamiento y las mediciones deben ser independientes.

Es una práctica común utilizar la muestra de residuos para comprobar los supuestos del modelo, ya que si los supuestos se cumplen, los residuos o también conocidos como residuales se pueden ver como una muestra aleatoria de una distribución normal con media cero y varianza constante. Los residuos, $e_{ij}$, se definen como la diferencia entre la respuesta observada $y_{ij}$, y la respuesta predicha por el modelo $(\hat{y}_{ij})$, lo cual permite hacer un diagnóstico más directo de la calidad del modelo, ya que su magnitud señala qué tan bien describe a los datos el modelo. Para comprobar cada supuesto existen pruebas analíticas y gráficas. Por sencillez, muchas veces se prefieren las pruebas gráficas. Éstas tienen el inconveniente de que son exactas, pero aun así, en la mayoría de las situaciones prácticas proporcionan le evidencia suficiente contra o a favor de los supuestos.

El uso de las pruebas gráficas requiere una fuerte evidencia visual para concluir que el supuesto en cuestión no se cumple, ya que se requiere que la evidencia en contra de un supuesto esté soportada por más de dos puntos.

Cuando son uno o dos los puntos que se salen del comportamiento esperado de las gráficas se puede tratar de un problema de puntos aberrantes, no de una violación del supuesto en cuestión. En este caso debe de investigarse la obtención de dichas mediciones atípicas ya que ese tipo de puntos pueden afectar sensiblemente los resultados del análisis.

*Normalidad*

Un procedimiento gráfico para verificar el cumplimiento del supuesto de normalidad de los residuos consiste en graficar los residuos en la gráfica de probabilidad normal. Esta gráfica tiene las escalas de tal manera que si los residuos siguen un distribución normal, al graficarlos tienden a quedar alineados en una línea recta; por lo tanto, si claramente no se alinean se concluye que el supuesto de normalidad no es correcto.

Para realizar esta gráfica se considera $N$ residuos $e_i$ que resultan del análisis de varianza de los cuales se quiere verificar su procedencia de una distribución normal. Los pasos en la construcción de la gráfica de probabilidad normal para los residuos son los siguientes:

1. Ordenar los $N$ valores en forma ascendente y asignarle rangos de 1 a $N$.
2. Calcular una posición de graficación para cada dato en función de su rango y del total de observaciones como $(i-0.5)/N, i = 1, 2, \dots, N$.
3. Obtener el valor norma estándar de $Z_i$ que cumple con la relación $(i-0.5)/N = P(Z < Z_i) = \Phi(Z_i)$

Donde $\Phi(Z_i)$ es la distribución normal estándar acumulada evaluad en$Z_i$. Las parejas a dibujar en el papel son $(r_i, Z_i)$. Vamos a calcular los residuos del ejemplo anterior, para comprobar este supuesto. Realizaremos los cálculos como se muestra en la tabla 5.

*Varianza constante*

Una forma de verificar el supuesto de varianza constante o que los tratamientos tengan la misma varianza, es graficando los valores predichos contra los residuos ($\hat{y}_{ij}\ vs\ e_i$) por lo general $\hat{y}_{ij}$ va en el eje horizontal y los residuos en el eje vertical. Si los puntos en esta gráfica se distribuyen de manera aleatoria en una banda horizontal, sin ningún patrón claro o contundente), entonces es señal de que se cumple el supuesto de que los tratamientos tienen igual varianza. Por el contrario, si se distribuyen con algún patrón claro, como por ejemplo en forma de corneta o embudo, entonces es señal de no se está cumpliendo dicho supuesto. Un claro embudo en los residuales indicará que el error de pronóstico del modelo tiene una relación directa con la magnitud del pronóstico. En la tabla 6 se presenta el valor predicho $\hat{y}_{ij}$ y el valor residual $e_{ij}$; ya que con estos valores se construya la gráfica para comprobar el supuesto de igualdad de varianzas, esta se muestra en la figura 6.

| Dato | Rango i | $(i-0.5)/N$ | $\Phi(Z_i)$ | Dato | Rango i | $(i-0.5)/N$ | $\Phi(Z_i)$ |
|---|---|---|---|---|---|---|---|
| 49 | 1 | 0.02 | -2.04 | 71 | 13 | 0.52 | 0.05 |
| 55 | 2 | 0.06 | -1.53 | 72 | 14 | 0.56 | 0.16 |
| 58 | 3 | 0.10 | -1.26 | 72 | 15 | 0.60 | 0.26 |
| 60 | 4 | 0.15 | -1.05 | 73 | 16 | 0.65 | 0.37 |
| 64 | 5 | 0.19 | -0.89 | 75 | 17 | 0.69 | 0.49 |
| 64 | 6 | 0.23 | -0.74 | 76 | 18 | 0.73 | 0.61 |
| 64 | 7 | 0.27 | -0.61 | 77 | 19 | 0.77 | 0.74 |
| 65 | 8 | 0.31 | -0.49 | 77 | 20 | 0.81 | 0.89 |
| 65 | 9 | 0.35 | -0.37 | 78 | 21 | 0.85 | 1.05 |
| 66 | 10 | 0.40 | -0.26 | 85 | 22 | 0.90 | 1.26 |
| 68 | 11 | 0.44 | -0.16 | 85 | 23 | 0.94 | 1.53 |
| 68 | 12 | 0.48 | -0.05 | 93 | 24 | 0.98 | 2.04 |

Tabla 5. Cálculos para realizar una gráfica de probabilidad normal

La gráfica se muestra en la figura 5, como podemos observar en esta gráfica, se cumple el supuesto de normalidad, ya que la mayoría de los puntos están sobre la línea recta.

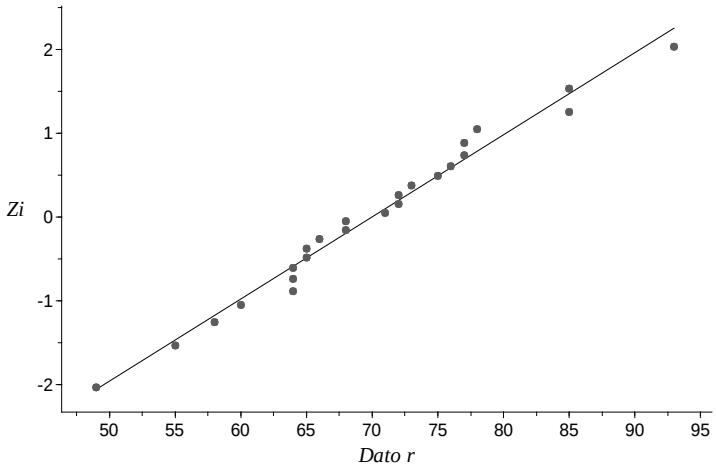


Figura 5.Grafica de probabilidad

*Independencia*

La suposición de independencia en los residuos se puede verificar si se grafica el orden en que se colectó un dato contra el residuo correspondiente, de esta manera, si al graficar en el eje horizontal el tiempo y en el vertical los residuos, se detecta una tendencia o patrón no aleatorio claramente definido, esto es evidencia de que existe una correlación entre los errores y por lo tanto, el supuesto de independencia no se cumple. Si el comportamiento de los puntos es aleatorio dentro de una banda horizontal, se puede decir que el supuesto se está cumpliendo. Si este supuesto no se cumple, generalmente indica que hay deficiencias en la planeación y ejecución del experimento; asimismo, puede ser un indicador de que no se aplicó en forma correcta el principio de aleatorización, o de que conforme se fueron realizando las pruebas experimentales aparecieron factores que afectaron la respuesta observada. Por ello, en caso de tener problemas con este supuesto, las conclusiones que se obtienen del análisis son endebles y por ello es mejor revisar lo hecho y tratar de investigar por qué no se cumplió con ese supuesto de independencia, a fin de reconsiderar la situación.

Para comprobar los supuestos de igualdad de varianzas y el supuesto de independencia en nuestro ejemplo, primero obtendremos los valores residuales, el valor predicho y el valor aleatorio, tal y como se muestra en la tabla 5.

| Observación | grasa | $e_{ij}$ | $\hat{y}_{ij}$ | Valor aleatorio | Observación | grasa | $e_{ij}$ | $\hat{y}_{ij}$ | Valor aleatorio |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 64 | -7.83 | 71.83 | 14 | 13 | 75 | -2.67 | 77.67 | 24 |
| 2 | 72 | 0.17 | 71.83 | 23 | 14 | 93 | 15.33 | 77.67 | 7 |
| 3 | 68 | -3.83 | 71.83 | 9 | 15 | 71 | -6.67 | 77.67 | 15 |
| 4 | 77 | 5.17 | 71.83 | 11 | 16 | 73 | -4.67 | 77.67 | 4 |
| 5 | 65 | -6.83 | 71.83 | 6 | 17 | 76 | -1.67 | 77.67 | 13 |
| 6 | 85 | 13.17 | 71.83 | 12 | 18 | 78 | 0.33 | 77.67 | 8 |
| 7 | 88 | 1.33 | 86.67 | 16 | 19 | 55 | -3.67 | 58.67 | 17 |
| 8 | 91 | 4.33 | 86.67 | 10 | 20 | 66 | 7.33 | 58.67 | 19 |
| 9 | 97 | 10.33 | 86.67 | 2 | 21 | 49 | -9.67 | 58.67 | 5 |
| 10 | 82 | -4.67 | 86.67 | 20 | 22 | 64 | 5.33 | 58.67 | 18 |
| 11 | 85 | -1.67 | 86.67 | 3 | 23 | 60 | 1.33 | 58.67 | 1 |
| 12 | 77 | -9.67 | 86.67 | 22 | 24 | 58 | -0.67 | 58.67 | 21 |

Tabla 6. Valores residuales, valores predichos y el valor aleatorio para cada observación

Tomando los valores de la tabla 5 graficaremos los residuos contra los valores predichos, para comprobar el supuesto de igualdad de varianza y contra el orden aleatorio con la finalidad de comprobar el supuesto de independencia, como se muestra en la figura 6.

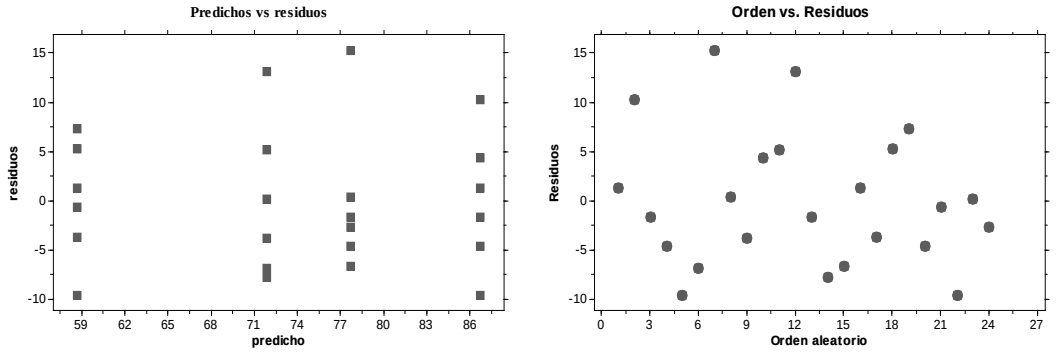


Figura 6.Gráficas de residuos para los diferentes tipos de grasas.

En la figura 5 y 6, se puede observar que se cumplen los 3 supuestos para el análisis de varianza ya que en la figura 6 no se sigue un patrón de comportamiento y en la figura si la mayoría de los valores están sobre la línea recta.

Que se puede hacer cuando los datos no satisfacen las suposiciones, por ejemplo, vamos a suponer que las varianzas de las respuestas para diferentes tratamientos no fueran iguales. En este caso, lo que se puede hacer es transformar los resultados, es decir, en lugar de utilizar los datos originales, estos se pueden cambiar por sus raíces cuadradas, logaritmos o alguna otra función. Se han encontrado transformaciones que tienden a estabilizar la varianza de las respuestas y al mismo tiempo hacen que las distribuciones de probabilidad de las respuestas transformadas estén más cerca de la normalidad.

Cundo es imposible lograr que se satisfagan las suposiciones del análisis de varianza, se deben utilizar procedimientos no paramétricos de pruebas de hipótesis; ya que estos se basan en las magnitudes comparativas de las mediciones y son casi tan eficientes y poderosos para detectar diferencias de tratamientos.

## Ventajas y desventajas al utilizar un Diseño completamente aleatorizado

Ventajas

a. Permite gran flexibilidad, es decir puede usarse cualquier número de tratamientos y repeticiones, además se puede variar el número de repeticiones de un tratamiento a otro.
b. El análisis estadístico es sencillo, aún si el número de repeticiones no es el mismo para cada tratamiento.
c. El análisis estadístico es fácil aun cuando los datos de algunas de las unidades experimentales o algunos tratamientos completos se hayan perdido o se rechacen por alguna causa.
d. Es el diseño que se basa en más grados de libertad para la estimación de los cuadrados medios.

Desventajas

Para usar este diseño se necesitan unidades experimentales muy homogéneas, porque de otra manera la variación entre ellas pasa a formar parte del error experimental

### Caso desbalanceado

En algunos experimentos con un solo factor, el número de observaciones que se hacen bajo cada tratamiento puede ser diferente. Se dice entonces que el diseño es desbalanceado. Al análisis de varianza sigue siendo válido, aunque es necesario hacer unas ligeras modificaciones en las fórmulas de las sumatorias de cuadrados. Por ejemplo si se hacen $n_j$ observaciones bajo el tratamiento $i(i = 1, 2, 3, \dots, k)$, y $N = \sum_{j=1}^{n_j} n_j$, el número total de observaciones. Las fórmulas para calcular las sumatorias de cuadrado para el análisis de varianza con tamaños de muestra $n_i$ diferentes en cada tratamiento quedan de la siguiente manera:

$$SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_j} y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N}$$

$$SC_{Trt} = \sum_{i=1}^{k} \frac{y_{i\bullet}^2}{n_j} - \frac{Y_{\bullet\bullet}^2}{N}$$

La sumatoria de cuadrados del error se obtiene por sustracción como:

$$SC_{error} = SC_{Tot} - SC_{Trt}$$

Entonces la tabla de análisis de varianza queda de la siguiente manera:

| Fuente de Variación | Grados de libertad (gl) | Suma de Cuadrados (SC) | Varianza o Cuadrados Medios (CM) | $F_{calc}$ |
|---|---|---|---|---|
| "entre" Tratamientos | $k-1$ | $SC_{Trt} = \sum_{i=1}^{k} \frac{y_{i\bullet}^2}{n_j} - \frac{Y_{\bullet\bullet}^2}{N}$ | $CM_{Trt} = \frac{SC_{Trt}}{k-1}$ | $F_{calc} = \frac{CM_{Trt}}{CM_{error}}$ |
| "dentro" Tratamientos o error residual | $N-k$ | $SC_{error} \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{k} \frac{y_{i\bullet}^2}{n}$ | $CM_{error} = \frac{SC_{error}}{N-K}$ | |
| Total | $N-1$ | $SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N}$ | | |

Elegir un diseño balanceado tiene dos ventajas importantes. Primera, el procedimiento de prueba es relativamente insensible a las desviaciones pequeñas del supuesto de la igualdad de varianzas si los tamaños de las muestras son iguales. Este no es el caso para tamaños de las muestras diferentes. Segunda, la potencia dela prueba se maximiza si las muestras son de tamaños iguales.

## Apéndice

Demostración de :

$$SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{Y_{\cdot\cdot}^2}{N}$$

Elevando al cuadrado

$$SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}^2 - 2y_{ij}\bar{Y}_{\cdot\cdot} + \bar{Y}_{\cdot\cdot}^2\right)$$

$$SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - 2\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}\bar{Y}_{\cdot\cdot} + \sum_{i=1}^{k}\sum_{j=1}^{n}\bar{Y}_{\cdot\cdot}^2$$

en donde

$$\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = Y_{\cdot\cdot} \qquad \text{y} \qquad \sum_{i=1}^{k}\sum_{j=1}^{n} = N$$

de acuerdo a la tabla 1, tenemos:

$$\bar{Y}_{\cdot\cdot} = \frac{Y_{\cdot\cdot}}{N}$$

Sustituyendo las dos ecuaciones anteriores en la sumatoria de cuadrados totales

$$SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - 2(Y_{\cdot\cdot})\left(\frac{Y_{\cdot\cdot}}{N}\right) + N\left(\frac{Y_{\cdot\cdot}}{N}\right)^2$$

$$SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - 2\left(\frac{Y_{\cdot\cdot}^2}{N}\right) + N\left(\frac{Y_{\cdot\cdot}^2}{N^2}\right)$$

$$SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - 2\left(\frac{Y_{\cdot\cdot}^2}{N}\right) + \left(\frac{Y_{\cdot\cdot}^2}{N}\right)$$

finalmente se tiene:

$$SC_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N}$$

Ahora desarrollaremos la sumatoria de cuadrados de los tratamientos

$$SC_{Trt} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet}^2 - 2\bar{y}_{i\bullet}\bar{Y}_{\bullet\bullet} + \bar{Y}_{\bullet\bullet}^2)$$

$$SC_{Trt} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\bar{y}_{i\bullet}^2 - 2\sum_{i=1}^{k}\sum_{j=1}^{n}\bar{y}_{i\bullet}\bar{Y}_{\bullet\bullet} + \sum_{i=1}^{k}\sum_{j=1}^{n}\bar{Y}_{\bullet\bullet}^2$$

se sabe que

$$\bar{y}_{i\bullet} = \frac{y_{i\bullet}}{n} \qquad \text{y} \qquad \bar{Y}_{\bullet\bullet} = \frac{Y_{\bullet\bullet}}{N}$$

además

$$\sum_{i=1}^{k} y_{i\bullet} = Y_{\bullet\bullet}\ ; \qquad \sum_{j=1}^{n} = n \quad \text{y} \quad \sum_{i=1}^{k}\sum_{j=1}^{n} = N$$

entonces:

$$SC_{Trt} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n\left(\frac{y_{i\bullet}}{n}\right)^2 - 2n\left(\frac{y_{i\bullet}}{n}\right)\left(\frac{Y_{\bullet\bullet}}{N}\right) + N\left(\frac{Y_{\bullet\bullet}}{N}\right)^2$$

$$SC_{Trt} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n\left(\frac{y_{i\bullet}^2}{n^2}\right) - 2n\left(\frac{Y_{\bullet\bullet}}{n}\right)\left(\frac{Y_{\bullet\bullet}}{N}\right) + N\left(\frac{Y_{\bullet\bullet}^2}{N^2}\right)$$

eliminando términos,

$$SC_{Trt} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n} - 2\left(\frac{Y_{\bullet\bullet}^2}{N}\right) + \left(\frac{Y_{\bullet\bullet}^2}{N}\right)$$

finalmente tenemos que:

$$SC_{Trt} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n} - \frac{Y_{\bullet\bullet}^2}{N} \qquad (14)$$

Se puede demostrar, como en las ecuaciones anteriores, que la sumatoria de cuadrado del error es igual a

$$SC_{error} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}^2 - \sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n} \qquad (15)$$

## EJERCICIOS PROPUESTOS

1.- En una industria química se prueban diferentes mezclas para ver si difieren en cuanto al peso molecular final. Se prueban cuatro diferentes mezclas, con cinco repeticiones cada una. A continuación se muestra una parte de la tabla del análisis de varianza y los promedios obtenidos para cada mezcla:

| Fuente de Variación | Valor *p* |
|---|---|
| Mezcla | 0.01 |
| Error | |

| Mezcla | Peso medio |
|---|---|
| A | 10000 |
| B | 7000 |
| C | 8000 |
| D | 7500 |

a) ¿Las mezclas difieren de manera significativa en cuanto a su peso molecular?
b) Con el análisis de varianza y de acuerdo al promedio, ¿se puede asegurar que con la mezcla *B* se logra un menor peso molecular? Argumente su respuesta.
c) Si al verificar los supuestos de varianza constante (igual varianza entre las mezclas), éstos no se cumplen, ¿qué significa eso? ¿Se puede seguir apoyando la conclusión del inciso a)?

2.- Se hace un estudio sobre la efectividad de tres marcas de spray para matar moscas. Para ello, cada producto se aplica a un grupo de 100 moscas, y se cuenta el número de moscas muertas expresado en porcentajes. Se hacen seis réplicas y los resultados obtenidos se muestran a continuación:

| | Número de réplica | | | | | |
|---|---|---|---|---|---|---|
| **Marca de spray** | 1 | 2 | 3 | 4 | 5 | 6 |
| **1** | 72 | 65 | 67 | 75 | 62 | 73 |
| **2** | 55 | 59 | 68 | 70 | 53 | 50 |
| **3** | 64 | 74 | 61 | 58 | 51 | 69 |

a) Formule la hipótesis adecuada y el modelo estadístico.
b) ¿Existe diferencia entre la efectividad promedio de los productos en spray?
c) ¿Hay algún spray mejor? Argumente su respuesta.
d) Dé un intervalo al 95% de confianza para la efectividad promedio (porcentaje) de cada una de las marcas.
e) Dibuje las gráficas de medias y los diagramas d caja simultáneos, después interprételos.
f) Verifique los supuestos de normalidad y de igual varianza entre las marcas.

3.- En un centro de investigación se realiza un estudio para comparar varios tratamientos que, al aplicarse previamente a los frijoles crudos, reducen su tiempo de cocción. Estos tratamientos, son a base de bicarbonato de sodio ($NaHCO_3$) y cloruro de sodio o sal común (NaCl). El primer tratamiento es el de control, que consiste en no aplicar ningún tratamiento. El tratamiento $T_2$ es el remojo en agua con bicarbonato de sodio, el $T_3$ es remojar en agua con sal común y el $T_4$ es remojar en agua con una combinación de ambos ingredientes en proporciones iguales. La variable de respuesta es el tiempo de cocción en minutos. Los datos se muestran en la siguiente tabla:

| Control | $T_2$ | $T_3$ | $T_4$ |
|---|---|---|---|
| 213 | 76 | 57 | 84 |
| 214 | 85 | 67 | 82 |
| 204 | 74 | 55 | 85 |
| 208 | 78 | 64 | 92 |
| 212 | 82 | 61 | 87 |
| 200 | 75 | 63 | 79 |
| 207 | 82 | 63 | 90 |

a) ¿De qué manera el experimentador debe aleatorizar los experimentos y el material experimental?

b) Dé ejemplos de factores que deben estar fijos durante las pruebas experimentales, para que no afecten los resultados y las conclusiones.
c) Formule y pruebe la hipótesis de que las medias de los tratamientos son iguales.
d) Obtenga el diagrama de caja y el gráfico de medias, después interprételos.
e) ¿Hay algún tratamiento mejor? ¿Cuál es el tiempo de cocción esperado para el mejor tratamiento?
f) Algo importante a cuidar en un experimento es que no haya efectos colaterales no deseados, causados por el tratamiento ganador; en este caso, piense en los posibles efectos colaterales que podría causar el mejor tratamiento.
g) ¿Se cumplen los supuestos del modelo? Verifique gráficamente.
h) Pruebe la hipótesis di igualdad de varianzas entre tratamientos (que corresponde a un supuesto).

4.- Para estudiar la confiabilidad de ciertos tableros electrónicos para carros, se someten a un envejecimiento acelerado durante 100 horas a determinada temperatura, y como variable de interés se mide la intensidad de corriente que circula entre dos puntos, cuyos valores aumentan con el deterioro. Se probaron 20 módulos repartidos de manera equitativa en cinco temperaturas y los resultados obtenidos fueron los siguientes:

| 20°C | 40°C | 60°C | 80°C | 100°C |
|---|---|---|---|---|
| **15** | 17 | 23 | 28 | 45 |
| **18** | 21 | 19 | 32 | 51 |
| **13** | 11 | 25 | 34 | 57 |
| **12** | 16 | 22 | 31 | 48 |

a) Formule la hipótesis y el modelo estadístico para el problema.
b) Realice el análisis de varianza para estos datos, a fin de estudiar si la temperatura afecta la intensidad de corriente promedio.
c) ¿La temperatura afecta la variabilidad de las intensidades? Es decir, verifique si hay igual varianza entre los diferentes tratamientos.

5.- En una empresa de manufactura se propone un tratamiento para reducir el porcentaje de productos defectuosos. Para validar esta propuesta se diseñó un experimento en el que se producía con o sin la propuesta de mejora. Cada corrida experimental consistió en producir un lote y la variable de respuesta es el porcentaje de producto defectuoso. Se hicieron 25 réplicas para cada tratamiento. Los datos obtenidos se muestran a continuación:

| Porcentaje de producto defectuoso | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Con tratamiento** | 5.3 | 4.0 | 4.0 | 4.0 | 2.6 | 2.1 | 5.1 | 4.1 | 4.1 | 3.2 | 5.1 | 2.2 | 4.1 |
| | 2.2 | 1.1 | 2.0 | 3.0 | 3.1 | 2.1 | 1.2 | 3.3 | 2.1 | 4.0 | 2.0 | 3.0 | |
| **Sin tratamiento** | 8.0 | 13.2 | 7.2 | 8.2 | 9.1 | 6.7 | 12.2 | 16.3 | 9.2 | 6.4 | 7.2 | 17.2 | 12.3 |
| | 8.7 | 11.3 | 4.5 | 6.6 | 9.2 | 10.2 | 10.6 | 13.3 | 5.2 | 6.2 | 8.0 | 4.8 | |

a) ¿Las diferencias son significativas estadísticamente?
b) ¿Cuál es el porcentaje de defectos que se espera con el nuevo tratamiento?
c) Cuantifique el nivel de reducción que se logró con el tratamiento propuesto.

6.- Una compañía farmacéutica desea evaluar el efecto que tiene la cantidad de almidón en la dureza de las tabletas. Se decidió producir lotes con una cantidad determinada de almidón, y que las cantidades de almidón a probar fueran 2%, 5% y 10%. La variable de respuesta sería el promedio de la dureza de 20 tabletas de cada lote. Se hicieron 4 réplicas por tratamiento y se obtuvieron los siguientes resultados:

| % de almidón | Dureza | | | |
|---|---|---|---|---|
| 2 | 4.3 | 5.2 | 4.8 | 4.5 |
| 5 | 6.5 | 7.3 | 6.9 | 6.1 |
| 10 | 9.0 | 7.8 | 8.5 | 8.1 |

a) ¿Hay evidencia suficiente de que el almidón influye en la dureza de las tabletas? Halle el ANOVA
b) Realice los análisis complementarios necesarios.
c) Si se desea maximizar la dureza de las tabletas, ¿Qué recomendaría al fabricante?
d) Verifique los supuestos

7.- Los datos que se presentan enseguida son rendimientos en toneladas por hectárea de un pasto con tres niveles de fertilización nitrogenada. El diseño fue completamente aleatorizado, con cinco repeticiones por tratamiento:

| Niveles de nitrógeno | | |
|---|---|---|
| **1** | 2 | 3 |
| **14.823** | 25.151 | 32.605 |
| **14.676** | 25.401 | 32.460 |
| **14.720** | 25.131 | 32.256 |
| **14.5141** | 25.031 | 32.669 |
| **15.065** | 25.267 | 32.111 |

a) ¿Las diferencias muestrales hacen obvia la presencia de diferencias poblacionales?
b) Obtenga el análisis de varianza e interprétalo.
c) Analice los residuos, ¿hay algún problema?

8.- Un químico del departamento de desarrollo de un laboratorio farmacéutico desea conocer cómo influye el tipo de aglutinante utilizado en tabletas de ampicilina de 500 mg en el porcentaje de friabilidad; para ello, se eligen los siguientes aglutinantes: polivinilpirrolidona (PVP), carboximetilcelulosa sódica (CMC) y grenetina (Gre). Los resultados del diseño experimental son los siguientes:

| Aglutinante | % de friabilidad | | | | |
|---|---|---|---|---|---|
| PVP | 0.485 | 0.250 | 0.073 | 0.205 | 0.161 |
| CMC | 9.64 | 9.37 | 9.53 | 9.86 | 9.79 |
| Gre | 0.289 | 0.275 | 0.612 | 0.152 | 0.137 |

a) Especifique el nombre del diseño experimental.
b) ¿Sospecha que hay algún efecto significativo del tipo de aglutinante sobre la variable de respuesta?
c) Escriba las hipótesis para probar la igualdad de medias y el modelo estadístico.
d) Realice el análisis adecuado para probar las hipótesis e interprete los resultados.
e) Revise los supuestos, ¿hay algún problema?

9.- Se cultivaron cuatro diferentes clonas de *agave tequilana* bajo un mismo esquema de manejo. Se quiere saber qué clona es la que responde mejor a dicho manejo, evaluando el nivel de respuesta con el porcentaje de azúcares reductores totales en base húmeda. Los datos se muestran a continuación:

| Clona | | | |
|---|---|---|---|
| **1** | 2 | 3 | 4 |
| **8.69** | 8.00 | 17.39 | 10.37 |
| **6.68** | 16.41 | 13.73 | 9.16 |
| **6.83** | 12.43 | 15.62 | 8.13 |
| **6.43** | 10.99 | 17.05 | 4.40 |
| **10.30** | 15.53 | 15.42 | 10.38 |

a) Mediante ANOVA, compare las medias de las clonas y verifique residuales.
b) ¿Hay una clona que haya respondido mejor al esquema de manejo? Argumente su respuesta.
c) En caso de que exista un empate estadístico entre 2 o más clonas, ¿Qué propondría para desempatar?

10.- Uno de los defectos que causan mayor desperdicio en la manufactura de discos ópticos compactos son los llamados "cometas". Típicamente, se trata de una partícula que opone resistencia al fluido en la etapa de entintado. Se quiere comprobar de manera experimental la efectividad de un tratamiento de limpieza de partículas que está basado en fuerza centrípeta y aire ionizado.

A 12 lotes de 50 CD se les aplica el tratamiento y a otros 12 lotes no se les aplica; en cada caso se mide el porcentaje de discos que presentan cometas, los resultados son los siguientes:

| Con tratamiento | Sin tratamiento |
|---|---|
| **5.30** | 8.02 |
| **4.03** | 13.18 |
| **4.03** | 7.15 |
| **4.00** | 8.23 |
| **2.56** | 9.11 |
| **2.05** | 6.66 |
| **5.06** | 12.15 |
| **4.06** | 16.3 |
| **2.08** | 9.20 |
| **4.03** | 6.35 |
| **2.04** | 7.15 |
| **1.18** | 8.66 |

a) Con el ANOVA vea si es efectivo el tratamiento de limpieza. ¿Debería implementarse?
b) ¿Es razonable suponer en el inciso a) que las varianzas son iguales?
c) ¿En qué porcentaje se reducen los discos de cometas?