
**Ricardo Alvarado Barrantes**

# Análisis de experimentos estadísticos usando R

EDITORIAL UCR

**Ricardo Alvarado Barrantes**

# Análisis de experimentos estadísticos usando R

EDITORIAL UCR
2025

CC.SIBDI.UCR - CIP/4304
Nombres: Alvarado Barrantes, Ricardo, autor.
Título: Análisis de experimentos estadísticos usando R /
Ricardo Alvarado Barrantes.
Descripción: Primera edición. | San José, Costa Rica :
Editorial UCR, 2025.

Identificadores: **ISBN 978-9968-02-276-7** (rústico)

Materias: LEMB: Diseño experimental. |
Estadística matemática – Métodos de simulación. |
Diseño experimental – Programas para computador. |
Estadística matemática – Programas para computador. |
Diseño experimental – Problemas, ejercicios, etc. |
Estadística matemática – Problemas, ejercicios, etc. |
LCSH: R (Lenguaje de programación para computadores).
Clasificación: CDD 519.570.285.513.3–ed. 23

Edición aprobada por la Comisión Editorial de la Universidad de Costa Rica.
Primera edición: 2025.


Ciudad Universitaria Rodrigo Facio. San José, Costa Rica.
Apdo.: 11501-2060 • Tel.: 2511 5310 • Fax: 2511 5257
administracion.siedin@ucr.ac.cr
www.editorial.ucr.ac.cr

# Índice general

# Prefacio

Durante varios años he venido enseñando el curso Diseño de Experimentos, el cual forma parte del programa de Bachillerato en Estadística de la Universidad de Costa Rica. Actualmente, este curso se ubica en el tercer semestre de la carrera de Estadística junto con otros dos, Modelos de Regresión Aplicados y Modelos Lineales Avanzados, son una secuencia que pretende desarrollar en los estudiantes las habilidades para planear y encaminar estudios observacionales o experimentales con validez estadística mediante modelos matemáticos apropiados para el análisis de los datos obtenidos en los estudios planteados.

En el presente libro se proporcionan los fundamentos de los diseños experimentales. En el primer capítulo, se inicia con diseños que contienen un solo factor y se introduce la técnica de análisis de varianza para realizar pruebas de hipótesis sobre igualdad de promedios. En el segundo capítulo, se profundiza en las comparaciones múltiples entre pares de promedios y en contrastes ortogonales; en este libro se da especial atención al uso de vectores para el cálculo de los estadísticos que se utilizan en las pruebas de hipótesis y en la construcción de intervalos de confianza. Los capítulos tercero y cuarto se dedican a los diseños factoriales, donde se da especial atención al concepto de interacción entre dos o más factores. El capítulo quinto se centra en diferentes aspectos del uso de bloques, tales como los bloques aleatorizados, las parcelas divididas y los bloques incompletos. También se hace mención a la relación que tiene este tipo de diseños con el uso de modelos mixtos. Más adelante, en el capítulo sexto, se estudia el uso de covariables que permiten reducir la variabilidad del error experimental. Finalmente, el último capítulo se concentra en el concepto de potencia de las pruebas estadísticas y presenta la técnica de simulaciones.

El propósito del libro es servir como un manual de referencia para el análisis de datos provenientes de experimentos básicos en el lenguaje de programación R (R Core Team, 2023); no obstante, se ha dedicado una sección al principio de cada capítulo para explicar los conceptos más relevantes en el tema que se desarrolla. Se explican los modelos matemáticos y su escritura adecuada. Por consiguiente, el lector puede llevar a cabo el análisis de datos y comprender esta parte matemática que sirve para el planteamiento adecuado de hipótesis y, eventualmente, como base para la generación de datos, tal como sucede en el último capítulo donde se utiliza el enfoque de simulación de datos para entender el concepto de potencia de una prueba estadística.

Cada capítulo contiene uno o varios ejercicios basados en un problema, los cuales se desarrollan usando R versión 4.3.1. Se espera que el estudiante realice los ejercicios y luego compare sus resultados con las respuestas. Para esto se hace una descripción del problema y luego se da una lista de preguntas con ayudas para resolver los ejercicios, se indican las funciones de R recomendadas para contestar cada pregunta y el código apropiado para usar esas funciones. Los ejercicios empiezan con un análisis descriptivo de los datos para que el estudiante los pueda visualizar antes de entrar en la formulación de un modelo. Después de la lista de preguntas se desarrolla cada ítem con el código de R y se agregan comentarios que ayuden a dar conclusiones. En la dirección http://editorial.ucr.ac.cr/documentos/DATOS1623.zip están disponibles todos los datos que se usan en los ejercicios.

Para el desarrollo de los ejercicios se usan varias librerías de R. La mayoría de las funciones están disponibles en el paquete básico `Stats`, cuyo uso no requiere descarga alguna, puesto que se activa directamente con la instalación de R. Se usan algunas librerías para hacer gráficos, tales como `car` (Fox y Weisberg, 2019), `lattice` (Sarkar, 2008) y `ggplot2` (Wickham, 2016). Se usa `dplyr` (Wickham, Francois, Henry y Müller, 2019) para obtener estadísticos de resumen por grupos. La librería `ibd` (Mandal, 2019) se emplea para el análisis de bloques incompletos, mientras que `pwr` (Champely, 2018) sirve para el cálculo de la potencia en pruebas de hipótesis.

## Reconocimientos

Los datos que se utilizan en los ejercicios se tomaron, en algunos casos, de trabajos realizados por estudiantes o se simularon para que se lograran demostrar las características deseadas desde el punto de vista didáctico. Los estudiantes Edwin Abarca Araya, Elsa Guillén Amador y Christopher Torres Rojas recolectaron los datos para el trabajo que se titula «Efecto de la forma de salida y el tipo de calentamiento en el tiempo de recorrido de la carrera de 100 metros planos». De forma similar, las estudiantes Erika Araya Cárdenas, María Jesús Castro Solís y Angélica Zúñiga Baldí recolectaron los datos para el trabajo titulado «Efecto de la proporción de glicerina y el tipo de agua en la resistencia de burbujas de jabón». Los datos sobre asfalto fueron recolectados por el estudiante de Ingeniería Civil, Yordy Esteban Morales Guzmán, como parte de su tesis de licenciatura no concluida. Otros conjuntos de datos se generaron a partir de consultorías realizadas por el autor o la profesora María Isabel González Lutz. En tales casos, el problema original inspiró la descripción de uno nuevo y luego se generaron datos que cumplieran con las características necesarias para desarrollar el ejercicio.

Durante el proceso de construcción de este libro tuve la retroalimentación de muchas personas, especialmente, tuve largas sesiones de reflexión sobre conceptos, enfoques y detalles de diversa índole con mi colega María Isabel González Lutz, quien me dio aportes sumamente valiosos, los cuales hicieron que el trabajo final fuera más comprensible y acertado. También conté con la revisión detallada de parte del profesor Johnny Madrigal Pana, quien tuvo la paciencia de leer todo el manuscrito. Obtuve respuestas provenientes de estudiantes de este mismo curso o alumnos de otras asignaturas dentro de la carrera de Estadística, ante mi solicitud de leer partes del texto. Agradezco los comentarios y sugerencias de Carlos Arrieta Elizondo, Shu Wei Chou, César Gamboa Sanabria, Susana García Calvo, Catalina Sandoval Alvarado, Rebeca Sura Fonseca y Pablo Vivas Corrales.

En los últimos dos años compartí con la profesora Shirley Rojas Salazar la enseñanza del curso de Diseño de Experimentos, por lo que pude obtener

de ella importantes observaciones que surgieron durante la experiencia de impartir el curso y conversar sobre los detalles que descubrimos en el camino. Quiero agradecer muy especialmente a la estudiante Andrea Vargas Montero, quien pacientemente tomó notas durante las lecciones para informarme de cambios que se debían realizar a los ejercicios. Finalmente, quiero expresar un agradecimiento muy especial al estudiante Brayan Monge Blanco, pues no solo leyó todo el borrador final para hacer importantes observaciones, sino que siempre ha sido una persona totalmente dispuesta a impulsar mis iniciativas, a Brayan mi sincera gratitud.

Ricardo Alvarado Barrantes

San José, Costa Rica
Marzo, 2024

# Capítulo 1

## Análisis de varianza de una vía

En los experimentos estadísticos se estudia una variable llamada respuesta y se busca verificar si el hecho de variar ciertas condiciones controlables repercute en cambios en el promedio de esa variable. Cada una de las condiciones que se cambia se denomina tratamiento. El diseño puede contar con varios factores o variables que se controlan. Cada una de las posibilidades que se estudian de un factor es un nivel de ese factor. La combinación de todos los niveles de los diferentes factores da origen a los tratamientos del diseño experimental.

Cuando se trabaja con un experimento donde solo hay un factor de diseño, los tratamientos coinciden con los niveles de ese factor e interesa comparar el efecto de los diferentes tratamientos sobre el promedio de la variable de interés. Para hacer estas comparaciones se recurre al concepto del efecto del tratamiento y se dice que si el tratamiento no tiene efecto sobre la respuesta, los promedios se mantendrán iguales para todos los tratamientos. En caso contrario, cuando alguno de los tratamientos tiene un efecto positivo (negativo), el promedio correspondiente será mayor (menor) que el promedio general de la respuesta (combinando los datos de todos los tratamientos). Por lo tanto, el efecto de un tratamiento representa la distancia del promedio dentro de ese tratamiento al promedio general de la variable respuesta.

## 1.1 Modelo con un factor

Matemáticamente se denota el efecto del j-ésimo tratamiento con $\tau_j$ y se define como $\tau_j = \mu_j - \mu$, donde $\mu_j$ representa el promedio del j-ésimo tratamiento y $\mu$ el promedio general. De esta forma la representación matemática del efecto coincide con el concepto explicado anteriormente, que es la distancia entre el promedio específico del tratamiento respecto al promedio general.

En la Figura 1.1 (superior) se ilustra un ejemplo en el que no existe efecto del factor sobre la respuesta promedio. En este caso, los promedios de los 4 tratamientos son iguales al promedio general y la distancia entre cada uno de los promedios específicos y el promedio general es cero, por lo que el valor del j-ésimo efecto es cero para todos los tratamientos ($\tau_j = 0$). En cambio, en el gráfico inferior se observan diferencias entre los promedios, unos están por encima del promedio general y otros por debajo. Por lo tanto, los valores de $\tau_j$ no son cero en todos los casos, de hecho algunos son positivos (como en los tratamientos 3 y 4 en que el promedio del tratamiento está sobre el promedio general) y otros negativos (como en los tratamientos 1 y 2 en que el promedio del tratamiento está debajo del promedio general), lo cual hace que la suma de todos los efectos siempre sea cero.

Cuando se cuenta con un solo factor se dice que se tiene un diseño de una vía y se utiliza un modelo matemático que se puede parametrizar de dos formas: 1) suma nula o 2) tratamiento referencia.

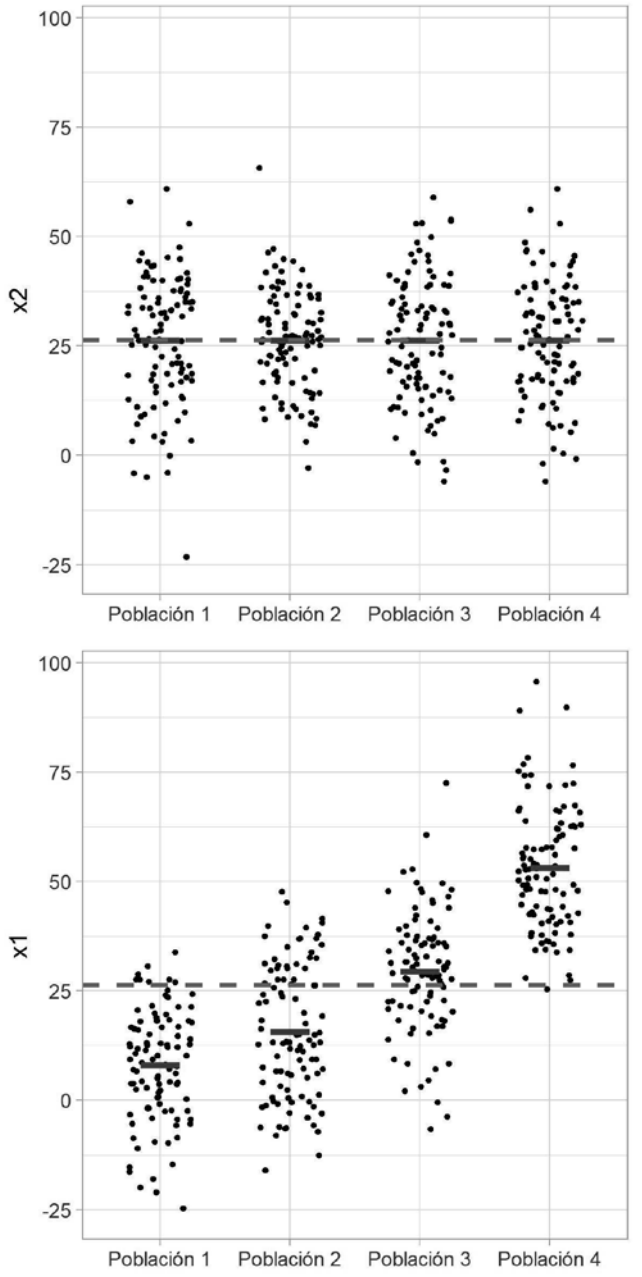


**Figura 1.1:** *Distribución de una variable con 4 tratamientos en dos situaciones*

*Nota: la línea discontinua es la media general, las líneas más angostas son las medias reales de cada población.*

### 1.1.1 Modelo de suma nula

Se asume que la suma de los coeficientes de todos los tratamientos es cero. En tal caso se introduce un coeficiente menos que la cantidad de niveles del factor, ya que el restante se obtiene por diferencia. Asumiendo que se toma el coeficiente del último tratamiento en función de los demás, y que hay $k$ tratamientos, esta restricción se puede expresar como:

$$\sum_{j=1}^{k} \tau_j = 0 \Rightarrow \tau_k = -\sum_{j=1}^{k-1} \tau_j.$$

El modelo se escribe de la siguiente forma:

$$\mu_j = \mu + \tau_j.$$

Cuando se usa esta parametrización para el modelo, el primer coeficiente representa la media general y los otros coeficientes son los efectos de los diferentes tratamientos. Tiene sentido pensar que la suma de todos los efectos sea cero, ya que el promedio general es la media de los promedios de todos los tratamientos y los efectos son las distancias de esos promedios respecto a la media general.

Se puede hacer uso de variables auxiliares para expresar el modelo de la forma que se usa para un modelo de regresión. Se requieren $k-1$ variables auxiliares $(C_1, ..., C_{k-1})$, una para cada uno de los primeros $k-1$ niveles del factor. Por ejemplo, si se tiene un caso de un factor con 4 niveles, y se cuenta con 2 observaciones en cada nivel, se requieren las variables auxiliares $C_1$, $C_2$ y $C_3$ definidas de la siguiente forma: $C_1$ toma valor 1 si la observación corresponde al nivel 1, $-1$ si corresponde al nivel $k$ (en este caso nivel 4) y 0 en otro caso (si corresponde al nivel 2 o 3); mientras que $C_2$ toma valor 1 si la observación es del nivel 2, sigue siendo $-1$ si es del nivel 4 y 0 si es del nivel 1 o 3. Finalmente, $C_3$ toma valor 1 si la observación es del nivel 3, sigue siendo $-1$ si es del nivel 4 y 0 si es del nivel 1 o 2. De esta forma, una observación que

es del nivel 4 va a tener valor $-1$ en las tres variables auxiliares. El cuadro 1.1 muestra la construcción de las variables auxiliares para este ejemplo.

**Cuadro 1.1:** *Variables auxiliares para un modelo de un factor con restricción de suma nula*

| Nivel | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 3 | 0 | 0 | 1 |
| 4 | -1 | -1 | -1 |
| 4 | -1 | -1 | -1 |

*Nota: el factor tiene cuatro niveles y hay dos observaciones por tratamiento.*

El modelo se escribe de la siguiente forma:

$$E[Y|Trat] = \mu + \sum_{j=1}^{k-1} \tau_j C_j = \mu + \tau_1 C_1 + \tau_2 C_2 + \tau_3 C_3.$$

### 1.1.2 Modelo de tratamiento referencia

Se toma uno de los tratamientos como referencia (vamos a tomar el primer tratamiento como referencia) y se define $\delta_j$ como la distancia del promedio del tratamiento j-ésimo al tratamiento de referencia, es decir, $\delta_j = \mu_j - \mu_1$, lo que hace que $\delta_1 = 0$. El modelo se escribe de la siguiente forma:

$$\mu_j = \mu_1 + \delta_j.$$

Esta forma de escribir el modelo mantiene el primer tratamiento como referencia y los coeficientes indican qué tanto se aleja cada promedio de $\mu_1$. Esta forma de parametrizar puede ser útil al realizar algunas comparaciones entre promedios.

Para obtener la expresión como un modelo de regresión, se requieren $k-1$ variables auxiliares $(D_2, ..., D_k)$ definidas de una forma diferente a la anterior. Para el mismo ejemplo, se requieren $D_2$, $D_3$ y $D_4$ definidas de la siguiente forma: $D_2$ toma valor 1 si la observación corresponde al nivel 2 y 0 si corresponde al nivel 1, 3 o 4; mientras que $D_3$ toma valor 1 si la observación es del nivel 3 y 0 si es del nivel 1, 2 o 4. Finalmente, $D_4$ toma valor 1 si la observación es del nivel 4 y 0 si es del nivel 1, 2 o 3. De esta forma, una observación que es del nivel de referencia, o sea, del nivel 1, va a tener valor 0 en las tres variables auxiliares. El cuadro 1.2 muestra la construcción de las variables auxiliares para este ejemplo.

**Cuadro 1.2:** *Variables auxiliares para un modelo de un factor con tratamiento 1 de referencia*

| Nivel | $D_2$ | $D_3$ | $D_4$ |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 |
| 4 | 0 | 0 | 1 |
| 4 | 0 | 0 | 1 |

*Nota: el factor tiene cuatro niveles y hay dos observaciones por tratamiento.*

El modelo se escribe de la siguiente forma:

$$E[Y|Trat] = \mu_1 + \sum_{j=2}^{k} \delta_j D_j = \mu_1 + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4.$$

## 1.2 Análisis de varianza

Cuando se hace una investigación basada en un experimento, se intenta llegar a conclusiones sobre el efecto que tiene un factor a partir de datos de muestras. Se puede pensar que los datos provienen de poblaciones

particulares, donde cada población tiene su propio promedio. Se intenta determinar si esos promedios podrían ser diferentes, lo cual estaría indicando que realmente existe un efecto del factor analizado sobre los promedios de esas poblaciones.

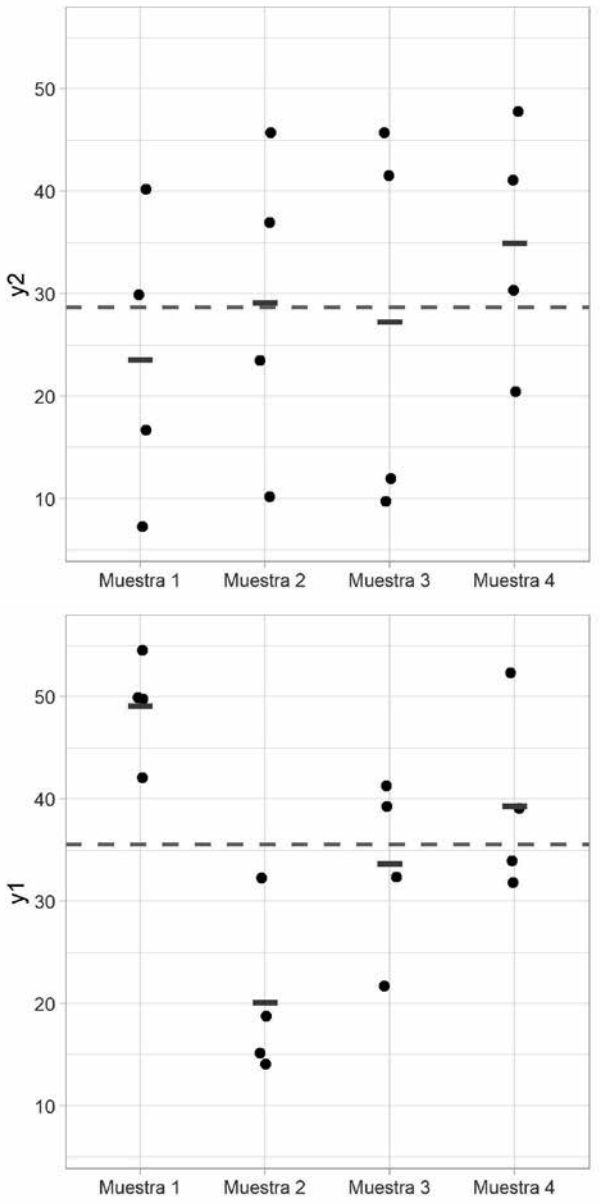


**Figura 1.2:** *Muestras de una variable con 4 tratamientos en dos situaciones*

*Nota: la línea discontinua es la media general, las líneas más angostas son las medias muestrales de cada tratamiento.*

Supongamos que se toman muestras de cada una de las poblaciones representadas en la Figura 1.1 (superior), donde se sabe que los promedios no son diferentes entre sí. En la Figura 1.2, se representan dos situaciones que podrían ocurrir a partir de dos experimentos realizados con esas poblaciones y con el mismo procedimiento, el cual consiste en extraer 4 valores de cada población. En el gráfico superior, se obtienen muestras que afortunadamente coinciden con la situación original y los promedios de cada una de esas muestras no se ven muy diferentes entre sí; sin embargo, en el inferior las observaciones dan promedios que son mucho más diferentes entre sí y llevan al investigador a pensar que el factor tiene un efecto sobre la respuesta promedio. Esta situación no es deseable, ya que el investigador estaría llegando a una conclusión errónea, la cual se conoce como error tipo I y que consiste en concluir que el factor tiene un efecto cuando en realidad no lo tiene. El error opuesto consiste en la situación en que las poblaciones sean como las del gráfico inferior de la Figura 1.1, en otras palabras, las medias son diferentes y las muestras resulten similares a las de la parte superior de la Figura 1.2. En ese caso, aunque el factor realmente tiene un efecto sobre la respuesta promedio, el investigador no logra demostrarlo porque los promedios observados no son muy diferentes e incurre en el error tipo II, que consiste en concluir que el factor no tiene un efecto cuando en realidad sí lo tiene.

Como en realidad no se sabe si los datos provienen de poblaciones con medias iguales o diferentes, se debe buscar un método que ayude a concluir si de verdad el factor tiene o no un efecto, pero basándose en la evidencia que le dan los datos recolectados. Para esto se plantea una hipótesis que establece que los promedios de todos los tratamientos son iguales y es equivalente a decir que los efectos de todos los tratamientos son iguales a cero, que es lo mismo que decir que el factor no tiene efecto sobre la respuesta promedio. Esto se puede expresar como:

$$H_0 : \mu_1 = \ldots = \mu_k \qquad \Leftrightarrow \qquad H_0 : \tau_1 = \ldots = \tau_k = 0$$

Esta hipótesis es independiente de cuál modelo se utilice (suma nula o tratamiento referencia) y se le llama hipótesis nula ($H_0$). Si los datos provienen de poblaciones como las de la Figura 1.1 (inferior), idealmente el investigador quisiera que sus datos le proporcionaran evidencia para rechazar esa hipótesis y de esta forma no cometería el error tipo II. En cambio, si provienen de poblaciones como las de la parte superior, querría no rechazar la hipótesis nula, pues si lo hace estaría cometiendo el error tipo I.

Para poner a prueba esta hipótesis se usa el análisis de varianza, método de descomposición de la variabilidad total de la respuesta en varias fuentes de variación. Para empezar, se considera la variabilidad total de la respuesta independientemente de los tratamientos a los que corresponde cada observación. Esta variabilidad se mide con la suma de cuadrados total (SCTot) y corresponde a la suma de las distancias al cuadrado de todas las respuestas respecto al promedio general. La SCTot se puede expresar como:

$$\text{SCTot} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

donde $\bar{y}$ representa la media general estimada de la respuesta y $n$ es la cantidad total de observaciones. La SCTot coincide con el numerador de la varianza de la respuesta, por lo que se puede obtener multiplicando la varianza de la respuesta por $n-1$. En el caso más simple, como es el diseño de una vía, el método de análisis de varianza consiste en descomponer la SCTot en dos partes: 1) suma de cuadrados de tratamiento (SCTrat) y 2) suma de cuadrados residual (SCRes):

$$\text{SCTot} = \text{SCTrat} + \text{SCRes}.$$

La SCTrat está relacionada con las distancias de los promedios observados entre sí. Para comparar estos promedios, lo que se hace más bien es comparar cada promedio con la media general. Si se resta el promedio de un tratamiento menos la media general, se obtiene una estimación del efecto

de ese tratamiento, es decir, $\hat{\tau}_j = \bar{y}_j - \bar{y}$, donde $\bar{y}_j$ es la media de la respuesta dentro del j-ésimo tratamiento. Se elevan al cuadrado los efectos estimados y se ponderan por el número de datos en el tratamiento correspondiente ($r_j$). La suma de todos estos efectos cuadráticos ponderados es la SCTrat y se puede expresar como:

$$\text{SCTrat} = \sum_{j=1}^{k} r_j \hat{\tau}_j^2 = \sum_{j=1}^{k} r_j (\bar{y}_j - \bar{y})^2.$$

La SCRes se obtiene al sumar todos los residuales elevados al cuadrado. Un residual es la distancia de una observación respecto a la media del tratamiento a la que ella pertenece, dicho de otro modo, $e_{ij}$ es el i-ésimo residual en el j-ésimo tratamiento y se obtiene mediante $e_{ij} = y_{ij} - \bar{y}_j$ que es la distancia de la i-ésima observación del j-ésimo tratamiento $(y_{ij})$ a la media de ese tratamiento. De forma similar a lo que sucede con la SCTot, la suma de los residuales cuadráticos en el j-ésimo tratamiento es equivalente a la varianza de la respuesta en ese tratamiento ($s_j^2$) multiplicada por los grados de libertad asociados ($r_j - 1$). De esta forma se obtiene que la suma de los residuales cuadráticos en todos los tratamientos es:

$$\text{SCRes} = \sum_{j=1}^{k} \sum_{i=1}^{r_j} e_{ij}^2 = \sum_{j=1}^{k} \sum_{i=1}^{r_j} (y_{ij} - \bar{y}_j)^2 = \sum_{j=1}^{k} (r_j - 1) s_j^2.$$

A partir de las sumas de cuadrados se obtienen los cuadrados medios, los cuales son medidas de variabilidad y se obtienen al dividir cada suma de cuadrados entre sus grados de libertad. El cuadrado medio de tratamiento (CMTrat) es una medida de la variabilidad entre las medias de los tratamientos, equivalente a decir que es una medida de la magnitud general de los efectos. Si los efectos son muy pequeños, el CMTrat va a dar muy cercano a cero. Para el CMTrat se tienen $k - 1$ grados de libertad. Después, el cuadrado medio residual (CMRes) es la medida de la variabilidad de la respuesta dentro de cada tratamiento y tiene $n - k$

grados de libertad, donde $n$ es el número total de observaciones en todos los tratamientos. La suma de los grados de libertad de las dos fuentes de variación es igual a $n-1$ $(k-1+n-k=n-1)$ . Por lo tanto, se tiene que:

$$\text{CMTrat} = \frac{\text{SCTrat}}{k-1} \qquad y \qquad \text{CMRes} = \frac{\text{SCRes}}{n-k}.$$

El CMRes también se puede calcular a partir de las varianzas de los tratamientos, como una media ponderada de las varianzas de cada tratamiento donde se pondera con los grados de libertad de cada varianza; en cambio, si se tiene el mismo número de observaciones en todos los tratamientos el CMRes se obtiene como un promedio simple de las varianzas.

El razonamiento detrás de la prueba de una hipótesis consiste en encontrar una probabilidad de cometer error tipo I (rechazar la hipótesis nula cuando es cierta) y compararla contra un máximo previamente establecido para esta probabilidad. Este máximo se conoce como nivel de significancia y se denomina $\alpha$. Si la probabilidad estimada de cometer error tipo I es suficientemente baja, en otras palabras, no supera el nivel de significancia, se decide rechazar la hipótesis nula, en caso contrario, no se rechaza.

Para rechazar la hipótesis nula de igualdad de medias debería observarse que las medias de los diferentes tratamientos estén bastante alejadas unas de otras; no obstante, esta lejanía es relativa y debe contrastarse con la variabilidad que tienen los datos dentro de cada tratamiento. En la Figura 1.3, se muestran dos casos que presentan la misma separación entre los promedios; sin embargo, en el gráfico superior hay poca variabilidad dentro de cada tratamiento, por lo que la separación entre las medias se hace más evidente y se pensaría que la hipótesis nula sea falsa. Al contrario, en la parte inferior esa misma separación no parece tan importante debido a la alta variabilidad dentro de cada tratamiento; por lo que posiblemente no se llegue a rechazar la hipótesis nula.

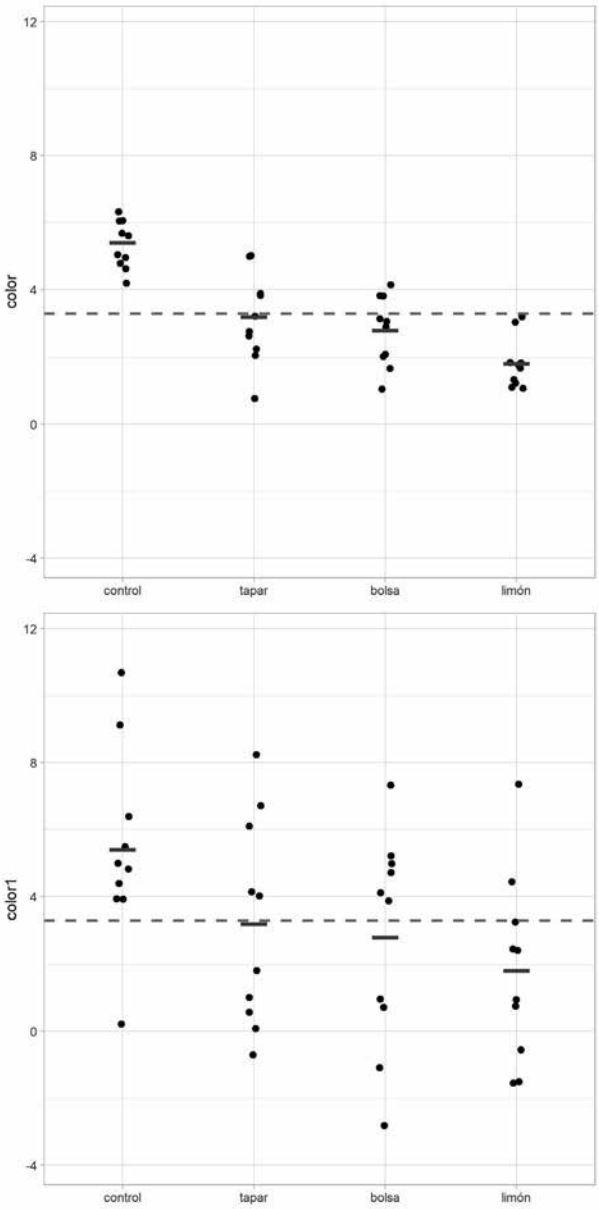


**Figura 1.3:** *Dos situaciones con igual separación entre medias*

*Nota: arriba con menos variabilidd y abajo con más variabilidad.*

El estadístico $F$ observado se construye con el fin de tener una medida objetiva de esta comparación y consiste en el cociente entre el CMTrat y el CMRes. Si este cociente es suficientemente alto se puede rechazar la hipótesis planteada, de lo contrario, no se rechaza. Con la ayuda de la distribución $F$

se obtiene una probabilidad asociada al error tipo I, que es la probabilidad de encontrar un valor mayor al estadístico $F$ observado en la distribución $F$ con $k-1$ y $n-k$ grados de libertad. Esta probabilidad se compara contra el nivel de significancia establecido previamente.

## 1.3 Caso de un factor con dos niveles

Cuando el factor que se analiza tiene solo dos niveles, la prueba de la hipótesis se puede realizar utilizando la distribución $t$ y se obtiene un resultado equivalente al obtenido con el análisis de varianza. En este caso, la hipótesis nula se reduce a:

$$H_0 : \mu_1 = \mu_2 \qquad \Leftrightarrow \qquad H_0 : \tau_1 = 0$$

La hipótesis alternativa es una hipótesis de dos colas:

$$H_0 : \mu_1 \neq \mu_2$$

Se debe obtener la estimación del efecto ($\hat{\tau_1}$) y su desviación estándar ($ee_{\hat{\tau_1}}$), con ellos se calcula el estadístico $t$ observado mediante:

$$t = \frac{\hat{\tau_1}}{ee_{\hat{\tau_1}}}.$$

Con la ayuda de la distribución $t$ se obtiene una probabilidad asociada al error tipo I, que es la probabilidad de encontrar un valor mayor al estadístico $t$ observado, en la distribución $t$ con $n-k$ grados de libertad. Esta probabilidad debe ser igual a la que se obtiene con el estadítico $F$ observado en la distribución $F$ explicada anteriormente.

## 1.4 Manzanas

Las manzanas tienen un compuesto llamado polifenol oxidasa, el cual hace que al cortarse y entrar en contacto con el aire se oscurezcan rápidamente. Para evitar el pardeamiento se probaron tres tratamientos: tapar (código 2), poner en bolsa plástica cerrada (código 3) y aplicar jugo de limón (código 4). Además, se incluyó un control sin aplicar nada (código 1). Se seleccionan 40 manzanas y a cada una se le aplica aleatoriamente uno de los 4 tratamientos, lo cual resulta en 10 manzanas para cada tratamiento. Una vez aplicado el tratamiento a cada manzana, se pide a 3 jueces que califiquen el color en una escala de 1 a 6, donde 1 es el color normal de la fruta y 6 es el más oscuro. Cada manzana recibe como calificación el promedio de los 3 jueces. El objetivo final es seleccionar el tratamiento que mantenga mejor el color original para una empresa que se encarga de banquetes.

En un primer análisis solo se va a investigar si existe alguna diferencia en el color promedio resultante con los cuatro tratamientos.

### 1.4.1 Ejercicios

1. Preparación:
   (a) Lea el archivo `manzanas.csv` en R.
   (b) Defina correctamente el factor y ponga las etiquetas correspondientes para cada uno de los tratamientos.
   (c) Guarde la base en un archivo llamado `manzanas.Rdata` para ser utilizado en futuros ejercicios.

2. Análisis gráfico:
   (a) Obtenga una tabla con los promedios de cada tratamiento y llámela `m`. Use:
   `tapply(y,x,mean)`.
   (b) Obtenga una tabla con las varianzas por tratamiento y llámela `v`.

(c) Obtenga la media general de la respuesta y llámela `media`.

(d) Haga un boxplot para analizar el efecto de los tratamientos sobre la respuesta promedio. Agregue la media general usando `abline(h=media,col=2)` y las medias de los tratamientos con `points(1:4,m,col=4,pch="-",cex=2)`.

(e) Obtenga los efectos muestrales de cada tratamiento a partir de la tabla de medias y compare estos resultados con lo que ve en el gráfico. Cada efecto se puede estimar como: $\hat{\tau}_j = \bar{y}_j - \bar{y}$.

(f) Explique el significado de cada uno de los valores obtenidos para los efectos muestrales.

(g) Obtenga la suma de los efectos anteriores.

(h) Obtenga una estimación de la varianza del error a partir de la tabla de varianzas. La estimación debe ser la media ponderada de las varianzas en los tratamientos, las cuales se ponderan con los grados de libertad de cada varianza; no obstante, en este caso se tiene el mismo número de réplicas en todos los tratamientos, por lo que basta hacer un promedio simple de las varianzas.

3. Análisis de varianza:

(a) Ajuste un modelo lineal. Use tanto la función `aov` como la función `lm`. La diferencia principal es que con `lm` se pueden obtener los coeficientes del modelo, mientras que con `aov` se puede obtener la tabla de efectos. En todo caso, cuando usa `lm`, por ejemplo `mod=lm(y~x)`, luego puede obtener `mod1=aov(mod)` de la misma forma que haciendo `mod1=aov(y~x)`.

(b) Obtenga los resultados del análisis de varianza mediante `anova(mod)` o `anova(mod1)`. Si usa la función `aov` da lo mismo usar `summary(mod1)` o `anova(mod1)`.

(c) Observe la línea de residuales para obtener el cuadrado medio residual y compárelo con la estimación de la varianza del error obtenida en el punto anterior.

(d) Observe los grados de libertad residuales y justifique por qué se obtiene ese número.

(e) Observe la línea del tratamiento y obtenga la suma de cuadrados de tratamiento.

(f) Haga la suma de los cuadrados de los efectos obtenidos anteriormente. Observe que estos cuadrados deben multiplicarse por el número de réplicas para obtener exactamente la suma de cuadrados de tratamiento. Justifique por qué esto debe ser así.

(g) Compare la variabilidad de los promedios con la variabilidad residual para determinar si hay alguna evidencia de diferencias entre las medias de la respuesta.

(h) Establezca adecuadamente la hipótesis que está poniendo a prueba y dé una conclusión.

4. Estimación de parámetros del modelo de tratamiento referencia:

(a) Obtenga las estimaciones de los parámetros del modelo. Por *default* R usa el modelo de tratamiento referencia. Esto se logra con el ajuste hecho con `lm` mediante `summary(mod)` o `mod$coef`.

(b) ¿Qué significa el intercepto en este modelo?

(c) ¿Qué representa cada uno de los coeficientes del modelo?

(d) Obtenga la matriz de estructura y observe la codificación de las variables auxiliares.

(e) A partir de los coeficientes obtenidos, obtenga los efectos muestrales y compárelos con los obtenidos en el punto 2(e).

(f) Obtenga los efectos directamente con `model.tables(mod)` (solo funciona si el modelo fue hecho con la función `aov`).

5. Modelo de suma nula:

   (a) Cambie al modelo de **suma nula** usando la siguiente instrucción: `options(contrasts=c("contr.sum","contr.poly")).`

   Para volver al modelo de **tratamiento referencia** se usa: `options(contrasts=c("contr.treatment","contr.poly")).`

   (b) Verifique la codificación con `contrasts(base$trat).`

   (c) Repita los pasos del punto 4. Compare los resultados.

6. Factor con dos niveles:

   (a) Para ilustrar el caso cuando el factor tiene solo dos niveles, haga una base que contenga solo los datos que corresponden al nivel 1 y 2, llámela `base1`.

   (b) Para eliminar los niveles que no tienen datos haga `base1$trat=factor(as.numeric(base1$trat).`

   (c) Ajuste el modelo con `lm`. Obtenga el análisis de varianza y observe la probabilidad asociada a la hipótesis de igualdad de medias.

   (d) Obtenga el `summary`, extraiga la estimación del efecto y su error estándar, verifique el valor de $t$ y obtenga la probabilidad asociada en la distribución $t$. Compare el resultado con el análisis de varianza.

### 1.4.2 Solución

1. Preparación:

   (a) Lectura:

```
base=read.csv("manzanas.csv",sep=";")
```

   (b) Definición de factor:

```
base$trat=factor(base$trat)
levels(base$trat)=c("control","tapar","bolsa","limón")
base$trat
```

```
## [1] tapar tapar tapar tapar tapar tapar tapar tapar
## [9] tapar tapar bolsa bolsa bolsa bolsa bolsa bolsa
## [17] bolsa bolsa bolsa bolsa limón limón limón limón
## [25] limón limón limón limón limón limón control control
## [33] control control control control control control control
## Levels: control tapar bolsa limón
```

(c) Almacenar la base:

```
save(base,file="manzanas.Rdata")
```

2. Análisis gráfico:

(a) Tabla con las medias:

```
(m=tapply(base$color,base$trat,mean))
```

```
## control tapar bolsa limón
##     5.4   3.2   2.8   1.8
```

(b) Tabla con las varianzas:

```
(v=tapply(base$color,base$trat,var))
```

```
## control tapar bolsa limón
##    0.49  1.73  1.07  0.62
```

(c) Media general:

```
(media=mean(base$color))
```

```
## 3.3
```

(d) Boxplot:

```
boxplot(color~trat,ylab="color",xlab="tratamiento",data=base)
abline(h=media,lty=2)
points(1:4,m,pch="-",cex=2)
```

En la Figura 1.4 se nota que los puntajes de color en esta muestra son más bajos en los tres tratamientos que en el control. Cuando se aplicó limón estos puntajes tienden a ser más bajos que cuando se cubrió de alguna forma. También se nota que los dos tratamientos en que se cubrió producen resultados muy similares.

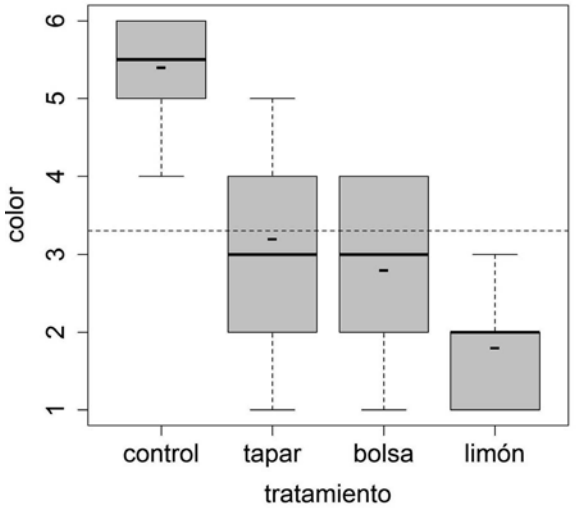


**Figura 1.4:** *Puntajes de color por tratamiento*

*Nota: la línea discontinua es la media general, las líneas más anchas dentro de cada caja son las medianas de cada tratamiento y las líneas más angostas son las medias.*

(e) Efectos muestrales:

```
(ef=m-media)
```

```
## control tapar bolsa limón
##     2.1  -0.1  -0.5  -1.5
```

Estos números coinciden con el gráfico, puesto que los valores negativos concuerdan con aquellas medias que están por debajo de la media general y el valor positivo del control concuerda con el gráfico en que su media está por encima de la media general.

(f) Significado:

El control tiene una media que está 2,1 puntos sobre la media general, por lo que se dice que el control tiene el efecto de subir la media de la escala de color 2,1 puntos. El limón produce una media 1,5 puntos por debajo de la media general, es decir, tiene el efecto de bajar la media 1,5 puntos. Similarmente, los dos tratamientos en que se cubre tienen un leve efecto sobre la media ya que la bajan muy poco.

(g) Suma de los efectos:

```
sum(ef)
```

```
## 1.110223e-15
```

Aunque no da exactamente cero, esto se debe a un asunto computacional pero la suma de los efectos debe ser siempre cero por su misma construcción.

(h) Estimación de la varianza del error:

Primero se obtiene el número de réplicas en cada tratamiento y se observa que el diseño es balanceado, o sea, que tiene el mismo número de réplicas en todos los tratamientos.

```
(r=table(base$trat))
```

```
## control   tapar   bolsa   limón
##      10      10      10      10
```

La estimación de la varianza del error se obtiene al ponderar las varianzas de los tratamientos por los grados de libertad de cada tratamiento (número de réplicas menos 1).

```
(v1=sum((r-1)*v)/(sum(r)-4))
```

```
## 0.98
```

Puesto que el diseño es balanceado se obtiene el mismo resultado si simplemente se promedian las varianzas de los cuatro tratamientos.

```
(v2=mean(v))
```

```
## 0.98
```

# Acerca del autor

**Ricardo Alvarado Barrantes** obtuvo la licenciatura en la Universidad de Costa Rica, la maestría en la Universidad de Michigan, EE.UU., y el doctorado en Estadística en la Universidad de Padua, Italia. En los últimos 17 años se ha dedicado a la docencia y a la investigación en la Escuela de Estadística de la Universidad de Costa Rica, donde se ha especializado en modelos estadísticos.


Corrección filológica: *Kendy Valverde V.* • Revisión de pruebas: *Sherlyn Jiménez B.*
Diseño de contenido y diagramación: *El autor* • Diseño de portada: *Boris Valverde G.*
Imagen de portada: creada a partir de inteligencia artificial. • Control de calidad: *Grettel Calderón A.*

Editorial UCR es miembro del Sistema Editorial Universitario Centroamericano (SEDUCA),
perteneciente al Consejo Superior Universitario Centroamericano (CSUCA).

Impreso bajo demanda en la Sección de Impresión del SIEDIN.
Setiembre, 2025.

CIENCIAS SOCIALES | Estadística

Este libro presenta los fundamentos de los diseños experimentales en estadística, partiendo del diseño con un solo factor y la comparación de pares de promedios. Se exponen diseños factoriales con dos o más factores. Además, se presenta el diseño de bloques aleatorizados y se mencionan los modelos mixtos con la inclusión de parcelas divididas. Luego, se estudia el uso de covariables que permiten reducir la variabilidad del error experimental. Finalmente, se profundiza en el concepto de potencia de las pruebas estadísticas y en la técnica de simulaciones. A lo largo del libro se usan datos reales para ilustrar cómo hacer el análisis con el software estadístico R.

EDITORIAL UCR

ISBN: 978-9968-02-276-7