Universidad Autónoma de Yucatán

Facultad de Ingeniería Química

Notas de Estadstica Aplicada

Elaborado por:

M. en C. Delta M. Sosa Cordero
MCM. Noé G. Chan Chí

Enero - Mayo 2020

Contenido

1 Diseño de Experimentos y optimización

1.1. Introducción

La Estadística interviene en la investigación a través de la experimentación y observación. El uso adecuado de la Estadística hace más eficiente la investigación y por lo tanto, como herramienta importante de investigación no puede ser ajena a la planeación general del proyecto de investigación.

El método estadístico que será usado en el análisis de los datos obtenidos, debe formar parte importante del diseño total.

Sin embargo, aunque parece evidente, es frecuente todavía observar proyectos en áreas de investigación que reportan muchos datos, obtenidos fortuitamente, sin idea clara de para qué les servirá o cómo los analizarán.

En esos casos, mucho esfuerzo se desperdicia porque no hay manera legítima de respaldar las hipótesis que los métodos suponen para que las conclusiones sean válidas.

Es necesario, por lo tanto, además de adquirir conocimientos generales de los métodos más importantes, tener en cuenta las etapas principales de una investigación estadística. Es una guía general, planteada de diferentes formas, manteniendo lineamientos comunes en casi todos los títulos sugeridos en libros de consulta.

1. **Formulación del Problema** Para investigar con éxito un problema se deben definir conceptos precisos, formular preguntas claras e imponer límites adecuados al problema, tomando en cuenta tiempo y dinero disponibles y la habilidad de los investigadores.

   Si no se logra una buena formulación, se pueden obtener datos incorrectos o irrelevantes.

   La calidad de las conclusiones estadísticas depende de la precisión de los datos y éstos a su vez dependen de la exactitud y correcta formulación del problema.

2. **Diseño del Experimento.** Siempre se desea obtener un máximo de información empleando un mínimo de costos y tiempo.

   Esto implica determinar entre otras cosas un tamaño de muestra , o la cantidad y tipo de datos que sean eficientes en la resolución del problema. Por lo que la selección del método de análisis de los datos y la técnica de muestreo son de gran importancia.

Algunas veces, no se logran obtener muestras completamente aleatorias. Sin embargo, la representación de la muestra es un punto esencial.

También es importante definir el tipo de datos, cuántos y cómo deben obtenerse. La planificación y el diseño experimental tienen por objetivo principal garantizar lo más posible que las conclusiones sean válidas.

3. **Recolección de datos o Experimentación.** Esta es la parte que implica mayor tiempo y trabajo en toda investigación que se realiza. Debe sujetarse a las reglas y acuerdos discutidos en los pasos previos. Una buena recolección de datos que logre obtener información pertinente y verídica producirá mejores y calidad en los resultados.

4. **Clasificación y descripción de resultados.** Es importante que los datos obtenidos se presenten y se representen en forma clara así como las medidas de descripción como promedios o medidas de dispersión adecuadas. La sencillez y claridad en un trabajo reflejan todo el tiempo invertido y la planeación efectuada.

5. **Inferencia estadística y conclusiones.** La aplicación de los métodos estadísticos, deben incluir una clara inferencia a la población estudiada, con el respaldo de la estadística se toman decisiones basadas en las observaciones obtenidas en la muestra y desde luego se formulan respuesta al problema planteado.

Clasificación de la estadística

La Estadística en general, puede considerarse en dos grandes ramas:

a) La Estadística Paramétrica, entre otros temas se refiere a técnicas para estimar y probar hipótesis acerca de parámetros de la población de estudio.

   Se requiere que los datos sean producto de mediciones y de naturaleza cuantitativa.

   Las técnicas están basadas en suponer que las poblaciones siguen distribución normal, t de Student o binomial, entre otras.

b) La Estadística No Paramétrica que se refiere a técnicas cuya validez no se basa en la distribución de probabilidades de las poblaciones. Las hipótesis que se prueban no incluyen por lo tanto, afirmaciones acerca de parámetros y los datos son resultado de conteos de frecuencia o se asignan rangos.

   Es importante mencionar que si es posible aplicar ambos enfoques, se recomienda la paramétrica, pues los métodos presentan mayor eficiencia.

Entre las escalas de medida se encuentran:

Escala Nominal. Son medidas cualitativas o nombres sin jerarquía, sólo indican diferencia de atributos. Ejemplos: enfermedades, números asignados a equipos, personal de trabajo, etc.

Escala Ordinal. Son medidas que indican cierta jerarquía y distinguen un objeto o persona con respecto a otras si tienen más de una característica, pero sin conocer la cantidad de diferencia entre éstos. Ejemplo: los pacientes en convalecencia pueden considerarse: sin mejoría, mejorando y bastante mejorados.

Escala de Intervalo. Cuando sí se conoce la diferencia o cantidad entre las medidas asignadas a los objetos o personas. El uso de una distancia unitaria arbitraria y el cero no necesariamente es un cero verdadero, en el sentido de indicar ausencia total de lo que se mide. Ejemplo: Temperatura en grados F, $50°F$, $55°F$, $70°F$.

Escala de Razón. Permite conocer cuántas veces es más grande la medida asignada a cada objeto o personas. Se caracteriza por considerar un punto cero verdadero. Ejemplo: El ingreso económico de A es el doble del de B. Cero es la ausencia de ingreso. Altura, peso, longitud pueden ser usados en este tipo de escala.

1.2. Principios del diseño de experimentos

El *diseño estadístico de experimentos*, es el proceso de planear un experimento para obtener datos apropiados que se analizan con métodos estadísticos con el objetivo de obtener conclusiones válidas y objetivas.

Un *experimento diseñado* es una prueba o secuencia de pruebas en las que se inducen cambios deliberados en las variables independientes o de entrada del proceso, para observar e identificar las causas de los cambios en la respuesta o variable dependiente.

Entre los *objetivos* del diseño de experimentos, se pueden considerar, desarrollar un proceso consistente o robusto; esto significa que las variables no controladas afectarán en forma mínima al proceso. Mejorar el rendimiento de un proceso. También desarrollar nuevos procesos o un nuevo producto. Un objetivo muy importante, es desear optimizar o caracterizar el proceso determinando qué factores, controlados o no, influyen en alguna característica.

Se les llama *tratamientos*, a las variables independientes o variables controladas.

En los experimentos, es frecuente considerar diferentes *niveles* de un tratamiento definidos en forma *cualitativa* (tipos de tratamientos) o en forma *cuantitativa* (cantidades de tratamiento).

De hecho, cada nivel es en sí un tratamiento.

La variable dependiente o respuesta es el efecto que ocasionan los tratamientos.

En cualquier experimento existen variables *exógenas*, son las que tienen posible efecto sobre la variable respuesta, sin embargo no son de interés en el estudio. El investigador, además de identificarlas, tiene un gran interés en controlarlas.

Se llama *unidad experimental*, al sujeto o la entidad más pequeña donde se aplica el tratamiento.

Las medidas son cada uno de los valores de la variable respuesta, o sea el efecto del tratamiento sobre cada una de las unidades experimentales.

Las unidades experimentales expuestas al mismo tratamiento presentan en general, diferentes medidas.

Se llama *error experimental* a la variabilidad de esas medidas.

Entre la causas más comunes del error experimental, se encuentra:

a) La variable respuesta es una variable aleatoria.

b) Diferencias inherentes de las unidades experimentales.

c) Ausencia de uniformidad durante el proceso del experimento.

Entre los puntos más importantes, en el diseño de experimentos, se pueden considerar los señalados en Guía para el Diseño de Experimentos. (Montgomery, 1991)

1. Comprensión y planteamiento del problema.

2. Elección de factores y niveles.

3. Selección de la variable respuesta.

4. Elección del Diseño Estadístico.

5. Realización del Experimento.

6. Análisis de Datos.

7. Conclusiones y Recomendaciones.

1.2.1. Directrices para el diseño de experimentos

El diseño estadístico de experimentos y el análisis estadístico de los datos son dos aspectos muy relacionados. Algunos autores incluyen la selección del método estadístico que se pretende usar en el análisis objetivo de datos sujetos a errores experimentales, en el diseño empleado.

Entre los principios básicos del diseño de experimentos, es importante tener en cuenta:

1) **Obtención de réplicas**. Esto es repeticiones del experimento básico. Si se tienen tres tratamientos, una réplica estará formada por tres unidades experimentales en las que se han aplicado cada uno de los tratamientos.

   Se dice que un tratamiento o nivel de tratamiento se repite cuando se aplica a más de una unidad experimental. Estas repeticiones se usan para calcular una estimación del error experimental.

2) La **aleatorización** en experimentos fundamenta el uso de los métodos estadísticos en el Diseño de Experimentos.

   Se refiere a la asignación aleatoria de las unidades experimentales a los tratamientos y al orden en la realización de las pruebas.

   Es importante cuando se requiere el supuesto de que las mediciones sean valores de variables aleatorias independientes.

   Permite promediar los efectos sistemáticos entre los diferentes niveles y reducir el error experimental.

   Reduce el efecto de variables exógenas.

3) **Análisis por Bloques**. Un bloque es un subconjunto o porción de material experimental más homogéneo, comparado con el total. Incrementa la precisión del experimento.

1.2.2. Diseño experimental completamente aleatorizado (DCA) de un factor

1.2.3. Análisis de varianza

En este capítulo se desarrollarán métodos para contrastar las diferencias entre las medias de más de dos poblaciones. Puede ser de interés comparar tres o más máquinas, varios proveedores, cuatro procesos, tres materiales, cinco dosis de un fármaco, etcétera.

Por ejemplo, una comparación de cuatro dietas de alimentación en la que se utilizan ratas de laboratorio, se hace con el fin de estudiar si alguna nueva dieta que se propone es mejor o igual que las ya existentes; en este caso, la variable de interés es el peso promedio alcanzado por cada grupo de animales después de ser alimentado con la dieta que le tocó.

Por lo general, el interés del experimentador está centrado en comparar los tratamientos en cuanto a sus medias poblacionales, sin olvidar que también es importante compararlos con respecto a sus varianzas.

Desde el punto de vista estadístico

Hipótesis.

La hipótesis fundamental a probar cuando se comparan varios tratamientos es

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu \quad \text{vs} \quad H_A : \mu_i \neq \mu_j \text{ para algún } i \neq j$$

con la cual se quiere decidir si los tratamientos son iguales estadísticamente en cuanto a sus medias, frente a la alternativa de que al menos dos de ellos son diferentes.

El método capaz de probar la hipótesis de igualdad de las k medias con un solo estadístico de prueba, es el denominado análisis de varianza

El análisis de varianza (ANOVA) se puede definir como una técnica mediante la cual la variación total presente en un conjunto de datos se divide en varias componentes, cada una de ellas tiene asociada una fuente de variación específica, de manera que en el análisis es posible conocer la magnitud de las contribuciones de cada fuente de variación a la variación total.

El desarrollo del análisis de varianza se debe principalmente a los trabajos de Sir R. A. Fisher realizados entre 1912 y 1962.

El análisis de varianza se aplica ampliamente en la investigación, pues está íntimamente relacionado con el diseño experimental. La relación entre estos dos tópicos se puede resumir diciendo, que cuando se diseña un experimento el cual queremos someter a un análisis, los investigadores pueden, antes de llevar a cabo su investigación, identificar aquellas fuentes de variación que consideran importantes y pueden seleccionar un modelo que les permita medir la extensión de la contribución de esas fuentes a la variación total.

En el análisis de varianza a las variables se les suele llamar *factores* y a los diferentes niveles de cada variable o factor se les llama *tratamientos* o categorías.

Las suposiciones en el análisis de varianza son fundamentalmente:

- la normalidad de las distribuciones,
- igualdad de varianzas,
- muestreo aleatorio y
- datos de tipo cuantitativo.

Muchas comparaciones, como las antes mencionadas, se hacen con base en el diseño completamente al azar (DCA), que es el más simple de todos los diseños que se utilizan para comparar dos o más tratamientos, dado que sólo consideran dos fuentes de variabilidad: los *tratamientos* y el *error aleatorio*.

Este diseño se llama *completamente al azar* porque todas las corridas experimentales se realizan en orden aleatorio completo. De esta manera, si durante el estudio se hacen en total N pruebas, éstas se corren al azar, de manera que los posibles efectos ambientales y temporales se vayan repartiendo equitativamente entre los tratamientos.

Supongamos que se tienen k poblaciones o tratamientos, independientes y con medias desconocidas μ_1, μ_2, ..., μ_k, así como varianzas también desconocidas pero que se suponen 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.

Los datos generados por un diseño completamente al azar para comparar dichas poblaciones se pueden escribir como en la tabla **1.1**

El elemento Y_{ij} en esta tabla es la j-ésima observación que se hizo en el tratamiento i; n_i es el tamaño de la muestra o las repeticiones observadas en el tratamiento i. Es recomendable

Tratamientos				
T_1	T_2	T_3	$\cdots$	T_k
Y_{11}	Y_{21}	Y_{31}	$\ldots$	Y_{k1}
Y_{12}	Y_{22}	Y_{32}	$\ldots$	Y_{k2}
Y_{13}	Y_{23}	Y_{33}	$\ldots$	Y_{k3}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Y_{1n_1}	Y_{2n_2}	Y_{3n_3}	$\ldots$	Y_{kn_k}

Tabla 1.1: Diseño completamente al azar

utilizar el mismo número de repeticiones ($n_i = n$) en cada tratamiento, a menos que hubiera alguna razón para no hacerlo[1]. Cuando $n_i = n$ para toda i se dice que el diseño es *balanceado*.

El número de tratamientos k es determinado por el investigador y depende del problema particular de que se trata. El número de observaciones por tratamiento (n) debe escogerse con base en la variabilidad que se espera observar en los datos, así como en la diferencia mínima que el experimentador considera que es importante detectar. Con este tipo de consideraciones, 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 **1.1** se podrán describir con el modelo estadístico lineal dado por:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

donde μ es el parámetro de escala común a todos los tratamientos, llamado *media global*, τ_i es un parámetro que mide el efecto del tratamiento i y ε_{ij} es el error atribuible a la medición Y_{ij}. Este modelo implica que en el diseño completamente al azar actuarían a lo más dos fuentes de variabilidad: los *tratamientos* y el *error aleatorio*.

El análisis de varianza (ANOVA) es la técnica central en el análisis de datos experimentales. La idea general de esta técnica es separar la variación total en las partes con las que contribuye cada fuente de variación en el experimento. En el caso del DCA se separan la variabilidad debida a los tratamientos y la debida al error. Cuando la primera predomina "claramente" sobre la segunda, es cuando se concluye que los tratamientos tienen efecto (figura **1.1** a)), o dicho de otra manera, las medias son diferentes. Cuando los tratamientos no dominan contribuyen igual o menos que el error, por lo que se concluye que las medias son iguales (figura **1.1** b))

Antes de comenzar con el análisis del Diseño Completasmente al Azar (DCA) se introduce la notación que simplifica la escritura de las expresiones involucradas en dicho análisis

[1] Si uno de los tratamientos resulta demasiado caro en comparación con los demás, se pueden plantear menos pruebas con éste. Por otra parte, cuando uno de los tratamientos es un control (tratamiento de referencia) muchas veces es el más fácil y económico de probar, y como es de interés comparar a todos los tratamientos restantes con el control, se recomienda realizar más corridas en éste para que sus parámetros queden mejor estimados.

(a) No hay efecto de tratamiento

(b) Si hay efecto de tratamiento

Figura 1.1: División de la variabilidad total en sus componentes

Notación de puntos

Sirve para representar de manera abreviada cantidades numéricas que se pueden calcular a partir de los datos experimentales, donde Y_{ij} representa la j-ésima observación en el tratamiento i, con $i = 1, 2, \ldots, k$ y $j = 1, 2, \ldots, n_i$. Las cantidades de interés son las siguientes:

- $Y_{i\bullet}$: Suma de las observaciones del tratamiento i.
- $\overline{Y}_{i\bullet}$: Media de las observaciones del tratamiento i.
- $Y_{\bullet\bullet}$: Suma total de las $N = n_1 + n_1 + \ldots + n_k$ mediciones
- $\overline{Y}_{\bullet\bullet}$: Media global o promedio de todas las observaciones

Note que el punto indica la suma sobre el correspondiente subíndice. Algunas de estas relaciones son:

$$Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}; \qquad \overline{Y}_{i\bullet} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}; \qquad Y_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}; \qquad \overline{Y}_{\bullet\bullet} = \frac{Y_{\bullet\bullet}}{N}$$

donde $N = \sum_{i_1}^{k} n_i$ es el total de observaciones.

Hipótesis.

El objetivo del análisis de varianza en el DCA es probar la hipótesis de igualdad de los tratamientos con respecto a la media de la correspondiente variable de respuesta:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu \quad vs \quad H_A : \mu_i \neq \mu_j \text{ para algún } i \neq j \tag{1.1}$$

Para probar la hipótesis dada por las relaciones 1.1 mediante la técnica de ANOVA, lo primero es descomponer la variabilidad total de los datos en sus dos componentes: la variabilidad debida a tratamientos y la que corresponde al error aleatorio, como se hace a continuación.

Una medida de la variabilidad total presente en las observaciones de la tabla **1.1** es la **suma total de cuadrados** dada por

$$SC_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{\bullet\bullet}}{N}$$

Las fuentes de variabilidad tratamiento y error surgen al hacer el siguiente arreglo algebraico: al sumar y restar adentro del paréntesis la media del tratamiento i, $(\overline{Y}_{i\bullet})$:

$$SC_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(Y_{ij} - \overline{Y}_{i\bullet}) + (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})\right]^2$$

y desarrollando el cuadrado, la SC_T se puede partir en dos componentes como:

$$SC_T = \sum_{i=1}^{k} n_i(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2$$

donde el primer componente es la *suma de cuadrados de tratamientos* ($\mathbf{SC_{trat}}$) y el segundo es la *suma de cuadrados del error* ($\mathbf{SC_e}$). Al observar con detalle estas sumas de cuadrados se aprecia que:

- La SC_{trat} mide la variación o diferencias entre tratamientos, ya que si éstos son muy diferentes entre sí, entonces la diferencia $\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}$ tenderá a ser grande en valor absoluto, y con ellos también será grande la SC_{trat}.

- Mientras que la SC_e mide la variación *dentro de tratamientos*, ya que si hay mucha variación entre las observaciones de cada tratamiento entonces $Y_{ij} - \overline{Y}_{i\bullet}$ tenderán a ser grande en valor absoluto.

- En forma abreviada, esta descomposición de la suma total de cuadrados se puede escribir como

$$SC_T = SC_{\text{trat}} + SC_e \tag{1.2}$$

- **Grados de libertad**. Como hay un total de $N = \sum_{i=1}^{n} n_i$ observaciones, la SC_T tiene $N-1$ grados de libertad. Hay k tratamientos o niveles del factor de interés, así que SC_{trat} tiene $k-1$ grados de libertad, mientras que SC_e tiene $N-k$, esto de la ecuación (1.2) pues de manera similar los grados de libertad cumplen $N-1 = (k-1)+(N-k)$.

- **Cuadrados medios**. Las sumas de cuadrados divididas entre sus respectivos grados de libertad se llaman *cuadrados medios*. Los dos que más interesan son el cuadrado medio de tratamientos y el cuadrado medio del error, que se denotan por

$$CM_{trat} = \frac{SC_{trat}}{k-1} \qquad \text{y} \qquad CM_e = \frac{SC_e}{N-k}$$

✍ **Análisis estadístico(Montgomery)**. Se presenta ahora como puede llevarse a cabo una prueba formal de la hipótesis de que no hay diferencia en las medias de los tratamientos. Puesto que se supone que los errores ε_{ij} siguen una distribución normal e independiente con media cero y varianza σ^2, las observaciones y_{ij} tienen una distribución normal e independiente con media $\mu+\tau_i$ y varianza σ^2. Por lo tanto, SC_T es una suma de cuadrados de variables aleatorias con distribución normal; por consiguiente puede demostrarse que SC_T/σ^2 tiene una distribución ji-cuadrada con $N-1$ grados de libertad. Además puede demostrarse que SC_E/σ^2 es una variable ji-cuadrada con $N-k$ grados de libertad y que SC_{trat}/σ^2 es una variable ji-cuadrada con $k-1$ grados de libertad si la hipótesis nula H_0 es verdadera.

Puesto que los grados de libertad de SC_{trat} y SC_E suman $N-1$, el número total de grados de libertad, el *teorema de Cochran* implica que SS_{trat}/σ^2 y SC_E/σ^2 son variables aleatorias ji-cuadrada con una distribución independiente. Por lo tanto, si la hipótesis nula de que no hay diferencias en las medias de los tratamientos es verdadera, el cociente

$$F_0 = \frac{SC_{trat}/(k-1)}{SC_E/(N-k)} = \frac{CM_{trat}}{CM_e}$$

sigue una distribución F con $(k-1)$ grados de libertad en el numerador y $(N-k)$ en el denominador. La expresión para F_0 es el **estadístico de prueba** para la hipótesis de que no hay diferencias en las medias de los tratamientos.

Por los cuadrados medios esperados se observa que, en general, CM_e es un estimados insesgado de σ^2. Así mismo, bajo la hipótesis nula, CM_{trat} es un estimador insesgado de $sigma^2$. Sin embargo, si la hipótesis nula es falsa, el valor esperado de CM_{trat} es mayor que σ^2. Por lo tanto, bajo la hipótesis alternativa, el valor esperado del numerador del estadístico de prueba es mayor que el valor esperado del denominador, y H_0 deberá rechazarse para valores del estadístico de prueba que son muy grandes. Esto implica una región crítica de una sola cola superior. Por lo tanto, H_0 deberá rechazarse y concluirse que hay diferencias entre las medias de los tramientos si

$$F_0 > F_{\alpha,k-1,N-k}$$

donde F_0 es el estadístico de prueba. De manera alternativa, podría usarse el enfoque de valor P para tomar la decisión.

Toda la información necesaria para calcular el estadístico F_0 hasta llegar al valor-p se escribe en la llamada *tabla de análisis de varianza* (ANOVA). En esta tabla, las abreviaturas significan lo siguiente: FV fuente de variabilidad (efecto), SC suma de cuadrados. GL grados de libertad, CM cuadrado medio, F_0 estadístico de prueba, valor-p significancia observada

Tabla 1.2: Tabla ANOVA para el DCA

FV	SC	GL	CM	F_0	valor-p
Tratamientos	$SC_{trat} = \sum_{i=1}^{k} \frac{Y_{i\bullet}^2}{n_i} - \frac{Y_{\bullet\bullet}^2}{N}$	$k-1$	$CM_{trat} = \frac{SC_{trat}}{k-1}$	$\frac{CM_{trat}}{CM_e}$	$P(F > F_0)$
Error	$SC_e = SC_T - SC_{trat}$	$N-k$	$CM_e = \frac{SC_e}{N-k}$		
Total	$SC_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} - \frac{Y_{\bullet\bullet}^2}{N}$	$N-1$			

Ejemplo 1 (Comparación de cuatro métodos de ensamble, (Gutiérrez & De La Vara))**.**

Un equipo de mejora investiga el efecto de cuatro métodos de ensamble A, B, C y D, sobre el tiempo de ensamble en minutos. En primera instancia, la estrategia experimental es aplicar cuatro veces los cuatro métodos de ensamble en orden completamente aleatorio (las 16 pruebas en orden aleatorio). Los tiempos de ensamble obtenidos se muestran en la tabla. Si se usa el diseño completamente al azar (DCA), se supone que, además del método de ensamble, no existe ningún otro factor que influya de manera significativa sobre la variable de respuesta (tiempo de ensamble).

Tabla 1.3: Diseño completamente al azar

Método de ensamble			
A	B	C	D
6	7	11	10
8	9	16	12
7	10	11	11
8	8	13	9

Solución.
Una manera de comparar los métodos de ensamble (tratamientos) es probar la hipótesis

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu \quad vs. \quad H_A : \mu_i \neq \mu_j \text{ para } i = A, B, C, D$$

En caso de no rechazar H_0 se concluye que los tiempos promedio de los cuatro métodos de ensamble son estadísticamente iguales; pero si se rechaza, se concluye que al menos dos de ellos son diferentes.

Para calcular la ANOVA, los cálculos se simplifican si identificamos los datos de la tabla **1.4**, con la información se pueden calcular las sumas de cuadrados, como se hace a continuación:

		Métodos de ensamble			
		A	B	C	D
Observaciones	⇒	6	7	11	10
		8	9	16	12
		7	10	11	11
		8	8	13	9
Total por tratamiento ($Y_{i\bullet}$)	⇒	29	34	51	42
Número de datos en cada tratamiento (n_i)	⇒	4	4	4	4
Media muestral por tratamiento ($\overline{Y}_{i\bullet}$)	⇒	7.25	8.50	12.75	10.50

Tabla 1.4: Detalles de los cálculos para el ANOVA

1. Suma total de cuadrados o variabilidad total de los datos:

$$SC_T = \sum_{i=1}^{4}\sum_{j=1}^{4} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N} = (6^2+7^2+\ldots+13^2+9^2) - \frac{(6+7+\ldots+13+9)^2}{4+4+4+4}$$
$$= 1620 - \frac{156^2}{16} = 99$$

2. Suma de cuadrados de tratamientos o variabilidad debida a la diferencia entre métodos de ensamble:

$$SC_{trat} = \sum_{i=1}^{4} \frac{Y_{i\bullet}^2}{n_i} - \frac{Y_{\bullet\bullet}^2}{N} = \frac{(29^2+34^2+51^2+42^2)}{4} - \frac{156^2}{16} = 69.5$$

3. Suma de cuadrados del error o variabilidad dentro de métodos de ensamble:

$$SC_e = SC_T - SC_{trat} = 99 - 69.5 = 29.5$$

4. Cuadrados medios de tratamiento y del error (efecto ponderado de cada fuente de variación):

$$CM_{trat} = \frac{SC_{trat}}{k-1} = \frac{69.5}{3}$$
$$CM_e = \frac{SC_e}{N-k} = \frac{29.5}{12} = 2.46$$

5. Estadístico de prueba:

$$F_0 = \frac{CM_{trat}}{CM_e} = \frac{23.17}{2.46} = 9.42$$

6. Con la información anterior podemos llenar la tabla **1.2** de ANOVA

FV	SC	GL	CM	F_0	valor-p (o F_{crit})
Tratamientos	69.5	3	23.17	9.42	$P(F > F_0)$ (o $F_{\alpha,k-1,N-k}$)
Error	29.5	12	2.46		
Total	99	15			

7. Conclusión: El valor de la significancia observada o valor-p es el área bajo la curva de la distribución $F_{3,12}$ a la derecha de $F_0 = 9.42$, lo cual es difícil de calcular de forma manual. Sin embargo, cuando esto no sea posible, recordemos que otra forma de rechazar o no una hipótesis es comparar el estadístico de prueba contra un número crítico de tablas. En el caso de las tablas de la distribución F, se lee que el valor crítico para $\alpha = 0.05$ es $F_{0.05,3,12} = 3.49$. Como $F_0 = 9.42 > F_{0.05,3,12} = 3.49$, entonces se rechaza H_0, con lo cual se concluye que sí hay diferencia o efecto de los métodos de ensamble en cuanto a su tiempo promedio

■

Después de que se rechazó la hipótesis nula en un análisis de varianza, es necesario ir a detalle y ver cuáles tratamientos son diferentes. A continuación veremos estrategias distintas para ir a ese detalle.

1.2.4. Pruebas de comparación múltiple de medias

Cuando no se rechaza la hipótesis nula $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu$, el objetivo del experimento está cubierto y la conclusión es que los tratamientos no son diferentes.

Si por el contrario se rechaza H_0, y por consiguiente se acepta la hipótesis alternativa $H_A : \mu_i \neq \mu_j$ para algún $i \neq j$, es necesario investigar cuáles tratamientos resultaron diferentes, o cuáles provocan la diferencia.

Método LSD (Diferencia Significativa Mínima)

Hipótesis.

Una vez que se rechazó H_0 en el ANOVA, el problema es probar la igualdad de todos los posibles pares de medias con la hipótesis:

$$H_0 = \mu_i = \mu_j \quad vs \quad H_A : \mu_i \neq \mu_j \tag{1.3}$$

Para toda $i \neq j$. Para k tratamientos se tiene en total $k(k-1)/2$ pares de medias. El estadístico de prueba para cada una de las hipótesis dadas en (1.3) es la correspondiente diferencia en valor absoluto entre sus medias muestrales $|\overline{Y}_{i\bullet} - \overline{Y}_{j\bullet}|$.

Se rechaza la hipótesis $H_0 : \mu_i = \mu_j$ si ocurre que:

$$|\overline{Y}_{i\bullet} - \overline{Y}_{j\bullet}| > t_{\alpha/2,N-k}\sqrt{CM_e\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = LSD \tag{1.4}$$

donde:

- el valor de $t_{\alpha,N-k}$ se lee de las tablas de la distribución T de Student con $N-k$ grados de libertad que corresponden al error,
- el CM_e es el cuadrado medio del error y se obtiene de la tabla ANOVA
- n_i y n_j son el número de observaciones para los tratamientos i y j respectivamente.

La cantidad LSD se llama diferencia significativa mínima (least significant difference), ya que es la diferencia mínima que debe existir entre dos medias muestrales para considerar que los tratamientos correspondientes son significativamente diferentes. Así, cada diferencia de medias muestrales en valor absoluto que sea mayor que el número LSD se declara significativa. Note que si el diseño es balanceado, es decir, si $n_1 = n_2 = \ldots = n_k = n$, la diferencia mínima significativa se reduce a

$$LSD = t_{\alpha/2,N-k}\sqrt{\frac{2CM_e}{n}}$$

En caso de rechazar H_0 de (1.3) se acepta la hipótesis alternativa H_A la cual nos dice que las medias de los tratamientos i y j son diferentes.

Ilustraremos esta prueba con el ejemplo 1

Ejemplo 2 (Método de Diferencia Significativa Mínima, (Gutiérrez & De La Vara)).

Determine que pares de medias son estadísticamente distintas entre si para el caso de

		Métodos de ensamble			
		A	B	C	D
Observaciones	⇒	6	7	11	10
		8	9	16	12
		7	10	11	11
		8	8	13	9
Media muestral por tratamiento ($\overline{Y}_{i\bullet}$)	⇒	7.25	8.50	12.75	10.50

Tabla 1.5: Datos del ejemplo 1

con ANOVA

FV	SC	GL	CM	F_0	valor-p (o F_{crit})
Tratamientos	69.5	3	23.17	9.42	3.49
Error	29.5	12	2.46		
Total	99	15			

Tabla 1.6: Tabla de diferencias de medias muestrales

	$\overline{Y}_{A\bullet}$	$\overline{Y}_{B\bullet}$	$\overline{Y}_{C\bullet}$	$\overline{Y}_{D\bullet}$
$\overline{Y}_{A\bullet}$	-	1.25	5.50*	3.25*
$\overline{Y}_{B\bullet}$	-	-	4.25*	2.00
$\overline{Y}_{C\bullet}$	-	-	-	2.25
$\overline{Y}_{D\bullet}$	-	-	-	-

Solución.
Para investigar cuales pares de medias son estadísticamente diferentes se prueban las seis posibles pares de hipótesis

$$\begin{aligned}
H_0 : \mu_A = \mu_B \quad vs \quad H_A : \mu_A \neq \mu_B \\
H_0 : \mu_A = \mu_C \quad vs \quad H_A : \mu_A \neq \mu_C \\
H_0 : \mu_A = \mu_D \quad vs \quad H_A : \mu_A \neq \mu_D \\
H_0 : \mu_B = \mu_C \quad vs \quad H_A : \mu_B \neq \mu_C \\
H_0 : \mu_B = \mu_D \quad vs \quad H_A : \mu_B \neq \mu_D \\
H_0 : \mu_C = \mu_D \quad vs \quad H_A : \mu_C \neq \mu_D
\end{aligned}$$

utilizando el método de LSD. En el ANOVA, los grados de libertad de CM_e y su valor son $N - k = 12$ y $CM_e = 2.46$, con $\alpha = 0.05$ y 12 grados de libertad de la distribución T de Student es $t_{0.025,12} = 2.18$, como en cada tratamiento son $n = 4$, entonces:

$$LSD = t_{\alpha/2,N-k}\sqrt{\frac{2CM_e}{n}} = 2.18\sqrt{\frac{2 \times 2.46}{4}} = 2.42$$

La decisión sobre cada una de las seis hipótesis listadas arriba se obtiene al comparar las correspondientes diferencias de medias muestrales en valor absoluto con el número $LSD = 2.42$. Se declaran significativas aquellas diferencias que son mayores a este número. Los resultados se muestran en la tabla de diferencias entre medias, tabla **1.6**, de donde se concluye que $\mu_A = \mu_B$, $\mu_B = \mu_D$, $\mu_C = \mu_D$, mientras que $\mu_A \neq \mu_C$, $\mu_B \neq \mu_C$ y $\mu_A \neq \mu D$ ■

Método de Tukey

Este método consiste en comparar las diferencias entre medias muestrales con el valor crítico dado por

$$T_\alpha = q_\alpha(k, N-k)\sqrt{\frac{CM_e}{n}}$$

donde:

- ✔ CM_e es el cuadrado medio del error
- ✔ n es el número de observaciones por tratamiento

- ✔ k es el número de tratamientos
- ✔ $N-k$ es igual a los grados de libertad del error (CM_e)
- ✔ α es el nivel de significancia prefijado
- ✔ $q_\alpha(k, N-k)$ son puntos porcentuales de la distribución del rango estudentizado, obtenido de tablas

Se declaran significativamente diferentes los pares de medias cuya diferencia muestral en valor absoluto sea mayor que T_α

Ejemplo 3 (Método de Tukey, (Gutiérrez & De La Vara)).

Emplee el método de Tukey con ejemplo de los métodos de ensamble (ejemplo 1)

Solución.
A partir del ANOVA, se toma la información pertinente y de las tablas del rango estudentizado, para $\alpha = 0.05$, se obtiene $q_{0.05}(4,12) = 4.20$, de manera que el valor crítico es

$$T_{0.05} = q_{0.05}(4,12)\sqrt{\frac{CM_e}{n}} = 4.20 \times \sqrt{\frac{2.46}{4}} = 3.29$$

al comparar con la tabla de diferencias de medias muestrales, los resultados sobre las hipótesis son

Tabla 1.7: Tabla de diferencias de medias muestrales

	$\overline{Y}_{A\bullet}$	$\overline{Y}_{B\bullet}$	$\overline{Y}_{C\bullet}$	$\overline{Y}_{D\bullet}$
$\overline{Y}_{A\bullet}$	-	1.25	5.50*	3.25
$\overline{Y}_{B\bullet}$	-	-	4.25*	2.00
$\overline{Y}_{C\bullet}$	-	-	-	2.25
$\overline{Y}_{D\bullet}$	-	-	-	-

De esta tabla se concluye que $\mu_a = \mu_B = \mu_C$, $\mu_C = \mu_D$, $\mu_A \neq \mu_C$ y $\mu_B \neq \mu_C$. ■

Observe que esta prueba no encuentra diferencia entre los métodos de ensamble A y D, la cual si detectó el método LSD. Esto es congruente con el hecho de que la prueba de Tukey es menos potente que la prueba LSD, por lo que pequeñas diferencias no son detectadas como significativas.

Asimismo, el riesgo de detectar una diferencia que no existe es menor con el método de Tukey.

En la práctica, después de que se ha rechazado H_0 con el ANOVA, conviene aplicar ambos métodos (LSD y Tukey) u otros, cuando haya dudas sobre cuál es el tratamiento ganador. Cuando la diferencia entre dos tratamientos es clara, ambos métodos coinciden.

Método de Duncan

En este método para la comparación de medias, si las k muestras son de igual tamaño, los k promedios se acomodan en orden ascendente y el error estándar de los promedios se estima con $S_{\bar{Y}_i} = \sqrt{CM_e/n}$. Si alguno o todos los tratamientos tienen tamaños diferentes, se reemplaza n por la media armónica de las $\{n_i\}$, que está dada por

$$n_{AR} = \frac{k}{\sum_{i=1}^{k} n_i}$$

Note que cuando $n_1 = n_2 = \ldots = n_k = n$, ocurre que $n_{AR} = n$. De la tabla de tangos significantes de Duncan (en plataforma), se obtienen los valores críticos $r_\alpha(p, l)$, $p = 2, 3, \ldots, k$ donde α es el nivel de significancia prefijado y l los grados de libertad del error (mismos que CM_e). Con estos $k-1$ valores se obtienen los rangos de significancia mínima dados por

$$R_p = r_\alpha(p, l) S_{\bar{Y}_i}; \quad p = 2, 3, \ldots, k$$

Las diferencias observadas entre las medias muestrales se comparan con los rangos R_p de la siguiente manera:

I) primero se compara la diferencia entre la media más grande y la más pequeña con el rango R_k

II) luego, la diferencia entre la media más grande y la segunda más pequeña se compara con el rango R_{k-1}

III) estas comparaciones continúan hasta que la media mayor se haya comparado con todas las demás

IV) enseguida, se compara la diferencia entre la segunda media más grande y la media más pequeña con el rango R_{k-1}.

V) Después, la diferencia entre la segunda más grande con la segunda más pequeña se compara con el valor R_{k-2}, y así sucesivamente hasta que se comparan los pares de medias posibles con el rango que corresponda

En las comparaciones donde la diferencia observada es mayor que el rango respectivo, se concluye que esas medias son significativamente diferentes. Si dos medias caen entre otras dos que no son muy diferentes, entonces esas dos medias poblacionales también son consideradas estadísticamente iguales.

Ejemplo 4 (Método de Dunca, (Gutiérrez & De La Vara)).

Emplee el método de Duncan con ejemplo de los métodos de ensamble (ejemplo 1)

En la tabla ANOVA se lee que $CM_e = 2.46$, con 12 g.l. Así el error estándar promedio $S_{\bar{Y}_i} = \sqrt{2.46/4} = 0.78$ Las medias ordenadas son: $\bar{Y}_C = 12.75$, $\bar{Y}_D = 10.50$, $\bar{Y}_B = 8.5$, $\bar{Y}_A = 7.25$ Las comparaciones son:

Diferencia pob.	Rango R_p
$\mu_C - \mu_A$	R_4
$\mu_C - \mu_B$	R_3
$\mu_C - \mu_D$	R_2
$\mu_D - \mu_A$	R_3
$\mu_D - \mu_B$	R_2
$\mu_B - \mu_A$	R_2

donde

$$R_4 = r_{0.05}(4,12)S_{\bar{Y}_i} = (3.313)(0.78) = 2.58$$
$$R_3 = r_{0.05}(3,12)S_{\bar{Y}_i} = (3.225)(0.78) = 2.5155$$
$$R_2 = r_{0.05}(2,12)S_{\bar{Y}_i} = (3.082)(0.78) = 2.4040$$

Diferencia pob.	Dif. muestral y Rango R_p
$\mu_C - \mu_A$	$12.75 - 7.25 = 5.5 > 2.58 = R_4$
$\mu_C - \mu_B$	$12.75 - 8.5 = 4.25 > 2.52 = R_3$
$\mu_C - \mu_D$	$12.75 - 2.25 < 2.40 = R_2$
$\mu_D - \mu_A$	$10.5 - 7.25 = 3.25 > 5.52 = R_3$
$\mu_D - \mu_B$	$10.5 - 2 < 2.40 = R_2$
$\mu_B - \mu_A$	$8.5 - 7.25 = 1.25 < 2.4 = R_2$

De la tabla anterior se concluye que $\mu_A = \mu_B$, $\mu_B = \mu_D$ y $\mu_C = \mu_D$, mientras que $\mu_A \neq \mu_C$, $\mu_B \neq \mu_C$ y $\mu_A \neq \mu_D$

Prueba de Scheffé

Similar a la prueba de Tukey, sin embargo en lugar de usar $q_\alpha(k, N-k)$ del rango studentizado, se usa la tabla de Fisher para calcular el valor

$$\sqrt{(k-1)F_{1-\alpha,k-1,N_k}}$$

Se define el estadístico

$$S = \sqrt{(k-1)F_{1-\alpha,k-1,N_k}}\sqrt{\frac{2CM_e}{n}}$$

El valor del estadístico se compara con las diferencias entre los pares de medias de los tratamientos, si la diferencia excede el valor, entonces se tiene diferencia significativa entre las medias comparadas.

Residuales

Se mencionó anteriormente, que, por fluctuaciones del muestreo al azar, a las desviaciones de cada observación con respecto a su propio grupo, se les llama errores residuales, son las cantidades $\varepsilon_{ij} = y_{ij} - \mu_j$.
Los supuestos del modelo admiten una verificación gráfica, cuando se solicitan las gráficas de los residuales estimados y se observa que no presentan ningún patrón o tendencia. No se realizarán por ahora estas verificaciones, sin embargo, existen algunas sugerencias muy útiles y usadas con frecuencia como usar transformaciones de los valores obtenidos, como $\sqrt{x}$, $\log x$ que tienden a estabilizar la varianza y a normalizar las distribuciones.

1.3. Diseño por bloques y de dos o más factores

El método de análisis de varianza para el diseño de bloques aleatorios completos, fue desarrollado por Sir R. A. Fisher en 1925. Con este método, se aisla y se elimina de la variación del error, la variación atribuida a los bloques. La eficacia del diseño depende de conseguir bloques adecuados y esto se relaciona con el conocimiento del investigador acerca del material experimental. Por ejemplo considere el caso presentado en (Gutiérrez & De La Vara):

Supongamos que se quieren comprar varias máquinas, si cada máquina es manejada por un operador diferente y se sabe que éste tiene una influencia en el resultado, entonces es claro que el factor operador debe tomarse en cuenta si se quiere comparar a las máquinas de manera justa. Un operador más hábil puede hacer ver a su máquina (aunque ésta sea la peor) como la que tiene el mejor desempeño, lo cual impide hacer una comparación adecuada de los equipos. Para evitar este sesgo hay dos maneras de anular el posible efecto del factor operador: la manera lógica es utilizar el mismo operador en las cuatro máquinas; sin embargo, tal estrategia no siempre es aconsejable, ya que utilizar al mismo sujeto elimina el efecto del factor operador pero restringe la validez de la comparación con dicho operador, y es posible que el resultado no se mantenga al utilizar a otros operadores. La otra forma de anular el efecto operador en la comparación consiste en que cada operador trabaje durante el experimento con cada una de las máquinas. Esta estrategia es la más recomendable, ya que utilizar a todos los operadores con todas las máquinas permite tener resultados de la comparación que son válidos para todos los operadores. Esta última forma de nulificar el efecto de operadores, recibe el nombre de *bloqueo*.

- ✍ El objetivo del Diseño de bloques aleatorios, es hacer comparaciones entre un conjunto de tratamientos dentro de bloques de material experimental relativamente homogéneo, para que las medias de los tratamientos estén libres de los efectos de los bloques.
- ✍ Las unidades experimentales se dividen en grupos homogéneos llamados bloques.
- ✍ Cada bloque contiene un número de unidades experimentales igual o un múltiplo del número de tratamientos analizados.

- Los tratamientos se asignan al azar dentro de cada bloque.
- Se enfatiza en la importancia de seleccionar los bloques adecuados. Una forma de garantizar la selección adecuada es el conocimiento profundo de los factores y los efectos de interés en el estudio, por lo tanto preguntarnos qué hace que las unidades experimentales puedan responder en forma diferente a un mismo tratamiento y encontrar razones o fuentes de bloqueo, logrará que se tengan resultados eficientes. También el cuadrado medio del error se reduce y mejora la probabilidad de detectar diferencia significativa.
- Ejemplos de bloques son:
  a) diferentes cepas
  b) grupos de edades
  c) diferentes laboratorios
  d) diferentes días
- A los factores adicionales al factor de interés que se incorporan de manera explícita en un experimento comparativo se les llama factores de bloque

1.3.1. Diseño aleatorizado por bloques completos

En un *diseño en bloques completos al azar* (DBCA) se consideran tres fuentes de variabilidad: el factor de tratamientos, el factor de bloque y el error aleatorio, es decir, se tienen tres posibles "culpables" de la variabilidad presente en los datos. La palabra completo en el nombre del diseño se debe a que en cada bloque se prueban todos los tratamientos, o sea, los bloques están completos. La aleatorización se hace dentro de cada bloque; por lo tanto, no se realiza de manera total como en el diseño completamente al azar. El hecho de que existan bloques hace que no sea práctico o que incluso sea imposible aleatorizar en su totalidad.

Supongamos una situación experimental con k tratamientos y b bloques. El aspecto de los datos para este caso se muestra en la tabla 1.3.1, y considera una repetición en cada combinación de tratamiento y bloque

	Bloque				
Tratamiento	1	2	3	$\cdots$	b
1	Y_{11}	Y_{12}	Y_{13}	$\cdots$	Y_{1b}
2	Y_{21}	Y_{22}	Y_{23}	$\cdots$	Y_{2b}
3	Y_{31}	Y_{32}	Y_{33}	$\cdots$	Y_{3b}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	Y_{k1}	Y_{k2}	Y_{k3}	$\cdots$	Y_{kb}

Tabla 1.8: Arreglo de datos en un diseño en bloques completos al azar

El modelo estadístico para este diseño está dado por

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

donde

- Y_{ij} es la i-ésima observación del j-ésimo tratamiento
- μ es la media global o gran media
- β_j es el efecto de los bloques
- τ_i es el efecto de los tratamientos
- ε_{ij} es el efecto del error residual

Entre los supuestos del modelo:

- Cada una de las kn poblaciones sigue una distribución normal con una media μ_{ij} y la misma varianza σ^2 . Por lo que los errores residuales son variables aleatorias con distribución normal, independientes con media cero y varianza σ^2.
- Los efectos del tratamiento y del bloque son aditivos, significa que no existe interacción entre ellos. Cuando esto se cumple la suma de los efectos correspondientes es cero.

Hipótesis.

Se quiere probar la hipótesis

$$H_0 = \mu_1 = \mu_2 = \cdots = \mu_k = \mu \qquad \text{vs} \qquad H_1 : \mu_i \neq \mu_j \text{ para algún } i \neq j$$

La afirmación a probar es que la respuesta media poblacional lograda con cada tratamiento es la misma para los k tratamientos y que, por lo tanto, cada respuesta media μ_i es igual a la media global poblacional, μ.

Análisis de varianza

La hipótesis se prueba con un análisis de varianza con dos criterios de clasificación, porque se controlan dos fuentes de variación: el factor de tratamientos y el factor de bloque.

$$SC_T = SC_B + SC_{TRAT} + SC_E$$

donde:

$$SC_T = \sum_{j=1}^{b}\sum_{i=1}^{k} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N}$$

$$SC_{TRAT} = \sum_{i=1}^{k} \frac{Y_{i\bullet}^2}{b} - \frac{Y_{\bullet\bullet}^2}{N}$$

$$SC_B = \sum_{j=1}^{b} \frac{Y_{\bullet j}^2}{k} - \frac{Y_{\bullet\bullet}^2}{N}$$

y la del error se obtiene por sustracción como:

$$SC_E = SC_T - SC_{TRAT} - SC_B$$

La tabla ANOVA es

Fuente de variabilidad	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	F_{tab}
Tratamientos	SC_{trat}	$k-1$	CM_{trat}	$F_0 = \frac{CM_{trat}}{CM_e}$	$F_{\alpha,k-1,(k-1)(b-1)}$
Bloques	SC_B	$b-1$	CM_B	$F_0 = \frac{CM_B}{CM_e}$	$F_{\alpha,b-1,(k-1)(b-1)}$
Error	SC_e	$(k-1)(b-1)$	CM_e		
Total	SC_T	$N-1$			

Ejemplo 5 (Diseño en bloques completos al azar(Gutiérrez & De La Vara)).

En el ejemplo 1, donde se planteó la comparación de cuatro métodos de ensamble, ahora se va a controlar activamente en el experimento a los operadores que realizarán el ensamble, lo que da lugar al siguiente diseño en bloques completos al azar.

	Operador			
Método	1	2	3	4
A	6	9	7	8
B	7	10	11	8
C	10	13	11	14
D	10	13	11	9

Solución.
Recordemos que la variable de respuesta son los minutos en que se realiza el ensamble. Para comparar los cuatro métodos se plantea la hipótesis:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu \text{ vs } H_a : \mu_i \neq \mu_j \text{ para algún } i \neq j = A, B, C, D$$

Para calcular estas sumas es necesario obtener antes la media global y los totales por tratamiento y por bloque, como se ilustra a continuación.

	Operador				
Método	1	2	3	4	Total por tratamiento
A	6	9	7	8	$Y_{1\bullet} = 30$
B	7	10	11	8	$Y_{2\bullet} = 36$
C	10	13	11	14	$Y_{3\bullet} = 51$
D	10	13	11	9	$Y_{4\bullet} = 43$
Total	$Y_{\bullet 1} = 33$	$Y_{\bullet 2} = 48$	$Y_{\bullet 3} = 40$	$Y_{\bullet 4} = 39$	Total global $Y_{\bullet\bullet} = 160$

Con estos totales las sumas de cuadrados se obtienen como:

$$SC_T = \sum_{j=1}^{b}\sum_{i=1}^{k} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N} = (6^2 + 7^2 \ldots + 9^2) - \frac{160^2}{16} = 108$$

$$SC_{TRAT} = \sum_{i=1}^{k} \frac{Y_{i\bullet}}{b} - \frac{Y_{\bullet\bullet}^2}{N} = \frac{30^2 + 36^2 + 51^2 + 43^2}{4} - \frac{160^2}{16} = 61.5$$

$$SC_B = \sum_{j=1}^{b} \frac{Y_{\bullet j}^2}{k} - \frac{Y_{\bullet\bullet}^2}{N} = \frac{33^2 + 48^2 + 40^2 + 39^2}{4} - \frac{160^2}{16} = 28.5$$

$$SC_E = SC_T - SC_{TRAT} - SC_B = 18$$

Los grados de libertad de la SC_T corresponden al número total de observaciones menos uno ($N-1 = 16-1 = 15$), mientras que los de las SC_{TRAT} y SC_B son el número de tratamientos menos uno y el número de operadores menos uno, respectivamente. En este caso ambas sumas tienen $4-1 = 3$ grados de libertad. Por último, la SC_E tiene $15-3-3 = 9$ grados de libertad. Con esta información se procede a llenar la tabla ANOVA

Fuente de variabilidad	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	F_{tablas}
Tratamientos	61.5	2	20.5	$F_0 = 10.25$	$F_{0.05,3,9} = 3.8625$
Bloques	28.5	3	9.5	$F_0 = 4.75$	$F_{0.05,3,9} = 3.8625$
Error	18	9	2		
Total	108	15			

De esta tabla de observa que $F_0 > F_{tablas}$, por lo que se rechaza la hipótesis H_0 de que el tiempo medio poblacional de los métodos de ensamble son iguales y se acepta que al menos dos de los métodos son diferentes en cuanto al tiempo promedio que requieren ■

Ejercicios

I. (Daniel) Completar el cuadro de ANOVA.
Un fisioterapeuta tiene interés en comparar tres métodos para enseñar a sus pacientes a

usar cierto mecanismo protésico. Consideró que el aprendizaje sería diferente en los pacientes según sus edades y diseño un experimento en el que la edad fuera tomada en cuenta. Los datos obtenidos se presentan en la tabla. Considerando un nivel de significancia del 5% realice el análisis de varianza para saber si existe diferencia entre los métodos de enseñanza.

	Método de enseñanza		
Grupo de edad	A	B	C
Menor que 20	7	9	10
De 20 a 29	7	9	10
De 30 a 39	9	9	12
De 40 a 49	10	9	12
De 50 o más	11	12	14

Tabla 1.9: Tiempo (en días) para el aprendizaje

II. (Montgomery) Una química desea probar el efecto que tienen cuatro agentes químicos sobre la resistencia de un tipo particular de tela. Como puede haber variabilidad entre un rollo de tela y otro, decide utilizar un diseño aleatorizado por bloques, considerando los rollos de tela como bloques. Ella selecciona 5 rollos y les aplica los cuatro agentes químicos en orden aleatorio. A continuación se proporcionan los resultados de la resistencia a la tensión. Analice estos datos y haga las conclusiones apropiadas con $\alpha = 5\%$

Agente Químico	Rollo de tela				
	1	2	3	4	5
1	73	68	74	71	67
2	73	67	75	72	70
3	75	68	78	73	68
4	73	71	75	75	69

1.3.2. Diseño completamente aleatorizado de dos o más factores

Se estudiarán los principios básicos y el método de análisis de varianza en el diseño factorial de dos o más factores, completamente aleatorizado.

Un experimento factorial es el experimento en que se investigan dos o más factores con el mismo interés. El objetivo del diseño factorial es analizar simultáneamente los efectos de dos o más variables.

Pueden ser factores fijos o factores aleatorios. Las diferentes categorías de los factores se llaman niveles o tratamientos.

En un experimento factorial, no sólo es posible estudiar los efectos de factores individuales, sino también la interacción entre los factores. Se dice que existe interacción entre los factores cuando la diferencia entre los niveles de un factor no es la misma para todos los niveles de los otros factores. Los diseños factoriales son más eficientes que los experimentos de un factor a la vez, con mayor importancia si puede existir interacción.

Diseños factorial de dos factores

Es el caso más simple del diseño factorial. Si A y B son factores, A tiene a niveles y B tiene b niveles entonces en un diseño factorial con n réplicas, se tiene que en cada réplica completa del experimento, se investigan todas las ab combinaciones de los tratamientos.
El orden en el que se realizan las abn observaciones es al azar por esto, se dice que es un diseño completamente aleatorizado.
El modelo en un diseño factorial completamente aleatorizado, con dos factores fijos es de la forma:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon$$
$$i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b;\ k = 1, 2, \ldots, n$$

donde μ es la media general, α_i es el efecto debido al i-ésimo nivel del factor A, β_j es el efecto del j-ésimo nivel del factor B, $(\alpha\beta)_{ij}$ representa al efecto de interacción en la combinación ij y ε_{ij} es el error aleatorio que se supone sigue una distribución normal con media cero y varianza constante σ^2

Entre los supuestos del modelo, se tienen:

- Las observaciones en cada una de las ab celdas son una muestra aleatoria independiente de tamaño n extraída de la población representada por la celda.

- Cada población sigue una distribución normal con la misma varianza.

Considere los factores A y B con a y b ($a, b \geq 2$) niveles de prueba, respectivamente. Con ellos se puede construir el diseño factorial $a \times b$, el cual consiste en $a \times b$ tratamientos. Algunos casos particulares de uso frecuente son: el factorial 2^2, el factorial 3^2 y el factorial 3×2. Se llama réplica a cada corrida completa del arreglo factorial. Si se hacen n réplicas, el número total de corridas experimentales es $n(a \times b)$

Se presenta un ejemplo completo, mostrando las definiciones y operaciones resumidas

Ejemplo 6 (Factorial 4 × 3, (Gutiérrez & De La Vara)).

Consideremos un experimento en el que se quiere estudiar el efecto de los factores A: profundidad de corte sobre el acabado de un metal y B: velocidad de alimentación. Aunque los factores son de naturaleza continua, en este proceso sólo se puede trabajar en 4 y 3 niveles, respectivamente. Por ello, se decide correr un factorial completo 4 × 3 con tres réplicas, que permitirá obtener toda la información relevante en relación al efecto de estos factores sobre el acabado. Al aleatorizar las 36 pruebas se obtienen los datos de la tabla.

		B: velocidad			
		0.20	0.25	0.30	Total $Y_{i..}$
A: profundidad	0.15	74 64 [198] 60	92 86 [266] 88	99 98 [299] 102	763
	0.18	79 68 [220] 73	98 104 [290] 88	104 99 [298] 95	808
	0.21	82 88 [262] 92	99 108 [302] 95	108 110 [317] 99	881
	0.24	99 104 [299] 96	104 110 [313] 99	114 111 [332] 107	944
	Total $Y_{.j.}$	979	1 171	1 246	$Y_{...} = 3\,396$

Solución.
Las hipótesis de interés para los tres efectos en el modelo anterior son:

$$H_0: \text{Efecto de profundidad } (A) = 0 \text{ vs } H_A: \text{Efecto de profundidad } (A) \neq 0$$
$$H_0: \text{Efecto de velocidad } (B) = 0 \text{ vs } H_A: \text{Efecto de velocidad } (B) \neq 0$$
$$H_0: \text{Profundidad} \times \text{velocidad } (AB) = 0 \text{ vs } H_A: \text{Profundidad} \times \text{velocidad } (AB) \neq 0$$

Estas hipótesis también se puede plantear como

$$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_4 \text{ vs } H_A: \alpha_i \neq 0 \text{ para algún } i$$
$$H_0: \beta_1 = \beta_2 = \ldots = \beta_3 \text{ vs } H_A: \beta_j \neq 0 \text{ para algún } j$$
$$H_0: (\alpha\beta)_{ij} = 0 \text{ para todo } ij \text{ vs } H_A: (\alpha\beta)_{ij} \neq 0 \text{ para algún } ij$$

■

Estas hipótesis se prueban mediante la técnica del análisis de varianza, que para un diseño factorial $a \times b$ con n réplicas resulta de descomponer la variación total como

$$SC_T = SC_A + SC_B + SC_{AB} + SC_E$$

donde los grados de libertad de cada una de ellas son:

$$nab = (a-1) + (b-1) + (a-1)(b-1) + ab(n-1)$$

El factor $n-1$ en los grados de libertad de la suma de cuadrados del error (SC_E) señala que se necesitan al menos dos réplicas del experimento para calcular este componente y, por ende, para construir una tabla de ANOVA. Recordemos que las sumas de cuadrados divididas entre sus correspondientes grados de libertad se llaman cuadrados medios (CM). Al dividir éstos entre el cuadrado medio del error (CM_E) se obtienen los estadísticos de prueba con distribución F

FV	SC	GL	CM	F_0	Valor - p
Efecto A	SC_A	$a-1$	CM_A	CM_A/CM_E	$P(F > F_0^A)$
Efecto B	SC_B	$b-1$	CM_B	CM_B/CM_E	$P(F > F_0^B)$
Efecto AB	SC_{AB}	$(a-1)(b-1)$	CM_{AB}	CM_{AB}/CM_E	$P(F > F_0^{AB})$
Error	SC_E	$ab(n-1)$	CM_E		
Total	SC_T	$abn-1$			

Si el valor p es menor al nivel de significancia α prefijado, se rechaza la hipótesis nula y se concluye que el correspondiente efecto está activo o influye en la variable respuesta. Recordando la notación para representar sumas y medias: $Y_{\bullet\bullet\bullet}$ es la suma de todas las observaciones; $\overline{Y}_{\bullet\bullet\bullet}$ es la media global; $Y_{i\bullet\bullet}$ es el total en el nivel i del factor A; $\overline{Y}_{i\bullet\bullet}$ es la media en el nivel i del factor A; $Y_{\bullet j\bullet}$ es el total en el nivel j del factor B y $\overline{Y}_{\bullet j\bullet}$ es la correspondiente media. Es decir:

$$Y_{\bullet\bullet\bullet} = \sum_{i=1}^{a}\sum_{j=i}^{b}\sum_{k=1}^{n} Y_{ijk}, \qquad \overline{Y}_{\bullet\bullet\bullet} = \frac{Y_{\bullet\bullet\bullet}}{abn}$$

$$Y_{i\bullet\bullet} = \sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}, \qquad \overline{Y}_{i\bullet\bullet} = \frac{Y_{i\bullet\bullet}}{bn} \qquad i = 1, 2, \ldots, a$$

$$Y_{\bullet j\bullet} = \sum_{i=1}^{a}\sum_{k=1}^{n} Y_{ijk}, \qquad \overline{Y}_{\bullet j\bullet} = \frac{Y_{\bullet j\bullet}}{an} \qquad i = 1, 2, \ldots, b$$

$$Y_{ij\bullet} = \sum_{k=1}^{n} Y_{ijk}, \qquad \overline{Y}_{ij\bullet} = \frac{Y_{ij\bullet}}{n}$$

Con esta notación la suma de cuadrados totales es:

$$SC_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}^2 - \frac{Y_{\bullet\bullet\bullet}}{n}$$

donde $N = abn$ es el total de observaciones en el experimento. Las sumas de cuadrados de

efectos son:

$$SC_A = \sum_{i=1}^{a} \frac{Y_{i\bullet\bullet}^2}{bn} - \frac{Y_{\bullet\bullet\bullet}}{n}$$

$$SC_B = \sum_{j=1}^{b} \frac{Y_{\bullet j\bullet}^2}{an} - \frac{Y_{\bullet\bullet\bullet}}{n}$$

$$SC_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{Y_{ij\bullet}^2}{n} - \frac{Y_{\bullet\bullet\bullet}}{n} - SC_A - SC_B$$

$$SC_E = SC_T - SC_A - SC_B - SC_{AB}$$

Para obtener el ANOVA de la tabla, los cálculos necesarios son:

$$SC_A = \sum_{i=1}^{4} \frac{Y_{i\bullet\bullet}^2}{3\times 3} - \frac{Y_{\bullet\bullet\bullet}^2}{4\times 3\times 3} = \frac{763^2+808^2+881^2+944^2}{3\times 3} - \frac{3396^2}{4\times 3\times 3} = 2\,125.1$$

$$SC_B = \sum_{j=1}^{3} \frac{Y_{\bullet j\bullet}^2}{4\times 3} - \frac{Y_{\bullet\bullet\bullet}^2}{4\times 3\times 3} = \frac{979^2+1\,171^2+1\,246^2}{4\times 3} - \frac{3396^2}{4\times 3\times 3} = 3\,160.5$$

$$SC_{AB} = \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{Y_{ij\bullet}}{3} - \frac{Y_{\bullet\bullet\bullet}^2}{4\times 3\times 3} - SC_A - SC_B = \frac{198^2+220^2+\ldots+332^2}{3} - \frac{3396^2}{4\times 3\times 3} - 2\,125.1 - 3\,160.5$$
$$= 557.07$$

La suma de cuadrados totales y la suma de cuadrados del error están dadas por

$$SC_T = \sum_{i=1}^{4}\sum_{j=1}^{3}\sum_{k=1}^{3} Y_{ijk}^2 - \frac{Y_{\bullet\bullet\bullet}^2}{4\times 3\times 3} = 6\,532$$

$$SC_E = SC_T - SC_A - SC_B - SC_{AB} = 6\,532 - 2\,125.1 - 3\,160.5 - 557.07 = 689.33$$

Con esta información se construye el análisis de varianza de la tabla

FV	SC	GL	CM	F_0	Valor-p
B: vel.	3 160.5	2	1 580.25	55.02	0.0000
A: prof.	2 125.10	3	708.37	24.66	0.0000
AB	557.07	6	92.84	3.23	0.0180
Error	689.33	24	28.72		
Total	6 532	35			

Del ANOVA se concluye que los tres efectos A : vel, B : prof y AB están activos o influyen en el acabado

Ejercicios

1. (Gutiérrez & De La Vara) De un diseño factorial de dos factores, completamente aleatorizado y considerando un nivel de significancia del 5 %, completar la tabla del análisis de

varianza usando los datos del enunciado.
Para mejorar la resistencia a la torsión de las adhesiones de componentes electrónicos sobre placas, se estudiaron dos tipos de pegamentos (A_1 y A_2) y tres temperaturas de curado (60, 80 y 100°C). En cada combinación se analizaron dos componentes y los resultados obtenidos son los siguientes:

	Curado		
	60	80	100
Pegamento A_1	2.5	3.8	4.0
	2.8	3.4	4.2
Pegamento A_2	1.6	3.2	4.3
	1.22	2.8	4.7

a) Plantee las hipótesis de interés en este problema y el modelo estadístico correspondiente.

b) Construya el ANOVA y decida cuáles efectos están activos

1.3.3. Diseño factorial 2^k

Se estudiará en esta sección, los principios básicos y el análisis de varianza del diseño factorial 2^k.

En general, el diseño factorial general, se usa en experimentos que involucran dos o más factores y el objetivo es analizar simultáneamente los efectos de estos factores sobre la variable respuesta.

- ✍ Un caso especial es el de k factores, cada uno con sólo dos niveles.
- ✍ Se llama diseño factorial 2^k.
- ✍ Los niveles pueden ser cuantitativos o cualitativos.
- ✍ Una réplica completa requiere $2 \times 2 \times \ldots \times 2 = 2^k$ observaciones.

La importancia de este caso especial de diseño factorial, se debe entre otras razones a:

a) Es útil en etapas iniciales de proyectos de investigación, principalmente cuando se analizan muchos factores.

b) Se usan en experimentos de tamizado o selección de factores, porque el número de corridas es mínimo al analizar k factores en un diseño factorial completo.

c) Constituye la base de otros diseños de uso práctico.

Los supuestos principales son:

1. Los factores son fijos.

2. El diseño es completamente aleatorizado.

3. Cada población sigue una distribución normal, con la misma varianza.

4. Como se tienen sólo dos niveles para cada factor, se supone que la respuesta es aproximadamente lineal en el rango de los niveles

Ejemplo 7 (Diseño Factorial 2^2, (Montgomery))**.**

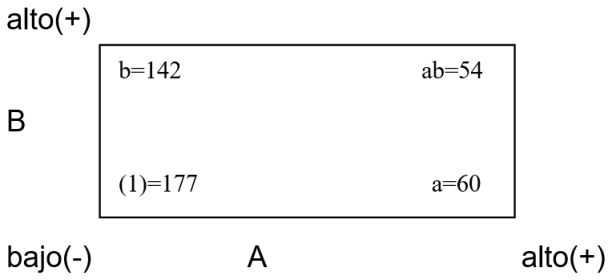
Un investigador está interesado en las diferencias en la fatiga potencial que resulta de dos diferentes tipos de botellas, de vidrio y de plástico, de 32 onzas en cajas de 12 botellas de producto. Se usan dos empleados para una tarea que consiste en mover 40 cajas de producto en una plataforma de carga estándar y acomodarlas en un estante de ventas. Se hacen cuatro réplicas del diseño factorial, como medida de la cantidad de esfuerzo se midió el aumento del ritmo cardíaco (pulso) inducido por la tarea.

		Empleado			
		1		2	
Tipo de Botella	Vidrio	39	45	20	13
		58	35	16	11
	Plástico	44	35	13	10
		42	21	16	15

Solución.
El diseño sólo tiene dos factores A: empleado y B: tipo de botella; cada uno con sólo dos niveles, llamados "bajo"se escribe "−" y "alto" o sea "+".
Se escriben los datos en notación geométrica del diseño 2^k:

A	B	Combinación de tratamientos	Réplica				Total
			I	II	III	IV	
−	−	(1)	39	58	45	35	177
+	−	a	20	16	13	11	60
−	+	b	44	42	35	21	142
+	+	ab	13	16	10	15	54

Una representación geométrica, es el siguiente cuadrado, en el cual se representan los totales de las n observaciones tomadas en cada tratamiento.



Para los efectos del diseño 22, se definen:

- El efecto principal de A:

$$A = \overline{y}_{A^+} - \overline{y}_{A^-} = \frac{a+ab}{2n} - \frac{b+(1)}{2n} = \frac{1}{2n}\left[a+ab-b-(1)\right]$$

- El efecto principal B:

$$B = \overline{y}_{B^+} - \overline{y}_{B^-} = \frac{1}{2n}\left[b+ab-a-(1)\right]$$

- La interacción AB se estima con la diferencia de las diagonales:

$$AB = \frac{1}{2n}\left[ab+(1)-a-b\right]$$

A las cantidades entre corchetes se les llama *contrastes*.
Los signos de los coeficientes de los contrastes $(+1$ o $-1)$

Tratamientos	Efecto factorial			
	I(total)	A	B	AB
(1)	$+$	$-$	$-$	$+$
a	$+$	$+$	$-$	$-$
b	$+$	$-$	$+$	$-$
ab	$+$	$+$	$+$	$+$

Por ejemplo, se obtienen:

$$A = \frac{\text{contraste}_A}{8} = -\frac{205}{8} = -25.625$$
$$B = \frac{\text{contraste}_B}{8} = -\frac{41}{8} = -5.125$$
$$AB = \frac{\text{contraste}_{AB}}{8} = 3.625$$

La tabla de análisis de varianza se obtiene, en forma similar al caso factorial general, sin embargo es importante señalar que los contrastes se usan, para calcular las sumas de cuadrados:

$$SC_A = \frac{[a+ab-b-(1)]^2}{4n} = \frac{42025}{16} = 2626.56$$
$$SC_B = \frac{[b+ab-a-(1)]^2}{4n} = \frac{1681}{16} = 105.063$$
$$SC_{AB} = \frac{[ab+(1)-a-b]^2}{4n} = \frac{841}{16} = 52.5625$$

La suma de cuadrados total se obtiene,

$$SC_T = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n} Y_{ijk}^2 - \frac{Y_{\bullet\bullet\bullet}^2}{4n} = 15\,197 - \frac{433^2}{16}$$

La suma de cuadrados del error,

$$SC_E = SC_T - SC_A - SC_B - SC_{AB} = 694.752$$

La tabla del ANOVA, queda entonces de la forma

FV	SC	GL	CM	F_0	F_{crit}
A	2626.56	1	2626.56	45.37	4.75
B	105.063	1	105.063	1.875	4.75
AB	52.5625	1	52.5625	0.9079	4.75
Error	694.752	12	57.896		
Total	3478.9375	15			

Se confirma las sugerencias de los efectos, observando que solo el efecto del factor A, presenta la diferencia significativa. ■

2 Regresión y Correlación Múltiple

El análisis de regresión tiene como objetivo modelar en forma matemática el comportamiento de una variable de respuesta en función de una o más variables independientes (Gutiérrez & De La Vara). Por ejemplo, suponga que el rendimiento de un proceso químico está relacionado con la temperatura de operación. Si mediante un modelo matemático es posible describir tal relación, entonces este modelo puede ser usado para propósitos de predicción, optimización o control.

Para estimar los parámetros de un modelo de regresión son necesarios los datos, los cuales pueden obtenerse de experimentos planeados, de observaciones de fenómenos no controlados o de registros históricos.

Dos técnicas estadísticas clásicas, muy útiles para analizar la naturaleza y la intensidad de la relación entre dos variables son la regresión y la correlación. Estas técnicas, aunque están relacionadas tienen propósitos diferentes.

El análisis de regresión puede ser lineal o no lineal, lineal simple para dos variables y lineal múltiple de tres o más variables.

2.1. Regresión y correlación simple

Sir Francis Galton (1822-1911) presentó las ideas de regresión, en sus investigaciones acerca de la herencia.

En su trabajo reportó que la estatura de un individuo adulto, sin importar si sus padres son altos o bajos, tiende hacia la estatura promedio de la población, a esta tendencia le llamó regresión.

El análisis de regresión se usa para investigar la forma probable de la relación entre variables (la ecuación que relaciona a ambas variables) y entre los objetivos principales que se tienen están predecir o estimar el valor de una variable que corresponde al valor dado de otra variable o bien optimizar o controlar un proceso.

Sean dos variables X y Y, suponga que se quiere explicar el comporta miento de Y con base en los valores que toma X. Para esto, se mide el valor de Y sobre un conjunto de n valores de X, con lo que se obtienen n parejas de puntos (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). A Y se le llama la *variable dependiente o la variable de respuesta* y a X se le conoce como *variable independiente o variable regresora*. La variable X no necesariamente es aleatoria, ya que en muchas ocasiones el investigador fija sus valores; en cambio, Y sí es una variable aleatoria.

Suponga que las variables X y Y están relacionadas linealmente y que para cada valor de X, la variable dependiente, Y, es una variable aleatoria. Es decir, que cada observación de Y puede ser descrita por el modelo:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.1}$$

donde ε es un error aleatorio con media cero y varianza σ^2. La ecuación 2.1 es conocida como el *modelo de regresión lineal simple*.

Para tener bien especificada la ecuación que relaciona las dos variables será necesario estimar los dos parámetros, que tienen los siguientes significados: β_0 es el punto en el cual la línea recta intercepta o cruza el eje y, y β_1 es la pendiente de la línea, es decir, es la cantidad en que se incrementa o disminuye la variable Y por cada unidad que se incrementa X.

El primer paso para estudiar la relación entre dos variables es efectuar un diagrama de dispersión, este diagrama consiste en representar los pares de valores obtenidos y que se quieren modelar.

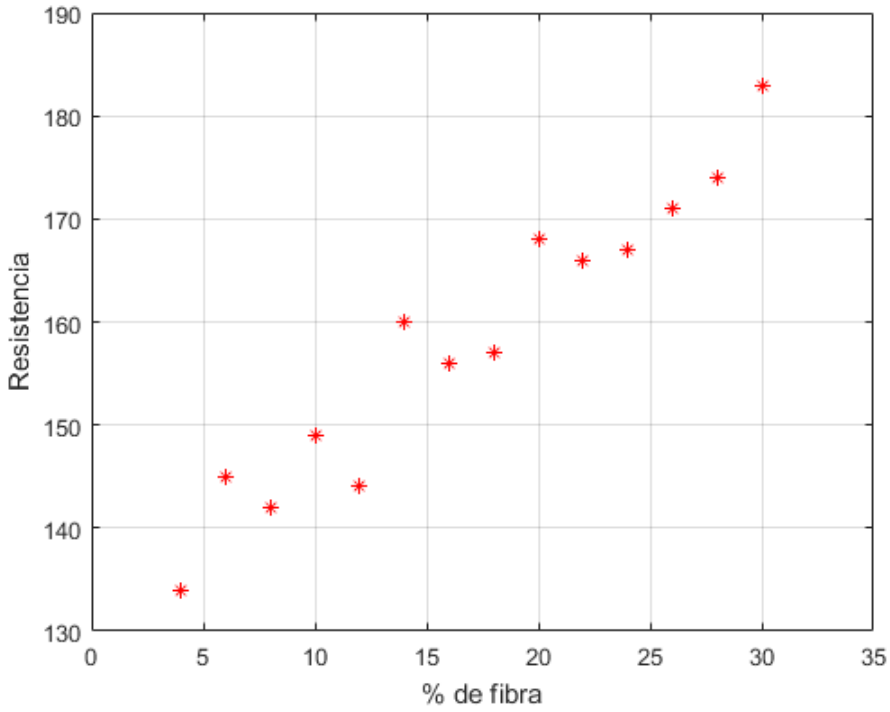
Ejemplo 8 (Diagrama de dispersión (Gutiérrez & De La Vara))**.**

En un laboratorio se quiere investigar la forma en que se relaciona la cantidad de fibra (madera) en la pulpa con la resistencia del producto (papel). Los datos obtenidos en un estudio experimental se muestran en la tabla

% Fibra	Resistencia	% Fibra	Resistencia
4	134	18	157
6	145	20	168
8	142	22	166
10	149	24	167
12	144	26	171
14	160	28	174
16	156	30	183

Solución.

En este caso, la variable de respuesta o variable dependiente es la resistencia, por eso se denota con Y. Para tener una idea de la relación que existe entre X y Y, los 14 pares de datos son representados en el plano en el diagrama de dispersión de la figura



■

Para estimar los coeficientes de regresión β_0 y β_1 se requiere usar la información de una muestra de tamaño n. Cada observación se expresa como

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2.2}$$

Un procedimiento para ajustar la mejor recta y, por lo tanto, para estimar β_0 y β_1 es mediante el método de mínimos cuadrados, el cual consiste en lo siguiente:
Si de la ecuación (2.2) despejamos los errores, los elevamos al cuadrado y los sumamos, obtenemos lo siguiente:

$$S = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} \left(y_i - \left[\beta_0 + \beta_1 x_i\right]\right)^2$$

El procedimiento matemático para minimizar los errores de la ecuación anterior y así encontrar los estimadores de mínimos cuadrados de β_0 y β_1, consiste en derivar a S con respecto a β_0, $\frac{\partial S}{\partial \beta_0}$ y derivar también a S respecto a β_1, $\frac{\partial S}{\partial \beta_1}$. Al igualar a cero las dos ecuaciones y resolverlas en forma simultánea con respecto a las dos incógnitas (β_0 y β_1), se obtiene la solución única:

$$\beta_1 = \frac{S_{xy}}{S_{xx}}$$
$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

donde:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$
$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x}) = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \tag{2.3}$$

$\overline{x}$, $\bar{y}$ son las medias muestrales de las dos variables: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Con estos valores sustituidos en la recta de regresión, se obtiene la ecuación de regresión o la ecuación de la recta de y sobre x, que sirve para estimar o predecir valores de y, dado algún valor de x

$$Y = \beta_0 + \beta_1 X$$

Un resultado importante que garantiza la calidad de los estimadores, es el **Teorema de Gauss Markov**: En un modelo lineal, los estimadores de mínimos cuadrados son los mejores estimadores lineales de mínima varianza.

Ejemplo 9 (Modelo de regresión (Gutiérrez & De La Vara))**.**

Encuentre la linea recta que mejor explica la relación entre el porcentaje de fibra y resistencia del papel del ejemplo 8.

Solución.
El procedimiento para realizar los cálculos se presenta en la tabla siguiente

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
4	134	16	17 956	536
6	145	36	21 025	870
8	142	64	20 164	1 136
10	149	100	22 201	1 490
12	144	144	20 736	1 728
14	160	196	25 600	2 240
16	156	256	24 336	2 496
18	157	324	24 649	2 826
20	168	400	28 224	3 360
22	166	484	27 556	3 652
24	167	576	27 889	4 008
26	171	676	27 241	4 446
28	174	784	30 276	4 872
30	183	900	33 489	5 490

Así

$$n = 14, \quad \sum x_i = 238, \quad \overline{x} = 17, \quad \sum y_i = 2216, \quad \overline{y} = 158.286,$$
$$\sum x_i^2 4956, \quad \sum y_i^2 = 353342, \quad \sum x_i y_i = 39150$$

Al usar las fórmulas para el modelo de regresión

$$S_{xy} = 39\,150 - \frac{(238)(2\,216)}{14} = 1\,478$$
$$S_{xx} = 4\,956 - \frac{238^2}{14} = 910$$
$$\beta_1 = \frac{1\,478}{910} = 1.6242$$
$$\beta_0 = 158.286 - (1.6242)(17) = 130.67$$

Por tanto, la línea recta que mejor explica la relación entre porcentaje de fibra y resistencia de papel, está dad por

$$Y = 130.67 + 1.6242X$$

■

Como interpretación del modelo anterior, por cada punto porcentual de incremento en el porcentaje de fibra, se espera un incremento de la resistencia de 1.6242 en promedio. La ecuación sirve para estimar la resistencia promedio esperada para cualquier porcentaje de fibra utilizada, claro que esa estimación será más precisa en la medida que X esté dentro del intervalo de los valores con los que se hizo la estimación.

Pruebas de hipótesis en la regresión lineal simple

En cualquier análisis de regresión no basta hacer los cálculos que se explicaron antes, sino que es necesario evaluar qué tan bien el modelo (la línea recta) explica la relación entre X y Y. Una primera forma de hacer esto es probar una serie de hipótesis sobre el modelo.

Hipótesis.

La hipótesis de mayor interés plantea que la pendiente es significativamente diferente de cero. Esto se logra al probar la hipótesis

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_a : \beta_1 \neq 0$$

Si la hipótesis nula es verdadera el estadístico t_0, definido a continuación, tiene una distribución *t-Student* con $n-2$ grados de libertad,

$$t_0 = \frac{\beta_1}{\sqrt{CM_E / S_{xx}}}$$

donde CM_E se conoce como cuadrado medio del error y se calcula con la expresión

$$CM_e = \frac{SC_E}{n-2} = \frac{1}{n-2}\left(S_{yy} - \frac{(S_{xy}^2)}{S_{xx}}\right), \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

y[1] S_{xx}, S_{xy} se han definido en (2.3). Se rechaza H_0 si

$$|t_0| > t_{\alpha/2,\, n-2}$$

En caso contrario no se rechaza H_0.

No rechazar $H_0 : \beta_1 = 0$, en el caso del modelo de regresión lineal simple, implica que no existe una relación lineal significativa entre X y Y; por lo tanto, no existe relación entre estas variables o ésta es de otro tipo.

Si se utiliza como criterio de rechazo la comparación de la significancia observada (p-value o valor-p) contra la significancia predefinida (α), entonces se rechaza H_0 si valor-$p < \alpha$.

En ocasiones, en lugar de probar que $\beta_1 = 0$, puede ser de interés probar que es igual a cierta constante ($H_0 : \beta_1 = c$), en ese caso en el numerador del estadístico se resta c, es decir, el estadístico queda de la siguiente manera $t_0 = (\beta_1 - c)/\sqrt{CM_E/S_{xx}}$ y el criterio de rechazo es el mismo.

Ejemplo 10.

Hacer el análisis de regresión para el modelo del ejemplo 8, considere $\alpha = 5\,\%$

Solución.
El modelo en esta caso fue $y = 1.62416x + 130.675$, es decir, $\beta_1 = 1.62418$, así

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 353\,342 - \frac{(2\,216)^2}{14} = 2580.9$$
$$SC_E = S_{yy} - \beta_1 S_{xy} = 2\,580.9 - (1.62418)(1\,478) = 180.36$$

así

$$CM_E = \frac{SC_E}{n-2} = \frac{180.36}{12} = 15.03$$

así $\sqrt{CM_E/S_{xx}} = \sqrt{15.03/910} = 0.12852$. Por tanto, el estadístico es

$$t_0 = \frac{\beta_1}{\sqrt{CM_E/S_{xx}}} = \frac{1.62748}{0.12852} = 12.638$$

Al comparar con el valor del crítico $t_{0.25,12} = 2.179$ se tiene que $t_0 > t_{0.25,12}$, entonces se rechaza la hipótesis nula, por lo que se concluye que β_1 es significativamente diferente de cero ■

[1] Note que $SC_E = S_{yy} - \beta_1 S_{xy}$

Análisis de varianza del modelo de regresión

Otro enfoque para analizar la significancia del modelo es descomponer la variabilidad observada para un análisis de varianza o ANOVA, considerando dos fuentes de variación: debida al modelo y la del error del ajuste.

Para el estadístico, se descompone S_{yy} es

$$S_{yy} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

$$SC_{total} = SC_{regresion} + SC_{error}$$

El primer componente de S_{yy} o SC_{total} denotado por SC_R o $SC_{regresion}$ mide la variabilidad explicada por la recta de regresión y se le conoce como *suma de cuadrados de regresión*. El segundo componente de S_{yy} corresponde a la *suma de cuadrados del error*, SC_e o SC_{error} y mide la variabilidad no explicada por la recta de regresión.

Los grados de libertad de S_{yy} son $n-1$, SC_R tiene un grado de libertad y SC_E tiene $n-2$. Al dividir las sumas de cuadrados entre sus grados de libertad obtenemos los cuadrados medios

$$CM_R = \frac{SC_R}{1} \qquad CM_E = \frac{SC_E}{n-2}$$

y de manera alternativa, se puede calcular

$$SC_R = \beta_1 S_{xy}$$

Hipótesis.

Todo lo anterior podemos utilizarlo para generar otra forma de pro bar la hipótesis sobre la significancia de la regresión:

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

Ya que si H_0 es verdadera, entonces el estadístico

$$F_0 = \frac{CM_R}{CM_E}$$

tiene una distribución F con 1 y $n-2$ grados de libertad en el numerador y denominador, respectivamente.

Por lo tanto, se rechaza $H_0 : \beta_1 = 0$, si el estadístico de prueba es mayor que el valor crítico correspondiente, es decir, se rechaza H_0 si $F_0 > F(\alpha, 1, n–2)$.

El análisis de varianza para probar la significancia del modelo de regresión se resume en la siguiente tabla

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	F_{tablas}
Regresión	$SC_R = \beta_1 S_{xy}$	1	CM_R	CM_R/CM_E	$F_{(\alpha,1,n-2)}$
Error o residual	$SC_E = S_{yy} - \beta_1 S_{xy}$	$n-2$	CM_E		
Total	S_{yy}	$n-1$			

La suma de cuadrados de la regresión, tiene sólo un grado de libertad, porque corresponde al número de variables independientes.

Se tienen n-1 grados de libertad para la suma de cuadrados del total, es el número de observaciones independientes.

Los grados de libertad de la suma de cuadrados del error resulta ser n-2, se calculan usando la propiedad aditiva de los grados de libertad o puede verse también porque se estiman dos parámetros al obtener la ecuación de la recta.

Ejemplo 11 ((Gutiérrez & De La Vara))**.**

Hacer el análisis de varianza para el modelo del ejemplo 8, considere $\alpha = 5\%$

Solución.

Los elementos para la tabla ANOVA, algunos ya calculados previamente, son:

$$\begin{aligned} \beta_1 = 1.62418, \quad S_{xy} = 1\,478 \quad &\Rightarrow SC_R = 2400.5 \\ &\quad S_{yy} = 2\,580.86 \\ &\Rightarrow SC_E = 2\,580.9 - 2400.5 = 180.32 \end{aligned}$$

Así la tabla para el análisis de varianza es

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	F_{tablas}
Regresión	2400.5	1	2400.5	159.71	$F_{(0.05,1,12)} = 4.75$
Error o residual	180.32	12	15.0271		
Total	2580.86	13			

Como $F_0 > F_{(0.05,1,12)} = 4.75$ se concluye que el modelo de regresión es significativo ■

Coeficiente de determinación

En la sección anterior estudiamos pruebas de hipótesis para verificar que hay una relación significativa entre X y Y; sin embargo, no hemos visto si tal relación permite hacer estimaciones con una precisión aceptable.

Un criterio cuantitativo para evaluar la calidad de ajuste es el *coeficiente de determinación* definido por

$$R^2 = \frac{\text{Variabilidad explicada por el modelo}}{\text{Variabilidad tota}}$$
$$= \frac{SC_R}{S_{yy}}$$

R^2 es una medida del ajuste del modelo lineal a los datos. Debe tenerse en cuenta que no es una medida de la adecuación del modelo.

Note que $0 \leq R^2 \leq 1$. En general R^2 se interpreta como la proporción de la variabilidad en los datos (Y) que es explicada por el modelo. En el caso de los datos del ejemplo 8, a partir de la tabla ANOVA (ejemplo 11) tenemos: $R^2 = (2400.5)/(2580.86) = 0.930$. Por lo tanto, podemos decir que 93 % de la variación observada en la resistencia es explicada por el modelo (línea recta), lo cual nos dice que la calidad del ajuste es satisfactorio, y que por ello, la relación entre X y Y es descrita adecuadamente por una línea recta.

Cuando R^2 es pequeño, se dice que la recta de regresión no proporciona un buen ajuste para los datos, esto es, no hay una relación lineal entre las variables.

Si R^2 es grande, entonces las variables están fuertemente relacionadas de manera lineal, pero esto no significa que necesariamente el modelo lineal simple sea el adecuado.

En el análisis de varianza, se definen las sumas de cuadrados

$$SC_T = SC_R + SC_E$$

Mientras mayor sea el ajuste mayor será la $SC_{regresion}$ y menor será la SC_E.

La proporción de la variabilidad total explicada por el modelo, la proporciona el coeficiente de determinación $SC_R = S_{yy} - SC_E$.

Si ambos términos de la ecuación anterior se dividen entre S_{yy}

$$\frac{SC_R}{S_{yy}} = 1 - \frac{SC_E}{S_{yy}}$$

El lado izquierdo de la igualdad representa la proporción de la variabilidad de Y explicada por el modelo de regresión y el lado derecho es uno menos la variabilidad no explicada por la regresión.

Coeficiente de correlación

Es bien conocido que el *coeficiente de correlación*, r, mide la intensidad de la relación lineal entre dos variables X y Y. Si se tiene n pares de datos de la forma (x_i, y_i), entonces este coeficiente se obtiene de la siguiente manera:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

El valor de r se encuentra en el intervalo $[-1, 1]$; si r es próximo a -1, entonces tendremos una relación lineal negativa fuerte, y si r es próximo a cero, entonces diremos que no hay correlación lineal, y finalmente si r es próximo a 1, entonces tendremos una relación lineal positiva fuerte. Por ejemplo, para los datos del ejemplo 8, el coeficiente de correlación es $r = (1\,478)/\sqrt{(910)(2\,580.9)}$, lo cual habla de una correlación positiva fuerte.

2.2. Regresión y correlación múltiple

En muchas aplicaciones, situaciones prácticas existen varias variables independientes que posiblemente influyen o están relacionadas con la variable respuesta Y , y por lo tanto deben tenerse en cuenta si se quiere predecir o entender mejor el comportamiento de Y.

Por ejemplo, si se quiere predecir o explicar el consumo de electricidad de una casa, es muy probable que sea necesario considerar el tamaño o tipo de casa, el número de personas que viven, la temperatura promedio de la región, el número o tipo de aparatos eléctricos que contiene la casa y otras variables que puedan impactar el consumo.

Modelo de Regresión Lineal Múltiple

Si X_1, X_2, …, X_k son variables independientes o regresoras, y Y es la variable respuesta, entonces el modelo de regresión lineal múltiple, con k variables independientes es el polinomio de primer orden:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

donde β_i son parámetros o coeficientes de regresión, ε es el error aleatorio con $E(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2$.

- Si $k = 1$, es el modelo de regresión lineal simple y representa una línea recta.
- Si $k = 2$, representa un plano.
- Si $k \geq 3$, no se puede graficar, se dice que representa un hiperplano en el espacio k-dimensional.

El término lineal del modelo de regresión, se debe a los parámetros desconocidos β_0, β_1, β_2, …, β_k son lineales en la ecuación del modelo.

En forma similar al caso de regresión lineal simple, se interpreta β_0 como la ordenada al origen y β_i mide el cambio esperado en Y por cambio unitario en X_i, cuando las demás variables regresoras se mantienen fijas o constantes.

Con frecuencia se requieren modelos de mayor orden para explicar el comportamiento de la variable respuesta en términos de las variables regresoras, por ejemplo si se considera

que algunas de las variables independientes son cuadráticas, se puede usar un polinomio de segundo orden como modelo de regresión:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \varepsilon$$

Es también un modelo de regresión lineal múltiple, pues la ecuación sigue teniendo una función lineal de los parámetros desconocidos β_0, β_1, β_2, β_{12}, β_{11}, β_{22} y además se pueden redefinir las variables regresoras y se obtiene entonces la misma forma del modelo de regresión lineal general, esto es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

donde $X_3 = X_1 X_2$, $X_4 = X_1^2$ y $X_5 = X_2^2$

2.2.1. Estimación de los Parámetros del Modelo de Regresión Lineal Múltiple

Para estimar los parámetros, se requieren $n > k$ datos o valores observados y_i de la variable dependiente, los cuales corresponden a cada combinación de valores de las variables regresoras $(x_{1i}, x_{2i}, \ldots, x_{ki})$, con una estructura como en el siguiente cuadro:

Y	X_1	X_2	$\cdots$	X_k
y_1	x_{11}	x_{21}	$\cdots$	x_{k1}
y_2	x_{12}	x_{22}	$\cdots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_n	x_{1n}	x_{2n}	$\cdots$	x_{kn}

En términos de los datos, el modelo de regresión múltiple puede expresarse como:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \varepsilon_i$$

con $i = 1, 2, \ldots, n$.

En forma similar, al proceso descrito para el modelo lineal simple se tiene que al despejar los errores, elevarlos al cuadrado y sumarlos, se obtiene la función:

$$f(\beta_0, beta_1, \beta_2, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{i=1}^{k} \beta_j x_{ji} \right]^2$$

Al aplicar el método de valores extremos, se calculan las derivadas parciales $\partial f / \partial \beta_j$ para $j = 1, \ldots, k$, se igualan a cero y al resolverse las $k+1$ ecuaciones simultáneas, se obtienen los estimadores de mínimos cuadrados $\hat{\beta}_j$, los cuales resultan estimadores insesgados de los parámetros del modelo.

El modelo ajustado está dado por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

Además de estimar los parámetros, con frecuencia resulta de interés probar hipótesis con respecto a los coeficientes de regresión individuales.

Estas pruebas se usan para valorar cada variable de regresión en el modelo.

El modelo podría ser más efectivo si se le introducen variables adicionales o si se eliminan una o más variables consideradas en el modelo.

Aumentar variables siempre provoca que la suma de cuadrados de la regresión aumente y la suma de cuadrados del error disminuye, por lo que si se agrega una variable de poca importancia para el modelo, podría aumentar la media de cuadrados del error y disminuir la utilidad del modelo.

Hipótesis.

La prueba de hipótesis para la significancia de cualquier coeficiente β_i es de la forma:

$$H_0\colon \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0$$

Se usa el estadístico de prueba

$$t_0 = \frac{\hat{\beta}_i}{\sqrt{CM_E \cdot C_{ii}}}$$

se rechaza la hipótesis nula si $|t_0| > t_{\alpha/2,\, n-k-1}$.

El término C_{ii} es $(i+1)$-ésimo elemento de la diagonal de $(X'X)^{-1}$, la cual surge en el proceso de los mínimos cuadrados, usando notación matricial que reduce en forma considerable los cálculos. La decisión también se puede tomar en función del valor-p y la significancia α

Ejemplo 12 (Regresión múltiple (Montgomery))**.**

Un distribuidor quiere analizar el sistema de reparto de un producto. Específicamente, está interesado en predecir el tiempo de servicio a un cliente particular. El ingeniero industrial (II) a cargo del estudio ha sugerido que los dos factores más importantes que intervienen en el tiempo de reparto son el número de cajas de producto que se entregan y la máxima distancia que debe recorrer el repartidor. El II, recopiló la muestra de reparto que se presenta a continuación. Encuentre un modelo lineal y argumente sobre lo significativo que resultan los factores considerados por el II.

X_1: núm. de paquetes	X_2: dist. máxima	Y : tiempo en minutos
10	30	24
15	25	27
10	40	29
20	18	31
25	22	25
18	31	33
12	26	26
14	34	28
16	29	31
22	37	39
24	20	33
17	25	30
13	27	25
30	23	42
24	33	40

Solución.
Se presenta una tabla con el resumen de la información obtenida con el apoyo de Software

Parámetro	Estimación	Error estándar	Estadístico T	Valor-p
β_0	-1.37362	7.25053	-0.189451	0.8529
β_1	0.928629	0.189436	4.90208	0.0004
β_2	0.530653	0.181671	2.92095	0.0128

Tabla 2.1: Análisis de regresión múltiple

Entonces el modelo ajustado está dado por

$$\hat{y} = -1.374 + 0.929X_1 + 0.531X_2$$

Puede observarse que ambas variables regresoras, tienen p-valor menor a 0.05, por lo que resultan estadísticamente significativas. No se sugiere simplificar el modelo, eliminando alguna de ellas. ■

Inferencias en el Modelo de Regresión Múltiple. Análisis de Varianza

Es importante tener en cuenta que estas pruebas se basan en la independencia de $\hat{\beta}_i$, por lo que si no son independientes, tampoco las pruebas de t de student lo serán y, en consecuencia el coeficiente podría ser aparentemente significativo cuando en realidad no lo es.

Para valorar la introducción de una variable nueva en el modelo de regresión, se usa de preferencia la prueba de significación del modelo de regresión, puesto que es un procedimiento

en el que puede determinarse la contribución a la suma de cuadrados de la regresión de un parámetro, dado que otros ya se encuentran en el modelo.

Por lo tanto, es una prueba importante verificar si el modelo de regresión es significativo.

Hipótesis.

En la regresión lineal múltiple, se prueban las hipótesis:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \text{ vs } H_a : \beta_i \neq 0 \text{ para al menos una } i$$

El rechazo de la hipótesis nula implica que al menos una variable contribuye en el modelo ajustado, en forma significativa.

El proceso en la prueba es una generalización del procedimiento usado en la regresión lineal simple.

La suma total de cuadrados S_{yy} se descompone en la suma de cuadrados de la regresión SC_R y en la suma de cuadrados del error SC_E:

$$S_{yy} = SC_R + SC_E$$

Si se supone $H_0 : \beta_i = 0$ verdadera, entonces $\frac{SC_R}{\sigma^2} \sim \chi^2_k$ donde el número de grados de libertad para la chi-cuadrada es igual al número de variables de regresión del modelo.
También se tiene que $\frac{SC_E}{\sigma^2} \sim \chi^2_{n-k-1}$ y que SC_R y SC_E son independientes. Por lo tanto, el estadístico de prueba resulta ser:

$$F_0 = \frac{\frac{SC_R}{k}}{\frac{SC_E}{n-k-1}} = \frac{CM_R}{CM_E}$$

Se rechaza $H_0 : \beta_i = 0$ si $F_0 > F_{\alpha,k,n-k-1}$ o también si valor-p= $P(F > F_0) < \alpha$.
Para completar el procedimiento anterior necesitamos una forma explícita para calcular SC_R.
Una fórmula para calcular la suma de cuadrado del error es:

$$SC_E = y'y - \hat{\beta}' X' y$$

Además, como la suma de cuadrados S_{yy} está dada por:

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n}\right)^2}{n}$$

La SC_E puede expresarse como:

$$SC_E = \left[y'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \right] - \left[\hat{\beta}' X' y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \right] = S_{yy} - SC_R$$

Así, hemos obtenido una forma explícita para la suma de cuadrados de la regresión:

$$SC_R = \hat{\beta}' X' y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Usando los datos del ejemplo, el procedimiento para la prueba de significancia del modelo de regresión múltiple, se resume en una tabla de análisis de varianza:

FV	SC	GL	CM	F_0	Valor-p
Modelo	379.014	2	189.507	12.54	0.0011
Residuo	181.386	12	15.1155		
Total	560.4	14			

Se observa que el valor-p de la prueba es menor que 0.05, por lo que se concluye que el modelo de regresión es significativo, esto es, al menos una variable contribuye significativamente a la regresión

En forma similar al modelo de regresión lineal simple, el **Coeficiente de Determinación** se define como la cantidad:

$$R^2 = \frac{SC_R}{S_{yy}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{SC_E}{S_{yy}}$$

Se usa como una evaluación de la adecuación del modelo de regresión.
En el caso del ejemplo $R^2 = 67.63\,\% = 0.6763$, como $0 \leq R^2 \leq 1$ se interpreta como la proporción de variabilidad de los datos explicada por el modelo de regresión.
Sin embargo, esta interpretación se usa en forma no muy estricta debido a que ese porcentaje aumenta con sólo agregar una variable más al modelo, aunque no necesariamente el nuevo modelo ajustado sea mejor que el anterior.
La R **cuadrada ajustada** se define:

$$R^2_{aj} = 1 - \frac{CM_E}{CM_{total}}$$

Tiene la ventaja de que no aumenta en forma automática, cada vez que se aumenta una variable regresora al modelo. Ambos coeficientes se interpretan de forma similar al caso de regresión lineal simple, es decir, como el porcentaje de variabilidad de los datos que son explicados por el modelo. Se cumple que $0 < R^2_{aj} < R^2 \leq 1$; en general, para hablar de un modelo que tiene un ajuste satisfactorio es necesario que ambos coeficientes tengan valores superiores a 0.7.
Con los datos del ejemplo, se observa que el estadístico R-cuadrado ajustado, que es más conveniente para comparar modelos con diferentes números de variables independientes, indica que el 62.2383 % de la variabilidad de la variable respuesta es explicada por el modelo.

Coeficiente de correlación múltiple. Si las variables regresoras son aleatorias, de tal forma que junto con la variable respuesta Y, puedan considerarse conjuntamente distribuidas, entonces se define Coeficiente de Correlación Mútiple como:

$$R = \sqrt{R^2}$$

Es una medida de la intensidad de la relación entre la variable dependiente Y y el conjunto de variables o términos del modelo: X_1, X_2, …, X_k.
Con los datos del ejemplo, R=0.8224, que indica una correlación lineal positiva fuerte.

Ejercicios

I. Considere la tabla presentada a continuación

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
0	4			
1	3			
2	6			
3	9			
4	9			
5	11			
6	12			
7	14			
$\sum_{i=1}^{n} x_i =$	$\sum_{i=1}^{n} y_i =$	$\sum_{i=1}^{n} x_i y_i =$	$\sum_{i=1}^{n} x_i^2 =$	$\sum_{i=1}^{n} y_i^2 =$

a) Realice los cálculos indicados en la tabla

b) Can base en lo anterior, construya la tabla del análisis de regresión para la recta de regresión y el análisis de varianza

c) A partir de lo anterior, obtenga conclusiones

d) Obtenga el coeficiente de determinación y valore la calidad del ajuste

II. Se midió el porcentaje de supervivencia de los espermatozoides de cierto tipo de semen animal, después de almacenarlo con distintas combinaciones de concentraciones de tres materiales que se emplean para incrementar la supervivencia. Se presentan los datos

y (% supervivencia)	x_1 (peso %)	x_2 (peso %)	x_3 (peso %)
25.5	1.74	5.3	10.8
31.2	6.32	5.42	9.4
25.9	6.22	8.41	7.2
38.4	10.52	4.63	8.5
18.4	1.19	11.6	9.4
26.7	1.22	5.85	9.9
26.4	4.1	6.62	8
25.9	6.32	8.72	9.1
32	4.08	4.42	8.7
25.2	4.15	7.6	9.2
39.7	10.15	4.83	9.4
35.7	1.72	3.12	7.6
26.5	1.7	5.3	8.2

Se ajusta el modelo de regresión múltiple: $y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

a) Se presentan los cálculos del análisis de regresión con el apoyo de Software. Escriba la ecuación de la recta ajustada e interprete la significancia cada uno de los coeficientes con $\alpha = 0.05$

	Coeficientes	Error típico	Estadístico	Valor-p (Probabilidad)
Intercepción	39.15734995	5.88705963	6.651427438	9.35855E-05
x_1: Peso %	1.016100441	0.190895196	5.322818294	0.0004792
x_2: Peso %	-1.861649203	0.267325499	-6.963979162	6.58179E-05
x_3: Peso %	-0.343260493	0.617052039	-0.556290995	0.591572145

b) Complete el ANOVA, interprete los resultados con $\alpha = 0.05$. ¿El modelo es significativo?

F.V.	Suma de C	Grad.Lib.	Cuadrados Medios	F calculada	F crit.
Regresión					4.49×10^{-5}
Residuos	38.67				
Total					

c) Estime e interprete el coeficiente de determinación y el coeficiente de correlación

d) Con el modelo obtenido de a), estimar el porcentaje de supervivencia para $x_1 = 4.6$, $x_2 = 6.3$, $x_3 = 8.9$ ¿considera que es adecuada la estimación? ¿por qué?

3 Estadística No Paramétrica

3.1. Pruebas no paramétricas

Los procedimientos no paramétricos o métodos de distribución libre no suponen que las muestras aleatorias se seleccionen de poblaciones normales. En estas técnicas en general, no se supone ningún conocimiento acerca de las distribuciones de las poblaciones de estudio, excepto tal vez que sean continuas.

Es importante, reconocer cuando las suposiciones de normalidad no se pueden justificar y que no siempre se tienen mediciones cuantitativas.

En muchas aplicaciones, los datos se obtienen en una escala ordinal y por lo tanto, se considera casi natural en estos métodos, la asignación de rangos a los datos.

Se dice que los métodos no paramétricos implican un análisis de rangos.

Los métodos son de gran importancia en los estudios que involucran clasificación de productos por su aceptación o calidad mediante asignación de jerarquías, por ejemplo grado 1 si se considera que el producto tiene el más alto nivel de calidad, grado 2 si tiene el segundo nivel y así sucesivamente; entonces sería de interés una prueba para determinar si existe alguna coincidencia en la opinión de los jueces.
Se deben señalar algunas desventajas asociadas con las pruebas no paramétricas:

a) No usan toda la información obtenida de la muestra.

b) Se requiere un tamaño de muestra grande, para obtener una buena potencia de la prueba.

Es importante tener en cuenta que, cuando ambos métodos son aplicables, la prueba no paramétrica es menos eficiente que el procedimiento paramétrico correspondiente y esta es la razón de que siempre se prefiera el procedimiento paramétrico.

3.1.1. Prueba del rango con signo de Wilcoxon

Esta prueba fué propuesta en 1945, por Frank Wilcoxon (1892–1965) químico y estadístico estadounidense conocido por el desarrollo de diversas pruebas estadísticas no paramétricas. Se trata de alternativas no paramétricas a la prueba t de Student.

La prueba del rango de Wilcoxon considera tanto la magnitud como el signo de las diferencias de dos muestras apareadas.
Las hipótesis en esta situación son:

Hipótesis.

H_0 : La media entre las dos poblaciones es igual vs H_a : La media entre dos poblaciones es distinta.

Los pasos a seguir para la prueba del rango con signo de Wilcoxon ((Marques de Cantú)) son:

- ✍ Calcular las diferencias D_i entre los pares de muestras.
- ✍ Ordenar los valores absolutos de las diferencias D_i en orden ascendente; si alguna diferencia es cero se elimina y si hay más de una diferencia con el mismo valor se saca la media aritmética de los rangos que le corresponden y se le asigna el valor medio a todos los D_i con el mismo valor
- ✍ Reasignar los signos a las diferencias D_i ya ordenadas
- ✍ Sumar los rangos con signo positivo, w^+, y los rangos con signo negativo, w^-, por separado.
- ✍ Sea T el menor de los valores absolutos de las sumas de los rangos positivos y negativos obtenidos en el paso anterior.
- ✍ Compara T con los valores de la tabla de rangos con signo de Wilcoxon. Si el valor calculado de T es menor o igual que el valor teórico (de tablas). Rechazamos la hipótesis nula.

Al seleccionar muestras repetidas esperaríamos que w_+ y w_- y, por lo tanto, w variará. De esta manera, consideramos a w_+, w_- y w como valores de las correspondientes variables aleatorias W_+, W_- y W. La hipótesis nula $\mu = \mu_0$ se puede rechazar a favor de la hipótesis alternativa $\mu < \mu_0$ sólo si w_+ es pequeña y w_- es grande. De igual manera, la hipótesis alternativa $\mu > \mu_0$ se puede aceptar sólo si w_+ es grande y w_- es pequeña. Para una alternativa bilateral se puede rechazar H_0 a favor de H_1 si w_+ o w_- y, en consecuencia, w son suficientemente pequeñas. Por lo tanto, no importa cuál sea la hipótesis alternativa, cuando el valor del estadístico adecuado W_+, W_- o W es suficientemente pequeño, se rechaza la hipótesis nula.

Ejemplo 13 (Ejemplo de una población (Walpole)).

Los siguientes datos representan el número de horas que funciona una desbrozadora antes de requerir una recarga:

1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2, 1.7.

A un nivel de significancia de 0.05 utilice la prueba de rango con signo de Wilcoxon para probar la hipótesis de que esta desbrozadora especíi ca funciona con una mediana de 1.8 horas antes de requerir una recarga

Solución.

La hipótesis a probar es, con $\alpha = 5\,\%$, es

$$H_0 : \hat{\mu} = 1.8 \qquad \text{vs} \qquad H_a : \hat{\mu} \neq 1.8$$

es decir, el compensador específico opera con una mediana de 1.8 horas antes de requerir una recarga.

En este caso, el procedimiento de la prueba es restar $\mu_0 = 1.8$ de cada uno de los datos, descartando todas las diferencias iguales a cero.

Datos	1.5	2.2	0.9	1.3	2.0	1.6	1.8	1.5	2.0	1.2	1.7
D_i	-0.3	0.4	-0.9	-0.5	0.2	-0.2	0	-0.3	0.2	-0.6	-0.1
$\lvert D_i \rvert$	0.3	0.4	0.9	0.5	0.2	0.2	-	0.3	0.2	0.6	0.1
Signo	-	+	-	-	+	-		-	+	-	-
Orden	6	7	10	8	4	3		5	2	9	1
Rango	5.5	7	10	8	3	3		5.5	3	9	1

Después de lo anterior, $n = 10$ debido a que se descarta una medición. Se determinan los valores

- W_+: suma total de rangos que corresponden a signos positivos, es este caso $W_+ = 13$
- W_-: suma total de rangos que correspondan a signos negativos, se tiene que $W_- = 42$

Por lo que $W = \text{mín}\{W_-, W_+\} = 13$. En tablas de valores críticos para la prueba bilateral de rangos con signo y nivel de significancia del 5 %, se observa una región crítica que corresponde a $W \leq 8$, por lo que no se rechaza la hipótesis nula y se concluye que la mediana del tiempo de operación no es significativamente diferente de 1.8 horas. ■

Ejemplo 14 (Caso: dos poblaciones (Marques de Cantú))**.**

En una prueba Brinell de dureza, una esfera de acero endurecido se comprime, bajo una carga especifica , sobre el material por probar. Se mide el diámetro de la indentación esférica. Se dispone de dos esferas (una por cada uno de dos fabricantes), cuyas indentaciones en 15 piezas de material se compararán. Cada pieza de material se probará dos veces, una con cada esfera. Los datos obtenidos se indican en la tabla. Use la prueba del rango con signo de Wilcoxon para determinar si los efectos de los tratamientos son iguales con $\alpha = 0.05$.

Muestra	Diámetro X	Diámetro Y
1	73	51
2	43	41
3	47	43
4	53	41
5	58	47
6	47	32
7	52	24
8	38	43
9	61	53
10	56	52
11	56	57
12	34	44
13	55	57
14	65	40
15	75	68

Solución.
En este caso la hipótesis es $H_0 : \hat{\mu}_1 = \hat{\mu}_2$ vs $H_a : \hat{\mu}_1 \neq \hat{\mu}_2$, es decir, los efectos de los tratamientos son iguales contrastado con que sean distintos. El procedimiento para calcular el valor W de Wilcoxon se muestra en la tabla de los cálculo anteriores tenemos $W_+ = 101.5$ y $W_- = 18.5$, por tanto, $W = \text{mín}\{W_+, W_-\} = 18.5$, al compararlo con el valor crítico de las tablas $W_{crit} = 25$, como $W < W_{crit}$ debemos rechazar H_0, por lo tanto los efectos de los tratamientos no son iguales ■

Aproximación para muestras grandes

Si $n \geq 25$ la distribución muestral de W_+ (W_-) se aproxima a la distribución normal, con media y varianza dadas por

$$\mu_W = \frac{n(n+1)}{4} \qquad \sigma_W^2 = \frac{n(n+1)(2n+1)}{24}$$

D_i	$\|D_i\|$	signo	Orden	Rango
22	22	+	15	13
2	2	+	2	2.5
4	4	+	4	4.5
12	12	+	11	11
11	11	+	10	10
15	15	+	12	12
28	28	+	15	15
-5	5	-	6	6
8	8	+	8	8
4	4	+	5	4.5
-1	1	-	1	1
-10	10	-	9	9
-2	2	-	3	2.5
25	25	+	14	14
7	7	+	7	7

Tabla 3.1: Cálculos de la prueba del signo de Wilcoxon del ejemplo 14

Por tanto, cuando n excede a los valores de la tabla se utiliza el estadístico

$$Z = \frac{W - \mu_W}{\sigma_W}$$

con la significancia adecuada para la región crítica de la prueba

3.1.2. Prueba U de Mann-Whitney

La prueba U de Mann-Whitney sirve para probar hipótesis acerca de dos medias de dos muestras independientes cuando los datos no alcanzan a ser de tipo cuantitativo sino ordinales. La metodología se describe a continuación:

- Arreglamos primero los valores de las muestras en orden creciente desde 1 a $n_1 + n_2$ (donde n_1 es el tamaño de la muestra 1 y n_2 es el tamaño de la muestra 2), conservando cada valor dentro de su propia muestra y asignamos rangos del mismo modo que en la prueba de Wilcoxon.

- La suma de los rangos de la muestra 1 la llamamos R_1 y la suma de los rangos de la muestra 2, R_2.

✍ Luego empleamos una de las siguientes fórmulas:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

✍ Una vez calculado **cualquiera de los dos** U_1 o U_2, sea $U = U_1$ (o $U = U_2$ según su elección de la U_i calculada en el paso previo y con ello obtenemos el valor

$$U' = n_1 n_2 - U$$

✍ Luego empleamos comparamos el mín$\{U, U'\}$ con el menor de los valores críticos de la tabla para los tamaños de muestras n_1 y n_2 correspondientes y el nivel de significancia α indicado. Si el valor calculados es menor que el valor crítico (tablas) se rechaza $H_0 : Md_1 = Md_2$

✍ En caso de que n_1 o n_2 tenga valor que no se incluye en la tabla, podemos usar la distribución normal con

$$Z = \frac{U - \mu_U}{\sigma_U}$$

donde $\mu_U = \dfrac{n_1 n_2}{2}$, $\sigma_U = \sqrt{\dfrac{n_1 n_2(n_1+n_2+1)}{12}}$

Ejemplo 15 (Prueba U de Mann-Whitney (Marques de Cantú))**.**

Se mide el peso del polvo (en miligramos) en el gasto de gas en dos sistemas de tuberías diferentes. Los valores que se obtienen son los siguientes

A	75	21	72	74	85	43	34	65	90	35
B	20	37	55	50	64	41				

Pruebe la hipótesis de que las poblaciones correspondientes son iguales al 5 % de nivel de significación.

Solución.
En este caso la hipótesis es $H_0 : \hat{\mu}_1 = \hat{\mu}_2$ vs $H_a : \hat{\mu}_1 \neq \hat{\mu}_2$, es decir, las poblaciones son iguales contrastado con que sean diferentes. El procedimiento para calcular el valor U de la prueba U de Mann-Whitney es

A	Rangos	B	Rangos
20	1	75	14
37	5	21	2
55	9	72	13
50	8	71	12
64	10	85	15
41	6	43	7
		34	3
		65	11
		90	16
		35	4
$n_1 = 6$	$R_1 = 39$	$n_2 = 10$	$R_2 = 97$

Así

$$U = U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 6(10) + \frac{6(7)}{2} - 39 = 60 + 21 - 39 = 42$$
$$U' = n_1 n_2 - U = 60 - 42 = 18$$

De la tabla para $\alpha = 5\%$ con $n_1 = 6$ y $n_2 = 10$, tenemos $U_{crit} = 11$, por lo tanto, se acepta H_0, lo que significa que las medias no son significativamente diferentes ■

3.1.3. Prueba de un factor con rangos o prueba de Kruskal-Wallis

Esta prueba fue publicada por W. H. Kruskal y W. A. Wallis, en 1952. Se llama también la prueba H de Kruskal-Wallis, y se considera una alternativa no paramétrica al análisis de varianza.

Se debe recordar, que una de las condiciones importantes para la aplicación de la técnica del análisis de varianza son los supuestos de la distribución normal de las poblaciones involucradas.

Debido a la falta de normalidad y por lo tanto, tampoco tiene sentido la condición de la igualdad de varianzas.

Los datos en general son rangos, por lo que esta prueba es llamada análisis de varianza para rangos.

Se emplea para probar igualdad de parámetros de posición.

Es una generalización de la prueba de la suma de rangos para casos de tres o más medias o medianas.

Supongamos el caso $k > 2$ muestras.

Sea n_i $(i = 1, 2, \ldots, k)$ el número de observaciones en la i-ésima muestra. Primero combinamos todas las k muestras y acomodamos las $n = n_1 + n_2 + \ldots + n_k$ observaciones en orden ascendente, y sustituimos el rango apropiado en para cada observación. En el caso de empates

(observaciones idénticas), seguimos el procedimiento acostumbrado de reemplazar las observaciones por la media de los rangos que tendrían las observaciones si fueran distinguibles. La suma de los rangos que corresponde a las n_i observaciones en la i-ésima muestra se denota mediante la variable aleatoria R_i. Consideremos ahora el estadístico

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

que se aproxima muy bien mediante una distribución χ^2 con $k-1$ grados de libertad, cuando la hipótesis nula es verdadera, siempre y cuando cada muestra conste de al menos 5 observaciones.

Teorema 1 (Prueba de Kruskal-Wallis (Walpole))**.**

Para probar la hipótesis nula H_0 de que k muestras independientes provienen de poblaciones idénticas se calcula

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(n+1)$$

donde r_i es el valor supuesto de R_i para $i = 1, 2, \ldots, k$. Si h cae en la región crítica $H < \chi^2_\alpha$ con $v = k-1$ grados de libertad, se rechaza H_0 al nivel de significancia α; de otra manera no se rechaza H_0

Ejemplo 16 (Prueba de Kruskal-Wallis (Walpole))**.**

En un experimento para determinar cuál de tres diferentes sistemas de misiles es preferible, se mide la tasa de combustión del propulsor. Los datos, después de codificarlos, se presentan en la tabla. Utilice la prueba de Kruskal-Wallis y un nivel de significancia de $\alpha = 0.05$ para probar la hipótesis de que las tasas de combustión del propulsor son iguales para los tres sistemas de misiles.

Sistema de misiles								
1			2			3		
24.0	16.7	22.8	23.2	19.8	18.1	18.4	19.1	17.3
19.8	18.9		17.6	20.2	17.8	17.3	19.7	18.9
						18.8	19.3	

Solución.

La hipótesis en este caso es

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad \text{vs} \qquad H_1 : \text{las tres medias son distintas}$$

En la tabla, convertimos las 19 observaciones a rangos y se suma los rangos de cada sistemas de misiles

Sistemas de misiles		
1	2	3
19	18	7
1	14.5	11
17	6	2.5
14.5	4	2.5
9.5	16	13
	5	9.5
		8
		12
$r_1 = 61$	$r_2 = 63.5$	$r_3 = 65.5$

Ahora calculamos el estadístico con $n_1 = 5$, $n_2 = 6$, $n_3 = 8$, $r_1 = 61$, $r_2 = 63.5$ y $r_3 = 65.5$,

$$h = \frac{12}{(19)(20)}\left(\frac{61^2}{5} + \frac{63.5^2}{6} + \frac{65.5^2}{8}\right) - (3)(20) = 1.66$$

Con $\alpha = 5\,\%$ la región crítica es $h > \chi^2_{0.05} = 5.991$ con $v = 2$ grados de libertad. Como $h = 1.66$ no cae en la región crítica, por tanto, no hay evidencia suficiente para rechazar la hipótesis de que las tasas de combustión del propulsor son iguales para los tres sistemas de misiles. ■

Ejercicios

I. La siguiente tabla muestra las calificaciones de un grupo de 15 estudiantes en matemáticas y artes. Use la prueba de rangos con signo de Wilcoxon para determinar si las medias de las calificaciones para estos estudiantes difieren de manera importante para las dos materias e indique la conclusión apropiada para $\alpha = 0.05$.

Estudiante	Matemáticas	Artes	Estudiante	Matemáticas	Artes
1	22	53	9	62	55
2	37	68	10	65	74
3	36	42	11	66	68
4	38	49	12	56	64
5	42	51	13	66	67
6	58	65	14	67	73
7	58	51	15	62	65
8	60	71			

II. Un fabricante de cigarrillos afirma que el contenido de alquitrán de la marca de cigarrillos B es distinta que la de la marca A. Probar esta afirmación se registraron las siguientes medidas del contenido de alquitrán, en miligramos:

Marca A	1	12	9	13	11	14
Marca B	8	10	7			

III. De una clase de matemáticas de 12 estudiantes que tienen las mismas capacidades y utilizan material programado se seleccionan 5 para proporcionarles enseñanza adicional. Los resultados del examen final son los siguientes

	Calificación						
Con enseñanza adicional	87	69	78	91	80		
Sin enseñanza adicional	75	88	64	82	93	79	67

4 Pruebas para datos de frecuencias

La distribución Ji- cuadrada se usará, en esta unidad, en técnicas estadísticas relacionadas con el análisis de conteo o datos de frecuencia.

Estas pruebas son adecuadas, para usar con variables de clasificación y se usan las frecuencias asociadas a cada categoría.

Por ejemplo, es posible saber para una muestra de estudiantes inscritos a la facultad, cuántos son varones y cuántos son mujeres. Para esta misma muestra, también es posible saber cuántos estudiaron un programa de prepa incorporado a la universidad, cuántos de un bachillerato técnico y cuántos de algún otro programa. Es posible saber, para la población de la cual se extrajo la muestra si el tipo de preparatoria es diferente de acuerdo al género.

Las hipótesis que se prueban no incluyen en general, afirmaciones sobre parámetros. Por ejemplo, H_0: la muestra proviene de una población con una distribución (de cierto tipo) vs H_1: la muestra no proviene de una población con una distribución (de cierto tipo).

Ya se ha estudiado y utilizado la distribución ji-cuadrada (χ^2), tanto para estimar como para probar hipótesis acerca de la varianza o desviación estándar de una población. La distribución ji-cuadrada se emplea también, (Marques de Cantú), para :

- Probar hipótesis acerca de datos de frecuencias, es decir, para comparar datos experimentales, obtenidos en forma de frecuencias o proporciones, con frecuencias esperadas. Esto es, probar estadísticamente si la distribución de frecuencias observadas es compatible ("se ajusta a") con alguna distribución teórica conocida: uniforme, multinomial, binomial, Poisson, Normal, etcétera. A estas pruebas se les denomina **pruebas de bondad de ajuste**.

- Para probar preferencias o pruebas de independencia, llamadas también tablas de contingencia.

4.1. Pruebas bondad de ajuste

Se usa bondad y ajuste para referirse a la comparación de la distribución de una muestra con alguna distribución teórica que se supone describe a la población de la cual se extrajo la muestra.

Se le atribuye a Karl Pearson (1857-1936), la justificación de emplear la Ji-cuadrada como prueba de la congruencia entre observación, siempre que los datos estén en forma de frecuencias.

Por ejemplo, podría ser necesario determinar si una muestra de valores observados para alguna variable aleatoria es compatible con la hipótesis de que la muestra se extrajo de una población de valores con distribución normal, o binomial, de Poisson o cualquier otra.

A continuación se establece la prueba χ^2 para bondad de ajuste (Marques de Cantú): Supóngase que al realizar un experimento aleatorio n veces, se presentan los resultados R_1, R_2, …, R_k con frecuencias observadas O_1, O_2, …, O_k, y que de acuerdo con las leyes de la probabilidad, se espera que los mismos resultados se presentan con frecuencias E_1, E_2, …, E_k.
Una medida de las diferencias entre frecuencias observadas y las esperadas está dad por el estadígrafo χ^2 definido por:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \qquad k > 2$$

donde $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = n$.

Si las frecuencias observadas coinciden o se aproximan mucho a las esperadas, el valor del estadígrafo χ^2 tiene mucho a cero. Por el contrario, si las frecuencias observadas difieren significativamente de las esperadas, el valor del estadígrafo χ^2 será positivo y tan grande cuanto mayor sean las diferencias entre las frecuencias. Bajo estas condiciones se tiene que la región de rechazo es **solo la región derecha** (*cola derecha o unilateral superior*), y las hipótesis son las siguientes

Hipótesis (Hipótesis para bondad de ajuste)**.**

H_0: los datos provienen de una muestra al azar de una población distribuida de acuerdo a un modelo teórico.
H_a: los datos no provienen de una población distribuida de acuerdo al modelo teórico

El estadígrafo de contraste es:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \qquad \text{o} \qquad \chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n$$

con $k - r$ grados de libertad, donde:

- ✔ k es el número de eventos o categorías
- ✔ r es el número de restricciones ($r \geq 1$, ya que $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = n$ es siempre una restricción, y cada parámetro que se tenga que estimar es una restricción más)

En ocasiones, las frecuencias esperadas dan resultados menores que 1, y los investigadores frecuentemente hacen notar en la literatura que el estadígrafo $\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ no se distribuye como χ^2 si las frecuencias esperadas son pequeñas. Algunos establecen que para poder usar la ji-cuadrada las E deben ser al menos 5 ($E_i \geq 5$). Otros opinan que se puede usar la χ^2 si las E_i tienen un valor de por lo menos 1 ($E_i \geq 1$). Si en la práctica resulta una o varias $E_i < 1$ se juntan

las categorías adyacentes para que se cumpla el supuesto ($E_i \geq 1$). Al combinar dos o más categorías adyacentes, el número de estas se reduce y por tanto los grados de libertad también se reducen

Ejemplo 17 (Distribución uniforme (Marques de Cantú))**.**

Una pequeña fabrica de dulces que elabora y empaca pequeños caramelos de azúcar, mezcla, con igual proporción, seis colores de los caramelos. Para probar la proposición de proporciones iguales, se toma una muestra al azar, con los siguientes resultados:

Color	Rosa	Lila	Amarillo	Anaranjado	Verde	Blanco
Frecuencia	20	17	18	14	10	11

Probar la proposición de proporciones iguales con $\alpha = 0.05$

Solución.
Las hipótesis en este caso son:
H_0: Los datos provienen de una distribución uniforme, esto es, proporciones iguales y la distribución es: $p_i = \frac{1}{6}$, $i = 1, 2, \ldots, 6$
H_a: Los datos no provienen de una distribución uniforme.

Las probabilidades son todas $p_i = \frac{1}{6}$ y las frecuencias esperadas son $E_i = p_i \cdot n = \frac{1}{6}(90) = 15$. Los cálculos para el estadístico son:

O_i	E_i	O_i^2/E_i
20	15	$20^2/15$
17	15	$17^2/15$
18	15	$18^2/15$
14	15	$14^2/15$
10	15	$10^2/15$
11	15	$11^2/15$
$\sum_{i=1}^6 O_i = n = 90$	$\sum_{i=1}^6 E_i = n = 90$	$\sum_{i=1}^6 O_i^2/E_i = 1430/15 = 95.33$

Así $\chi^2 = \sum_{i=1}^{6} \frac{O_i^2}{E_i} - n = 95.33 - 90 = 5.33$

La región crítica con $\alpha = 0.05$ está definida como $\chi > \chi^2_{0.05} = 11.07$ con $\nu = 6 - 1 = 5$ g.l. En consecuencia, se acepta H_0, la muestra observada no presenta suficiente evidencia que indique que los datos no provienen de una distribución uniforme. ■

Ejemplo 18 (Distribución multinomial (Marques de Cantú)).

Una caja contiene gran cantidad de tornillos de 4 diferentes tamaños. Se supone que están en la proporción: 8 : 7 : 3 : 2. Una muestra de 400 contiene 180 del tamaño 1, 120 del 2, 40 del 3 y 60 del 4.

a) Estime las proporciones para cada tipo de tornillos en la caja.

b) Pruebe la hipótesis de que los tornillos están en la proporción 8 : 7 : 3 : 2 con $\alpha = 0.005$

Solución.

a) $p_1 = \frac{8}{20}$, $p_2 = \frac{7}{20}$, $p_3 = \frac{3}{20}$, $p_4 = \frac{2}{20}$

b) Este caso las hipótesis son:
H_0: Los tornillos tienen la distribución 8 : 7 : 3 : 2, es decir, $p_1 = \frac{8}{20}$, $p_2 = \frac{7}{20}$, $p_3 = \frac{3}{20}$, $p_4 = \frac{2}{20}$.
H_a: Los tornillos no están en las proporciones esperadas.

El estadístico de prueba es $\chi^2 = \sum_{i=1}^{4} \frac{O_i^2}{E_i} - n$

La región de rechazo con $\alpha = 0.005$, $\nu = 4 - 1 = 3$ g.l. es $\chi^2 > \chi^2_{0.005} = 12.838$

Los cálculos para el estadístico se muestran en la tabla, con $E_i = p_i \cdot n$, con $n = 400$ (el tamaño de muestra)

O_i	E_i	O_i^2/E_i
180	160	202.5
120	140	102.8571
40	60	26.667
60	40	90
$\sum O_i = 400$	$\sum E_i = 400$	$\sum O_i^2/E_i = 422.0238$

Así $\chi^2 = \sum \frac{O_i^2}{E_i} - n = 422.0328 - 400 = 22.0238$

Debido a que $\chi^2 = 22.03 > \chi^2_{0.005} = 12.83$ se rechaza H_0, por tanto, los tornillos no están en las proporciones esperadas 8 : 7 : 3 : 2. ■

Ejemplo 19 (Distribución de Poisson (Marques de Cantú))**.**

El administrador de un hospital ha estado estudiando el número de urgencias que llegan al hospital por día y sospecha que estas se distribuyen según Poisson. También ha determinado que el número medio de urgencias por día es de 3 ($\lambda = 3$). Para determinar si efectivamente el número de urgencias por día que llegan al hospital siguen la ley de Poisson, tomamos una muestra al azar de 90 días de los archivos del hospital. Los datos se resumen en la siguiente tabla.

Número de urgencias por día	0	1	2	3	4	5	6	7	8	9	10 o más
Número de días con este número de urgencias	5	14	15	23	16	9	3	3	1	1	0

Solución.
Con $\lambda = 3$ y la tabla de distribución de Poisson, determinamos las probabilidades para $x = 0, 1, 2, \ldots, 9$ y para $x \geq 10$, restamos 1 de la suma de las anteriores, esto es:

$$f(X=0) = 0.0498, \quad f(X=1) = 0.1494, \quad \ldots. \quad f(x \geq 10) = 1 - 0.9988 = 0.0011$$

Para obtener las frecuencias esperadas E_i multiplicamos las probabilidades por $n = 90$

X_i	O_i	$f(X_i)$	E_i
0	5	0.049787	4.481
1	14	0.149361	13.442
2	15	0.224042	20.164
3	23	0.224042	20.164
4	16	0.168031	15.123
5	9	0.100819	9.074
6	3	0.050409	4.537
7	3	0.021604	1.944
8	1	0.008102	0.729
9	1	0.002701	0.243
10	0	0.001102	0.099
Total	90	1	90

$\Rightarrow$

X_i	O_i	$f(X_i)$	E_i
0	5	0.049787	4.481
1	14	0.149361	13.442
2	15	0.224042	20.164
3	23	0.224042	20.164
4	16	0.168031	15.123
5	9	0.100819	9.074
6	3	0.050409	4.537
7	3	0.021604	1.944
8 o más	2	0.011905	1.071
Total	90	1	90

Tabla 4.1: Los valores para $X = 8, 9, 10$ se juntan pues E_i eran menores a 1

Así el valor calculado de χ^2 es

$$\chi^2 = \sum_{i=1}^{9} \frac{O_i^2}{E_i} - n = 93.7563 - 90 = 3.7563$$

El valor teórico para $\nu = 9-1 = 8$ grados de libertad (puesto que λ fue dada) y $\alpha = 0.05$ es $\chi^2_{0.05} = 15.507$.
Concluimos por tanto que el número de urgencias por día que llegan al hospital sigue una distribución de Poisson con $\lambda = 3$. ■

Ejemplo 20 (Distribución normal (Marques de Cantú))**.**

Supongamos que se tiene una tabla con datos agrupados de 100 observaciones como la que se muestra

Clases	Frec.
(80, 90]	8
(90, 100]	15
(100, 110]	21
(110, 120]	23
(120, 130]	16
(130, 140]	9
(140, 150]	8
Total	100

Determine si la muestra provino de una distribución normal con $\alpha = 0.05$,

Solución.
Se procede de la siguiente manera

✍ Debemos calcular $\bar{x}$ y s puesto que no se conocen μ y σ, es decir, la media y la desviación estándar para datos agrupados. El calcular o estimar dos parámetros como $\bar{x}$ y s esto genera "restricciones" que disminuyen los grados de libertad para el valor crítico de las tablas.

Las marcas de clase de cada intervalo son

Clases	Frec.	m_i
(80, 90]	8	85
(90, 100]	15	95
(100, 110]	21	105
(110, 120]	23	115
(120, 130]	16	125
(130, 140]	9	135
(140, 150]	8	145
Total	100	

donde

$$\bar{x} = \frac{\sum f_i m_i}{n} = 113.30 \qquad s = \sqrt{\frac{\sum f_i m_i^2 - n\bar{x}^2}{n-1}} = 16.64$$

✍ Ahora calculamos las probabilidades de cada intervalo, primero normalizamos los extremos de los intervalos con $Z = \dfrac{X - \bar{x}}{s}$

Clases	Frec.	$(z_i, z_{i+1}]$
< 90	8	$(-\infty, -1.40]$
(90, 100]	15	$(-1.40, -0.80]$
(100, 110]	21	$(-0.80, -0.20]$
(110, 120]	23	$(-0.20, 0.40]$
(120, 130]	16	$(0.40, 1.0]$
(130, 140]	9	$(1.0, 1.60]$
$140 <$	8	$(1.60, \infty]$
Total	100	

Para determinar la probabilidad de la clase $(-\infty, 90]$ calculamos la probabilidad $P(z < -1.40) = 0.0808$. La probabilidad de la segunda clase $(90, 100]$ la obtenemos calculando $P(-1.40 < Z < -0.80) = P(Z < -0.8) - P(Z < -1.4) = 0.1311$, de manera análoga encontramos las probabilidades de los otros intervalos, como se muestra en la tabla

Clases	Frec.	$(z_i, z_{i+1}]$	p_i
< 90	8	$(-\infty, -1.40]$	0.0808
(90, 100]	15	$(-1.40, -0.80]$	0.1311
(100, 110]	21	$(-0.80, -0.20]$	0.2088
(110, 120]	23	$(-0.20, 0.40]$	0.2347
(120, 130]	16	$(0.40, 1.0]$	0.1859
(130, 140]	9	$(1.0, 1.60]$	0.1039
$140 <$	8	$(1.60, \infty]$	0.0548
Total	100		

✍ Calculamos la frecuencia esperada de cada clase como $E_i = p_i\, n$ donde $n = 100$ (total de datos),

Clases	Frec.	$(z_i, z_{i+1}]$	p_i	E_i
< 90	8	$(-\infty, -1.40]$	0.0808	8.08
(90, 100]	15	$(-1.40, -0.80]$	0.1311	13.11
(100, 110]	21	$(-0.80, -0.20]$	0.2088	20.88
(110, 120]	23	$(-0.20, 0.40]$	0.2347	23.47
(120, 130]	16	(0.40, 1.0]	0.1859	18.59
(130, 140]	9	(1.0, 1.60]	0.1039	10.39
$140 <$	8	$(1.60, \infty]$	0.0548	5.48
Total	100			

El valor del estadístico es:

$$\chi^2 = \sum_{i=1}^{7} \frac{O_i^2}{E_i} - n = 101.989 - 100 = 1.989$$

✍ Para la conclusión, consideramos el valor χ^2_{tab} con 4 grados de libertar (7 clases $-$ 1 $-$ 2 restricciones (estimaciones de $\bar{x}$ y s)), es decir

$$\chi^2_{tab} = \chi^2_{0.05,4} = 9.488$$

Por tanto, concluimos que la muestra si tiene una distribución normal ya que $\chi^2 < \chi^2_{tab}$

■

Ejercicios

I. La teoría de Mendel dice que el número de un tipo de chícharos que cae en las clasificaciones redonda y amarilla, arrugada y amarilla, redonda y verde, y arrugada y verde debe estar en la proporción 9:3:3:1. Suponga que 100 de estos chícharos revelaron 56, 19, 17 y 8 en las respectivas categorías. ¿Estos datos son consistentes con el modelo? Use a = .05. (La expresión 9:3:3:1 significa que 9/16 de los chícharos deben ser redondos y amarillos, 3/16 deben ser arrugados y amarillos, etc.)

II. Durante un periodo fijo se observó el número de accidentes sufridos por mecánicos, con los resultados que se ven en la siguiente tabla. Pruebe, con un nivel de significancia de 5%, la hipótesis de que los datos provienen de una distribución de Poisson. Considere el parámetro de la distribución como el promedio de los datos agrupados.

Accidentes mecánico	0	1	2	3	4	5	6	7	8
Frecuencia de observación	296	74	26	8	4	4	1	0	1

4.2. Tablas de contingencia

Otro uso de la distribución ji-cuadrada es la prueba de hipótesis de que dos criterios de clasificación, cuando son aplicados a las mismas unidades elementales, son independientes.

La clasificación en dos criterios de los mismos individuos se hace en las llamadas **tablas de contingencia** en la cual los renglones representan los niveles de un criterio y las columnas representan los niveles del otro criterio de clasificación. Por ejemplo, un criterio de clasificación puede ser los ingresos anuales por familia y el otro criterio las zonas donde viven los habitantes de una ciudad. Si los ingresos anuales por familia y las zonas donde viven **son independientes**, entonces tendríamos que en todas las zonas de la ciudad vivirían la misma proporción de familias de bajos, medios y altos ingresos.

Hipótesis (Hipótesis de independencia)**.**

Las hipótesis de prueba son:
H_0: los dos criterios no están relacionados (o los criterios son independientes) vs
H_a: si existe relación entre los criterios (o los criterios no son independientes, sino dependientes)

La clasificación de los elementos de una muestra, de acuerdo a dos criterios se representa mediante las llamadas Tablas de Contingencia.

Las filas o renglones: $R_1, R_2, \ldots, R_r$ representan los diferentes r niveles del primer criterio y las columnas $C_1, C_2, \ldots, C_c$ representan los diferentes c niveles del segundo criterio.

Estos niveles o categorías se definen mutuamente excluyentes y se registran las frecuencias observadas.

El arreglo general de una tabla de contingencia $r \times c$ es de la forma

		Segundo Criterio					
		C_1	C_2	C_3	$\ldots$	C_c	Total
Primer Criterio	R_1	a_{11}	a_{12}	a_{13}	$\cdots$	a_{1c}	$a_{1\bullet}$
	R_2	a_{21}	a_{22}	a_{23}	$\cdots$	a_{2c}	$a_{2\bullet}$
	R_3	a_{31}	a_{32}	a_{33}	$\cdots$	a_{3c}	$a_{3\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	R_r	a_{r1}	a_{r2}	a_{r3}	$\cdots$	a_{rc}	$a_{r\bullet}$
	Total	$a_{\bullet 1}$	$a_{\bullet 2}$	$a_{\bullet 3}$	$\cdots$	$a_{\bullet c}$	n

Teorema 2 (Prueba de independencia (Walpole)).

En cada ij-celda se tiene la frecuencia observada O_{ij}, es decir, $O_{ij} = a_{ij}$ formada por la interacción del nivel i del primer criterio y el nivel j del segundo criterio.
Las frecuencias esperadas E_{ij} se calculan, suponiendo la hipótesis nula cierta, por lo que se definen como el producto de la frecuencia relativa de la fila i, la frecuencia de la columna j por el tamaño de la muestra, esto es, $E_{ij} = \dfrac{a_{i\bullet}a_{\bullet j}}{n}$
El estadístico de prueba se define por

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Si $\chi^2 > \chi^2_\alpha$ con $(c-1)(f-1)$ grados de libertad, rechace la hipótesis nula de independencia al nivel de significancia α; en otro caso no la rechace

Ejemplo 21 (Prueba de independencia (Daniel)).

En la siguiente tabla se muestran los resultados obtenidos en una investigación realizada en una muestra de 300 adultos residentes en cierta área metropolitana. A cada individuo se le pidió que indicara cuál de las políticas sobre fumar en lugares públicos preferían y su nivel máximo de educación. ¿Es posible concluir a partir de estos datos, que en la población muestreada existe una relación entre el nivel de educación y la actitud hacia el hábito de fumar en lugares públicos? Usar $\alpha = 5\,\%$.

	Sin restricción	En áreas especiales	Prohibición	Sin opinión	Total
Licenciatura	5	44	23	3	75
Preparatoria	15	100	30	5	150
Primaria	15	40	10	10	75
Total	35	184	63	18	300

Solución.
Las hipótesis de prueba son:
H_0: El nivel de educación y las políticas sobre fumar son independientes (es decir, los dos criterios no están relacionados)
H_1: El nivel de educación y las políticas sobre fumar en lugares públicos no son independientes (es decir, si existe relación entre los dos criterios)

Suponiendo la hipótesis nula verdadera calculamos las frecuencias esperadas con la fórmula $E_{ij} = \dfrac{a_{i\bullet}a_{\bullet j}}{n}$

$E_{11} = \frac{(75)(35)}{300} = 8.75$	$E_{12} = \frac{(75)(184)}{300} = 46$	$E_{13} = \frac{(75)(63)}{300} = 15.75$	$E_{14} = \frac{(75)(18)}{300} = 4.5$
$E_{21} = \frac{(150)(35)}{300} = 17.5$	$E_{22} = \frac{(150)(184)}{300} = 92$	$E_{23} = \frac{(150)(63)}{300} = 31.5$	$E_{24} = \frac{(150)(18)}{300} = 9$
$E_{31} = 8.75$	$E_{32} = 46$	$E_{33} = 15.75$	$E_{34} = 4.5$

El valor del estadístico es por tanto

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{O_{ij}^2}{E_{ij}} - n = 322.50 - 300 = 22.5$$

Considerando $\alpha = 5\,\%$, el valor en las tablas de ji-cuadrada con $\nu = (3-1)(4-1) = 6$ grados de libertad, se tiene $\chi^2_{0.05} = 12.59$. Puesto que el valor calculado es mayor que el valor crítico, esto es $\chi^2 = 22.5 > 12.592 = \chi^2_\alpha$.
La conclusión de la prueba es que H_0 se rechaza. Esto significa entonces, que sí existe relación entre el nivel de educación y las políticas sobre fumar en lugares públicos, es decir no son independientes. ■

Ejercicios

I. En un experimento diseñado para estudiar la dependencia de la hipertensión con respecto a los hábitos de fumar se tomaron los datos que se muestran en la tabla de 180 individuos. Pruebe la hipótesis de que la presencia o ausencia de hipertensión es independiente de los hábitos de tabaquismo. Utilice un nivel de significancia de 0.05.

	No fumadores	Fumadores moderados	Fumadores empedernidos
Con hipertensión	21	36	30
Sin hipertensión	48	26	19

II. Joseph Jacobson y Diane Wille realizaron un estudio para determinar el efecto del cuidado temprano de niños con patrones de apego entre hijo y madre. En el estudio, 93 infantes fueron clasificados como “seguro” o “ansioso” usando el paradigma Ainsworth de situación extraña. Además, los infantes fueron clasificados de acuerdo con el número promedio de horas por semana que recibían cuidado. Los datos aparecen en la tabla. ¿Los datos indican dependencia entre patrones de apego y el número de horas de atención al niño? Pruebe usando $\alpha = 0.05$

	Horas de cuidado de infantes		
Patrón de afecto	Bajo (0-3 hrs)	Moderado (4-19 hrs)	Alto (20 - 54 hrs)
Seguro	24	35	5
Ansioso	11	10	8

III. Un estudio de la cantidad de violencia en televisión en lo que respecta a la edad del televidente dio los resultados que se muestran en la siguiente tabla, para 81 personas. (Cada una

de las personas del estudio fue clasificada, de acuerdo con los hábitos de ver TV de la persona, como que ve poca violencia o mucha violencia.) ¿Los datos indican que ver violencia no depende de la edad del televidente con un nivel de signif cancia de 5 %?

	Edad		
Ve	16-34	35-54	55 y más
Poca violencia	8	12	21
Mucha violencia	18	15	7

Bibliografía

[Daniel] Daniel W. W. (2010). Bioestadística. (3ra ed).México: Limusa Wiley

[Montgomery] Montgomery D. C. (2005). Diseño y Análisis de Experimentos. México: Iberoamérica

[Walpole] Walpole, R. E., Myers, R. H, Myers, S,L, y Ye K. (2012) Probabilidad y Estadística para Ingeniería y Ciencias. México: Pearson Education.

[Gutiérrez & De La Vara] Gutiérrez Pulido, H & De La Vara Salazar, R. (2008) Análisis y Diseño de Experimentos (2da edición) Mc Graw Hill

[Marques de Cantú] Marques de Cantú, M. J. (1991). Probabilidad y Estadística para ciencias químico-biológicas. McGrawl-Hill