Diseño y conducción de Experimentos

en Arquitectura de Computadores -Terminología y Conceptos

Contenido

1. Diseño de Experimentos-Terminología y Conceptos

1.1 INTRODUCCION

En la conducción de experimentos DoE (Design of Experiments), un ingeniero realiza cambios deliberados o intencionados en las variables controlables del sistema o proceso, observa entonces los datos de salida del sistema y luego hace una inferencia o decisión acerca de qué variables son responsables de los cambios observados en el rendimiento de salida.

Como cualquier disciplina altamente técnica, DoE está lleno de siglas y terminología que pueden intimidar al estudiante que lo estudia por primera vez. El propósito de este capítulo es entonces proporcionar de una manera no tan compliacada, una introducción a los muchos términos, conceptos y cuestiones de conducción de DoE. No espere entender todos los aspectos del DoE presentados en este capítulo la primera vez que lo lee. Muchos de los matices solo se harán evidentes después de su aplicación, ya sea en el trabajo final en éste curso o luego en la aplicación en otras materias o tal vez en su trabajo de grado y vida profesional. Después de leer este capítulo, debe tener una comprensión suficiente de la terminología y lenguaje de DoE para pasar a la consideración de diseño de experimentos y sus aplicaciones.

En esta sección se intenta explicar los términos que se utilizan en DoE mediante el ejemplo, suponiendo un estudio o análisis de un sistema de cómputo. Suponga que el problema es determinar la configuración de un sistema de cómputo, donde varias opciones pueden aplicarse. En primer lugar, escoger el microprocesador o tipo de CPU. Las alternativas podrían ser unas versiones de los microprocesadores Intel i7, ARM v11 o AMD FX. En segundo lugar, supóngase que debe determinar el tamaño de memoria principal entre 2, 4 y 8GBytes. En tercer lugar, diferentes tamaños de memoria caché L3. En cuarto lugar, la carga de trabajo que el sistema puede ejecutar, que puede ser uno de tres tipos: trabajo de oficina, video juego y transacciones de un servidor Web.

1.2 DISEÑO DE EXPERIMENTOS: DEFINICIÓN, ÁMBITO DE APLICACIÓN, Y MOTIVACION

Los diseños y conducción de experimentos desempeñan un papel muy importante en el diseño y desarrollo de ingeniería y en la fase de validación de un servicio, modelo o desarrollo de un componente software o hardware de un sistema en condiciones más cercanas a los escenarios de trabajo. Generalmente, cuando los productos y procesos se diseñan y desarrollan con DoE, presentan de un mejor rendimiento, mayor confiabilidad y menores costos generales. Desempeñan un papel crucial en la reducción del tiempo de espera para el diseño de ingeniería y las actividades de desarrollo.

El diseño de experimentos (DoE) es una técnica formal y estructurada para estudiar cualquier situación. Eso implica una respuesta que varía en función de una o más variables independientes. El DoE está diseñado específicamente para abordar problemas complejos donde más de una variable puede afectar una respuesta y dos o más variables pueden interactuar entre sí.

El DoE reemplaza los métodos inferiores, como el tradicional pero desafortunadamente todavía común método de estudio del efecto de una variable a la vez (OVAT, One Variable at a Time). Comparado con el DoE, el método OVAT es un uso ineficiente de los recursos y es incapaz de detectar la presencia de o la cuantificación de las interacciones entre variables.

DoE se utiliza donde se recopilan y analizan los datos experimentales. Su uso es esperado en todas las ramas de la investigación científica, pero el DoE está cada vez más extendido en ingeniería, manufacturing, biología, medicina, economía, sociología, psicología, marketing, agricultura, etc.

La popularidad del DoE se debe a su tremendo poder y eficiencia. Cuando se utiliza correctamente, el DoE puede proporcionar las respuestas a preguntas específicas sobre el comportamiento de un sistema utilizando un número óptimo de observaciones experimentales. Es decir, experimentos diseñados y estructurados para responder preguntas específicas con rigor estadístico, pero de experimentos con muy pocas observaciones no se podrá

obtener la confianza deseada en los resultados y experimentos con demasiadas observaciones se desperdiciarían los recursos. DoE da las respuestas que buscamos con un gasto mínimo de tiempo y recursos.

1.3 DEFINICIÓN DE UN EXPERIMENTO

Un modelo simple de un proceso se muestra en la Figura 1.1. Los procesos tienen entradas que determinan cómo el proceso opera y salidas que son producidas por dicho proceso.
El propósito de un experimento es determinar cómo las entradas afectan las salidas. Los experimentos pueden ser realizados para documentar con propósitos científicos, el comportamiento de las entradas y salidas, pero el objetivo de la experimentación de ingeniería, que es lo que nos compete, es descubrir o validar algo acerca de un proceso o sistema particular, utilizando modelos y herramientas estadísticas y software computacional. Las entradas de proceso se llaman variables, factores, o predictores y salidas de proceso se llaman respuestas.

Figura 1.1 Modelo simple de un proceso.

Cada experimento implica la observación de las entradas (las variables) y las salidas (las respuestas). La acción tomada por el experimentador en las entradas determina si el experimento es pasivo o activo. Cuando el experimentador simplemente observa el sistema y registra cualquier cambio que ocurra en las entradas y las salidas correspondientes, el experimento es pasivo. Este tipo de experimentación puede ser costoso, requiere mucho tiempo y es improductivo. Cuando el experimentador varía intencionalmente las entradas, entonces el experimento es activo. La experimentación activa, realizada bajo condiciones controladas de forma lógica de manera estructurada, es una herramienta tremendamente poderosa y es el tipo de experimentación utilizado en el DoE.

En éste punto vale la pena reflexionar sobre qué tipo de respuestas (salida) del sistema de cómputo, por ejemplo, estaríamos interesados en estudiar, y especialmente ¿para qué? ¿En qué casos, nosotros como ingenieros debemos usar un diseño estadístico experimental?

la mayoría de las veces, llevar a cabo un experimento nace de la necesidad de realizar una evaluación de desempeño de uno o varios sistemas de cómputo y donde las variables de entrada son algunos de los muchísimos factores (de hardware, de software, ambientales, sociales, etc.) que pueden afectar la respuesta la cual estamos interesados en estudiar. ¿Pero en qué situaciones o escenarios nos vemos en la necesidad de evaluar el desempeño?
Y la respuesta la podemos ver de dos formas, la primera, es cuándo necesitamos comparar dos o más sistemas para decidir cuál es el mejor o la opción más adecuada, de acuerdo a una métrica de desempeño establecida, o en análisis costo beneficio de varias alternativas que dispongamos en la fase de diseño dentro del ciclo de vida de un producto, servicio o sistema. También en la fase de validación de una mejora de un componente o arquitectura respecto a otra.
La otra situación donde un análisis de los experimentos también ayuda es en la separación de los efectos de diversos factores que pueden afectar el rendimiento. Permite la determinación de si un factor tiene un efecto significativo o si la diferencia observada es simplemente debido a variaciones aleatorias causadas por errores de medición y parámetros que no fueron controlados.

El objetivo de un adecuado diseño experimental es obtener el máximo de información con el mínimo número de experimentos. Un buen diseño ahorra una considerable cantidad de trabajo que se habría gastado en la recopilación de datos. El diseño de experimentos se apoya en los métodos estadísticos porque nos ayudan a describir y comprender la variabilidad. Por variabilidad, se quiere decir, que las observaciones sucesivas de un sistema o fenómeno no producen exactamente el mismo resultado. Todos encontramos variabilidad en nuestra vida cotidiana, y el pensamiento estadístico puede brindarnos una manera útil de incorporar esta variabilidad en nuestros procesos de toma de decisiones.

Por ejemplo, considere el tiempo de autonomía o duración de la batería de su Smartphone. ¿Siempre obtiene exactamente el mismo rendimiento de duración de la batería? Por supuesto que no, de hecho, a veces el rendimiento de la batería varía considerablemente. Esta variabilidad observada en tiempo de duración de la batería depende de muchos factores, el tipo de batería (Li-ion), la intensidad de la iluminación y tiempo de uso de la pantalla, o el tipo aplicaciones que se usan, los servicios activos como GPS, conexión WiFi, uso de bluetooth, volumen y tiempo de reproducción de sonido. La distancia en la que se encuentra a la antena celular del operador de telefonía móvil, los cambios en la utilización o tiempo de carga de la batería (que podrían influir en los ciclos de carga de la batería), incluso la temperatura ambiente y la humedad pueden alargar o reducir la autonomía de un Smartphone. Estos factores representan fuentes potenciales de variabilidad en el sistema. Las estadísticas nos brindan un marco para describir esta variabilidad y para conocer qué fuentes potenciales de variabilidad son las más importantes o cuáles tienen el mayor impacto en el tiempo de autonomía de la batería del smartphone.

1.4 IDENTIFICACIÓN DE VARIABLES Y RESPUESTAS

Quizás la mejor manera de identificar y documentar las muchas variables y respuestas de un proceso es construir un diagrama de causa y efecto. Considere el ejemplo que se muestra en la Figura 1.2. El problema de tiempo de inactividad(fuera de servicio) de un computador.

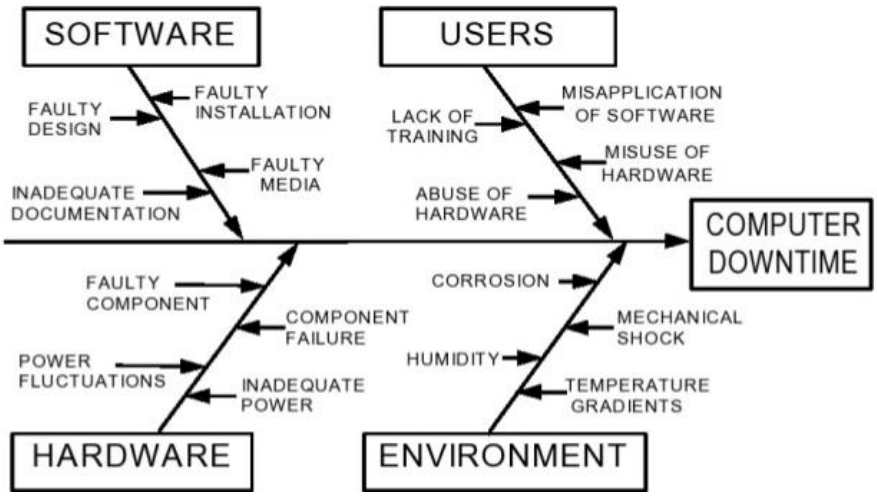


Figura 1.2 Diagrama de causa y efecto de tiempo de inactividad (fuera de servicio) del computador. Imagen tomado de: https://www.slideshare.net/johnpadua/production-and-quality-tools-seven-quality-tools-and-introduction-to-statistics

Las entradas del proceso se muestran a la izquierda y las respuestas se muestran a la derecha. Tenga en cuenta que este diagrama es sólo una elaboración de un modelo simple como el proceso en la figura 1.1. Es útil tener una clasificación de factores, en el ejemplo hay 4 categorías de factores, los referentes al software, los debidos al hardware, los ambientales y los problemas causados por los usuarios. En el artículo: "A Factor Framework for Experimental Design for Performance Evaluation of Commercial Cloud Services", (Li, 2013), puede darse cuenta que antes de iniciar el experimento para la evaluación de desempeño, se debe definir por un lado la(s) métrica(s) de medición y por otro la selección de factores para la evaluación de los servicios de nube, pero esa no es una tarea tan elemental. Teniendo el objetivo claro y entendiendo muy bien el problema un diagrama de causa y efecto sería de mucha ayuda. El artículo propone un framework general de factores para el diseño experimental con el objetivo de evaluar las alternativas de plataformas de servicio en la nube. En el siguiente diagrama puede darse una idea de la diversidad de factores y características de desempeño:

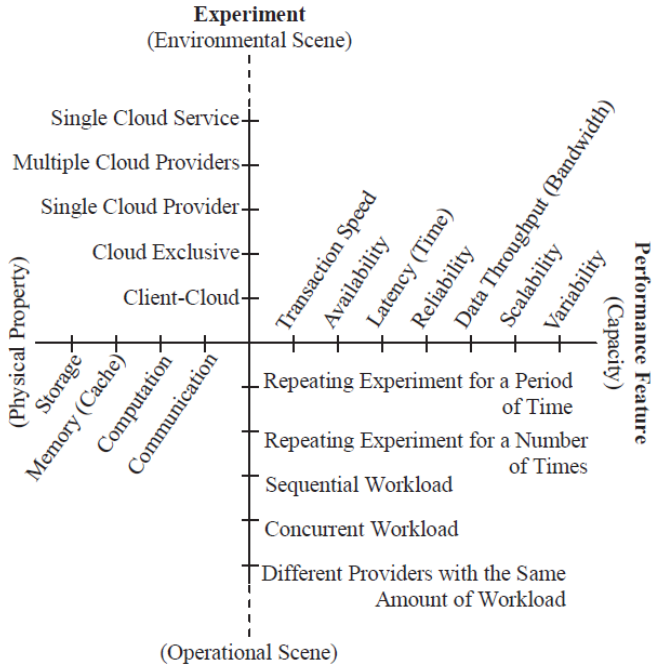


Figura 1.3 Two-dimensional taxonomy of performance evaluation of commercial Cloud services.

MINITAB tiene la capacidad de crear diagramas causa-efecto, en la barra de menú **Stat> Quality Tools> Cause and Effect**.

Es muy importante crear un diagrama de causa y efecto para un experimento. Los diagramas de causa y efecto:

- Proporciona un lugar conveniente para recopilar ideas para nuevas variables.
- Sirve como una herramienta de referencia rápida cuando las cosas salen mal y cuando son necesarios decisiones rápidas.
- Resume todas las consideraciones hechas de las variables durante la vida del experimento.
- Proporciona una excelente fuente de información para planificar nuevos experimentos.

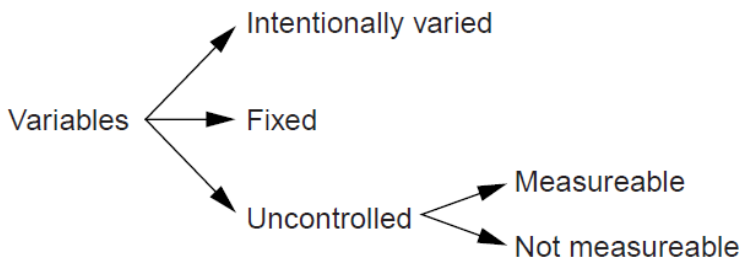


Figure 1.4 Disposition of the experimental variables.

Cada proceso tiene una multitud de variables y todas ellas juegan un papel en un experimento. Algunas variables son intencionalmente variadas por el experimentador para determinar su efecto en las respuestas, otras se mantienen constantes para asegurar que no afecten las respuestas y algunas variables se ignoran con la esperanza de que tendrán poco o ningún efecto en las respuestas. Si el diseño del experimento es bueno, si se lleva a cabo con cuidado, y si se cumplen los supuestos sobre las variables no controladas, el experimentador puede aprender algo sobre el problema. Esta clasificación de las variables de proceso se resume en la Figura 1.3. Un diagrama de causa y efecto se puede aumentar utilizando resaltadores de diferentes colores para clasificar las variables en clases intencionalmente variadas, fijas y no controladas.

1.5 TIPOS DE VARIABLES

Las entradas a un proceso se conocen como variables, ***factores*** o predictores. Cada variable en un experimento tiene sus propios ajustes únicos denominados niveles o tratamientos. La relación entre los niveles de una variable determina si la variable es cualitativa o cuantitativa. Los niveles de una variable cualitativa difieren en tipo. Por ejemplo, en el problema de un sistema de cómputo, la variable tipo de procesador puede tener tres niveles cualitativos determinados por el fabricante: AMD FX-6300, ARM Cortex-A53 e Intel Core i7-3930K por ejemplo. Una variable cuantitativa tiene niveles que difieren en tamaño. Por ejemplo, la variable de tamaño de cache de nivel 3, puede aparecer en el experimento en cuatro niveles cuantitativos: 2, 4, 6 y 12 MB. Una ventaja de una variable cuantitativa es que los resultados del experimento se pueden usar para interpolar entre los niveles de la variable incluida en el experimento. Por ejemplo, el comportamiento de los tamaños de la cache podría usarse para predecir cómo se comportaría con un tamaño de 8MB.
Algunos experimentos incluyen solo una única variable de diseño, pero muchos de los experimentos en los que estaremos interesados contendrán dos o más variables. Aunque un experimento con más de una variable puede contener una mezcla de variables cualitativas y cuantitativas, los experimentos creados con solo variables cuantitativas generalmente ofrecen más posibilidades de diseño. A veces es posible, y generalmente es deseable, redefinir una variable cualitativa para que se vuelva cuantitativa. Esto puede requerir algo de imaginación, pero con la práctica y por necesidad, a menudo se hace posible.

1.6 VARIABLE RESPUESTA

Siempre que sea posible, la respuesta de un experimento debe ser cuantitativa. Se puede usar cualquier sistema de medición apropiado, pero debe ser repetible y reproducible.
En arquitectura de computadores, generalmente, la variable de respuesta es el rendimiento medido del sistema de cómputo. Por ejemplo, la variable de respuesta podría ser el rendimiento expresado por las tareas completadas por unidad de tiempo, o tiempo de respuesta por programa, o cualquier otra métrica. Dado que las técnicas de diseño experimental son aplicables para cualquier tipo de mediciones, no sólo las mediciones de rendimiento son válidas, el término respuesta es el término más general que se utiliza en lugar de performance o tiempo de ejecución.

La mayoría de los experimentos se realizan con el propósito de aprender sobre una sola respuesta; sin embargo, se pueden considerar múltiples respuestas. Por ejemplo, en el problema del rendimiento, suponga que el objetivo principal del experimento es aumentar el desempeño medido en MIPS (Millones de Instrucciones por Segundo) pero sin comprometer las características de consumo de energía. Esto requerirá que ambas respuestas se registren durante la ejecución del experimento y que los modelos se ajusten a cada respuesta. Una solución aceptable al problema debe satisfacer todos estos requisitos simultáneamente. Este tipo de problema es común y está completamente dentro de la capacidad del método DoE.
Un modelo un poco más detallado para un proceso o sistema bajo estudio puede representarse por medio del de la figura 1.4:

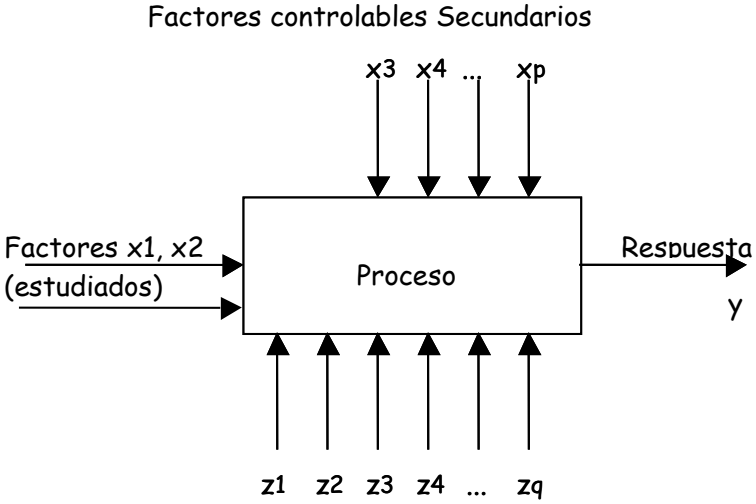


Figura 1.5 Modelo general de un proceso o

Algunas de las variables del sistema o proceso x1, x2,..., xk son controlables, mientras que otras z1, z2,...,zk son no controlables (aunque su efecto se puede "ocultar" o pueden ser controlables para los fines de prueba). Entre los objetivos del experimento pueden incluirse:

- Determinar cuáles variables tiene mayor influencia en la respuesta, y.
- Determinar el mejor valor de las x que influyen en y, de modo que y tenga casi siempre un valor cercano a valor nominal deseado. (menor tiempo de ejecución, por ejemplo)
- Determinar el mejor valor de las x que influyen en y, de modo que la variabilidad de y sea pequeña.
- Determinar el mejor valor de las x que influyen en y, de modo que se minimicen los efectos de las variables incontrolables z1, z2,...zq.

1.7 INTERACCIONES

Cuando un proceso contiene dos o más variables, es posible que algunas variables interactúen entre sí. Existe una interacción entre las variables cuando el efecto de una variable en la respuesta depende del nivel de otra variable. Las interacciones pueden ocurrir entre dos, tres o más variables, pero generalmente se asume que las interacciones de tres variables y de orden superior son insignificantes. Esta es generalmente una suposición segura, aunque hay ciertos sistemas donde las interacciones de orden superior son importantes. Con la práctica, las interacciones de dos factores entre las variables a menudo se pueden identificar mediante gráficos simples de la respuesta experimental representados en función de las dos variables involucradas. Estas gráficas generalmente muestran la respuesta representada en función de una de las variables con los niveles de la otra variable que se distinguen por diferentes tipos de líneas o símbolos. Los gráficos de múltiples variables también son útiles para identificar interacciones. Para que se haga una idea, recuerde el experimento en la práctica de medición del tiempo de ejecución de dos versiones de un programa (uno que aprovecha la localidad espacial y el otro no), y se graficó la respuesta (tiempo en ns) en función de los tamaños del arreglo. Con ello se pudo evidenciar algún tipo de interacción entre la variable tipo de algoritmo y el nivel de la variable tamaño del arreglo de datos.

El manejo de las interacciones entre variables es una fortaleza del método DoE y una debilidad del método de una variable a la vez (OVAT). Mientras que el DOE reconoce y cuantifica las interacciones entre variables para que puedan usarse para comprender y gestionar mejor la respuesta, el método OVAT ignora las interacciones y, por lo tanto, fallará en ciertos casos cuando los efectos de esas interacciones sean relativamente grandes. El éxito del DOE proviene de su consideración de todas las combinaciones posibles de niveles variables. OVAT falla porque se basa en un algoritmo simple pero defectuoso para determinar cómo las variables afectan la respuesta. En algunos casos, OVAT obtendrá el mismo resultado que DoE, pero en muchos otros casos su resultado será muy inferior.

Ejemplo 1.1 Dos variables, A y B, se pueden establecer en dos estados indicados por –1 y +1. Las Figuras 1.5 a y b muestran cómo las respuestas Y1 e Y2 dependen de las variables A y B. Use estas figuras para determinar si hay una interacción entre las dos variables y para demostrar cómo el DoE es superior a OVAT si el objetivo es maximizar las respuestas.

Solución: En la Figura 1.5a, los segmentos de línea que conectan los dos niveles de la variable B son sustancialmente paralelos, lo que indica que los niveles de B no causan un cambio en la forma en que A afecta la respuesta, por lo que probablemente no haya interacción entre A y B en este caso.
En la Figura 1.4b, los segmentos de línea que conectan los niveles de B divergen, lo que indica que el nivel elegido de B determina cómo A afecta la respuesta, por lo que probablemente haya una interacción entre A y B.

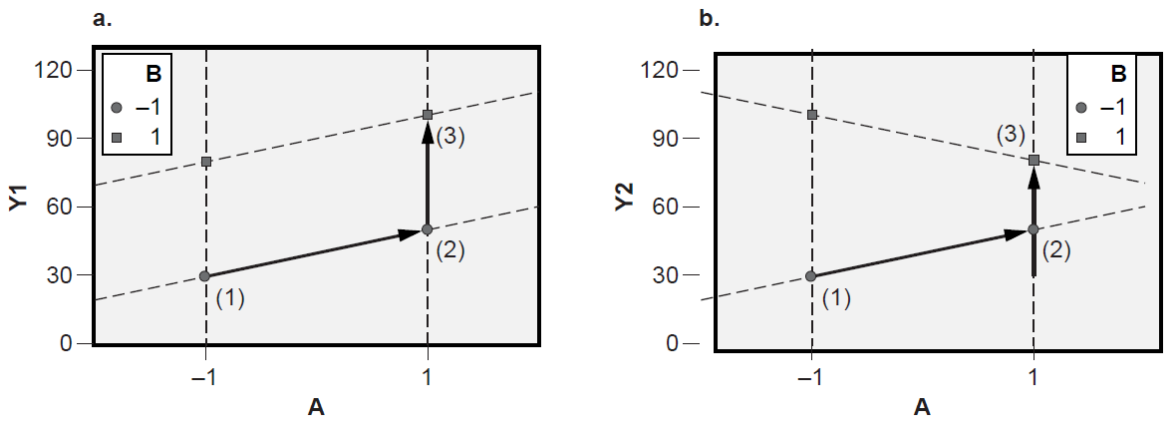


Figura 1.6 Ejemplos de dos variables sin interacción (a) y con interacción (b).

La debilidad del método OVAT es que sigue una ruta de decisión limitada a través del espacio de diseño que puede o no conducir a la solución óptima. En un experimento para estudiar las situaciones en las Figuras 1.5a y b, si el punto de inicio en el proceso OVAT es el punto (1) en ambos casos, y el primer paso en el experimento es investigar el efecto de la variable A, seguido un segundo paso para investigar B, luego se obtiene la respuesta máxima deseada en la Figura 1.5a (punto 3) pero no en b. En la Figura 1.5a, donde no hay interacción entre las variables A y B, se obtiene la solución óptima independientemente de qué variable se estudie primero. Pero en la Figura 1.5b, donde hay una interacción entre A y B, la solución máxima se obtiene solo si esas variables se estudian en el orden correcto. Por el contrario, el método DoE investiga las cuatro (cuatro porque tenemos 2 variables y cada una con dos niveles) configuraciones posibles en el espacio de diseño, por lo que se garantiza que encontrará la solución máxima independientemente de que A y B interactúen o no.

1.8 TIPOS DE EXPERIMENTOS

Dos de las consideraciones principales que distinguen los diseños de experimentos son la cantidad de variables de diseño que incluyen y la complejidad del modelo que proporcionan. Para un número específico de variables de diseño, podría haber muchos diseños de experimentos para elegir, pero los diseños extremos que abarcan todos los demás se denominan: *experimentos de detección* y *experimentos de superficie de respuesta*. Los experimentos de detección se utilizan para estudiar un gran número de variables de diseño con el fin de identificar las más importantes. Algunos experimentos de detección pueden evaluar muchas variables con muy pocas ejecuciones experimentales. Los experimentos de detección utilizan solo dos niveles de cada variable de diseño y no pueden resolver las interacciones entre pares de variables, una característica que puede hacer que estos diseños sean bastante riesgosos.

Los experimentos de superficie de respuesta son más complejos y difíciles de administrar que los experimentos de detección, por lo que generalmente involucran solo de dos a cinco variables. Cada variable en un diseño de superficie de respuesta debe ser cuantitativa y se requerirán tres o más niveles de cada variable. La ventaja de usar tantos niveles variables es que los diseños de superficie de respuesta proporcionan modelos muy complejos que incluyen al menos efectos principales, interacciones de dos factores y términos para medir la curvatura inducida en la respuesta por cada variable de diseño. Existe un conjunto intermedio de diseños de experimentos que se ubica entre los experimentos de detección y los experimentos de superficie de respuesta en términos de su complejidad y capacidad. Estos experimentos suelen utilizar dos niveles de cada variable de diseño y pueden resolver los efectos principales, las interacciones de dos factores y, a veces, las interacciones de orden superior, por eso tal vez escuchará términos como 2^k refiriéndose a diseños de k factores, cada uno de dos niveles.
Cuando las variables del diseño son todas cuantitativas, en estos diseños se puede incluir un conjunto selecto de ejecuciones adicionales con niveles de variable intermedios para proporcionar una prueba de la curvatura en la respuesta, pero no una resolución completa. La existencia de esta familia de diseños intermedios debería hacer evidente que en realidad existe un espectro discreto de diseños experimentales para un número dado de variables experimentales, donde el espectro está limitado por los diseños de detección y de superficie de respuesta.
Cuando se enfrenta a una nueva situación en la que hay poco conocimiento o experiencia previa, la mejor estrategia puede ser emplear una serie de experimentos previos más pequeños en lugar de dedicar todo el tiempo y los recursos disponibles a un experimento grande. El primer experimento que debe considerarse es un experimento de selección para determinar las variables más influyentes de entre las muchas variables que podrían afectar el proceso. Un experimento de selección para muchas variables usualmente identificará las dos o tres variables significativas que dominan el proceso. Por ejemplo, imagine un experimento sencillo de selección para determinar las variables más influyentes en el tiempo de convergencia de un algoritmo de ordenamiento de datos numéricos. Además del tipo de algoritmo que use, usted debería pensar en escoger al menos otros dos o tres

factores que piense afecta de manera significativa la respuesta, o descartar otras que en principio pensaba que eran relevantes y no afectaban la respuesta. Algunos factores estarán en el dominio del hardware de la máquina (características el procesador, tamaño de la memoria principal, tipo de tarjeta de video) y otros a nivel del software (sistema operativo, lenguaje de programación, formato de la variable, entre otras). El siguiente paso en la serie de experimentos sería construir un experimento más complejo que involucre las variables clave identificadas por el experimento de selección o las que son objetos de estudio. Este diseño debería al menos ser capaz de resolver interacciones de dos factores, pero a menudo el diseño elegido es un diseño de superficie de respuesta que puede caracterizar más completamente el proceso que se está estudiando.
En ocasiones, cuando se planifica una serie de experimentos de este tipo, los conocimientos proporcionados por los primeros experimentos son suficientes para indicar una solución efectiva al problema que inició el proyecto y entonces se puede suspender el programa experimental, pero si se requiere un análisis más profundo de las interacciones entre las variables y su efecto en la respuesta requerirá un diseño de experimento mucho más completo.

1.9 TIPOS DE MODELOS

Los modelos juegan un papel importante en el análisis de casi todos los problemas de ingeniería. Gran parte de la educación formal de los ingenieros implica aprender sobre los modelos relevantes para campos específicos y las técnicas para aplicar estos modelos en la formulación y solución de problemas. Como un simple ejemplo, supongamos que estamos midiendo el flujo de corriente en un cable de cobre delgado. Nuestro modelo para este fenómeno podría ser la ley de Ohm:

$$Corriente = \frac{Voltaje}{Resistencia} \text{ o } I = \frac{V}{R}$$

Llamamos a este tipo de modelo un modelo mecánico porque se construye a partir de nuestro conocimiento de la física mecánica básica que relaciona estas variables. Sin embargo, si realizamos este proceso de medición más de una vez, quizás en diferentes momentos o incluso en días diferentes, la corriente observada podría diferir ligeramente debido a pequeños cambios o variaciones en factores que no están completamente controlados, como los cambios en la temperatura ambiente, fluctuaciones en la precisión del medidor, pequeñas impurezas presentes en diferentes ubicaciones en el cable y inestabilidad en la fuente de voltaje, en consecuencia, un modelo más realista de la corriente observada podría ser

$$I = \frac{V}{R} + \epsilon$$

donde ϵ es un término agregado al modelo para tener en cuenta el hecho de que los valores observados del flujo de corriente no se ajustan perfectamente al modelo mecanicista. Podemos pensar en ϵ como un término que incluye los efectos de todas las fuentes de variabilidad no modificadas que afectan a este sistema.

Ha habido otras muchas referencias a un modelo que se construirá a partir de datos experimentales. La palabra modelo se refiere a la descripción matemática de cómo se comporta la respuesta como una función de la variable o variables de entrada. Un buen modelo explica el comportamiento sistemático de los datos originales de una manera concisa. La forma específica del modelo depende del tipo de variable de diseño utilizada en el experimento.
Si un experimento contiene una única variable de diseño cualitativa establecida en diferentes niveles de tratamiento, entonces el modelo consiste en la media de tratamiento. Habrá tantas medias para el modelo como tratamientos en el experimento. Si un experimento contiene una única variable cuantitativa que cubre un rango

de valores, entonces el modelo consistirá en una ecuación que relaciona la respuesta con el predictor cuantitativo. Los experimentos que involucran predictores cualitativos generalmente se analizan mediante análisis de varianza (ANOVA). Los experimentos que involucran predictores cuantitativos generalmente se analizan por regresión. Los experimentos que combinan variables tanto cualitativas como cuantitativas se analizan mediante un modelo de regresión especial denominado modelo lineal general.
Cualquier modelo debe ir acompañado de una descripción correspondiente de los errores o discrepancias entre los valores observados y previstos. Estas cantidades están relacionadas por:

$$y_i = \hat{y}_i + \varepsilon_i$$

donde y_i representa el último valor observado de la respuesta, $\hat{}y_i$ representa el valor predicho correspondiente del modelo, y ε_i representa la diferencia entre ellos. Las ε_i suelen denominarse residuos. En general, la relación entre los datos, el modelo y el error se puede expresar como:

Dato→ Modelo + Error

Como mínimo, la declaración de error debe incluir una descripción de la forma o distribución de los residuos y una medida resumida de su variación. La cantidad de error o variación residual generalmente se informa como una desviación estándar llamada error estándar del modelo indicado con el símbolo $\hat{\sigma}_\varepsilon$ o s_ϵ. Cuando la respuesta se mide bajo varias condiciones o tratamientos diferentes, como es el caso habitual en un experimento diseñado, puede ser necesario describir la forma y el tamaño de los errores en cada condición.
La mayoría de las técnicas de análisis estadístico que usaremos para analizar experimentos diseñados exigen que los errores cumplan con algunos requisitos muy específicos. Los métodos más comunes usados en la unidad, la regresión para predictores cuantitativos y ANOVA para predictores cualitativos, requieren que la distribución de los errores sea normal y con una desviación estándar constante en todas las condiciones experimentales. No se preocupe, dichas pruebas de normalidad la realizaremos con MINITAB. Por eso será importante prestar atención en la clase y recordar los aspectos fundamentales vistos en el curso de inferencia estadística.

Como se verá, una declaración de error completa para una situación que se analizará mediante regresión o ANOVA es, "La distribución de errores es normal y la distribución de las varianzas de los errores es constante con error estándar igual a s_ϵ", donde s_ϵ es algún valor numérico. Si la distribución de errores no cumple con los requisitos de normalidad y homogeneidad de las varianzas del error, entonces los modelos obtenidos por regresión y ANOVA pueden ser incorrectos. En consecuencia, es muy importante verificar los supuestos sobre el comportamiento de la distribución de errores antes de aceptar un modelo. Cuando no se cumplen estas condiciones, es posible que se requieran métodos especiales para analizar los datos.

1.10 SELECCIÓN DE NIVELES VARIABLES

La selección de los niveles de las variables para un diseño de experimento es un problema muy serio. Muchos experimentos fallan porque los niveles de una o más variables se eligen incorrectamente.

1.10.1 Niveles Variables Cualitativos

Para las variables cualitativas, la elección de niveles no es tan crítica. Solo asegúrese de que cada nivel sea práctico y que proporcione datos válidos. Por ejemplo, en el tema del algoritmo de ordenamiento, no considere el lenguaje Python en su experimento si sabe que la ejecución del algoritmo va tener un problema inherente en el manejo de

los datos o en los requerimientos adicionales de memoria que a la final afectan negativamente la respuesta. Sin embargo, si no sabe por qué usar Python es un problema siendo a su vez un lenguaje mucho más intuitivo y fácil de usar, es posible que desee utilizarla en su experimento de todos modos. El experimento puede mostrar que, con la elección correcta de otras variables, dicho lenguaje de programación será el adecuado para el trabajo y le ahorrará dinero y tiempo de desarrollo.
A veces es posible redefinir una variable cualitativa como una variable cuantitativa. Por ejemplo, la anterior clasificación del fabricante del procesador AMD FX-6300, ARM Cortex-A53 e Intel Core i7-3930K, cambiaría si se determinara que la única diferencia entre ellos es cuantitativa, como si los procesadores solo difirieran en la frecuencia de reloj, por ejemplo, 3.5, 3.1 y 3.2 GHz. (Aunque ya sabemos que existen otros muchos aspectos de los diferencian: números de estados de pipeline, tamaño de las cache L1, L2 y L3. Si son procesadores de 32 o 64 bits, si utilizan tecnología hyperthreading, si tiene ejecución fuera de ordeno no, entre otras características). Si este fuera el caso, un experimento diseñado para resolver los efectos de la frecuencia de reloj podría predecir un mejor rendimiento. Siempre trate de redefinir una variable cualitativa para hacerla cuantitativa. Incluso si no elige o no está interesado en analizarlo o interpretarlo de esta manera, pero proporciona una mayor comprensión de cómo se comporta la variable.

1.10.2 Niveles de variables cuantitativos

La selección de niveles para variables cuantitativas puede llegar a ser bastante complicada. El tema más importante es la elección de los niveles más altos y más bajos. Estos niveles deben ser seguros, es decir, el producto obtenido en estos niveles debe ser útil o al menos el proceso o el sistema debe poder operar en estos niveles. Esto tiende a forzar la elección de los niveles para que sean más estrechos, por lo que hay menos riesgo de perder corridas o causar algún daño. Sin embargo, si los niveles se eligen demasiado juntos, es posible que no vea diferencias entre ellos y que pueda perder algo importante fuera del rango de experimentación. Por ejemplo, en el experimento que se realizó en clase, donde se definió el tamaño del arreglo N de 250 hasta 1500, tal vez el valor superior debió ser mucho mayor si se quería estudiar el efecto de todos los niveles de la jerarquía de memoria, al menos hasta la cache L3 o el efecto del acceso a la memoria principal, o tal vez la diferencia entre niveles debió ser más espaciada, ya que con esos valores cercanos no se notó mucho el efecto en la variable respuesta, por ejemplo.
Los experimentadores siempre están tratando de adivinar los niveles de seguridad más altos y más bajos para las variables, de modo que tengan una alta probabilidad de ver efectos medibles en las respuestas. Esta es a menudo una tarea difícil y angustiosa, y es muy importante incluir expertos, investigadores y gerentes que tengan conocimiento, experiencia o que sean responsables del proceso o sistema porque son los que tienen más probabilidades de ofrecer una guía valiosa.

1.11 COVARIABLES

La Figura 1.3 muestra que todas las variables en un experimento pueden clasificarse como controladas intencionalmente en los niveles deseados, en las mantenidas constantes o no controladas. Una variable cuantitativa no controlada intencionalmente que se puede medir durante el experimento se denomina covariable. Las covariables comunes son variables como la temperatura, la presión atmosférica, la humedad y el voltaje de línea. Si la covariable no tiene influencia en la respuesta, entonces no tiene ninguna consecuencia, pero en muchos casos no está claro si la covariable es importante o no. Todas las variables conocidas que no están controladas durante el experimento son covariables y deben medirse y registrarse. Luego, cuando se realiza el análisis estadístico de los datos experimentales, el efecto de estas covariables se puede eliminar de la respuesta. En general, el efecto de la covariable debe tener un efecto muy pequeño, si no, un efecto no medible, en la respuesta.

Si el efecto de una covariable se vuelve demasiado grande, puede interferir con las estimaciones de los efectos de otras variables. Las covariantes deben ser variables continuas (es decir, cuantitativas). Siempre se analizan utilizando métodos de regresión. Por esta razón, la palabra covariable también se usa para referirse a variables cuantitativas que se han modificado intencionalmente en el experimento. En conclusión, al agregar covariables se puede mejorar considerablemente la exactitud del modelo y se puede afectar significativamente los resultados del análisis final. Incluyendo una covariable en el modelo se puede reducir el error en el modelo para incrementar la potencia de las pruebas de los factores. Por ejemplo, un ingeniero desea estudiar el efecto del overcloking de un procesador sobre el tiempo de ejecución de un programa. El ingeniero ejecuta el programa con diferentes valores de velocidad de reloj para acelerar el tiempo de respuesta. pero no puede controlar la temperatura del procesador, pero si se puede monitorear. La temperatura es una covariable que se debería considerar en el modelo.

1.12 DEFINICIÓN DE DISEÑO EN DISEÑO DE EXPERIMENTOS

La palabra diseño en la frase diseño de experimentos se refiere a la forma en que las variables o factores se varían intencionalmente a lo largo de muchas ejecuciones en un experimento. Una vez que se identifican las variables experimentales y se eligen los niveles de cada variable, se puede diseñar el experimento. Por lo general, el diseño del experimento se expresa en forma de dos matrices: una matriz de variables y una matriz de diseño. Consideremos el ejemplo de la ejecución de un programa en sistema de cómputo. Supongamos que las variables a considerar son: el tipo de algoritmo, el sistema operativo y el lenguaje de programación. Para cada uno de dichos factores se ha decidido usar solo dos niveles. La siguiente matriz de variables muestra una forma posible de seleccionar niveles de variables:

Nivel	X1: Tipo de algoritmo	X2: Sistema Operativo	X3: Lenguaje de programación
–	[i,j]	Windows 10	C++
+	[j,i]	Ubuntu	C#

El propósito de esta matriz es definir claramente las variables experimentales y sus niveles. Tenga en cuenta el uso de los nombres de las variables genéricas x1, x2 y x3. Su uso permite referencias a variables sin conocer sus nombres o el contexto. (A veces, se usan las letras A, B y C en lugar de x1, x2 y x3. Algunas personas prefieren usar x1, x2, ... para indicar variables cuantitativas y A, B, ... para indicar variables cualitativas, pero no hay una convención estandarizada para asignar nombres genéricos a las variables.) Ahora, con la notación - y +, se muestra un diseño de experimento en la matriz de diseño:

Std	Run	X1	X2	X3
1	4	-	-	-
2	6	-	-	+
3	2	-	+	-
4	7	-	+	+
5	8	+	-	-
6	1	+	-	+
7	3	+	+	-
8	5	+	+	+

Este experimento tiene ocho corridas. La columna Std o estándar usa un número entero para identificar cada configuración única de x1, x2 y x3. Cada fila, llamada corrida o celda del experimento, define un conjunto diferente de condiciones para la prueba de ejecución. Por ejemplo, la ejecución número 3 debe realizarse con niveles (x1, x2, x3) = (-, +, -) o es decir, con el tipo de algoritmo [i,j], sistema operativo Ubuntu y lenguaje de programación C++. Este diseño en particular se llama diseño factorial completo 2^3 porque hay tres variables, cada una en dos niveles, y el experimento requiere $2^3 = 8$ ejecuciones.

La columna de orden estándar Std identifica el orden lógico de las ejecuciones experimentales. Las ejecuciones reales del experimento no se deben realizar en este orden debido a la posibilidad de confundir una de las variables del estudio con una variable oculta, es decir, una variable no controlada y no observada que cambia durante el experimento y puede afectar la respuesta. El orden de las ejecuciones experimentales siempre es aleatorio, como el orden de ejecución en la columna Run u orden de ejecución. La asignación aleatoria no proporciona una protección perfecta contra las variables ocultas, pero a menudo es efectiva, por lo que siempre haremos una asignación aleatoria. La asignación al azar es tan importante que se considera parte del diseño del experimento; cualquier diseño está incompleto si no se ha identificado un plan de asignación al azar adecuado.
La matriz de ejecuciones experimentales a menudo se organiza por orden estándar en las etapas de planificación y análisis del DOE, pero se organiza mejor por orden de ejecuciones aleatorias cuando se está construyendo el experimento. Esto simplifica el trabajo de la persona que realmente tiene que construir el experimento y disminuye las posibilidades de cometer un error.

1.13 TIPOS DE DISEÑOS

Hay muchos tipos diferentes de diseños de experimentos. En general, se pueden clasificar en grandes grupos con nombres extraños: factoriales, factoriales 2^n, parcelados, compuestos centrales, Box-Behnken y Plackett-Burman. También hay diseños híbridos que combinan características de dos o más de estos grupos. Pero tan complicado como suena todo esto, pero en realidad solo se utiliza un puñado de estos diseños para la mayoría de los experimentos. Nunca lo adivinarías mirando un libro de texto de DoE. Los libros están siempre llenos de todo tipo de diseños de experimentos grandes y elaborados porque son los más divertidos para que los autores hablen y escriban. La Figura 1.6 es una captura de pantalla de los diseños específicos ofrecidos en MINITAB.

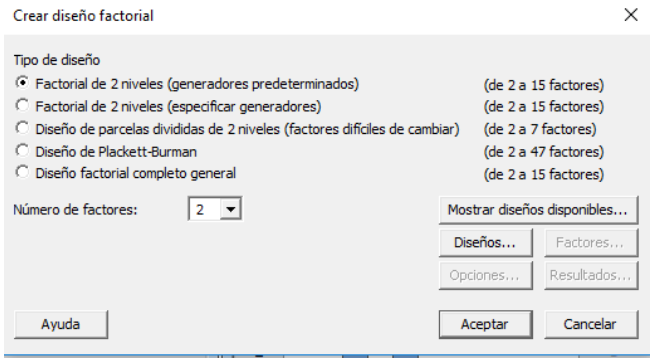


Figura 1.7: Tipos de Diseño factorial de MINITAB

1.14 ALEATORIZACIÓN

Por lo general, es imposible construir todas las ejecuciones de un experimento simultáneamente, por lo que las ejecuciones se realizan típicamente una tras otra. Dado que las condiciones experimentales no controladas pueden cambiar de una ejecución a otra, se debe considerar la influencia del orden de las ejecuciones.

Incluso un experimento simple con una variable en dos niveles sería más fácil de construir si todas sus ejecuciones se realizaran en un orden conveniente (por ejemplo, 11112222); sin embargo, tales planes de orden de ejecución corren el riesgo de atribuir erróneamente el efecto de una variable no observada que cambia durante el experimento a la variable experimental. El método aceptado para protegerse contra este riesgo es aleatorizar el orden de ejecución de los niveles de la variable experimental (por ejemplo, 21121221). Al aleatorizar, los efectos de cualquier cambio sistemático no observado en el proceso no relacionado con la variable experimental se distribuyen de manera uniforme y aleatoria en todos los niveles de la variable experimental. Aleatoriedad significa que tanto el nivel de factor asignado a un experimento en particular, como el orden en que se efectúan las pruebas se efectué de una manera aleatoria.

Supongamos, por ejemplo, que se desea saber cómo el tamaño de la memoria cache influye en el nivel el tiempo de ejecución de un programa, específicamente con el tiempo promedio de acceso a memoria (AMAT), para esto, primero se efectúa cuatro pruebas donde se mide el tiempo de ejecución (variable respuesta) de un programa para 5 valores distintos del tamaño N de un arreglo y los resultados promedio de las repeticiones para cada corrida son: (supongamos que la medida está en ns):

N (tamaño del arreglo)	100	200	300	400	500
Cache 1	6.2	2.8	2.7	3.0	2.95
Cache 2	7.4	3.9	4.0	4.1	4.0

A primera vista se nota que la primera corrida de la prueba 1 tiene una respuesta muy alta respecto a las corridas que le siguen, eso mismo ocurre con la prueba con la cache 2. Para descartar que dicho efecto no es por el orden en que se hace la prueba (primera vez n=100) y si tal vez por el tamaño del arreglo, o por procesos iniciales de asignación de recursos del sistema operativo cuando se ejecuta el programa o por otro factor no identificado, lo que se puede hacer es confundir, anular o igualar este efecto, realizando las pruebas en orden aleatorio" bajo los dos tipos de cache y un orden aleatorio de n (tamaño del arreglo). La aleatoriedad por lo tanto es importante por al menos dos razones: confunde el efecto de factores no controlables y valida las pruebas estadísticas al hacer que los errores experimentales sean estadísticamente independientes.

1.15 REPLICACIÓN Y REPETICIÓN

La matriz de diseño de un experimento determina qué términos será capaz de resolver el modelo, pero la sensibilidad del análisis a los efectos de las pequeñas variables está determinada por la cantidad de veces que se construye cada ejecución experimental. Generalmente, cuantas más veces se construyen las ejecuciones de un diseño de experimento, mayor será la sensibilidad del experimento. Hay dos formas diferentes de repetir las

ejecuciones de un diseño de experimento. Cuando se hacen unidades consecutivas sin cambiar los niveles de las variables de diseño entre unidades, estas unidades similares se llaman *repeticiones*. Cuando se producen dos o más unidades similares en un experimento, pero en diferentes momentos espaciados a lo largo del experimento y no como unidades consecutivas, estas unidades similares se denominan *réplicas*.

Al principio, podría parecer que el uso de repeticiones y réplicas daría resultados similares, si no idénticos, pero ese no suele ser el caso. De hecho, los valores de la respuesta para las repeticiones y las repeticiones serán casi idénticos si el proceso es estable, pero la replicación casi siempre conduce a una mayor variación en la respuesta debido a cambios en las variables no controladas. Es como si ejecutara un algoritmo en diferentes horas del día, y que existiera una variación considerable en las respuestas. A pesar de esta aparente desventaja de la replicación sobre la repetición, la replicación generalmente proporciona una medida más realista del ruido inherente en el proceso y es la forma preferida de aumentar el número de ejecuciones en un experimento. La diferencia en los valores asociados con las repeticiones y las repeticiones se aclara por la forma en que se tratan en los análisis estadísticos; las ejecuciones repetidas se promedian mientras que las observaciones replicadas individuales se conservan, por lo que las repeticiones hacen comparativamente poco para aumentar la sensibilidad de un experimento a pequeños efectos variables.

Es importante entender las diferencias entre las mediciones de respuesta de repetición y de réplica. Estas diferencias afectan la estructura de la hoja de trabajo y las columnas en las que usted ingresa los datos de respuesta, lo que a su vez afecta la forma en que Minitab interpretará los datos. Las repeticiones se ingresan de forma horizontal en las filas de múltiples columnas, mientras que las réplicas se ingresan de forma vertical en una sola columna.

1.16 BLOQUEO

A menudo, mientras se prepara un experimento diseñado para estudiar una o más variables, se identifica otra variable importante, una variable molesta, que no se puede mantener constante o aleatoria. Si durante el curso del experimento cambia el nivel de esta variable que tiene un efecto correspondiente en la respuesta, estos cambios en la respuesta inflarán el ruido en el experimento haciéndolo menos sensible a pequeñas, pero posiblemente importantes diferencias entre los niveles de las variables de estudio. En lugar de tolerar el ruido adicional introducido por esta variable molesta, las ejecuciones experimentales se deben construir en subconjuntos llamados bloques, donde cada bloque usa un solo nivel de la variable molesta. El método habitual para asignar ejecuciones a bloques es crear una o más réplicas completas del diseño del experimento dentro de cada bloque. Luego, cuando se realiza el análisis estadístico de los datos, la variable de bloqueo se incluye en el modelo para que las diferencias entre los niveles de la variable de molestia se tengan en cuenta. Este enfoque aísla la variación causada por la variable molesta y recupera toda la sensibilidad del diseño del experimento.

Aunque se incluye una variable de bloqueo en el análisis estadístico de un experimento, generalmente no hacemos pruebas para ver si hay diferencias entre sus niveles. Tales pruebas serían inseguras porque, dado que los niveles de la variable de bloqueo no se ejecutan normalmente en orden aleatorio, puede haber otras variables no identificadas que cambien durante el experimento y que sean la causa real de las diferencias aparentes entre los niveles. Cuando una variable de estudio no puede ser aleatoria y debe ejecutarse en bloques, debemos ser muy cuidadosos para garantizar que las condiciones experimentales sean tan constantes como sea posible y debemos estar conscientes del riesgo de que nuestras conclusiones sobre las diferencias entre bloques puedan ser erróneas. Si realmente necesitamos determinar si hay diferencias entre los niveles de alguna variable, entonces no tenemos

otra opción: sus niveles deben ejecutarse en orden aleatorio. Si no necesitamos determinar si hay diferencias entre los niveles de alguna variable, entonces podemos tratarla como una variable de bloqueo.

Imagínese que usted reserva la sala de cómputo 501L para hacer corridas sobre un computador con las características idénticas a las que tiene los computadores de la sala actualmente. Y como se supone que todas las máquinas tienes la misma configuración hardware y software, entonces usted se le ocurre dividir todas las corridas ejecutando las pruebas usando 4 de los computadores. Eso para poder terminar más rápido todas las corridas que ejecutándolas en una sola máquina. Pero luego reflexiona y piensa que el ejecutar el experimento en las 4 máquinas eso pueda generar mucho ruido, ya que sospecha que entre las máquinas puede haber diferencias. Eso sería una variable molesta la cual no está interesado en estudiar (Diferencias entre las máquinas). Entonces en los 4 computadores (que también hace las veces de unidades experimentales) las ejecuciones experimentales se deben construir en subconjuntos llamados bloques. Es decir que cada computador sería un bloque de análisis.

Otra situación de bloqueo se da cuando un experimento es tan grande que no se puede completar de una vez, se debe construir en bloques definidos, por ejemplo, por días o turnos y dependiendo de la disponibilidad del algún recurso. Luego, si hay diferencias entre los bloques, las diferencias pueden explicarse en el análisis sin disminuir la sensibilidad del diseño original.

2.0 METODOLOGÍA GENERAL PARA REALIZ*AR* UN EXPERIMENTO

Existen varias metodologías en la literatura, una de ellas es la desarrollada por Douglas C. Montgomery:

2.1 **Identifique claramente el problema o situación a resolver.** Antes de poder planear un experimento necesitamos definir claramente que es la que estamos buscando, aun cuando esto puede parecer trivial en ocasiones es tanta la presión para tomar decisiones que corremos a experimentar sin por lo menos definir claramente nuestros objetivos. Es necesario desarrollar todas las ideas sobre los objetivos del experimento. Suele ser importante solicitar la opinión de todas las partes implicadas. Un planteamiento claro del problema contribuye a menudo en forma sustancial a un mejor conocimiento del fenómeno y de la solución final del problema.
En este paso es necesario definir qué tipo de información es exactamente la que nos interesa, ya que no podemos medir o variar todos y cada uno de los componentes de un experimento.

Antes de planear un experimento se debe de investigar y analizar todos los aspectos y datos que ya se tengan sobre este problema o el sistema de cómputo en éste caso.

En conclusión, como resultado de este paso, la hipótesis a probar debe quedar bien definida.

2.2 **Identificar Factores.** En este paso dos tipos de variables se deben de identificar, factores primarios y factores o variables secundario y de ruido.

El experimentador debe elegir los factores que variarán en el experimento, los intervalos de dicha variación y los niveles específicos de interés a los cuales se hará el experimento. También debe considerarse la forma en que se controlarán estos factores para mantenerlos en los valores deseados, y cómo se les medirá. Para ello es necesario conocer el sistema de manera práctica y teórica.
La variable dependiente o variable de respuesta es la característica de desempeño que queremos mejorar y cuyo comportamiento deseamos conocer, la métrica a usar puede ser el performance o el tiempo de ejecución de un programa, o la configuración optima de un sistema computacional.

Los factores primarios representan aquellas causas o factores cuyo efecto sobre la variable respuesta se quiere analizar. Cada uno de estos factores se deberá probar al menos a dos valores diferentes para evaluar su efecto, a cada uno de estos valores o niveles se les llama **tratamientos**.

Al seleccionar la respuesta, el experimentador debe estar seguro de que la respuesta que se va a medir realmente provea información útil acerca del proceso de estudio. Con mayor frecuencia, el promedio o la desviación estándar (o ambos) de la característica medida serán la variable de respuesta. No son raras las respuestas múltiples. La capacidad de medición (o el análisis de errores y precisión y rango de las variables de medición) también es un factor importante. Si la capacidad de medición es deficiente, sólo puede esperarse que el experimento detecte efectos relativamente grandes de los factores (por ejemplo, medir en nano segundos una respuesta cuya media está en una decena de ns); en caso contrario deben hacerse repeticiones.

2.3 Experimentación preliminar

A menudo, la única forma de llenar los vacíos de conocimiento identificados en la declaración del problema del DOE es realizar algunos experimentos preliminares, ya sea en el laboratorio o en el proceso real a estudiar. La experimentación preliminar exitosa es fundamental para disminuir los riesgos de un experimento, especialmente cuando se deben considerar muchos factores. Estos experimentos preliminares usualmente toman la forma de pequeños conjuntos de ejecuciones para investigar una variable o procedimiento a la vez.
El propósito de la experimentación preliminar es:

- Adquirir experiencia con nuevas variables experimentales.
- Confirmar que no hay variables no identificadas.
- Identificar los límites superior e inferior seguros para las variables experimentales.
- Confirmar que los procedimientos utilizados para operar las variables son precisos.
- Confirme que los operadores y el equipo funcionan correctamente y como se espera.
- Estimar la desviación estándar de la respuesta para que se pueda realizar un cálculo del tamaño de la muestra.

2.3. Elegir el diseño experimental. Esto implica definir de qué manera se efectuarán las pruebas y que modelo que describe mejor el experimento. En el resto de este documento se describen varios tipos de experimentos de los cuales se debe tomar el que mejor se ajuste a la situación particular.

Para elegir el diseño es necesario considerar el tamaño muestral (número de repeticiones), seleccionar un orden adecuado para los ensayos experimentales, y determinar si hay implicado bloqueo u otras restricciones de aleatorización.

Es importante tener presente los objetivos experimentales al seleccionar el diseño, se tiene interés en identificar qué factores causan diferencias en estimar la magnitud del cambio de la respuesta. En otras situaciones habrá más interés en verificar la uniformidad. Por ejemplo, pueden compararse dos configuraciones de sistema de cómputo A y B, siendo *A* la estándar y B una alternativa de menor costo. El experimentador estará interesado en demostrar que no hay diferencia en cuanto a la relación costo desempeño (por ejemplo), entre las dos condiciones, o la razón tiempo promedio de ejecución y periodo de reloj del procesador.

2.4. Efectuar el experimento. Esto de acuerdo a lo que se defina en el paso 2.3.

Cuando se realiza el experimento, es vital vigilar el sistema cuidadosamente para asegurar que todo se haga conforme a lo planeado. En esta fase, los errores en el procedimiento suelen anular la validez experimental. La planeación integral es decisiva para el proceso. En un complejo entorno de investigación y desarrollo, es fácil subestimar los aspectos logísticos y de planeación de la realización de un experimento diseñado.

2.5. Análisis de los datos. Estos son básicamente análisis estadísticos similares a los vistos en el curso de inferencia estadística, aunque en DoE pueden tratarse más de 2 factores.
Deben emplearse métodos estadísticos para analizar los datos, de modo que los resultados y conclusiones sean objetivos más que apreciativos. Existen muchos excelentes paquetes de software para el análisis de datos como Matlab o Minitab, y varios métodos gráficos sencillos son importantes en la interpretación de tales datos (diagrama de cajas). El análisis de residuos y la verificación de la idoneidad del modelo son también técnicas de análisis de gran utilidad.

Hay que recordar que los métodos estadísticos sólo proporcionan directrices para la veracidad y validez de los resultados. Los métodos estadísticos, aplicados adecuadamente, no permiten probar algo experimentalmente, sólo hacen posible obtener el probable error de una conclusión, o asignar un nivel de confiabilidad a los resultados. La principal ventaja de los métodos estadísticos es que agregan objetividad al proceso de toma de decisiones. Las técnicas estadísticas, articuladas con un buen conocimiento técnico del sistema y al sentido común, suelen llevar a conclusiones razonables.

2.6. Conclusiones y toma de decisiones.

Una vez que se han analizado los datos, él experimentador debe extraer conclusiones prácticas de los resultados y recomendar un curso de acción. En esta fase a menudo son útiles los métodos gráficos, en especial al presentar los resultados a otras personas. También deben realizarse corridas de seguimiento y pruebas de confirmación para validar las conclusiones del experimento.

3.0 EXPERIMENTOS PARA LAS CLASIFICACIONES DE UNA VÍA

3.1 INTRODUCCION

Cuando aparece una variable en dos niveles en un experimento (por ejemplo, el fabricante de procesadores A versus B), la prueba t de dos muestras se puede usar para probar una diferencia entre las medias de los dos niveles. Cuando un experimento similar tenga tres o más niveles (por ejemplo, el fabricante A, B, C, etc.), podría tener la tentación de usar la prueba t de dos muestras para probar todos los pares de niveles posibles (A versus B, A versus C, B versus C, etc.) pero es arriesgado y se prefiere un mejor método de análisis, llamado análisis de varianza (ANOVA).

Las pruebas t son ciertamente factibles, pero existe un grave riesgo asociado con este enfoque. Si el nivel de significación para cada prueba t de dos muestras es α, entonces la probabilidad de no cometer un error de Tipo 1 en una prueba es 1 - α. Como tenemos que hacer $\binom{k}{2}$ pruebas, la probabilidad de no cometer ningún error de Tipo 1 viene dada por:

$$1 - \alpha_{total} = (1 - \alpha)^{\binom{k}{2}}$$

donde α_{total} es la probabilidad de cometer un error de Tipo 1 en al menos una de las pruebas. α_{total} puede llegar a ser muy grande, especialmente cuando hay muchas pruebas para realizar, por lo que se debe reconsiderar el uso de múltiples pruebas t. Se requiere entonces un método alternativo, uno que reemplace las múltiples pruebas t de dos muestras por una prueba única o que reduzca la tasa de error general a un valor razonable. Para que se haga una idea, suponga que se realizará un experimento con cinco tratamientos para probar para diferencias entre sus medias, y se maneja un $\alpha = 0.05$. Entonces la probabilidad general de error de Tipo 1 para comparaciones múltiples para cada prueba t de dos muestras sería:

$$1 - \alpha_{total} = (1 - 0.05)^{\binom{5}{2}} = 1 - 0.40$$

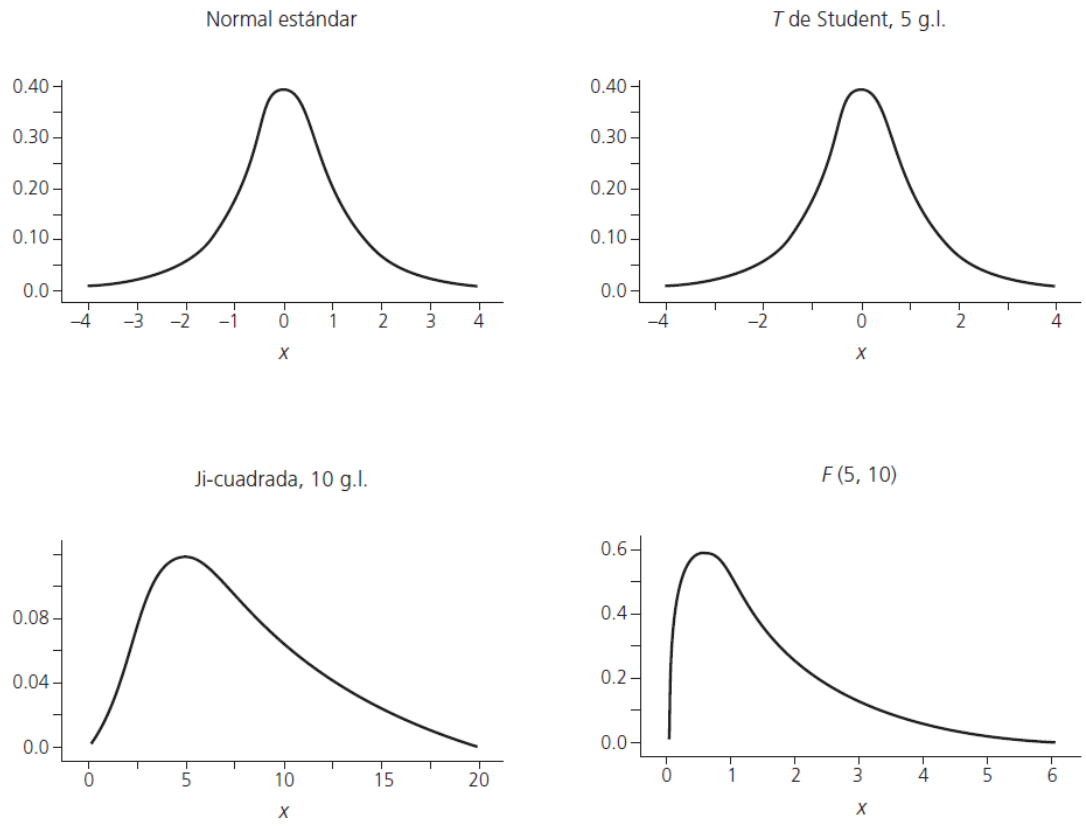
o α_{total}= 0,40. α_{total} se incrementa seriamente al hacer tantas comparaciones múltiples en relación con el seleccionado para una única prueba t. Este nivel de riesgo es totalmente inaceptable.

ANOVA emplea una única prueba estadística para comparar simultáneamente todos los pares de medias posibles para ver si hay diferencias entre ellos. Puede parecer extraño que la palabra varianza aparezca en una prueba para medias, pero la técnica realmente usa la prueba de F para dos variaciones para probar las diferencias entre medias de tres o más de tratamientos de una manera inteligente y segura.
Un experimento para probar las diferencias entre las medias de varios niveles diferentes de una sola variable implica una clasificación de una vía y analizamos los datos de clasificación de una vía utilizando ANOVA de una vía, que es el tema de este capítulo. Como es de esperar, los experimentos más complejos que involucran dos o más variables de clasificación se denominan diseños de clasificación bidireccionales y multidireccionales y se analizan utilizando ANOVA de dos vías y multidireccional.

3.2 DISTRIBUCIONES DE PROBABILIDAD E INFERENCIA

La distribución de probabilidad o distribución de una variable aleatoria X relaciona el conjunto de valores posibles de X, con la probabilidad asociada a cada uno de estos valores y los representa a través de una tabla o por medio de una función planteada como una fórmula.

Las distribuciones de probabilidad que más se usan en intervalos de confianza y pruebas de hipótesis son las distribuciones; normal, T Student, ji-cuadrada y F. En la figura se representan las formas típicas de estas cuatro distribuciones. La distribución normal está completamente definida por sus parámetros, que son, μ, y la desviación estándar, σ. Por ejemplo, en la figura se muestra la distribución normal estándar es decir con μ=0 y σ=1.



Muestra de las distribuciones de probabilidad de mayor uso en inferencia

Se puede observar que la distribución normal estándar como la T de Student son simétricas y centrales en cero, mientras que la ji-cuadrada y F son sesgadas y solo toman valores positivos. Lo interesante es que las 4 distribuciones están relacionadas entre sí, ya que las distribuciones T de Student, ji-cuadrada y F se definen en términos de la distribución normal estándar.

Los parámetros que definen por completo las distribuciones T de Student, ji –cuadrada y F, reciben el nombre de grados de libertad, que tienen que ver con los tamaños muestrales involucrados. Por ejemplo, si se tiene una muestra de tamaño 20, será de interés una distribución T de Student con 19 grados de libertad para hacer inferencias sobre la varianza poblacional. Dicha distribución tiene a la distribución normal estándar cuando el tamaño de muestra crece y prácticamente es la misma para n>45. La distribución estándar es una curva única, por ello existen tablas que proporcionan cualquier área o probabilidad de interés. No pasa lo mismo con las otras distribuciones, ya que para cada tamaño muestral es una curva diferente. Recordemos que las distribuciones normal y T student sirven para hacer inferencia sobre las medias, la distribución ji-cuadrada será de utilidad para hacer inferencias sobre varianzas y la distribución F se emplea para comparar varianzas, por eso es que la distribución F es la de mayor relevancia en diseño de experimentos, dado que el análisis de variabilidad que se observa en un experimento se hace comparando varianzas.

3.3 EL ENFOQUE GRÁFICO DE ANOVA.

Suponga que desea determinar el tiempo promedio de ejecución del siguiente programa escrito en lenguaje C, es decir estamos interesados en analizar varias muestras del tiempo que dura la iteración interna del ciclo *for*, esto se determina dividiendo el tiempo total entre N^2. En éste caso el tamaño del arreglo es N=500.

```
1   typedef int array_t[N][N];
2
3   int sumA(array_t a)
4   {
5       int i, j;
6       int sum = 0;
7       for (i = 0; i < N; i++)
8           for (j = 0; j < N; j++) {
9               sum += a[i][j];
10          }
11      return sum;
12  }
13
```

Considere los datos de clasificación de una vía de k = 8 tratamientos, donde cada una de las 8 muestras del tiempo de ejecución del programa (k = 8) tiene un tamaño n = 20. Las muestras fueron tomadas en la misma unidad experimental, un mismo computador donde se ejecutó el programa en diferentes momentos. En la Tabla 2.1 se muestran los datos obtenidos y sus diagramas de caja correspondientes en la Figura 2.1. Si elegimos utilizar múltiples pruebas *t* de dos muestras para verificar las diferencias entre las medias de los tratamientos, habría $\binom{8}{2}$ = 28 pruebas para realizar. Con α = 0.05 para pruebas individuales, la tasa de error general de Tipo 1 es total = 1 - 0.95^{28} o aproximadamente el 76 por ciento, lo que obviamente es inaceptable.

Tabla 2.1 Datos muestra de 8 tratamientos de tamaño 20. Respuesta en ns

	Tratamientos (tiempo de ejecución [ns])							
Observación	**t1**	**t2**	**t3**	**t4**	**t5**	**t6**	**t7**	**t8**
1	10,30	11,37	10,16	10,46	11,52	8,99	8,72	11,51
2	12,03	10,88	9,67	9,67	12,56	12,07	10,31	9,71
3	9,94	9,07	9,34	11,29	10,88	12,50	11,92	10,36
4	11,32	11,45	10,34	11,80	11,00	10,87	9,05	11,07
5	12,19	12,42	10,57	11,50	10,63	11,60	11,84	9,63
6	12,67	11,31	10,76	10,68	11,24	10,60	10,39	10,36
7	9,95	11,65	8,73	10,26	10,25	10,11	11,38	10,27
8	11,88	9,49	10,03	9,57	10,80	11,99	11,83	9,95
9	8,19	10,06	9,78	11,23	9,10	10,80	9,88	8,54
10	10,64	10,36	10,09	11,44	9,32	9,87	10,14	9,71
11	10,83	10,73	11,14	11,17	8,94	10,25	10,46	10,00
12	9,57	8,76	8,43	10,60	10,46	8,67	10,26	11,31
13	11,80	11,89	11,10	10,57	9,47	10,89	10,39	10,74
14	8,48	12,63	9,91	10,51	10,32	9,90	11,92	12,16
15	12,47	10,98	9,90	10,66	12,78	10,33	9,52	10,87
16	12,15	11,53	8,68	10,33	10,52	10,99	10,69	10,96
17	10,30	8,93	10,71	9,92	8,91	12,15	11,09	10,76
18	11,49	11,53	9,19	8,65	10,82	10,69	11,97	9,14
19	10,17	11,59	10,64	11,28	9,01	10,66	12,04	10,32
20	7,97	10,59	11,35	8,26	12,21	8,03	10,44	10,28

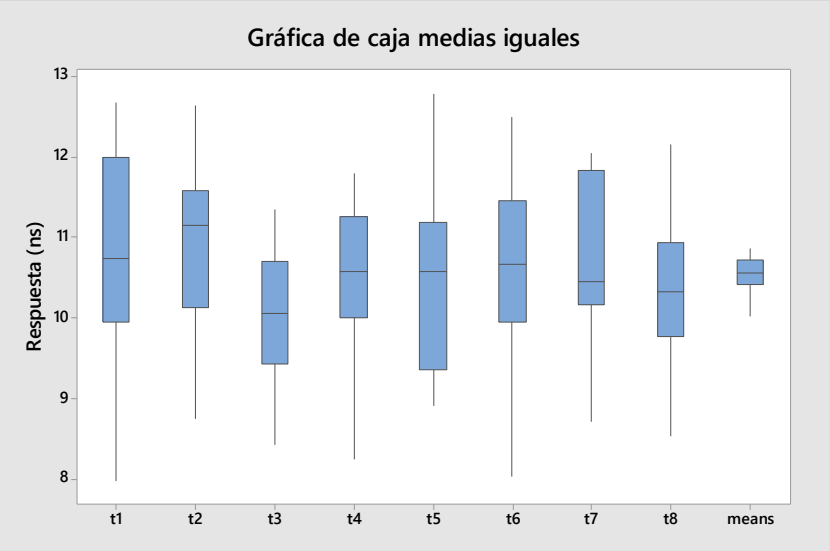


Figura 2.1 Diagrama de cajas de tratamientos con todas las medias iguales

Ahora inspeccione el último diagrama de caja en la Figura 2.1. Este diagrama de caja se construyó a partir de las medias de tratamiento k = 8. Una de las características de los diagramas de caja es que tienden a agrandarse con el tamaño de la muestra, pero en este caso, cada uno de los nueve diagramas de caja en la figura se construye a partir de las ocho observaciones. El diagrama de caja de las 8 medias muestrales es obviamente mucho más pequeño que los otros ocho diagramas de caja, pero esto es solo una consecuencia del teorema del límite central*, que dice que la distribución de las medias muestrales para muestras extraídas de una sola población se contrae por un factor de A la distribución de esa población. Es decir:

$$\sigma_{\overline{y}} = \frac{\sigma_y}{\sqrt{n}} \quad (3.1)$$

Entonces, como se esperaba, el diagrama de caja de los medias de muestra en la Figura 2.1 parece ser más pequeño que la mayoría de los otros diagramas de caja en aproximadamente un factor de $\sqrt{8}$. Ahora, considere los diagramas de caja en la Figura 2.2. De nuevo, se muestrean k = 8 tratamientos diferentes, todos de tamaño n = 20 y con aproximadamente la misma cantidad de variación que las gráficas de caja en la Figura 2.1, pero estas gráficas de caja claramente tienen medias que son diferentes entre sí. Observe el efecto de las diferentes medias en el último diagrama de caja, que nuevamente muestra la distribución de las medias de muestra; está inflado a un tamaño mucho mayor que el de cualquiera de los diagramas de caja individuales, lo que contradice la predicción del teorema del límite central. Esta contradicción, que se basa en el tamaño relativo del diagrama de caja de las medias muestrales, confirma nuestra observación de que existen diferencias significativas entre las medias de los tratamientos.

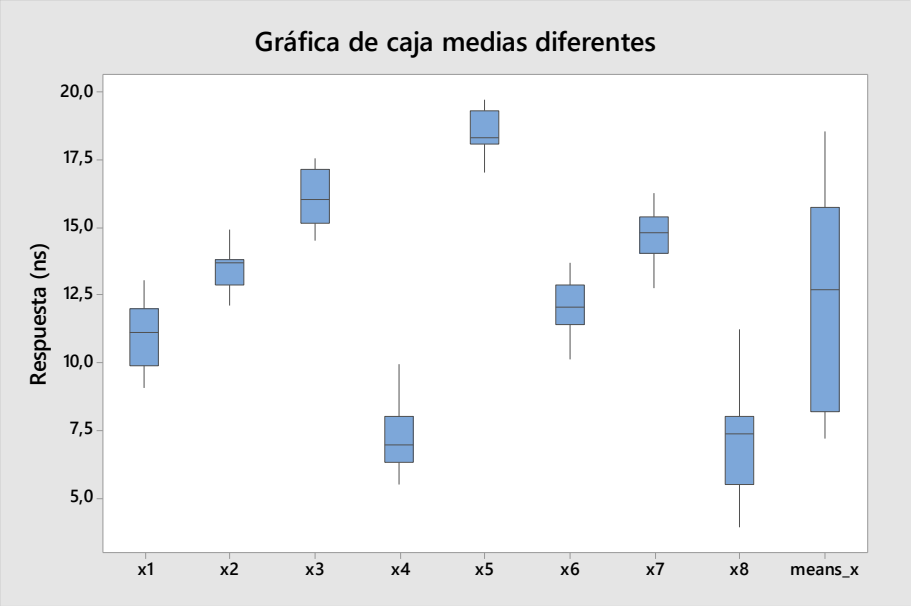


Figura 2.2 Diagrama de cajas de tratamientos con medias diferentes

Las figuras 2.1 y 2.2 sugieren un nuevo método gráfico para verificar las diferencias entre pares de medias para muchos tratamientos diferentes. Si la distribución de las medias muestrales se contrae por el factor esperado según lo predicho por el teorema del límite central, entonces probablemente no haya diferencias significativas entre las medias poblacionales. Sin embargo, si la distribución de medias muestrales no se contrae con la cantidad esperada, entonces una o más de las poblaciones que se muestrean deben tener una media que sea significativamente diferente de las demás. Esta prueba gráfica es la base del método ANOVA.

3.4 INTRODUCCIÓN A ANOVA

3.3.1 La justificación de ANOVA

La falla del método de comparaciones múltiples para probar las diferencias entre k medias de tratamiento se debió a la aplicación de $\binom{k}{2}$ pruebas separadas. Se requiere una prueba única que compare todas las medias simultáneamente para evitar este problema. Tal análisis es proporcionado por el análisis de varianza (ANOVA).

Para entender cómo funciona ANOVA, considere la siguiente situación. Supongamos que k muestras aleatorias, todas de tamaño n, se extraen de la *misma* población normal, de modo que se garantiza que las muestras tengan la misma media y varianza de la población. Sea x = {1, 2,. . . , k} para distinguir las diferentes muestras y ***y*** indica la respuesta medida. Si la varianza de la población es $\sigma_y{}^2$, entonces el teorema del límite central describe la distribución de las medias muestrales, por lo que la varianza de la distribución de la muestra $\overline{ys}$ es:

$$\sigma_{\overline{y}}{}^2 = \frac{\sigma_y{}^2}{n} \qquad (3.2)$$

Por supuesto, $\sigma_{\overline{y}}{}^2$ y $\sigma_y{}^2$ son generalmente desconocidas, pero podemos estimarlos a partir de los datos de muestra. Se puede hacer una estimación de $\sigma_y{}^2$ y a partir de la varianza muestral de cualquiera de las k muestras, pero se puede hacer una mejor estimación promediando todas las k varianzas muestrales.

Esta técnica, que combina las varianzas de diferentes tratamientos, se denomina agrupación, y decimos que las varianzas de las *k* muestras se agrupan para estimar la varianza de la población común. La varianza muestral agrupada también se denomina varianza de error porque mide la variabilidad de error aleatorio dentro de las muestras. Si la varianza de la muestra i-ésima ${S_i}^2$ está dada por:

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2 = \frac{1}{n-1}\sum_{j=1}^{n}\varepsilon_{ij}^2$$

entonces la varianza agrupada o de error es:

$$s_\varepsilon^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2$$

La cantidad $\varepsilon_{ij} = y_{ij} - \overline{y}_i$ es la discrepancia o error entre una observación y su media del tratamiento. Dado que ${S_i}^2$ mide la cantidad de variación de error dentro del tratamiento i-ésimo, entonces ${S_\varepsilon}^2$ mide la cantidad de variación que ocurre dentro de los tratamientos, basándose en la información de todos los tratamientos. Tenga en cuenta que ${S_\varepsilon}^2$ es una estimación para ${\sigma_y}^2$. Es esta cantidad la que determina el tamaño nominal de los diagramas de caja en las Figuras 2.1 y 2.2.

Una segunda estimación de ${\sigma_y}^2$ se puede determinar a partir de la varianza de la distribución de las medias de *k* muestras. Si la media de la muestra i es $\overline{y}_i$, entonces la gran media es:

$$\overline{\overline{y}} = \frac{1}{k}\sum_{i=1}^{k}\overline{y}_i$$

y la varianza de las $\overline{y}_i s$ está dada por:

$$s_{\overline{y}}^2 = \frac{1}{(k-1)}\sum_{i=1}^{k}(\overline{y}_i - \overline{\overline{y}})^2 \quad (3.6)$$

Esta es la cantidad que determina el tamaño de los diagramas de caja de las medias de muestra en las Figuras 2.1 y 2.2. Combinada con la ecuación 3.2, la ecuación 3.6 nos da una segunda estimación de ${\sigma_y}^2$ dada por ${\sigma_y}^2 \cong n{s_{\overline{y}}}^2$, determinada por la variación que ocurre entre los niveles de tratamiento.

¡Ahora por la magia! Tenemos dos estimaciones para ${\sigma_y}^2$, una dada por la variación *dentro* de las muestras:

$${\hat{\sigma}_y}^2 = {S_\varepsilon}^2$$

y otro dado por la variación *entre* muestras:

$${\hat{\sigma}_y}^2 = n{s_{\overline{y}}}^2$$

La relación de las dos estimaciones de ${\sigma_y}^2$:

$$F = \frac{\hat{\sigma}_{\bar{y}}^2}{\hat{\sigma}_{y}^2} = \frac{n s_{\bar{y}}^2}{s_{\varepsilon}^2}$$

sigue una distribución F con el numerador k - 1 y los grados de libertad del denominador k (n - 1). Hay k - 1 grados de libertad para el numerador porque se utilizan k - 1 medias muestrales para calcular $\sigma_y{}^2$. Hay k (n - 1) grados de libertad para el denominador porque k varianzas muestrales, cada una con n - 1 grados de libertad, se utilizan para calcular la varianza de error $S_\varepsilon{}^2$. Si se cumplen todas las condiciones necesarias, las que hacen funcionar el teorema del límite central y la afirmación de que todas las k muestras tienen la misma media poblacional, entonces esperamos que esta relación de F sea F = 1. Si relajamos la condición de que todas las muestras provienen de la misma población, entonces si una o más de las medias poblacionales son diferentes de las otras, $s_{\overline{y}}{}^2$ se inflará y nuestra proporción F será mayor que F = 1. En la práctica, si $F \cong 1$, entonces aceptamos la afirmación de que todos los medios de tratamiento son iguales (o el juicio se reserva) y si F >> 1 rechazamos la afirmación de que todos los medios de tratamiento son iguales.

Ejemplo 2.1

Las medias y las variaciones de las k = 8 muestras de la Figura 2.1 se muestran en la siguiente tabla. Utilice estas estadísticas para realizar el ANOVA para probar la afirmación de que las nueve muestras provienen de poblaciones con la misma media.

i	1	2	3	4	5	6	7	8
$\boldsymbol{n_i}$	20	20	20	20	20	20	20	20
$\boldsymbol{\overline{y}_i}$	10,72	10,86	10,03	10,49	10,54	10,60	10,71	10,38
$\boldsymbol{S_i^2}$	2,02	1,25	0,71	0,86	1,38	1,36	1,01	0,71

Solución:

La varianza del error viene dada por:

$$s_\varepsilon^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2$$

$$s_\varepsilon^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2 = \frac{1}{8}\sum_{i=1}^{8} s_i^2 = \frac{1}{8}(2.02 + 1{,}25 + 0.71 + \cdots + 0.71)$$

$$s_\varepsilon^2 = \frac{1}{8}(9.29) = 1.16$$

La gran media es:

$$\bar{\bar{y}} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{1}{8}(84.33) = 10.54$$

y la varianza de las medias de la muestra viene dada por:

$$s_{\bar{y}}^2 = \frac{1}{(8-1)}\sum_{i=1}^{8}(\bar{y}_i - \bar{\bar{y}})^2$$

$$\frac{1}{7}((10.72 - 10.54)^2 + (10.86 - 10.54)^2 + \cdots + (10.38 - 10.54)^2)$$

$$s_{\bar{y}}^2 = \frac{1}{7}(0.46) = 0.065$$

La relación F para el ANOVA es:

$$F = \frac{n s_{\bar{y}}^2}{s_\varepsilon^2} = \frac{20x0.065}{1.16} = 1.126$$

La siguiente tabla es la arrojada por Minitab

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Factor	7	9,153	1,308	1,13	0,350
Error	152	176,557	1,162		
Total	159	185,710			

F = 1.126 está muy cerca del valor F = 1 esperado, lo que sugiere que no hay diferencias entre ninguna de las medias de k = 8, pero para ser rigurosos necesitamos encontrar el valor crítico de F que determina el límite de aceptación / rechazo Para las hipótesis.

La distribución F tiene $df_{numerator}$ = k - 1 = 8 - 1 = 7 grados de libertad y $df_{denominator}$ = k (n - 1) = 8 (8 - 1) = 56 denominadores de libertad. A un nivel de significación de α = 0.05 tenemos $F_{0.05,7,56}$ = 2,178[1]. Dado que (F = 1.126) <($F_{0.05}$ = 2.178) debemos concluir que no hay diferencias significativas entre las medias de las nueve poblaciones.

El **valor p** es una probabilidad que mide la evidencia en contra de la hipótesis nula. Las probabilidades más bajas proporcionan una evidencia más fuerte en contra de la hipótesis nula. En resumen:

Valor p ≤ α: Las diferencias entre algunas de las medias son estadísticamente significativas
Valor p > α: Las diferencias entre las medias no son estadísticamente significativas

Si el valor p es menor

[1] =DISTR.F.INV(0,05;7;56) en Excell

que o igual al nivel de significancia, usted rechaza la hipótesis nula y concluye que no todas las medias de población son iguales. Utilice su conocimiento especializado para determinar si las diferencias son significativas desde el punto de vista práctico.

Si el valor p es mayor que el nivel de significancia, puede pasar que usted no cuenta con suficiente evidencia para rechazar la hipótesis de que las medias de población son todas iguales, eso depende de la experticia y conocimiento del proceso, pero se intenta usar una muestra más grande, o mejorar el proceso de obtención de las mediciones o mejorar las condiciones durante el experimento. O al usar un nivel de significancia más alto, aumenta la probabilidad de que usted rechace la hipótesis nula. Sin embargo, sea cauteloso porque no conviene rechazar una hipótesis nula que en realidad sea verdadera. En nuestro ejemplo, como los tratamientos se realizaron en condiciones muy similares en el mismo computador, podemos entonces esperar que las diferencias entre las 8 medias no son estadísticamente significativas, al tratarse de la misma población.

Recuerde las condiciones requeridas para validar el uso del método ANOVA son:

1. Las poblaciones que se muestrean están normalmente distribuidas.
2. Las poblaciones que se muestrean presentan homogeneidad de varianzas.
3. Las observaciones son independientes.

Qué hacer con datos no normales

Tiene varias opciones si desea realizar una prueba de hipótesis con datos no normales.

1. Continuar con el análisis si la muestra es lo suficientemente grande
2. Usar una prueba no paramétrica
3. Transformar los datos

3.4.1 La tabla de ANOVA

Hay una varianza de muestra adicional que se puede calcular a partir de los datos. La varianza total en el conjunto de datos puede calcularse considerando todas las desviaciones de las unidades kn de la gran media $\bar{\bar{y}}$:

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

Existe una relación muy importante entre las tres variaciones: s^2_{total}, s^2_{ε} y $s^2_{\bar{y}}$. La varianza total s^2_{total} se divide realmente en dos partes. La varianza total dentro de las muestras que se mide por s^2_{ε} y la varianza entre las muestras que se mide por $s^2_{\bar{y}}$. Ésta es la clave de ANOVA, la partición de la varianza total dentro y entre los componentes es de importancia fundamental.

Las diferentes variaciones y sus grados de libertad asociados se resumen en una tabla ANOVA:

Fuente	***df***	$\widehat{\sigma}_y^{\,2}$	***F***
Tratamiento	*K-1*	$ns^2_{\bar{y}}$	$\frac{ns_{\bar{y}}^2}{s^2_{\varepsilon}}$
Error	*K(n-1)*	s^2_{ε}	
Total	*Kn-1*	s^2_{total}	

3.4.2 Test de Rango Múltiple de Duncan

Una prueba post-ANOVA conocida por su sensibilidad a pequeñas diferencias entre las medias de tratamiento es la prueba de rango múltiple de Duncan. Esta prueba requiere que los tamaños de muestra para los tratamientos sean iguales o al menos aproximadamente iguales. Potencialmente prueba todos los pares posibles de medios de tratamiento; sin embargo, las pruebas individuales se realizan en una secuencia que ayuda a protegerse contra los errores de Tipo 1.

3.4.3 Test de comparaciones múltiples de Tukey

La prueba de rango múltiple de Duncan es poderosa pero difícil de realizar porque hay muchos conjuntos de medios de tratamiento a considerar y diferentes valores críticos en cada paso del análisis. Un compromiso popular con la prueba de Duncan es la prueba HSD (Honestly Significantly Different) de Tukey (también llamada prueba de Tukey-Kramer o prueba de comparaciones múltiples de Tukey). Al igual que las otras pruebas descritas, la prueba HSD de Tukey considera todos los posibles pares de medios de tratamiento. Aunque la prueba de Tukey es menos poderosa que la de Duncan (es decir, menos sensible a las pequeñas diferencias entre los medios de tratamiento), implica menos cálculos, es más fácil de informar y es bastante popular.

3.5 ANOVA CON MINITAB

MINITAB puede aceptar datos ANOVA de una vía en dos formatos diferentes. El primer formato, requiere que todos los valores de respuesta se apilen en una sola columna con una columna asociada que distinga los diferentes tratamientos. La columna de identificador de tratamiento puede contener valores numéricos, de texto o de fecha / hora; sin embargo, hay algunos diagnósticos de ANOVA que no aceptan datos de texto, por lo que se prefieren los valores numéricos o de fecha / hora. El segundo formato requiere que cada tratamiento aparezca en su propia columna. El formato apilado se prefiere al formato no apilado porque se puede generalizar fácilmente a los diseños de clasificación de múltiples vías y permite que las observaciones se ingresen en la Hoja de trabajo en su orden de ejecución aleatoria. El último punto es importante porque le permite a MINITAB mostrar una gráfica de diagnóstico adicional de los residuos frente al orden de ejecución que se puede usar para evaluar el supuesto de independencia.
Para mostrarle el manejo de la herramienta, retomaremos un ejemplo de un experimento donde se ejecutó dos versiones de un algoritmo en 3 computadores del laboratorio, donde además se varió el tamaño del arreglo de datos el cuál el algoritmo calcula la suma total de todos los elementos del arreglo. Por tal razón existen 3 factores (Computador, algoritmo, tamaño), de los cuales analizaremos solo el factor computador. La variable respuesta en el tiempo de ejecución del ciclo interno medido en ns. Los datos los puede copiar desde Excel y pegarlos en la hoja de trabajo de Minitab, en resumen, haga los siguiente:

1. Abra los datos de muestra, DatosLabSem8.xls. y copie los datos en Minitab
2. Elija **Estadísticas** > **ANOVA** > **Un solo factor**.
3. Seleccione **Los datos de respuesta están en una columna para todos los niveles de factores**.
4. En **Respuesta**, ingrese *Tiempo*.

5. En **Factor**, ingrese *Computador*.

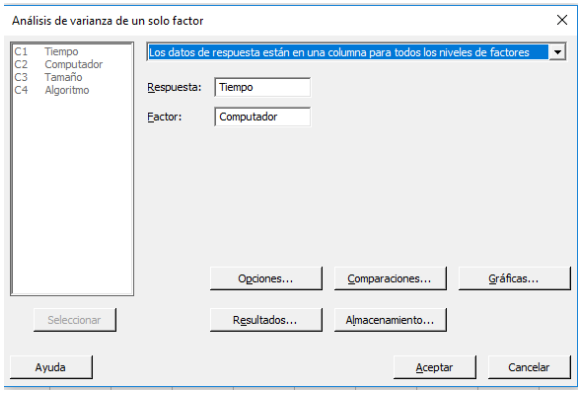
6. Haga clic en el botón **Comparaciones** y luego seleccione **Tukey y Fisher**.

7. Haga clic en **Aceptar** en cada cuadro de diálogo.

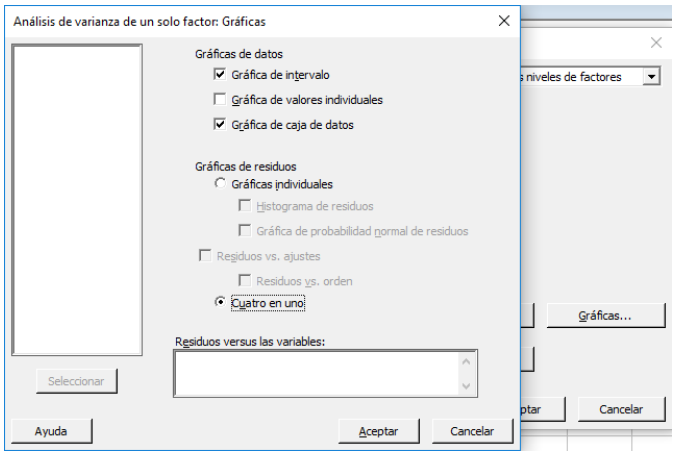
Hoja de trabajo 1 ***

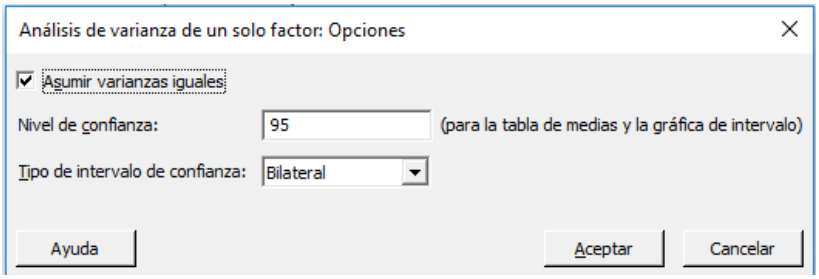
	C1	C2	C3	C4-T
	Tiempo	Computador	Tamaño	Algoritmo
1	11,056	2	500	I-J
2	10,432	2	500	I-J
3	10,117	2	500	I-J
4	10,154	2	500	I-J
5	10,906	2	500	I-J
6	10,030	2	500	I-J
7	12,093	2	500	I-J
8	14,866	2	500	I-J

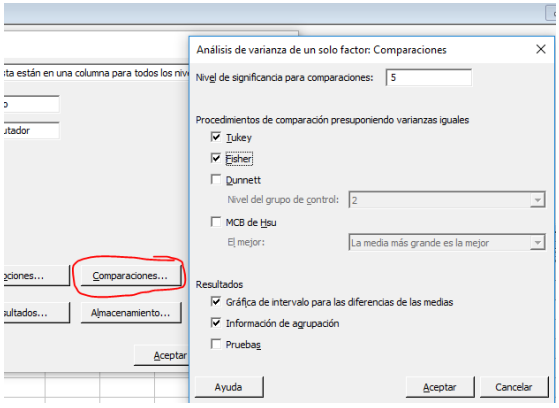
Configure MINITAB para ANOVA de una vía desde el menú Estadísticas> ANOVA> Un solo factor… y escoja según el formato de sus datos. Para los datos apilados, deberá especificar la respuesta experimental en el cuadro Respuesta: Tiempo y la columna de identificador de tratamiento en el cuadro Factor: Computador. En el caso sin apilar, deberá especificar las columnas que contienen las respuestas experimentales en el cuadro de entrada Respuestas (en columnas separadas). En nuestro ejemplo sería la opción *los datos de respuesta están en una sola columna para todos los niveles del factor*



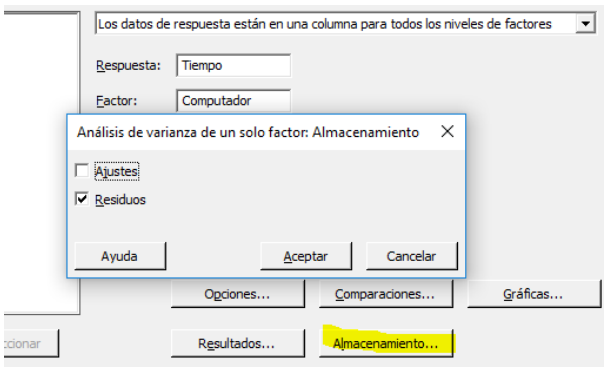
Use el menú de Gráficas para seleccionar los diagnósticos gráficos de ANOVA: Diagramas de caja de datos, Gráfica de intervalos y cuatro en uno (que incluye Histograma de residuos, Gráfico de residuos normal, Residuos contra ajustes, Residuos Vs. orden).



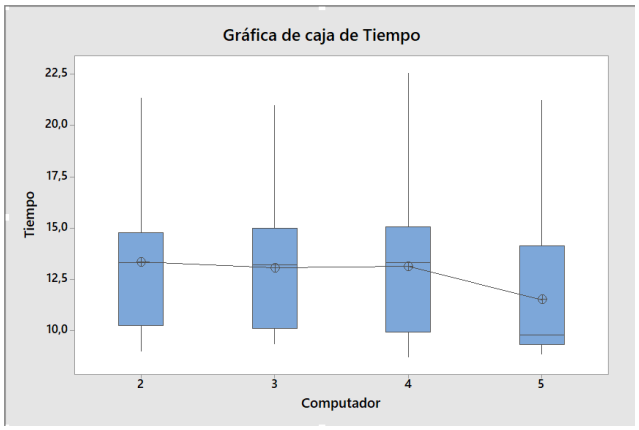


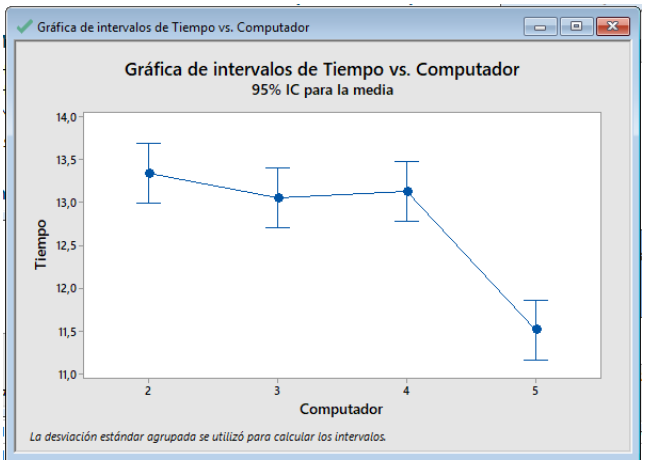


Si sospecha (como es el caso) que hay otra estructura en los residuos que no será evidente en los gráficos de diagnóstico predeterminados, active Almacenar residuos para que pueda hacer su propio análisis de seguimiento después de ejecutar el ANOVA.



https://www.youtube.com/watch?v=w62VPUXxu4Q





Hipótesis nula	Todas las medias son iguales
Hipótesis alterna	No todas las medias son iguales
Nivel de significancia	α = 0,05

Se presupuso igualdad de varianzas para el análisis.

Información del factor

Factor	Niveles	Valores
Computador	4	2; 3; 4; 5

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Computador	3	840,9	280,29	22,23	0,000
Error	1596	20125,5	12,61		
Total	1599	20966,3			

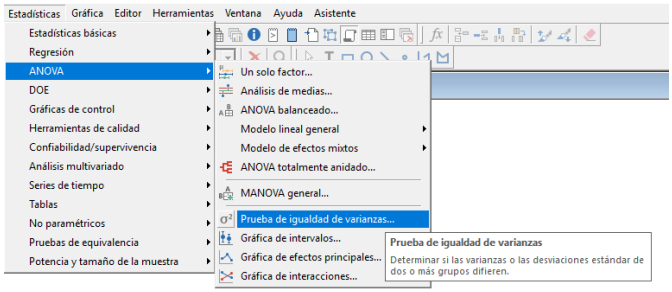
https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/anova/how-to/one-way-anova/methods-and-formulas/analysis-of-variance/

¿Se acepta o se rechaza la hipótesis nula?

Prueba de igualdad de varianzas

Si las gráficas de residuos dejan alguna duda sobre la validez de la suposición de que los residuos son homocedásticos, use el menú Estadísticas> ANOVA> Prueba de igualdad de varianzas para realizar las pruebas de Bartlett y Levene para la homogeneidad de las varianzas del error.

MINITAB informa los resultados cuantitativos de las pruebas en la ventana Sesión y también crea un gráfico de los intervalos de confianza para las desviaciones estándar de la población. Al igual que en otros casos, si dos intervalos de confianza se traslapan el uno del otro, es probable que existan razones para creer que las poblaciones son heterocedasásticas y que el ANOVA puede verse comprometido.



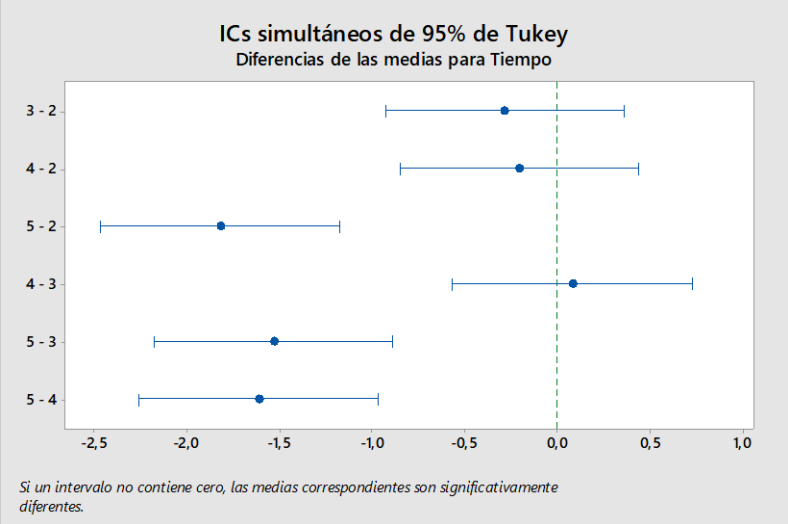
¿Qué son las comparaciones múltiples?

Las comparaciones múltiples de las medias permiten examinar cuáles medias son diferentes y estimar el grado de diferencia. Usted puede evaluar la significancia estadística de las diferencias entre las medias usando un conjunto de intervalos de confianza, un conjunto de pruebas de hipótesis o ambos. Los intervalos de confianza permiten evaluar la significancia práctica de las diferencias entre las medias, además de la significancia estadística. Como es habitual, la hipótesis nula de no diferencia entre medias se rechaza si y solo si el intervalo de confianza no contiene el cero.

¿Qué es el método de Tukey para comparaciones múltiples?

El método de Tukey se utiliza en ANOVA para crear intervalos de confianza para todas las diferencias en parejas entre las medias de los niveles de los factores mientras controla la tasa de error por familia en un nivel especificado. Es importante considerar la tasa de error por familia cuando se hacen comparaciones múltiples, porque la probabilidad de cometer un error de tipo I para una serie de comparaciones es mayor que la tasa de error para cualquier comparación individual. Para contrarrestar esta tasa de error más elevada, el método de Tukey ajusta el nivel de confianza de cada intervalo individual para que el nivel de confianza simultáneo resultante sea igual al valor que usted especifique.

Considere el resultado de la prueba de tukey del caso en cuestión:



Usted decide examinar las 10 comparaciones entre los cuatro computadores para determinar específicamente cuáles medias de tiempo son diferentes. Usando el método de Tukey, usted especifica que todo el conjunto de comparaciones debe tener una tasa de error por familia de 0.05 (equivalente a un nivel de confianza simultáneo de 95%). Minitab calcula que los 10 niveles de confianza individuales para obtener el nivel de confianza conjunto de 95%. Entendiendo este contexto, usted puede examinar entonces los intervalos de confianza para determinar si alguno de ellos no incluye el cero, lo que indica una diferencia significativa. Los intervalos de confianza que contienen cero indican que no hay diferencia. ¿Que concluye entonces con ésta prueba? ¿Qué recomendación puede hacer?

4. EXPERIMENTOS PARA CLASIFICACIONES MULTI FACTORIAL

Durante la preparación para un experimento de clasificación de una sola vía, a menudo se identifica una segunda variable de clasificación que podría afectar la respuesta que se estudia. Esta segunda variable podría tomar la forma de:

• Día/Hora. Las ejecuciones experimentales pueden tener que realizarse durante varios días o en diferentes momentos en el día.

• Usuarios. Las ejecuciones experimentales deben ser recolectadas por diferentes usuarios.

• Tipo de Memoria. Las ejecuciones experimentales podrían tener que hacerse con memorias RAM o cache de diferentes tamaño y tecnologías.

• Temperatura. Puede haber razones para considerar investigar el efecto de diferentes temperaturas.

• Equipos de cómputo. Puede haber una razón para probar una diferencia entre dos o Más máquinas.

• Métodos. Puede haber dos o más formas de operar el proceso, por ejemplo, grados de localidad espacial, o definición de variables locales o globales.

Una estrategia para gestionar una segunda variable inevitable en un experimento sería ignorarla, pero esto sería ingenuo y arriesgado. Si la segunda variable afecta la respuesta, entonces al ignorarla podríamos confundir su efecto con la primera variable o su efecto podría inflar el error estándar del modelo, lo que haría más difícil detectar diferencias entre los niveles de la primera variable. Sin embargo, si dos variables experimentales se incorporan a un experimento correctamente, estos y otros problemas pueden evitarse y, a veces, podemos obtener el beneficio adicional de aprender sobre las diferencias.

4.1 Realizar un ANOVA de dos factores

Considere la clasificación general de dos vías que se muestra en la Tabla 4.1 donde hay a diferentes niveles de la primera variable A indicada en columnas y b diferentes niveles de la segunda variable B indicada en filas. yij es la observación tomada en el i-ésimo nivel de A y el j-ésimo nivel de B. Para mantener el análisis simple, no hay replicación, aunque la replicación es posible y probablemente probable. Por ahora, no se preocupe por la replicación, deje que su software se encargue de ello.

		A				
	y_{ij}	**1**	**2**	**3**	**. . .**	***a***
B	1	y_{11}	y_{21}	y_{31}	. . .	y_{a1}
	2	y_{12}	y_{22}	y_{32}	. . .	y_{a2}
	3	y_{13}	y_{23}	y_{33}	. . .	y_{a3}
	⋮	⋮	⋮	⋮	⋮	⋮
	b	y_{1b}	y_{2b}	y_{3b}	. . .	y_{ab}

Una de las primeras técnicas de análisis de datos exploratorios para probar con los datos clasificados de dos vías es construir diagramas de caja usando solo una variable a la vez. Se requieren dos conjuntos de diagramas de caja. El primer conjunto se construye a partir de los datos clasificados de acuerdo con A y el segundo conjunto se construye a partir de los mismos datos clasificados de acuerdo con B. Se podrían agregar a cada diagrama de caja adicionales que muestren la distribución de las medias de la muestra.

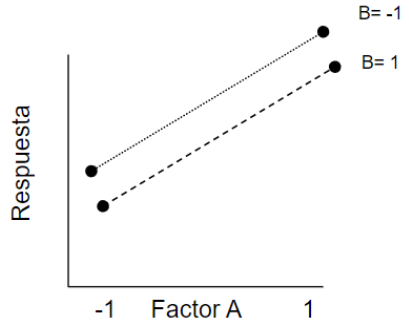
Esta modalidad de ANOVA tiene la propiedad de poder incorporar dos, factores, cada uno con “a” y “b” cantidad de niveles. La ANOVA de Dos Vías es un modelo experimental más complejo comparado con la de un solo factor. La ANOVA de dos vías, permite entender el efecto del factor A, el factor B, además de su interacción AB, lo que la convierte en una herramienta muy poderosa.

La hipótesis nula en una ANOVA de dos vías, se expresa como:

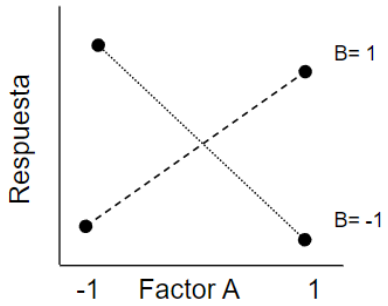
H0: No hay efecto significativo del Factor A, No hay efecto significativo del Factor B, No hay efecto significativo por la interacción entre AB

Las interacciones suceden cuando el efecto de un factor, sobre la variable de respuesta depende del nivel de otro factor.

No interacción significativa



interacción significativa

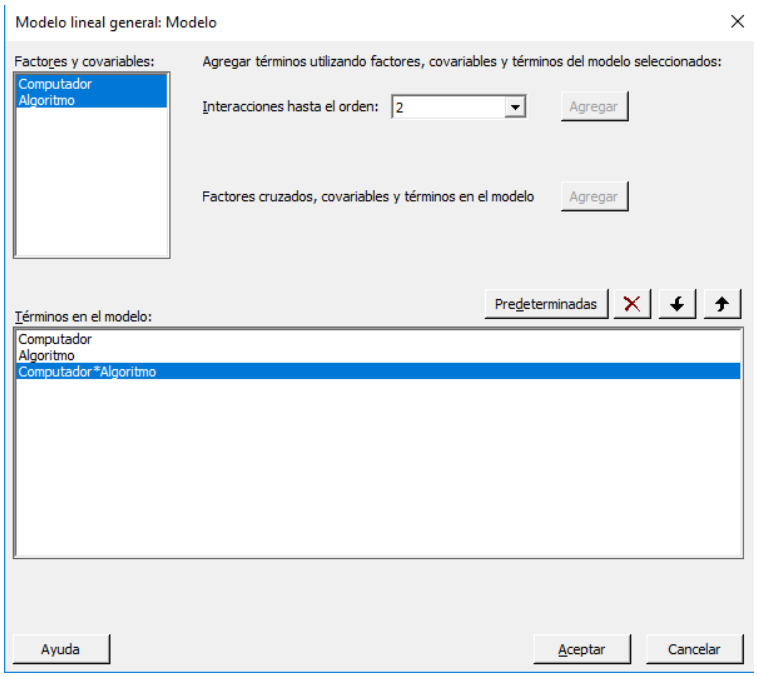


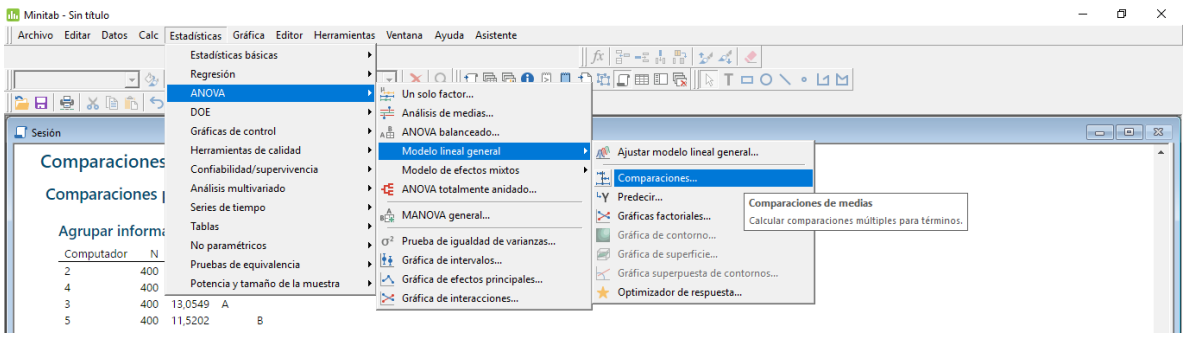
ANOVA dos vías: Tabla Sumaria

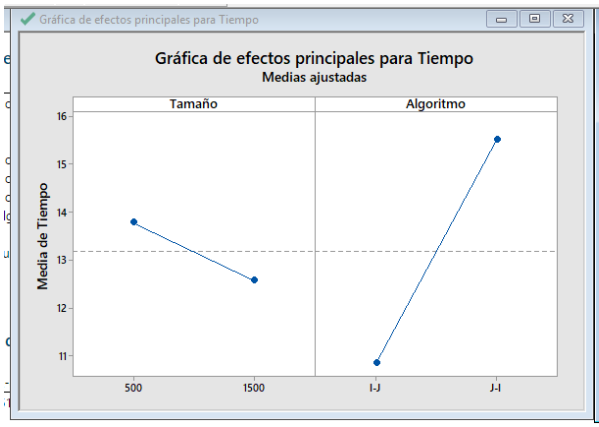
Fuente de Variación	Suma Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
A	SS_A	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$F_0 = \frac{MS_A}{MS_E}$
B	SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$F_0 = \frac{MS_B}{MS_E}$
Interacci{on	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$F_0 = \frac{MS_{AB}}{MS_E}$
Error	SS_E	$ab(n-1)$	$MS_E = \frac{SS_E}{ab(n-1)}$	
Total	SS_T	$abn-1$		

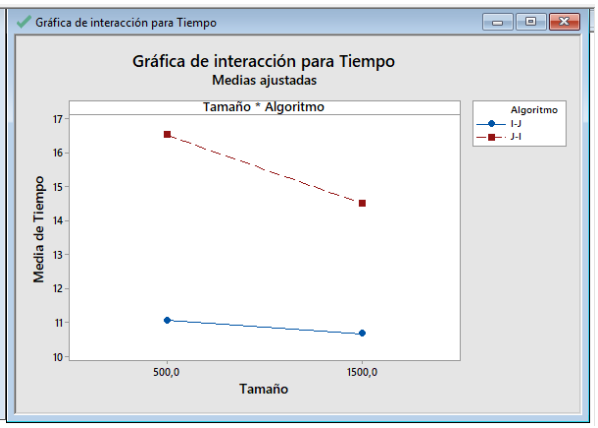
Para realizar un ANOVA de dos factores en Minitab, utilice **Estadísticas** > **ANOVA** > **Modelo lineal general** > **Ajustar modelo lineal general**. Supongamos que la respuesta se denomina Tiempo y sus factores son Computador y Tipo de algoritmo.

1. Elija **Estadísticas** > **ANOVA** > **Modelo lineal general** > **Ajustar modelo lineal general**.
2. En **Respuestas**, ingrese Tiempo.
3. En **Factores**, ingrese *Computador Algoritmo*.
4. Haga clic en **Modelo**.
5. En **Factores y covariables**, seleccione tanto **Computador** como **Algoritmo**. A la derecha de **Interacciones hasta el orden**, elija **2** y haga clic en **Agregar**.
6. Haga clic en **Aceptar** en cada cuadro de diálogo

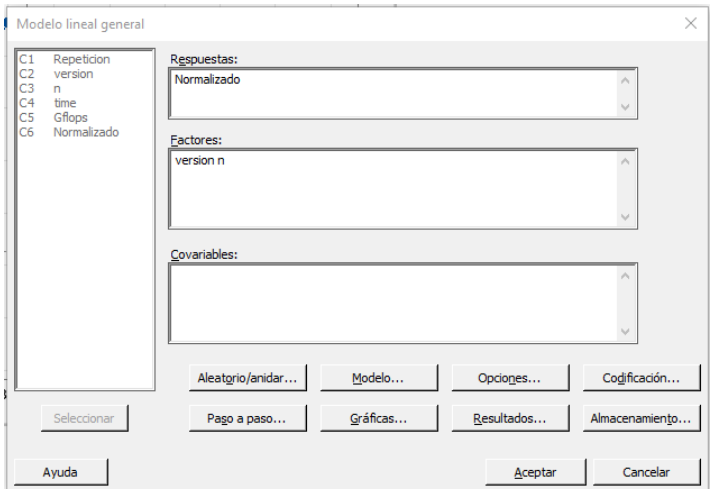

Modelo lineal general: Modelo
Factores y covariables:
Computador
Algoritmo
Agregar términos utilizando factores, covariables y términos del modelo seleccionados:
Interacciones hasta el orden:
2
Agregar
Factores cruzados, covariables y términos en el modelo
Agregar
Predeterminadas
Términos en el modelo:
Computador
Algoritmo
Computador*Algoritmo
Ayuda
Aceptar
Cancelar

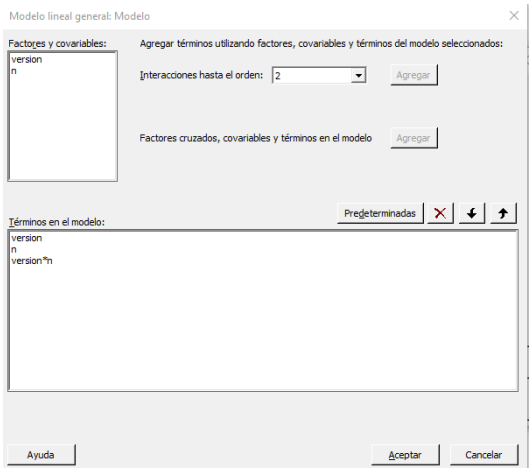

Minitab - Sin título
Archivo Editar Datos Calc Estadísticas Gráfica Editor Herramientas Ventana Ayuda Asistente
Estadísticas básicas
Regresión
ANOVA
DOE
Gráficas de control
Herramientas de calidad
Confiabilidad/supervivencia
Análisis multivariado
Series de tiempo
Tablas
No paramétricos
Pruebas de equivalencia
Potencia y tamaño de la muestra
Un solo factor...
Análisis de medias...
ANOVA balanceado...
Modelo lineal general
Modelo de efectos mixtos
ANOVA totalmente anidado...
MANOVA general...
Prueba de igualdad de varianzas...
Gráfica de intervalos...
Gráfica de efectos principales...
Gráfica de interacciones...
Ajustar modelo lineal general...
Comparaciones...
Predecir...
Gráficas factoriales...
Gráfica de contorno...
Gráfica de superficie...
Gráfica superpuesta de contornos...
Optimizador de respuesta...
Comparaciones de medias
Calcular comparaciones múltiples para términos.
Sesión
Comparaciones
Comparaciones
Agrupar informa
Computador N
2 400
4 400
3 400 13,0549 A
5 400 11,5202 B





Ajustando otro modelo:





Información del factor

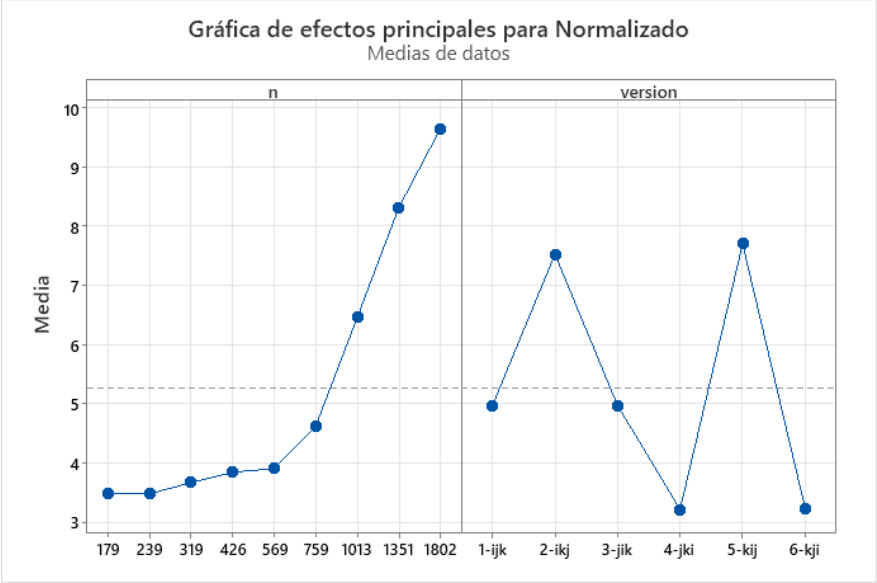
Factor	Tipo	Niveles	Valores
version	Fijo	6	1-ijk; 2-ikj; 3-jik; 4-jki; 5-kij; 6-kji
n	Fijo	9	179; 239; 319; 426; 569; 759; 1013; 1351; 1802

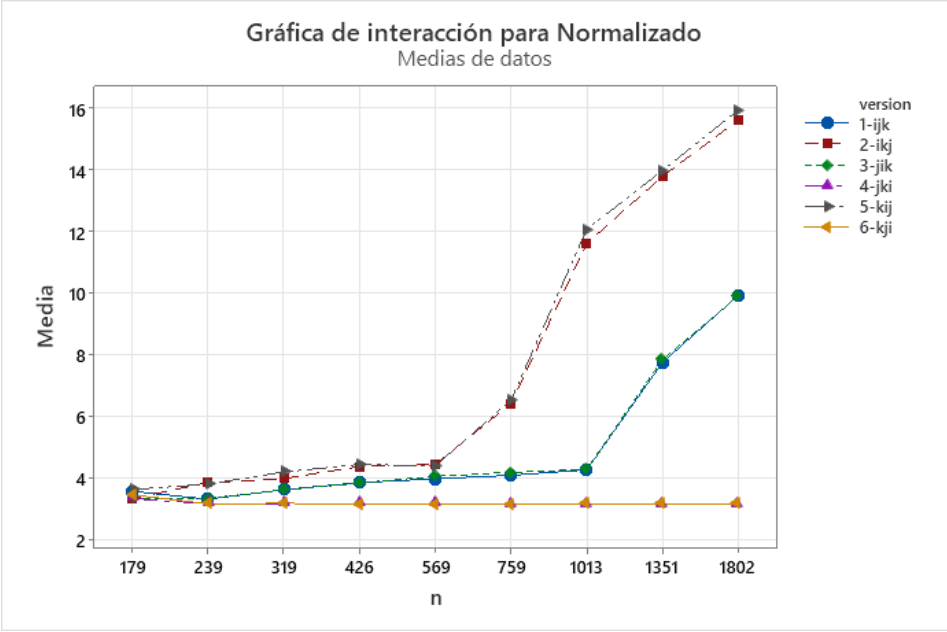
Análisis de Varianza

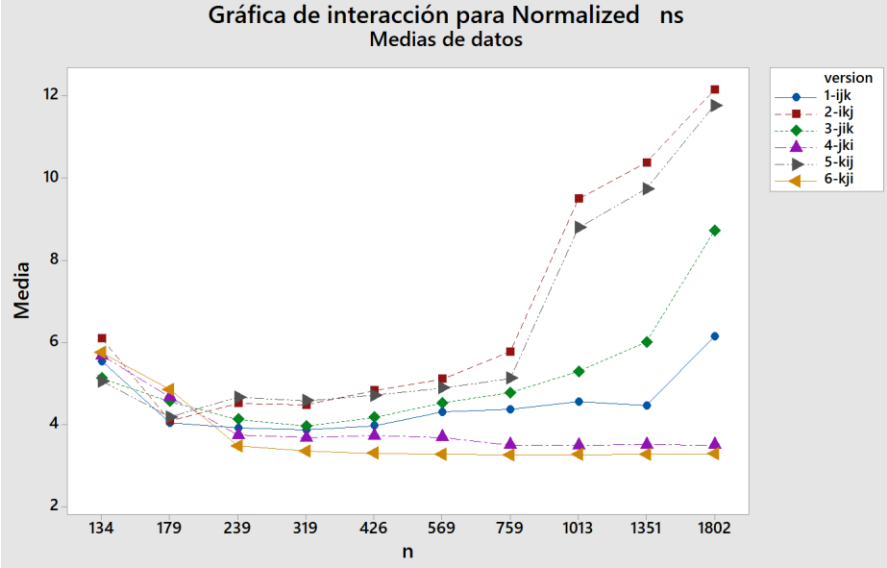
Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
version	5	526,62	105,324	4753,43	0,000
n	8	775,99	96,999	4377,71	0,000
version*n	40	612,50	15,312	691,07	0,000
Error	108	2,39	0,022		
Total	161	1917,50			

Resumen del modelo

S	R-cuadrado	R-cuadrado(ajustado)	R-cuadrado (pred)
0,148854	99,88%	99,81%	99,72%


Gráfica de efectos principales para Normalizado
Medias de datos
n
version
Media
10
9
8
7
6
5
4
3
179 239 319 426 569 759 1013 1351 1802
1-ijk 2-ikj 3-jik 4-jki 5-kij 6-kji


Gráfica de interacción para Normalizado
Medias de datos
Media
16
14
12
10
8
6
4
2
179 239 319 426 569 759 1013 1351 1802
n
version
1-ijk
2-ikj
3-jik
4-jki
5-kij
6-kji



4.2 Diseño de bloques aleatorizados

Un diseño de bloques aleatorizados es un diseño frecuentemente utilizado para minimizar el efecto de la variabilidad cuando se asocia con unidades discretas (por ejemplo, ubicación, computador, planta, tiempo). El caso usual consiste en distribuir aleatoriamente una réplica de cada combinación de tratamientos dentro de cada bloque. Por lo general, no hay un interés intrínseco en los bloques, y se considera que éstos son factores aleatorios. La suposición habitual es que el bloque por interacción de tratamiento es cero, y esta interacción pasa a ser el término de error para probar los efectos del tratamiento. Si designa la variable de bloqueo como Bloque, los términos en el modelo serían entonces Bloque, A, B y A*B. También especificaría el Bloque como el factor aleatorio.

Bibliografía

- MONTGOMERY, D. C. (2005). *DISEÑO Y ANALISIS DE EXPERIMENTOS* (2a. ed.). MEXICO: LIMUSA WILEY.
- MATHEWS, Paul G. ASQ Quality Press, [2005] Design of experiments with MINITAB
- Li, Zheng (Eddie) & O'Brien, Liam & Zhang, He & Cai, Rainbow. (2013). A Factor Framework for Experimental Design for Performance Evaluation ofCommercial Cloud Services. CloudCom 2012 - Proceedings: 2012 4th IEEE International Conference on Cloud Computing Technology and Science. 10.1109/CloudCom.2012.6427525.