UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN
DIVISIÓN SISTEMA UNIVERSIDAD ABIERTA Y EDUCACIÓN A DISTANCIA

APUNTE ELECTRÓNICO

Estadística Inferencial

Licenciatura en Administración

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

COLABORADORES

DIRECTOR DE LA FCA
Mtro. Tomás Humberto Rubio Pérez

SECRETARIO GENERAL
Dr. Armando Tomé González

COORDINACIÓN GENERAL
Mtra. Gabriela Montero Montiel
Jefa del Centro de Educación a Distancia y Gestión del Conocimiento

COORDINACIÓN ACADÉMICA
Mtro. Francisco Hernández Mendoza
FCA-UNAM

AUTORES
Lic. Manuel García Minjares
Mtra. Adriana Rodríguez Domínguez

REVISIÓN PEDAGÓGICA
Lic. Laura Antonia Fernández Lapray

CORRECCIÓN DE ESTILO
Mtro. José Alfredo Escobar Mellado

DISEÑO DE PORTADAS
L.CG. Ricardo Alberto Báez Caballero
Mtra. Marlene Olga Ramírez Chavero

DISEÑO EDITORIAL
Mtra. Marlene Olga Ramírez Chavero

Dr. Enrique Luis Graue Wiechers
Rector

Dr. Leonardo Lomelí Vanegas
Secretario General

Mtro. Tomás Humberto Rubio Pérez
Director

Dr. Armando Tomé González
Secretario General

Mtra. Gabriela Montero Montiel
Jefa del Centro de Educación a Distancia
y Gestión del Conocimiento / FCA

Estadística Inferencial
Apunte electrónico

Edición: agosto de 2017.

Facultad de Contaduría y Administración
Circuito Exterior s/n, Ciudad Universitaria
Delegación Coyoacán, C.P. 04510, México, Ciudad de México.

ISBN: 978-970-32-5314-2
Plan de estudios 2012, actualizado 2016.

Hecho en México

OBJETIVO GENERAL

Al finalizar el curso, el alumno será capaz de inferir las características de una población con base en la información contenida, así como de contrastar diversas pruebas para la toma de decisiones.

TEMARIO DETALLADO

(96 horas)

	Horas
1. Introducción al muestreo	4
2. Distribuciones muestrales	8
3. Estimación de parámetros	10
4. Pruebas de hipótesis	10
5. Pruebas de hipótesis con la distribución ji cuadrada	8
6. Análisis de regresión lineal simple	10
7. Análisis de series de tiempo	8
8. Pruebas estadísticas no paramétricas	6

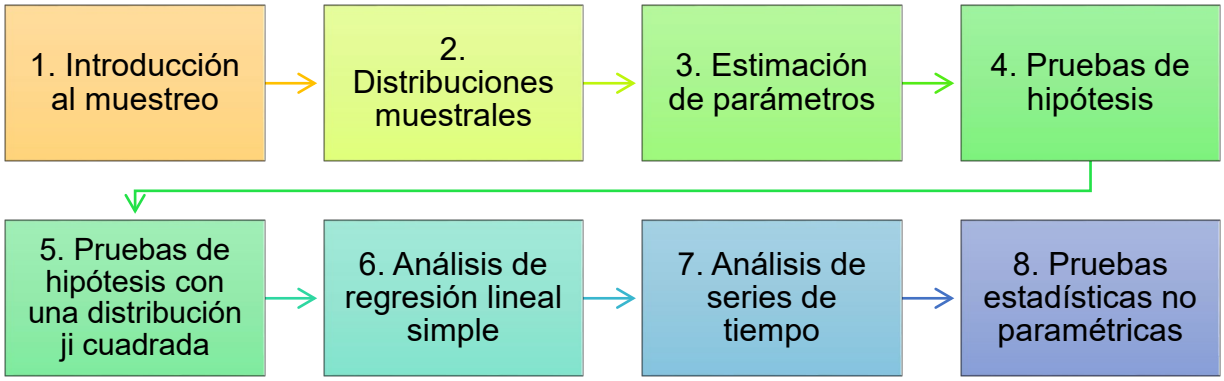
INTRODUCCIÓN

El plan de estudios vigente de las carreras ofrecidas por la Facultad de Contaduría y Administración de la UNAM pretende que en su ejercicio profesional el egresado sea capaz de analizar situaciones, evaluar acciones y decidir rumbos de acción. Esto es imposible si no dispone de información.

A fin de proveer al estudiante de herramientas para analizar información, dentro del mapa curricular de las carreras de la Facultad de Contaduría y Administración están las asignaturas de Estadística Descriptiva y Estadística Inferencial, materias de conocimientos fundamentales porque contribuyen a desarrollar capacidades de análisis y síntesis que el alumno necesita para una toma de decisiones adecuada.

A diferencia de la estadística descriptiva, donde la toma de decisiones descansa en la descripción de la información de una muestra, en la estadística inferencial el fundamento son las pruebas estadísticas que permiten inferir alguna característica de interés de una población con base en la información de una muestra.

El objetivo general de la materia Estadística Inferencial, establecida en el plan 2012, es que al término del curso el alumno sea capaz de inferir las características de una población con base en la información contenida en una muestra, y pueda contrastar diversas pruebas para la toma de decisiones. Para alcanzar este propósito, el programa comprende las siguientes unidades:



El estudio de las unidades 1-3 permitirá alcanzar la primera parte del objetivo general. La unidad 1 tiene la finalidad de que el estudiante conozca de forma global cómo se obtiene una muestra. La unidad 2 presenta las distribuciones muestrales más empleadas en inferencia estadística. Y la unidad 3 se enfoca a la realización de estimaciones de los parámetros de una población a través de la información de una muestra.

Una vez entendido cómo recolectar y obtener información de las muestras, lo siguiente que plantea el objetivo general es contrastar hipótesis con base en pruebas estadísticas realizadas con la información de una muestra. De esto tratan, en conjunto con la unidad 2, las unidades 4 y 5.

En la unidad 6, se muestra cómo analizar la regresión lineal simple para explicar el comportamiento de una variable a partir de otra. En este tipo de análisis, el contraste de hipótesis juega un papel central en la determinación de la existencia de esta relación, al igual que el tema de estimación para entender por qué son empleados los estimadores de mínimos cuadrados.

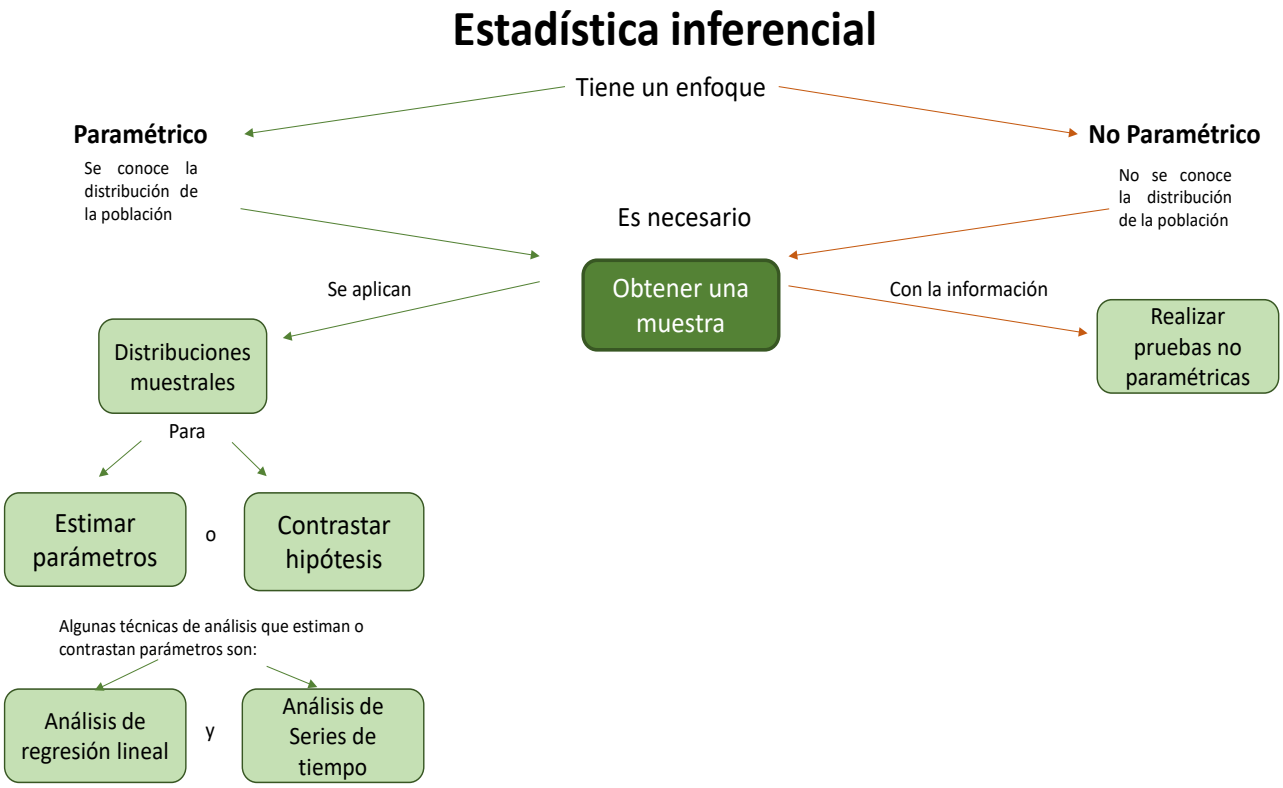
En la unidad 7, se busca que el alumno explique el comportamiento de una variable a lo largo del tiempo y realice un pronóstico de ella.

En la unidad 8, último tema del programa, se enseña al alumno a realizar análisis inferencial con métodos no paramétricos.

Como valor agregado, se plantea cómo emplear Microsoft Excel (2013) para aplicar algunas técnicas que se expondrán a lo largo de esta obra.

Este material está pensado para que el estudiante del SUAyED tenga un primer acercamiento a la estadística inferencial, cuyo aprendizaje autodidacta requiere de un contenido que facilite su comprensión y fomente profundizar en los temas con la consulta de la bibliografía sugerida. También puede aprovecharlo el estudiante del sistema escolarizado.

ESTRUCTURA CONCEPTUAL



UNIDAD 1

Introducción al muestreo

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

OBJETIVO PARTICULAR

Al terminar la unidad, el alumno reconocerá los diferentes tipos de muestreo y sus características.

TEMARIO DETALLADO

(4 horas)

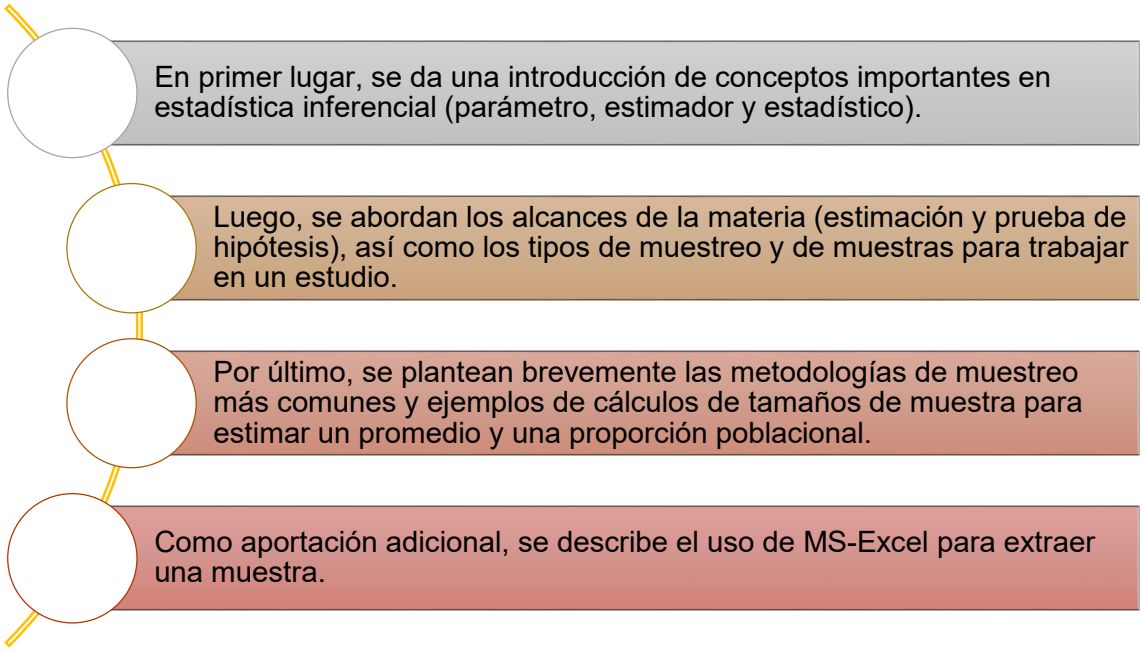
1. Introducción al muestreo

1.1. Parámetros estadísticos y estimadores
1.2. Estimación de parámetros y pruebas de hipótesis
1.3. Muestreo aleatorio y muestreo de juicio
1.4. Muestras únicas y muestras múltiples
1.5. Muestras independientes y muestras relacionadas
1.6. Tipos de muestreo aleatorio

INTRODUCCIÓN

El éxito de cualquier toma de decisiones depende de la calidad de información que se tenga. Hoy día, uno de los retos de las organizaciones es disponer de información accesible, detallada y actualizada que promueva una acertada toma de decisiones. En el desempeño profesional es común encontrar situaciones donde no se posee la suficiente información para tomar una decisión, por lo que se vuelve necesario realizar un esfuerzo extraordinario para recabarla; el dilema, entonces, es determinar cuánta información se requiere.

En esta unidad, se expone la importancia de la metodología del muestreo para extraer información que garantice resultados confiables.



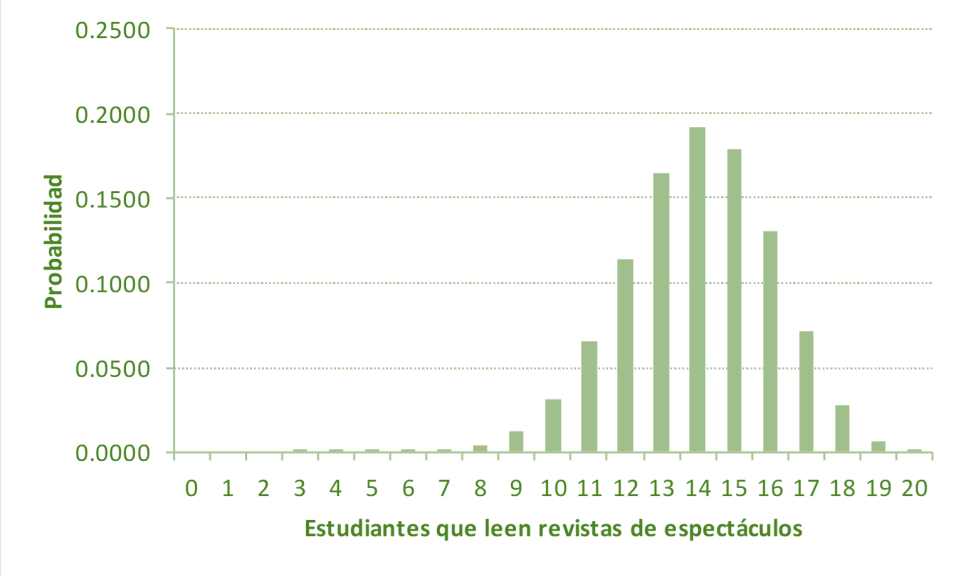
1.1. Parámetros, estadísticos y estimadores

En el curso de Estadística Descriptiva, se brindaron las herramientas para describir el comportamiento de un conjunto de datos con el empleo de tablas, gráficas y medidas descriptivas. Así, después de llevar a cabo los procedimientos para generarlos, se puede concluir acerca de la distribución de los datos su valor medio y variabilidad, y con base en ello tomar decisiones. Sin embargo, con frecuencia, la información descrita es un subconjunto o muestra proveniente de un conjunto mayor del que se desea conocer su comportamiento. Entonces, surge la pregunta si la información descrita en la muestra se puede generalizar a la población. Por ejemplo, si el promedio del porcentaje de aciertos de un examen de conocimientos de matemáticas aplicado a un grupo de Contaduría de primer semestre del turno matutino de la Facultad de Contaduría y Administración de la UNAM es 56%, ¿se podría decir que este resultado es generalizable a toda la población de la Facultad de Contaduría y Administración de la UNAM? El curso de Estadística Inferencial proporcionará los fundamentos para responder esta pregunta.

De acuerdo con lo estudiado en el curso de Estadística Descriptiva, el comportamiento de la distribución de una variable se encuentra relacionado con un valor denominado *parámetro*. Como ejemplo, supóngase que la proporción de personas que leen revistas sobre noticias de espectáculos es de 0.7 entre los estudiantes de primer semestre de Administración de la Facultad de Contaduría y Administración, y se desea estudiar el número de estudiantes que leen este tipo de publicaciones en una muestra de 20

estudiantes. La distribución de probabilidades de la variable asociada al ejercicio sería como se muestra en la figura 1.

Figura 1. Distribución de probabilidad del número de alumnos que leen revistas de espectáculos con una proporción de 0.7 y 20 encuestas aplicadas

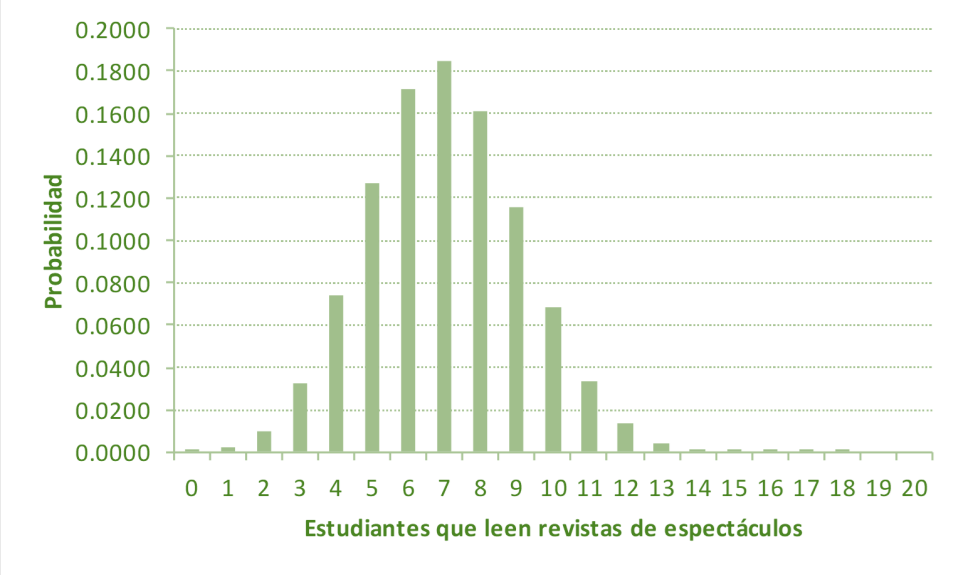


Fuente: elaboración propia con empleo de Microsoft Excel (2013)

En la figura anterior, se muestra la distribución de probabilidades de la variable asociada al experimento: número de estudiantes que leen revistas de espectáculos en 20 entrevistas. La variable estudiada en este experimento tiene una distribución binomial con n = 20 y p = 0.7. Las mayores probabilidades se observan entre 13 y 15 estudiantes. Es decir, es más probable que en el experimento resulte ese número de estudiantes quienes leen revistas sobre espectáculos.

Continuando con este ejemplo, ¿cómo sería la distribución de la variable asociada al experimento si la proporción de personas que leen publicaciones con contenidos de espectáculos fuera de 0.35 en vez de 0.7? La respuesta se muestra en la figura 2.

Figura 2. Distribución de probabilidad del número de alumnos que leen revistas de espectáculos con una proporción de 0.35 y 20 encuestas aplicadas



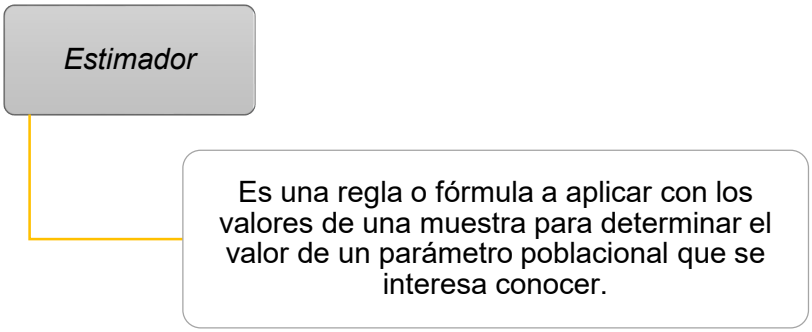
Fuente: elaboración propia con empleo de Microsoft Excel (2013).

La figura anterior muestra un patrón distinto al de la figura 1. En este caso, se observa una probabilidad mayor de que entre 6 y 8 de las 20 personas entrevistadas lean revistas de espectáculos.

En este ejemplo, modificar la proporción de alumnos que leen revistas de espectáculos cambió la distribución de probabilidades de la variable asociada al experimento. Y esta proporción es un parámetro cuyo valor condiciona la distribución de la variable de interés.

El ejemplo anterior muestra el efecto del valor de un parámetro en la distribución de una variable de interés, pero normalmente se ignora el valor de este parámetro y debe fijarse su valor. Supóngase que en el ejemplo anterior el problema de interés hubiera sido determinar la proporción de estudiantes de primer semestre de Administración de la Facultad de Contaduría y Administración de la UNAM que leen revistas de espectáculos a partir de entrevistar a 20 estudiantes. Supóngase que, de los 20 entrevistados, 8 leen esta clase de revistas. Entonces, de acuerdo con los resultados de esta muestra, la proporción de estudiantes que leen revistas de espectáculos es $\frac{8}{20} = \mathbf{0.4}$. La división realizada, $\frac{\acute{\boldsymbol{e}}\boldsymbol{xitos}}{\boldsymbol{tamaño\ de\ muestra}}$, es un *estimador* de la proporción de estudiantes de la población de interés que leen revistas de espectáculos, y el valor obtenido es una *estimación*.

Supóngase además que en vez de un valor se quisiera tener un rango de valores donde fuera más probable que se encuentre la proporción real (con base en los valores de la muestra, se analizará en la unidad 3 que la proporción real se encuentra entre 0.18 y 0.61).



En la unidad 3 de este curso, se mostrarán los estimadores más utilizados, así como la manera de realizar estimaciones, ya sea con valores puntuales o con un rango de valores posibles.

Regresando al ejemplo, ahora supóngase que, de acuerdo con la experiencia de estudios anteriores, se sabe que la proporción de alumnos que leen revistas de espectáculos es de 0.37, y se sospecha que esta proporción es mayor en esta generación. ¿El resultado obtenido en la muestra (0.4) nos permite afirmar que la proporción es mayor? En la unidad 4, se podrá contestar esta pregunta con el empleo de estadísticos de prueba, valores basados en la distribución y valores muestrales que permiten tomar una decisión sobre si apoyar o no una hipótesis. En este caso, el estadístico es de 0.274, por lo que no existe evidencia estadística para apoyar que la proporción de alumnos que leen revistas de espectáculo es mayor a 0.37.

Se puede afirmar que:

La estadística inferencial
Pretende determinar el valor de parámetros poblacionales utilizando estimadores o estadísticos de prueba con los valores de una muestra.

1.2. Estimación de parámetros y pruebas de hipótesis

En la sección anterior, se comentó que la estadística inferencial busca determinar el valor de parámetros poblacionales a partir de una muestra con el empleo de estimadores o estadísticos de prueba. Así, la estadística inferencial afronta dos problemáticas: estimación de parámetros y pruebas de hipótesis.

Estimación de parámetros

- Se pretende fijar el valor de un parámetro poblacional que se interesa conocer a través de una regla o fórmula basada en los valores de la muestra.

En el apartado precedente, se planteó el caso donde se deseaba determinar la proporción de estudiantes que leen revistas de espectáculos, que toma el papel de parámetro poblacional. Luego de entrevistar a 20 estudiantes, se obtuvo que 8 de ellos (0.4) leen revistas de este tipo. Aquí, el estimador es la división de los 8 casos que leen revistas entre el total de casos. En la unidad 3, se revisará cómo realizar estimaciones puntuales o por intervalos.

Pruebas o contrastes de hipótesis

Consisten en apoyar o rechazar una hipótesis acerca del valor de un parámetro poblacional a través del uso de un estadístico de prueba.

En la unidad 4, esto se abordará con mayor profundidad. En el ejemplo del subtema anterior, se contrastaron dos hipótesis al final: la proporción es 0.37; la proporción se ha incrementado. Después de aplicar un estadístico de prueba, se concluye que no existe evidencia para rechazar que la proporción es 0.37.

1.3. Muestreo aleatorio y muestreo de juicio

Como se ha mencionado, en estadística inferencial se intenta determinar el valor de un parámetro poblacional a partir de los valores de una muestra: tanto el tamaño como la manera de extraer esta muestra determinará la validez de los resultados. Antes de enfocarnos a los tipos de muestreo, es importante mencionar algunos conceptos básicos relacionados con el muestreo.

Población	Se llama así al total de unidades que cumplen con ciertas características medibles a las cuales se les aplicarán métodos estadísticos para su estudio. El tamaño de la población es denotada con la letra N.

Como ejemplo, supóngase que se desea estudiar los hábitos de estudio de los alumnos vigentes de la Facultad de Contaduría y Administración de la UNAM de la modalidad a distancia. Así, la población son los alumnos vigentes de la modalidad a distancia de la Facultad de Contaduría y Administración de la UNAM.

Censo	•Es la medición realizada a todas de unidades que conforman la población.

Supóngase que se desea conocer el número de empleados que tienen las 10 tiendas de conveniencia ubicadas en cierta colonia. Si se verifica la información de las 10 tiendas, entonces se realiza un censo.

Muestreo	Es la metodología con la que se determina el número de elementos que serán seleccionados de la población para formar un subconjunto llamado muestra.
Muestra	Es un subconjunto de la población cuyos elementos son elegidos mediante alguna metodología de muestreo; su estudio permitirá realizar inferencias respecto a la población. El tamaño de muestra se denota con la letra *n*.
Muestra representativa	Se dice que una muestra es representativa cuando las unidades que la conforman contienen las diferentes características de la población en una proporción semejante, de manera que es una imagen de ella.
Unidad muestral	Unidad más pequeña de la que se recaban las mediciones.
Marco muestral	Fuente de referencia de donde se selecciona la muestra. Como ejemplo, supóngase que se desea obtener información de los empleados de una empresa a través de una muestra, el marco muestral es la nómina de la última quincena.

Conveniencia de realizar un censo o levantar una muestra

Cuando se necesita levantar información, en ocasiones, surge el dilema de si es conveniente recabar la información a través de un censo o de una muestra. El censo es recomendable si el tamaño de la población no es demasiado grande o cuando los resultados tienen trascendencia. Por ejemplo, si un profesor imparte su clase a un grupo de 50 alumnos y desea conocer cuántos van a faltar un día previo a una fecha festiva, puede obtener la información preguntando a todo su grupo. Otro caso es el proceso de admisión a licenciatura en la UNAM, donde alrededor de 150,000 estudiantes aplican un examen de admisión. La asignación es realizada una vez que se han calificado todos los exámenes, y no a través de una muestra.

El muestreo conviene si no se cuenta con suficientes recursos para llevar a cabo un censo, y cuando los resultados permitan tener cierto margen de error. Una de sus principales ventajas es que se logra ahorrar costos y tiempos, y se tiene un mejor control (véase figura 3).

Figura 3. Ventajas del muestreo

VENTAJAS DEL MUESTREO		
Menor costo	**Menor tiempo**	**Mayor control**
En la capacitación del personal En la recolección, el análisis y obtención de resultados En el control de personal		

La figura anterior ilustra las ventajas del muestreo: menor costo, menor tiempo y mayor control en capacitar al personal, recolectar y analizar la información, y el control de campo. Todo esto conlleva una disminución del riesgo de cometer errores.

Muestreo aleatorio y muestreo de juicio

Para obtener una muestra, puede emplearse un muestreo aleatorio (probabilístico) o uno de juicio (no probabilístico). En el aleatorio, la selección de un elemento de la población depende del azar; mientras que en uno de juicio, la selección se basa en el criterio del investigador.

En la figura 4, se contrastan las principales diferencias entre el muestreo aleatorio (probabilístico) y el de juicio (no probabilístico).

Figura 4. Características del muestreo aleatorio (probabilístico) y de juicio (no probabilístico)

MUESTREO	
Probabilístico	**No probabilístico**
Considera la aleatoriedad para la selección de cada unidad de la población.	No considera el azar para la selección.
Se emplean métodos estadísticos.	Se realiza a juicio personal.
Los resultados se extrapolan a la población estudiada	Los resultados tienen validez solo para los elementos de la muestra.

En este curso, cuando se hable de los *resultados de una muestra*, se estará haciendo referencia a un muestreo aleatorio (probabilístico). De igual manera, cuando se mencione *muestreo probabilístico*, se estará refiriendo a un muestreo aleatorio.

1.4. Muestras únicas y muestras múltiples

En la sección anterior, se habló acerca de los tipos de muestreo que pueden emplearse para seleccionar una muestra. Normalmente, se requiere una muestra única para realizar inferencias de la población.

Como ejemplo, supóngase que se desea conocer las horas de estudio que los estudiantes de primer ingreso de la Facultad de Contaduría y Administración de la UNAM dedican a materias de matemáticas después del horario de clase. Para conocer este dato, es suficiente una muestra de alumnos a quienes se pregunte sobre qué tiempo dedican a estudiar matemáticas luego del horario de clase. En este ejemplo, el estudio se centra en una población, pero cuando interesa estudiar más de una población, se necesitará extraer muestras de cada una, por lo que el estudio requiere *muestras múltiples*. Para ilustrar esta situación, supóngase que se desea dar seguimiento a los egresados de posgrado de la UNAM, tanto de maestría como de doctorado. Dado que las poblaciones de maestría y doctorado son diferentes, se procede a extraer una muestra de los egresados de maestría y otra de los egresados de doctorado.

1.5. Muestras independientes y muestras relacionadas

En estadística inferencial, es frecuente querer realizar un comparativo entre grupos para confirmar si existe una diferencia significativa entre ellos.

Muestras independientes.

Cuando los grupos son muestras de poblaciones independientes, el estudio contempla.

Por ejemplo, se quiere conocer si los alumnos de Administración tienen mejor aprovechamiento en la asignatura Estadística Descriptiva en comparación con los de Contaduría. Para tal fin, se compara un grupo de estudiantes de Administración con uno de Contaduría y se realizan las mediciones correspondientes.

Muestra relacionada.

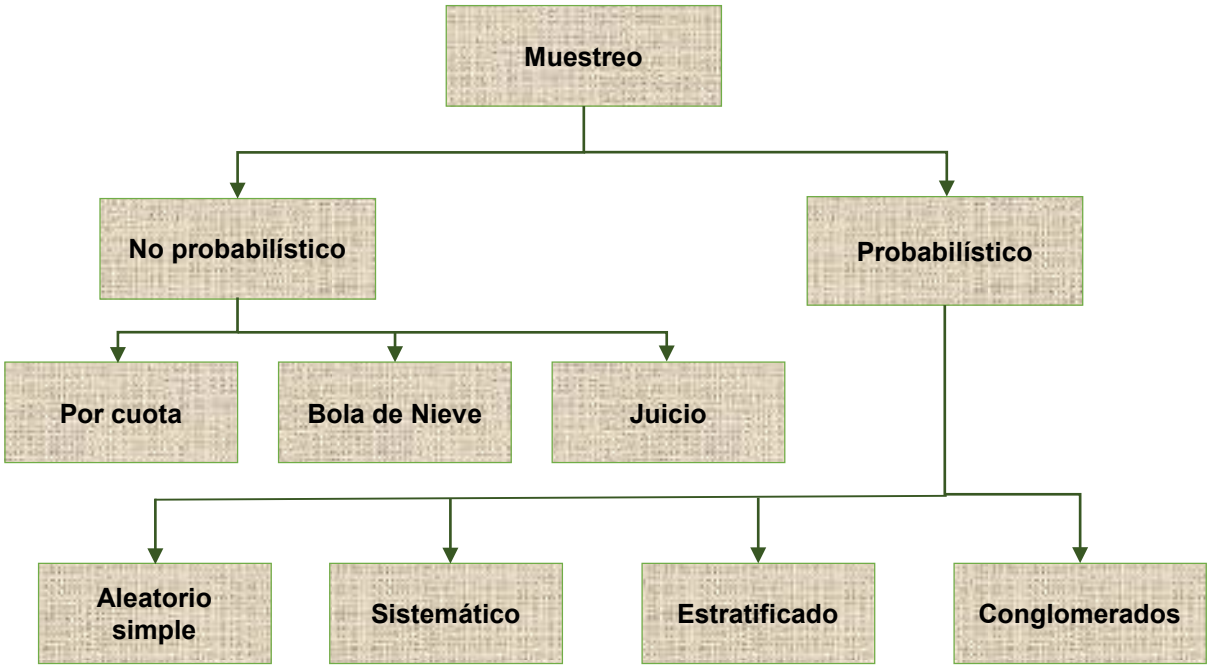
Cuando se efectúan mediciones de la misma muestra, pero en condiciones diferentes.

Por ejemplo, para complementar su estudio de Matemáticas Financieras, a un grupo de alumnos de Contaduría se les imparte un taller: ingresan a un portal donde resuelven problemas relacionados con la materia y se les aplica una evaluación al comienzo y final del semestre para medir la mejora de su aprovechamiento. Al mismo tiempo, se da seguimiento a un grupo control, el cual recibe la impartición tradicional del curso para contrastar la mejora. En este caso, como se trata del mismo grupo en diferentes momentos, el estudio trabaja con una muestra relacionada.

1.6. Tipos de muestreo aleatorio

En el subtema 1.3, se mencionó que el muestreo puede ser aleatorio (probabilístico) y de juicio (no probabilístico). Ahora, en la figura 5 se desglosan los principales tipos de muestreo de cada uno.

Figura 5. Principales tipos de muestreo aleatorio (probabilístico) y de juicio (no probabilístico)



Fuente: elaboración propia.

En este apartado, se expondrán los tipos de muestreo aleatorio: aleatorio simple, sistemático, estratificado y de conglomerados. Y los de juicio (no probabilístico): por cuota, juicio y bola de nieve.

A. Tipos de muestreo por juicio (no probabilísticos)

Muestreo por cuota

En este tipo de esquema de muestreo, predomina el criterio del investigador. Por lo general, se aplica cuando la persona encargada del estudio conoce bien las características de las unidades en estudio, por lo que fija el número de unidades que serán consideradas.

Por ejemplo, en un estudio de mercado, el gerente encargado en la venta de pañales quiere identificar la aceptación de un nuevo pañal con olor a chocolate, por lo que pide a la gente de campo que en cada supermercado muestre y dé a oler el pañal para conocer la reacción de las primeras 20 mamás que vayan a comprar algún pañal de la marca.

Muestreo por juicio o intencional

A criterio del investigador, son elegidos los elementos que pueden aportar al estudio.

Ejemplo: se quiere saber el estilo de liderazgo del Lic. José Luis Domínguez, gerente de ventas de la empresa ABDE, por lo que el área de recursos humanos entrevista a cinco personas que han trabajado con él.

Muestreo de bola de nieve

Este método se aplica para eventos donde es difícil recabar información, por tal razón, al encontrar una unidad que cumpla con las características que se buscan en el estudio, se espera que éste nos contacte con otro y ése con otro, y así sucesivamente hasta conseguir una muestra suficiente.

Por ejemplo, se quiere realizar un estudio de resistencia a alguna enfermedad en personas que su alimentación sea a base de insectos; o una psicoanalista desea probar que los reclusos que han asesinado más de cinco veces pueden ser buenos padres.

B. Tipos de muestreo aleatorio (probabilísticos)

En el muestreo aleatorio, la selección de la muestra considera el azar, de manera que cada elemento de la población tiene una probabilidad de ser incluido en la muestra.

A continuación, se exponen brevemente los tipos de muestreo aleatorio: aleatorio simple, sistemático, estratificado y de conglomerados. Después, se aborda un tema de mucha importancia: la determinación del tamaño de muestra en un muestreo aleatorio simple; y se termina con un ejemplo de cómo se obtiene una muestra con MS-Excel.

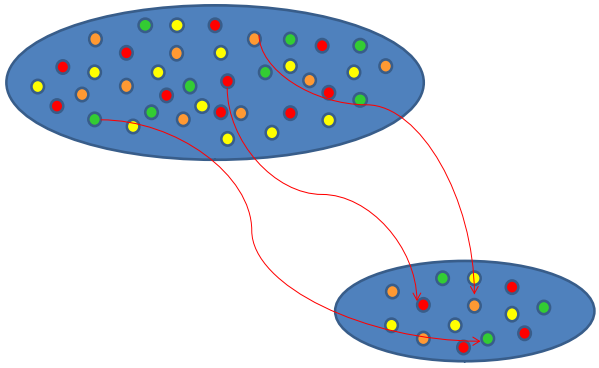
Muestreo aleatorio simple

En este método, las unidades de población tienen la misma probabilidad de ser elegidas. Cada elemento es seleccionado aleatoriamente.

En la figura 6, se ilustra cómo funciona esta metodología, se esquematiza la manera cómo funciona el muestreo aleatorio simple. El óvalo de mayor tamaño representa la población de interés; y los puntos contenidos, las unidades muestrales. En tanto, el óvalo de menor tamaño simboliza la muestra extraída de la población.

Las flechas indican que las unidades muestrales contenidas en la muestra provienen de la población. La elección de las unidades muestrales se realizó de manera aleatoria.

Figura 6. Funcionamiento del muestreo aleatorio simple



Fuente: elaboración propia.

Por ejemplo, en la comida de fin de año de una empresa se realiza una rifa con 20 premios. Se meten todos los nombres de los empleados en una tómbola y se van extrayendo los ganadores uno a uno de forma aleatoria.

Muestreo sistemático

A diferencia del anterior, en este método los elementos de la población son seleccionados cada *K* números, donde *K* es un valor constante que se determina a través de dividir el tamaño de la población entre el tamaño de la muestra deseada:

$$K = \frac{N}{n}$$

Se presenta a continuación la aplicación de este tipo de muestreo.

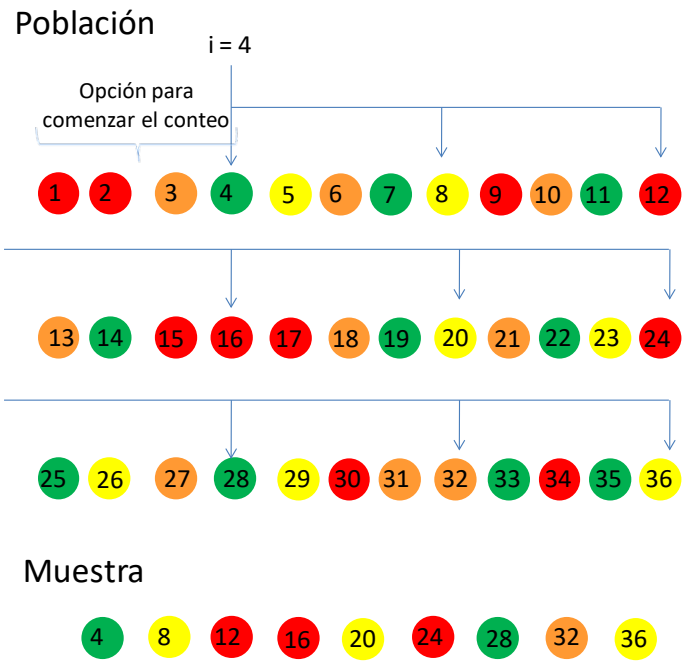
Una universidad cuenta con 36 alumnos de excelencia y desea extraer de ellos una muestra de 9 para aplicarles una evaluación psicométrica. ¿Cómo se debe seleccionar la muestra con un muestreo sistemático?

En este problema, el tamaño de la población (N) es 36; y el de la muestra (n), 9. Por tanto, la constante *K* es

$$K = \frac{36}{9} = 4$$

Este resultado indica que, de cada 4 alumnos, se escogerá uno para que sea parte de la muestra. Este resultado también apunta que se pueden extraer 4 muestras sistemáticas de tamaño 9. El método funcionaría de la siguiente manera: se numeran del 1 al 36 a los alumnos de excelencia; posteriormente, se elige un número aleatorio entre 1 y K (4), y a partir de ahí se selecciona cada *K* elemento. Supóngase que se escoge como primer alumno de la muestra al que se encuentra numerado con 4, entonces la muestra se conformaría con los alumnos numerados con 4, 8, 12, 16, 20, 24, 28, 32 y 36. En la figura 7, se ilustra esta metodología para el ejemplo.

Figura 7. Selección de una muestra sistemática para una población de tamaño 36 y una muestra de tamaño 9 con la unidad 4 como primer elemento de la muestra



Fuente: elaboración propia.

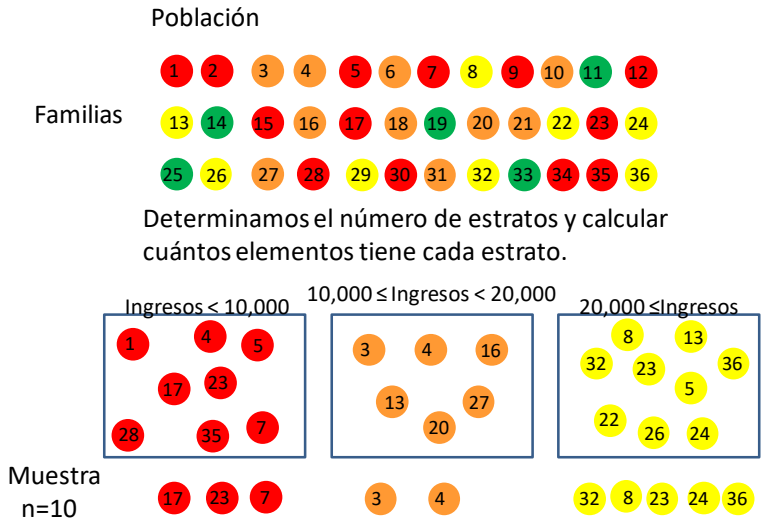
Muestreo estratificado

En el muestreo estratificado, la población es dividida en categorías diferentes entre sí, llamadas estratos, que poseen gran homogeneidad respecto a alguna característica (por ejemplo, profesión, sexo, estado civil, etcétera). Lo que se pretende con este tipo de muestreo es asegurar que todos los estratos de interés estarán representados adecuadamente en la muestra, además de ganar precisión.

Cada estrato funciona independientemente, pudiendo aplicarse dentro de ellos un muestreo aleatorio simple o sistemático para elegir los elementos que formarán parte de la muestra. Para ejemplificar esta metodología, supóngase que se quiere conocer la periodicidad con que 36 familias acuden al supermercado. A fin de estudiar mejor

la población, se decidió segmentarla en tres estratos de acuerdo con su nivel de ingreso mensual: con ingresos menores a $10,000; con ingresos entre $10,000 y $20,000; con ingresos mayores de $20,000. Dado lo anterior, se decidió tomar una muestra de tamaño 10, donde estuvieran representados los tres estratos. En la figura 8, se ilustra este tipo de muestreo.

Figura 8. Ilustración de un muestreo estratificado para una población de 36 familias dividida en tres estratos de ingreso, de la que se extrae una muestra de tamaño 10



Fuente: elaboración propia.

La figura anterior ilustra cómo se agrupa a la población original en tres estratos de ingreso, y de cada uno se extraen elementos para conformar el tamaño total de muestra que se necesita. Es práctica común que el número de elementos de la muestra de cada estrato sea proporcional al tamaño del estrato con respecto al total poblacional.

Muestreo por conglomerados

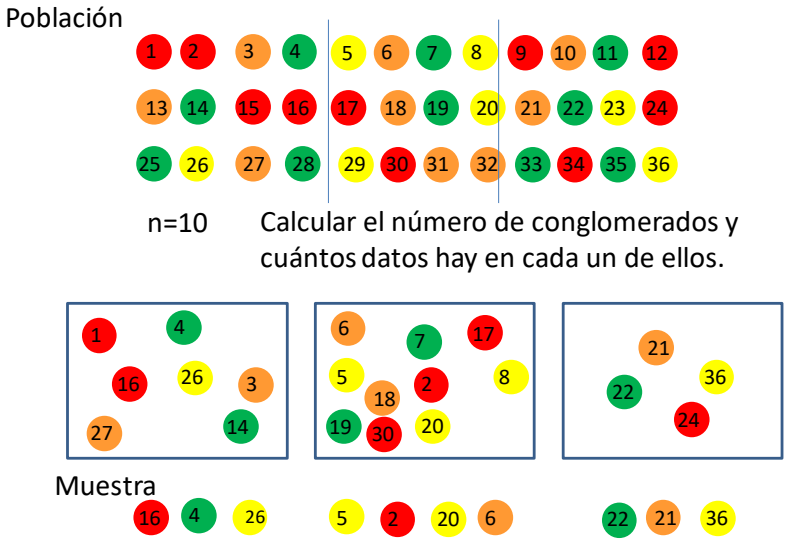
En este tipo de muestreo, cada unidad de la muestra está formada por un grupo de elementos, al que se le llama *conglomerado*. Este grupo contiene representantes de toda la población (de acuerdo con la característica que se mida).

Muestreo por conglomerados

El muestreo por conglomerados consiste en seleccionar aleatoriamente el número de conglomerados necesario para alcanzar el tamaño muestral, donde se investigan a todos los elementos que componen los conglomerados elegidos, o a una muestra.

La figura 9 ejemplifica esta metodología para una población de 36 elementos agrupados en tres conglomerados de 12 elementos cada uno.

Figura 9. Ilustración de un muestreo por conglomerados donde se extrae una muestra de tamaño 10 de una población de 36 elementos agrupados en tres conglomerados de tamaño 12



Fuente: elaboración propia.

En la figura anterior, el diseño de muestreo por conglomerados aplicado es el siguiente: se consideran los tres conglomerados y de cada uno se extrae una muestra.

En una segunda etapa, se extrae otra muestra de la anterior conformando los 10 elementos que se necesitaban.

Se pueden presentar variantes en el muestreo por conglomerados de acuerdo con el contexto de la situación, pero en esencia la metodología consiste en seleccionar una muestra de conglomerados y escoger de cada uno una muestra de las unidades que lo conforman.

Errores de estimación

Al aplicar un muestreo, existirá un error en las estimaciones porque no se está recabando información de toda la población, por ello el arte del muestreo consiste en determinar la muestra que minimice ese error. Cuando se recaba información de una muestra, se pueden presentar dos tipos de errores:

Atribuibles al muestreo	No atribuibles al muestreo
Son por la diferencia entre el valor del estimador muestral y el valor del parámetro poblacional considerando la información de la muestra con la que se trabajó.	Se explican, entre otras causas, por un mal diseño del instrumento, la logística implementada o una elevada tasa de no respuesta.

Cálculo del tamaño de muestra en un muestreo aleatorio simple

Como se mencionó en el apartado anterior, en todo ejercicio de muestreo va a existir un error de estimación, por lo que de antemano debe fijarse el límite de error permitido,

así como garantizar que ese error no sea mayor a lo permitido en un cierto número de repeticiones. Para lograr lo anterior, el tamaño de muestra juega un papel central, ya que, a medida que se tenga mayor información de un parámetro, se incrementa la probabilidad de realizar una estimación certera.

En la siguiente tabla, se exponen las fórmulas para calcular el tamaño de una muestra para estimar una media y una proporción (parámetros) cuando se tiene conocimiento del tamaño de la población *N* y cuando no es así[1].

Tabla 1. Fórmulas para calcular el tamaño de muestra para estimar una media y proporción poblacional cuando se conoce o no el tamaño de la población

Parámetro	N conocida	N desconocida
Media	$n = \frac{Z^2 S^2 N}{Ne^2 + Z^2 S^2}$	$n = \frac{Z^2 S^2}{e^2}$
Proporción	$n = \frac{Z^2 pqN}{Ne^2 + Z^2 pq}$	$n = \frac{Z^2 pq}{e^2}$

Donde:

[1] Para efectos de este curso, se asumirá que la fracción $\frac{n}{N}$ no es importante.0.

- **n** • tamaño de la muestra
- **N** • tamaño de la población
- **S** • desviación estándar
- **p** • proporción muestral
- **q** • 1 – p
- **e** • error permitido
- **Z** • Nivel de confianza, expresado como valor del cuantil *z* de una distribución normal estándar que separa la curva en dos áreas de tamaño 1 – α/2 y α/2 (0<α<1).

En la tabla 2, se muestran los valores de *z* para niveles de confianza de 90%, 95% y 99%.

Tabla 2. Valores de *z* para niveles de confianza de 90%, 95% y 99%

Nivel de confianza	z
90%	1.64
95%	1.96
99%	2.58

Como se mencionó, estos valores *z* son los cuantiles de una distribución normal estándar que separa la curva en dos áreas de tamaño 1 – α/2 y α/2 (0<α<1). Por ejemplo, para un nivel de confianza de 95%, α = 1 – 0.95 = 0.05 y α/2 = 0.05/2 = 0.025. El cuantil z = 1.96 separa la curva normal estándar en dos regiones de tamaño 1 – 0.025 = 0.975 y 0.025.

Ejemplos de cálculo de tamaño de muestra

A continuación, se muestran ejemplos de cómo calcular el tamaño de muestra para estimar una media o proporción poblacional.

1. Calcular el tamaño de muestra que se requiere para estimar el ingreso medio de un despacho de consultoría de 90 empleados en nómina, donde se conoce que existe una desviación de $15,000. El tamaño de muestra debe garantizar un error de estimación máximo de $5,000, con un nivel de significancia del 95%.

¿Qué variables se conocen?

N = 90

S = 15,000

e = 5,000

Z = 1.96 (véase tabla 2)

¿Se conoce o no el valor de N? Sí, N = 90.

¿Es un cálculo para un promedio o una proporción? Promedio, ya que se pide estimar el gasto administrativo medio.

Fórmula que se aplica:	Sustituyendo los valores:
$n = \frac{Z^2 S^2 N}{Ne^2 + Z^2 S^2}$	$n = \frac{1.96^2 \cdot 15,000^2 \cdot 90}{90 \cdot 5,000^2 + 1.96^2 \cdot 15,000^2}$ $n = \frac{3.8416 \cdot 225,000,000 \cdot 90}{90 \cdot 25,000,000 + 3.8416 \cdot 225,000,000}$ $n = \frac{77,789,541,119}{2,250,000,000 + 864,328,234.7}$ $n = \frac{77,789,541,119}{3,114,328,235}$ $n = 24.9779$ $n = 25$ Es decir, se tomará una muestra de 25 empleados.

2. Se desea conocer cuál es el grado de satisfacción de los 3582 alumnos de primer ingreso de la Facultad de Contaduría y Administración de la UNAM con respecto al servicio de las ventanillas. En las últimas tres generaciones, esta aceptación fue del 40%. Es necesario determinar a cuántos alumnos hay que entrevistar para garantizar un error máximo de 10 puntos porcentuales con un nivel de significancia del 90%.

¿Qué variables se conocen?

N = 3582 alumnos

P = 40% = 0.4

e = 10%, es decir, 0.10

Z = 1.64 (véase tabla 2)

Dado que el parámetro que se busca estimar es una proporción, el tamaño de muestra se determina con la siguiente fórmula:

Fórmula que se aplica:	**Sustituyendo los valores:**
$$n = \frac{Z^2pqN}{Ne^2 + Z^2pq}$$	Para este caso, falta calcular *q*, se sabe que q = 1– p, entonces: $$q = 1 - 0.4 = 0.6.$$ Así: $$n = \frac{(1.64^2)(0.4)(0.6)(3,582)}{(3,582)(0.1)^2 + (1.64^2)(0.4)(0.6)}$$ $$n = \frac{2,312.195328}{35.82 + 0.645504}$$ $$n = \frac{2,312.195328}{36.465504}$$ $$n = 63.41 = 64$$ Con 64 entrevistas, se garantiza una estimación de *P* con un error de 10% y un nivel de confianza de 90%.

3. Una empresa que comercializa aparatos electrónicos desea estimar el número promedio de aparatos que adquieren anualmente sus principales clientes. Se conoce que la desviación estándar es de 90 aparatos. Es necesario calcular el tamaño de muestra que garantice un nivel de confianza de 99% con un error permitido de 10 piezas.

¿Qué variables se conocen?

S = 90

e = 10

Z = 2.58 (véase tabla 2)

Dado que no se conoce el tamaño poblacional y que el parámetro que se busca estimar es un promedio, el tamaño de muestra se determina con la siguiente fórmula:

Fórmula que se aplica:	Sustituyendo los valores:
$$n = \frac{Z^2 S^2}{e^2}$$	Así: $$n = \frac{(2.58^2)(90^2)}{10^2}$$ $$n = \frac{53,916.84}{100}$$ $$n = 539.17 = 540$$ Con 540 entrevistas, se garantiza una estimación del promedio con un error de 10 piezas y un nivel de confianza de 99%.

4. Históricamente, la proporción de vuelos demorados de una aerolínea es de 10%. Los responsables de la aerolínea desean revisar los itinerarios de una muestra de vuelos del último año para comprobar si se sigue observando la misma proporción de demora. Se pide calcular el tamaño de muestra que permita estimar la proporción de vuelos demorados en un año con un nivel de confianza de 95% y un error de 3 puntos porcentuales.

¿Qué variables se conocen?

P = 10% = 0.1

e = 3%, es decir, 0.03

Z = 1.96 (véase tabla 2)

Como se desconoce el tamaño de la población y el parámetro que se busca estimar es una proporción, el tamaño de muestra se determina con la siguiente fórmula:

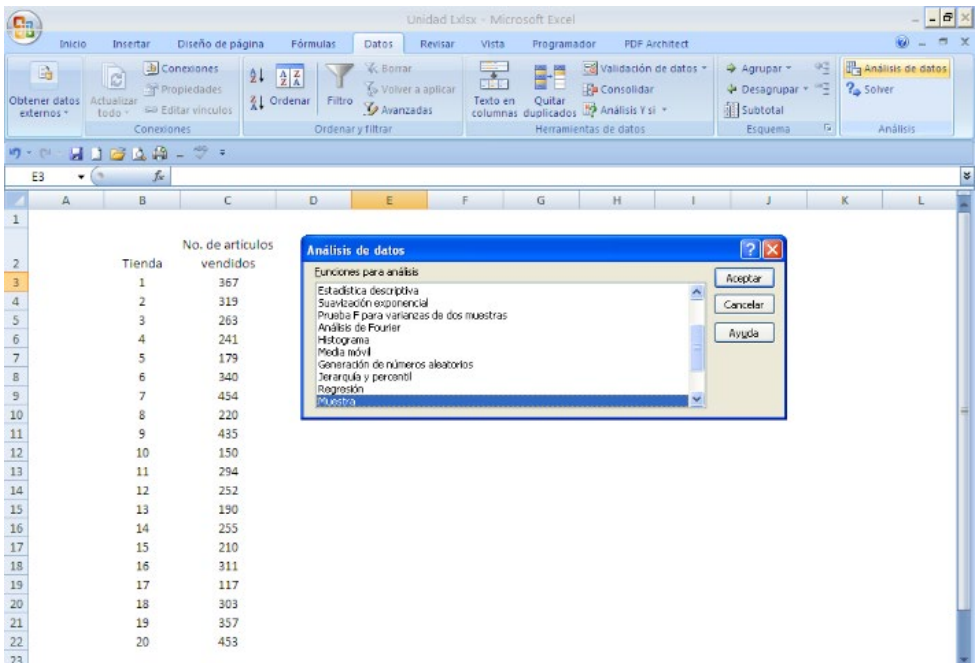
Fórmula que se aplica:	Sustituyendo los valores:
$n = \frac{Z^2 pq}{e^2}$	Donde q = 1 – p = 1 – 0.1 = 0.9. Así: $n = \frac{1.96^2(0.1)(0.9)}{0.03^2}$ $n = \frac{(3.8416)(0.09)}{0.0009}$ $n = \frac{0.345744}{0.0009}$ $n = 384.16 = 385$ Con 385 entrevistas, se garantiza una estimación de *P* con un error de 3% y un nivel de confianza de 95%.

Selección de una muestra con MS-Excel

MS-Excel en su módulo de análisis de datos que permite extraer una muestra de un conjunto de datos. Para valorar su utilidad, se trabajará con el siguiente ejemplo.

Supóngase que cierta marca de ropa cuenta con 20 establecimientos y se quiere elegir al azar cinco de ellos para realizarles una visita y auditar que las ventas reportadas coinciden con las que se realizan realmente.

Antes de emplear la herramienta de Excel, se sugiere numerar las 20 tiendas. A continuación, ir al menú Datos y elegir la opción Análisis de datos. Se desplegará una caja de diálogo con las opciones de análisis que se pueden ejecutar en el módulo, elegir la opción Muestra.



Fuente: Microsoft Excel (2013).

Se desplegará otro cuadro de diálogo que se divide en tres partes: Entrada, Método de muestreo y Opciones de salida. A continuación, se explica cada una.

Entrada.	En esta sección, se introduce la región donde se encuentra la numeración asignada a las tiendas (región de entrada).	
Método del muestreo.	En esta sección, se elige el tipo de muestreo a implementar. Excel considera dos:	**Periódico.** Se refiere al muestreo sistemático. En caso de elegir esta opción, se activa la casilla donde se indica el periodo de selección (K).
		Aleatorio. Se refiere al muestreo aleatorio simple. Si se opta por este tipo de muestreo, el paquete solicita el tamaño de la muestra.

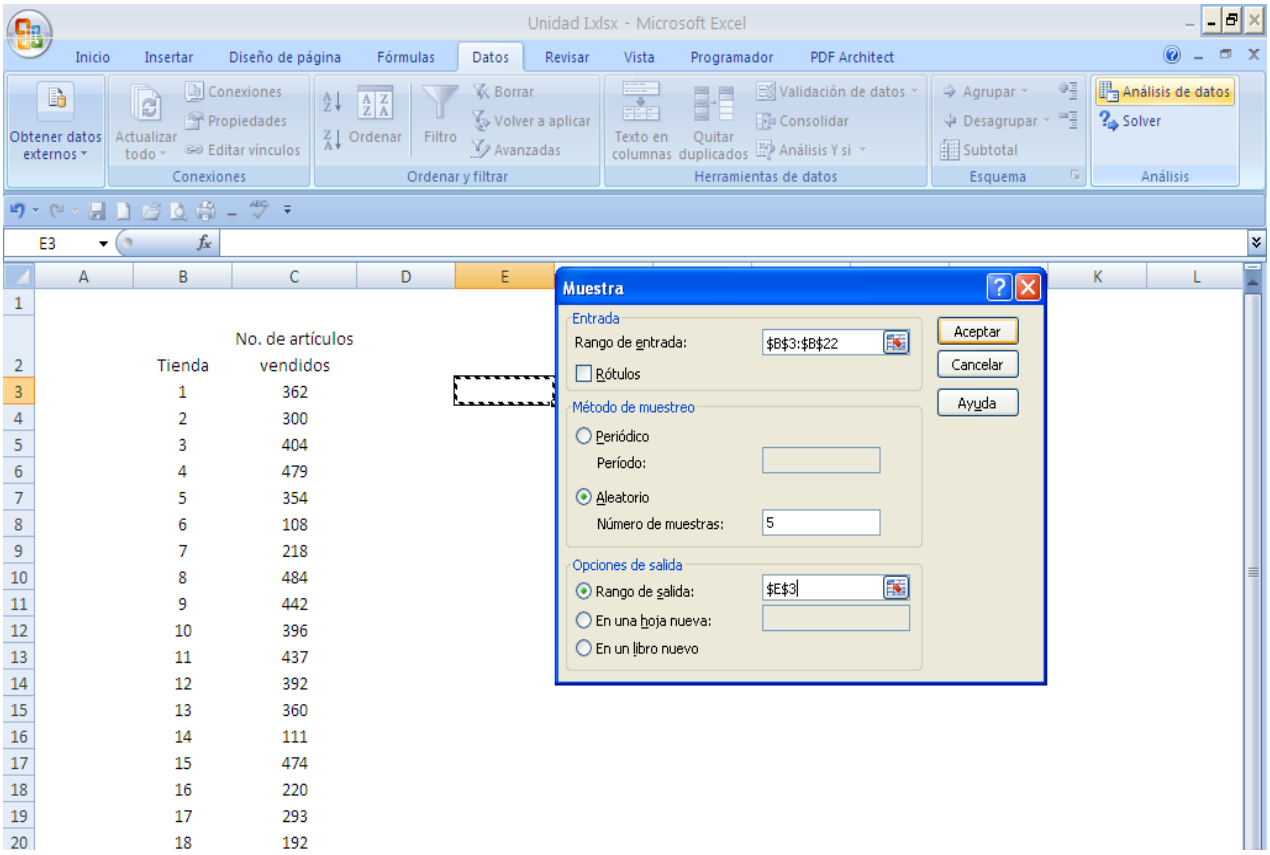
Para este ejemplo, se elige la opción de Aleatorio, y en la casilla de Número de muestras se captura el número de unidades que tendrá la muestra (5).

Opciones de salida.	En esta sección, se indica dónde se va a escribir la muestra: en un rango de salida, una nueva hoja o un nuevo libro.

En este ejemplo, se elige Rango de salida y se ingresa la coordenada de la celda en la cual se desea que comience a escribir la muestra, en este caso, la celda es E3.

Si se elige como alternativa en una nueva hoja, la muestra se escribe en una hoja nueva del mismo archivo. En caso de optar por Libro nuevo, la muestra se escribirá en un archivo nuevo.

Una vez completadas las secciones, oprimir Aceptar.

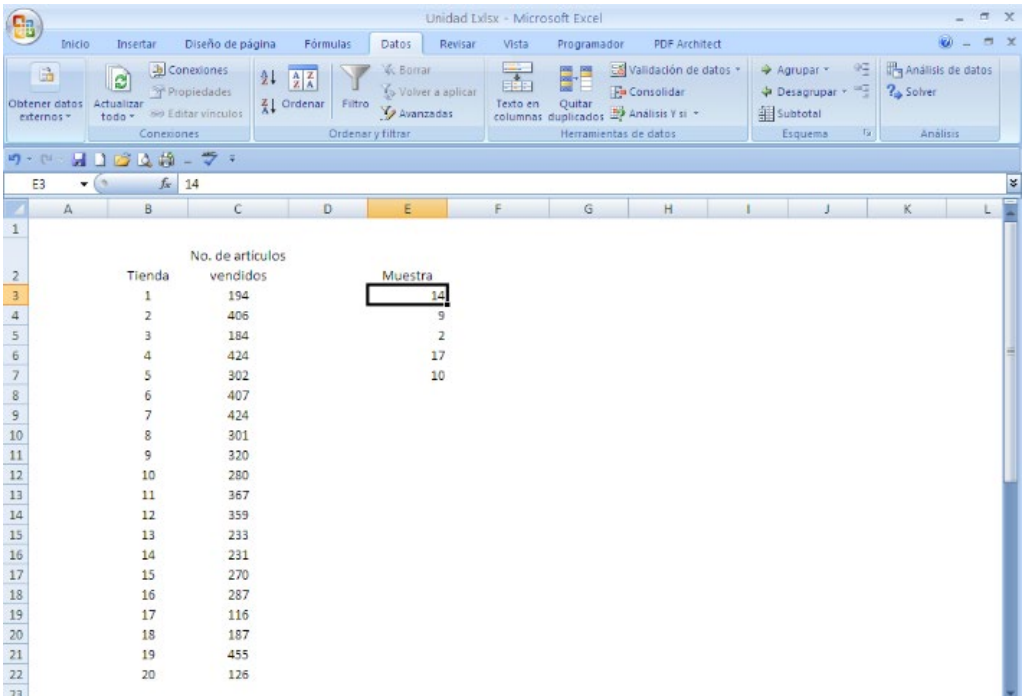


Fuente: Microsoft Excel (2013).

Excel mostrará los elementos de la muestra en donde se le indicó. En este ejemplo, Excel seleccionó las tiendas 14, 9, 2, 17 y 10.[2]

Se recomienda revisar que no existan números repetidos; de ser así, se puede volver a escoger una nueva muestra del tamaño de los elementos que se desean reemplazar.

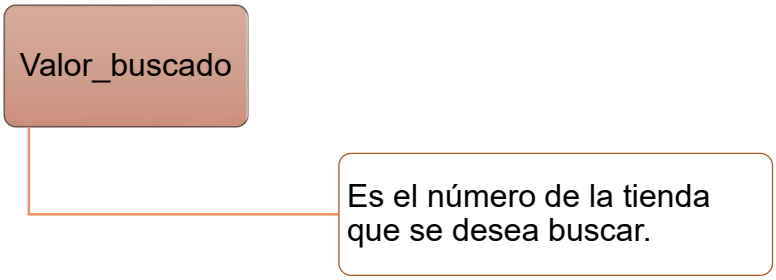
[2] Como se eligió un muestreo aleatorio, los resultados no necesariamente deben coincidir.



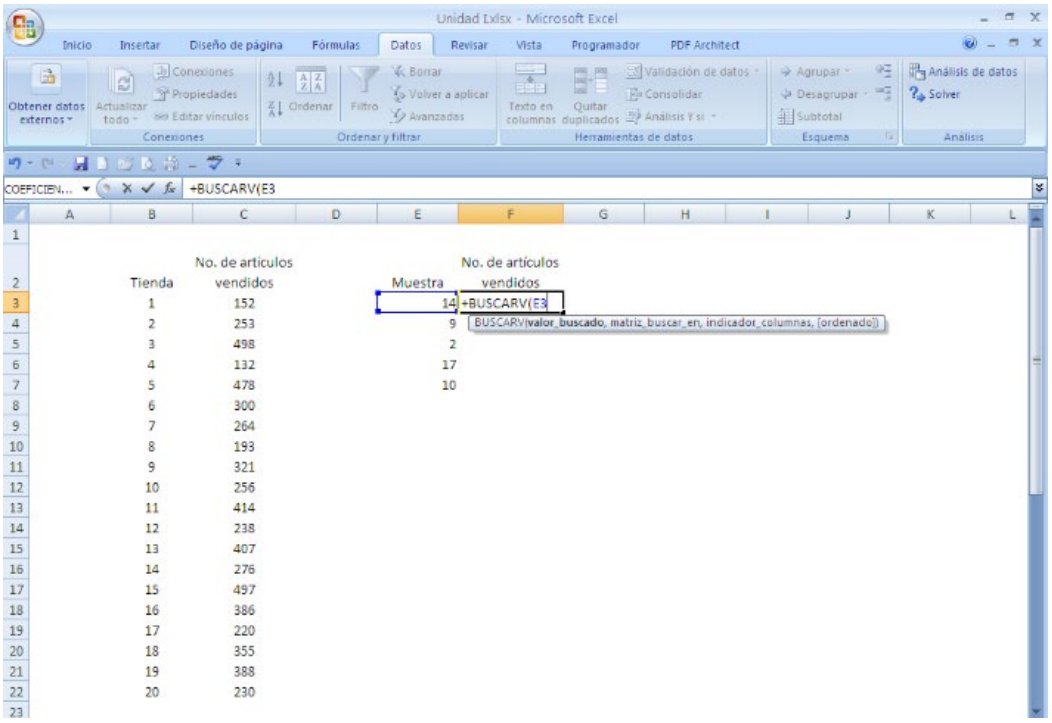
Fuente: Microsoft Excel (2013).

Supóngase que a los elementos de la muestra se quiere agregar el número de artículos vendidos. Para hacerlo, se puede emplear la función Buscarv, que tiene la siguiente estructura:

Buscarv (valor_buscado, matriz_buscar_en, indicador_columnas, [ordenado])



En este ejemplo, son los valores arrojados de la muestra 14, 9, 2, 17 y 10.

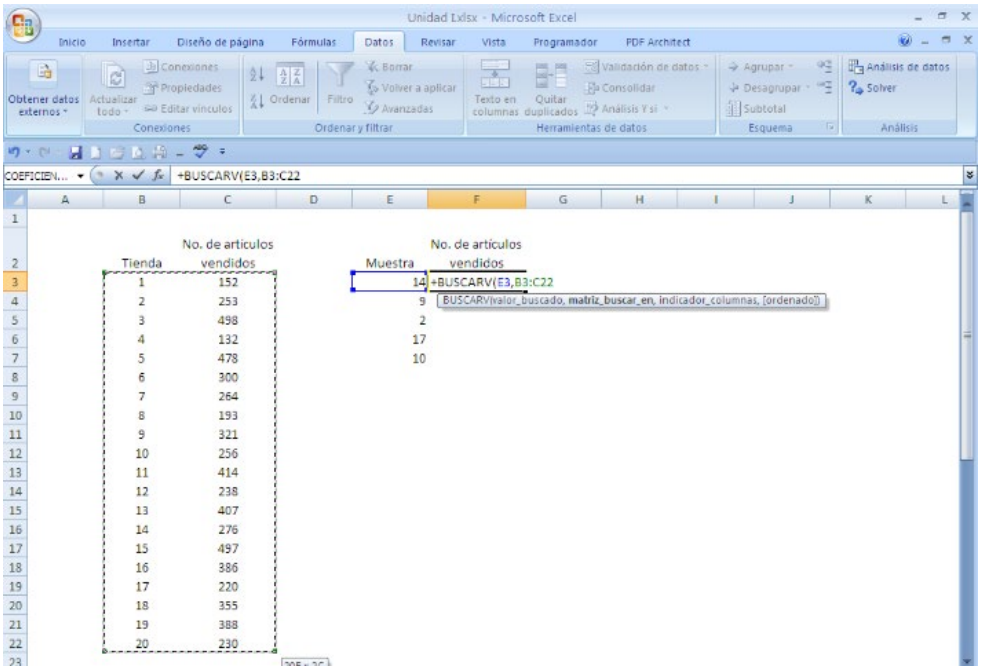


Fuente: Microsoft Excel (2013).

Matriz_buscar_en

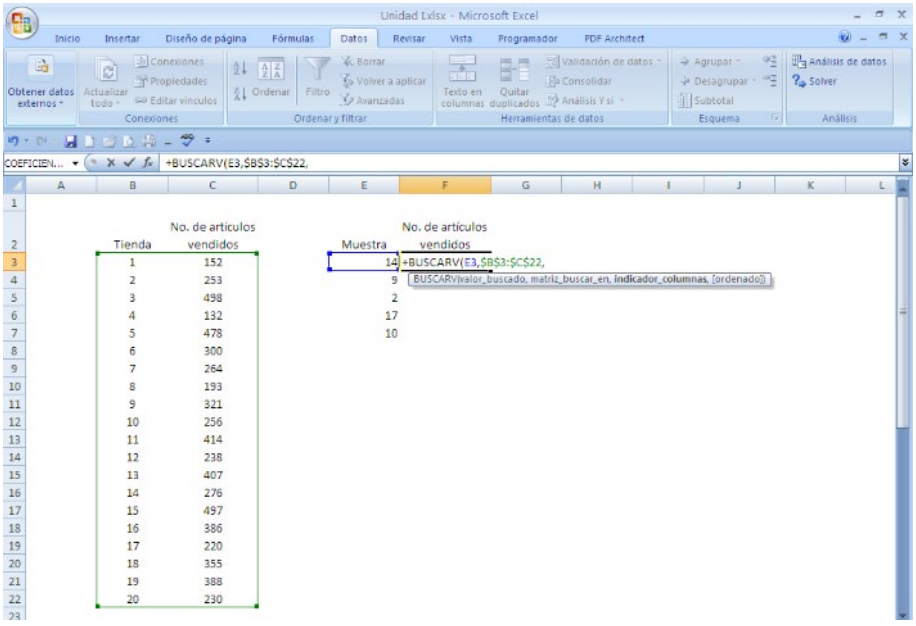
Este parámetro se refiere al rango donde se buscará la información.

En este caso, las dos columnas completas de Tienda y No. de artículos vendidos.[3]



Fuente: Microsoft Excel (2013).

Escogidas las columnas, fijar el rango oprimiendo una vez la tecla F4. Aparecerán signos de $ que indican que ya está fija la matriz.

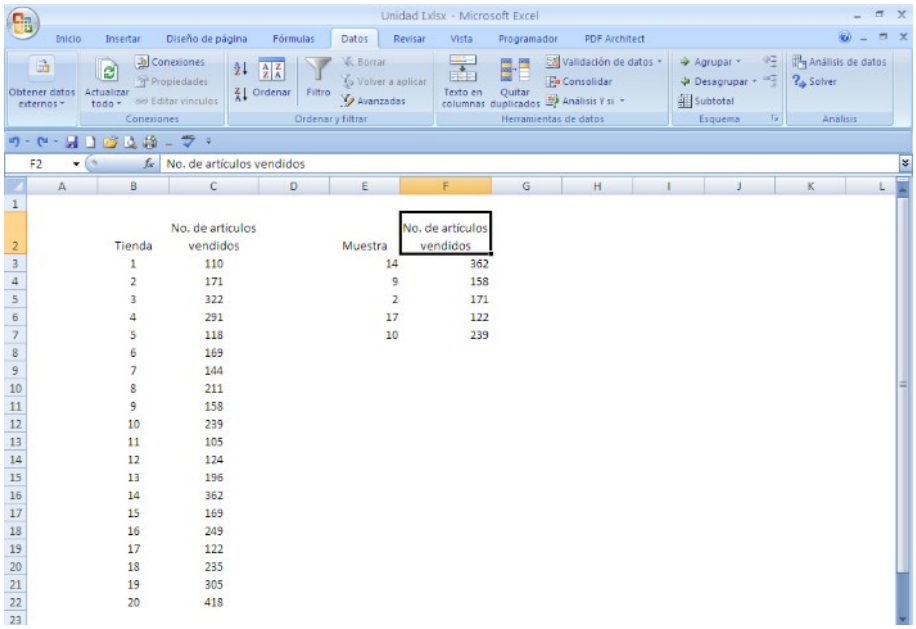


Fuente: Microsoft Excel (2013).

[3] En este rango de búsqueda, la primera columna debe tener los valores buscados; de lo contrario, no trabajará correctamente la función.

Indicador de columnas	Ordenados
• En este parámetro, se presenta el número de columna del rango de búsqueda donde se encuentra la información que se desea agregar. En este ejemplo, la información que se quiere agregar es el número de unidades vendidas que se halla en la columna 2 del rango de búsqueda.	• Es un valor lógico. Si se escribe 0 (cero), se está indicando que se requieren valores de búsqueda coincidentes. Si se pone 1, significa que los valores de búsqueda pueden ser parecidos.

Completados los parámetros de la función, oprimir la tecla Intro, y automáticamente aparecerán las ventas de cada una de las tiendas. Por ejemplo, la tienda 14 tiene 362 artículos vendidos.



Tienda	No. de artículos vendidos
1	110
2	171
3	322
4	291
5	118
6	169
7	144
8	211
9	158
10	239
11	105
12	124
13	196
14	362
15	169
16	249
17	122
18	235
19	305
20	418

Muestra	No. de artículos vendidos
14	362
9	158
2	171
17	122
10	239

Fuente: Microsoft Excel (2013).

Uso de números aleatorios en MS-Excel

También se puede extraer una muestra generando números aleatorios. Un número aleatorio es una cifra producida al azar a través de un algoritmo interno y que tiene la misma probabilidad de ser elegido respecto a otro número. Excel permite seleccionar números aleatorios enteros entre un rango de valores con la siguiente función:

ALEATORIO.ENTRE (inferior,superior)	*Inferior*	Es el valor mínimo aleatorio permitido
	Superior	Es el valor máximo aleatorio permitido

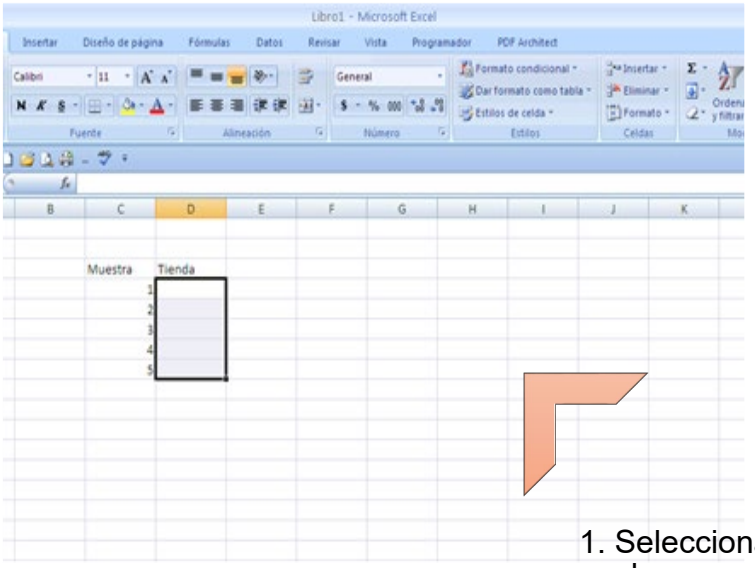
Supóngase que se desea obtener un número aleatorio entre 1 y 10. Aplicando la función ALEATORIO.ENTRE(inferior,superior), se tiene:

ALEATORIO.ENTRE(1,10)

En este ejemplo, al presionar la tecla Intro, se generó el número 7.[4]

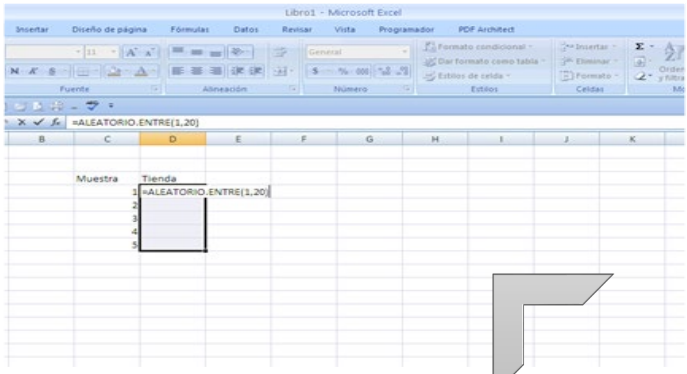
Regresando al ejemplo de las 20 tiendas, supóngase que se desea determinar las tiendas que serán auditadas utilizando números aleatorios. Se procederá de la siguiente manera.

[4] Si se volviera a presionar la tecla Intro, se generaría otro número aleatorio.



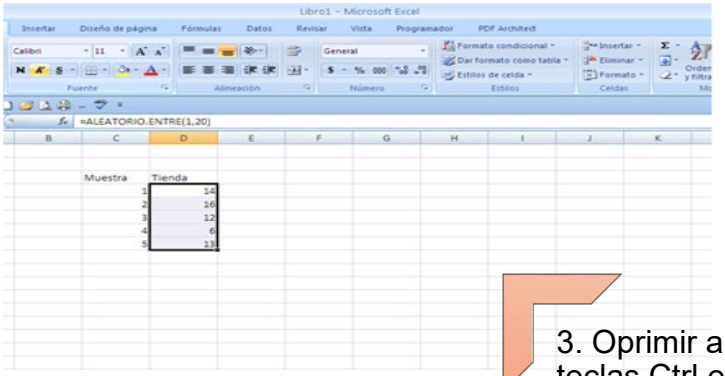
1. Seleccionar toda el área en la cual se generarán los números aleatorios.

Fuente: Microsoft Excel (2013).



2. Escribir la función Aleatorio.Entre, utilizando un rango de 1 a 20.

Fuente: Microsoft Excel (2013).



3. Oprimir al mismo tiempo las teclas Ctrl e Intro. Se generarán los números aleatorios.

- Los datos conservan la fórmula. Por ello se recomienda copiar y pegar los datos como valores (pegar – pegado especial – valores) para que no cambien cada vez que se realice una acción.

Fuente: Microsoft Excel (2013).

Para efectos de este ejemplo, las tiendas 6, 12, 13, 14 y 16 son las elegidas para auditarlas (el resultado no necesariamente debe ser el mismo si se replica el ejercicio, debido a que se eligen números aleatorios). De esta manera, se obtiene una muestra empleando números aleatorios.

Si se quisiera generar un número aleatorio entre 0 y 1, hay que hacerlo con la función ALEATORIO(). Esta función no cuenta con parámetros después de escribir su nombre; solamente se abre y cierra paréntesis, y al dar Intro se genera un número entre 0 y 1.

RESUMEN

Las metodologías empleadas en estadística inferencial tienen como insumo la información recabada de una muestra, por ello su obtención cobra relevancia, pues la manera de hacerlo garantizará la validez de los resultados.

Esta unidad ha presentado una introducción al muestreo. En primer lugar, se abordaron tres conceptos que se utilizarán a lo largo del curso: parámetros, estadísticos y estimadores.

La estadística inferencial busca determinar el valor o comportamiento de parámetros poblacionales con el empleo de estimadores y estadísticos aplicados con información de una muestra.

Si se requiere estimar el valor de un parámetro, se emplean estimadores; y cuando se busca contrastar hipótesis sobre el comportamiento de algún parámetro poblacional, se recurre a pruebas de hipótesis.

Se estudió también el tipo de muestras que puede utilizarse en un estudio, ya sea por:

El método empleado (aleatorio o juicio), → número de muestras (únicas o múltiples) → o su independencia (independientes o relacionadas).

Además se expuso de manera breve las características de tipos de muestreo aleatorio (aleatorio simple, sistemático, estratificado y de conglomerados) y se explicó la manera de calcular tamaños de muestra para un muestreo aleatorio simple asumiendo una fracción de muestreo ($\frac{n}{N}$) sin importancia.

Al final, se planteó un ejemplo de cómo utilizar Microsoft Excel (2013) para obtener muestras tanto con el módulo de análisis de datos como con números aleatorios.

BIBLIOGRAFÍA

SUGERIDA

Autor	Capítulo	Páginas
Anderson, S. (2012)	7	265-272
Levin, R. (2010)	6	236-250
Lind, D. (2012)	8	266-274

UNIDAD 2

Distribuciones muestrales

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

OBJETIVO PARTICULAR

Al terminar la unidad, el alumno identificará e interpretará los diferentes tipos de distribuciones muestrales.

TEMARIO DETALLADO

(8 horas)

2. Distribuciones muestrales

2.1. La distribución muestral de la media
2.2. El teorema central del límite
2.3. La distribución muestral de la proporción
2.4. La distribución muestral de la varianza

INTRODUCCIÓN

El insumo de la estadística tanto descriptiva como inferencial es la información, por lo que la obtención de la muestra juega un papel central en la validez de los resultados. En estadística inferencial, con los valores recabados en una muestra se puede deducir el valor de un parámetro de interés, lo que permitirá determinar el comportamiento de una población.

Al trabajar con muestras, los parámetros presentan comportamientos que se aproximan a distribuciones teóricas de probabilidad. Esto permite evaluar la congruencia de los resultados y la calidad de las inferencias a realizar.

En esta unidad, se expondrán algunas distribuciones muestrales que serán utilizadas en el resto del curso. Primero, la distribución normal y t de Student, asociadas a medias o proporciones; y al final de la unidad, la χ^2 (ji – cuadrada) y *F*, asociadas con varianzas.

En la parte intermedia de la unidad, se destina una sección para exponer uno de los resultados más importantes de la teoría de la probabilidad: el teorema del límite central, el cual garantiza que un promedio muestral tiene una distribución que se aproxima a una normal conforme aumenta el tamaño de la muestra.

2.1. La distribución muestral de la media

Durante el curso de Estadística Descriptiva, en la sección dedicada a probabilidad, se abordaron las variables aleatorias.

Variable aleatoria

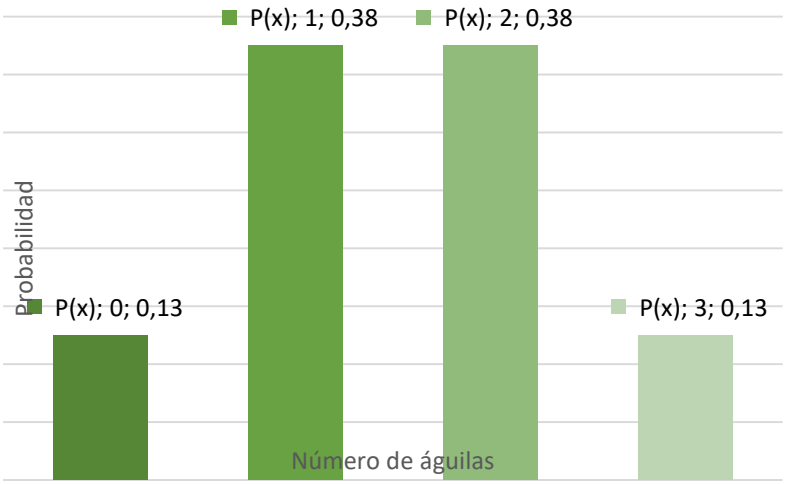
Una variable aleatoria es una función que mapea los elementos del espacio muestral al conjunto de los números reales; es decir, una variable aleatoria representa de forma numérica todos los resultados posibles de un experimento.

Asimismo, cada valor de la variable aleatoria tiene asociada una probabilidad de ocurrencia, que en conjunto conforman la distribución de probabilidades o simplemente la distribución de la variable aleatoria.

Para ejemplificar lo anterior, supóngase que se tiene el siguiente experimento: número de águilas que se observan en tres lanzamientos de una moneda de diez pesos. El espacio muestral de este experimento lo conforman $2^3 = 8$ eventos que son AAA, AAS, ASA, SAA, ASS, SAS, SSA y SSS: *A* representa un resultado de águila; y *S*, de sol.

El número de águilas que pueden aparecer en tres lanzamientos son 0, 1, 2 o 3, por lo que la variable aleatoria *X* asociada al experimento toma estos valores. La probabilidad de ocurrencia de cada valor de la variable aleatoria es 1/8 para X = 0 y X = 3; 3/8 para X = 1 y X = 2. La distribución de *X* se muestra en la siguiente figura.

Figura 1. Distribución de probabilidades de la variable aleatoria asociada al número de águilas observadas en tres lanzamientos de una moneda de diez pesos



Fuente: elaboración propia.

Es habitual que de una muestra aleatoria de tamaño *n* se calcule el promedio con los valores extraídos, donde el resultado dependerá de la muestra:

el promedio muestral es una variable aleatoria que cuenta con una distribución de probabilidades.

Supóngase que al área de planeación de cierta organización la conforman cinco empleados, los cuales cuentan con la siguiente antigüedad en el trabajo.

Tabla 1. Antigüedad de los empleados del área de planeación en la organización

Empleado	Antigüedad en años
1	7
2	3
3	4
4	5
5	2

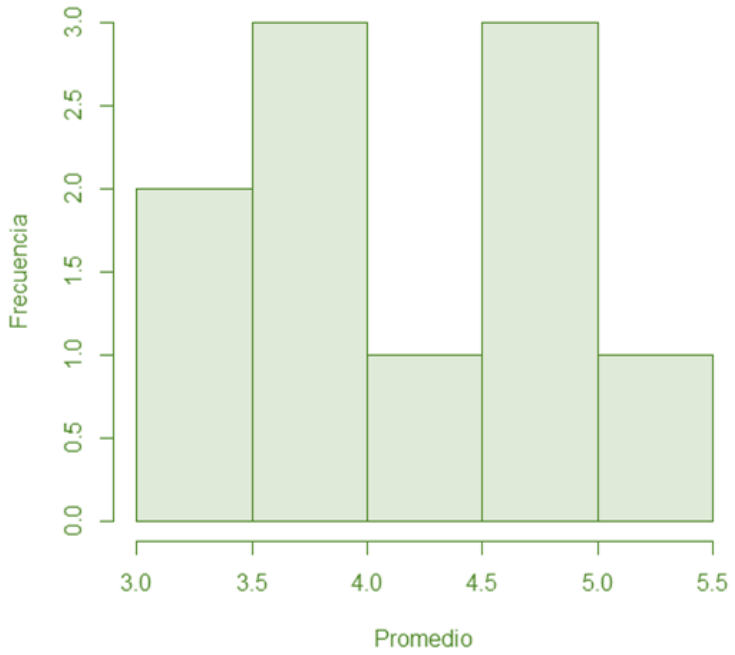
Si se extrae una muestra de tres empleados (sin reemplazo) y se calcula su promedio de antigüedad, hay $\binom{5}{3} = 10$ posibles resultados, los cuales se detallan en la tabla 2.

Tabla 2. Valores posibles del promedio de antigüedad de una muestra de dos empleados del área de planeación

Muestra	Empleados en la muestra	Promedio de antigüedad
1	1,2,3	$\frac{7+3+4}{3} = 4.7$
2	1,2,4	$\frac{7+3+5}{3} = 5.0$
3	1,2,5	$\frac{7+3+2}{3} = 4.0$
4	1,3,4	$\frac{7+4+5}{3} = 5.3$
5	1,3,5	$\frac{7+4+2}{3} = 4.3$
6	1,4,5	$\frac{7+5+2}{3} = 4.7$
7	2,3,4	$\frac{3+4+5}{3} = 4.0$
8	2,3,5	$\frac{3+4+2}{3} = 3.0$
9	2,4,5	$\frac{3+5+2}{3} = 3.3$
10	3,4,5	$\frac{4+5+2}{3} = 3.7$

En cuanto a la distribución de frecuencias, se muestra en la figura 2.

Figura 2. Distribución de frecuencias de los promedios de antigüedad de una muestra de tres empleados del área de planeación



Fuente: elaboración propia.

En la figura anterior, se muestra la distribución de frecuencias de los posibles promedios. Obsérvese que es más factible tener un resultado entre 3.5 y 4.0 o entre 4.5 y 5.0.

La distribución de todos los promedios posibles de una muestra de tamaño *n* se conoce como *distribución muestral de la media.*

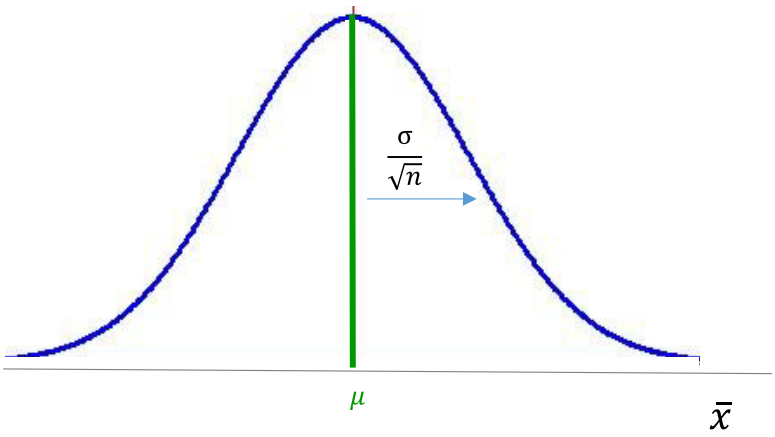
En el ejemplo anterior, la distribución muestral de la media es bimodal, lo que se debe a la poca información y dispersión de datos. ¿Si la población hubiera sido de mayor tamaño o la muestra hubiera permitido repeticiones, la distribución se habría conservado? La respuesta es no.

En la siguiente sección, se analizará un resultado que garantiza que la distribución muestral de la media se aproxima a una distribución normal conforme se incrementa el tamaño de la muestra. Por lo pronto, solamente se hará mención de este resultado.

Distribución muestral de la media

Supóngase que se tiene una población de tamaño *N* con media μ y varianza σ^2 de la que se extrae una muestra de tamaño *n*. La distribución de la media muestral ($\bar{x}$) se aproxima a una normal con media μ y varianza σ^2/n (figura3) en la medida que se incrementa el tamaño de la muestra (n).[5]

Figura 3. Distribución muestral de la media



Fuente: elaboración propia.

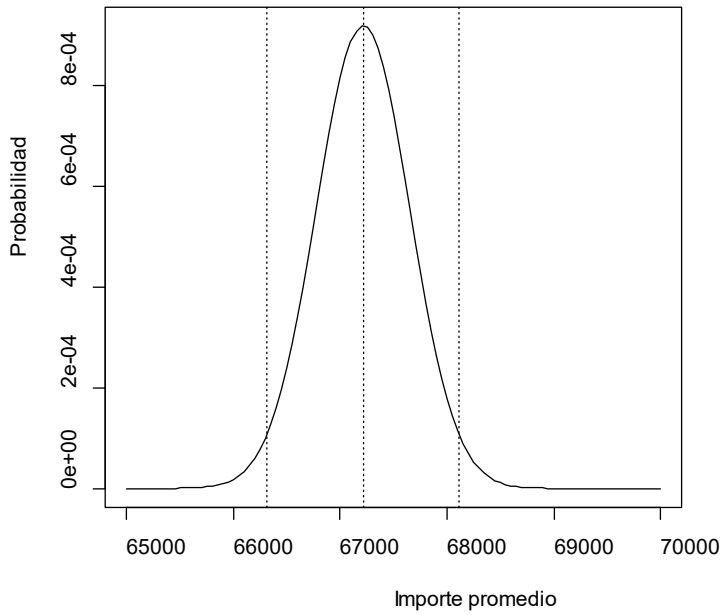
Conociendo lo anterior, puede estandarizarse esta distribución y utilizar el cálculo de una probabilidad para medir la calidad de la muestra, lo cual se ejemplifica a continuación.

[5]Cuando la fracción $\frac{n}{N} > 0.05$ se multiplica por el factor de ajuste $\sqrt{\frac{N-n}{N-1}}$

Supóngase que una organización realizó 8620 movimientos bancarios durante el último ejercicio fiscal, con un importe promedio de $67,213.49 y una desviación de $5,315.22. Se contrató un despacho de auditores para validar estas operaciones. Ante la premura con la que se requieren los resultados, se determinó auditar una muestra de 150 movimientos. Se considera que los resultados son satisfactorios si el promedio muestral difiere del real en $900. Entonces, ¿cuál es la probabilidad de que el promedio muestral difiera del real $900?

Conforme a lo expuesto, la distribución muestral del promedio se aproxima a una distribución normal con media de $67,213.49 y una desviación de $\frac{\$5{,}315.22}{\sqrt{150}}$. Se busca la probabilidad de que el promedio muestral se encuentre entre $67,213.49 ± $900. En la figura 4 se muestra la región de interés.

Figura 4. Distribución del promedio muestral de los movimientos bancarios



Fuente: elaboración propia.

La figura anterior presenta la distribución de todos los promedios obtenidos con muestras de 150 movimientos bancarios. La línea al centro de la distribución es el promedio real y las otras dos líneas verticales alrededor del promedio real limitan la región de los resultados considerados satisfactorios ($66,313.49 y $68,113.49).

Para calcular la probabilidad, se procede a estandarizar los valores para trabajar con una distribución normal con media cero y desviación estándar uno (Z).

De esta manera:

$$\boldsymbol{P(66,313.49 < X < 68,113.49)}$$

$$\boldsymbol{P\left(\frac{66,313.49 - 67,213.49}{\frac{5,315.22}{\sqrt{150}}} < \frac{X - 67,213.49}{\frac{5,315.22}{\sqrt{150}}} < \frac{68,113.49 - 67,213.49}{\frac{5,315.22}{\sqrt{150}}}\right)}$$

$$\boldsymbol{P(-2.073 < Z < 2.073)}$$

Para calcular esta probabilidad, se utilizará la probabilidad acumulada hasta 2.073 y se restará la acumulada a –2.073. Se aplicará la siguiente función de Excel: DISTR.NORM.ESTAND(z), donde *z* es el cuantil de la distribución normal estándar en donde se desea calcular la probabilidad acumulada.

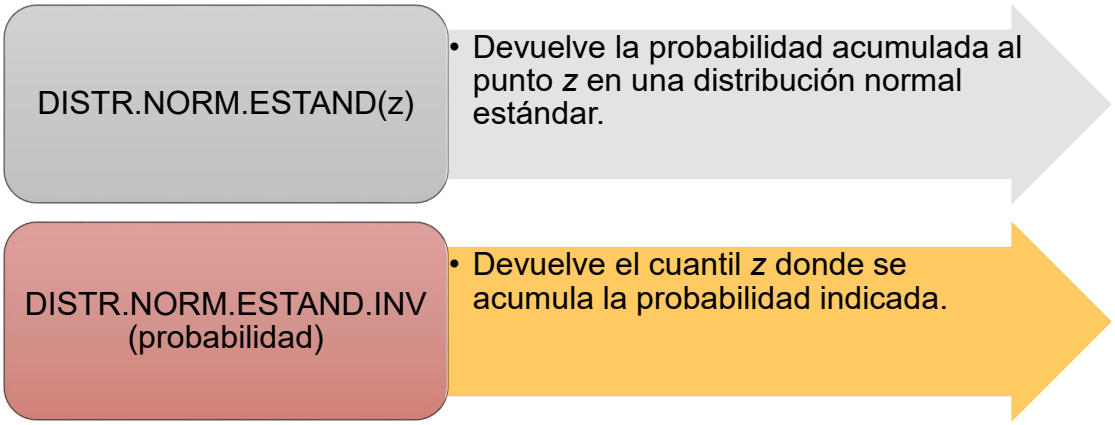
Entonces, la probabilidad buscada se calcula así:

DISTR.NORM.ESTAND(2.073) – DISTR.NORM.ESTAND(–2.073)
= 0.9809 – 0.0191 = 0.9618

Este resultado indica que la probabilidad de que la muestra proporcione un resultado satisfactorio es de 0.9618: los resultados de la muestra son confiables.

Observación

Al trabajar una distribución normal estandarizada en Excel, se pueden utilizar las siguientes funciones:



Distribución muestral de la media cuando se desconoce σ^2

Aunque resulta sencillo determinar la distribución muestral de la media cuando se tiene la varianza o la desviación estándar poblacional, no siempre es posible conocerla. Al presentarse esta situación, se utilizan los valores de la muestra para estimarla de la siguiente manera:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- Donde:

s^2 = varianza muestral
x_i = valor del i-ésimo elemento de la muestra
$\overline{x}$ = promedio muestral
N = tamaño de la muestra

Y la distribución muestral de la media no es una normal, sino una t de Student con n – 1 grados de libertad.

La distribución t de Student es también una distribución acampanada alrededor de cero. A diferencia de una distribución normal estándar (Z), sus extremos tardan en tomar una forma asintótica, por lo que se dice que es “pesada en las colas”.

La distribución t de Student depende de un parámetro conocido como *grados de libertad*. La distribución t de Student es única para cada grado de libertad y conforme aumenta se aproxima más a una distribución normal estándar.

Los grados de libertad se refieren al número de valores independientes en el cálculo de la varianza muestral. Como se sabe que la suma de las desviaciones alrededor de la media es cero, se necesita conocer n – 1 valores para determinar el restante.

Con tamaños de muestra grandes (n>30), la distribución t de Student se comporta similar a una normal estandarizada, debido a lo cual se sugiere su uso en muestras de tamaño menor a 30.

Función de densidad de la distribución t de Student:

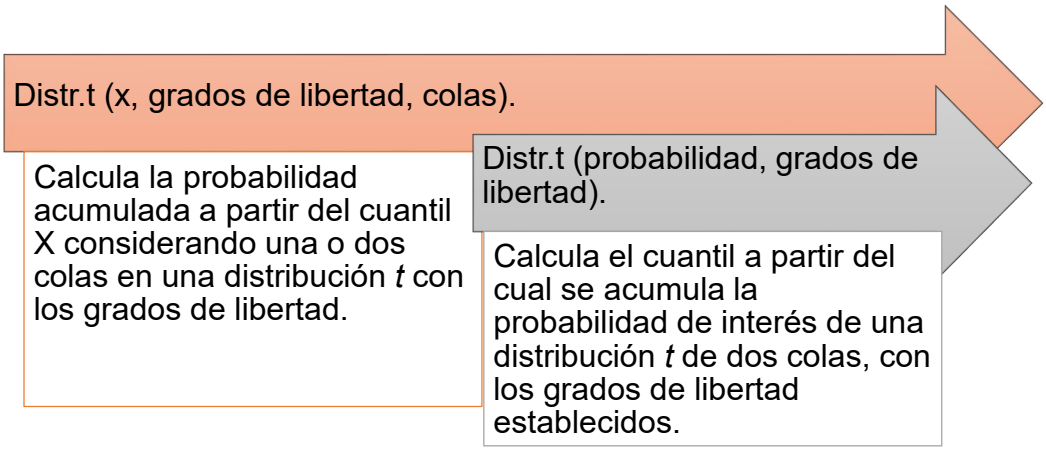
$$t_n = \frac{1}{\sqrt{n\pi}} \cdot \frac{\Gamma\frac{(n+1)}{2}}{\Gamma\frac{n}{2}} \cdot \left(1 + \frac{x^2}{n}\right)^{\left(-\frac{n+1}{2}\right)}$$

Para $x \in (-\infty, \infty)$

- Donde:

t_n = valor *t* con *n* grados de libertad
Γ = función gamma
N = grados de libertad

Cuando se trabaja con una distribución t en Excel, se utilizan las siguientes funciones:



Para ilustrar el uso de la distribución t de Student, supóngase que en el ejemplo anterior se desconoce el valor de la varianza poblacional, además el auditor decidió utilizar una muestra de cinco movimientos con los siguientes valores: $65,128, $69,310, $68,501, $66,920 y $67,821.

El primer paso es calcular el promedio muestral:

$$\overline{x} = \frac{65,128 + 69,310 + 68,501 + 66,920 + 67,821}{5} = 67,536$$

A continuación, se calcula la varianza muestral:

$$s^2 = \frac{(65,128 - 67,536)^2 + (69,310 - 67,536)^2 + (68,501 - 67,536) + (66,920 - 67,536)^2 + (67,821 - 67,536)^2}{5 - 1}$$

= 2,584,361.5

Por tanto, la desviación muestral es:

$$\sqrt{2,584,361.5} = 1,607.59$$

A continuación, se estandarizan los datos:

$$\boldsymbol{P(66,313.49 < X < 68,113.49)}$$

$$\boldsymbol{P\left(\frac{66,313.49 - 67,213.49}{\frac{1,607.59}{\sqrt{5}}} < \frac{X - 67,213.49}{\frac{1,607.59}{\sqrt{5}}} < \frac{68,113.49 - 67,213.49}{\frac{1,607.59}{\sqrt{5}}}\right)}$$

$$\boldsymbol{P(-1.252 < t_4 < 1.252)}$$

Para calcular esta probabilidad, se utilizará la probabilidad contenida entre –1.252 y 1.252, con la función de Excel Distr.t(x,grados de libertad, colas), explicada anteriormente.

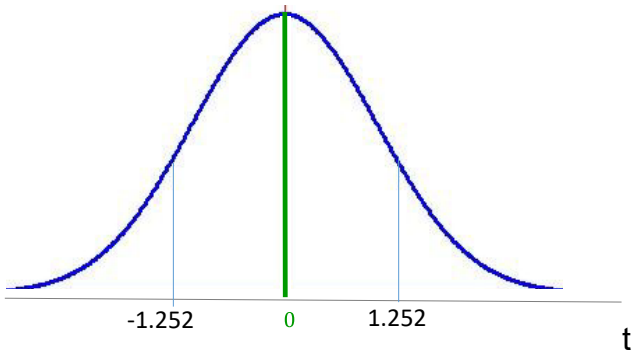
Entonces, la probabilidad buscada se calcula así:

(1-Distr.t(1.252,4, 2)**)** = 0.7212

Este resultado indica que la probabilidad de que la muestra proporcione un resultado satisfactorio es de 0.7212, por lo que es recomendable incrementar el tamaño de la muestra.

Observación:
La función Distr.t(1.252,4, 2)

Figura 5. Segmentación de la distribución t con cuatro grados de libertad considerada en el problema



Fuente: elaboración propia.

Calcula la probabilidad acumulada en las colas, es decir, la suma del área acumulada de menos infinito a –1.252, y desde 1.252 a infinito. Como la región de interés se encuentra entre –1.252 y 1.252, se utiliza el complemento.

2.2. El teorema central del límite

En la sección anterior, se mencionó que la distribución muestral de una media es una normal, pero ¿cuál es el sustento teórico de esta afirmación? En la teoría de probabilidad existen dos resultados muy importantes: la ley de los grandes números y el teorema del límite central, este último garantiza que el promedio de una muestra siga una distribución normal. A continuación, se expone este teorema.

Teorema del límite central

El teorema del límite central establece que, si se cuenta con un conjunto de variables aleatorias $X_1, X_2, \ldots, X_n$, las cuales son independientes e idénticamente distribuidas con valor esperado

$$E(X_1) = E(X_2) = \ldots = E(X_n) = \mu$$

y varianza

$$V(X_1) = V(X_2) = \ldots = V(X_n) = \sigma^2$$

entonces, a medida que se incrementa el número de variables (n),

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Donde:

$\bar{X}_n$ = Promedio de n variables

$N\left(\mu, \frac{\sigma^2}{n}\right)$ = Distribución normal con media μ y varianza σ^2/n

El resultado indica que la distribución del promedio del conjunto de variables se aproxima a una normal con media μ y varianza σ^2conforme el tamaño de la muestra se incrementa.

Este resultado es aplicable al muestreo, donde los elementos de la muestra pueden considerarse como variables aleatorias independientes con la misma distribución de la población de la que proceden con media μ y varianza σ^2. Así, el promedio muestral conforme el tamaño de la muestra se incrementa se aproxima a una distribución normal con media μ y varianza σ^2/n.

Para entender mejor este resultado, supóngase que de una población con media μ y varianza σ^2 se extraen *N* muestras aleatorias de tamaño *n* y con cada una se calcula el promedio. Si se construye un histograma con los *N* promedios, tendría una forma acampanada alrededor del punto μ y su varianza se aproxima a σ^2/ n.

Para ejemplificar lo anterior, supóngase que se desea conocer el comportamiento del promedio del lanzamiento de un dado. Asumiendo que el dado no se encuentra cargado en ningún número, cualquier valor tiene la misma probabilidad de ser elegido (1/6), por lo que el valor esperado (μ) es el siguiente:

$$\boldsymbol{\mu = E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5}$$

Y la varianza (σ^2):

$$\boldsymbol{\sigma^2 = E(X^2) - E^2(X)}$$

Donde:

$$\boldsymbol{E(X^2) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = 15.2}$$

Así:

$$\boldsymbol{\sigma^2 = E(X^2) - E^2(X) = 15.2 - 3.5^2 = 2.9}$$

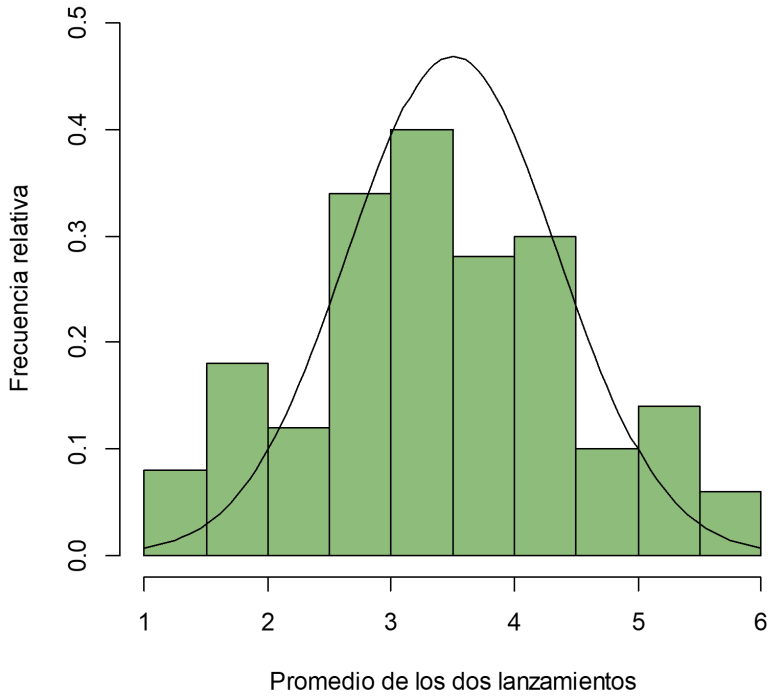
Supóngase que se lanza el dado dos veces (n = 2) y se calcula el promedio de los dos resultados y se repite este experimento 100 ocasiones (N = 100). Se obtienen los resultados que se muestran en la tabla siguiente.

Tabla 3. Resultados de dos lanzamientos de un dado en 100 ocasiones

	Lanzamiento				Lanzamiento				Lanzamiento				Lanzamiento			
Muestra	1	2	Promedio	Muestra	1	2	Promedio	Muestra	1	2	Promedio	Muestra	1	2	Promedio
1	2	4	3	26	5	6	5.5	51	4	3	3.5	76	5	4	4.5
2	6	3	4.5	27	6	3	4.5	52	6	5	5.5	77	2	6	4
3	6	6	6	28	6	5	5.5	53	3	1	2	78	4	2	3
4	6	3	4.5	29	5	1	3	54	3	6	4.5	79	3	5	4
5	5	2	3.5	30	5	6	5.5	55	5	4	4.5	80	1	6	3.5
6	2	4	3	31	2	1	1.5	56	2	4	3	81	6	2	4
7	5	2	3.5	32	2	2	2	57	4	6	5	82	4	3	3.5
8	4	2	3	33	1	1	1	58	5	2	3.5	83	5	6	5.5
9	3	6	4.5	34	5	5	5	59	2	3	2.5	84	3	3	3
10	2	4	3	35	4	3	3.5	60	4	1	2.5	85	1	6	3.5
11	1	3	2	36	4	4	4	61	6	4	5	86	4	2	3
12	2	6	4	37	5	1	3	62	2	2	2	87	4	5	4.5
13	3	5	4	38	5	1	3	63	3	3	3	88	6	5	5.5
14	1	4	2.5	39	3	4	3.5	64	2	4	3	89	5	1	3
15	1	6	3.5	40	2	5	3.5	65	5	3	4	90	6	4	5
16	1	5	3	41	6	1	3.5	66	1	3	2	91	3	1	2
17	6	2	4	42	4	5	4.5	67	2	6	4	92	4	5	4.5
18	3	6	4.5	43	4	4	4	68	4	2	3	93	2	3	2.5
19	4	3	3.5	44	2	5	3.5	69	3	5	4	94	6	6	6
20	3	2	2.5	45	3	6	4.5	70	1	2	1.5	95	6	3	4.5
21	5	6	5.5	46	1	1	1	71	5	2	3.5	96	5	1	3
22	3	4	3.5	47	4	3	3.5	72	4	3	3.5	97	5	2	3.5
23	4	4	4	48	6	6	6	73	4	5	4.5	98	5	3	4
24	4	5	4.5	49	4	3	3.5	74	4	1	2.5	99	1	3	2
25	3	1	2	50	1	3	2	75	2	6	4	100	5	5	5
													Promedio:		3.6
													Varianza:		1.3

La tabla anterior muestra los resultados de las 100 muestras de dos lanzamientos y sus respectivos promedios. Obsérvese que el promedio de los promedios es 3.6 (cercano a 3.5, el valor esperado) y la varianza de los promedios (1.3), que se acerca a 2.9/2 = 1.45. La siguiente figura muestra el histograma de la distribución del promedio de dos lanzamientos junto con la distribución teórica a la que debería aproximarse.

Figura 6. Distribución del promedio de dos lanzamientos de un dado



Fuente: elaboración propia con empleo del paquete estadístico R.[6]

Se debe tomar en cuenta que el paquete estadístico donde se graficó la figura anterior muestra la frecuencia relativa modificada por un factor calculado por 10 entre el número de intervalos.

Ahora, supóngase que en vez de realizar dos lanzamientos se hicieran cinco, se calculara el promedio y se repitiera este experimento 100 ocasiones. En la siguiente tabla, se muestran los resultados.

[6] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org/

.

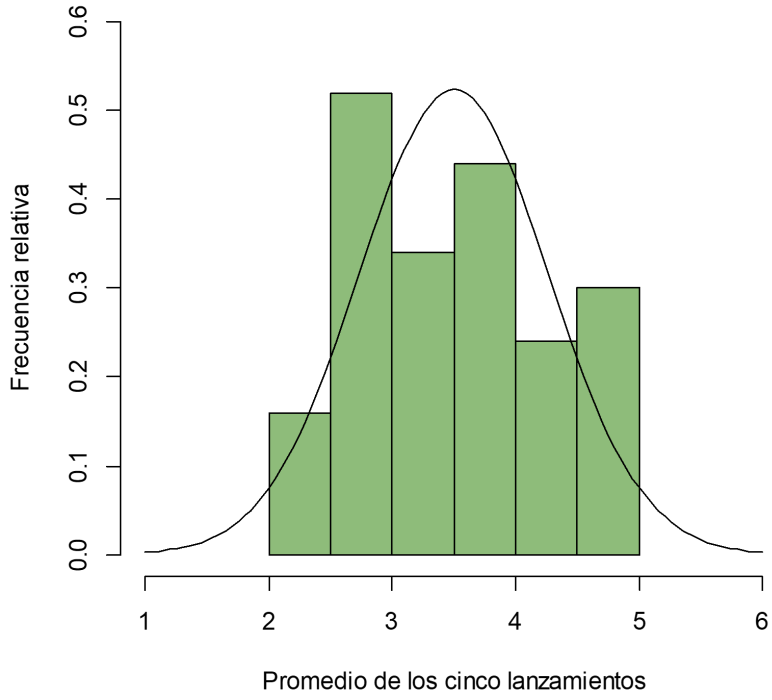
Tabla 4. Resultados de cinco lanzamientos de un dado en 100 ocasiones

	Lanzamiento					
Muestra	1	2	3	4	5	Promedio
1	3	3	5	2	3	3.2
2	4	4	3	2	5	3.6
3	1	1	5	2	6	3
4	1	5	6	6	3	4.2
5	3	2	3	2	3	2.6
6	5	4	4	5	5	4.6
7	3	6	5	1	2	3.4
8	5	6	3	4	6	4.8
9	3	3	2	2	5	3
10	3	3	3	3	4	3.2
11	3	4	5	2	1	3
12	1	5	4	4	3	3.4
13	3	2	2	5	3	3
14	2	5	6	1	1	3
15	1	6	1	1	5	2.8
16	2	3	3	2	5	3
17	2	1	3	1	6	2.6
18	6	5	2	6	3	4.4
19	1	5	5	3	5	3.8
20	3	3	1	4	2	2.6
21	4	6	4	5	1	4
22	5	1	4	4	1	3
23	6	3	5	4	1	3.8
24	5	1	5	4	6	4.2
25	2	4	5	3	1	3
26	1	5	6	5	6	4.6
27	1	3	4	3	5	3.2
28	6	5	3	6	2	4.4
29	4	6	4	5	4	4.6
30	5	6	2	4	6	4.6
31	6	6	2	3	2	3.8
32	4	6	5	4	2	4.2
33	2	3	1	4	6	3.2
34	4	3	2	5	2	3.2
35	2	2	5	1	3	2.6
36	2	6	5	1	1	3
37	4	4	2	4	4	3.6
38	6	1	1	3	2	2.6
39	4	4	6	2	3	3.8
40	5	1	1	4	5	3.2
41	1	3	2	4	1	2.2
42	6	1	2	5	2	3.2
43	6	3	3	4	6	4.4
44	6	5	1	4	2	3.6
45	4	4	6	6	5	5
46	3	5	1	2	4	3
47	5	3	6	2	6	4.4
48	6	4	4	4	2	4
49	4	2	6	6	2	4
50	3	5	6	6	4	4.8

	Lanzamiento					
Muestra	1	2	3	4	5	Promedio
51	1	4	6	2	1	2.8
52	5	5	1	1	2	2.8
53	4	3	5	1	2	3
54	5	4	4	1	6	4
55	6	1	4	1	4	3.2
56	5	3	5	2	2	3.4
57	2	6	5	2	6	4.2
58	3	1	6	3	3	3.2
59	4	4	3	5	6	4.4
60	2	1	4	2	3	2.4
61	1	6	4	1	3	3
62	3	6	6	4	4	4.6
63	5	1	1	2	3	2.4
64	1	3	2	1	5	2.4
65	6	1	6	1	4	3.6
66	5	6	1	5	1	3.6
67	2	4	3	5	5	3.8
68	3	4	2	6	4	3.8
69	3	1	6	3	3	3.2
70	4	4	6	6	4	4.8
71	2	4	4	2	1	2.6
72	6	5	6	3	4	4.8
73	2	6	5	6	6	5
74	5	3	2	2	3	3
75	1	5	5	2	3	3.2
76	6	2	6	4	5	4.6
77	5	1	6	3	3	3.6
78	5	5	1	4	1	3.2
79	5	5	2	1	5	3.6
80	3	3	1	2	3	2.4
81	2	5	2	5	6	4
82	2	4	6	5	6	4.6
83	1	6	3	1	4	3
84	6	2	6	2	5	4.2
85	1	1	2	6	1	2.2
86	2	5	5	1	1	2.8
87	3	2	5	2	1	2.6
88	2	3	2	3	6	3.2
89	3	1	1	6	1	2.4
90	4	6	4	3	6	4.6
91	1	1	2	2	5	2.2
92	3	6	6	1	6	4.4
93	5	1	1	5	6	3.6
94	4	1	1	6	6	3.6
95	1	1	3	5	5	3
96	6	5	4	1	4	4
97	6	3	5	4	5	4.6
98	3	3	6	6	4	4.4
99	5	3	2	6	1	3.4
100	1	4	4	6	3	3.6
					Promedio:	3.5
					Varianza:	0.6

En el caso de 100 muestras de tamaño cinco, el promedio de los promedios es 3.5, el valor esperado del lanzamiento de un dado; y la varianza de los promedios es 0.6, la cual es casi 2.9/5 = 0.58. La siguiente figura es la gráfica de la distribución de los promedios de las 100 muestras con la distribución teórica a la que debe aproximarse.

Figura 7. Distribución del promedio de cinco lanzamientos de un dado

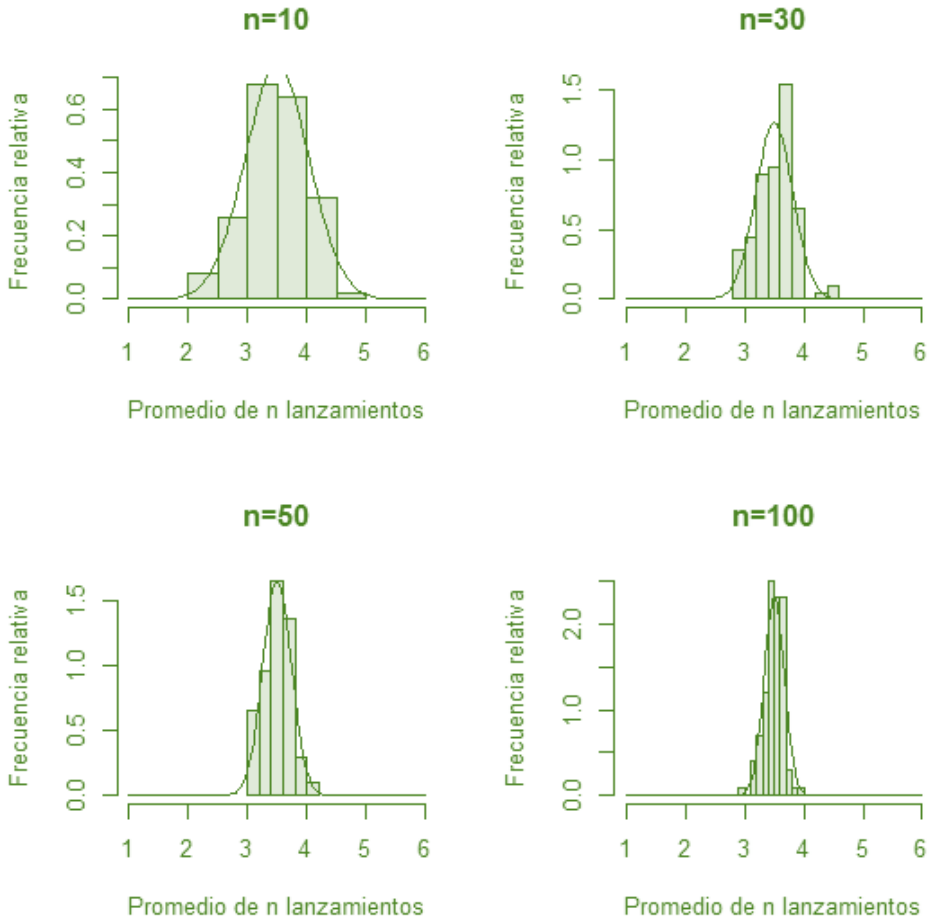


Fuente: elaboración propia con empleo del paquete estadístico R.

Obsérvese que la dispersión va disminuyendo: ahora el promedio se sitúa entre 2 y 5, y ya no incluye los valores extremos.

Conforme se incrementa el número de lanzamientos, la distribución de frecuencias se concentra cada vez más alrededor de 3.5 y se asemeja más a una distribución normal con media 3.5 y varianza 2.9/n. En la siguiente figura, se expone la distribución de frecuencias de 100 muestras de tamaño de 10, 30, 50 y 100 lanzamientos.

Figura 8. Distribución del promedio de cien muestras de 10, 30, 50 y 100 lanzamientos de un dado



Fuente: elaboración propia con empleo del paquete estadístico R.

De esta manera, se ha expuesto el teorema del límite central.

2.3. La distribución muestral de la proporción

Con frecuencia, la proporción poblacional *P* es uno de los parámetros que interesa conocer al extraer una muestra. Para hacerlo, se emplea la proporción muestral *p*, cuyo cálculo se realiza de la siguiente manera:

$$p = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Donde:

x_i = valor del i-ésimo elemento de la muestra
n = tamaño de la muestra

La proporción es un caso del promedio donde los valores que toman los elementos de la muestra son 1 si cumple con el criterio de interés, y 0 en caso contrario. De esta manera, cada elemento tiene una distribución Bernoulli con parámetro P y varianza P·(1 – P) debido a que los elementos de la muestra son independientes:

$$E(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} E(x_i) = \sum_{i=1}^{n} P = n \cdot P$$

y

$$V(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} V(x_i) = \sum_{i=1}^{n} P \cdot (1 - P) = n \cdot P \cdot (1 - P)$$

Que es el valor esperado y la varianza de una distribución binomial.

Con lo anterior:

$$\boldsymbol{E(p) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n} \cdot E(\sum_{i=1}^{n} x_i) = \frac{1}{n} \cdot \sum_{i=1}^{n} E(x_i) = \frac{1}{n} \cdot \sum_{i=1}^{n} P = \frac{1}{n} \cdot n \cdot P = P}$$

Y

$$\boldsymbol{V(p) = V\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n^2} \cdot V(\sum_{i=1}^{n} x_i) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} V(x_i) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} P \cdot (1 - P) = \frac{1}{n^2} \cdot n \cdot P \cdot (1 - P) = \frac{P \cdot (1 - P)}{n}}$$

Según la estadística descriptiva, si una variable *X* tiene una distribución binomial con parámetros *n* y *p*, entonces puede aproximarse a una normal con media *n*·*p* y varianza *n*·*p*·(1 – p) si $\boldsymbol{np \geq 5}$ y $\boldsymbol{n(1 - p) \geq 5}$.

Otro resultado importante, propiedad de la distribución normal, es que, si una variable *X* se distribuye como una normal con media μ y varianza σ^2 y si se define la variable Y como Y = a·X + b donde a y b son constantes, entonces *Y* tiene una distribución normal con media a·μ + b y varianza $a^2 \cdot \sigma^2$.

Aplicando los resultados anteriores, para *n* considerablemente grande la distribución de $\sum_{i=1}^{n} x_i$ se aproxima a una normal con media *n*·P y varianza *n*·P·(1 – P).
Si se define la siguiente variable Y = a·$\sum_{i=1}^{n} x_i$ + b, donde $\boldsymbol{a = \frac{1}{n}}$ y b=0, entonces:

$$\boldsymbol{Y = \frac{\sum_{i=1}^{n} x_i}{n} + 0 = p}$$

Tiene una distribución normal con media $\boldsymbol{\frac{1}{n} \cdot n \cdot P = P}$

y varianza $\boldsymbol{\frac{1}{n^2} \cdot n \cdot P \cdot (1 - P) = \frac{P \cdot (1 - P)}{n}}$

Observaciones

1. Cuando la proporción poblacional *P* es conocida y la población es finita con $\frac{n}{N} \leq 0.05$, la desviación de la proporción muestral será así:

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}}$$

Pero si $\frac{n}{N} > 0.05$, la desviación de la proporción muestral será ajustada de la siguiente manera:

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Donde *N* es el tamaño de la población y *n* el tamaño de muestra.

2. Cuando se desconoce la proporción poblacional *P*, se utiliza la proporción muestral. Si la población es finita con $\frac{n}{N} \leq 0.05$, la desviación de la proporción muestral será así:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n-1}}$$

Pero si $\frac{n}{N} > 0.05$, la desviación de la proporción muestral será ajustada de la siguiente manera:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n-1}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Donde *N* es el tamaño de la población y *n* el tamaño de muestra.

Para mostrar la utilidad de la distribución muestral de la proporción, se expone el siguiente ejemplo.

De acuerdo con una encuesta realizada a una población de 2919 egresados de licenciatura de la Facultad de Contaduría y Administración, el 80.4% considera excelentes o buenas las técnicas de enseñanza que utilizaron sus profesores durante la carrera[7]. Con la intención de conocer a mayor profundidad la metodología de enseñanza de sus docentes, la Dirección de la Facultad decide contactar a una muestra aleatoria de 100 egresados que contestaron la encuesta. ¿Cuál es la probabilidad de que el porcentaje de egresados en la muestra que juzgue excelentes o buenas las técnicas de enseñanza de sus profesores de licenciatura sea mayor a 90%?

Previo a establecer la distribución muestral de la proporción, se identifica que en este problema se está dando la proporción poblacional (80.4%) y el tamaño de la población (2,919) y de la muestra (100). Con esta información se puede calcular la fracción de muestreo ($\frac{n}{N}$), la cual es $\frac{100}{2,919} = \mathbf{0.03}$. En este caso, como es menor a 0.05, no es necesario realizar algún ajuste al cálculo de la desviación estándar de la proporción muestral.

De esta manera:

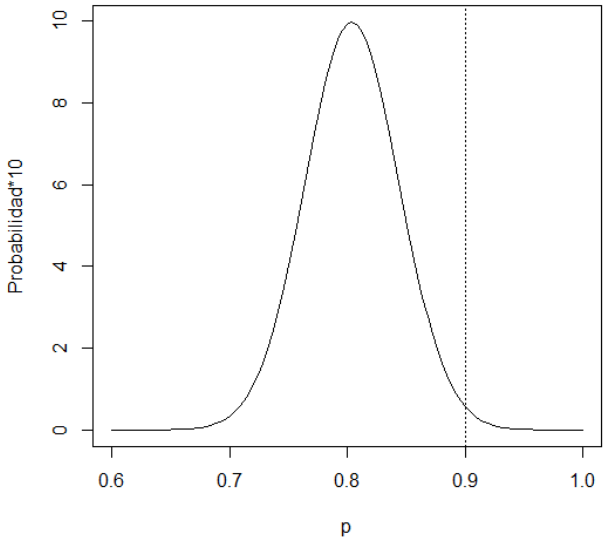
$$E(p) = P = 0.804$$

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.804(1-0.804)}{100}} = 0.04$$

Ahora, como n·P = (100)·(0.804) = 80.4 y n·(1 – P) = (100)·(1 – 0.804) = 19.6 son mayores a 5, entonces la distribución muestral de la proporción se aproxima a una normal con media 0.804 y desviación 0.04. (Véase figura 9).

[7]UNAM. Dirección General de Planeación. *Perfiles de alumnos egresados del nivel licenciatura de la UNAM 2012-2013,* p. 71. http://www.planeacion.unam.mx/Publicaciones/pdf/perfiles/egresados/p_eg2012-2013.pdf. Consultado el 13 de julio de 2015.

Figura 9. Distribución muestral de una proporción calculada con muestras de cien elementos



Fuente: elaboración propia con empleo del paquete estadístico R.

La figura anterior enseña la distribución muestral de la proporción para tamaños de muestra de 100 elementos. La región que se pide calcular se encuentra a la derecha de la línea punteada.

$$\boldsymbol{P(X > 0.9) = 1 - P(X \leq 0.9) = 1 - P\left(Z \leq \frac{0.9 - 0.804}{0.04}\right) = 1 - P(Z \leq 2.4)}$$

Utilizando la función de Excel DISTR.NORM.ESTAND(z), se obtiene:

$$\boldsymbol{1 - P(Z \leq 2.4) = 1 - 0.9918 = 0.0082}$$

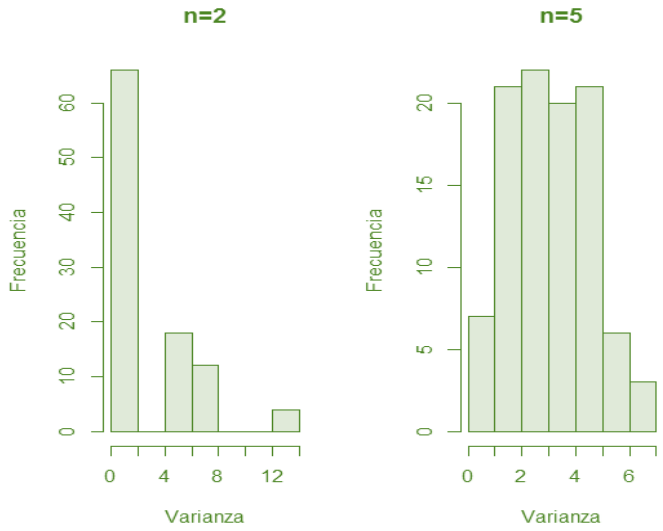
Este resultado indica que es prácticamente imposible tener en la muestra un porcentaje mayor a 90% de egresados que consideren excelentes o buenas las técnicas de enseñanza de sus profesores de licenciatura.

2.4. La distribución muestral de la varianza

En las secciones anteriores, se estudiaron las distribuciones muestrales de la media y de la proporción, dos parámetros que frecuentemente se desea conocer al extraer una muestra. Otro parámetro que también se busca identificar a través de un muestreo es la varianza, a partir de la cual se llega a la desviación estándar.

En el ejemplo del subtema 2.2, se plantearon lanzamientos de un dado para mostrar el comportamiento del promedio muestral, ¿cómo sería la distribución de la varianza de 100 muestras de dos y cinco lanzamientos? (Tablas 3 y 4). En este orden, la figura 10 presenta la distribución de frecuencias de las varianzas de las 100 muestras de dos y cinco lanzamientos.

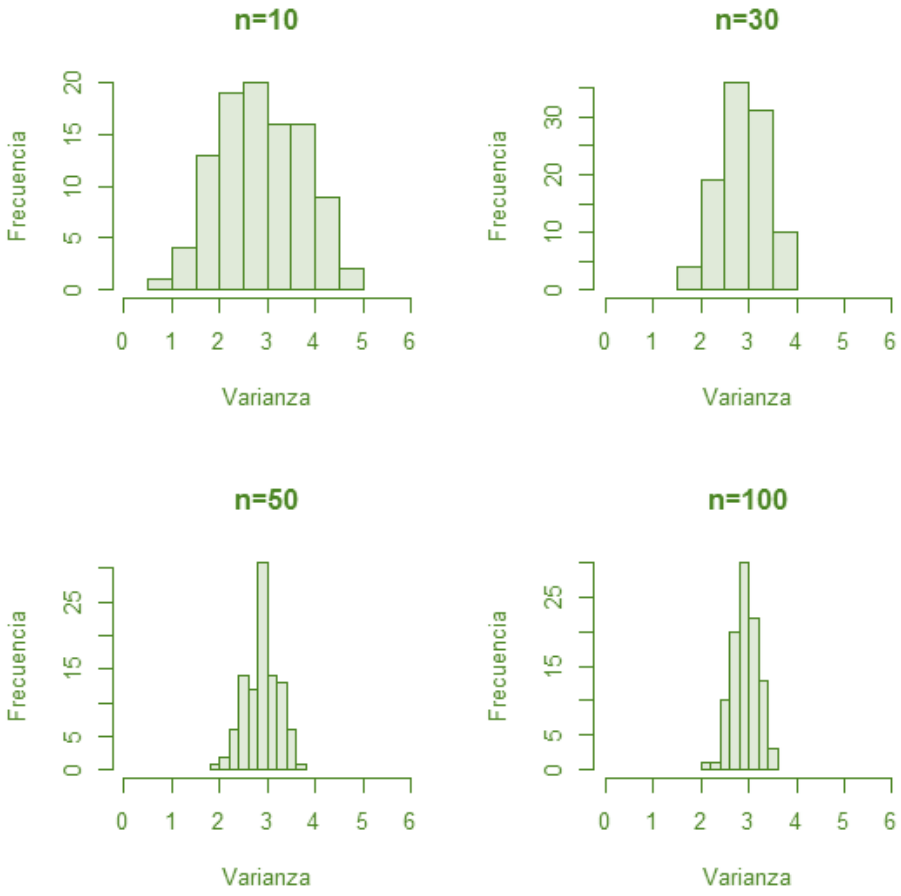
Figura 10. Distribución de frecuencias de las varianzas de dos y cinco lanzamientos de un dado



Fuente: elaboración propia con empleo del paquete estadístico R.

En la figura anterior, se expresan las distribuciones de las varianzas de dos y cinco lanzamientos, ambas sesgadas a la derecha. Obsérvese que con muestras de dos elementos la distribución de frecuencias de la varianza se asemeja a una exponencial, y al aumentar la muestra a cinco lanzamientos la distribución presenta una curvatura y menor variación. Si se aumentara la muestra a 10, 30, 50 y 100 lanzamientos, la varianza tendría el comportamiento que ilustra la figura 11.

Figura 11. Distribución de la varianza para muestras de 10, 30, 50 y 100 elementos



Fuente: elaboración propia con empleo del paquete estadístico R.

Nótese que, a medida que el tamaño de muestra se incrementa, la distribución de la varianza pierde su sesgo y tiene un comportamiento acampanado.

La distribución empleada para modelar la varianza muestral es χ^2(ji-cuadrada), cuya función de densidad es

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{\frac{-x}{2}}$$

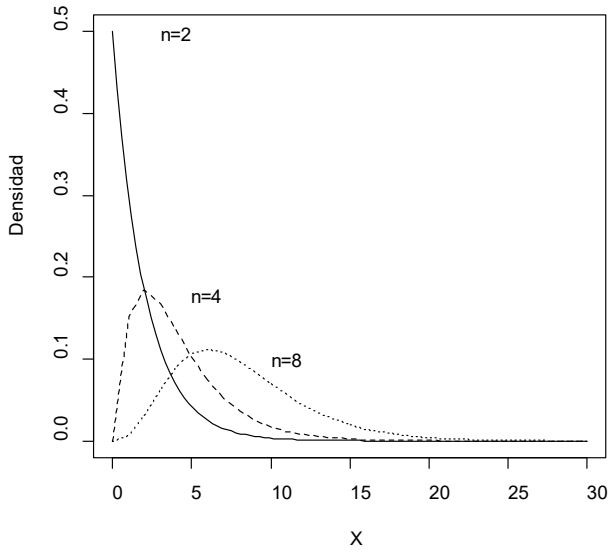
Para x >0

Donde *n* son los grados de libertad, que se definen de la misma forma como se hizo con la distribución t de Student.

Las características de esta distribución son las siguientes:

- Está definida para valores positivos.
- Es sesgada a la derecha.
- La forma de la distribución varía de acuerdo con los grados de libertad.
- Cuando n > 2, la media de la distribución es *n* y la varianza es 2n.
- El valor modal de la distribución se observa en n – 2.

Figura 12. Ejemplo del comportamiento de una distribución χ^2 con 2, 4 y 8 grados de libertad



Fuente: elaboración propia.

En la figura anterior, se distingue que, conforme aumentan los grados de libertad, la distribución tiende a aplanarse y el sesgo disminuye.

Resultados importantes

Al trabajar con esta distribución, se deben considerar los siguientes resultados importantes:

El valor esperado de la varianza muestral $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ es σ^2.

Si de una población normal se toma una muestra aleatoria simple de tamaño *n*, la variable $x^2 = \frac{(n-1)s^2}{\sigma^2}$, tendrá una distribución ji cuadrada con n – 1 grados de libertad.

Si una variable aleatoria *X* tiene una distribución χ^2 con *n* grados de libertad, entonces, con una *n* suficientemente grande, la variable aleatoria $\sqrt{2X}$ se aproxima a tener una distribución normal con media $\sqrt{2n-1}$ y varianza 1.

Funciones en Excel para trabajar la distribución χ^2

Excel dispone de las siguientes funciones para trabajar con la distribución:

Distr.chi(x,grados_de_libertad).	Calcula la probabilidad que se acumula en una distribución x^2 con los grados de libertad establecidos a partir del punto *x*.
Prueba.chi.inv(probabilidad, grados de libertad).	Calcula el cuantil a partir del cual se acumula la probabilidad buscada en una distribución x^2 con los grados de libertad establecidos a partir del punto *x*.

Para ejemplificar el uso de la distribución, supóngase que las transacciones bancarias de una organización en el último ejercicio fiscal se distribuyen como una distribución normal con una desviación estándar de $8,500. Si se elige al azar una muestra de 15 transacciones a fin de auditar al departamento responsable, ¿cuál es la probabilidad de que la desviación muestral exceda a la poblacional?

Para resolver el problema, se requiere calcular

$$P(s > \sigma) = P(s^2 > \sigma^2) = P(\frac{s^2}{\sigma^2} > 1) = P((n-1)\cdot\frac{s^2}{\sigma^2} > n-1)$$

Como la variable $\frac{(n-1)s^2}{\sigma^2}$ tiene una distribución χ^2 con n – 1 grados de libertad, entonces, la región que se está solicitando se encuentra a la derecha del valor esperado, es decir, se requiere calcular P(X > 14). Utilizando la función de Excel Distr.chi (14,14) = 0.4497, se calcula la probabilidad solicitada. Este resultado indica que es más probable que la variabilidad muestral sea menor a la poblacional.

En caso de no conocerse la varianza poblacional, el problema se resuelve de la misma manera.

Distribución para comparar dos varianzas

En este curso de estadística inferencial, a veces será necesario comparar la variabilidad de dos muestras, por lo que se empleará la distribución conocida como F, la cual tiene la siguiente función de densidad:

$$f(x) = \frac{\Gamma(\frac{n+d}{2})}{\Gamma(\frac{n}{2}) \cdot \Gamma(\frac{d}{2})} \cdot (\frac{n}{d})^{\frac{n}{2}} \cdot \frac{x^{\frac{n}{2}-1}}{(1+\frac{n}{d}x)^{\frac{n+d}{2}}}$$

Para $x > 0$

Donde *n* y *d* son los grados de libertad de cada una de las muestras a comparar.

Características de la distribución F:

- Es una distribución continua.
- Está definida para valores positivos.
- Tiene un sesgo positivo.
- Es asintótica.

Funciones en Excel para trabajar la distribución F

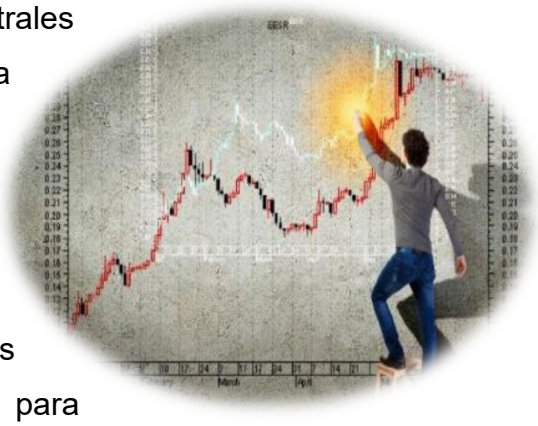
Excel tiene las siguientes funciones para trabajar con la distribución:

Distr.f(x,grados de libertad, grados de libertad2)	• Calcula la probabilidad que se acumula en una distribución F con los grados de libertad de cada muestra a partir del punto *x*.
Distr.f.inv(probabilidad, grados de libertad)	• Calcula el cuantil a partir del cual se acumula la probabilidad buscada en una distribución F con los grados de libertad de cada muestra a partir del punto *x*.

En la unidad 4, se mostrará con mayor detenimiento el empleo de la distribución F.

RESUMEN

Se analizó la importancia del muestreo para inferir sobre un parámetro de la población de interés. Al obtener una muestra aleatoria, se busca conocer los valores de los parámetros poblacionales por medio de los valores que arroja la muestra. Los parámetros muestrales son variables aleatorias porque dependen de los valores de los elementos en la muestra, por lo que resulta necesario identificar sus distribuciones para medir la calidad de los resultados.



También se expusieron las distribuciones muestrales principales para inferir sobre el promedio, una proporción y la varianza poblacional. Los dos primeros siguen una distribución normal y la varianza muestral puede modelarse con una distribución ji cuadrada. Además se mencionaron de forma general las características de la distribución F, la cual se empleará para comparar dos varianzas.

De igual manera, se explicó el teorema del límite central utilizando como ejemplo el lanzamiento de un dado, lo que garantiza que la distribución muestral del promedio se acerca a una normal conforme la muestra se incrementa.

Como valor agregado, se presentaron las funciones de Excel para trabajar con las distribuciones muestrales del promedio, de una proporción y de la varianza, que se aplicarán en las siguientes unidades.

BIBLIOGRAFÍA

SUGERIDA

Autor	Capítulo	Páginas
Anderson, S.	7	265-307
Levin, R.	6	247-272
Lind, D.	8	275-296

Licenciatura: **Administración**

UNIDAD 3

Estimación de parámetros

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 **2016** actualizado

OBJETIVO PARTICULAR

Al terminar la unidad, el alumno aprenderá los métodos de estimación de parámetros y su interpretación.

TEMARIO DETALLADO

(10 horas)

3. Estimación de parámetros

- 3.1. Estimaciones por punto y estimaciones por intervalo
- 3.2. Error de muestreo y errores que no son de muestreo
- 3.3. Propiedades de los estimadores
- 3.4. Estimación de una media con muestras grandes
  - 3.4.1. Determinación del tamaño de muestra necesario para estimar una media
- 3.5. Estimación de una media con muestras pequeñas
- 3.6. Estimación de una proporción
  - 3.6.1. Determinación del tamaño de muestra para estimar una proporción
- 3.7. Otros intervalos de confianza

INTRODUCCIÓN

Con frecuencia, las organizaciones requieren tener indicios del comportamiento de cierta variable de interés. Por ejemplo, el área de mercadotecnia de un banco pudiera estar interesada en conocer qué proporción de tarjetahabientes del producto *premium* responden a una promoción relacionada con un viaje a pagar en plazos sin intereses. O una organización no gubernamental dedicada a implementar programas para mejorar la nutrición de los niños entre seis y 12 años de comunidades rurales querría conocer el promedio de ingesta calórica de esta población.

Como se ha explicado en la primera unidad de este material, el comportamiento de la población está determinado por el valor de un parámetro. Este parámetro, normalmente desconocido, se calculará con la información de una muestra.

En esta unidad, se estudiará uno de los temas básicos de la materia: la estimación de parámetros, en particular, la media, el promedio y la varianza poblacional.

En primer lugar, se mostrarán los tipos de estimación empleados: puntual y de intervalo. El siguiente tema corresponde a los tipos de errores de estimación, los cuales son atribuibles al muestreo u otra causa, continuando con las propiedades de los estimadores. Una vez explicados los aspectos más importantes que se deben tomar en cuenta en la estimación, el siguiente paso es mostrar cómo realizar estimaciones del promedio poblacional (tanto con muestras grandes como con pequeñas), estimaciones de una proporción, y finalmente cómo construir un intervalo de confianza para la varianza y desviación poblacional.

3.1. Estimaciones por punto y estimaciones por intervalo

Como se mencionó en la unidad de introducción al muestreo, la finalidad de la estadística inferencial es realizar estimaciones de parámetros poblacionales con los valores de una muestra. Supóngase que en una organización se realizará un evento deportivo donde se ofrecerán bebidas energéticas a 800 participantes: los organizadores se preguntan qué cantidad del líquido adquirir. Para resolver este problema, encuestan a una muestra de 50 posibles asistentes acerca de la cantidad de bebida que consumen en un evento similar. La encuesta arrojó que en promedio consumen cuatro litros por persona; así, los organizadores estiman que deberán adquirirse 800 x 4 = 3200 litros. Los organizadores creen que no necesariamente se tendría que consumir esa cantidad, por lo que prefieren manejar un intervalo., y después de un análisis de la información estiman que el consumo será entre 3000 y 3400 litros. ¿Qué diferencia hubo entre ambas estimaciones? En este subtema, se responderá esta pregunta.

Notación y conceptos

Un parámetro es un valor de la población que determina su comportamiento; por ejemplo, el comportamiento de una población con un promedio de cinco unidades es diferente a otra con promedio de ocho unidades. Para hacer referencia a un parámetro poblacional, se utilizará la letra θ.

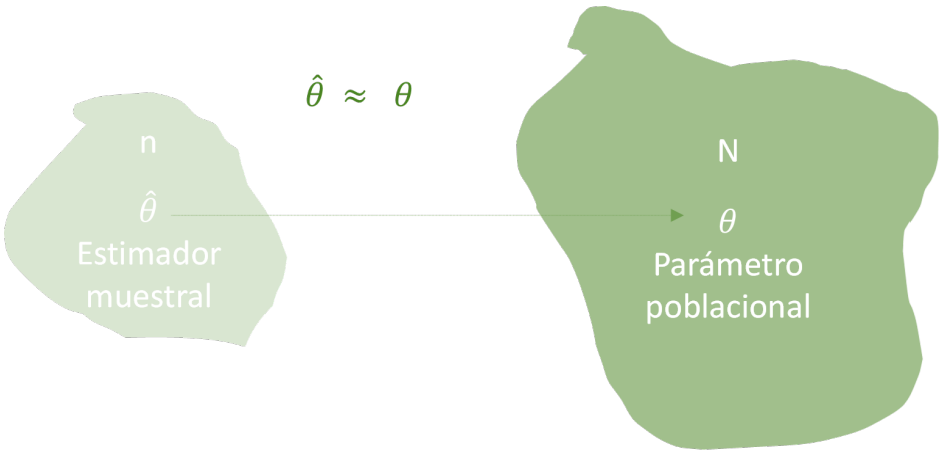
El *estimador* es la regla que indica cómo realizar el cálculo de una estimación a través de una fórmula que involucre los valores de una muestra; se denota como $\widehat{\boldsymbol{\theta}}$. El símbolo "^" significa que la fórmula es un estimador del parámetro θ. Por ejemplo, si el parámetro poblacional a estimar (θ) es el promedio poblacional $\overline{\boldsymbol{X}}$, el estimador del parámetro se denotará como $\widehat{\overline{\boldsymbol{X}}}$.

Como observación adicional, para referirse a parámetros poblacionales se utilizan letras mayúsculas o letras del alfabeto griego.

Se define como *estimación* al valor resultante de aplicar el estimador con los datos de la muestra.

En la figura 1 se ilustra el objetivo de un estimador.

Figura 1. Objetivo de un estimador

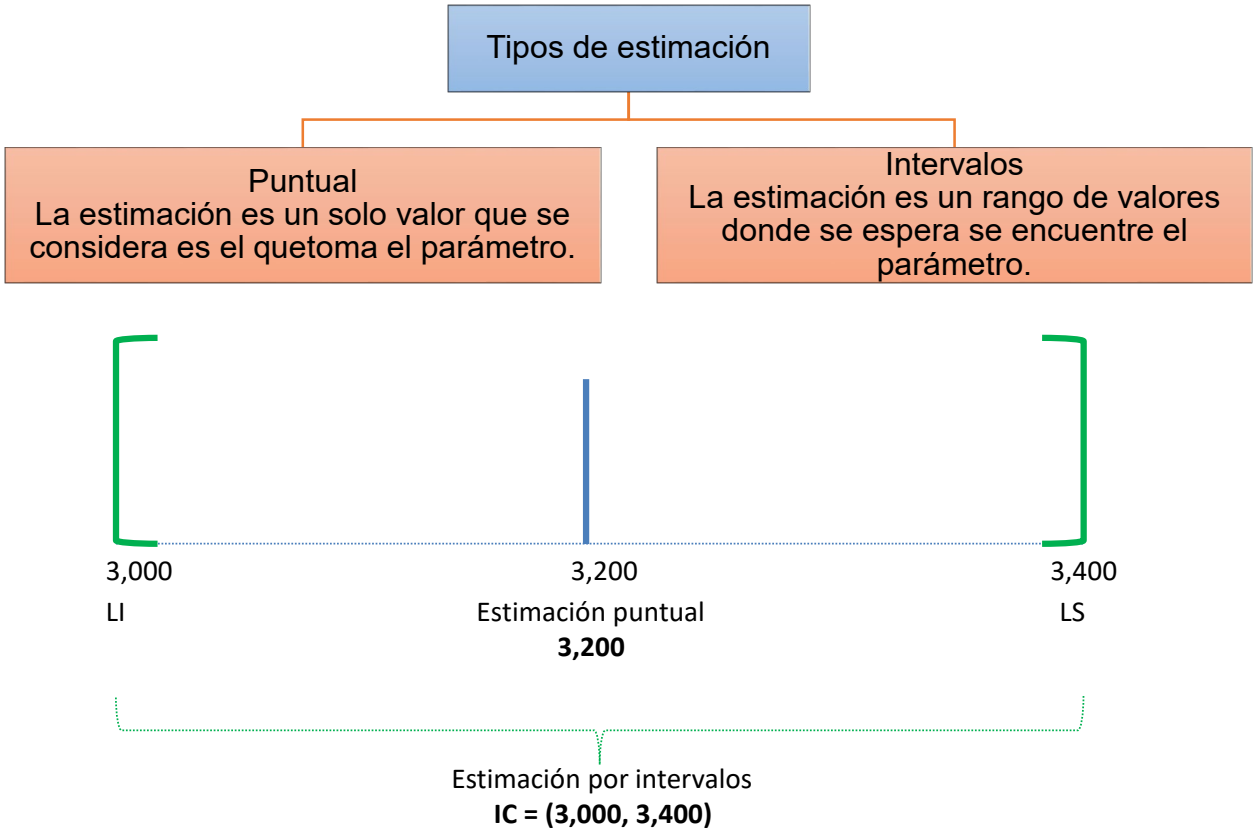


La figura anterior presenta dos conjuntos de diferente tamaño. El menor ejemplifica una muestra de tamaño *n* tomada del conjunto mayor, que es la población con *N* elementos. Dentro de la muestra, se obtiene el estimador $\widehat{\boldsymbol{\theta}}$, el cual busca estimar el valor del parámetro poblacional θ, que normalmente se desconoce. Se espera que la estimación se aproxime al valor real, lo cual se representa con el símbolo ≈.

Tipos de estimación

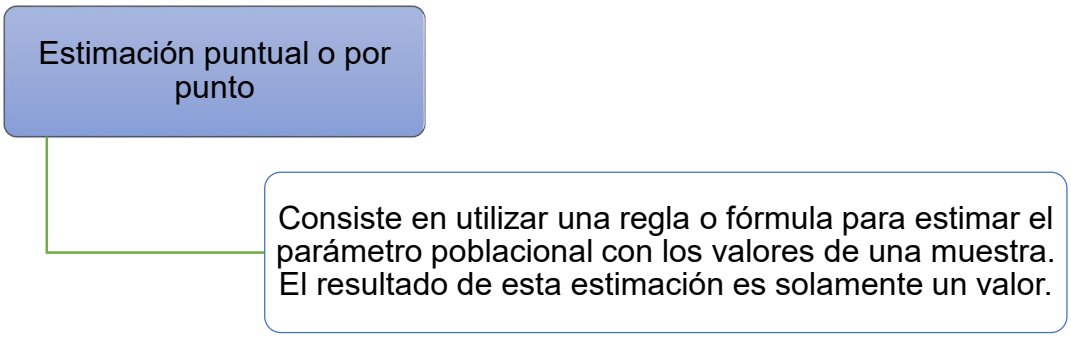
En el ejemplo narrado al comienzo de esta sección, los organizadores del evento estimaron la cantidad de bebida energética de dos maneras: a través de un valor puntual y mediante un rango. Lo anterior ejemplifica que la estimación de un parámetro puede hacerse de forma puntual o por intervalo. La figura 2 explica ambos tipos de estimación.

Figura 2. Tipos de estimación



La figura anterior define los tipos de estimación (puntual y de intervalo). La parte inferior de la figura representa esos tipos de estimación: la línea central de color azul señala la estimación puntual del parámetro (3200 litros de bebida energética); y las líneas en color verde, el rango de valores donde se espera que se encuentre el valor del parámetro (3000, 3400).

Estimación puntual (por punto)



Los parámetros poblacionales que habitualmente interesa estimar son el promedio y la proporción poblacional. La tabla siguiente presenta los estimadores para estos parámetros.

Tabla 1. Parámetros y estimadores más usados

Parámetro poblacional Θ	Estimador $\widehat{\boldsymbol{\theta}}$	
	Notación	Fórmula
Promedio M	Promedio $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$	$\bar{\boldsymbol{x}} = \frac{\sum \boldsymbol{x_i}}{\boldsymbol{n}}$
Proporción P	Proporción $\widehat{\mathbf{P}} = \mathrm{p}$	$\boldsymbol{p} = \frac{\sum x_i}{\boldsymbol{n}} \boldsymbol{x_i}$ = 0 o 1

En la tabla anterior, la primera columna muestra el nombre y la notación del parámetro. Las siguientes dos columnas hacen referencia al estimador del parámetro: una indica cómo denotarlo; la otra, la fórmula que lo define.

Cuando se desconoce la varianza poblacional, se recurre a estimarla con la muestral:

$$s^2 = \frac{\sum(x_1 - \bar{x})^2}{n-1}$$

Estimación por intervalos

Estimación por intervalos	Consiste en calcular un rango de valores en los que se espera, con cierto nivel de confianza, que se encuentre contenido el parámetro. El resultado de esta estimación es un intervalo. Es común llamar a este rango de valores *intervalo de confianza*.

Fórmula general para construir el intervalo de confianza:

$$IC = \hat{\theta} \pm \delta\sigma_{\hat{\theta}}$$

- Donde:

$IC = Intervalo\ de\ confianza$
$\hat{\theta} = Estimador\ puntual\ del\ parámetro$
$\delta = nivel\ de\ confianza$. Probabilidad de que el intervalo de confianza contenga al parámetro de la población
$\sigma_{\hat{\theta}} = desviación\ estandar\ del\ estimador$

Como un estimador emplea los valores de una muestra aleatoria, el resultado es también aleatorio, por lo que un estimador es una variable aleatoria con un valor esperado $E[\hat{\theta}]$ y una varianza $Var\ [\hat{\theta}]$. El nivel de confianza de la fórmula más que entenderse como una probabilidad debe considerarse una proporción de éxito en un número muy grande de repeticiones.

Para construir un intervalo de confianza, es necesario conocer la estimación puntual del parámetro y la desviación del estimador, y determinar el nivel de confianza.

La siguiente tabla muestra, para los parámetros promedio y proporción, su estimador, la fórmula para realizar la construcción del intervalo de confianza para muestras grandes y pequeñas, la fórmula para realizar una estimación puntual y la desviación estándar del estimador.

Tabla 2. Elementos para construir un intervalo de confianza para el promedio y proporción poblacional

Parámetro población	Estimador	Tamaño de la muestra	Fórmula		
			Intervalo de confianza	Estimador puntual	Desviación estándar del estimador
Promedio μ	$\bar{x}$	n>30	$IC = \bar{x} \pm Z\frac{\sigma}{\sqrt{n}}$	$\bar{x} = \frac{\sum x_i}{n}$	$s_{\bar{x}} = \frac{s}{\sqrt{n}}$
		n<30	$IC = \bar{x} \pm t\frac{\sigma}{\sqrt{n}}$		
Proporción P	p	n>30	$IC = p \pm Z\sqrt{\frac{pq}{n}}$	$p = \frac{\sum x_i}{n}$	$s_p = \sqrt{\frac{pq}{n}}$
		n<30	$IC = p \pm t\sqrt{\frac{pq}{n}}$	$x_i = 0 \; o \; 1$	q=1-p

3.2. Error de muestreo y errores que no son de muestreo

En todo ejercicio de estimación se asume la ocurrencia de un error, por lo que desde el diseño se debe buscar disminuirlo. Los errores que se pueden presentar son atribuibles al muestreo o a otras causas, como se explica a continuación.

Error de muestreo

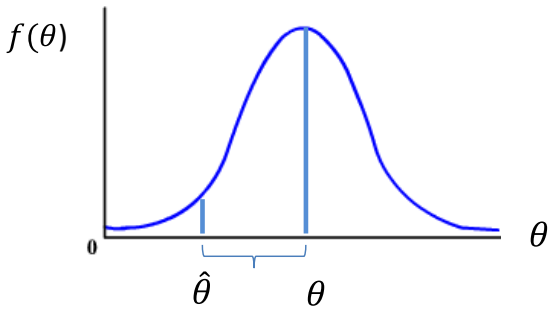
Toda estimación tiene un error debido a que se conoce una parte de la información. Al comienzo de cualquier ejercicio de estimación se debe fijar el límite de error permitido, como un porcentaje o como una desviación de unidades.

El error de muestreo se refiere a un error de la estimación atribuible a la muestra.

Por ejemplo, supóngase que se determinó manejar un error de cinco puntos porcentuales en la estimación de la proporción de alumnos que reprueban un curso de matemáticas financieras; supóngase además que la proporción real es de 36% y la muestra obtenida arroja una estimación de 15%. El error en la estimación más que a la metodología se debe a los alumnos que fueron seleccionados: el error es atribuible a la muestra.

En la figura 3 se ilustra este error de muestreo.

Figura 3. Error de muestreo



Como se sabe, el valor del parámetro determina la distribución de la población, por eso en el eje horizontal se relaciona con el valor del parámetro, por tanto, la distribución se encuentra asociada a este valor.

En la figura, la muestra consiste en elementos con valores ubicados principalmente en la parte izquierda de la distribución:

> La estimación resultó estar alejada del parámetro real, aunque de acuerdo con la distribución es menos probable que ocurra (esto no significa que no pueda ocurrir).

La distancia entre el valor real del parámetro y su estimación es el error. Para manejar este error, se buscará un tamaño de muestra que garantice

$$P\left[\left|\theta - \hat{\theta}\right| < B\right] \geq 1 - \alpha$$

Es decir, la probabilidad de que ocurra un error máximo B debe ser al menos 1 – α, donde alfa (α) es un valor entre 0 y 1. Entonces, esta probabilidad es el nivel de significancia.[8]

[8] La fórmula anterior es resultado de la ley de los grandes números, uno de los principales resultados de probabilidad: entre mayor información se tenga de un parámetro, la probabilidad de que la estimación se acerque al valor real se incrementa.

Error no atribuible al muestreo

El error no atribuible al muestreo se debe, entre otras causas, a un mal diseño del instrumento, la logística implementada o una elevada tasa de no respuesta.

Un buen diseño que considere estas eventualidades ayudará a reducir y controlar el riesgo de error.

Un ejemplo de error no atribuible al muestreo es el siguiente. Una empresa desea conocer el número de tazas de café que toma cierto segmento de interés, y en vez de utilizar una variable cuantitativa en la respuesta de su pregunta emplea una cualitativa.

3.3. Propiedades de los estimadores

Para estimar un parámetro, puede existir en ocasiones más de un estimador, por lo que es necesario utilizar aquellos que tengan las propiedades que se explican a continuación.

Propiedades deseables de los estimadores

Insesgado
• La primera propiedad de un estimador es que estime lo que se quiere estimar; por ejemplo, si se realizara una estimación con muchas muestras aleatorias, el valor esperado del estimador es el parámetro poblacional de interés. Cuando esto ocurre, el estimador es *insesgado*.

Un estimador es insesgado si satisface la siguiente condición:

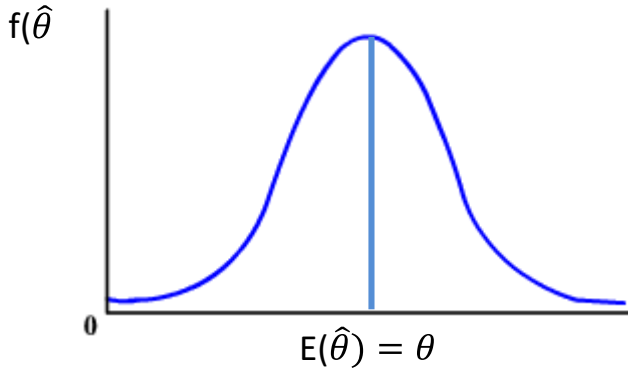
$$\boldsymbol{E[\hat{\theta}] = \theta}$$

Si esto se cumple, entonces:

$$\boldsymbol{ECM[\hat{\theta}] = Var\,[\hat{\theta}]}$$

En la figura 4 se ilustra esta propiedad.

Figura 4. Distribución de un estimador insesgado



La figura anterior ilustra la distribución de un estimador insesgado cuyo valor esperado es el parámetro. Es importante mencionar que la distribución acampanada de la figura solamente es con fines ilustrativos, ya que un estimador no necesariamente tiene esta distribución de probabilidades.

Con menos variabilidad
• La siguiente característica que se busca en un estimador es que sus estimaciones varíen lo menos posible del parámetro poblacional. Un estimador así es más eficiente o con menos variabilidad.

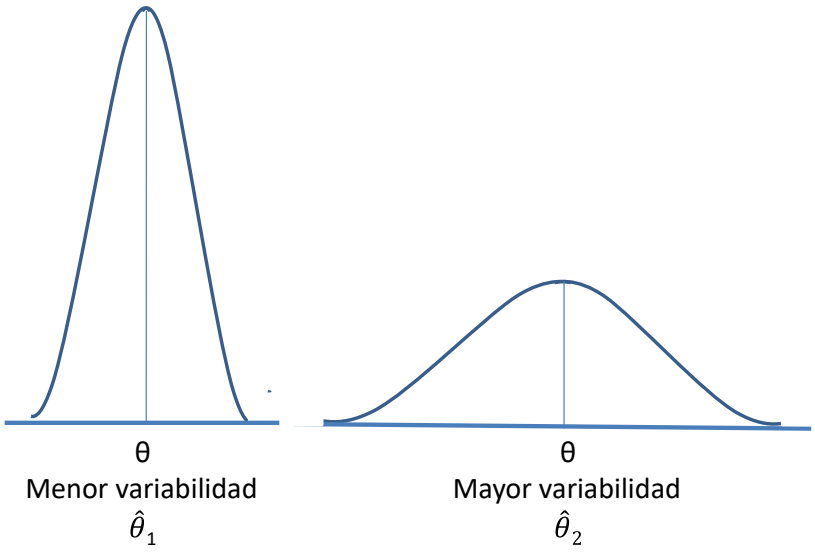
Sean $\widehat{\boldsymbol{\theta}}_{\boldsymbol{1}}$ y $\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}$ dos estimadores del parámetro θ:

$$\text{Si } \boldsymbol{Var}\left[\widehat{\boldsymbol{\theta}}_{\boldsymbol{1}}\right] < \boldsymbol{Var}\left[\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}\right]$$

Entonces, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{1}}$ es más eficiente que $\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}$

La figura 5 ilustra esta característica.

Figura 5. Eficiencia de dos estimadores insesgados

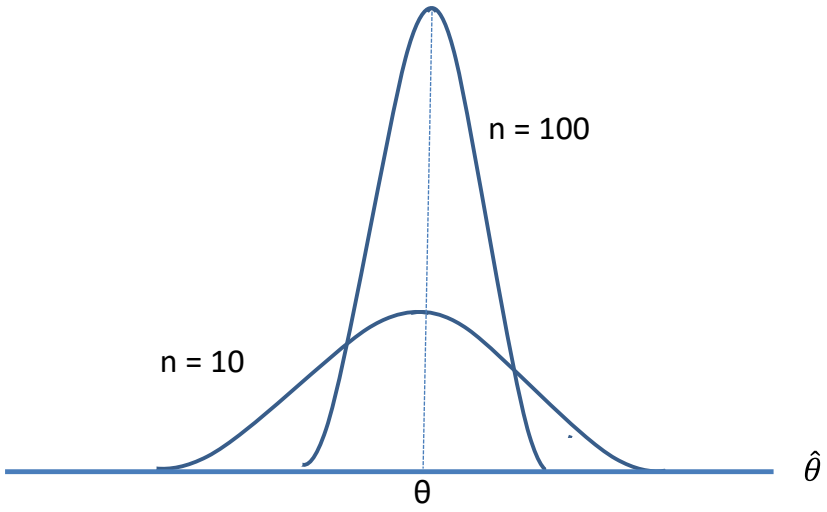


La figura 5 ilustra la distribución de dos estimadores $\widehat{\boldsymbol{\theta}}_{\boldsymbol{1}}$ y $\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}$ del parámetro poblacional θ. Aunque ambos estimadores son insesgados, el primero da mejores estimaciones, en tanto es más probable que arroje un valor más cercano al parámetro real respecto del segundo. Por tanto, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{1}}$ es más eficiente que $\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}$.

Consistente
• La última propiedad esperada en un estimador es que, a medida que utilice mayor información de la población, su estimación sea cada vez más cercana al parámetro poblacional. Cuando esto ocurre, el estimador es *consistente.*

La figura 6 ilustra el comportamiento de un estimador consistente.

Figura 6. Comportamiento de un estimador consistente



La figura anterior ilustra el comportamiento de un estimador consistente. Conforme aumenta el tamaño de muestra, la variabilidad del estimador disminuye: las estimaciones son cada vez más cercanas al valor real del parámetro.

3.4. Estimación de una media con muestras grandes

El teorema del límite central garantiza que, conforme aumenta el tamaño de la muestra, la distribución del promedio muestral se acerca a una distribución normal cuya media es el promedio poblacional, y la varianza es la varianza poblacional entre el tamaño de la muestra. Como regla general:

se considera que con un tamaño de muestra al menos de 30 elementos la distribución del promedio muestral sigue una distribución normal.

Teniendo presente esta regla, en muestras grandes (al menos de 30 elementos) se empleará una distribución normal para realizar una estimación por intervalo de la media.

La tabla siguiente muestra el parámetro medio, su estimador, el tamaño de la muestra, la fórmula para el intervalo de confianza de la media, la fórmula para calcular el estimador de la media y la fórmula del estimador de la media muestral.

Tabla 3. Elementos para realizar la estimación puntual y por intervalo de la media (promedio) con muestras grandes

Parámetro población	Estimador	Tamaño de la muestra	Fórmula		
			Intervalo de confianza	Estimador puntual	Desviación estándar del estimador
Promedio μ	$\bar{x}$	n>30	$IC = \bar{x} \pm z\frac{s}{\sqrt{n}}$	$\bar{x} = \frac{\sum x_i}{n}$	$s_{\bar{x}} = \frac{s}{\sqrt{n}}$

En la tabla anterior, columna 5, se muestra el estimador puntual de la media poblacional, que es el promedio muestral. En la columna 4, se presenta cómo calcular el intervalo de confianza. En este caso, cuando se conoce la varianza poblacional, se utiliza en los cálculos; de no ser así, se estima este valor empleando la varianza muestral. El valor *Z* representa el nivel de confianza buscado; en este planteamiento, z es el cuantil de una distribución normal estándar (Z) que parte la curva en dos áreas, una con valor $1-\frac{\alpha}{2}$ y otra de $\frac{\alpha}{2}$, siendo α un valor entre 0 y 1.

En la columna 6, se muestra la desviación del estimador, acorde con el teorema del límite central.

Para calcular el valor de *z*, se puede recurrir a tablas o algún paquete. A continuación, se plantea cómo calcularlo en MS-Excel.

En Excel se emplea la función

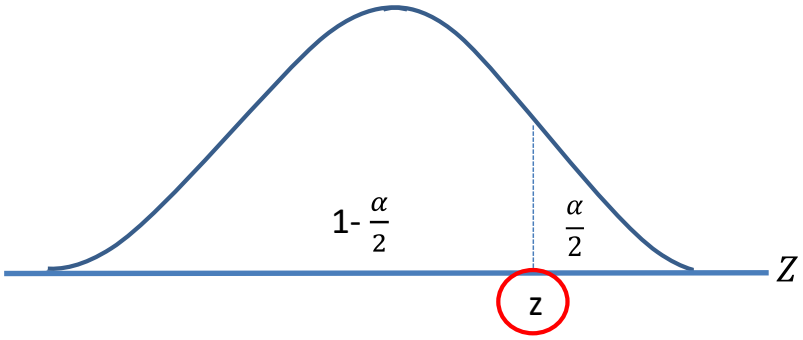
DISTR.NORM.ESTAND.INV(probabilidad)

Para calcular el cuantil z donde se acumula una probabilidad de $1-\frac{\alpha}{2}$ $(0<\alpha<1)$

En la figura 7, se ilustra el valor que calcula la fórmula.

Figura 7.Valor calculado con la fórmula de Excel

DISTR.NORM.ESTAND.INV(probabilidad)



En la figura anterior, en el valor encerrado en el círculo rojo se encuentra el punto que estima la fórmula. La información que se introduce en la fórmula es el área acumulada del lado izquierdo de *z*.

Supóngase que se desea realizar una estimación con un nivel de confianza del 95%, entonces:

1 – α = 0.95 α = 1-0.95 α = 0.05	Por tanto:	$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

Entonces, el valor *z* es el siguiente:

DISTR.NORM.ESTAND.INV(1-0.025) = 1.96

En la tabla 4 se muestran los valores de *z* para los niveles de confianza más usados.

Tabla 4. Valores de z obtenidos para los niveles de confianza más usados empleando Excel

Nivel de confianza	α 1-nivel de confianza	Función en MS-Excel *DISTR.NORM.ESTAND.INV*$(1-\frac{\alpha}{2})$	z
90%	10%	DISTR.NORM.ESTAND.INV(1-0.10/2)	1.64
95%	5%	DISTR.NORM.ESTAND.INV(1-0.05/2)	1.96
99%	9%	DISTR.NORM.ESTAND.INV(1-0.01/2)	2.58

Ejemplos de estimación de una media con muestras grandes

Primer ejemplo

El director financiero de una agencia de publicidad desea conocer el gasto promedio de la organización, pues está preocupado por el nivel de gasto registrado recientemente. Por tal motivo realiza una auditoría a 30 facturas elegidas al azar.

La información de las erogaciones seleccionadas se muestra a continuación.

Monto de las facturas auditadas

Monto de facturas combinadas

Factura	Gasto en miles	Factura	Gasto en miles
1	99	16	96
2	15	17	79
3	59	18	71
4	14	19	56
5	72	20	51
6	59	21	72
7	68	22	25
8	22	23	71
9	40	24	52
10	79	25	99
11	97	26	70
12	82	27	82
13	93	28	47
14	76	29	35
15	48	30	93

Con la información de esta muestra, procede lo siguiente:

- Estimar el gasto promedio de la organización con una estimación puntual.
- Estimar un intervalo de confianza con un nivel de confianza del 99%.
- Interpretar los resultados.

Respuestas

Para solucionar este problema, se sugiere realizar lo siguiente.

1. Determinar el parámetro a estimar Promedio μ	**2.** Determinar el estimador del Parámetro $\bar{x}$
3. Calcular el estimador puntual a través de la fórmula correspondiente: $$\bar{x} = \frac{\sum x_i}{n}$$ $$\bar{x} = \frac{99 + 15 + \cdots + 35 + 93}{30}$$ $$\bar{x} = \frac{1,922}{30} = 64.06$$	**4.** Determinar la fórmula para realizar el cálculo de la estimación por intervalo del estimador: $$IC = \bar{x} \pm Z\frac{s}{\sqrt{n}}$$
5. Establecer el nivel de confianza para calcular *z*, a través del nivel de confianza (1-α) Nivel de confianza 99%, es decir, 0.99 Determinar el valor de α, donde α = 1 – nivel de confianza α = 1 – 0.99 α = 0.01 Calcular el valor de *z*, utilizando la función en Excel DISTR. NORM. ESTAND. INV (1- α /2) DISTR.NORM. ESTAND. INV (1-0.01/2) Z = 2.5758 = 2.58	**6.** Determinar la fórmula para calcular la desviación estándar del estimador $$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$ Calcular la desviación estándar *s* y definir el valor de *n* n = 30 $$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$ $$s = \sqrt{\frac{(99 - 64.06)^2 + (15 - 64.06)^2 + \cdots (35 - 64.06)^2 + (93 - 64.06)^2}{30 - 1}}$$ $$s = \sqrt{\frac{18,339.86}{29}}$$ $$s = \sqrt{632.409}$$ $$s = 25.14$$

7. Calcular la desviación del estimador a través de la fórmula correspondiente:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$s_{\bar{x}} = \frac{25.14}{\sqrt{30}}$$

$$s_{\bar{x}} = \frac{25.14}{5.477}$$

$$s_{\bar{x}} = 4.59$$

8. Sustituir los valores en la fórmula general y calcular el límite inferior (LI) y límite superior (LS) del intervalo de confianza (IC):

$$IC = \bar{x} \pm Z\frac{s}{\sqrt{n}}$$

$$IC = 64.06 \pm 2.58 \cdot 4.59$$

$$IC = 64.06 \pm 11.845$$

$$LI = 64.06 - 11.845$$

$$LI = 52.22$$

$$LS = 64.06 + 11.845$$

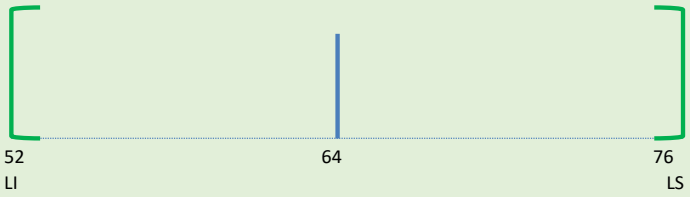
$$LS = 75.91$$

9. Construir el intervalo de confianza:

$$IC = (LI, LS)$$

$$IC = (52.22, 75.91)$$

Conforme a la estimación puntual, el promedio de gasto de la organización es de $64.06 (miles). De acuerdo con la estimación por intervalo, el gasto promedio de la organización a un nivel de confianza del 99% se sitúa entre $52.22 y $75.91 (miles).



Segundo ejemplo

Una farmacéutica cuenta con 500 representantes médicos. Con la intención de diseñar un plan de incentivos, se quiere conocer el promedio de visitas que realizan los representantes, para lo cual se analizó una muestra de 35 representantes médicos elegidos al azar.

En la siguiente tabla, se muestran las visitas realizada en un día por 35 representantes seleccionados.

a. Estimar el promedio de visitas que realizan los representantes médicos, con una estimación puntual.

b. Estimar un intervalo de confianza con un nivel de confianza del 95%.

c. Interpretar los resultados.

Representante	Número de visitas realizadas	Representante	Número de visitas realizadas
1	8	19	5
2	4	20	5
3	7	21	8
4	8	22	7
5	6	23	7
6	6	24	7
7	5	25	5
8	8	26	6
9	6	27	8
10	6	28	7
11	7	29	5
12	7	30	5
13	6	31	7
14	4	32	4
15	5	33	7
16	7	34	7
17	8	35	6
18	6		

Respuestas

1. Determinar el parámetro a estimar: Promedio μ	**2.** Determinar el estimador del parámetro $\bar{x}$
3. Calcular el estimador puntual a través de la fórmula correspondiente: $\bar{x} = \frac{\sum x_i}{n}$ $\bar{x} = \frac{8+4+\cdots+7+6}{35}$ $\bar{x} = \frac{220}{35} = 6.28$	**4.** Determinar la fórmula para realizar el cálculo de la estimación puntual del estimador: $IC = \bar{x} \pm Z\frac{s}{\sqrt{n}}$

5. Establecer el nivel de confianza para calcular *z*:

95%, es decir, 0.95

Determinar el valor de α:

α = 1 – 0.95

α = 0.05

Calcular *z* con la función de Excel:

DISTR. NORM. ESTAND. INV (1- α /2)

DISTR. NORM. ESTAND. INV (1- 0.05/2)

Z = 1.959 = 1.96

6. Determinar la fórmula para calcular la desviación estándar del estimador:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Calcular la desviación estándar *s* y definir el valor de *n*:

n = 35

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$s = \sqrt{\frac{(8-6.28)^2 + (4-6.28)^2 + \cdots (7-6.28)^2 + (6-6.28)}{35-1}}$$

$$s = \sqrt{\frac{51.14}{34}}$$

$$s = \sqrt{1.504}$$

$$s = 1.226$$

7. Calcular la desviación del estimador a través de la fórmula correspondiente:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$s_{\bar{x}} = \frac{1.226}{\sqrt{35}}$$

$$s_{\bar{x}} = \frac{1.226}{5.916}$$

$$s_{\bar{x}} = 0.2073$$

8. Sustituir los valores en la fórmula general y calcular el límite inferior (LI) y límite superior (LS) del intervalo de confianza (IC):

$$IC = \bar{x} \pm Z\frac{s}{\sqrt{n}}$$

$$IC = 6.28 \pm 1.96 \cdot 0.207$$

$$IC = 6.28 \pm 0.4063$$

$$LI = 6.28 - 0.4063$$

$$LI = 5.87$$

$$LS = 6.28 + 0.4063$$

$$LS = 6.69$$

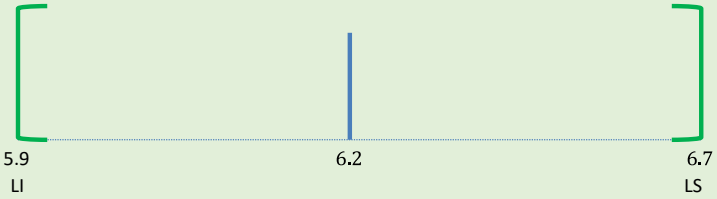
9. Construir el intervalo de confianza:

$$IC = (LI, LS)$$

$$IC = (5.87, 6.69)$$

Con base en la estimación puntual, el promedio de visitas diarias efectuadas por un representante médico es de 6.

Conforme a la estimación por intervalo, el promedio de visitas que realiza un representante médico al día con un nivel de confianza del 95% se sitúa entre 6 y 7.



3.4.1. Determinación del tamaño de muestra necesario para estimar una media

En la unidad dos se mostró que conforme el tamaño de la muestra se incrementa, el promedio muestral se aproxima a una distribución normal cuyo valor esperado es la media poblacional, y su desviación es la desviación poblacional dividida entre la raíz cuadrada del tamaño de la muestra. Dado lo anterior, conforme el tamaño de la muestra se incrementa, disminuye el error de estimación. Ahora bien, ¿de qué tamaño debe ser la muestra para garantizar una estimación confiable?

Para responder a lo anterior, se debe tener claridad sobre dos aspectos: el error máximo permitido y el nivel de confianza deseado. Cualquier resultado de un muestreo va a presentar un error de estimación, pero se busca que el riesgo α de que la distancia entre la estimación y el valor real supere un límite de error B predefinido sea pequeño, es decir:

$$P(|\bar{x} - \bar{X}| > B) < \alpha$$

Lo cual es equivalente a:

$$\boldsymbol{P(|\bar{x} - \bar{X}| \leq B) \geq 1 - \alpha \quad 0 \leq \alpha \leq 1} \qquad \textbf{(1)}$$

Donde:

$\bar{X}$: Promedio poblacional
$\bar{x}$: Promedio muestral
B: Error permitido

Entonces, como el promedio muestral sigue una distribución normal, si se expresa el error B en múltiplos de la desviación estándar del estimador, digamos $\boldsymbol{z\sigma_{\bar{x}} = z\frac{\sigma}{\sqrt{n}}}$ (1), queda de la siguiente forma:

$$\boldsymbol{P(|\bar{x} - \bar{X}| \leq B) \geq 1 - \alpha}$$

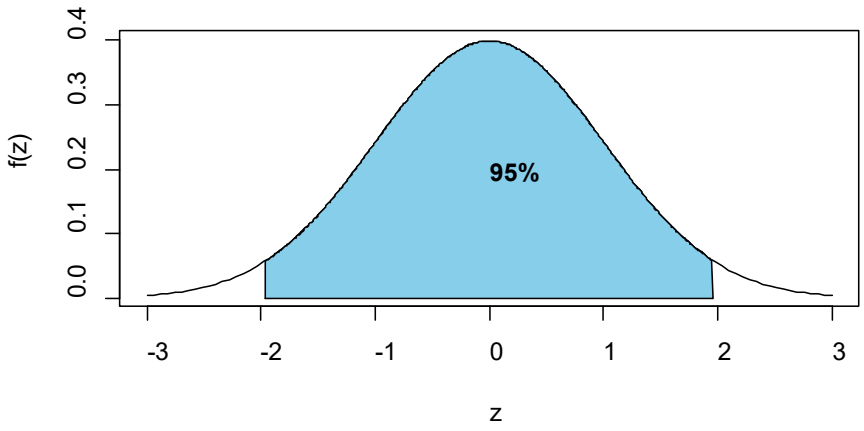
$$\boldsymbol{\rightarrow \; P\left(|\bar{x} - \bar{X}| \leq z\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \alpha}$$

$$\boldsymbol{\rightarrow \; P(|Z| \leq z\,) \geq 1 - \alpha}$$

Lo cual es una región de una distribución normal estandarizada limitada por los cuantiles **±z**.

A manera de ejemplo, si α=0.05, la región que se encuentra entre **±**1.96 en una distribución normal estandarizada tiene 95% de probabilidad, como lo ilustra la figura.

Figura 8. Región con probabilidad de 95% con un alfa de 0.05



Fuente: elaboración propia con uso de R. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Obsérvese que la confiabilidad de la estimación, 1-α, se encuentra directamente asociada al valor de z. Por otro lado, si:

$$\boldsymbol{B = z\sigma_{\bar{x}}}$$

Y además se está trabajando con una población finita de tamaño *N* entonces:

$$\boldsymbol{B = z\sigma_{\bar{x}} = z\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}$$

Al despejar *n* se obtiene:

$$\boldsymbol{B^2 = \frac{z^2\sigma^2}{n}\left(1-\frac{n}{N}\right)}$$

$$\boldsymbol{B^2 = \frac{z^2\sigma^2}{n} - \frac{z^2\sigma^2}{N}}$$

$$\boldsymbol{\frac{z^2\sigma^2}{n} = B^2 + \frac{z^2\sigma^2}{N} = \frac{NB^2 + z^2\sigma^2}{N}}$$

$$\boldsymbol{n = \frac{Nz^2\sigma^2}{NB^2 + z^2\sigma^2}} \qquad \textbf{(2)}$$

En caso de estar trabajando con una población infinita o con una donde el tamaño *N* es desconocido, el error *B* es:

$$B = z\sigma_{\bar{x}} = z\frac{\sigma}{\sqrt{n}}$$

Al despejar *n* de esta ecuación se obtiene:

$$n = \frac{z^2\sigma^2}{B^2} \qquad \textbf{(3)}$$

Las fórmulas (2) y (3) son las mismas que se presentaron en la sección 1.6.

Como observación adicional, en caso de desconocerse el valor de σ^2 se utiliza el estimador insesgado s^2.

Regresando al punto inicial, para calcular el tamaño de muestra se debe establecer el nivel de error permitido, *B*, así como el riesgo *α* de obtener una estimación que difiera del valor real en una cantidad mayor a *B,* la cual se encuentra asociada a *z*. A continuación, se muestran ejemplos de cálculo de tamaño de muestra para estimar una media.

Ejemplo 1. En un estudio acerca del gasto en los hogares, se desea conocer el monto mensual promedio destinado a transportación en una colonia de 600 familias. Se ha determinado entrevistar una muestra aleatoria de familias que garantice un error máximo de estimación de $20, con un riesgo *α* de sobrepasar dicho error de 5%. Se conoce por estudios en poblaciones semejantes que la desviación estándar del ingreso destinado a transportación es de $100, ¿de qué tamaño debe ser la muestra?

Respuesta

El problema indica que el parámetro de interés es un promedio (monto mensual promedio destinado a transporte) además brinda la siguiente información:

N=600

B=20

α=0.05

σ=100

Derivado de α=0.05, *z* es 1.96. El valor *z* se obtiene aplicando la fórmula de Excel:

DISTR.NORM.ESTAND.INV(1-0.05/2)

Donde el 0.05 en el interior de la fórmula se refiere al valor de *α*.

Sustituyendo estos valores en (2) se tiene:

$$\boldsymbol{n = \frac{Nz^2\sigma^2}{NB^2 + z^2\sigma^2}}$$

$$\boldsymbol{n = \frac{(600)(1.96)^2(100)^2}{(600)(20)^2 + (1.96)^2(100)^2}}$$

$$\boldsymbol{n = 82.8}$$

Es decir, se requiere una muestra de 83 hogares para garantizar una estimación que difiera en $20 del valor real con una confiabilidad de 1-α = 1 - 0.05 = 0.95 = 95%.

Ejemplo 2. ¿De qué tamaño sería la muestra en el ejemplo anterior si se desconociera el tamaño de la población (N) y el resto de los parámetros se mantuviera igual?

Respuesta

Del ejemplo anterior se tiene la siguiente información:

B=20

α=0.05

σ=100

z=1.96

Al sustituir estos valores en la fórmula (3) se obtiene:

$$\boldsymbol{n = \frac{z^2\sigma^2}{B^2}}$$

$$\boldsymbol{n = \frac{(1.96)^2(100)^2}{20^2}}$$

$$\boldsymbol{n = 96.04}$$

Es decir, se requiere una muestra de 96 hogares para garantizar una estimación que difiera en $20 del valor real con una confiabilidad de 1-α = 1 - 0.05 = 0.95 = 95%.

Estimación del tamaño de muestra considerando el error como una proporción del parámetro real

Es frecuente encarar situaciones donde se desea fijar el error de estimación como un porcentaje de la media, por ejemplo 5% o 10%, en este caso el error *B* es expresado como $r\bar{X}$ donde *r* es un número entre 0 y 1, al proceder de esta manera:

$$B = z\sigma_{\bar{x}}$$

$$r\bar{X} = z\sigma_{\bar{x}}$$

Si se está trabajando con una población finita de tamaño *N:*

$$r\bar{X} = z\frac{\sigma}{\sqrt{n}}\sqrt{\left(1 - \frac{n}{N}\right)}$$

Al despejar *n* se obtiene:

$$r^2\bar{X}^2 = z^2\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$$

$$r^2\bar{X}^2 = z^2\frac{\sigma^2}{n} - \frac{z^2\sigma^2}{N}$$

$$z^2\frac{\sigma^2}{n} = r^2\bar{X}^2 + \frac{z^2\sigma^2}{N}$$

$$z^2\frac{\sigma^2}{n} = \frac{Nr^2\bar{X}^2 + z^2\sigma^2}{N}$$

$$n = \frac{Nz^2\sigma^2}{Nr^2\bar{X}^2 + z^2\sigma^2}$$

Si se divide tanto el numerador como denominador por $\bar{X}^2$ se obtiene:

$$n = \frac{Nz^2 \frac{\sigma^2}{\bar{X}^2}}{Nr^2 + z^2 \frac{\sigma^2}{\bar{X}^2}}$$

$$n = \frac{Nz^2CV^2}{Nr^2+z^2CV^2} \qquad (4)$$

Donde CV es el coeficiente de variación definido como la razón de la desviación estándar con la media.

En caso de no tener conocimiento del tamaño *N* de la población:

$$r\bar{X} = z\frac{\sigma}{\sqrt{n}}$$

$$r^2\bar{X}^2 = z^2\frac{\sigma^2}{n}$$

$$n = \frac{z^2}{r^2}\frac{\sigma^2}{\bar{X}^2}$$

$$n = \frac{z^2}{r^2}CV^2 \qquad (5)$$

Las fórmulas (4) y (5) tienen como ventaja adicional que no es necesario conocer el valor de los parámetros poblacionales y es suficiente con definir la relación entre la desviación respecto a la media.

Ejemplo 3. Un auditor desea determinar el pago promedio que una organización realizó a sus proveedores en el último ejercicio fiscal. El número total de comprobantes por este concepto es de $4,000 y se desea tener una estimación que no difiera en más de 10% del valor real con una confiabilidad de 95%. Por su experiencia, el auditor considera razonable asumir que la desviación estándar de los pagos es 1.2 veces mayor al promedio.

Respuesta

El problema indica que el parámetro de interés es un promedio (pago promedio a proveedores), además brinda la siguiente información:

N=4,000
r =10%
1-α=0.95 es decir α= 1- 0.95 = 0.05
CV=1.2

Derivado que α=0.05, *z* es 1.96.

Al sustituir estos valores en la fórmula (4) se obtiene:

$$\boldsymbol{n = \frac{Nz^2CV^2}{Nr^2 + z^2CV^2}}$$

$$\boldsymbol{n = \frac{(4,000)(1.96)^2(1.2)^2}{(4,000)(0.1)^2 + (1.96)^2(1.2)^2}}$$

$$\boldsymbol{n = 485.98}$$

Es decir, se requiere una muestra de 486 comprobantes para garantizar una estimación que difiera en 10% del promedio real con una confiabilidad de 95%.

Ejemplo 5. ¿De qué tamaño sería la muestra en el ejemplo anterior si se desconociera el tamaño de la población (N) y el resto de los parámetros se mantuviera igual?

Respuesta

Del ejemplo anterior se conoce que *r* =10%, *z* = 1.96 y CV = 1.2. Sustituyendo estos valores en la fórmula (5) se obtiene:

$$\boldsymbol{n = \frac{z^2}{r^2} CV^2}$$

$$\boldsymbol{n = \frac{1.96^2}{0.1^2} 1.2^2}$$

$$\boldsymbol{n = 553.19}$$

Es decir, se requiere una muestra de 553 comprobantes para garantizar una estimación que difiera en 10% del promedio real con una confiabilidad de 95%.

3.5. Estimación de una media con muestras pequeñas

En la sección anterior, se mostró cómo realizar estimaciones de la media con muestras grandes; sin embargo, en la práctica es común enfrentar situaciones donde el tamaño de la muestra es menor a 30 elementos. ¿Cómo realizar estimaciones para este caso? Sabemos que la distribución de la media muestral tiende a ser una normal conforme aumenta el tamaño de muestra; para muestras pequeñas, donde además se desconoce la varianza poblacional, la distribución del estimador muestral se asemeja: puede modelarse con una distribución t de Student, con n – 1 grados de libertad. La distribución *t* se aproxima a una normal estándar en la medida que aumenta el tamaño de la muestra.

La tabla 5 presenta los elementos para realizar una estimación de la media poblacional con muestras pequeñas.

Tabla 5. Elementos para realizar una estimación de la media (promedio) poblacional con muestras pequeñas

Parámetro población	Estimador	Tamaño de la muestra	Fórmula		
			Intervalo de confianza	Estimador puntual	Desviación estándar del estimador
Promedio μ	$\bar{x}$	n<30	$IC = \bar{x} \pm t\frac{s}{\sqrt{n}}$	$\bar{x} = \frac{\sum x_i}{n}$	$s_{\bar{x}} = \frac{s}{\sqrt{n}}$

La tabla 5 tiene la misma descripción que la tabla 4, con la diferencia que el nivel de confianza está expresado en el cuantil *t* de una distribución t de Student con n – 1 grados de libertad que parte la curva en dos áreas, una con valor 1-$\frac{\alpha}{2}$ y la otra de $\frac{\alpha}{2}$, siendo α un valor entre 0 y 1.

Para calcular el valor de *t*, se puede recurrir a tablas o algún paquete. A continuación, se muestra cómo calcularlo en MS-Excel.

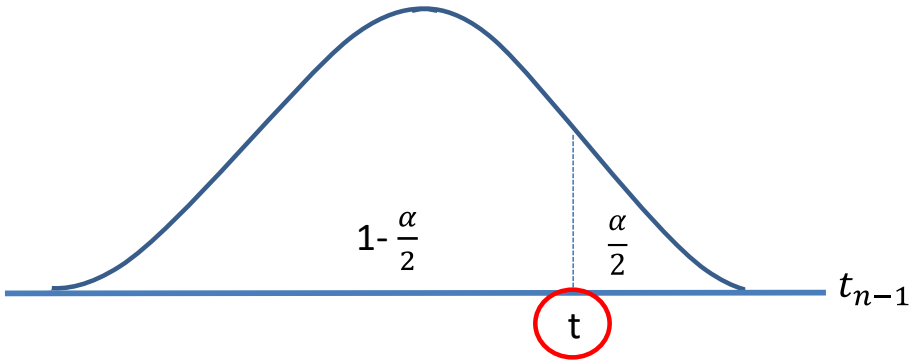
Excel presenta esta función:

DISTR.T.INV(probabilidad, grados_de_libertad

- En esta función, el parámetro *probabilidad* se refiere a $\frac{\alpha}{2}$ y el resultado es el cuantil *t*, que separa la curva en dos regiones, una con área 1-$\frac{\alpha}{2}$ y la otra de $\frac{\alpha}{2}$, siendo α un valor entre 0 y 1.)

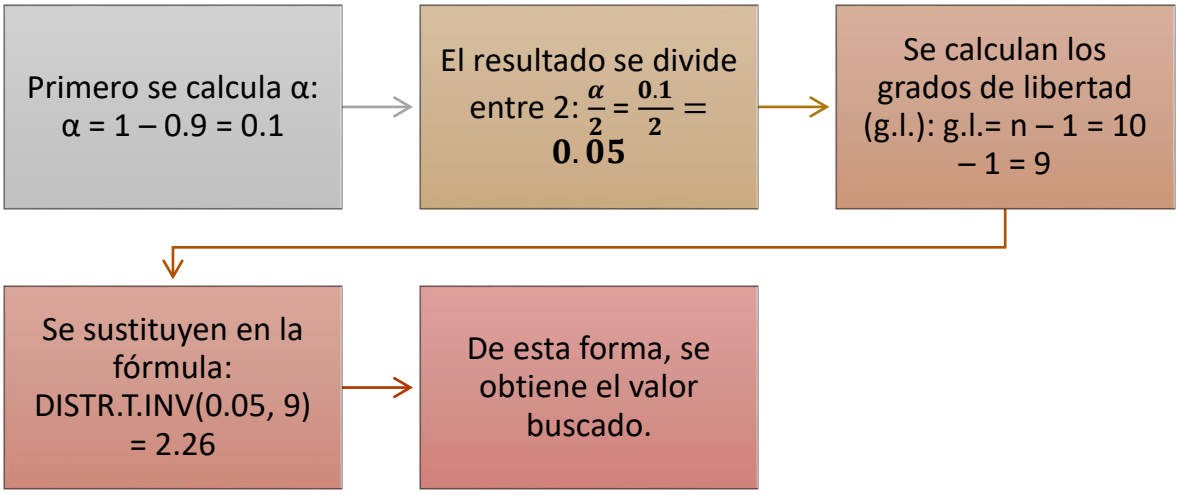
En la figura 9, se ilustra el valor obtenido con la fórmula DISTR.T.INV(α/2, n-1)

Figura 9. Valor calculado con la fórmula de Excel DISTR.T.INV(α/2, n-1)



En la figura anterior, el valor que se encuentra encerrado en el círculo rojo es el resultado de la fórmula.

Supóngase que se desea realizar una estimación con un nivel de confianza de 90% con una muestra de 10 elementos.



Ejemplos de estimación de la media poblacional con muestras pequeñas

A continuación, se muestran algunos ejemplos de esta estimación:

Primer ejemplo

1. Se desea estimar el número de horas promedio de capacitación de 350 empleados de una empresa fabricante de refrescos. Ante la dificultad de recabar información, se eligió una muestra de 10 empleados del área operativa y se registraron las horas de capacitación recibidas durante el mes de julio.

La siguiente tabla muestra la información.

Horas de capacitación de 10 empleados durante el mes de julio

Empleado	Horas de capacitación Julio
1	21
2	40
3	19
4	30
5	40
6	36
7	28
8	28
9	33
10	28

a. Realizar una estimación puntual del promedio de horas de capacitación de los empleados recibidas en el mes de julio.

b. Realizar una estimación por intervalo para el promedio de capacitación de los empleados recibida en el mes de julio con un nivel de confianza del 95%.

c. Interpretar los resultados.

Respuestas

1. Determinar el estimador del parámetro $\bar{x}$	**2.** Calcular el estimador puntual a través de la fórmula correspondiente: $$\bar{x} = \frac{\sum x_i}{n}$$ $$\bar{x} = \frac{21 + 40 + \cdots + 33 + 28}{10}$$ $$\bar{x} = \frac{303}{10} = 30.3$$
3. Determinar la fórmula para realizar el cálculo de la estimación por intervalo: $$IC = \bar{x} \pm t\frac{s}{\sqrt{n}}$$	**4.** Establecer el nivel de confianza para calcular el valor del punto de corte a través de α: Nivel de confianza = 95%, es decir, 0.95 Determinar el valor de α: α = 1 – nivel de confianza α = 1 – 0.95 α = 0.05 Calcular el valor del punto de corte *t* con la función Excel: DISTR.T.INV(α /2, n-1) DISTR.T.INV(0.05/2, 10-1) DISTR.T.INV(0.025, 9) t = 2.685 = 2.69
5. Sustituir los valores en la fórmula general para calcular el límite inferior (LI) y límite superior (LS) del intervalo: $$IC = \bar{x} \pm Z\frac{s}{\sqrt{n}}$$ $$IC = 30.3 \pm 2.69 \cdot 2.25$$ $$IC = 30.3 \pm 6.041$$ $$LI = 30.3 - 6.041$$ $$LI = 24.25$$ $$LS = 30.3 + 6.041$$ $$LS = 36.34$$	**6.** Sustituir los valores en la fórmula general para calcular el límite inferior (LI) y límite superior (LS) del intervalo: $$IC = \bar{x} \pm Z\frac{s}{\sqrt{n}}$$ $$IC = 30.3 \pm 2.69 \cdot 2.25$$ $$IC = 30.3 \pm 6.041$$ $$LI = 30.3 - 6.041$$ $$LI = 24.25$$ $$LS = 30.3 + 6.041$$ $$LS = 36.34$$

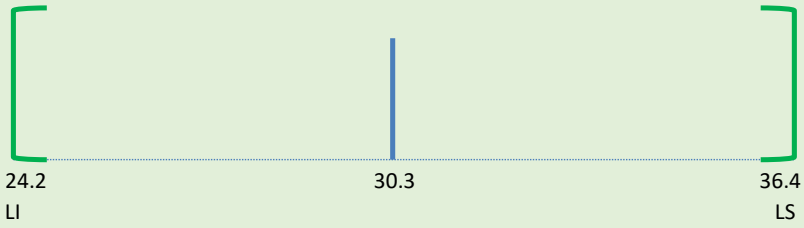
7. Construir el intervalo de confianza:

$$IC = (LI, LS)$$

$$IC = (24.25, 36.34)$$

Con base en la estimación puntual, el promedio de horas de capacitación fue de 30.3 horas en el mes de julio.

De acuerdo con la estimación por intervalo, el promedio de horas de capacitación recibidas en el mes de julio por los empleados del área operativa de la empresa de refrescos está entre 24.25 y 36.34 horas, con un 95% de confianza.



Segundo ejemplo

Con la intención de producir propaganda destinada a jóvenes de 18 años, en cierto estado del país con 500 poblados, se extrajo una muestra de 16 localidades, donde se realizó un censo de jóvenes residentes de 18 años. La siguiente tabla muestra la población de 18 años en 16 localidades.

Localidad	Población de 18 años
1	255,226
2	245,317
3	251,149
4	259,036
5	269,143
6	279,054
7	286,484
8	292,889
9	299,688
10	305,969
11	314,557
12	316,589
13	324,413
14	330,382
15	337,431
16	342,457

a. Realizar una estimación puntual del promedio de jóvenes de 18 años.

b. Realizar una estimación por intervalo para el promedio de jóvenes de 18 años con un nivel de confianza del 99%.

c. Interpretar los resultados.

Respuestas

1. Determinar el parámetro a estimar: Promedio μ	**2.** Determinar el estimador del parámetro: $\bar{x}$
3. Calcular el estimador puntual a través de la fórmula correspondiente: $\bar{x} = \frac{\sum x_i}{n}$ $\bar{x} = \frac{255,226 + 245,317 + \cdots + 337,431 + 342,457}{16}$ $\bar{x} = \frac{4,709,784}{16} = 294,361.5$	**4.** Determinar la fórmula para realizar el cálculo de la estimación por intervalo: $IC = \bar{x} \pm Z\frac{s}{\sqrt{n}}$
5. Establecer el nivel de confianza para calcular el valor del punto de corte a través de α: Nivel de confianza = 99%, es decir, 0.99 Determinar el valor de α: α = 1 – 0.99 α = 0.01 Calcular el valor del punto de corte *t* utilizando la función Excel: DISTR. T. INV (α /2, n-1) DISTR.T. INV (0.01/2, 16-1) DISTR. T. INV (0.005,15) t = 3.286 = 2.29	**6.** Determinar la fórmula para calcular la desviación estándar del estimador: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ Calcular la desviación estándar *s* y definir el valor de *n*: n = 16 $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$ $s = \sqrt{\frac{(255,226 - 294,361.5)^2 + (245,317 - 294,361.5)^2 + \cdots (337,431 - 294,361.5)^2 + (342,457 - 294,361.5)^2}{16-1}}$ $s = \sqrt{\frac{15,420,430,666}{15}}$ $s = \sqrt{102,8028,711.06}$ $s = 32,062.88$

7. Calcular la desviación del estimador a través de la fórmula correspondiente:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$s_{\bar{x}} = \frac{32,062.88}{\sqrt{16}}$$

$$s_{\bar{x}} = \frac{32,062.88}{4}$$

$$s_{\bar{x}} = 8,015.72$$

8. Sustituir los valores en la fórmula para calcular el límite inferior (LI) y límite superior (LS) del intervalo (IC):

$$IC = \bar{x} \pm Z\frac{s}{\sqrt{n}}$$

$$IC = 294,361.5 \pm 3.29 \cdot 8,015.72$$

$$IC = 294,361.5 \pm 26,339.96$$

$$LI = 294,361.5 - 26,339.96$$

$$LI = 268,021.535$$

$$LS = 294,361.5 + 26,339.96$$

$$LS = 320,701.465$$

9. Construir el intervalo de confianza:

$$IC = (LI, LS)$$

$$IC = (268,022, 320,701)$$

Con base en la estimación puntual, el promedio de jóvenes de 18 años por localidad es de 294 362.

De acuerdo con la estimación por intervalo, el promedio de jóvenes de 18 años por localidad se encuentra entre 268 022 y 320 701, con un 99% de confianza.

3.6. Estimación de una proporción

Como se ha comentado, la media muestral se acerca a una distribución normal que tiene como valor esperado el promedio poblacional, y su varianza es la varianza poblacional entre el tamaño de la muestra. Una proporción muestral es, en cierta manera, un promedio donde los valores son ceros y unos, por lo que su distribución se acerca a una normal cuya media es la proporción poblacional, y la varianza es la proporción por su complemento entre el tamaño de muestra.

La estimación de una proporción poblacional es parecida a la realizada para una media. Se debe aclara que, si el producto de la proporción muestral por el tamaño de muestra es al menos cinco, entonces el uso de una distribución normal es adecuado; en caso contrario, se deberán utilizar otras técnicas de estimación diferentes a las de este curso.

La tabla 6 presenta los elementos requeridos para efectuar la estimación de una proporción.

Tabla 6. Elementos requeridos para realizar una estimación de la proporción poblacional

Parámetro población	Estimador	Fórmula		
		Intervalo de confianza	Estimador puntual	Desviación estándar del estimador
Proporción P	p	$IC = p \pm Z\sqrt{\frac{pq}{n}}$	$p = \frac{x_i}{n}$ $x_i = 0 \; o \; 1$	$s_p = \sqrt{\frac{pq}{n}}$ q=1-p

A continuación, se presenta un ejemplo de estimación de una proporción poblacional con la información de una muestra.

Se retoma el primer ejemplo de la sección de estimación de medias para muestras grandes.

El director financiero de una agencia de publicidad desea conocer el gasto promedio de la organización debido a que está preocupado por el nivel de gasto registrado recientemente. Entonces, realiza una auditoría a 30 facturas elegidas al azar. La información de las erogaciones seleccionadas se muestra a continuación.

Factura	Gasto en miles	Factura	Gasto en miles
1	99	16	96
2	15	17	79
3	59	18	71
4	14	19	56
5	72	20	51
6	59	21	72
7	68	22	25
8	22	23	71
9	40	24	52
10	79	25	99
11	97	26	70
12	82	27	82
13	93	28	47
14	76	29	35
15	48	30	93

a. Realizar una estimación puntual de la proporción de gastos mayores a $80 mil.

b. Realizar una estimación por intervalo para la proporción de gastos mayores a $80 mil con un nivel de confianza del 95%.

c. Interpretar los resultados.

Respuestas

1. Determinar el parámetro a estimar:

Proporción P

2. Determinar el estimador del parámetro:

P

3. Calcular el estimador puntual a través de la fórmula correspondiente:

$$p = \frac{\sum x_i}{n}$$

Se definen n y $\sum x_i$:

n = 30

$\sum x_i$: valores mayores a 80

$$x_i = 8$$

Factura	Gasto en miles	Factura	Gasto en miles
1	99	16	96
2	15	17	79
3	59	18	71
4	14	19	56
5	72	20	51
6	59	21	72
7	68	22	25
8	22	23	71
9	40	24	52
10	79	25	99
11	97	26	70
12	82	27	82
13	93	28	47
14	76	29	35
15	48	30	93

$$p = \frac{8}{30}$$

$$p = 0.266$$

4. Determinar la fórmula para realizar el cálculo de la estimación puntual del estimador:

$$IC = p \pm Z\sqrt{\frac{pq}{n}}$$

Establecer el nivel de confianza para calcular el valor del punto de corte a través del valor de α:

Nivel de confianza = 95%, es decir, 0.95

Determinar α:

α = 1 – nivel de confianza

α = 1 – 0.95

α = 0.05

Calcular el valor del punto de corte *z* con la función Excel:

DISTR. NORM. ESTAND. INV (1-α/2)

DISTR.NORM. ESTAND. INV (1-0.05/2)

Z = 1.959 = 1.96

6. Determinar la fórmula para calcular la desviación estándar del estimador:

$$s_p = \sqrt{\frac{pq}{n}}$$

Donde $p = 0.266$

$$q = 1 - p$$
$$q = 1 - 0.266$$
$$q = 0.733$$
$$n = 30$$

7. Calcular la desviación del estimador a través de la fórmula correspondiente:

$$s_p = \sqrt{\frac{0.266 \cdot 0.733}{30}}$$
$$s_p = \sqrt{\frac{0.1955}{30}}$$
$$s_p = \sqrt{0.0065}$$
$$s_p = 0.807$$

8. Sustituir los valores en la fórmula para calcular el límite inferior (LI) y límite superior (LS) del intervalo de confianza (IC):

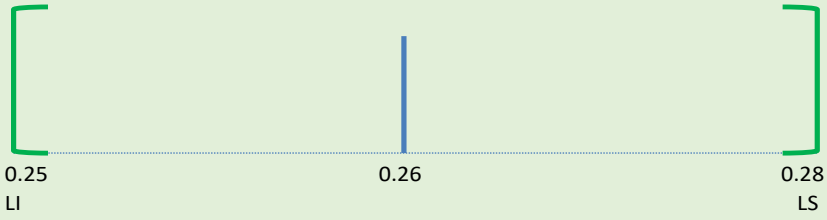
$$IC = p \pm Z\sqrt{\frac{pq}{n}}$$
$$IC = 0.266 \pm 1.96 \cdot 0.0807$$
$$IC = 0.266 \pm 0.0157$$
$$LI = 0.266 - 0.01571$$
$$LI = 0.2508$$
$$LS = 0.266 + 0.0157$$
$$LS = 0.2824$$

9. Construir el intervalo de confianza:

$$IC = (LI, LS)$$
$$IC = (0.25, 0.28)$$

Con base en la estimación puntual, la proporción de gastos mayores a $80 mil en la organización es de 26%.

De acuerdo con la estimación por intervalo, la proporción de gastos mayores a $80 mil en la organización se encuentra entre el 25% y 28%, con un 95% de confianza.



3.6.1. Determinación del tamaño de muestra para estimar una proporción

Como se mencionó en la unidad dos, la proporción es un caso de un promedio donde los valores son ceros y unos. En la sección 2.3 se mostró que la proporción muestral para un tamaño de muestra *n* considerablemente grande se aproxima a una distribución normal con media *P* y varianza $\frac{P\cdot(1-P)}{n}$. Si se busca calcular un tamaño de muestra que garantice una estimación de P con un error máximo *B* con confiabilidad 1-α, se parte de la siguiente igualdad:

$$B = z\sigma_p$$

$$B^2 = z^2\sigma_p^2$$

Para poblaciones finitas, se aproxima a:

$$B^2 \cong \frac{z^2 p\cdot(1-p)}{n}\left(1-\frac{n}{N}\right)$$

Al despejar *n* se obtiene:

$$n = \frac{Nz^2pq}{NB^2+z^2pq} \qquad \textbf{(6)}$$

Donde q = 1-p

Obsérvese que las fórmulas (2) y (6) de la sección 3.4.1 son semejantes, sólo cambia σ^2 por *pq*.

Sustituyendo *pq* en vez de σ^2, la fórmula (3) de la sección 3.4.1 queda así:

$$n = \frac{z^2pq}{B^2} \qquad (7)$$

Que es la manera de calcular el tamaño de muestra cuando se desconoce el tamaño de la población o la proporción $\frac{n}{N} < 0.05$.

Como sugerencia, cuando se desconoce la proporción poblacional y no se cuenta con información muestral, considerar que P = 0.5.

Ejemplo 1. De un listado de 8,500 establecimientos comerciales se desea extraer una muestra para conocer la proporción de negocios que aplica el proceso administrativo. Se desea que la estimación no difiera del valor real en más de diez puntos porcentuales con una confiabilidad de 95%. ¿De qué tamaño debe ser la muestra?

Respuesta

Se desea estimar una proporción (proporción de negocios que aplica el proceso administrativo), la información que proporciona el problema es:

N = 8,500
B = 10% = 0.1
1-α = 0.95, es decir, α=0.05

En consecuencia, *z* = 1.96. Como no se cuenta con información de *P* se asumirá que es 0.5.

Aplicando la fórmula (6) se llega al resultado.

$$n = \frac{(8,500)(1.96)^2(0.5)(0.5)}{(8,500)(0.1)^2 + (1.96)^2(0.5)(0.5)}$$

$$n = 94.97$$

Es decir, se requiere una muestra de 95 establecimientos para garantizar una estimación que difiera en 10% de la proporción real con una confiabilidad de 95%.

Ejemplo 2. ¿De qué tamaño sería la muestra en el ejemplo anterior si se desconociera el tamaño de la población (N) y el resto de los parámetros se mantuviera igual?

Respuesta

Del ejemplo anterior se conoce que *B*=10%, *z*= 1.96 y *P* = 0.5. Sustituyendo estos valores en la fórmula (7) se obtiene:

$$\boldsymbol{n = \frac{(1.96)^2(0.5)(0.5)}{0.1^2}}$$

$$\boldsymbol{n = 96.04}$$

Es decir, se requiere una muestra de 96 establecimientos para garantizar una estimación que difiera en 10% de la proporción real con una confiabilidad de 95%.

Obsérvese que los resultados de los ejemplos 1 y 2 son semejantes, esto debido a que la proporción $\boldsymbol{\frac{n}{N} = \frac{95}{8{,}500} = 0.01 < 0.05}$.

3.7. Otros intervalos de confianza

En los subtemas anteriores, se realizaron estimaciones de la media y proporción poblacionales, ahora, se mostrará cómo hacer estimaciones de la varianza poblacional a partir de una muestra.

Intervalo de confianza para la varianza de la población

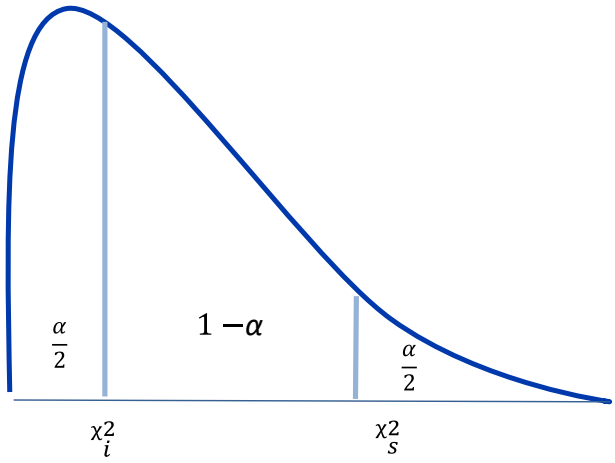
Para estimar el valor de la varianza poblacional a partir de una muestra, se utilizará el siguiente resultado:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Este resultado indica que el cociente $\frac{(n-1)S^2}{\sigma^2}$ tiene una distribución ji cuadrada con n – 1 grados de libertad. De esta manera, para construir un intervalo de confianza para la varianza poblacional, se empleará una distribución ji cuadrada con n – 1 grados de libertad.

Para construir un intervalo de confianza que contenga a la varianza poblacional, se deberán encontrar dos cuantiles de una distribución ji cuadrada tal que el área contenida entre ambos puntos sea 1– α, donde α es un valor entre 0 y 1, tal como lo muestra la figura 10.

Figura 10. Ubicación de los cuantiles de una distribución ji cuadrada que permitan construir un intervalo que contenga a σ^2 con un nivel de confianza 1-α



La figura anterior ilustra la ubicación de los cuantiles que encierran una región cuya área es 1 – α. El valor superior (χ^2_s) separa la curva en dos regiones, donde la derecha tiene un área de $\frac{\alpha}{2}$). El valor inferior (χ^2_i), por su parte, separa la curva en dos regiones, donde la izquierda tiene un área de $\frac{\alpha}{2}$.

Para obtener estos valores, se puede recurrir a tablas o utilizar un *software*. A continuación se explica cómo calcular estos cuantiles en MS-Excel.

En Excel se emplea la fórmula:

PRUEBA.CHI.INV $(probabilidad, grados\_de\_libertad)$

- El parámetro de *probabilidad* se refiere a la región que se acumula a la derecha del cuantil, y *grados_de_libertad* es el tamaño de la muestra menos uno.

Supóngase que se desea encontrar los puntos críticos que garantizan un nivel de confianza del 90% con una muestra de 30 elementos.

Se establece el nivel de confianza:

Nivel de confianza = 90%, es decir, 0.90

Se determina α:

α = 1 – nivel de confianza

α = 1 – 0.90

α = 0.1

Se divide alfa entre 2:

$$\frac{\mathbf{0.1}}{\mathbf{2}} = \mathbf{0.05}$$

Se calcula el punto que corta la curva en dos regiones: una de área $\frac{\alpha}{2}$ (región derecha) y otra de $\mathbf{1} - \frac{\alpha}{2}$ (región izquierda), con la fórmula:

PRUEBA.CHI.INV ($\frac{\alpha}{2}$, n-1)

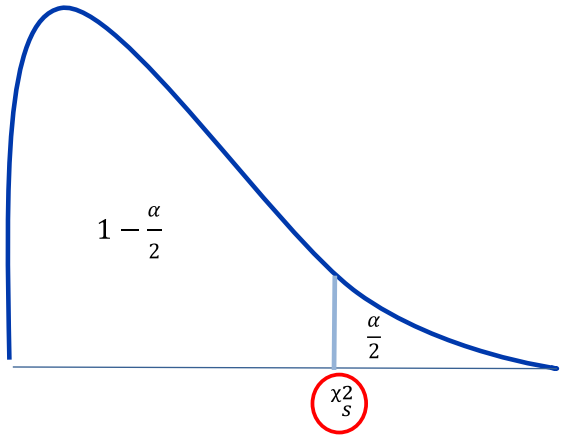
PRUEBA.CHI.INV ($\frac{0.1}{2}$, 30-1)

PRUEBA.CHI.INV (0.05,29)

$\chi_s^2 = 42.556$

La figura 11 representa el punto que se está obteniendo.

Figura 11. Ilustración del punto que se obtiene con la fórmula PRUEBA.CHI.INV (α/2, n – 1)



En la figura 11 se representa el resultado obtenido al aplicar la fórmula. Nótese que el cuantil divide la curva en dos regiones: la de la derecha tiene un área de $\frac{\alpha}{2}$; y la de la izquierda, de $\mathbf{1} - \frac{\alpha}{2}$.

A continuación se calcula el punto que corta la curva en dos regiones: una de área $\mathbf{1} - \frac{\alpha}{2}$ (región derecha) y otra de $\frac{\alpha}{2}$ (región izquierda), con la fórmula:

PRUEBA.CHI.INV ($\mathbf{1} - \frac{\alpha}{2}$, n-1)

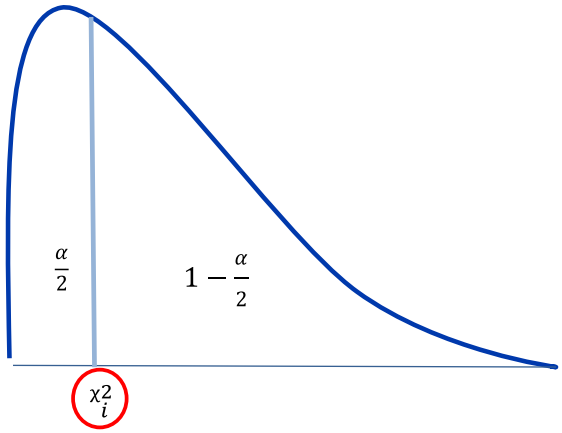
PRUEBA.CHI.INV (0.975, 30-1)

PRUEBA.CHI.INV (0.975,29)

$\mathbf{X_i^2} = 16.047$

La figura 12 ilustra el punto que se obtiene.

Figura 12. Ilustración del punto que se obtiene con la fórmula PRUEBA.CHI.INV (1-α/2, n-1)



La figura anterior ilustra el resultado al aplicar la fórmula. Nótese que el cuantil divide la curva en dos regiones: la de la derecha tiene un área de $\mathbf{1} - \frac{\alpha}{2}$; y la de la izquierda, de $\frac{\alpha}{2}$.

Obtenidos los cuantiles, se procede a plantear la siguiente desigualdad:

$$X_i^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq X_s^2$$

Despejando el parámetro poblacional σ^2, el intervalo queda de la siguiente manera:

$$\frac{(n-1)S^2}{X_s^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{X_i^2}$$

La tabla 7 presenta los elementos requeridos para realizar una estimación de intervalo de una varianza poblacional.

Tabla 7. Elementos requeridos para realizar una estimación de intervalo de una varianza poblacional

Parámetro población	Estimador	Fórmula		
		Intervalo de Confianza	Estimador puntual	Varianza muestral
Varianza σ^2	$\hat{\sigma}^2$	$LI = \frac{(n-1)S^2}{X_s^2}$ $LS = \frac{(n-1)S^2}{X_i^2}$	$\hat{\sigma}^2 = \frac{(n-1)S^2}{X^2}$	$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

A diferencia de los estimadores de intervalo empleados en secciones anteriores, para el intervalo de una varianza poblacional se calculan directamente los límites del intervalo sin necesidad de utilizar el estimador puntual.

A continuación, se ejemplifica cómo estimar un intervalo para la varianza poblacional.

Un parque de diversiones es visitado en promedio por 50 000 personas al mes. Con la finalidad de diseñar una promoción que incentive el consumo de productos ofrecidos por el parque, el gerente quiere conocer la variabilidad del gasto de las familias que lo visitan en un día. Para tal efecto, se entrevistó a 30 familias elegidas al azar y se registró su consumo durante su estadía.

En la siguiente tabla se muestra la información recabada.

Factura	Gasto al día	Factura	Gasto al día
1	645	16	470
2	1,177	17	1,264
3	524	18	436
4	1,192	19	645
5	746	20	409
6	803	21	709
7	1,612	22	1,009
8	382	23	1,180
9	571	24	1,410
10	697	25	377
11	792	26	1,283
12	442	27	1,321
13	959	28	1,534
14	881	29	675
15	1,506	30	1,625

a. Realizar una estimación por intervalo para la desviación poblacional con un nivel de confianza del 90%.

b. Interpretar los resultados.

Respuestas:

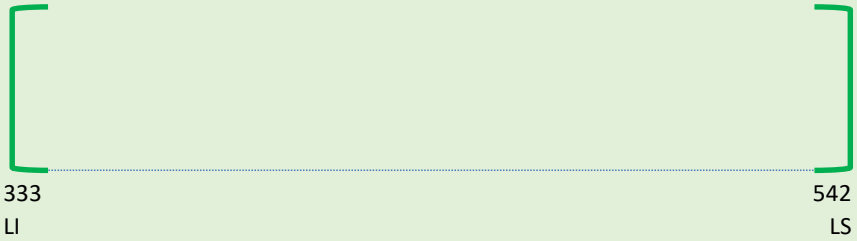
Determinar la fórmula para realizar la estimación por intervalo de la varianza poblacional: $$\frac{(n-1)S^2}{X_s^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{X_i^2}$$	Después se calcula el valor de la varianza muestral S^2 y los valores de X_i^2 y X_s^2 para poder obtener los límites inferior y superior. Varianza muestral S^2: $$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$
Para realizar el cálculo, se requiere calcular previamente el promedio de los datos de la muestra: $$\bar{x} = \frac{\sum x_i}{n}$$ $$\bar{x} = \frac{645 + 1,177 + \cdots + 675 + 1625}{30}$$ $$\bar{x} = \frac{27,276}{30}$$ $$\bar{x} = 909$$	De esta manera: $$S^2 = \frac{(645-909)^2 + (1,177-909)^2 + \cdots + (675-909)^2 + (1,625-909)^2}{30-1}$$ $$S^2 = \frac{4,717,528.8}{29}$$ $$S^2 = 403.32$$ Establecer el nivel de confianza para calcular los puntos de corte: Nivel de confianza = 90%, es decir, 0.90
Determinar el valor de α: α = 1 – nivel de confianza α = 1 – 0.90 α = 0.1 Calcular el valor del punto de corte X_s^2 con la función de MS-Excel: *PRUEBA.CHI.INV* ($\frac{\alpha}{2}$, *n-1)* PRUEBA.CHI.INV ($\frac{0.1}{2}$, 30-1) PRUEBA.CHI.INV (0.05,29) $X_S^2 = 42.556$	Calcular el valor del punto de corte X_i^2 con la función de MS-Excel: PRUEBA.CHI.INV ($1 - \frac{\alpha}{2}$, n-1) PRUEBA.CHI.INV (0.975, 30-1) PRUEBA.CHI.INV (0.975,29) $X_i^2 = 16.047$ Calcular el límite inferior (LI) y superior (LS) del intervalo (IC): n = 30 $S^2 = 403.32$ $X_i^2 = 16.04$ $X_s^2 = 42.55$ $$\sigma_i^2 = \frac{(n-1)S^2}{X_i^2}$$
Límite inferior: $$LI = \frac{(30-1)(403.32)}{42.55}$$ $$LI = \frac{(29)(162,673.407)}{42.55}$$ $$LI = \frac{4,717,528.8}{42.55}$$	Límite superior: $$LS = \frac{(30-1)(403.32)}{42.55}$$ $$LS = \frac{(29)(162,673.407)}{16.04}$$ $$LS = \frac{4,717,528.8}{16.04}$$

$LI = 110,852.08$ Calcular la raíz cuadrada para obtener el valor inferior de la desviación poblacional: $\sigma_{LI} = 332.94$	$LS = 293,980.664$ Calcular la raíz cuadrada para obtener el valor superior de la desviación poblacional: $\sigma_{LS} = 542.19$

Construir el intervalo de confianza para la desviación poblacional:

$$IC = (LI, LS)$$

$$IC = (333, 542)$$



Así, la desviación poblacional se ubica en un rango de entre $333 y $542 con un 90% de confianza. La promoción que establezca el parque de diversiones deberá estar en este rango respecto al gasto promedio familiar.

RESUMEN

Se expusieron las bases para llevar a cabo la estimación de parámetros, uno de los principales temas de la estadística inferencial. Se mostraron los tipos de estimaciones que pueden realizarse de un parámetro (puntual y por intervalo) y se analizaron las posibles fuentes de error de estimación, que pueden ser atribuibles a la muestra recabada o a otras causas. Es factible que el riesgo de error de estimación por causas no atribuibles al muestreo disminuya a través de un buen diseño del instrumento, de la logística y considerando la posible no respuesta.

Por otro lado, se expusieron las propiedades que se buscan en los estimadores: que sean insesgados, eficientes y consistentes.

Se mostró también cómo realizar estimaciones puntuales y por intervalo de una media poblacional utilizando muestras grandes y pequeñas. De igual manera, la forma de efectuar estimaciones de una proporción poblacional. Finalmente, se dieron los elementos para construir intervalos de confianza para la varianza y desviación poblacional.

El valor agregado de este material fue presentar el uso de fórmulas de MS-Excel para calcular los cuantiles de las distribuciones normal estándar, t de Student y χ^2, que expresan los niveles de confianza deseados para realizar estimaciones por intervalo.

BIBLIOGRAFÍA

SUGERIDA

Autor	Capítulo	Páginas
Anderson, S.	8	308-332
Levin, R.	7	273-308
Lind, D.	9	297-319

UNIDAD 4

Pruebas de hipótesis

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

OBJETIVO PARTICULAR

Al terminar la unidad, el alumno conocerá las pruebas de hipótesis y su aplicación.

TEMARIO DETALLADO

(10 horas)

4. Pruebas de hipótesis

- 4.1. Planteamiento de las hipótesis
- 4.2. Errores tipo I y tipo II
- 4.3. Pruebas de uno y de dos extremos, y regiones de aceptación y de rechazo
- 4.4. Pruebas de hipótesis para una media poblacional
- 4.5. Tres métodos para realizar pruebas de hipótesis
  - 4.5.1. El método del intervalo
  - 4.5.2. El método estadístico de prueba
  - 4.5.3. El método del valor de la P
- 4.6. Prueba de hipótesis sobre una proporción poblacional
- 4.7. Pruebas de hipótesis sobre la diferencia entre dos medias
- 4.8. Pruebas de hipótesis sobre la diferencia entre dos proporciones
- 4.9. Prueba para la diferencia entre dos varianzas

INTRODUCCIÓN

Hay dos temas que tienen mayor importancia en estadística inferencial: estimación de parámetros y pruebas de hipótesis. En la unidad anterior, se expuso el tema de estimación; ahora, se abordará el de pruebas de hipótesis.

En la práctica profesional es habitual enfrentar situaciones donde se requiere apoyar o no un supuesto sobre el comportamiento de una distribución. Por ejemplo, al área de mercadotecnia de una empresa de cosméticos le gustaría comprobar si el 75% de las mujeres entre 30 y 40 años que laboran utilizan un labial de color café. O un auditor desearía comprobar que el 60% de las licitaciones de cierta organización no cumplieron con las bases de la convocatoria. O al coordinador del departamento de Matemáticas le interesaría conocer si existe diferencia en el aprovechamiento de la materia de Estadística entre los alumnos de Administración y Contaduría.

En esta unidad se expone cómo plantear hipótesis, los tipos de errores que se pueden cometer y las clases de pruebas. Además, se muestra cómo realizar contrastes de hipótesis para una y dos medias, una y dos proporciones, y dos varianzas.

4.1. Planteamiento de las hipótesis

En todas las áreas del conocimiento donde se aplica el método científico, el planteamiento de hipótesis desempeña un papel central. Después de observar una situación, toda investigación parte de una hipótesis, la cual buscará apoyarse o no con la evidencia recabada en una muestra. Por ejemplo, un investigador de las ciencias administrativas podría estar interesado en demostrar que la proporción de PYMES que fracasan los primeros cinco años de vida es mayor en el sector comercial que en el de servicios. El gerente de marca de un producto desearía demostrar que las ventas de su producto aumentan 10% si el tiempo de promoción en radio es mayor a 25 minutos al día. O el coordinador de Matemáticas pretendería demostrar que no hay diferencia en el desempeño de los alumnos del turno vespertino respecto al matutino.

En este apartado, se mostrará qué es una hipótesis estadística, sus partes y cómo plantearla.

Hipótesis estadística

Una hipótesis estadística es un enunciado sobre el comportamiento de un parámetro poblacional o de la distribución de una variable aleatoria, la cual se confirmará a través de una *prueba de hipótesis* que utiliza la información de una muestra.

Elementos de una prueba de hipótesis:

Planteamiento de hipótesis

El planteamiento de hipótesis consiste en definir tanto la hipótesis nula como la alternativa de forma que involucre el parámetro a inferir. Es deseable plantear en la hipótesis nula que el parámetro de interés es igual a cierto valor ($\boldsymbol{\theta = \theta_0}$); y en la alternativa, que es menor, mayor o diferente.

La notación que se emplea para plantear hipótesis es identificar la hipótesis nula o alternativa *(H_o o H_a)* y separar con ":" el enunciado.

Como ejemplo, supóngase que el gerente de producto de cierta organización desea comprobar a los directivos que el principal competidor ocupa en promedio más de 30 minutos al día en sus promocionales insertados en radio. En este caso, el parámetro $\boldsymbol{\theta}$ de interés es el promedio poblacional (μ), y la hipótesis que el gerente pretende probar es que μ > 30, lo que la convierte en una hipótesis alternativa, y en consecuencia la hipótesis nula es que μ ≤ 30 (como esta región contiene la igualdad, es suficiente utilizar un solo punto de esta región). De esta manera, el planteamiento de hipótesis queda así:

$$H_0 : \mu = 30$$
$$H_a : \mu > 30$$

Para plantear una hipótesis, se sugiere dar estos pasos:

1. Identificar el parámetro poblacional que se desea inferir.
2. Identificar la hipótesis alternativa.
3. Identificar la hipótesis nula.

Una vez que se ha planteado la hipótesis, el siguiente paso es definir la precisión de la prueba, lo cual se explicará en la siguiente sección.

4.2. Errores tipo I y tipo II

Para realizar una prueba de hipótesis, se requiere recabar una muestra. Esto implica asumir que la inferencia tendrá una desviación respecto al comportamiento real.

Se corre el riesgo de dos situaciones que provoquen inferencias equivocadas al realizar un contraste de hipótesis.

1. Primera situación consiste en rechazar la hipótesis nula cuando es verdadera, a este caso se le conoce como el *error tipo I* y se le denota como α (probabilidad de que se presente esta situación).

2. Segundo caso se refiere a rechazar la hipótesis alternativa cuando es cierta, a este error se le denomina *error tipo II* y se denota como β (probabilidad de caer en esta situación).

En la tabla 1, se ilustran los escenarios posibles al contrastar hipótesis.

Tabla 1. Escenarios posibles al contrastar hipótesis

		Resultado de la prueba	
		Se acepta H_o	Se rechaza H_o
Situación real de H_o	Verdadera	**Decisión correcta** Se acepta H_o y es verdadera	**Error tipo I α** Se rechaza H_o y es verdadera
	Falsa	**Error tipo II β** Se acepta H_o y es falsa	**Decisión correcta** Se rechaza H_o y es falsa

Fuente: elaboración propia.

La tabla anterior muestra los escenarios posibles al realizar una prueba de hipótesis. El primero ocurre cuando se toma una decisión correcta, ya sea aceptando o rechazando la hipótesis cuando lo amerite. Los otros escenarios son los errores mencionados previamente.

Al trabajar con pruebas de hipótesis en la práctica, se fija el error tipo I, que se permite, lo cual se conoce como “nivel de significancia de la prueba”. Los valores más comunes son de 0.05 y 0.01. Cuando los resultados de la prueba no tienen consecuencias importantes, puede incrementarse el nivel de significancia. Es factible controlar este error desde el diseño del muestreo.

Este material se centra en controlar el error I. Para el manejo del error tipo II, se recomienda consultar a Anderson (2012, pp. 382-387).

4.3. Pruebas de uno y de dos extremos y regiones de aceptación y de rechazo

Existen dos pruebas:

Prueba de un estremo	Prueba de dos extremos
• Se contrasta que el valor del parámetro sea notablemente mayor o menor al fijado en la hipótesis nula.	• Se contrasta que el valor sea diferente al establecido en la hipótesis nula, es decir, puede ser notablemente mayor o menor.

El tipo de prueba se define con la hipótesis alternativa.

En la prueba de hipótesis se buscará un valor crítico a partir del cual se rechazará todo resultado en la prueba que se encuentra en esta zona (*región de rechazo*).

Cuando la prueba es de una cola, la región de rechazo se localizará en un extremo de la distribución del estadístico de prueba; y si es de dos extremos, en los extremos de la distribución.

En la tabla 2 se resume esta información.

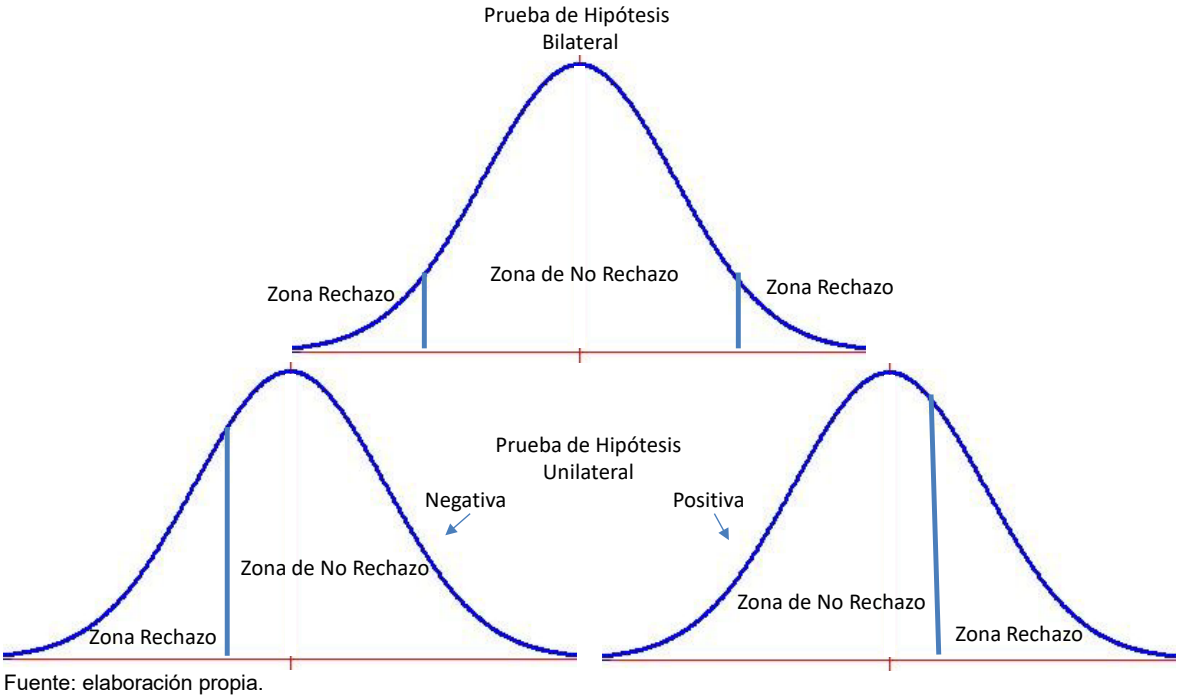
Tabla 2. Características de los tipos de pruebas de hipótesis

Signo de la H_a	**Prueba de hipótesis**	**Número de extremos**	**Ubicación de la zona de rechazo en la distribución**
≠	Bilateral	Dos	Extremos positivo y negativo
>	Unilateral	Una	Extremo positivo
<			Extremo negativo

Fuente: elaboración propia.

En la figura 1, se ilustran los tipos de pruebas de hipótesis y sus zonas de rechazo.

Figura 1. Tipos de pruebas de hipótesis y zonas de rechazo



Fuente: elaboración propia.

La figura anterior muestra que las regiones de rechazo se encuentran a partir de un valor que se denominará *crítico.*

El valor crítico

Se calcula a partir de la significancia que se defina en la prueba. Supóngase que la distribución del estadístico de prueba es una normal estandarizada. Para calcular el valor crítico, se emplea la siguiente función de Excel:

- DISTR.NORM.ESTAND.INV(probabilidad)

Donde la probabilidad es 1 – α. El resultado será el cuantil que separa la curva en dos partes: 1 – α y α.

Como ejemplo, supóngase que se estableció un nivel de significancia (α) de 1%. Se pide determinar el valor crítico para una prueba de una o dos colas cuando la distribución del estadístico de prueba es una distribución normal estandarizada.

a. Para una prueba de una cola (derecha)	α = 0.01 Probabilidad = 1 – 0.01 Probabilidad = 0.99 DISTR.NORM.ESTAND.INV(0.99) = 2.32
b. Para una prueba de una cola (izquierda)	Por la propiedad de simetría de una distribución normal, el punto crítico es –2.32
c. Para una prueba de dos colas, se buscarán dos puntos críticos y la región de rechazo se dividirá en dos partes iguales. Partiendo de:	α = 0.01 α se divide entre 2: $\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$ Probabilidad = 1 – 0.005 Probabilidad = 0.995 DISTR.NORM.ESTAND.INV(0.995) = 2.575

Por la propiedad de simetría de la distribución normal, los puntos críticos son ± 2.575

Para realizar la prueba de hipótesis, además de una muestra, se requiere un estadístico de prueba, regla que arroja un valor aleatorio para determinar el resultado de la prueba.

Fórmula general de un estadístico de prueba:

$$EP = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

- Donde:

EP = Estadístico de prueba
θ_0 = Parámetro poblacional asumiendo cierta la hipótesis nula
$\hat{\theta}$ = Estimador del parámetro
$\sigma_{\hat{\theta}}$ = Desviación estándar del estimador

El estadístico de prueba es una variable aleatoria debido a que su valor dependerá de los elementos que conforman la muestra. En el resto de la unidad, se expondrán las distribuciones muestrales asociadas a esos elementos.

Se cuenta con un estadístico de prueba para realizar una prueba con algún parámetro. La tabla 3 contiene un resumen de los estadísticos de prueba a emplear en esta unidad.

Tabla 3. Resumen de los elementos que conforman los estadísticos de prueba para los casos a estudiar en esta unidad

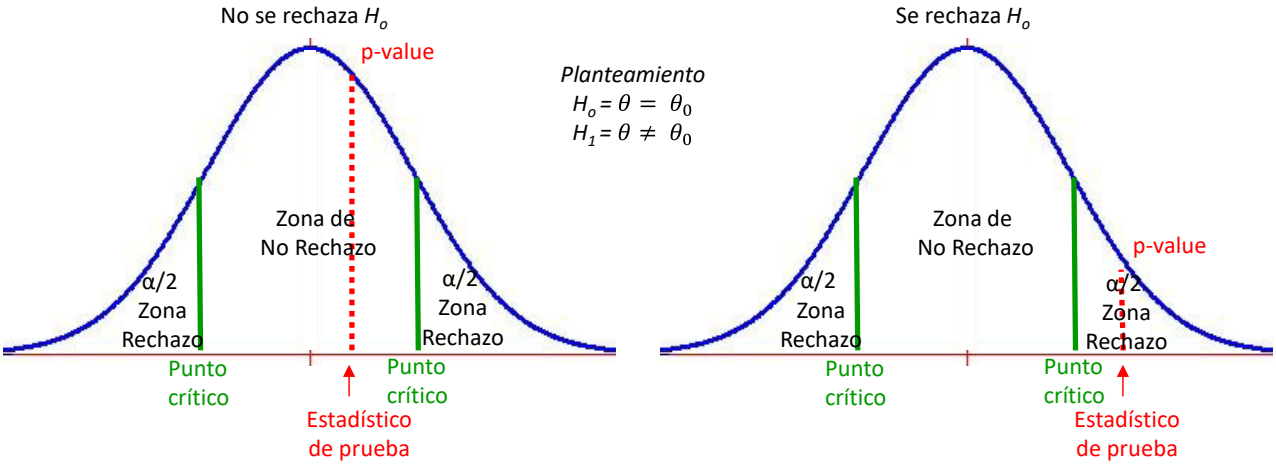
Parámetro poblacional	Estimador	Fórmula	
		Desviación estándar del estimador	Estadístico de prueba
Promedio μ	$\bar{x}$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ Si no se conoce σ, se sustituye por s	$EP = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ Si no se conoce σ, se sustituye por s
Proporción P	p	$\sigma_p = \sqrt{\frac{PQ}{n}}$ *q=1-p*	$EP = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$
Diferencia de medias $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ Si no se conocen σ_1, σ_2, se sustituyen por s_1 y s_2	$EP = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ Si no se conocen σ_1, σ_2, se sustituyen por s_1 y s_2
Diferencia de proporciones $p_1 - p_2$	$p_1 - p_2$	$\sigma_{p_1 - p_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$	$EP = \frac{p_1 - p_2}{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Comparación de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$	$\frac{s_1^2}{s_2^2}$		$EP = F = \frac{S_1^2}{S_2^2}$

Fuente: elaboración propia.

En las siguientes secciones se expondrá cada caso.

Para terminar este apartado, en la figura 2 se ilustran los elementos que conforman una prueba de hipótesis de dos colas con un estadístico de prueba con distribución normal estándar.

Figura 2. Elementos de una prueba de hipótesis



Fuente: elaboración propia.

La figura anterior muestra dos situaciones: la aceptación de la hipótesis nula (figura izquierda) y su rechazo (figura derecha).

4.4. Pruebas de hipótesis para una media poblacional

La primera prueba que se abordará en esta unidad se relaciona con el promedio poblacional (μ). Para estimar este parámetro, se emplea el promedio muestral ($\overline{x}$). En la segunda unidad, se mencionó que la distribución muestral de la media se acerca a una normal cuando la varianza poblacional es conocida o si la muestra es de tamaño de 30 o más elementos. Si se desconoce la varianza, la distribución muestral es una t con n – 1 grados de libertad, la cual se aproxima a una normal estandarizada conforme se incrementa la muestra.

Estadístico de prueba a utilizar:

$$EP = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Donde:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Cuando se conoce la varianza poblacional, en caso contrario, se sustituye por *s.*

A continuación, se muestran ejemplos sobre la realización de pruebas con la media.

Ejemplo 1

El gerente del área médica de una empresa farmacéutica sabe por experiencia que el promedio de visitas diarias a los médicos efectuadas por los representantes es de 6 con una desviación estándar de 1.22. Entonces, el gerente del área llama la atención a los representantes médicos, pues considera que el promedio de visitas ha disminuido en los últimos tres meses. A fin de comprobarlo, toma una muestra de 30 representantes, donde encuentra que el promedio es de 6.06 con una desviación de 1.41. Considerando una significancia de 0.05, ¿tiene razón el gerente?

Solución:

Para resolver la prueba, se darán los siguientes pasos.

1. Identificar los datos

- El parámetro a probar es μ:
  - $\mu = 6$
  - $n = 30$
  - $\bar{x} = 6.06$
  - $s = 1.41$
  - α = 5% (0.05)

2. Definir las hipótesis

- La hipótesis nula se establece con el valor histórico (μ = 6). La hipótesis alternativa se encuentra en este segmento del enunciado del problema: "el gerente del área llama la atención a los representantes médicos, pues considera que el promedio de visitas ha disminuido en los últimos tres meses…".

La prueba queda planteada así:

$H_0 : \mu = 6$

$H_1 : \mu < 6$

Es una prueba de un extremo (izquierdo).

3. Como el estadístico de prueba sigue una distribución normal estandarizada, se determinará el valor crítico y la zona de rechazo con las fórmulas que corresponden en Excel.

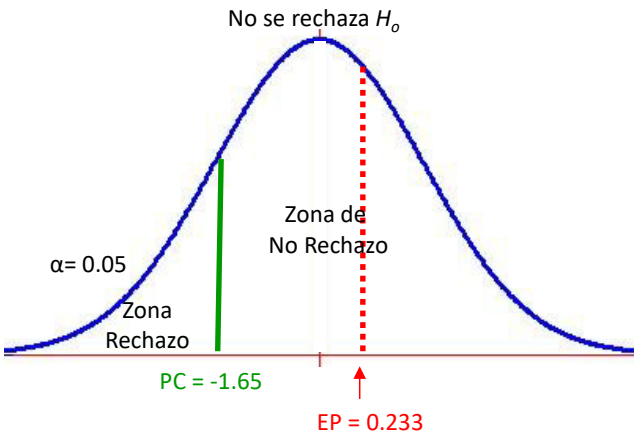
- Primero, se calculará el parámetro *probabilidad*:
- Probabilidad = 1 – 0.05 = 0.95
- De esta manera, se calcula el punto crítico (PC) con la siguiente fórmula:
- PC = DISTR.NORM.ESTAND.INV(0.95)
- PC = 1.645 = 1.65
- Como la prueba es de un extremo a la izquierda por la propiedad de simetría de la distribución normal, el valor crítico es –1.65

4. Se realizan los cálculos del estadístico de prueba con la fórmula correspondiente:

- $EP = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- $EP = \frac{6.06 - 6}{\frac{1.41}{\sqrt{30}}}$
- $EP = \frac{0.06}{\frac{1.41}{5.477}}$
- $EP = \frac{0.06}{0.2574}$
- $EP = 0.233$

5. Se construye una gráfica donde se ubica la zona de rechazo y el valor arrojado por el estadístico de prueba.

- El EP se sitúa en la región de no rechazo, así que H_0 no se rechaza a una significancia del 0.05. Es decir, no hay elementos para apoyar la idea del gerente de área de que el promedio de visitas médicas ha disminuido en los últimos tres meses.



Ejemplo 2

El promedio de un examen de conocimientos aplicado por una universidad cada año ha sido de 7.25. El director sospecha que el promedio de calificaciones disminuyó el último año, por lo que solicitó realizar un estudio tomando una muestra de 40 alumnos con una significancia del 10%.

Los resultados de calificaciones obtenidas por los 40 alumnos seleccionados se muestran a continuación.

Alumno	Calificación	Alumno	Calificación
1	2	21	7
2	7	22	0
3	6	23	1
4	2	24	0
5	1	25	6
6	0	26	5
7	10	27	10
8	9	28	8
9	9	29	10
10	5	30	3
11	7	31	8
12	5	32	8
13	3	33	1
14	5	34	2
15	5	35	8
16	10	36	3
17	4	37	8
18	8	38	8
19	7	39	10
20	3	40	9

Con una significancia de 10%, ¿se apoya la hipótesis del director?

Solución:

Para resolver la prueba, se darán los siguientes pasos.

1. Identificar los datos

El parámetro a probar es μ:

$\mu = 7.25$

$n = 40$

α = 10% (0.10)

2. Calcular el promedio y desviación muestral. Se pueden emplear las fórmulas promedio() y desvest()de Excel, o calcularlos con sus respectivas fórmulas.

Promedio muestral:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{223}{40}$$

$$\bar{x} = 5.57$$

Desviación estándar muestral:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{407.775}{40-1}} = \sqrt{\frac{407.775}{39}} = \sqrt{10.455}$$

$s = 3.23$

3. Definir las hipótesis

En este ejemplo, la hipótesis nula se establece con el valor histórico (μ = 7.25). La hipótesis alternativa se encuentra en este segmento del enunciado del problema: “El director sospecha que el promedio de calificaciones disminuyó el último año”.

La prueba se plantea así:

$$H_0 : \mu = 7.25$$

$$H_1 : \mu < 7.25$$

La prueba es de un extremo (izquierdo).

4. Como el estadístico de prueba sigue una distribución normal estandarizada, se determinará el valor crítico y la zona de rechazo con las fórmulas que correspondan en Excel.

Primero, se calcula el parámetro *probabilidad*:

Probabilidad = 1 – α= 0.9

De esta manera, se calcula el punto crítico:

PC = DISTR.NORM.ESTAND.INV(0.90)

PC = 1.281 = 1.28

Como la prueba es de un extremo a la izquierda por la propiedad de simetría de la distribución normal, el valor crítico es –1.28.

5. Se realizan los cálculos del estadístico de prueba con la fórmula siguiente:

$$EP = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$EP = \frac{5.57 - 7.25}{\frac{3.23}{\sqrt{40}}}$$

$$EP = \frac{-1.68}{\frac{3.23}{6.3245}}$$

$$EP = \frac{-1.68}{0.5112}$$

$$EP = -3.28$$

6. Se construye la gráfica para determinar la zona donde se ubica el valor del estadístico de prueba.

El EP se sitúa en la región de rechazo, así que H_0 no se acepta a una significancia de 0.10. Es decir, hay elementos para apoyar al director de que el promedio del examen de conocimientos disminuyó el último año.

Ejemplo 3

El gerente de producto de una marca de ropa conoce que a nivel nacional los hogares de su segmento de mercado destinan en promedio al mes $2,045 en la compra de ropa y calzado. El gerente piensa que los miembros de su programa de CRM (Customer Relationship Management) no gastan esa cantidad. Entonces, para diseñar una estrategia de venta con los miembros de su programa de CRM, entrevista a una muestra elegida al azar de 20 hogares, a quienes pregunta la cantidad de dinero destinada a vestido y calzado al mes. La muestra arrojó que en promedio un hogar miembro del CRM gasta al mes $1,930 con una desviación del $680. ¿Los resultados anteriores apoyan la hipótesis del gerente, con una significancia del 5%?

Solución:

Nuevamente, se siguen los pasos de los ejemplos anteriores.

1. Establecer los datos

- El parámetro a probar es μ:

μ = $2,045
n = 20
α = 5% (0.05)
$\bar{x}$ = $1,930
s = $680

2. Definir las hipótesis

- La hipótesis nula se establece con el valor conocido de la población (μ = 2,045). La hipótesis alternativa se encuentra en este segmento del enunciado del problema: "El gerente piensa que los miembros de su programa de CRM (Customer Relationship Management) no gastan esa cantidad".

La prueba queda planteada así:
H_0: $\mu = 2{,}045$
H_1: $\mu \neq 2{,}045$
Es una prueba de dos extremos.

3. Como se desconoce la desviación poblacional y además el tamaño de la muestra es inferior a 30 elementos, el estadístico de prueba sigue una distribución t de Student con 19 grados de libertad. Se determinará el valor crítico y la zona de rechazo con las fórmulas que correspondan en Excel.

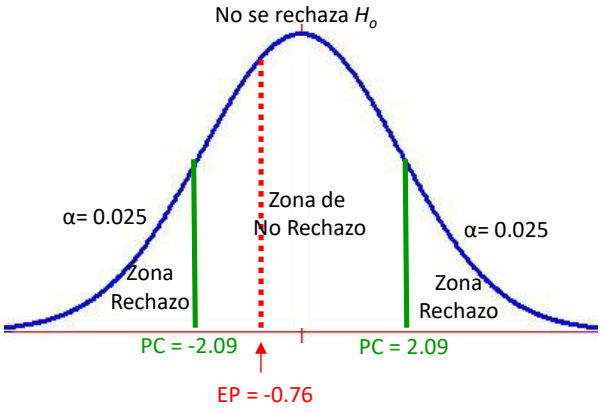
- En este caso, los puntos críticos se obtienen con la siguiente fórmula:
- PC = DISTR.T.INV(0.05,19) = 2.0930
- Como la prueba es de dos extremos y la fórmula considera esta situación, los puntos críticos son ± 2.09.

4. Se realizan los cálculos del estadístico de prueba con la fórmula correspondiente:

- $EP = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
- $EP = \frac{1{,}930 - 2{,}045}{\frac{680}{\sqrt{20}}}$
- $EP = \frac{-115}{\frac{680}{4.4721}}$
- $EP = \frac{-115}{152.05}$
- $EP = -0.7563$

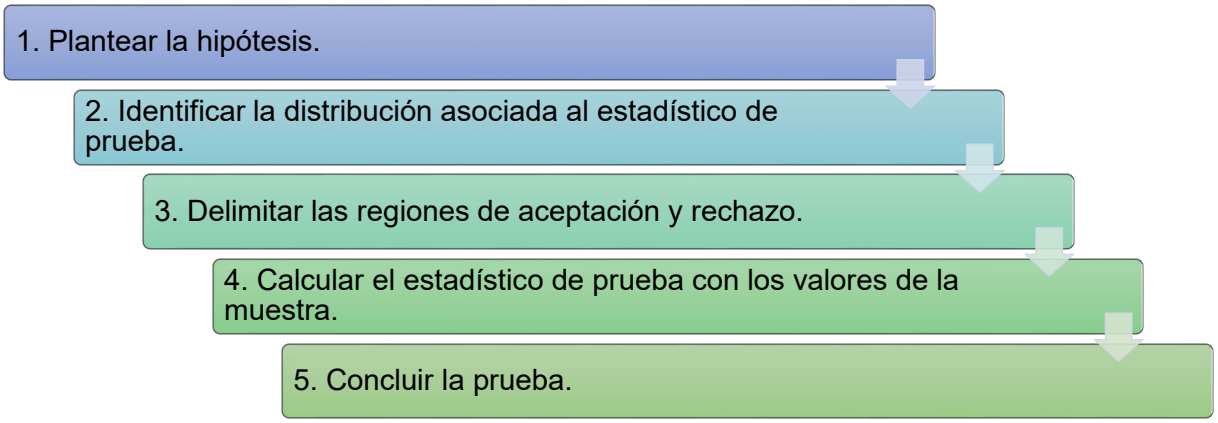
5. Se dibuja la gráfica para determinar la zona donde se encuentra el valor del estadístico de prueba.

El EP no se encuentra en la región de rechazo, así que H_0 no se rechaza a un nivel de significancia de 5%. Es decir, no hay elementos para no apoyar que los hogares miembros del CRM destinan al mes $2,045 en vestido y calzado.



4.5. Tres métodos para realizar pruebas de hipótesis

En la sección anterior, se expuso cómo realizar pruebas de hipótesis para una media, siguiendo estos pasos:



Para el punto 5, se pueden aplicar tres criterios: basarse en el valor del estadístico de prueba, utilizar el p-value o emplear un intervalo de confianza. La siguiente sesión muestra la manera de concluir una prueba de hipótesis utilizando los tres diferentes métodos.

4.5.1. El método del intervalo

En la unidad anterior, se aprendió a hacer estimaciones de un parámetro poblacional a través de un intervalo de confianza. En este apartado, dicho cálculo servirá para concluir una prueba de hipótesis. Se aplicará la siguiente regla:

Si el valor de μ_0 se encuentra en el intervalo de confianza, entonces no se rechaza H_0

- Este criterio del intervalo de confianza se emplea en pruebas bilaterales donde la confiabilidad del intervalo será 1 – α.

Para ejemplificar el uso del método del intervalo de confianza, se retomará el ejemplo 3 de la sesión anterior, donde con una significancia de 5% se deseaba probar que los hogares miembros del programa de CRM destinaban al mes en promedio una cantidad diferente a $2,045 en vestido y calzado. Para realizar la prueba, se empleó una muestra aleatoria de 20 familias, quienes en promedio destinaban mensualmente $1,930 en adquirir vestido y calzado con una desviación del $680.

El planteamiento de la hipótesis fue el siguiente:

H_0: μ = $2,045
H_1: μ ≠ $2,045

El estimador empleado para determinar el valor de μ es el promedio muestral, con una distribución t de Student con 19 grados de libertad, porque se desconoce la varianza poblacional.

Al conocer la distribución del promedio muestral, se puede estimar un intervalo de confianza que contenga a μ.

Con el método de intervalo de confianza, se acepta la hipótesis nula si $2,045 se encuentra contenido. Como la significancia de la prueba (α) es de 5%, la confiabilidad del intervalo es 1 – α = 1 – 0.05 = 0.95.

De esta manera, el intervalo de confianza (IC) con 95% de confiabilidad resulta:

$$IC = \bar{x} \pm t\frac{s}{\sqrt{n}}$$

$$IC = 1,930 \pm 2.09 \cdot \frac{680}{\sqrt{20}}$$

$$IC = 1,930 \pm 318.25$$

$$IC = (1,611.75 - 2,248.25)$$

Como el valor del parámetro bajo la hipótesis nula (2,045) lo contiene el intervalo, no se rechaza H_0

4.5.2. El método estadístico de prueba

En los ejemplos utilizados en las pruebas sobre la media se procedía a delimitar las regiones de aceptación y rechazo de acuerdo con la distribución del estadístico de prueba. Luego, con los valores de la muestra, se calculaba el valor del estadístico de prueba y se observaba la región donde caía este valor:

si caía en la región de aceptación, se aceptaba la hipótesis nula; de lo contrario, se rechazaba.

Esta metodología se empleó en la sección de pruebas de hipótesis para una media poblacional.

4.5.3. El método del valor de la *p*

Otro criterio para determinar si se acepta o no una hipótesis es a través del valor de la *p*, conocido como p-value. Este valor es la probabilidad de que el estadístico de prueba sea el que arroje la muestra o un valor mayor.

Como regla práctica, si p-value es mayor a la significancia de la prueba, se acepta la hipótesis nula; en caso contrario, se rechaza.

Para ejemplificar este método, se plantea nuevamente el ejemplo 3 de la sección anterior, donde el estadístico de prueba resultó $-\mathbf{0.7563}$. Como la distribución del estadístico de prueba es una *t* con 19 grados de libertad, la probabilidad de observar el valor del estadístico de prueba es el siguiente:

DISTR.T(–7563,19,2) = 0.4587

Como el p-value es mayor a 0.05, no se rechaza la hipótesis nula.

4.6. Prueba de hipótesis sobre una proporción poblacional

La segunda prueba que se abordará en esta unidad es la relacionada con la proporción poblacional (P). Para estimar este parámetro, se emplea la proporción muestral ($\boldsymbol{p}$). En la segunda unidad, se estudió que la distribución muestral de la proporción se acerca a una normal; en este caso, el estadístico de prueba a utilizar es el siguiente:

$$\boldsymbol{EP} = \frac{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}}{\boldsymbol{\sigma_{\hat{\theta}}}} = \frac{\boldsymbol{p} - \boldsymbol{P}}{\boldsymbol{\sigma_p}}$$

Donde:

$$\boldsymbol{\sigma_p} = \sqrt{\frac{\boldsymbol{P(1-P)}}{\boldsymbol{n}}}$$

Esto es cuando se conoce la proporción poblacional. En caso contrario, en vez de dividir entre *n*, se hace entre n – 1.

A continuación, se exponen ejemplos sobre la realización de pruebas con la proporción.

Ejemplo 1

Históricamente, las entregas a domicilio de productos adquiridos en una cadena de tiendas departamentales en el tiempo establecido en sus políticas, es de 85%. A fin de solicitar un bono de desempeño, el gerente de logística selecciona una muestra aleatoria de 50 casos para demostrar que la proporción de entregas hechas en tiempo se ha incrementado. En la muestra, se realizaron en tiempo 44 entregas. Con una significancia de 0.05, se pide confirmar si se apoya lo dicho por el gerente de logística.

Solución

Se realiza lo mismo de los ejemplos anteriores.

1. Establecer los datos

- Parámetro solicitado P:

P = 85%, 0.85 entregas

$$Q = 1 - P$$
$$Q = 1 - 0.85$$
$$Q = 0.15$$

n = 50

α = 0.05

- Calcular la proporción muestral (p), donde hay 44 entregas realizadas en tiempo:

$$p = \frac{44}{50} = 0.88$$

2. Definir las hipótesis

- En este ejemplo, la hipótesis nula se establece con el valor conocido de la población (P = 0.85). La hipótesis alternativa se encuentra en este segmento del enunciado del problema: "la proporción de entregas hechas en tiempo se ha incrementado".
- La prueba queda planteada así:

$H_0 : P = 0.85$

$H_1 : P > 0.85$

Es una prueba de un extremo (derecho).

3. Se determina la zona de rechazo calculando el punto crítico. Es necesario calcular el valor de la probabilidad que se sustituirá en la fórmula de Excel. Este valor se obtiene utilizando el valor de α:

- Probabilidad = 1 – 0.05 = 0.95
- PC = DISTR.NORM.ESTAND.INV(0.95)
- PC = 1.644 = 1.64

4. Se realizan los cálculos del estadístico de prueba con la fórmula correspondiente:

- $$EP = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$EP = \frac{0.88 - 0.85}{\sqrt{\frac{0.85 \cdot 0.15}{50}}}$$
$$EP = \frac{0.03}{\sqrt{\frac{0.1275}{50}}}$$
$$EP = \frac{0.03}{\sqrt{0.00255}}$$
$$EP = \frac{0.03}{0.05049}$$
$$EP = 0.594$$

5. Se dibuja la gráfica para determinar la zona donde se encuentra el valor del estadístico de prueba:

- El EP se halla en la región de no rechazo, así que H_0 no se rechaza con un nivel de significancia de 0.05. Es decir, no hay elementos para apoyar al gerente cuando afirma que la proporción de entregas hechas en tiempo ha incrementado.

Ejemplo 2

Las cifras oficiales reportadas por un país ante un organismo internacional señalan que el 72% de la población mayor a 18 años en posibilidad de generar un ingreso se encuentra en informalidad laboral. Un investigador considera que estas cifras no coinciden con la realidad nacional y levanta una encuesta entre 300 personas elegidas al azar, mayores de 18 años y en posibilidad de generar ingresos: obtiene que 262 se encuentran en informalidad. ¿Hay la suficiente evidencia con un nivel de significancia del 10% para apoyar las cifras oficiales?

Solución:

Se procede como en los ejemplos anteriores.

1. Establecer los datos

- Parámetro solicitado P: P = 72%, 0.72
- $Q = 1 - P$
- $Q = 1 - 0.72$
- $Q = 0.28$
- n = 300
- α = 0.10
- Probabilidad = 1 – 0.10 = 0.9
- $p = \frac{262}{300} = 0.87$

2. Definir las hipótesis

- En este ejemplo, la hipótesis nula se establece con el valor conocido de la población (P = 0.72). La hipótesis alternativa se encuentra en este segmento del enunciado del problema: "estas cifras no coinciden con la realidad nacional".

La prueba queda definida así:

$H_0 : P = 0.72$

$H_1 : P \neq 0.72$

Es una prueba de dos extremos.

3. Se determina la zona de rechazo calculando el punto crítico. Es necesario calcular el valor de la probabilidad que se sustituirá en la fórmula de Excel, el cual parte del valor de α.

- α = 0.10
- Como es de dos colas la significancia, se dividen entre dos: $\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$

La probabilidad será 1 – 0.05 = 0.95

Entonces, el punto crítico es el siguiente:

PC = DISTR.NORM.ESTAND.INV(0.95)

PC = 1.644 = 1.64

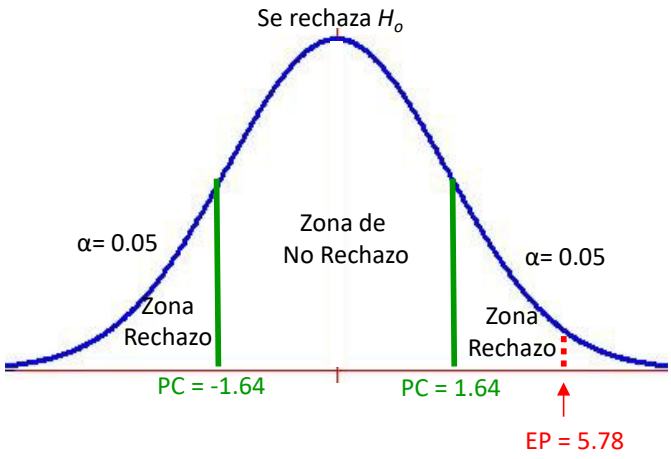
Por tratarse de una prueba de dos colas y por la simetría de la distribución normal, los puntos críticos son $\pm$ 1.64

4. Se realizan los cálculos del estadístico de prueba utilizando la fórmula correspondiente:

- $EP = \frac{p-P}{\sqrt{\frac{PQ}{n}}}$
- $EP = \frac{0.87-0.72}{\sqrt{\frac{0.72 \cdot 0.28}{300}}}$
- $EP = \frac{0.15}{\sqrt{\frac{0.2016}{300}}}$
- $EP = \frac{0.15}{\sqrt{0.000672}}$
- $EP = \frac{0.15}{0.02592}$
- $EP = 5.78$

5. Se dibuja la gráfica para determinar la zona donde se encuentra el valor del estadístico de prueba:

El EP se halla en la región de rechazo, así que H_0 se rechaza a un nivel de confianza de 90%. Es decir, hay elementos para apoyar al investigador cuando afirma que las cifras oficiales no coinciden con la realidad nacional.



4.7. Pruebas de hipótesis sobre la diferencia entre dos medias

Es común enfrentarse a situaciones donde se desea comparar los parámetros de dos poblaciones. Por ejemplo, el director de mercadotecnia de una organización podría estar interesado en conocer el nivel de ingreso de cierto segmento de interés en el Distrito Federal y en Tijuana; o el director de la FCA se interesaría en conocer el nivel de matemáticas de los alumnos de primer ingreso provenientes del concurso de selección en comparación con los de pase reglamentario.

En este apartado se muestra cómo realizar estos comparativos. La prueba que se abordará es la diferencia entre dos medias de poblaciones diferentes ($\mu_1 - \mu_2$). En esencia, la prueba establece que no existe diferencia importante entre las medias de estas poblaciones ($\mu_1 = \mu_2$). Para estimar esta diferencia, se recurre a la diferencia de los promedios muestrales ($\bar{x}_1 - \bar{x}_2$).

El estadístico de prueba a utilizar es el siguiente:

$$EP = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\mu_1 - \mu_2}}$$

Donde:

$$\sigma_{\mu_1 - \mu_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Esto es cuando se conoce la varianza poblacional; en caso contrario, se sustituye por *s*.

Cuando se conocen las varianzas poblacionales, el estadístico de prueba tiene una distribución normal estandarizada. De lo contrario, la distribución del estadístico de prueba es una *t* con grados de libertad definidos con la siguiente fórmula:

$$gl = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1 - 1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2 - 1}(\frac{s_2^2}{n_2})^2}$$

A continuación, se muestran ejemplos sobre la realización de este tipo de pruebas.

Ejemplo 1

Un inversionista planea desarrollar un gimnasio destinado a mujeres. Para definir a qué segmento enfocarse, opta por encuestar a dos grupos de 50 mujeres elegidas aleatoriamente que realizan ejercicio, para conocer el tiempo destinado a ello. El primer grupo lo integran mujeres de 30 años o menos; y el segundo, mujeres mayores de 30 años. Los resultados de las encuestas arrojaron que las mujeres de 30 años o menos destinan en ejercitarse un promedio de 3.1 horas al día con una varianza muestral de 1.43 horas; las mujeres mayores de 30 años, destinan en promedio 2.78 horas con una varianza muestral de 1.34 horas. Con una significancia del 1%, se pide determinar si existe diferencia entre los grupos al tiempo promedio destinado a ejercitarse.

Solución:

Se procede como en los ejemplos anteriores.

1. Establecer los datos

- Parámetro solicitado $\mu_1 - \mu_2$:
- $\bar{x}_1 = 3.1$
- $\bar{x}_2 = 2.78$
- $s_1^2 = 1.43$
- $s_2^2 = 1.34$
- $n_1 = 50$
- $n_2 = 50$
- $\alpha = 0.01$
- Como no se conocen las varianzas poblacionales, es necesario calcular los grados de libertad:

$$gl = \frac{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{s_1^2}{n_1})^2+\frac{1}{n_2-1}(\frac{s_2^2}{n_2})^2}$$

$$gl = \frac{(\frac{1.43}{50}+\frac{1.34}{50})^2}{\frac{1}{50-1}(\frac{1.43}{50})^2+\frac{1}{50-1}(\frac{1.34}{50})^2}$$

$$gl = \frac{(0.0286+0.0268)^2}{\frac{1}{49}(0.0286)^2+\frac{1}{49}(0.0268)^2}$$

$$gl = \frac{(0.0554)^2}{\frac{1}{49}(0.0008)+\frac{1}{49}(0.0007)}$$

$$gl = \frac{0.00309}{\frac{1}{49}(0.0008+0.0007)}$$

$$gl = \frac{0.00309}{\frac{1}{49}(0.0008+0.0007)}$$

$$gl = \frac{0.00309}{\frac{1}{49}(0.0008+0.0007)}$$

$$gl = \frac{0.00309}{\frac{1}{49}(0.0015)}$$

$$gl = \frac{0.00309}{\frac{1}{49}(0.0015)}$$

$$gl = \frac{0.00309}{0.0000313}$$

$$gl = 97.9$$

$$gl = 98$$

2. Definir las hipótesis

- En este ejemplo, la hipótesis nula es que no existe diferencia entre los grupos ($\mu_1 = \mu_2$, equivalente a $\mu_1-\mu_2 = 0$). La hipótesis alternativa se encuentra en este segmento del enunciado del problema: "si existe diferencia entre los grupos al tiempo promedio destinado a ejercitarse".
- La prueba queda planteada así:
- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$
- La prueba es de dos extremos.

3. Se determina la zona de rechazo calculando los puntos críticos de un estadístico de prueba con una distribución t de Student con 98 grados de libertad. Para calcularlos en Excel, es suficiente contar con la significancia de la prueba (0.01):

- PC= DISTRIR.T.INV(0.01,98)
- PC = 2.63
- Por la simetría de la distribución, los puntos críticos son ± 2.63.

4. Se realizan los cálculos del estadístico de prueba con la fórmula correspondiente:

- $EP = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
- $EP = \frac{3.1-2.78}{\sqrt{\frac{1.43}{50}+\frac{1.34}{50}}}$
- $EP = \frac{0.32}{\sqrt{0.0286+0.0268}}$
- $EP = \frac{0.32}{\sqrt{0.0408+0.0414}}$
- $EP = \frac{0.32}{\sqrt{0.0554}}$
- $EP = \frac{0.32}{0.2354}$
- $EP = 1.36$

5. Se dibuja la gráfica para determinar la zona donde se encuentra el valor del estadístico de prueba:

- El EP se localiza en la región de no rechazo, así que H_0 no se rechaza con una significancia de 0.01. Es decir, no hay elementos para apoyar que existe diferencia entre los grupos al tiempo promedio destinado a ejercitarse.

Ejemplo 2

De acuerdo con una encuesta de origen-destino, el tiempo de traslado al centro de la ciudad tiene una desviación de 0.24 horas para los habitantes del norte y de 0.19 horas para los del sur. En una organización ubicada en el centro de la ciudad, el director de recursos humanos desea proponer una política de contratación basada en el tiempo que toma el aspirante para trasladarse de su domicilio a la organización, para ello toma una muestra aleatoria de 30 empleados que viven al norte y otra de 35 que viven al sur: obtiene que el tiempo promedio de traslado es de 1.56 horas para los empleados que viven en el norte y de 2.08 horas para los que habitan en el sur. Entonces, con una significancia de 0.05, ¿existe evidencia estadística que apoye la promoción de una política de contratación basada en el tiempo de traslado del aspirante?

Solución:

Se procede como en los ejemplos anteriores.

1. Establecer los datos

Parámetro solicitado

$\mu_1 - \mu_2$:

$\bar{x}_1 = 1.56$

$\bar{x}_2 = 2.08$

$s_1 = 0.24$

$s_2 = 0.19$

$n_1 = 30$

$n_2 = 35$

$\alpha = 0.05$

2. Definir las hipótesis

La hipótesis nula es que no existe diferencia entre los grupos ($\mu_1 = \mu_2$, lo cual es equivalente a $\mu_1 - \mu_2 = 0$). La hipótesis alternativa se encuentra implícita en este segmento del enunciado del problema: "¿Existe evidencia estadística que apoye la promoción de una política de contratación basada en el tiempo de traslado del aspirante?". Si hay una diferencia importante, entonces se tiene el sustento para promover la política. La prueba queda planteada de la siguiente manera:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Es una prueba de dos extremos.

3. Se determina la zona de rechazo calculando los puntos críticos. Es necesario calcular el valor de la probabilidad que se sustituirá en la fórmula de Excel utilizando el valor de α. Como la prueba es de dos extremos, la significancia se divide entre 2:

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

La probabilidad es 1 – 0.025 = 0.975

PC = DISTR.NORM.ESTAND.INV(0.975)

PC = 1.96

4. Se realizan los cálculos del estadístico de prueba con la fórmula correspondiente:

$$EP = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$EP = \frac{1.56 - 2.08}{\sqrt{\frac{0.24}{30} + \frac{0.19}{35}}}$$

$$EP = \frac{-0.52}{\sqrt{0.008 + 0.0054}}$$

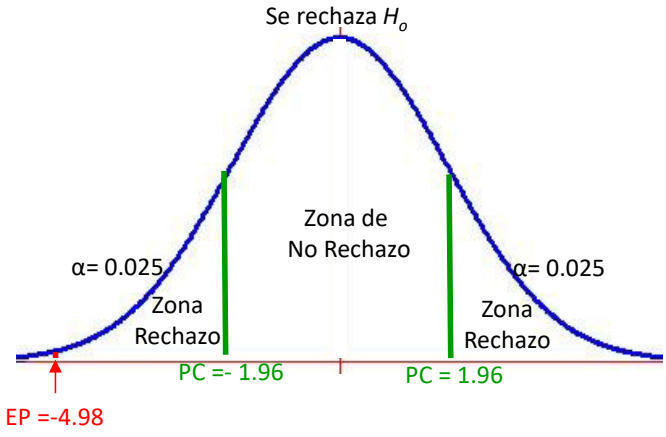
$$EP = \frac{-0.52}{\sqrt{0.0134}}$$

$$EP = \frac{-0.52}{0.1158}$$

$$EP = -4.48$$

Se dibuja la gráfica para determinar la zona donde se encuentra el valor del estadístico de prueba:

El EP se sitúa en la región de rechazo, así que H_0 se rechaza a un nivel de significancia de .05. Es decir, hay elementos para apoyar que existe evidencia estadística que apoye la promoción de una política de contratación basada en el tiempo de traslado del aspirante.



4.8. Pruebas de hipótesis sobre la diferencia entre dos proporciones

En este apartado, se muestra la prueba que realiza la diferencia entre dos proporciones poblacionales ($P_1 – P_2$). En esencia, la prueba establece que no existe diferencia importante entre las proporciones de estas poblaciones ($P_1 = P_2$). Para estimar esta diferencia, se emplea la diferencia de las proporciones muestrales $(\boldsymbol{p_1 - p_2})$.

El estadístico de prueba a utilizar es el siguiente:

$$\boldsymbol{EP = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}}}$$

- Donde:

$$\boldsymbol{\sigma_{p_1 - p_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}}$$

El estadístico de prueba tiene una distribución normal estandarizada.

A continuación, se muestran ejemplos sobre la realización de esta prueba.

Ejemplo 1

El SUAYED de la FCA de la UNAM ofrece dos modalidades para cursar las carreras impartidas en esa Facultad: universidad abierta y educación a distancia. En ambas modalidades, se cuida la calidad de sus profesores para garantizar la excelencia académica. Se sospecha que en la materia de Estadística II existe diferencia en la reprobación, por lo que se seleccionaron al azar dos muestras: una de 80 alumnos de educación a distancia y otra de 60 de universidad abierta, para comprobar si hay diferencia en las modalidades. Los resultados de las muestras se presentan en la siguiente tabla.

Alumnos de la muestra que aprueban y reprueban Estadística II

Modalidad	Aprueba	No aprueba	Total
Educación a distancia	55	25	80
Universidad abierta	32	28	60
Total	87	53	140

Con una significancia del 5%, ¿se apoya que no existe diferencia entre modalidades en la materia de Estadística II?

Solución:

Prueba de hipótesis:

$H_0 = p_1 = p_2$

$H_1 = p_1 \neq p_2$

Modalidad	Proporción de alumnos que	
	aprueba p_i	no aprueba q_i
Educación a distancia	$\frac{55}{80} = 0.69$	$\frac{25}{80} = 0.31$
Universidad abierta	$\frac{32}{60} = 0.53$	$\frac{28}{60} = 0.47$
Total	$\frac{87}{140} = 0.62$	$\frac{53}{140} = 0.38$

Estadístico de prueba:

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

$$Z = \frac{0.69 - 0.53}{\sqrt{\frac{0.69(1-0.69)}{80} + \frac{0.53(1-0.53)}{60}}}$$

$$Z = \frac{0.15}{\sqrt{\frac{0.69(0.31)}{80} + \frac{0.53(0.47)}{60}}}$$

$$Z = \frac{0.15}{\sqrt{\frac{0.2139}{80} + \frac{0.2491}{60}}}$$

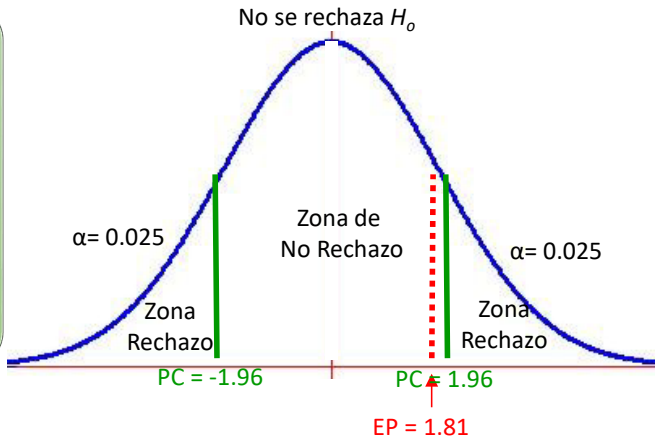
$$Z = \frac{0.15}{\sqrt{0.0027 + 0.0041}}$$

$$Z = \frac{0.15}{\sqrt{0.0068}}$$

$$Z = \frac{0.15}{0.0826}$$

$$Z = 1.81$$

Punto crítico:
DISTR.NORM.ESTAND.INV(1-0.05/2) = 1.959
El EP se halla en la región de no rechazo, así que H_0 no se rechaza con una significancia del 5%. Entonces, no hay evidencia para rechazar la hipótesis nula de que no existe diferencia entre las modalidades de acuerdo con la proporción de estudiantes aprobados en la materia de Estadística II.



Ejemplo 2

Una compañía dedicada a la venta de tiempos compartidos quiere lanzar una campaña de publicidad para captar más clientes y desea saber si la proporción de matrimonios que realiza la compra del tiempo compartido es igual a la proporción de parejas en unión libre. Se toma una muestra de 100 matrimonios y otra de 100 parejas en unión libre. La información del resultado de la venta se muestra en la siguiente tabla.

Resultado de la venta de tiempos compartidos en la muestra de matrimonios y parejas en unión libre

Muestra	Compra	No compra	Total
Matrimonios	63	37	100
Parejas en unión libre	47	53	100
Total	**110**	**90**	**200**

Con una significancia del 0.1, ¿se apoya que hay diferencia en el resultado de la venta de acuerdo con la situación marital?

Solución:

Muestra	Resultado de la venta	
	Compra	No compra
Matrimonios	$\frac{63}{100} = 0.63$	$\frac{37}{100} = 0.37$
Parejas en unión libre	$\frac{47}{100} = 0.47$	$\frac{53}{100} = 0.53$
Total	$\frac{110}{200} = 0.55$	$\frac{90}{200} = 0.45$

Prueba de hipótesis:

$H_0 = p_1 = p_2$

$H_a = p_1 \neq p_2$

Estadístico de prueba:

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

$$Z = \frac{0.63 - 0.47}{\sqrt{\frac{0.63(1-0.63)}{100} + \frac{0.47(1-0.47)}{100}}}$$

$$Z = \frac{0.16}{\sqrt{\frac{0.63(0.37)}{100} + \frac{0.47(0.53)}{100}}}$$

$$Z = \frac{0.16}{\sqrt{\frac{0.2331}{100} + \frac{0.2491}{100}}}$$

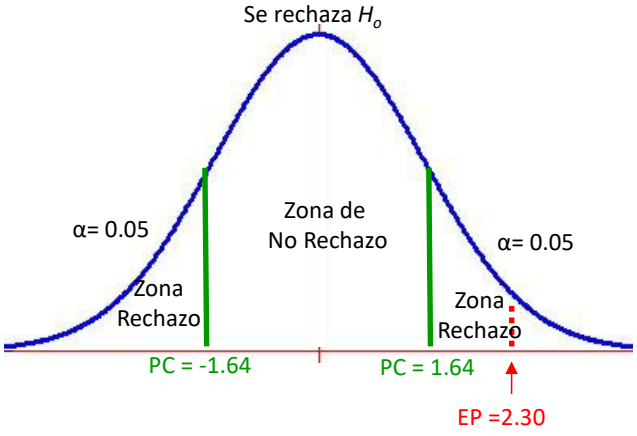
$$Z = \frac{0.16}{\sqrt{0.002331 + 0.002491}}$$

$$Z = \frac{0.16}{\sqrt{0.004822}}$$

$$Z = \frac{0.16}{0.0694}$$

$$Z = 2.30$$

Punto crítico:
DISTR.NORM.ESTAND.INV(1-0.10/2) = 1.644
El EP se sitúa en la región de rechazo, por lo que H_0 se rechaza con una significancia del 10%. Luego, hay evidencia para no apoyar que no existe diferencia en la proporción de matrimonios y parejas en unión libre que realizan la compra del tiempo compartido.



4.9. Prueba para la diferencia entre dos varianzas

La última prueba que se abordará en esta unidad se utilizará para comparar dos varianzas. A diferencia de las pruebas para comparar dos medias o dos proporciones, la distribución del estadístico de prueba es sesgada a la derecha, la cual es la distribución F (mencionada al final de la segunda unidad). Para emplear esta distribución, se parte del supuesto de que las muestras provienen de poblaciones con distribución normal y que las dispersiones son las mismas. Si no se cumple el supuesto, la prueba caerá en la región de rechazo.

El estadístico de prueba es el siguiente:

$$F = \frac{S_1^2}{S_2^2}$$

- Donde:

S_1^2 = la varianza muestral de la población 1

S_2^2 = la varianza muestral de la población 2

Al igual que las distribuciones t de Student y χ^2, la distribución F depende del número de elementos de las muestras extraídas de cada población, así que esta distribución tiene como parámetros los grados de libertad: el tamaño de la muestra de la primera población menos uno y el tamaño de la muestra de la segunda población menos uno. Se acostumbra colocar la varianza más grande de las muestras en el numerador. A continuación, se desglosa un ejemplo.

Ejemplo

Una escuela tiene como política aplicar exámenes departamentales de cada materia para comprobar que los conocimientos de los alumnos es el mismo independientemente del grupo al que pertenezcan, con una significancia del 0.1. El coordinador del área de matemáticas quiere saber si hay variación entre las calificaciones obtenidas de los dos grupos de Estadística Inferencial, y aplica el examen parcial de la material a una muestra de diez alumnos de cada grupo. Las calificaciones obtenidas en cada muestra son las siguientes:

Calificaciones

Grupo 1	Grupo 2
4	7
8	9
9	6
3	6
5	8
5	7
8	8
5	7
8	10
9	7

Respuesta:

1. Parámetros: σ_1^2, σ_2^2
2. Calcular las varianzas muestrales y los grados de libertad.
Utilizando las fórmulas de Excel se obtiene:

$$s_1^2 = 4.93$$
$$s_2^2 = 1.61$$

Como $s_1^2 > s_2^2$, se ubicará el valor de la primera muestra en el numerador del estadístico de prueba.
Los grados de libertad asociados al numerador son 10 – 1 = 9, el cual es el mismo para los del denominador, ya que ambos tamaños de muestra constan de 10 elementos.

3. Definir las hipótesis

- En este ejemplo, la hipótesis nula es que no existe diferencia entre los grupos ($\sigma_1^2 = \sigma_2^2$). La hipótesis alternativa se encuentra en este segmento del enunciado del problema: "hay variación entre las calificaciones obtenidas de los dos grupos de Estadística Inferencial".

La prueba queda planteada así:
$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$
Esta prueba es de dos extremos.

4. Como se trata de una prueba de dos colas y además está sesgada, se deben determinar dos valores críticos a través de la siguiente fórmula de Excel, que requiere del valor de α y de los grados de libertad:

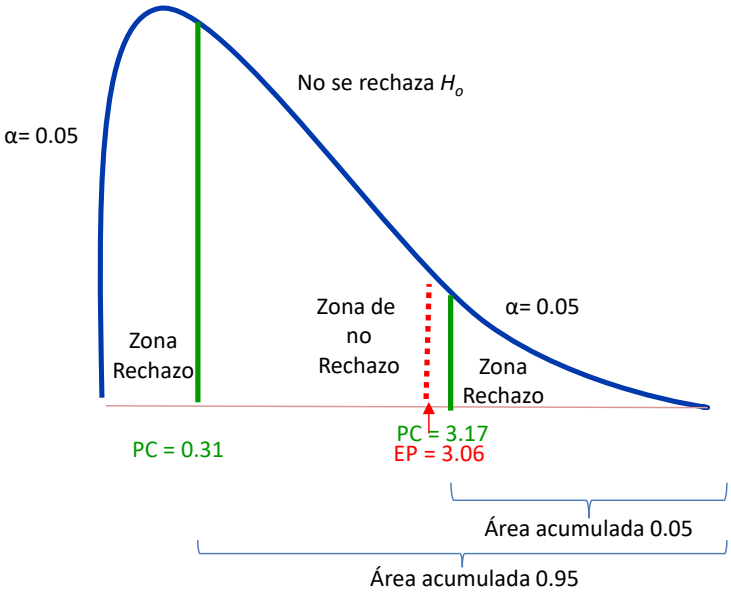
$$\frac{\alpha}{2} = \frac{0.1}{2} = 0.05$$

PC_1 = DISTR.F.INV(0.05,9,9)=3.17
PC_2 = DISTR.F.INV(0.95,9,9)=0.31

5. Se realizan los cálculos del estadístico de prueba utilizando la fórmula correspondiente:

$$F = \frac{S_1^2}{S_2^2}$$
$$F = \frac{4.93}{1.61}$$
$$F = 3.06$$

6. Se dibuja la gráfica para determinar las zonas de rechazo y comparar el estadístico de prueba: El EP se sitúa en la región de no rechazo, así que H_0 no se rechaza con una significancia del 10%. Luego, no hay evidencia para rechazar que no hay variación entre las calificaciones obtenidas de los dos grupos de Estadística Inferencial.



RESUMEN

En esta unidad, se trató el tema de prueba de hipótesis, consistente en un contraste de dos supuestos sobre el valor de un parámetro, el cual se prueba con los resultados de una muestra. Se analizó cómo plantear hipótesis, se mencionaron los tipos de errores que pueden cometerse, los tipos de pruebas que pueden realizarse y la forma de delimitar las regiones de aceptación y rechazo. Se explicó también cómo efectuar pruebas de hipótesis con los métodos de intervalo, estadístico de prueba y p-value. Además, con el apoyo de Excel, se trabajaron ejercicios para realizar pruebas con la media y la proporción, así como con la diferencia de medias, proporciones y varianzas.

BIBLIOGRAFÍA

SUGERIDA

Autor	Capítulo	Páginas
Anderson, S.	9	348-405
Levin, R.	8	319-358
	9	359-402
Lind, D.	10	333-370
	11	371-409

Licenciatura: **Administración**

UNIDAD 5

Pruebas de hipótesis con la distribución ji cuadrada

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 **2016** actualizado

OBJETIVO PARTICULAR

Al terminar la unidad, el alumno relacionará los conceptos de prueba de hipótesis con la distribución ji cuadrada.

TEMARIO DETALLADO

(10 horas)

5. Pruebas de hipótesis con la distribución ji cuadrada

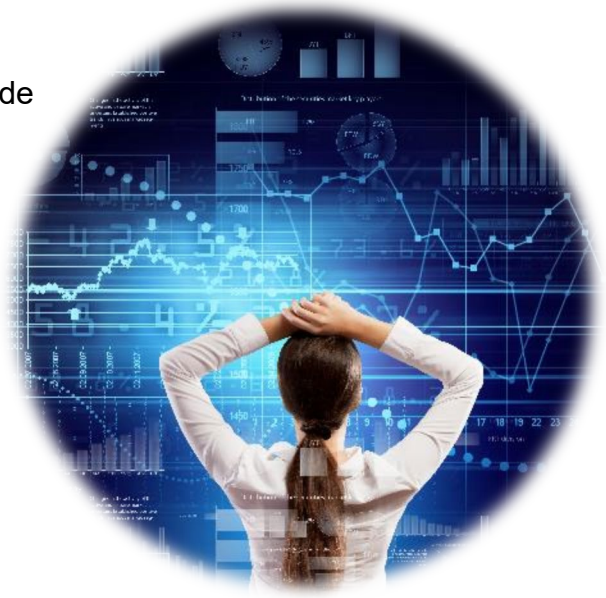
- 5.1. La distribución ji cuadrada, χ^2
- 5.2. Pruebas de hipótesis para la varianza de una población
- 5.3. Prueba para la diferencia entre *n* proporciones
- 5.4. Pruebas de bondad de ajuste a distribuciones teóricas
  - 5.4.1. Ajuste a una distribución Normal
  - **5**.4.2. Ajuste a una distribución Poisson
  - 5.4.3. Ajuste a una distribución Binomial
- 5.5. Pruebas sobre la independencia entre dos variables
- 5.6. Pruebas de homogeneidad

INTRODUCCIÓN

En la unidad anterior, se dieron las bases para realizar pruebas de hipótesis para contrastar valores de parámetros de una población, como la media y una proporción. Posteriormente, se contrastaron medias, proporciones y varianzas de poblaciones independientes utilizando estadísticos de prueba con distribuciones normal, t de Student y F. Ahora, en esta unidad, se empleará otra distribución muestral, la ji cuadrada (χ^2), útil no solamente para realizar pruebas relacionadas con una varianza poblacional, sino también para validar si una muestra se ajusta a una distribución teórica, si hay un cambio en una distribución, si dos variables son independientes o si dos muestras proceden de la misma población.

Primero, se expondrá la distribución χ^2; después, se mostrará su uso para contrastar hipótesis relacionadas con la varianza poblacional, diferencia de proporciones, bondad de ajuste, independencia y homogeneidad.

Para el profesional egresado de la Facultad de Contaduría y Administración, el conocimiento y manejo de esta distribución le dará una herramienta adicional para una mejor toma de decisiones.



5.1. La distribución ji cuadrada, χ^2

En la última sección de la tercera unidad, se utilizó la distribución χ^2 (ji cuadrada) para estimar un intervalo para una varianza poblacional. Teóricamente, esta distribución es un caso de otra distribución conocida como *gamma*; el parámetro que determina su distribución son los grados de libertad, es decir, el número de observaciones que pueden variar libremente. Las características de esta distribución son las siguientes:

- La distribución se encuentra definida para valores positivos.
- La forma de una distribución χ^2 depende de los grados de libertad (gl), por lo que hay un número infinito de distribuciones.
- El área bajo la curva es uno.
- La distribución es sesgada a la derecha.

En distribuciones muestrales, se emplea el estadístico

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

- Donde:

n = tamaño de muestra
σ^2 = varianza poblacional
s^2 = varianza muestral

El estadístico tiene una distribución χ^2 con n – 1 grados de libertad.

Este resultado es válido si la muestra proviene de una población con distribución normal.

5.2. Pruebas de hipótesis para la varianza de una población

En la unidad anterior, se realizaron pruebas de hipótesis relacionadas con una media, una proporción, diferencia de medias y diferencia de proporciones, y se finalizó con pruebas entre dos varianzas. En este capítulo, se expone cómo efectuar una prueba para la varianza de una población.

Como se ha mencionado en las unidades pasadas, en ocasiones se requiere hacer inferencias sobre la varianza poblacional. Así como en la unidad anterior, en este caso se plantea una hipótesis nula y otra alternativa que involucra a la varianza, pero el estadístico de prueba es:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Y la distribución asociada es una χ^2 con n – 1 grados de libertad.

A continuación, se analizan dos ejemplos.

Ejemplo 1.

Un *call center* tiene como criterio de calidad que la duración de sus llamadas tenga una desviación estándar de 1.5 respecto al promedio de cinco minutos El gerente del *call center* sospecha que la desviación es mayor, para confirmarlo elige una muestra de 50 llamadas y obtiene una desviación de 1.37 minutos. ¿Se puede afirmar con un nivel de confianza del 95% que la sospecha del gerente es correcta?

Parámetro solicitado:	Datos:
σ	$\sigma = 1.5$ $n = 50$ $s = 1.37$ Nivel de confianza: 95% = 0.95 Significancia: $\alpha = 1 - 0.95 = 0.05$ Grados de libertad: n – 1 = 50 – 1 = 49

Hipótesis:

$$H_0 = \sigma^2 = (1.5)^2$$
$$H_1 = \sigma^2 > (1.5)^2$$

Cálculo del estadístico de prueba:

$$\chi^2 = \frac{(50-1)\cdot(1.37)^2}{(1.5)^2}$$
$$\chi^2 = \frac{(49)\cdot(1.37)^2}{(1.5)^2}$$
$$\chi^2 = \frac{(49)\cdot 1.8967}{2.25}$$

Cálculo del punto crítico
Con el empleo de la función de Ms-Excel:

PRUEBA.CHI.INV(probabilidad,grados_de_libertad)

Se obtiene:

PRUEBA.CHI.INV(0.05,49) = 66.3386

En la figura 1, se ilustra la región donde cae el estadístico de prueba:

Figura 1. Resultado de la prueba de hipótesis *Ho*: $\sigma^2 = 1.5$ contra *Ho*: $\sigma^2 > 1.5$

La figura anterior muestra la distribución del estadístico de prueba asumiendo que la hipótesis nula es cierta. Como la prueba es unilateral, en este caso la región de rechazo se encuentra en el extremo derecho de la curva, a partir del punto crítico (66.33), ello significa que, si la prueba tiene un valor mayor a este punto, la hipótesis nula se rechaza. En la figura, se observa que el resultado de la prueba (40.87) es menor al punto crítico, por tanto, no se rechaza la hipótesis nula.
En conclusión, no existe evidencia estadística para rechazar la hipótesis nula, es decir, no se apoya la sospecha del gerente que la desviación estándar sea mayor a 1.5 minutos.

Ejemplo 2.

Una empresa realiza periódicamente una encuesta de clima laboral entre los empleados. Recientemente, varios departamentos solicitan que esta encuesta ya no se realice con la misma periodicidad, pues distrae las labores de los subordinados. En defensa de la encuesta, el director de recursos humanos sostiene que una variabilidad de 7 minutos no afecta el desempeño. Para comprobar que la variabilidad es de 7, elige una muestra de 20 empleados y obtiene un resultado de 6.7 minutos. ¿Se puede afirmar, con un nivel de confianza del 90%, que el director está en lo correcto?

Parámetro solicitado:	Datos:
$\boldsymbol{\sigma}$	$\boldsymbol{\sigma = 7}$ $\boldsymbol{n = 20}$ $\boldsymbol{s = 6.7}$ Nivel de confianza: 90% = 0.90 Significancia: $\boldsymbol{\alpha = 1 - 0.9 = 0.1}$ $\boldsymbol{\alpha = \frac{0.1}{2} = 0.05}$ Grados de libertad: n – 1 = 20 – 1 = 19

Hipótesis:

$$\boldsymbol{H_0 = \sigma^2 = (7)^2}$$
$$\boldsymbol{H_1 = \sigma^2 \neq (7)^2}$$

Cálculo del estadístico de prueba:

$$\boldsymbol{\chi^2 = \frac{(20-1)\cdot(6.7)^2}{(7)^2}}$$
$$\boldsymbol{\chi^2 = \frac{(19)\cdot 44.89}{49}}$$
$$\boldsymbol{\chi^2 = \frac{852.91}{49}}$$
$$\boldsymbol{\chi^2 = 17.4}$$

Cálculo del punto crítico
Con Excel, se obtienen los puntos críticos. Valor crítico superior:

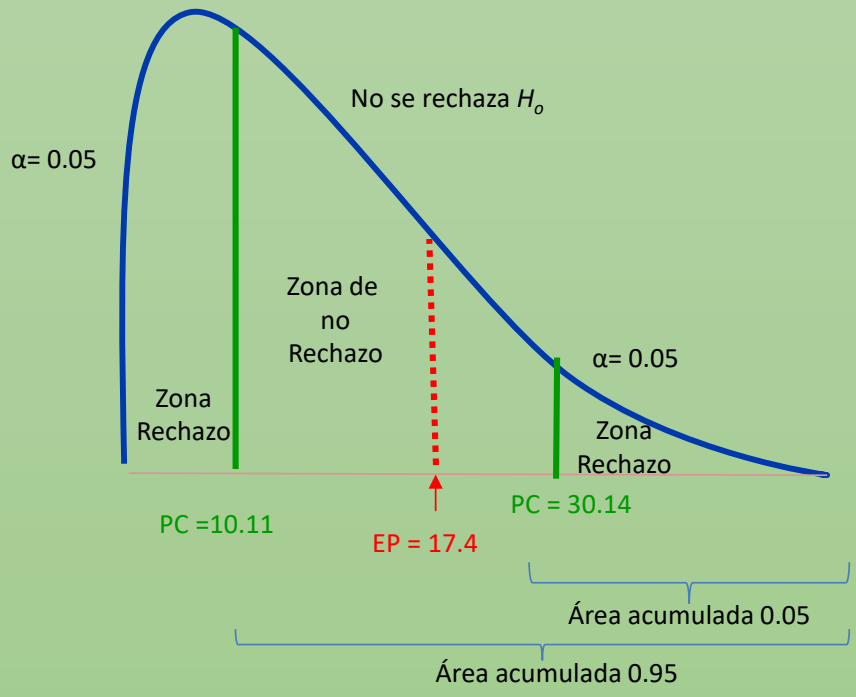
PRUEBA.CHI.INV (0.05,19) = 30.14

Valor crítico inferior:

PRUEBA.CHI.INV (0.95,19) = 10.11

En la figura 2, se ilustra la región donde cae el estadístico de prueba:

Figura 2. Resultado de la prueba de hipótesis *Ho*: = 7 contra *Ho*: ≠ 7



La figura anterior muestra la distribución del estadístico de prueba asumiendo que la hipótesis nula es cierta. Como la prueba es bilateral, la región de rechazo se encuentra en ambos extremos de la curva. La región de aceptación se halla entre los puntos críticos (10.11 y 30.14), esto significa que, si la prueba tiene un valor en esta región, la hipótesis nula se acepta. En la figura, se observa que el resultado de la prueba (17.4) se encuentra en la zona de aceptación, por tanto, no se rechaza la hipótesis nula. En conclusión, no existe evidencia estadística para rechazar la hipótesis nula: se apoya la defensa del director de recursos humanos.

5.3. Prueba para la diferencia entre *n* proporciones

En la sección anterior, se mostró el empleo de la distribución χ^2 para hacer un contraste de hipótesis de una varianza poblacional. A partir de esta sección, se analizará su utilidad en la comparación de datos observados contra esperados, y de esta manera apoyar o no un comportamiento teórico.

Estadístico de prueba que se empleará a partir de esta sección:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Donde:

o_i = valor observado
e_i = valor esperado
k = número de categorías

Este estadístico tendrá una distribución χ^2. Los grados de libertad varían según el contexto.

En esta sección, se aplicará el estadístico mencionado para apoyar o no que un conjunto de datos tiene una distribución multinomial.

En el curso de Estadística Descriptiva, se presentó la distribución binomial, la cual tiene como una de sus características que cada uno de los *n* ensayos independientes solamente ofrece dos resultados posibles manteniéndose constante la probabilidad de éxito. Cuando existen al menos tres resultados posibles, los cuales son mutuamente

excluyentes y cada uno con una probabilidad de ocurrencia de manera que su suma da uno, se está frente a una distribución multinomial.

Supóngase que históricamente la proporción de estudiantes de Administración que obtiene una calificación mayor a 9 en Estadística Inferencial es 0.05; entre 8 y 9, 0.15; entre 7 y 8, 0.55; y el resto, menor a 7. Se ha propuesto una estrategia de enseñanza que se espera mejore el aprovechamiento de la materia en los estudiantes de Administración. Un grupo piloto de 140 alumnos registró los siguientes resultados:

Nivel	Rango de calificación	Alumnos
A	9.1-10	15
B	8.1-9.0	35
C	7.1-8.0	50
D	Hasta 7.0	40
Total		140

¿Se podría apoyar con un nivel de confianza de 95% que la estrategia modificó el aprovechamiento de los estudiantes de Administración en Estadística Inferencial?
Obsérvese que el tratamiento de la información se ajusta al de una distribución multinomial porque hay más de dos resultados y cada alumno nada más puede estar en una categoría. Se denotará como p_A, p_B, p_C y p_D a la proporción de alumnos en cada nivel, y se aplicará una prueba de hipótesis para determinar si la nueva estrategia modifica el desempeño.

La hipótesis nula y alternativa para probar si la estrategia modifica o no el desempeño es la siguiente:

$$H_0: p_A = 0.05;\ p_B = 0.15;\ p_C = 0.55;\ \ p_D = 0.25$$
$$H_a: las\ proporciones\ poblacionales\ no\ son\ las\ de\ la\ hipótesis\ nula$$

Asumiendo como cierta la hipótesis nula, se esperaría que los 140 alumnos se distribuyeran de la siguiente manera:

Nivel	Rango de calificación	Proporción bajo H_o	Alumnos esperados
A	9.1-10	0.05	$\mathbf{140 \cdot 0.05 = 7}$
B	8.1-9.0	0.15	$\mathbf{140 \cdot 0.15 = 21}$
C	7.1-8.0	0.55	$\mathbf{140 \cdot 0.55 = 77}$
D	Hasta 7.0	0.25	$\mathbf{140 \cdot 0.25 = 35}$
Total			**140**

Se calcula el estadístico de prueba que tendrá una distribución χ^2 con k – 1 grados de libertad, en este caso, k = 4:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 = \frac{(15-7)^2}{7} + \frac{(35-21)^2}{21} + \frac{(50-77)^2}{77} + \frac{(40-35)^2}{35}$$

$$\chi^2 = 9.1 + 9.3 + 9.5 + 0.7$$

$$\chi^2 = 28.7$$

Se realiza una prueba bilateral. Con Microsoft Excel (2013), se calcula el punto crítico superior:

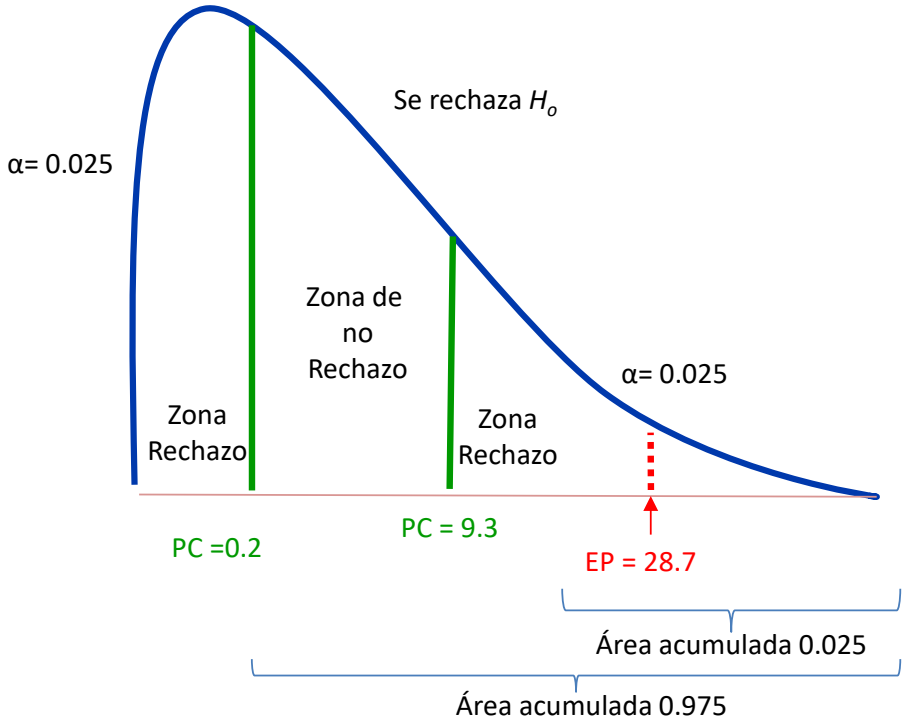
PRUEBA.CHI.INV(0.05/2,3) = 9.3

Y el inferior:

PRUEBA.CHI.INV(1-0.05/2,3) = 0.2

En la figura 1, se ilustra la región donde cae el estadístico de prueba.

Figura 1. Resultado de la prueba de hipótesis



Fuente: elaboración propia.

La figura anterior muestra la distribución del estadístico de prueba asumiendo que la hipótesis nula es cierta. Debido a que la prueba es bilateral, la región de rechazo se encuentra en ambos extremos de la curva. La región de aceptación se halla entre los puntos críticos (0.2 y 9.3), lo cual significa que, si la prueba tiene un valor en esta región, la hipótesis nula se acepta. En la figura se observa que el resultado de la prueba (28.7) se sitúa en la zona de rechazo, por tanto, se rechaza la hipótesis nula.

En conclusión, hay evidencia estadística para rechazar la hipótesis nula: la estrategia modificó el aprovechamiento de los estudiantes de Administración en Estadística Inferencial.

5.4. Pruebas de bondad de ajuste a distribuciones teóricas

Como se ha estudiado hasta este punto, tanto las técnicas de estimación como las de contraste de hipótesis se realizan con la información de una muestra. A veces, se pretende conocer si la población de la que proviene la muestra se ajusta a una distribución teórica. En esta sección, se utilizará la distribución χ^2 para probar si un conjunto de información se ajusta a una distribución Normal, Poisson o Binomial. En las tres distribuciones el proceso para realizar la prueba es similar:

Se forman categorías.

Se realizan conteos en cada categoría.

Se estima el valor esperado de elementos en cada categoría.

Se contrasta la hipótesis.

- H_0: los datos se ajustan a la distribución
- H_1: los datos no se ajustan a la distribución

Con el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Donde:

o_i = valor observado
e_i = valor esperado
k = número de categorías

Asumiendo cierta la hipótesis nula, este estadístico tendrá una distribución χ^2, con k – p – 1 grados de libertad, donde *k* es el número de categorías y *p* los parámetros de la distribución teórica.

La hipótesis nula se rechaza si el valor del estadístico de prueba resulta mayor al punto crítico de la distribución teórica.

A continuación, se muestra cómo realizar la prueba de bondad de ajuste para una distribución normal.

5.4.1. Ajuste a una distribución normal

Para explicar la prueba para el ajuste a una distribución normal, se utilizará el siguiente ejemplo.

Los resultados de una prueba realizada a 110 aspirantes a ocupar una plaza laboral se muestran a continuación.

80.0	60.0	56.7	54.2	52.5	50.8	48.3	46.7	45.0	42.5	36.7
69.2	59.2	56.7	53.3	51.7	50.0	48.3	46.7	45.0	42.5	36.7
69.2	59.2	56.7	53.3	51.7	50.0	48.3	45.8	44.2	41.7	36.7
69.2	59.2	56.7	53.3	51.7	49.2	47.5	45.8	44.2	41.7	34.2
69.2	58.3	56.7	53.3	51.7	49.2	47.5	45.8	44.2	40.8	34.2
68.3	58.3	55.8	53.3	50.8	49.2	47.5	45.0	44.2	40.8	33.3
64.2	57.5	55.8	53.3	50.8	49.2	47.5	45.0	43.3	40.0	32.5
63.3	57.5	55.8	52.5	50.8	49.2	46.7	45.0	43.3	38.3	32.5
61.7	56.7	55.0	52.5	50.8	48.3	46.7	45.0	43.3	37.5	29.2
60.8	56.7	54.2	52.5	50.8	48.3	46.7	45.0	42.5	36.7	37.2

A fin de precisar los puntajes que deben tener los candidatos para pasar a la siguiente etapa, se quiere probar primeramente que los datos provienen de una distribución normal con un nivel de confianza de 95%.

Como la distribución normal es continua, para probar el ajuste a esta distribución, se categorizará la información en deciles.

En primer lugar, se estimarán los parámetros de la distribución (media y desviación estándar) con la información de la muestra.

El estimador de la media (μ) es el promedio muestral; y el de la desviación estándar (σ), la desviación muestral.

Así:

$$\hat{\mu} = \frac{80.0 + 69.2 + \cdots + 29.2 + 37.2}{109} = 49.7$$

- Y:

$$\hat{\sigma} = \sqrt{\frac{(80.0 - 49.7)^2 + (69.2 - 49.7)^2 + \cdots + (29.2 - 49.7)^2 + (37.2 - 49.9)^2}{109 - 1}} = 8.9$$

Se va a probar, entonces, si la información se ajusta a una distribución normal con media 49.7 y desviación estándar de 8.9.

Para realizar la prueba, se formarán 10 categorías y cada una concentrará una probabilidad de 10%. Estas categorías se determinarán con los cuantiles *z* de una distribución normal estándar; una vez conocido este valor, se procede a convertirlo en la métrica de la prueba.

En la siguiente tabla se muestran los puntos de corte.

Tabla. Cálculo de los puntos de corte para formar las categorías que se utilizarán en la prueba de bondad de ajuste a una distribución normal

Corte	z	Puntaje 49.7+z·8.9
1	-1.28	38.29
2	-0.84	42.21
3	-0.52	45.03
4	-0.25	47.45
5	0.00	49.70
6	0.25	51.95
7	0.52	54.37
8	0.84	57.19
9	1.28	61.11

La tabla anterior consta de tres columnas: corte, z y puntaje. En la primera columna solamente se enumeran los puntos de corte que se requieren para dividir la distribución teórica en 10 partes iguales. La segunda (z) es el cuantil de una distribución normal estándar que acumula un área de 0.1 desde el último corte a la izquierda. Y la tercera es la conversión del valor del cuantil *z* a la métrica del examen. Esta conversión se fundamenta en que la distribución normal estándar se calcula así:

$$Z = \frac{X - \mu}{\sigma}$$

- Donde:

Z = variable estandarizada
X = variable original
μ = media de *X*
σ = desviación de *X*

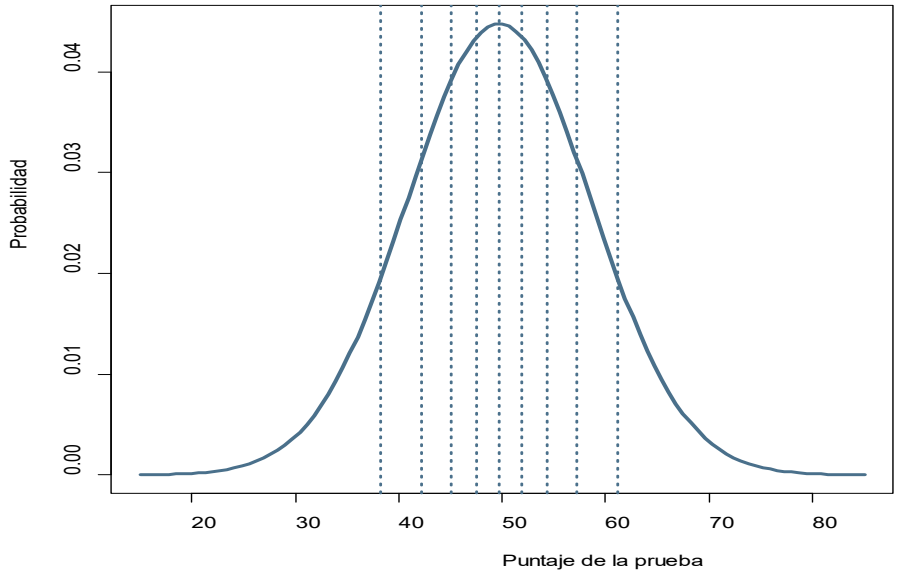
- Al despejar *X*, se obtiene:

$$X = \mu + Z \cdot \sigma$$

Así, el punto de corte en la métrica del examen se obtiene sumando al promedio (49.7) el producto del cuantil por la desviación estándar (8.9).

En la figura 2, se ilustra la segmentación de la distribución teórica con el empleo de los puntos de corte calculados.

Figura 2. Segmentación de la distribución teórica en 10 áreas iguales



Fuente: elaboración propia con empleo del paquete estadístico R[9]

[9]R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

La figura anterior muestra la segmentación en 10 áreas del mismo tamaño (0.1) de una distribución normal con media de 49.7 y desviación estándar de 8.9. El siguiente paso consiste en realizar un conteo de los aspirantes que caen en cada categoría (área) y compararlo con su número esperado: (110) (0.1) = 11. La siguiente tabla presenta las frecuencias observadas y esperadas para cada categoría.

Tabla. Frecuencias observadas y esperadas por categoría

	Frecuencia	
Categoría	**Observada**	**Estimada**
1	12	11
2	6	11
3	17	11
4	8	11
5	14	11
6	12	11
7	12	11
8	11	11
9	9	11
10	9	11
Total	**110**	**110**

Una vez que se cuenta con las frecuencias observadas y estimadas para cada categoría, se procede a realizar la prueba con el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Sustituyendo los valores, se tiene:

$$\chi^2 = \frac{(12-11)^2}{11} + \frac{(6-11)^2}{11} + \cdots + \frac{(9-11)^2}{11} = 8.2$$

A partir de la hipótesis nula, el estadístico de prueba tiene una distribución χ^2 con k – p – 1 grados de libertad. En este caso, k = 10 y p = 2 porque la distribución normal tiene dos parámetros (media y desviación estándar): se comparará el valor del

estadístico de prueba con el punto crítico de una distribución χ^2 con 10 – 2 – 1 = 7 grados de libertad que corta la curva en dos zonas: una con área de 0.05 a su derecha y la otra de 0.95.

Con Microsoft Excel (2013), se calcula el punto crítico de esta distribución así:

PRUEBA.CHI.INV(0.05, 7) = 14.07

Como el punto crítico es mayor al valor del estadístico de prueba, no se tiene evidencia estadística para rechazar la hipótesis nula. Luego, se apoya la hipótesis de que la muestra proviene de una población con distribución normal.

5.4.2. Ajuste a una distribución Poisson

En este apartado, se muestra un ejemplo donde se prueba la bondad de ajuste a una distribución Poisson.

En un establecimiento comercial, se han incrementado las quejas respecto a que no hay suficiente personal para atender a la clientela. Por su parte, los empleados solicitan al gerente que contrate más personal debido a que la demanda los supera. Con la intención de justificar la contratación de más personal, el gerente, durante una semana, tomó una muestra aleatoria de 60 periodos de 15 minutos y registró el número de clientes que acuden al establecimiento. Los registros son los siguientes:

10	6	9	8	12	9
20	15	1	20	16	1
14	16	18	0	19	9
17	1	5	4	10	4
10	20	13	10	16	19
8	17	13	9	1	6
5	10	15	10	14	9
10	15	8	3	11	8
18	17	14	17	12	9
3	2	14	15	16	1

Para realizar simulaciones, se debe estar convencido de que la distribución de las llegadas sigue una distribución Poisson. Con un nivel de confianza del 95%, se apoya la hipótesis de que las llegadas se ajustan a una distribución Poisson.

Para realizar la prueba, primero se construye una tabla de frecuencia de llegadas:

Llegadas	Casos
1	6
2	1
3	2
4	2
5	2
6	2
7	0
8	4
9	6
10	7
11	1
12	2
13	2
14	4
15	4
16	4
17	4
18	2
19	2
20	3
Promedio	**10.7**

En la primera columna de la tabla anterior, se muestra el número de llegadas registradas en periodos de 15 minutos, estas llegadas oscilan entre 1 y 20. En promedio, se registran 10.7 llegadas cada 15 minutos (este promedio se calculó utilizando el criterio de datos agrupados).

Con el propósito de no trabajar con frecuencias menores a cinco, se agruparán categorías y la tabla quedará de la siguiente forma:

Llegadas	Casos
1	6
2 a 7	9
8 a 9	10
10 y más	35

La agrupación utilizada es un tanto subjetiva (normalmente, queda al criterio del investigador).

Se busca probar que la muestra proviene de una población con distribución Poisson con parámetro $\lambda = 10.7$, por lo que el siguiente paso es calcular el valor esperado de cada categoría.

Llegadas	Casos	Probabilidad	Esperado
1 a 7	15	0.1624	10
8 a 9	10	0.2096	13
10 y más	35	0.6245	37

En la tabla anterior, se agruparon la primera y segunda categorías debido a que la frecuencia esperada de una llegada resultó cero. Para calcular la frecuencia esperada, primero se utiliza la distribución teórica Poisson con $\lambda = 10.7$. Después, la probabilidad obtenida en cada categoría se multiplica por el tamaño de la muestra (60). Una vez que se tienen los datos observados y esperados, se realiza la prueba con el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Sustituyendo los valores, se obtiene:

$$\chi^2 = \frac{(15-10)^2}{10} + \frac{(10-13)^2}{13} + \frac{(35-37)^2}{37}$$
$$\chi^2 = 3.52$$

A partir de la hipótesis nula, el estadístico de prueba tiene una distribución χ^2 con k – p – 1 grados de libertad. En este caso k = 3 y p = 1 porque la distribución Poisson tiene un parámetro, por lo que se comparará el valor del estadístico de prueba con el punto crítico de una distribución χ^2 con 3 – 1 – 1 = 1 grados de libertad que corta la curva en dos zonas: una con área de 0.05 a su derecha y la otra de 0.95.

Con Microsoft Excel (2013), se calcula el punto crítico de esta distribución así:

PRUEBA.CHI.INV(0.05, 1) = 3.84

Como el punto crítico es mayor al valor del estadístico de prueba, no se tiene evidencia estadística para rechazar la hipótesis nula: se apoya la hipótesis de que la muestra proviene de una población con distribución Poisson.

5.4.3. Ajuste a una distribución binomial

Para finalizar el empleo de la χ^2para ajustar a una distribución teórica, a continuación se presenta un ejercicio donde se desea probar que un conjunto de datos proviene de una distribución Binomial.

El expediente de un trámite se compone de cuatro documentos; si un documento está mal llenado, el expediente se clasifica como erróneo.

La auditoría realizada a la organización que elabora los expedientes mostró los siguientes resultados:

Documentos erróneos	Expedientes
0	130
1	150
2	200
3	120
4	50
Total	**650**

Antes de establecer alguna métrica, el auditor desea verificar que los expedientes con errores siguen una distribución binomial con un nivel de confianza del 95%.

La distribución binomial tiene dos parámetros: la probabilidad de éxito (p) y el número de ensayos (k). Si se define la variable teórica como el número de documentos con error de los cuatro que forman el trámite, k = 4 y p es la probabilidad de que un documento tenga error. Como *k* ya se conoce, el siguiente paso es estimar *p* minúscula. Para hacerlo, se calcula el promedio bajo un criterio de datos agrupados y se divide entre *k*. Realizando estas operaciones, la estimación de p = 0.42692.

Estimados los parámetros de la distribución teórica, se procede a calcular los valores esperados. Primero, se calculan las probabilidades de cada categoría y después la probabilidad calculada se multiplica por el total de expedientes.

En la siguiente tabla, se muestran las frecuencia observadas y estimadas:

Documentos erróneos	Expedientes	Probabilidad	Esperados
0	130	0.108	70
1	150	0.321	209
2	200	0.359	233
3	120	0.178	116
4	50	0.033	22
Total	**650**	**1**	**650**

Por último, se realiza la prueba con el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Sustituyendo los valores, se obtiene:

$$\chi^2 = \frac{(130-70)^2}{70} + \frac{(150-209)^2}{209} + \frac{(200-233)^2}{203} + \frac{(120-116)^2}{116} + \frac{(50-22)^2}{22}$$

$$\chi^2 = 110.1$$

A partir de la hipótesis nula, el estadístico de prueba tiene una distribución χ^2 con k – p – 1 grados de libertad. En este caso, k = 5 y p = 2 porque la distribución binomial tiene dos parámetros; entonces, se comparará el valor del estadístico de prueba con el punto crítico de una distribución χ^2 con 5 – 2 – 1 = 2 grados de libertad que corta la curva en dos zonas: una con área de 0.05 a su derecha, y la otra de 0.95.

Con Microsoft Excel (2013), se calcula el punto crítico de esta distribución así:

PRUEBA.CHI.INV(0.05, 2) = 5.99

Como el punto crítico es menor al valor del estadístico de prueba, no se tiene evidencia estadística para apoyar la hipótesis nula, es decir, se rechaza la hipótesis de que la muestra proviene de una población con distribución binomial.

5.5. Pruebas sobre la independencia entre dos variables

En las secciones 5.3 y 5.4, se mostró el uso de la distribución χ^2 para realizar pruebas acerca de la distribución de una población. Otra aplicación de la distribución es para determinar independencia entre dos variables cualitativas. Por ejemplo, podría ser de interés para el gerente de marca de una bebida gaseosa determinar si existe asociación entre el apego emocional a la marca respecto al consumo del producto; o al gerente de recursos humanos de una organización le sería de utilidad identificar la asociación entre el nivel de puntualidad de los empleados respecto a su zona de residencia. A continuación, se expone el empleo de la distribución χ^2 para determinar asociación entre variables.

Antes de entrar en materia, conviene repasar algunos conceptos revisados en el curso de Estadística Descriptiva referentes a probabilidad.

Independencia de eventos

Con frecuencia, es necesario determinar la probabilidad de dos eventos independientes. Los eventos A y B son independientes si P(A y B) = P(A)·P(B), es decir, si dos eventos son independientes, entonces, la probabilidad de que ocurran al mismo tiempo es el producto de sus probabilidades.

Como una extensión, si las variables X_1 y X_2 son independientes, su función conjunta

$$f(x_1, x_2) = f(x_1) \cdot f(x_2)$$

Para ilustrar lo anterior, se expone el siguiente ejemplo. Supóngase que la variable X_1 está asociada al resultado de un curso de estadística (aprobado, reprobado), donde la probabilidad de aprobar es 0.3 y la variable X_2 el sexo del alumno (mujer, hombre), siendo la probabilidad que una mujer tome el curso de 0.2.

En la tabla 1, se ilustra la distribución de ambas variables.

Tabla 1. Distribución de las variables X_1 y X_2

Género	Aprueba	Reprueba	$f(x_2)$
Mujer			0.2
Hombre			0.8
$f(x_1)$	**0.3**	**0.7**	**1.0**

En la tabla anterior, se presentan las variables de interés: por fila se muestra los valores de la variable X_2 (género del alumno); y en las columnas, los valores asociados a X_1 (resultado del curso). En los márgenes de la tabla se encuentran las distribuciones de probabilidad de las variables X_1 y X_2, denominadas distribuciones marginales.

Si X_1 y X_2 fueran independientes, su distribución conjunta $f(x_1,x_2)$ serían los valores de las celdas de la tabla, resultado de multiplicar las distribuciones marginales.

En la tabla 2, aparece el cálculo de la distribución conjunta.

Tabla 2. Cálculo de la distribución conjunta de X_1 y X_2

Género	Aprueba	Reprueba	$f(x_2)$
Mujer	0.2·0.3 = 0.06	0.2·0.7 = 0.14	0.2
Hombre	0.3·0.8 = 0.24	0.8·0.7 = 0.56	0.8
$f(x_1)$	**0.3**	**0.7**	**1.0**

Los valores de cada celda de la tabla son el resultado de multiplicar el valor de la distribución marginal en la fila por el de la columna.

Con lo anterior, si el grupo se compone de 60 alumnos, ¿cuántos se esperaría observar en cada categoría? Para responder esta pregunta, se multiplica los 60 por la probabilidad conjunta correspondiente, como se muestra a continuación.

Tabla 3. Distribución esperada de 60 alumnos conforme a la distribución conjunta de X_1 y X_2

Género	Aprueba	Reprueba	Total
Mujer	60·0.06 = 4	60·0.14 = 8	12
Hombre	60·0.24 = 14	60·0.56 = 34	48
Total	**18**	**42**	**60**

Por último, se muestra el uso de la distribución χ^2 para determinar independencia entre dos valores.

Tablas cruzadas

Una tabla cruzada se utiliza para clasificar observaciones de una muestra de acuerdo con dos o más características (variables cualitativas). Si las variables involucradas en la tabla son independientes, la distribución conjunta tiene una distribución χ^2 con (r – 1)·(c – 1) grados de libertad, donde *r* es el número de renglones de la tabla y *c* sus columnas.

De nuevo, el estadístico de prueba es:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

A continuación, se muestra un ejemplo.

La opinión de los alumnos asignados a la licenciatura de la UNAM sobre su nivel de preparación precedente se muestra a continuación por tipo de ingreso.

Opinión de los alumnos asignados a licenciaturas de la UNAM sobre su preparación precedente por tipo de ingreso

Tipo de ingreso	Excelente	Buena	Regular	Deficiente	Total
Pase reglamentado	6,160	15,184	1,276	80	22,700
Concurso de selección	4,012	9,007	1,298	139	14,456
Total	**10,172**	**24,191**	**2,574**	**219**	**37,156**

Fuente: Perfiles de Aspirantes y Asignados a Bachillerato, Técnico y Licenciatura de la UNAM 2013-2014. Dirección General de Planeación. UNAM.

¿Con un nivel de confianza de 95% se apoyaría la hipótesis de que la opinión del alumno respecto a su preparación previa a la licenciatura es independiente del tipo de ingreso?

Prueba de hipótesis:

H_0: *La opinión sobre la preparación precedente es independiente del tipo de ingreso*

H_a: *Existe asociación entre la opinión sobre la preparación precedente y el tipo de ingreso*

Para responder la pregunta, primero se calculan los valores esperados: se calculan las distribuciones marginales dividiendo los totales por fila y columna entre el total general. Por ejemplo, la proporción de alumnos que respondió excelente es $\frac{10,172}{37,156} = \mathbf{0.27}$; y la proporción de alumnos que ingresó por pase reglamentado, $\frac{22,700}{37,156} = \mathbf{0.61}$. El resto de las proporciones se muestra a continuación.

Tipo de ingreso	Excelente	Buena	Regular	Deficiente	Total
Pase reglamentado					0.61
Concurso de selección					0.39
Total	**0.27**	**0.65**	**0.07**	**0.01**	**1.00**

El siguiente paso consiste en calcular el valor esperado de cada celda de la tabla, al que se llega multiplicando el total general (37,156) por el producto de la probabilidad de la fila y de la columna. Por ejemplo, el valor esperado de alumnos de pase reglamentado que respondieron Excelente es el siguiente:

$$\mathbf{37,156 \cdot (0.61 \cdot 0.27) = 6,214}$$

De esta manera, los valores esperados se muestran a continuación.

Tipo de ingreso	Excelente	Buena	Regular	Deficiente	Total
Pase reglamentado	6,214	14,779	1,573	134	22,700
Concurso de selección	3,958	9,412	1,001	85	14,456
Total	**10,172**	**24,191**	**2,574**	**219**	**37,156**

Obtenidos los valores esperados, sigue calcular el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 = \frac{(6,160 - 6,214)^2}{6,214} + \frac{(15,184 - 14,779)^2}{14,779} + \frac{(1,276 - 1,573)^2}{1,573} + \frac{(80 - 134)^2}{134}$$
$$+ \frac{(4,012 - 3,958)^2}{3,958} + \frac{(9,007 - 9,412)^2}{9,412} + \frac{(1,298 - 1,001)^2}{1,001} + \frac{(139 - 85)^2}{85}$$
$$\chi^2 = 229.0608$$

Se rechazará la hipótesis nula si el estadístico de prueba es mayor al punto crítico.

Si se asume que la hipótesis nula es cierta, el estadístico de prueba tiene una distribución χ^2 con (r – 1) (c – 1), donde r = 2 renglones y c = 4 columnas, por tanto, tiene (2 – 1)·(4 – 1) = 3 grados de libertad.
Con Excel, se obtiene como punto crítico:

Como el valor de la prueba es notablemente mayor al punto crítico, se rechaza la hipótesis nula: se apoya que la opinión del estudiante sobre su preparación previa se encuentra asociada a su procedencia (tipo de ingreso).

5.6. Pruebas de homogeneidad

En la sección precedente, se utilizó la distribución χ^2 para determinar si dos variables son independientes; ahora, se empleará para comprobar que dos o más muestras son homogéneas.

Que dos o más muestras sean homogéneas significa que provienen de la misma población, por lo que es de esperarse que presenten un comportamiento similar. Supóngase que se desea realizar un estudio para determinar las causas por las que los alumnos de la carrera de Administración no tienen un buen desempeño en la materia de Estadística Inferencial. Se escogen al azar cuatro grupos (dos del turno matutino y dos del vespertino) y se obtiene la distribución de calificaciones en la materia, como se muestra a continuación.

Tabla. Distribución de las calificaciones del curso de Estadística Inferencial en cuatro grupos de Administración

Grupo	Calificación del curso				Total
	5	6.0 a 7.5	7.6 a 8.5	8.6 a 10	
Matutino$_1$	7	50	9	6	72
Matutino$_2$	9	55	8	7	79
Vespertino$_1$	6	40	6	5	57
Vespertino$_2$	7	35	7	6	55
Total	29	180	30	24	263

La tabla anterior presenta la distribución de calificaciones de Estadística Inferencial de 263 alumnos provenientes de los cuatro grupos seleccionados: 180 (68%) tiene calificaciones entre 6 y 7.5; y 24 (9%), notas mayores a 8.5.

Antes de continuar, los académicos responsables de la investigación quieren verificar que las muestras (los grupos) sean homogéneas con un nivel de confianza de 95% para generalizar los resultados que se obtengan, por lo que realizan una prueba de homogeneidad de muestras.

Hipótesis que contrastan:

H_0: las muestras son homogéneas

H_1: las muestras no son homogéneas

Así como se procedió para probar si dos variables son independientes, en este caso se utilizará el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Su distribución bajo la hipótesis nula es χ^2 con (r – 1)·(c – 1) grados de libertad. Para este ejemplo, la tabla cuenta con cuatro renglones (r) y cuatro columnas (c), por lo que la distribución tendrá (4 – 1)·(4 – 1) = 9 grados de libertad.

El cálculo de los valores esperados se realiza de la misma manera que la sección anterior.

Tabla. Valores esperados conforme a la hipótesis nula

Grupo	Calificación del curso				Total
	5	6.0 a 7.5	7.6 a 8.5	8.6 a 10	
Matutino$_1$	7	50	8	7	72
Matutino$_2$	8	55	9	7	79
Vespertino$_1$	6	39	7	5	57
Vespertino$_2$	5	36	6	5	52
Total	26	180	30	24	260

Estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 = \frac{(7-7)^2}{7} + \frac{(50-51)^2}{51} + \frac{(9-9)^2}{9} + \frac{(6-7)^2}{7} + \cdots + \frac{(6-5)^2}{5}$$

$$\chi^2 = 1.1$$

Se rechazará la hipótesis nula si el estadístico de prueba es mayor al punto crítico.

El punto crítico de una distribución χ^2 con 9 grados de libertad que separa la curva en dos regiones, una de 0.95 (izquierda del punto crítico) y otra de 0.05 (derecha del punto crítico), es el siguiente.

PRUEBA.CHI.INV(0.05, 9)=16.9

Como el valor de la prueba es menor al punto crítico, se acepta la hipótesis nula y se apoya que las muestras son homogéneas.

Como última observación, al utilizar el estadístico de prueba

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

se debe cuidar que los valores observados sean al menos de cinco. De no ser así, se sugiere juntar categorías para que se cumpla esta condición; de lo contrario, la prueba pierde precisión.

RESUMEN

En esta unidad, se expuso la distribución χ^2, su uso para contrastar hipótesis relacionadas con la varianza poblacional, diferencia de proporciones, bondad de ajuste, independencia y homogeneidad.

Se utilizaron dos estadísticos de prueba: $\boldsymbol{\chi^2 = \frac{(n-1)s^2}{\sigma^2}}$ para contrastar hipótesis relacionadas con la varianza poblacional, y $\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}}$ para el resto de las pruebas expuestas. Para que este último estadístico de prueba arroje resultados confiables, se debe observar que tanto la frecuencia observada como la esperada de las categorías sean al menos de cinco.

Como valor agregado, se utilizó Excel para el cálculo de los puntos críticos, que se ha venido practicando en unidades anteriores.

BIBLIOGRAFÍA

SUGERIDA

Autor	Capítulo	Páginas
Anderson, S.	11 12	449-471 472-505
Levin, R.	11	447-468
Lind, D.	17	648-679

Licenciatura: **Administración**

UNIDAD 6

Análisis de regresión lineal simple

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 **2016** actualizado

OBJETIVO PARTICULAR

El alumno conocerá el método de regresión lineal simple, así como su aplicación e interpretación.

TEMARIO DETALLADO

(10 horas)

6. Análisis de regresión lineal simple

6.1. Ecuación y recta de regresión

6.2. El método de mínimos cuadrados

6.3. Determinación de la ecuación de regresión

6.4. El modelo de regresión y sus supuestos

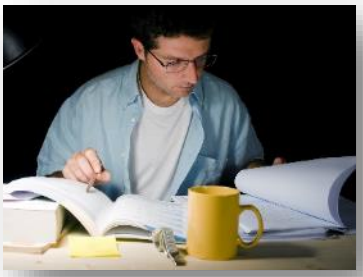
6.5. Inferencias estadísticas sobre la pendiente de la recta de regresión

6.6. Análisis de correlación

INTRODUCCIÓN

Existen situaciones donde se requiere determinar si el comportamiento de cierto suceso se explica con el conocimiento de otra información. Por ejemplo, puede ser de interés conocer el impacto del número de horas de preparación para un examen de admisión a una institución de educación superior en el porcentaje de aciertos; o la afectación de los ingresos de una organización en función del presupuesto destinado a publicidad; o la duración de la batería de un dispositivo electrónico de acuerdo con el tiempo destinado a descargar tutoriales.

Para los problemas descritos en el párrafo anterior, se emplea el análisis de regresión lineal simple, técnica que trata de explicar una variable de interés o respuesta (y) en función de otra (x), mediante un modelo lineal.



En esta unidad, se mostrarán las características del modelo de regresión y el método para estimar sus parámetros. Una vez obtenidos los parámetros, se expone cómo determinar la ecuación del modelo, los supuestos que debe cumplir y la manera de realizar inferencias sobre la pendiente de la recta de regresión. La unidad concluye con el análisis de correlación lineal entre dos variables continuas.

6.1. Ecuación y recta de regresión

En este apartado, se tratarán los conceptos del modelo de regresión lineal simple. Para entender mejor este modelo, se repasará brevemente la ecuación de la recta.

Ecuación de la recta

En el plano cartesiano, la forma de describir una recta es mediante la ecuación

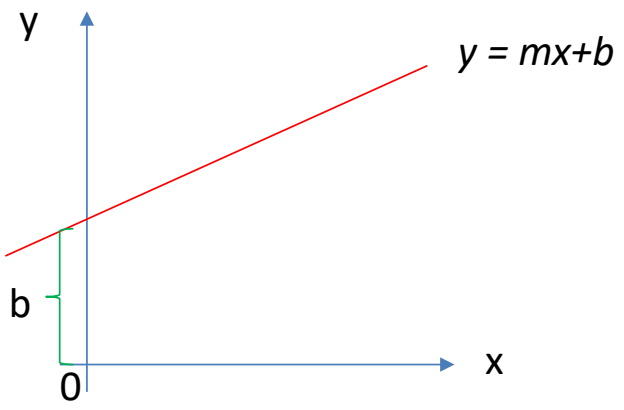
$$y = mx + b$$

- Donde:

m = pendiente de la recta
b = ordenada al origen o el punto donde intersecta la recta al eje Y, cuando x = 0

En la figura 1, se muestra una representación gráfica de la línea recta.

Figura 1. Representación gráfica de la línea recta



Fuente: elaboración propia.

La figura anterior ilustra la función de una línea recta con parámetros *m* y *b*. La pendiente *m* indica las unidades que se mueve *y* por cada unidad de cambio en *x*, y *b* es la intersección de la recta con el eje de las ordenadas.

Si m > 0, la recta tiene un ángulo de inclinación positivo; es decir, cada que aumenta *x*, aumenta *y*.

Si m < 0, la recta tiene un ángulo de inclinación negativo; es decir, cada que aumenta *x*, disminuye *y*.

Si m = 0 la recta es horizontal; es decir, cada que aumenta *x*, se mantiene constante *y* en *b*.

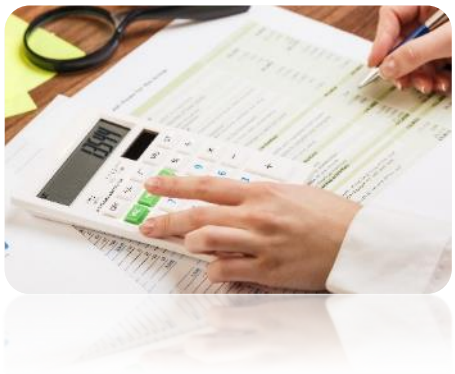
Para determinar la pendiente, es suficiente conocer dos puntos por donde atraviesa la recta $(x_1, y_1),(x_2, y_2)$ y aplicar la fórmula:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

Teniendo presente lo anterior, se presenta a continuación el modelo de regresión lineal simple.

Modelo de regresión lineal

El modelo de regresión lineal explica la relación entre una variable dependiente, a la que se denotará *y*, con otra(s) explicativa(s) a través de una ecuación de primer orden. Tanto las variables dependientes como las explicativas son observables.



Supóngase que una organización con 20 empleados realizó una evaluación del desempeño de cada empleado, y de acuerdo con el resultado se determinó un ajuste en el sueldo. Un auditor quiere explicar el incremento salarial conforme al desempeño del empleado.

En este ejemplo, el incremento salarial es la variable dependiente (y), ya que es resultado del desempeño de cada empleado (x). El incremento salarial observado del i-ésimo empleado (i = 1,2,..,20) se puede plantear de la siguiente manera:

Incremeto observado = Incremento esperado + variación (i= 1, 2, ..., 20)

Es decir, el incremento salarial observado del i-ésimo empleado tiene una parte explicable por la variable explicativa (nivel de desempeño observado) y otra no explicable, como puede ser una distracción del evaluador o su estado de salud al momento de la reunión.

Si denotamos como *y* al incremento salarial, como *x* al desempeño y como ε a la variación entre el incremento observado y estimado, entonces el incremento salarial del i-ésimo empleado (i = 1,2,..,20) se puede expresar así:

$$y_i = \mu(x_i) + \epsilon_i$$

Donde $\mu(x_i)$ representa el incremento esperado del i-ésimo empleado con su desempeño observado.

También $\mu(x_i)$ es un estimador de y_i cuya estimación depende del valor de x_i. En el modelo de regresión lineal, la regla para estimar *y* consiste en relacionarla con *x* a través de una ecuación lineal.

Regresando al ejemplo, $\mu(x_i)$ puede expresarse así:

$$\mu(x_i) = \hat{y}_i = \beta_0 + \beta_1 x_i$$

- Donde:

$\mu(x_i)$ = estimador del incremento salarial del i-ésimo empleado (i=1,2,..,20) en función del desempeño observado
β_0 = ordenada al origen de la recta de estimación
β_1 = pendiente de la recta de estimación

Entonces, el auditor puede partir del siguiente modelo para determinar el criterio de incremento salarial de los empleados de la organización:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Es el modelo de regresión lineal simple.

Ahora, cuando solamente se emplea una variable explicativa, al modelo de regresión lineal se le denomina *simple* y se modela con la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Donde:

Y_i = variable dependiente o respuesta de la i-ésima observación
β_0 = intersección con el eje Y
β_1 = pendiente de la recta
X_i = variable independiente o explicativa de la i-ésima observación
ε_i = error no observable de la i-ésima observación
i = 1,2,.., n.

Cuando hay más de una variable explicativa, el modelo de regresión lineal es múltiple y se modela con la siguiente ecuación:

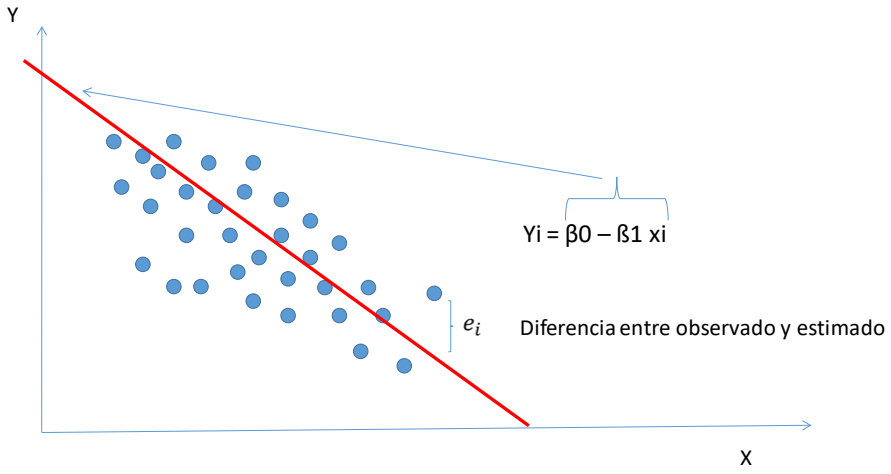
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Donde:

Y = variable dependiente o respuesta con *n* observaciones
β_0: intersección con el eje Y
$\beta_1, \beta_2, \ldots, \beta_p$ = razón de cambio de Y respecto a cada variable explicativa manteniendo el resto sin cambio.
X_1, X_2 y X_p = variables independientes o explicativas, cada una de *n* observaciones
ε: error entre Y observada y estimada

Este material de estudio se enfocará al modelo de regresión lineal simple, en el cual se estima una recta que cruce a lo largo de la información con la intención de explicar el comportamiento de la variable de interés, como lo ilustra la figura 2.

Figura 2. Ilustración del modelo de regresión lineal simple



Fuente: elaboración propia.

La figura anterior ilustra un gráfico de dispersión donde cada punto azul representa el valor de la variable respuesta (Y) observado con el valor de la variable explicativa (X), la línea roja es la recta estimada que se ajusta al conjunto de datos, cuya ecuación es $\boldsymbol{Y_i = \beta_0 - \beta_1 X_i}$, y la diferencia entre el valor observado y el estimado con la ecuación de regresión lineal es el error.

En el ejemplo de los incrementos salariales de la organización de 20 empleados, en el eje X se representaría el desempeño del empleado; y en el eje Y, el incremento salarial. Los puntos azules serían el incremento salarial observado de cada empleado asociado a su desempeño; y la línea roja, el modelo de regresión lineal simple. En el siguiente apartado, se explica cómo calcular la recta de regresión lineal simple.

6.2. El método de mínimos cuadrados

En la parte final de la sección anterior, en la figura 2 se ilustró cómo la recta de regresión lineal simple atraviesa el conjunto de datos; sin embargo, el número de rectas que se pueden trazar es infinito, por lo que surge la pregunta sobre cuál es la recta conveniente. La respuesta no es difícil, dado que lo deseable es que la diferencia entre el valor estimado y observado de una observación sea la menor posible.

Partiendo del modelo para una observación cualquiera:

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Entonces, el error es la diferencia entre los valores observados y estimados:

- $y_i - \beta_0 - \beta_1 x_i = \varepsilon_i$

Error de todas las observaciones (n):

- $\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i = \sum_{i=1}^{n} \varepsilon_i$

Como se explicó en la sección anterior, la recta $\beta_0 + \beta_1 x_i$ es un valor esperado de y_i, por lo que la suma de las diferencias entre los valores estimados y observados se espera sea cero. Para superar este inconveniente, se procede a trabajar con los errores al cuadrado, los cuales quedan expresados así:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

La recta que se busca es de parámetros $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$ y minimiza la expresión del lado derecho. A esta metodología para obtener la recta que garantiza el menor error de estimación se le conoce como *mínimos cuadrados*.

Los valores de los parámetros $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$, por el método de mínimos cuadrados, son los siguientes:

$$\boldsymbol{\beta_0} = \bar{y} - \boldsymbol{\beta_1}\bar{x}$$

$$\boldsymbol{\beta_1} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

- Donde:

$\boldsymbol{\beta_0}$: intersección con el eje Y
$\boldsymbol{\beta_1}$: pendiente de la recta de regresión lineal simple
$\bar{y}$: promedio de la variable dependiente
$\bar{x}$: promedio de la variable independiente
n: número de observaciones
x_i: i-ésima observación de la variable independiente (i = 1,..,n)
y_i: i-ésima observación de la variable dependiente (i = 1,..,n)

A continuación, se muestra a manera de ejemplo cómo estimar una recta de regresión lineal simple por mínimos cuadrados.

Una PYME que imparte clases de manejo a personas de entre 30 y 65 años, para negociar las condiciones de su póliza de accidentes con la compañía de seguros que les ofrece el servicio, quiere conocer la relación entre el número de accidentes automovilísticos en la localidad donde se encuentra el negocio. La información se presenta a continuación.

Accidentes automovilísticos por edad del conductor

ID	Edad	Accidentes	ID	Edad	Accidentes
1	30	1,004	19	48	504
2	31	946	20	49	432
3	32	914	21	50	456
4	33	742	22	51	346
5	34	714	23	52	382
6	35	842	24	53	334
7	36	744	25	54	298
8	37	792	26	55	252
9	38	844	27	56	240
10	39	722	28	57	244
11	40	982	29	58	288
12	41	644	30	59	218
13	42	594	31	60	208
14	43	604	32	61	146
15	44	480	33	62	130
16	45	570	34	63	130
17	46	440	35	64	122
18	47	410	36	65	104

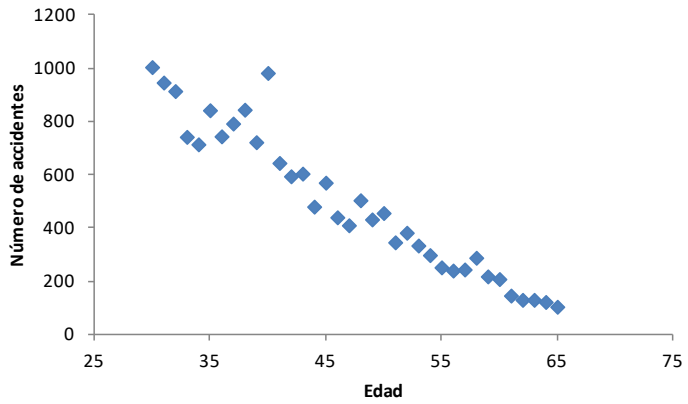
Para obtener la recta de regresión por mínimos cuadrados, se dan los siguientes pasos:

1. Determinar las variables dependientes *(Y)* e independiente(*X*).

En este problema, *Y* es el número de accidentes y *X* la edad del conductor debido a que el número de accidentes será explicado por la edad del conductor.

2. Graficar las variables *X* y *Y.*

Gráfica 1. Número de accidentes por edad del conductor



Fuente: elaboración propia con empleo de Microsoft Excel (2013).

En la gráfica 1, se ilustra el número de accidentes (Y) respecto a la edad del conductor (X). Se aprecia como patrón que, conforme el conductor es mayor, el riesgo de tener un accidente disminuye.

3. Calcular los parámetros de la recta de regresión que atraviesa el conjunto de datos por mínimos cuadrados.

A continuación, se calcula la pendiente de la recta:

$$\beta_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Obsérvese que en la fórmula se requieren cinco sumas, cuyo cálculo se muestra en la siguiente tabla.

Tabla 1. Memoria de cálculo de los elementos de la fórmula para calcular β_1 mediante mínimos cuadrados

	1 X_i Edad	2 Y_i Número de accidentes		1-2 X_iY_i		$(1)^2$ X_i^2	n
	30	1004		30120		900	36
	31	946		29326		961	
	32	914		29248		1024	
	33	742		24486		1089	
	34	714		24276		1156	
	35	842		29470		1225	
	36	744		26784		1296	
	37	792		29304		1369	
	38	844		32072		1444	
	39	722		28158		1521	
	40	982		39280		1600	
	41	644		26404		1681	
	42	594		24948		1764	
	43	604		25972		1849	
	44	480		21120		1936	
	45	570		25650		2025	
	46	440		20240		2116	
	47	410		19270		2209	
	48	504		24192		2304	
	49	432		21168		2401	
	50	456		22800		2500	
	51	346		17646		2601	
	52	382		19864		2704	
	53	334		17702		2809	
	54	298		16092		2916	
	55	252		13860		3025	
	56	240		13440		3136	
	57	244		13908		3249	
	58	288		16704		3364	
	59	218		12862		3481	
	60	208		12480		3600	
	61	146		8906		3721	
	62	130		8060		3844	
	63	130		8190		3969	
	64	122		7808		4096	
	65	104		6760		4225	
$\sum X_i$	**1710**	$\sum Y_i$	**17822**	$\sum X_i Y_i$	**748570**	$\sum X_i^2$	**85110**
$(\sum X_i)^2$	**2924100**						
		$\sum X_i \sum Y_i$	**30475620**				

Fuente: elaboración propia con empleo de Microsoft Excel (2013).

La tabla anterior presenta el cálculo de los elementos de la fórmula de la pendiente de la recta de regresión de mínimos cuadrados. La primera columna contiene la edad del conductor (X); la segunda, el número de accidentes reportados para cada edad (Y). La tercera columna se obtiene multiplicando las dos primeras, por ejemplo, el primer elemento de esta columna (30,120) es resultado de multiplicar el primer valor de la primera (30) por el primer valor de la segunda (1,004). La cuarta columna es resultado de multiplicar la primera por sí misma. Regresando a analizar el primer elemento (900), este se obtuvo de multiplicar por sí mismo el primer elemento de la primera columna (30). En la parte final, se encuentran las sumas y multiplicaciones que se requiere sustituir en la fórmula.

Sustituyendo, la pendiente es la siguiente:

$$\beta_{1=} \frac{(36 \cdot 748570) - 30475620}{(36 \cdot 85110) - 2924100}$$

$$\beta_{1=} \frac{26948520 - 30475620}{3063960 - 2924100}$$

$$\beta_{1=} \frac{-3527100}{139860}$$

$$\beta_{1=} -25.218$$

Y la ordenada al origen:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\bar{Y} = \frac{17822}{36}$$

$$\bar{Y} = 495.055$$

$$\bar{X} = \frac{17822}{36}$$

$$\bar{X} = 47.5$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_0 = 495.055 - (-25.218 \cdot 47.5)$$

$$\beta_0 = 495.055 - 1197.892$$

$$\beta_0 = 1692.948$$

De esta manera, se obtienen los parámetros de la recta de regresión lineal simple con el método de mínimos cuadrados. En la siguiente sección, se expone cómo determinar la ecuación de regresión lineal simple.

6.3. Determinación de la ecuación de regresión

Como se ha mencionado, el modelo de regresión lineal simple estima el valor observado de la variable dependiente (Y) a partir de la explicativa (X) con la ecuación de una recta. Una vez determinados los valores de los parámetros mediante mínimos cuadrados, la estimación de los valores de *Y* se realiza con la ecuación de regresión lineal simple:

$$\widehat{Y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} X_i$$

En el ejemplo anterior, $\boldsymbol{\beta_0} = \mathbf{1692.948}$ (1,693) y $\boldsymbol{\beta_1} = \mathbf{-25.218}$ (-25.2) por lo que la ecuación de regresión lineal simple es la siguiente:

$$\widehat{Y_i} = \mathbf{1,693 - 25.2} X_i$$

- Donde:

$\widehat{Y_i}$ = estimación del número de accidentes para conductores en la i-ésima observación. (i=1,2,…,36)
X_i = edad del conductor en la i-ésima observación. (i=1,2,…,36)

En esta ecuación, $\boldsymbol{\beta_0}$ indica que, cuando X = 0, se espera observar 1693 accidentes, lo que en el contexto del problema no tiene sentido, porque la edad de interés es entre 30 y 65. Por otro lado, la pendiente de la ecuación tiene una dirección negativa, esto significa que, conforme se avance en edad, se espera observar menos accidentes. El valor de la pendiente (–25.2) indica que, por cada año que aumenta la edad del conductor, el número de accidentes disminuye en 25.

6.4. El modelo de regresión y sus supuestos

Un aspecto fundamental cuando se trabaja con esta técnica es que el modelo de regresión lineal simple es estimado con los valores de una muestra, por lo que los valores obtenidos de $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$ son estimaciones de los parámetros de la recta con toda la población[10]. Así, el propósito del modelo no es solamente calcular los parámetros, sino realizar inferencia sobre los verdaderos valores de esos parámetros. Por lo anterior, es necesario considerar los siguientes supuestos al emplear una regresión lineal simple.

1. En el modelo de regresión lineal simple

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i} (i = 1,.., n)$$

tanto la variable dependiente (Y) como la explicativa (X) son observables.

[10]Los estimadores de β_0 y β_1 son insesgados.

2. El modelo es lineal en los parámetros no en las variables. Esto significa que se pueden realizar transformaciones sobre las variables originales para que haya una relación lineal, y la esencia del modelo no se pierde.

3. El error de estimación $\boldsymbol{\varepsilon_i}$ es una variable aleatoria cuyo valor esperado es cero y su varianza es $\boldsymbol{\sigma^2}$, la cual se mantiene constante en todas las observaciones y es desconocida.

4. Los errores $\boldsymbol{\varepsilon_i}$ son independientes. Esto significa que, dados dos valores cualesquiera de X, x_i, x_j ($i \neq j$), los errores $\boldsymbol{\varepsilon_i}$, $\boldsymbol{\varepsilon_j}$ son independientes.[11]

5. El error $\boldsymbol{\varepsilon_i}$ es una variable aleatoria con distribución normal. Al ser *y* una función lineal del error, también se distribuye normalmente.

Uno de los aspectos que más se descuida al ajustar un modelo de regresión lineal simple es revisar que se cumplan los supuestos del modelo (esta revisión implica analizar el comportamiento de los residuos). Como este tema no está incluido en el plan de estudios, no se abordará; sin embargo, se sugiere profundizarlo en Anderson (2012), parte de la bibliografía citada al término de la unidad.

[11]O al menos no correlacionados.

6.5. Inferencias estadísticas sobre la pendiente de la recta de regresión

Como se mencionó en la sección anterior, el propósito del modelo de regresión lineal simple no se reduce a calcular los parámetros de la recta, sino que implica realizar inferencia sobre ellos. Cuando se ajusta un modelo de regresión, la primera prueba efectuada es referente a si un modelo lineal es el adecuado para los datos, y posteriormente se hacen inferencias sobre la pendiente. En este apartado, se expondrá como llevar a cabo inferencias sobre la pendiente de la recta de regresión.

Para establecer inferencias con la pendiente del modelo, se contrastan las siguientes hipótesis:

$$H_0: \boldsymbol{\beta_1} = 0$$
$$H_1: \boldsymbol{\beta_1} \neq 0$$

La hipótesis nula significa que el valor de la pendiente del modelo no es importante: la variable X no tiene efecto sobre Y, es decir, X no es una variable explicativa de Y.

La hipótesis alternativa plantea que el valor de la pendiente sí es importante: X tiene efecto sobre Y.

Rechazar la hipótesis nula significa que la variable X es una variable explicativa de Y. Esto implica que el modelo puede aplicarse.

El estadístico de prueba empleado para contrastar la hipótesis nula es el siguiente:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s} \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- Donde:

$\widehat{\beta_1}$ = estimador de la pendiente de la recta de regresión
β_1 = pendiente de la recta de regresión asumiendo cierta la hipótesis nula
s = estimador de la desviación estándar, el cual es

$$S = \sqrt{\frac{\sum (Y_i - \hat{Y})^2}{n - 2}}$$

El estadístico de prueba tiene una distribución t de Student con n – 2 grados de libertad.

En la figura 3, se ilustra una prueba ubicada en zona de rechazo.

Figura 3. Ilustración de una prueba donde se rechaza la hipótesis nula

Fuente: elaboración propia.

La figura 3 ilustra una prueba donde el estadístico de prueba se ubica en la zona de rechazo, lo que significa que la pendiente tiene un valor significativo. Al final de la unidad, se muestra un ejemplo de cómo realizar inferencias de la pendiente con Microsoft Excel (2013).

En el ejemplo de los accidentes, se mencionó que el modelo ajustado es

$$\widehat{Y_i} = 1,693 - 25.2X_i$$

La pregunta es, entonces, si los coeficientes son significativos. Para responder esto, se realiza la prueba de hipótesis, donde H_0 es que los coeficientes son cero (no tienen un valor significativo). El resultado de la prueba se muestra a continuación.

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	1692.9	58.6	28.9	1.64442E-25
Edad	-25.2	1.2	-20.9	5.35232E-21

Fuente: Microsoft Excel (2013). Módulo de análisis de datos.

La tabla anterior muestra los valores de los coeficientes del modelo, su error, su estadístico de prueba y resaltado. Se ve la significancia de la prueba (p value), y como esta prueba es menor a 0.05, se rechaza H_0: los coeficientes son significativos.

6.6. Análisis de correlación

En el análisis de regresión lineal simple, si la variable *X* es explicativa de *Y*, entonces el modelo muestra el efecto de un cambio en *X* sobre *Y*. Un análisis complementario es el de correlación, el cual determina el grado de asociación lineal entre dos variables.

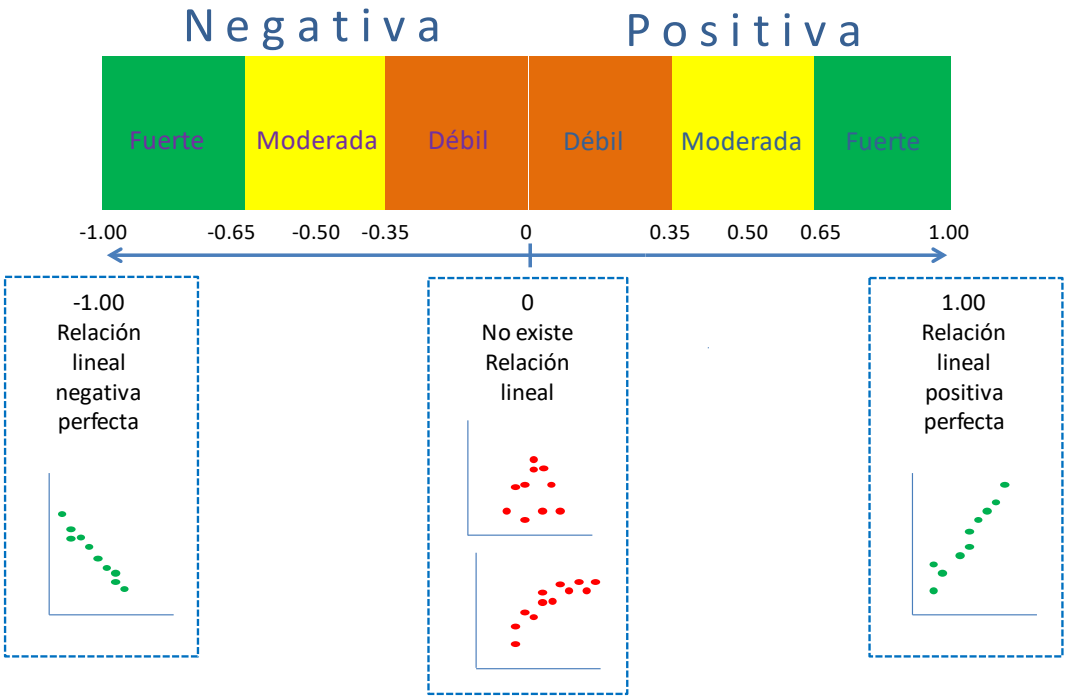
La correlación entre las variables *X* y *Y* se denota como ρ_{xy}, y se define como el grado en que se encuentran asociadas estas variables. El estimador de esta correlación es conocido como *coeficiente de correlación*, denotado como *r*, y su fórmula es

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

El coeficiente de correlación es un valor independiente de las unidades de las variables, lo que permite que pueda ser empleado en comparativos; toma valores entre –1 y 1 (en –1 significa que existe una asociación lineal perfecta negativa, es decir, el incremento de la variable explicativa resultará una disminución en la variable respuesta; y en 1, la asociación lineal entre las variables es perfecta y positiva, lo que implica que un aumento de la variable explicativa hará que aumente el valor de la variable respuesta). Cuando el coeficiente de correlación es cero, significa que las variables no están asociadas o que su asociación no es lineal.

La figura 4 muestra una categorización de la asociación entre dos variables en función del valor del coeficiente de correlación.

Figura 4. Nivel de asociación de dos variables de acuerdo con el valor del coeficiente de correlación



Fuente: elaboración propia.

En la figura anterior, se muestra cómo interpretar los niveles de asociación entre dos variables de acuerdo con el valor del coeficiente de correlación. Un valor mayor a cero indica que existe una correlación positiva; en caso contrario, la correlación es negativa. Las variables se considerarán con una asociación débil si su correlación tiene un valor absoluto entre 0 y 35; moderada, entre 35 y 65; y fuerte, mayor a 65.

Para el ejemplo del número de accidentes por edad del conductor, la correlación entre las dos variables es de –0.9633, lo que significa que la asociación entre las variables es casi negativa perfecta.

La tabla 2 muestra la memoria de cálculo de los elementos que forman parte de la fórmula de la correlación de las variables. En la parte superior de la tabla, se numera la columna (del 1 al 9) y en algunos casos, debajo de este número, se indican las

columnas involucradas en la obtención de sus cifras. Por ejemplo, los valores de la columna 5 se obtienen de restarle a la edad (columna 1) el promedio de edad (columna 2). Los valores involucrados en la fórmula del coeficiente de correlación son los dos que se hallan en la parte inferior derecha, y al sustituirlos se obtiene lo siguiente:

$$r = \frac{-97075}{\sqrt{103444464748}}$$

$$r = \frac{-97075}{101707.7418}$$

$$r = -0.9633$$

Es decir, el resultado comentado.

Tabla 2. Memoria de cálculo de los elementos de la fórmula para calcular *r* entre el número de accidentes y la edad del conductor

1	2	3	4	5	6	7	8	9
X_i	$\bar{X}$	Y_i	$\bar{Y}$	(1-2) $(X_i-\bar{X})$	(1-2)^2 $(X_i-\bar{X})^2$	(3-4) $(Y_i-\bar{Y})$	(3-4)^2 $(Y_i-\bar{Y})^2$	5-7
Edad	Promedio de X	Número de accidentes	Promedio de Y					
30	47.5	1004	495.06	-17.5	306.25	508.94	259024.45	-8906.52778
31		946		-16.5	272.25	450.94	203350.89	-7440.58333
32		914		-15.5	240.25	418.94	175514.45	-6493.63889
33		742		-14.5	210.25	246.94	60981.56	-3580.69444
34		714		-13.5	182.25	218.94	47936.67	-2955.75
35		842		-12.5	156.25	346.94	120370.45	-4336.80556
36		744		-11.5	132.25	248.94	61973.34	-2862.86111
37		792		-10.5	110.25	296.94	88176.00	-3117.91667
38		844		-9.5	90.25	348.94	121762.23	-3314.97222
39		722		-8.5	72.25	226.94	51503.78	-1929.02778
40		982		-7.5	56.25	486.94	237114.89	-3652.08333
41		644		-6.5	42.25	148.94	22184.45	-968.138889
42		594		-5.5	30.25	98.94	9790.00	-544.194444
43		604		-4.5	20.25	108.94	11868.89	-490.25
44		480		-3.5	12.25	-15.06	226.67	52.6944444
45		570		-2.5	6.25	74.94	5616.67	-187.361111
46		440		-1.5	2.25	-55.06	3031.11	82.5833333
47		410		-0.5	0.25	-85.06	7234.45	42.5277778
48		504		0.5	0.25	8.94	80.00	4.47222222
49		432		1.5	2.25	-63.06	3976.00	-94.5833333
50		456		2.5	6.25	-39.06	1525.34	-976388889
51		346		3.5	12.25	-149.06	22217.56	-521.694444
52		382		4.5	20.25	-113.06	12781.56	-508.75
53		334		5.5	30.25	-161.06	25938.89	-885.805556

54	298	6.5	42.25	-197.06	38830.89		-1280.86111	
55	252	7.5	56.25	-243.06	59076.00		-1822.91667	
56	240	8.5	72.25	-255.06	65053.34		-2167.97222	
57	244	9.5	90.25	-251.06	63028.89		-2385.027778	
58	288	10.5	110.25	-207.06	42872.00		-2174.08333	
59	218	11.5	132.25	-277.06	76759.78		-3186.13889	
60	208	12.5	156.25	-287.06	82400.89		-3588.19444	
61	146	13.5	182.25	-349.06	121839.78		-4712.25	
62	130	14.5	210.25	-365.06	133265.56		-5293.30556	
63	130	15.5	240.25	-365.06	133265.56		-5658.36111	
64	122	16.5	272.25	-373.06	139170.45		-6155.41667	
65	104	17.5	306.25	-391.06	152924.45		-6843.47222	
			3885		2662667	$\sum(X_i-\bar{X})(Y_i-\bar{Y})$	-97975	$\sum(X_i-\bar{X})^2(Y_i-\bar{Y})^2$ **10344464748**

Fuente: elaboración propia con empleo de Microsoft Excel (2013).

Coeficiente de determinación R^2

Para valorar el ajuste del modelo de regresión lineal simple, se considera otro coeficiente llamado *coeficiente de determinación*, denotado como R^2, que mide la variabilidad explicada por el modelo. Para calcular el coeficiente de determinación, se utiliza la siguiente fórmula:

$$R^2 = \frac{\sum(\widehat{Y_i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}$$

Donde:

R^2: coeficiente de determinación
$\widehat{Y_i}$: i-ésima estimación de Y
Y_i: i-ésima observación de Y
$\overline{Y}$: promedio de Y

Para el ejemplo del número de accidentes por edad del conductor, el coeficiente de determinación del modelo ajustado entre las dos variables 0.9279, esto significa que el modelo explica en un 93% la variabilidad de la información. La tabla 3 muestra el cálculo de los elementos que intervienen en la fórmula de R^2.

Tabla 3. Memoria de cálculo de los elementos de la fórmula para calcular R^2 entre el número de accidentes y la edad del conductor

1	2	3	4	5	6	7	8
Xi	Yi	$\hat{Y}$i	$\overline{Y}$	(3-4) ($\hat{Y}$i-$\overline{Y}$)	(5)2 ($\hat{Y}$i-$\overline{Y}$)2	(2-4) (Yi-$\overline{Y}$)	(7)2 (Yi-$\overline{Y}$)2
Edad	Número de accidentes	(-25.22 edad conductor)	Promedio de Y				
30	1004	936	495.06	441	194771.14	508.94	259024.448
31	946	911		416	173147.56	450.94	203350.892
32	914	886		391	152795.97	418.94	175514.448
33	742	861		366	133716.35	246.94	60981.5586
34	714	836		340	115908.70	218.94	47936.6698
35	842	810		315	99373.03	346.94	120370.448
36	744	785		290	84109.33	248.94	61973.3364
37	792	760		265	70117.61	296.94	88176.0031
38	844	735		240	57397.86	348.94	121762.225
39	722	709		214	45950.09	226.94	51503.7809
40	982	684		189	35774.29	486.94	237114.892
41	644	659		164	26870.47	148.94	22184.4475
42	594	634		139	19238.62	98.94	9790.00309
43	604	609		113	12878.74	108.94	11868.892
44	480	583		88	7790.85	-15.06	226.669753
45	570	558		63	3974.92	74.94	5616.66975
46	440	533		38	1430.97	-55.06	3031.1142
47	410	508		13	159.00	-85.06	7234.44753
48	504	482		-13	159.00	8.94	80.0030864

49	432	457	-38		1430.97	-63.06	3976.00309
50	456	432	-63		3974.92	-39.06	1525.33642
51	346	407	-88		7790.85	-149.06	22217.5586
52	382	382	-113		12878.74	-113.06	12781.5586
53	334	356	-139		19238.52	-161.06	25938.892
54	298	331	-164		26870.47	-197.06	38830.892
55	252	306	-189		35774.29	-243.06	59076.0031
56	240	281	-214		45950.09	-255.06	65053.3364
57	244	255	-240		57397.86	-251.06	63028.892
58	288	230	-265		70117.61	-207.06	42872.0031
59	218	205	-290		84109.33	-277.06	76759.7809
60	208	180	-315		99373.03	-287.06	82400.892
61	146	155	-340		115908.70	-349.06	121839.781
62	130	129	-366		133716.35	-365.06	133265.559
63	130	104	-391		152795.97	-365.06	133265.559
64	122	79	-416		173147.56	-373.06	139170.448
65	104	54	-441		194771.14	-391.06	152924.448
				$\sum(\hat{Y}i-\bar{Y})2$	2470810.97		$\sum(Yi-\bar{Y})2$ 2662667

Fuente: elaboración propia con empleo de Microsoft Excel (2013).

Así como en la tabla 2, en la parte superior de la tabla 3 se numera la columna (del 1 al 8), y en algunos casos, debajo de este número, se indican las columnas involucradas en la obtención de sus cifras. Por ejemplo, los valores de la columna 5 se obtienen de restarle a los accidentes estimados (columna 3) el promedio observado de accidentes (columna 4). Los valores involucrados en la fórmula del coeficiente de determinación son los dos que se sitúan en la parte inferior de la tabla; y al sustituirlos se obtiene:

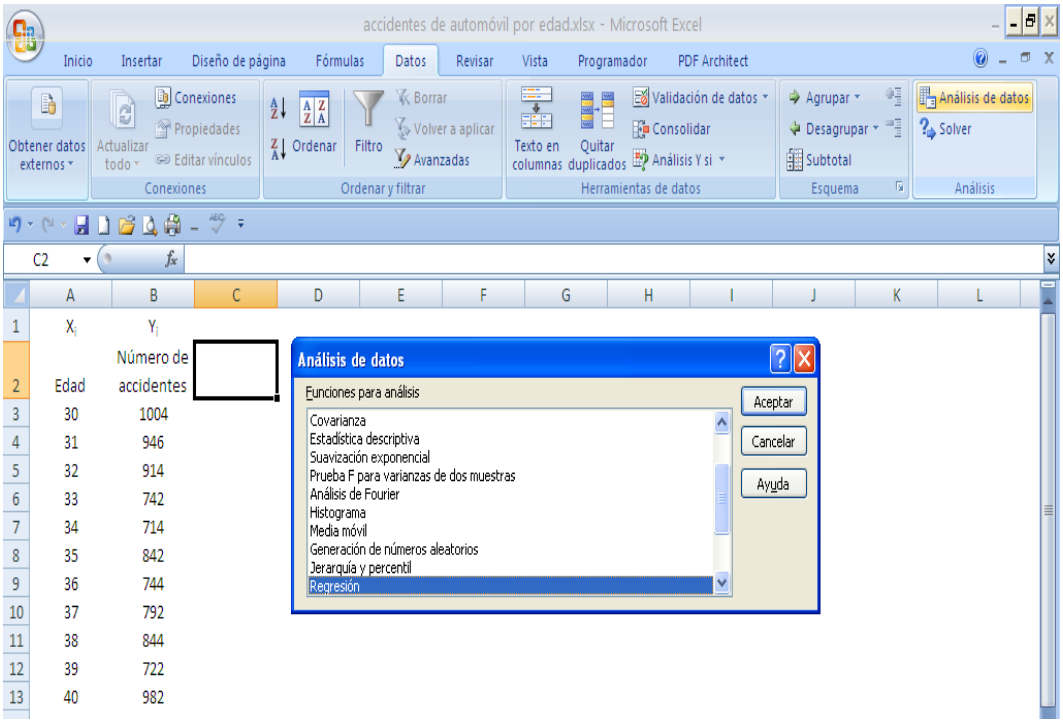
$$R^2 = \frac{2470810.97}{2662667.89}$$
$$R^2 = 0.927945$$

Es decir, el resultado comentado.

Análisis de regresión lineal simple con MS -Excel

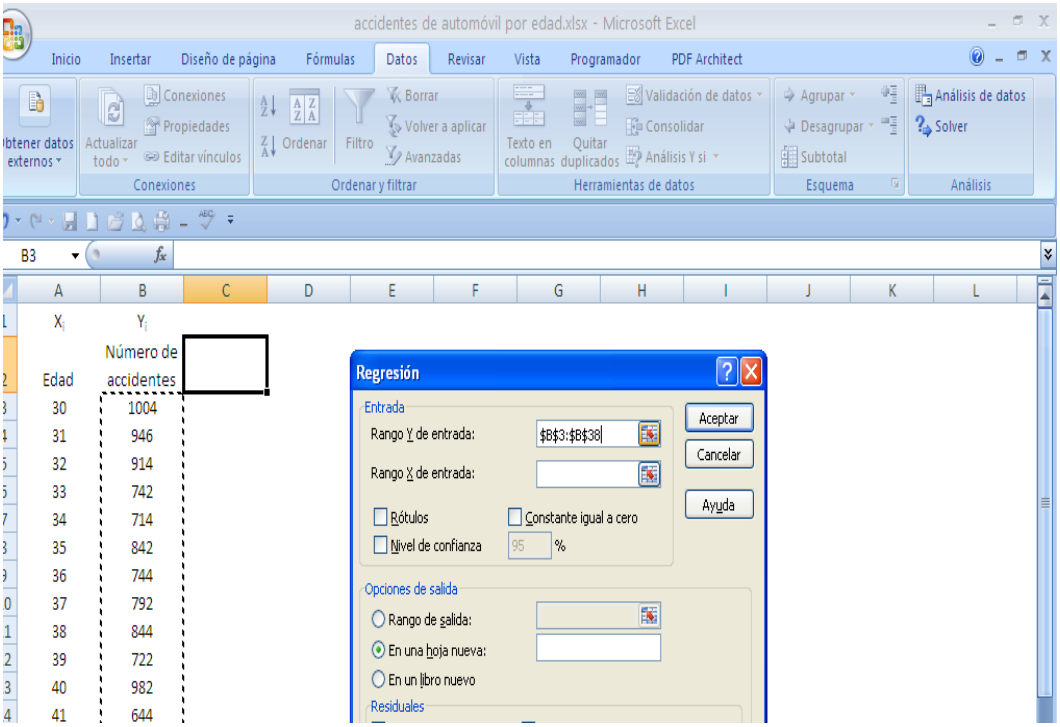
Al igual que otras técnicas de análisis, Microsoft Excel (2013) permite realizar regresión lineal simple en el módulo de *análisis de datos*. A continuación, se muestra el uso de esta herramienta con los datos del ejemplo de los accidentes registrados por edad del conductor.

Capturada la información en Excel, ir al menú de Datos, y seleccionar la opción Análisis de datos, previamente cargada. Se desplegará una ventana de diálogo de funciones para análisis, seleccionar Regresión.



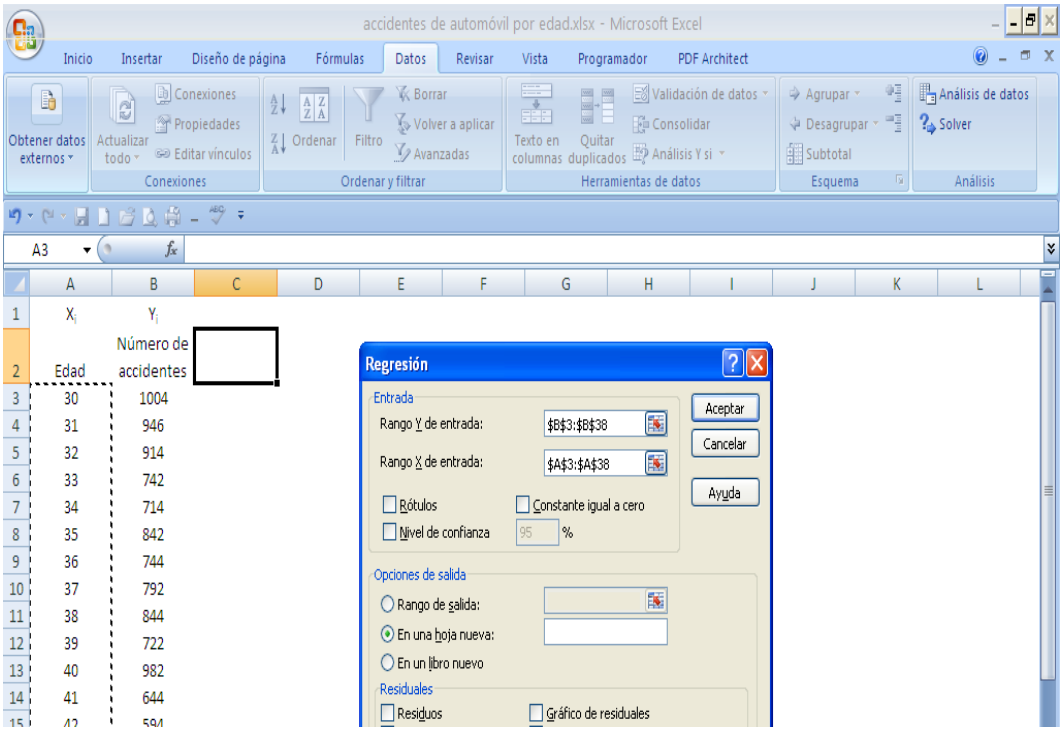
Fuente: Microsoft Excel (2013).

Se despliega una nueva ventana de diálogo, donde se ingresa la información y se determina la salida que se desea obtener. En el rango Y de entrada, seleccionar los datos de la variable dependiente, es decir, el número de accidentes.



Fuente: Microsoft Excel (2013).

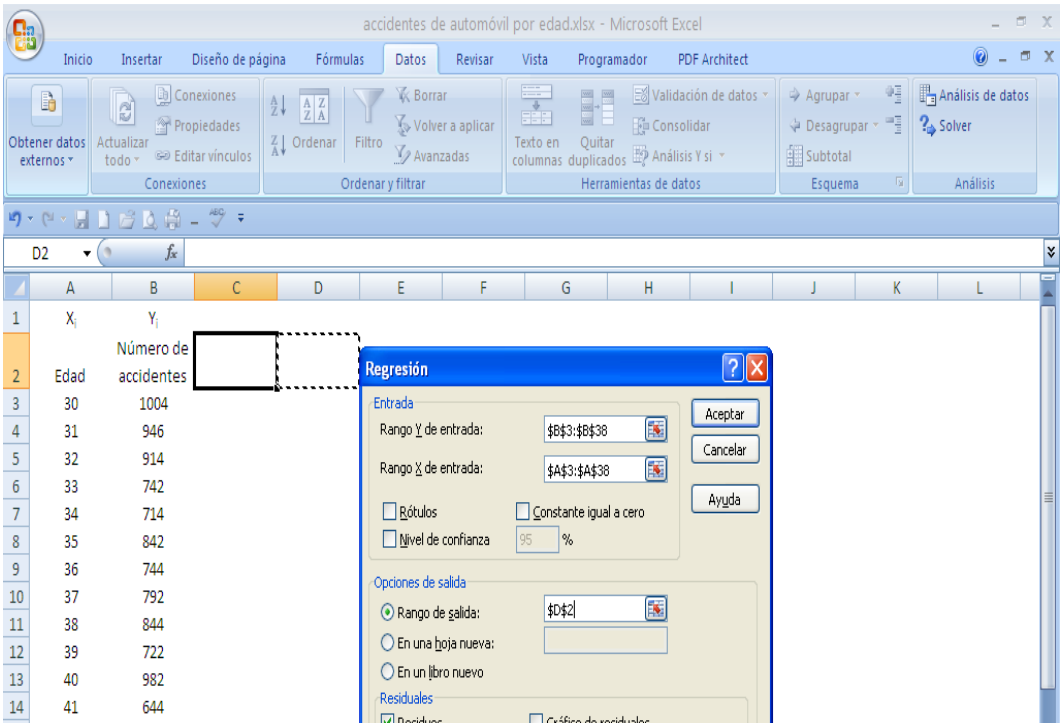
En el rango X de entrada, seleccionar los datos de la variable independiente, es decir, la edad.



Fuente: Microsoft Excel (2013).

En opciones de Salida, indicar el rango de salida; y en la sección de Residuales, la opción de Residuos. Finalmente, dar en Aceptar.

Es importante señalar que el nivel de significancia no se modifica. Excel toma por defecto el 95% de confianza valor, medida base para determinar si nuestro modelo y los parámetros de la ecuación lineal son o no significativos.



Fuente: Microsoft Excel (2013).

El cuadro-resumen que se proporciona es el siguiente. Algunas de las medidas señaladas con azul fueron calculadas previamente.

RESUMEN

Estadísticas dele regresión		
Coeficiente de correlación múltiple	0.963299334	**r**
Coeficiente de determinación R^2	0.927945607	**R^2**
R^2 ajustado	0.925826361	
Error típico	75.11890911	**S**
Observaciones	36	**n**

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	2470810.972	2470810.972	437.8657505	5.35232E-21
Residuos	34	191856.9172	5642.850506		
Total	35	2662667.889			

	Coeficientes	*Error Típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	1692.948091	58.59935094	28.89021915	1.64442E-25	1573.859882	1812.036299
Variable X1	-25.21879022	1.20518512	-20.92524195	5.35232E-21	-27.66802105	-22.76955939

Fuente: elaboración propia con empleo de Microsoft Excel (2013).

Los resultados señalados con morado indican la significancia del modelo y de cada uno de los parámetros. El primero (valor crítico de F) señala que el modelo lineal es adecuado para la información que se analiza, pues es significativo por ser menor a 0.05. En el caso de los parámetros, dado que las probabilidades son menores a 0.05, se rechaza la hipótesis nula de que los parámetros no son significativos y pueden emplearse sin inconveniente en la ecuación.

Otra manera de calcular los parámetros β_0 y β_1 es con las funciones

intersección.eje ()
pendiente()

El empleo de estas funciones se ilustrará en la siguiente unidad.

RESUMEN

Se expusieron las bases para realizar un análisis de regresión lineal simple con la información de dos variables observadas. En primer lugar, se mostró la ecuación empleada en el modelo de regresión lineal simple partiendo de un repaso de la ecuación general de la recta, y siguiendo con la metodología de mínimos cuadrados para estimar la recta que garantiza el menor error de estimación.

Calculados los parámetros del modelo, se planteó con un ejemplo la interpretación de la pendiente y se enunciaron los supuestos que debe cumplir el modelo (es habitual no comprobar esto en la práctica, por lo cual se sugiere profundizar en el análisis de los residuos).

Después se revisó la forma de realizar inferencia sobre la pendiente, y el cálculo de los coeficientes de correlación y determinación, los cuales indican, respectivamente, el grado de asociación entre las variables y la variabilidad explicada por el modelo de regresión lineal simple.

La unidad finaliza con un ejemplo de cómo ajustar un modelo de regresión lineal simple con el módulo de análisis de datos de Microsoft Excel (2013).

BIBLIOGRAFÍA

SUGERIDA

Autor	Capítulo	Páginas
Anderson, S.	14	560-641
Levin, R.	12	509-564
Lind, D.	13	461-511

UNIDAD 7

Análisis de series de tiempo

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

OBJETIVO PARTICULAR

Al terminar la unidad, el alumno conocerá los métodos para el análisis de series de tiempo, así como su aplicación e interpretación.

TEMARIO DETALLADO

(8 horas)

7. Análisis de series de tiempo

- 7.1. Los cuatro componentes de una serie de tiempo
- 7.2. Análisis gráfico de la tendencia
- 7.3. Tendencia secular
- 7.4. Variaciones estacionales
- 7.5. Variaciones cíclicas
- 7.6. Fluctuaciones irregulares
- 7.7. Modelos autorregresivos de promedios móviles

INTRODUCCIÓN

A lo largo del curso, se ha insistido en que la estadística inferencial contribuye a la toma de decisiones que, frecuentemente, deben realizarse con información recabada en el tiempo. Por ejemplo, para un inversionista, el conocimiento de los estados de resultados de una empresa durante los últimos cinco años le ayudaría a decidir si invierte en acciones de esa compañía. O la disposición de dinero en los cajeros automáticos permitiría determinar la cantidad de efectivo que la institución bancaria debe abastecer cada semana para garantizar el servicio de sus cuentahabientes. O el historial reciente de pagos de una persona facilitaría a una micro financiera dedicada a dar créditos de autos a determinar si el individuo es sujeto de crédito.

Los ejemplos anteriores ilustran la aplicación del análisis de series de tiempo. En esta unidad, se expondrá de manera básica el empleo de esta técnica (es labor del estudiante profundizar en otras fuentes). En primer lugar, se define qué es una serie de tiempo y se exponen los componentes que suelen integrarla. Después, se muestra cómo realizar un análisis exploratorio con el apoyo de una gráfica que permita visualizar la tendencia de la serie. El siguiente punto describe algunas metodologías para trabajar la tendencia de una serie de tiempo a partir del manejo de variaciones estacionales, cíclicas y fluctuaciones irregulares. Por último, se abordan de manera breve las series estacionales y los modelos auto regresivos y de medias móviles.

7.1. Los cuatro componentes de una serie de tiempo

Una serie de tiempo es el registro de una variable a lo largo del tiempo realizado con una periodicidad constante, por ejemplo, de forma diaria, semanal, mensual o anual. La observación tomada en el tiempo *t* de una variable se denotará como Y_t.

Las series de tiempo son aplicables por lo regular en todas las áreas de conocimiento: en el índice nacional de precios al consumidor (INPC), tasa de desempleo, cotización diaria del dólar norteamericano, evolución de los niveles de colesterol de un paciente sometido a un estudio clínico en el que se estudia el efecto de un medicamento, o las calificaciones de un alumno que periódicamente es sometido a evaluaciones.

De acuerdo con la forma como se registra su información, las series se dividen en discretas o continuas. Una serie de tiempo es discreta si las observaciones son realizadas en momentos específicos, normalmente con una misma periodicidad (por ejemplo, el número anual de suscriptores a una publicación). Y es continua si las observaciones se registran de forma continua en el tiempo (como el ritmo cardiaco de un paciente durante un examen médico).

Para facilitar el estudio de las series de tiempo, se dividen en cuatro partes:

a) Componente de tendencia (T)
b) Componente estacional (E)
c) Componente cíclico (C)
d) Componente de fluctuaciones irregulares (I)

Consideremos que no siempre se encuentran presentes los cuatro componentes en una serie de tiempo. En las siguientes secciones, se explicarán cada uno de estos componentes y su manejo.

Hay dos enfoques para asociar la serie de tiempo con sus componentes: aditivo y multiplicativo. En el primero, la serie de tiempo se considera que es resultado de la suma de sus componentes. De esta manera, la serie de tiempo Y_t queda expresada así:

$$Y_t = T_t + E_t + C_t + I_t$$

- Donde:

Y_t = valor de la serie al tiempo *t*
T_t = componente de tendencia al tiempo *t*
E_t = componente estacional al tiempo *t*
C_t = componente de cíclico al tiempo *t*
I_t = componente irregular o aleatorio al tiempo *t*

Y en el enfoque multiplicativo, la serie de tiempo se considera que es resultado de ajustar la tendencia con factores asociados a los otros componentes, por lo que la serie de tiempo Y_t queda expresada así:

$$Y_t = T_t * E_t * C_t * I_t$$

- Donde:

Y_t = valor de la serie al tiempo *t*
T_t = componente de tendencia al tiempo *t*
E_t = factor estacional al tiempo *t*
C_t = factor cíclico al tiempo *t*
I_t = factor irregular o aleatorio al tiempo *t*

7.2. Análisis gráfico de la tendencia

El primer paso para analizar una serie de tiempo es realizar, a modo de análisis exploratorio, una gráfica de líneas, donde en el eje X se ubicará el tiempo y en el eje Y el valor de la serie a lo largo del periodo. El análisis gráfico permitirá visualizar los componentes de la serie (por lo regular, la tendencia es el componente más evidente).

Una serie de tiempo muestra una tendencia si existe un crecimiento o disminución durante el periodo que se está analizando. Si la gráfica de la serie muestra un crecimiento continuo a lo largo del tiempo, se dice que la serie tiene una tendencia positiva (véase figura 1).

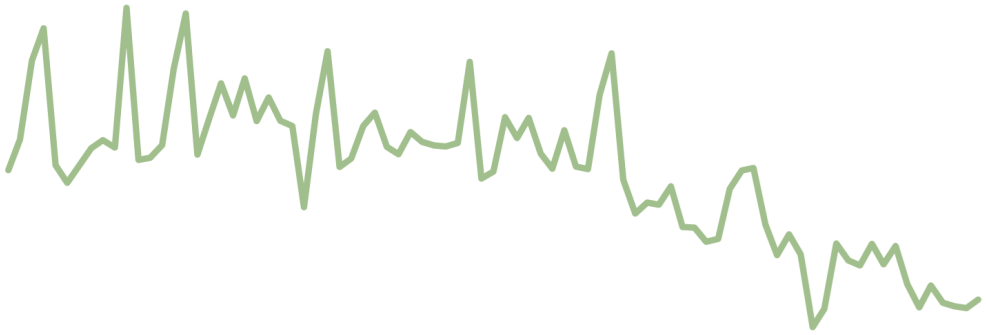
Figura 1. Serie de tiempo con tendencia positiva

Fuente: elaboración propia.

La figura anterior muestra una serie cuyo valor en general se incrementa a medida que va transcurriendo el tiempo.

Si la gráfica expresa un decrecimiento continuo a lo largo del tiempo, se dice que la serie presenta una tendencia negativa (véase figura 2).

Figura 2. Serie de tiempo con tendencia negativa



Fuente: elaboración propia.

La figura anterior muestra una serie cuyo valor, en general, decrece conforme transcurre el tiempo.

Una serie sin tendencia presentará variaciones alrededor de un solo valor a lo largo del tiempo, similar a lo que la presenta la figura 3.

Figura 3. Serie de tiempo sin tendencia

Fuente: elaboración propia.

En el análisis de series de tiempo, la realización de una gráfica es un paso casi forzado, en tanto permite conocer de forma visual su comportamiento y determinar el tratamiento que se dará a la serie. En la siguiente sección, se explicará cómo trabajar con la tendencia.

7.3. Tendencia secular

En el apartado anterior, se mencionó que el análisis de series de tiempo comienza con una exploración gráfica en donde se identifican los componentes más notables. Ahora, en este subtema, se explicará el componente de tendencia, que normalmente destaca más en una serie de tiempo; y para estimarla se aplicarán los métodos de regresión lineal y de promedios móviles.

La tendencia de una serie es la trayectoria o dirección que toma esa tendencia conforme avanza el tiempo. La importancia de este componente radica en que permite estimar el valor de una serie en un momento futuro. Por ejemplo, supóngase que el área de finanzas de cierta organización dedicada a realizar estudios de mercado se encuentra evaluando el presupuesto del siguiente año destinado a proporcionar un apoyo económico a los encuestadores asignados a la ciudad para traslado. Un análisis del precio del transporte público durante los últimos veinte años mostraría la manera como se ha ido incrementando, lo que permitiría establecer una estimación del precio en que se encontraría el servicio para el siguiente año.

A fin de estimar la tendencia, se acostumbra utilizar el modelo de regresión lineal simple o los promedios móviles. A continuación, se muestra en un ejemplo la aplicación de estos métodos.

Estimación de la tendencia con el modelo de regresión lineal simple

Con el método de regresión lineal simple, se estima una tendencia lineal al considerar que la variable dependiente es la serie y la independiente el tiempo. A continuación, se plantea un ejemplo.

Desde enero de 2013, la fábrica ABC requiere, para la producción de cierta tinta, un insumo químico, cuyo precio varía cada mes. Con la intención de diseñar un plan de adquisiciones, el área de finanzas desea estimar cuál será el precio al final del 2014, con la información de enero de 2013 a agosto de 2014.

Se muestra a continuación la información con la que cuenta el área de finanzas.

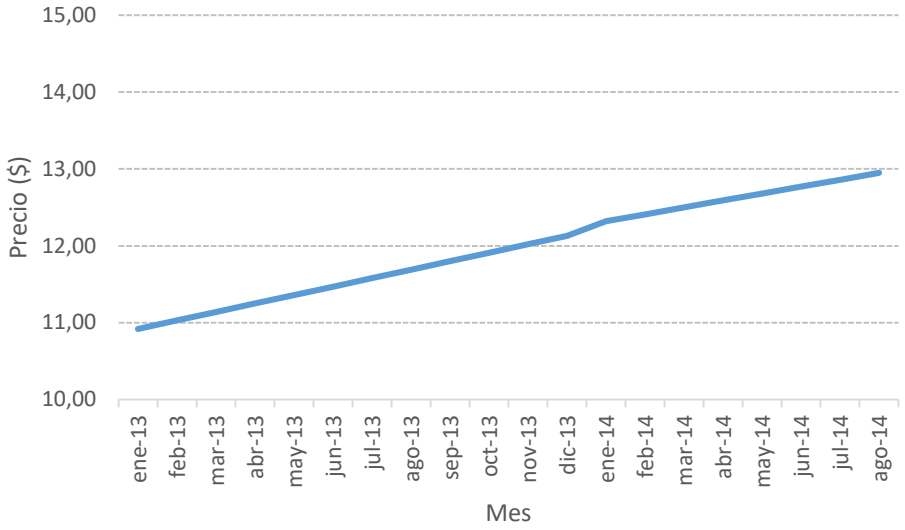
Precio promedio del insumo durante enero de 2013 a agosto de 2014

Mes	Precio	Mes	Precio	Mes	Precio	Mes	Precio
ene-13	10.92	jun-13	11.47	nov-13	12.02	abr-14	12.59
feb-13	11.03	jul-13	11.58	dic-13	12.13	may-14	12.68
mar-13	11.14	ago-13	11.69	ene-14	12.32	jun-14	12.77
abr-13	11.25	sep-13	11.80	feb-14	12.41	jul-14	12.86
may-13	11.36	oct-13	11.91	mar-14	12.50	ago-14	12.95

Precio por unidad de medida en pesos.

Se muestra en la siguiente gráfica el comportamiento del precio del insumo durante el periodo de análisis.

Precio del insumo de enero de 2013 a agosto de 2014



La gráfica muestra que, al comienzo del periodo de análisis, el precio de la unidad del insumo era casi de 11 pesos, y al finalizar se encuentra cerca de los 13 pesos. Se observa que, conforme han transcurrido los meses, el precio se asciende: la serie muestra una tendencia creciente. Para estimar la tendencia lineal que muestra la serie, se recurrirá al método de mínimos cuadrados (en este caso, la variable dependiente será el precio del insumo y la independiente el mes).

Antes de aplicar el método de mínimos cuadrados, se deberá realizar una adecuación a la variable independiente, que consiste en asignarle un valor numérico a cada mes. En este ejemplo, como la producción de la tinta comenzó a partir de enero de 2013, a esa observación se le asigna el valor 1, al siguiente mes el valor 2, y así sucesivamente hasta el valor 20, como se advierte en la tabla siguiente.

X	Mes	Precio (Y)	X	Mes	Precio (Y)
1	ene-13	10.92	11	nov-13	12.02
2	feb-13	11.03	12	dic-13	12.13
3	mar-13	11.14	13	ene-14	12.32
4	abr-13	11.25	14	feb-14	12.41
5	may-13	11.36	15	mar-14	12.50
6	jun-13	11.47	16	abr-14	12.59
7	jul-13	11.58	17	may-14	12.68
8	ago-13	11.69	18	jun-14	12.77
9	sep-13	11.80	19	jul-14	12.86
10	oct-13	11.91	20	ago-14	12.95

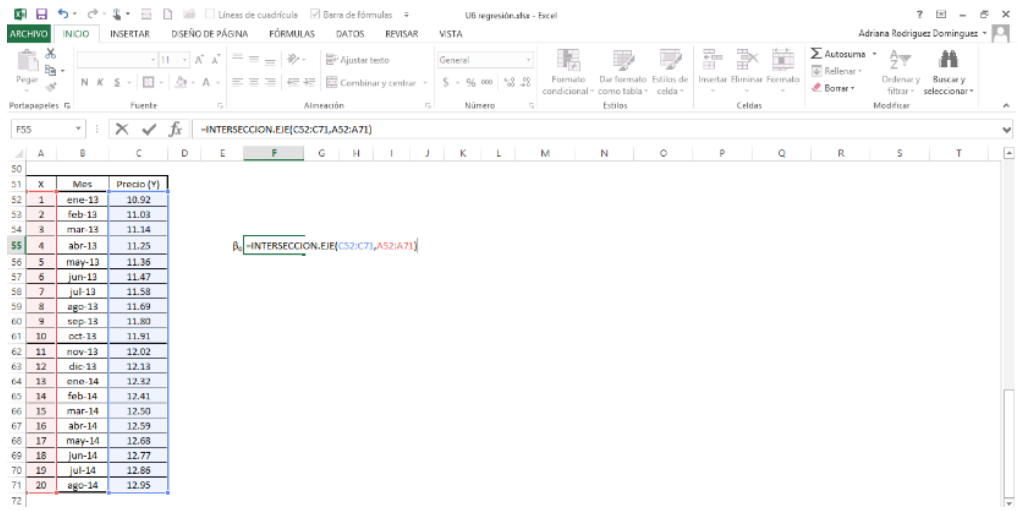
De esta manera, en el modelo se utilizarán las variables precio (Y) y X.

En la unidad anterior, se estudió cómo correr un modelo de regresión lineal simple en el módulo de análisis de datos en MS-Excel, a continuación, se utilizarán las funciones

intersección.eje()
pendiente()

para obtener los estimadores de los parámetros β_0 y β_1, respectivamente.

Para calcular β_0, en la función *intersección.eje()* se ingresan los valores de *Y*, se pone una coma y se procede a ingresar los valores de *X.*

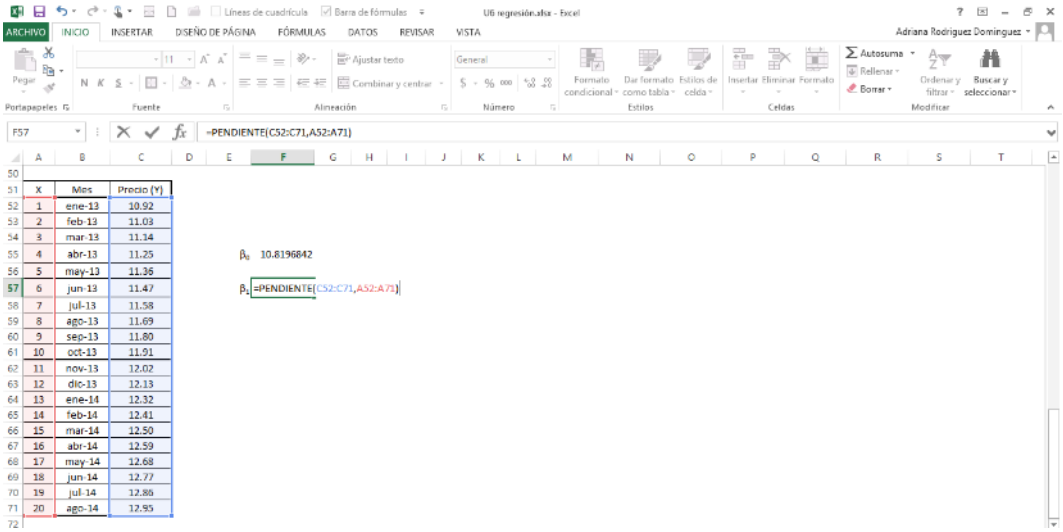


X	Mes	Precio (Y)
1	ene-13	10.92
2	feb-13	11.03
3	mar-13	11.14
4	abr-13	11.25
5	may-13	11.36
6	jun-13	11.47
7	jul-13	11.58
8	ago-13	11.69
9	sep-13	11.80
10	oct-13	11.91
11	nov-13	12.02
12	dic-13	12.13
13	ene-14	12.32
14	feb-14	12.41
15	mar-14	12.50
16	abr-14	12.59
17	may-14	12.68
18	jun-14	12.77
19	jul-14	12.86
20	ago-14	12.95

Fuente: Microsoft Excel (2013).

Al dar *enter*, se despliega el resultado (10.82).

Para calcular β_1, en la función *pendiente ()* se ingresan también los valores de la variable *Y* y *X* (en ese orden) separados por una coma.



X	Mes	Precio (Y)
1	ene-13	10.92
2	feb-13	11.03
3	mar-13	11.14
4	abr-13	11.25
5	may-13	11.36
6	jun-13	11.47
7	jul-13	11.58
8	ago-13	11.69
9	sep-13	11.80
10	oct-13	11.91
11	nov-13	12.02
12	dic-13	12.13
13	ene-14	12.32
14	feb-14	12.41
15	mar-14	12.50
16	abr-14	12.59
17	may-14	12.68
18	jun-14	12.77
19	jul-14	12.86
20	ago-14	12.95

Fuente: Microsoft Excel (2013).

Al dar *enter*, se despliega el resultado (0.11).

Entonces, la ecuación de mínimos cuadrados es

$$\text{Precio} = 10.82 + 0.11x$$

El modelo indica que, antes de comenzar a producir la tinta (en X = 0), el precio del insumo se encontraba en $10.82, y desde ese momento, por cada mes que transcurre, el precio del insumo se eleva 11 centavos. Luego, esta ecuación es la tendencia de la serie.

Determinada la tendencia, se puede estimar el precio del insumo para los meses de septiembre a diciembre del año actual sustituyendo en la ecuación el número que corresponde al mes (21, 22, 23 o 24). De esta manera, se espera que en diciembre el precio del insumo se encuentre en 10.82 + (0.11)·(24) = 13.46.

Para calcular los pronósticos, añadimos el número de periodos que se van a pronosticar. Si se desea conocer el precio de la gasolina de los meses 21, 22 y 23 (septiembre, octubre, noviembre), aplicamos la fórmula obtenida de la regresión lineal para dichos meses.

Como una observación final a este apartado, es importante definir si, de acuerdo con el contexto de la serie a analizar, es necesario identificar un punto donde la variable independiente (X) tome el valor de 0.

Estimación de la tendencia con el método de promedios móviles

El método de promedios móviles (PM) consiste en construir una nueva serie con los promedios de los datos establecidos por el orden.

El orden de un promedio móvil se refiere al número de datos consecutivos a promediar. Por ejemplo, en un promedio móvil de orden dos (PM_2), se promedia cada conjunto de dos datos consecutivos; en uno de orden tres (PM_3), cada conjunto de tres datos consecutivos, y así sucesivamente.

Un promedio móvil de orden n (PM_n) se obtiene así:

$$PM_n = \frac{\textit{suma de los valores de los n datos más recientes}}{n}$$

Supóngase que un profesor de Estadística aplica evaluaciones mensuales a sus alumnos. Las calificaciones de los cinco exámenes realizados por un estudiante de la clase son los siguientes:

Mes	Calificación
1	7
2	8
3	7
4	9
5	6

El promedio móvil de orden dos (PM_2) se obtiene de la siguiente manera:

1. Se promedian las dos primeras calificaciones (7 y 8) y se coloca el resultado en el segundo valor de la nueva serie (PM_2):

Mes	Calificación	PM2	
1	7		
2	8	7.5	$\frac{7+8}{2}$
3	7		
4	9		
5	6		

2. El siguiente valor de la nueva serie (PM_2) se obtiene de promediar las calificaciones de los meses 2 y 3 (8 y 7):

Mes	Calificación	PM2	
1	7		
2	8	7.5	$\frac{8+7}{2}$
3	7	7.5	
4	9		
5	6		

3. Seguir con el procedimiento hasta realizar el promedio de las últimas calificaciones (9 y 6):

Mes	Calificación	PM2	
1	7		
2	8	7.5	
3	7	7.5	
4	9	8	
5	6		$\frac{7+9}{2}$

Mes	Calificación	PM2	
1	7		
2	8	7.5	
3	7	7.5	
4	9	8	
5	6	7.5	$\frac{9+6}{2}$

El promedio móvil de orden tres (PM_3) se obtiene de la siguiente manera.

1. Se promedian las primeras tres calificaciones (7, 8 y 7) y el resultado se coloca en la nueva serie PM_3, centrado en la segunda posición:

Mes	Calificación	PM3	
1	7		
2	8	7.3	$\frac{7+8+7}{3}$
3	7		
4	9		
5	6		

2. El siguiente valor de la nueva serie PM_3 se obtiene de promediar las calificaciones de los meses 2, 3 y 4 (8, 7 y 9):

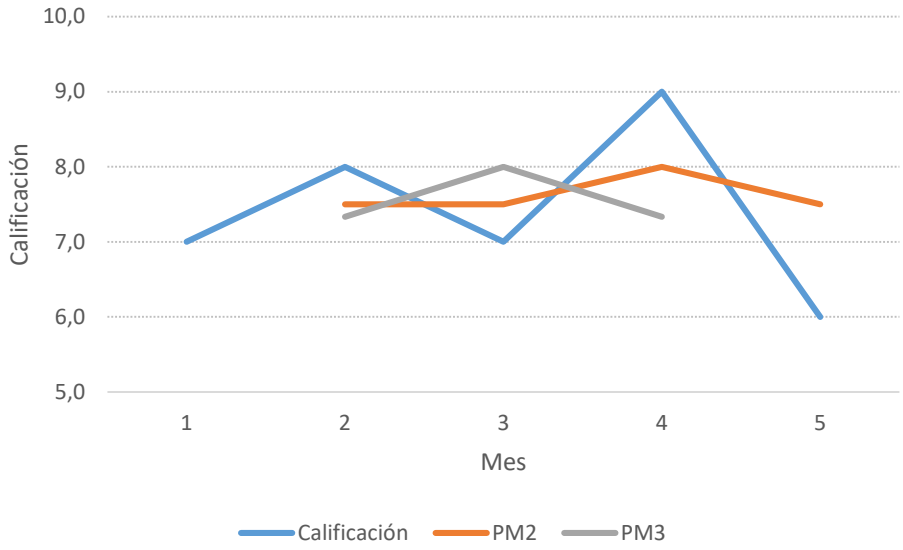
Mes	Calificación	PM3	
1	7		
2	8	7.3	
3	7	8.0	$\frac{8+7+9}{3}$
4	9		
5	6		

3. Finalmente, se calcula el promedio de los últimos tres valores (7, 9 y 6) y se registra el resultado en la nueva serie:

Mes	Calificación	PM3	
1	7		
2	8	7.3	
3	7	8.0	$\frac{7+9+6}{3}$
4	9	7.3	
5	6		

Como es imposible seguir promediando tres valores, la nueva serie PM_3 solamente tendrá tres elementos. Es importante mencionar que, conforme aumenta el orden, la nueva serie va teniendo menos valores respecto a la serie original.

La siguiente gráfica muestra el comportamiento de las calificaciones del estudiante y los promedios móviles de orden 2 y 3.



La serie de color azul de la gráfica representa el comportamiento del estudiante en las cinco evaluaciones realizadas en el curso; la serie de color rojo es el promedio móvil de orden dos, y la serie de color gris el promedio móvil de orden tres. Los promedios móviles son un *suavizamiento* de la serie original y muestran la tendencia de la serie. Para este ejemplo, el promedio móvil de orden dos explica mejor la tendencia de las calificaciones del estudiante, y refleja que sus calificaciones se encuentran alrededor de 7.5.

Supóngase ahora que restan dos evaluaciones al curso, ¿qué calificaciones se esperan de este estudiante? Para realizar el pronóstico, se utilizará el promedio móvil de orden dos, que para este ejemplo describe mejor la tendencia, y se procederá de la siguiente manera.

1. Asumir que el último valor del promedio móvil se observará en el siguiente mes:

Mes	Calificación	PM_2
1	7.0	
2	8.0	7.5
3	7.0	7.5
4	9.0	8.0
5	6.0	7.5
6	7.5	

2. Promediar las calificaciones de los meses 5 y 6 (6 y 7.5), y colocar el resultado en la posición 6 del promedio móvil:

Mes	Calificación	PM_2
1	7.0	
2	8.0	7.5
3	7.0	7.5
4	9.0	8.0
5	6.0	7.5
6	7.5	6.8

3. Repetir el procedimiento descrito en los puntos anteriores para estimar la última calificación:

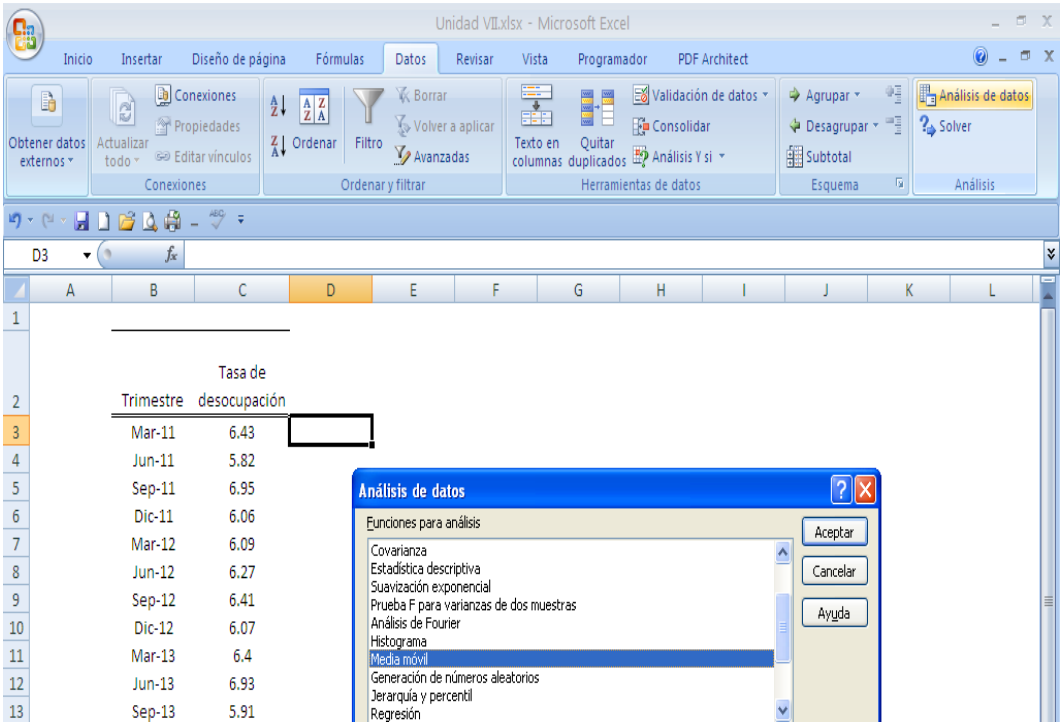
Mes	Calificación	PM_2
1	7.0	
2	8.0	7.5
3	7.0	7.5
4	9.0	8.0
5	6.0	7.5
6	7.5	6.8
7	6.8	7.2

De esta manera, de acuerdo con la tendencia mostrada por el promedio móvil de orden dos, se espera que el estudiante obtenga calificaciones de 6.8 y 7.2 en las evaluaciones faltantes.

Obtención de un promedio móvil con MS-Excel

MS-Excel permite obtener un promedio móvil al utilizar el módulo de análisis de datos, para hacerlo se procede así.

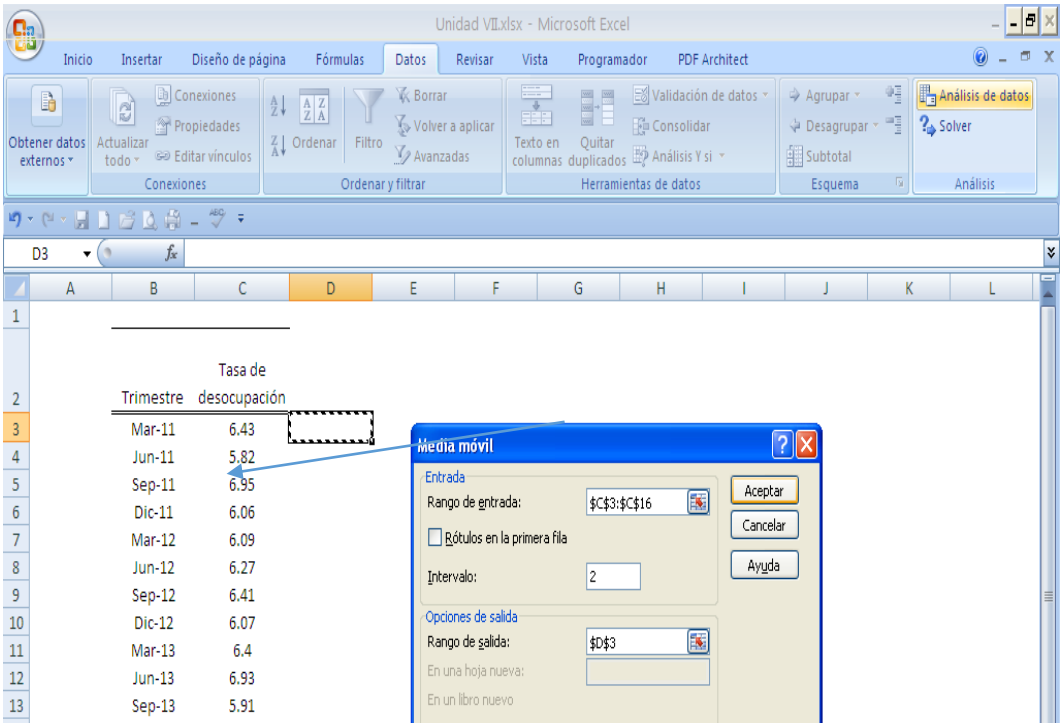
1. Acceder al menú Datos, seleccionar Análisis de datos y Media Móvil.



Fuente: Microsoft Excel (2013).

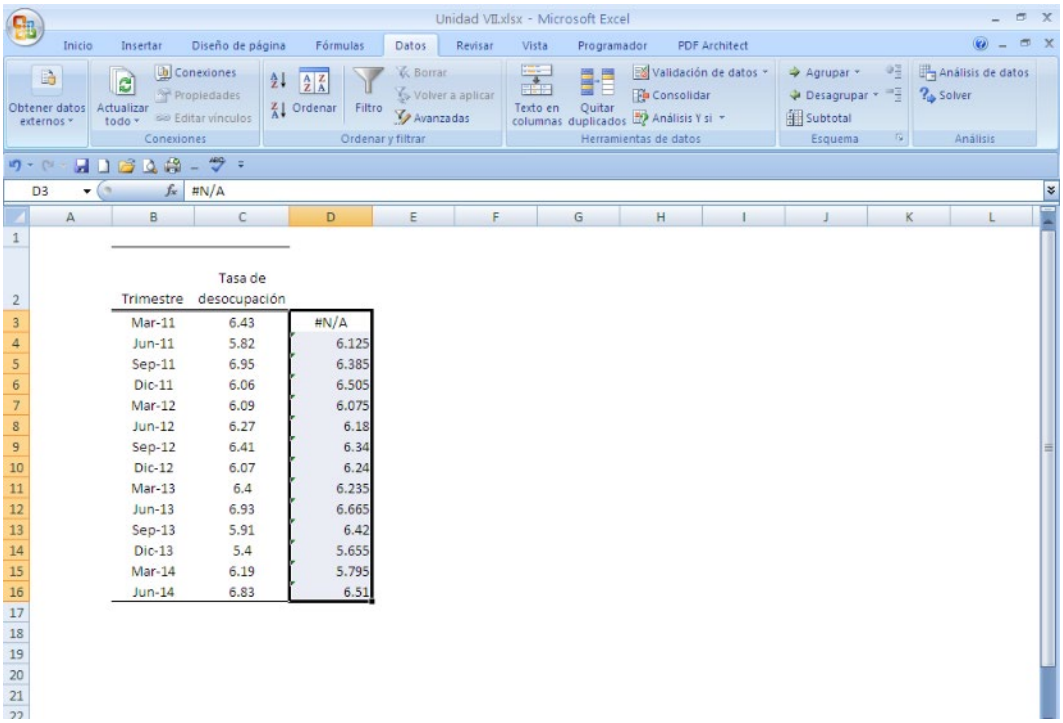
Se desplegará una ventana de diálogo que solicitará información de entrada y de salida.

En la sección Rango de entrada, seleccionar los datos de la variable; en Intervalo, indicar el rango deseado; y seleccionar el sitio en la hoja de cálculo en donde quiere que se despliegue el resultado en Rango de salida. Dar Aceptar.



Fuente: Microsoft Excel (2013).

Aparecerá una columna con los datos del promedio móvil.



Trimestre	Tasa de desocupación	
Mar-11	6.43	#N/A
Jun-11	5.82	6.125
Sep-11	6.95	6.385
Dic-11	6.06	6.505
Mar-12	6.09	6.075
Jun-12	6.27	6.18
Sep-12	6.41	6.34
Dic-12	6.07	6.24
Mar-13	6.4	6.235
Jun-13	6.93	6.665
Sep-13	5.91	6.42
Dic-13	5.4	5.655
Mar-14	6.19	5.795
Jun-14	6.83	6.51

Fuente: Microsoft Excel (2013).

7.4. Variaciones estacionales

En esta sección, se expondrá otro componente de una serie de tiempo: la estacionalidad. Una serie de tiempo tiene un comportamiento estacional si de forma periódica registra cambios a lo largo de un año. Por ejemplo, las ventas de una papelería muestran un comportamiento estacional caracterizado por un incremento durante los meses de julio y agosto, previo al comienzo del ciclo escolar del nivel básico. O la venta de pescados y mariscos crece un mes previo a las festividades de Semana Santa.

En la práctica, la manera de trabajar el componente de estacionalidad es calculando factores que se aplican a la tendencia.

A continuación, se analizará un ejemplo del tratamiento de este componente. Se muestra el indicador de comercio al por menor en México referente a artículos de papelería, libros, revistas y periódicos en el periodo, de enero de 2010 a diciembre de 2013.

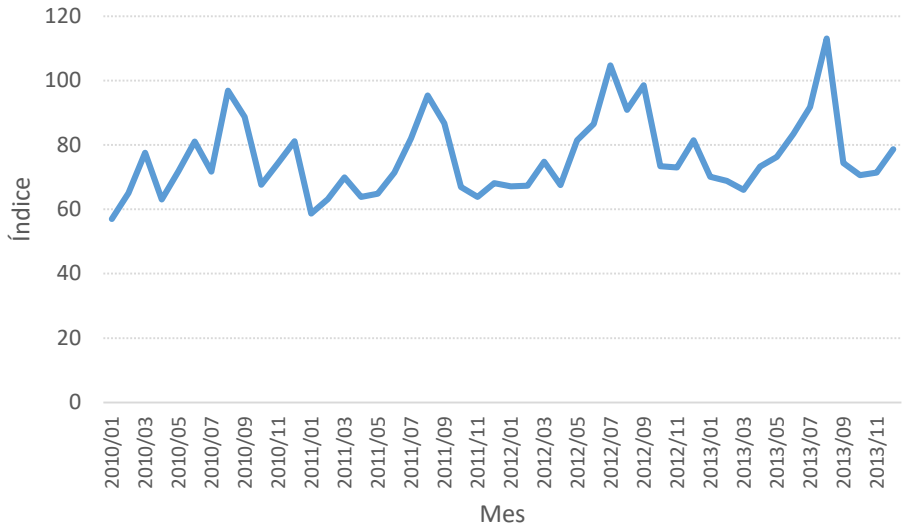
Indicador de comercio al por menor en artículos de papelería, libros, revistas y periódicos, de enero de 2010 a diciembre de 2013

Mes	2010	2011	2012	2013
Enero	57.0	58.7	67.2	70.1
Febrero	65.1	63.2	67.4	68.9
Marzo	77.6	69.9	74.8	66.0
Abril	63.1	63.9	67.5	73.3
Mayo	71.8	64.9	81.4	76.2
Junio	81.1	71.4	86.5	83.5
Julio	71.7	82.0	104.7	91.8
Agosto	96.9	95.4	90.9	113.1
Septiembre	88.7	86.7	98.5	74.4
Octubre	67.7	67.0	73.4	70.6
Noviembre	74.3	63.9	73.0	71.4
Diciembre	81.2	68.2	81.4	78.7

Base 2008.

Fuente: inegi.org.mx. fecha de consulta 7/06/2015

En la siguiente gráfica, se muestra el comportamiento de la serie.



La gráfica muestra que el índice en el periodo de análisis tiene una tendencia creciente, y alrededor de ella se aprecia que hay meses en que disminuye y meses donde se incrementa. En consecuencia, la serie cuenta con un componente de estacionalidad. Ahora bien, los factores de estacionalidad se calcularán de la siguiente manera.

A partir de la serie original, se construye un promedio móvil de orden 12, centrado de tal manera que se pierden los primeros y últimos seis meses.

Mes	Índice	PM12
ene-10	57.0	
feb-10	65.1	
mar-10	77.6	
abr-10	63.1	
may-10	71.8	
jun-10	81.1	
jul-10	71.7	74.7
ago-10	96.9	74.8
sep-10	88.7	74.7
oct-10	67.7	74.0
nov-10	74.3	74.1
dic-10	81.2	73.5
ene-11	58.7	72.7
feb-11	63.2	73.6
mar-11	69.9	73.4
abr-11	63.9	73.3
may-11	64.9	73.2
jun-11	71.4	72.3
jul-11	82.0	71.3
ago-11	95.4	72.0
sep-11	86.7	72.3
oct-11	67.0	72.7
nov-11	63.9	73.0
dic-11	68.2	74.4
ene-12	67.2	75.7
feb-12	67.4	77.6
mar-12	74.8	77.2
abr-12	67.5	78.2
may-12	81.4	78.7
jun-12	86.5	79.5
jul-12	104.7	80.6
ago-12	90.9	80.8
sep-12	98.5	80.9
oct-12	73.4	80.2
nov-12	73.0	80.7
dic-12	81.4	80.3
ene-13	70.1	80.0
feb-13	68.9	78.9
mar-13	66.0	80.8
abr-13	73.3	78.8
may-13	76.2	78.5
jun-13	83.5	78.4
jul-13	91.8	
ago-13	113.1	

sep-13	74.4	
oct-13	70.6	
nov-13	71.4	
dic-13	78.7	

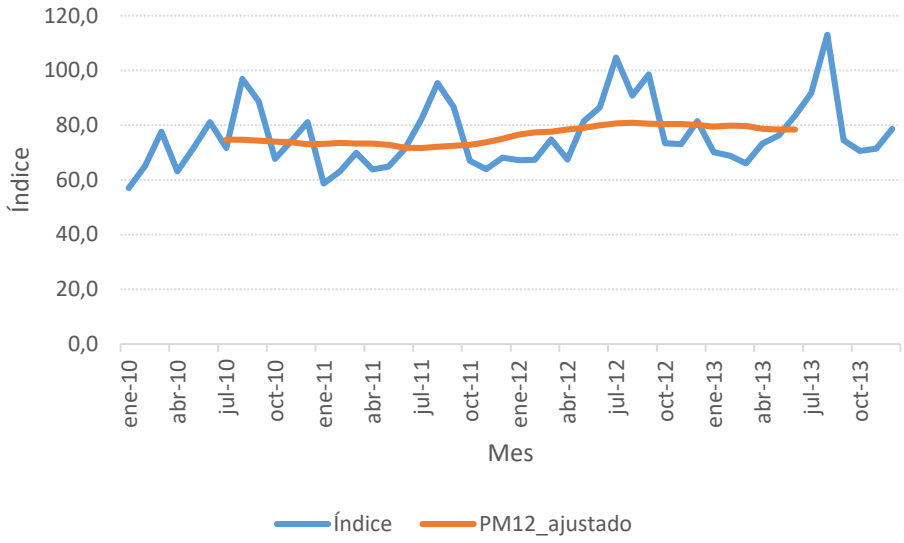
El primer punto del promedio móvil se obtiene al promediar los primeros 12 valores de la serie y se encontrará ubicado de manera que separa seis meses antes y después de él; es decir, se halla en el 15 de junio y el siguiente en el 15 de julio, para llevarlo al primero de julio se vuelve a construir un promedio móvil de orden 2.

Mes	Índice	PM12	PM12_ajustado
ene-10	57.0		
feb-10	65.1		
mar-10	77.6		
abr-10	63.1		
may-10	71.8		
jun-10	81.1		
jul-10	71.7	74.7	74.7
ago-10	96.9	74.8	74.7
sep-10	88.7	74.7	74.3
oct-10	67.7	74.0	74.0
nov-10	74.3	74.1	73.8
dic-10	81.2	73.5	73.1
ene-11	58.7	72.7	73.1
feb-11	63.2	73.6	73.5
mar-11	69.9	73.4	73.4
abr-11	63.9	73.3	73.2
may-11	64.9	73.2	72.8
jun-11	71.4	72.3	71.8
jul-11	82.0	71.3	71.6
ago-11	95.4	72.0	72.1
sep-11	86.7	72.3	72.5
oct-11	67.0	72.7	72.9
nov-11	63.9	73.0	73.7
dic-11	68.2	74.4	75.0
ene-12	67.2	75.7	76.6
feb-12	67.4	77.6	77.4
mar-12	74.8	77.2	77.7
abr-12	67.5	78.2	78.4
may-12	81.4	78.7	79.1
jun-12	86.5	79.5	80.0
jul-12	104.7	80.6	80.7
ago-12	90.9	80.8	80.9
sep-12	98.5	80.9	80.6
oct-12	73.4	80.2	80.4
nov-12	73.0	80.7	80.5
dic-12	81.4	80.3	80.1

ene-13	70.1	80.0	79.5
feb-13	68.9	78.9	79.8
mar-13	66.0	80.8	79.8
abr-13	73.3	78.8	78.7
may-13	76.2	78.5	78.5
jun-13	83.5	78.4	78.4
jul-13	91.8		
ago-13	113.1		
sep-13	74.4		
oct-13	70.6		
nov-13	71.4		
dic-13	78.7		

El primer valor de la última serie se obtuvo al promediar los primeros dos valores del promedio móvil de orden 12. El segundo valor de la nueva serie es resultado de promediar el segundo y tercero del promedio móvil de orden 12, y así sucesivamente.

La siguiente gráfica muestra un comparativo entre el comportamiento de la serie original y el promedio móvil ajustado.



En la gráfica anterior, se plantea el comportamiento de la serie y del promedio móvil, que en este caso funciona como un eje alrededor del cual varía la serie original.

El siguiente paso es calcular la variación de cada punto respecto al promedio móvil dividiendo el valor original de la serie entre el promedio móvil.

Mes	Índice	PM12	PM12_ajustado	Variación
ene-10	57.0			
feb-10	65.1			
mar-10	77.6			
abr-10	63.1			
may-10	71.8			
jun-10	81.1			
jul-10	71.7	74.7	74.7	0.96
ago-10	96.9	74.8	74.7	1.30
sep-10	88.7	74.7	74.3	1.19
oct-10	67.7	74.0	74.0	0.91
nov-10	74.3	74.1	73.8	1.01
dic-10	81.2	73.5	73.1	1.11
ene-11	58.7	72.7	73.1	0.80
feb-11	63.2	73.6	73.5	0.86
mar-11	69.9	73.4	73.4	0.95
abr-11	63.9	73.3	73.2	0.87
may-11	64.9	73.2	72.8	0.89
jun-11	71.4	72.3	71.8	0.99
jul-11	82.0	71.3	71.6	1.15
ago-11	95.4	72.0	72.1	1.32
sep-11	86.7	72.3	72.5	1.20
oct-11	67.0	72.7	72.9	0.92
nov-11	63.9	73.0	73.7	0.87
dic-11	68.2	74.4	75.0	0.91
ene-12	67.2	75.7	76.6	0.88
feb-12	67.4	77.6	77.4	0.87
mar-12	74.8	77.2	77.7	0.96
abr-12	67.5	78.2	78.4	0.86
may-12	81.4	78.7	79.1	1.03
jun-12	86.5	79.5	80.0	1.08
jul-12	104.7	80.6	80.7	1.30
ago-12	90.9	80.8	80.9	1.12
sep-12	98.5	80.9	80.6	1.22
oct-12	73.4	80.2	80.4	0.91
nov-12	73.0	80.7	80.5	0.91
dic-12	81.4	80.3	80.1	1.02
ene-13	70.1	80.0	79.5	0.88
feb-13	68.9	78.9	79.8	0.86
mar-13	66.0	80.8	79.8	0.83
abr-13	73.3	78.8	78.7	0.93
may-13	76.2	78.5	78.5	0.97
jun-13	83.5	78.4	78.4	1.06
jul-13	91.8			
ago-13	113.1			
sep-13	74.4			
oct-13	70.6			

nov-13	71.4			
dic-13	78.7			

El primer valor de la serie de variaciones se obtuvo de dividir 71.7 (valor original de la serie) entre 74.7 (promedio móvil ajustado). El valor resultante de 0.96 significa que el índice observado en julio de 2010 se encontró 4% debajo del promedio. De manera similar, se procedió con el resto de los valores.

Se llega a los factores estacionales mensuales promediando todas las variaciones obtenidas en el mismo mes.

Mes	2010	2011	2012	2013	Promedio
Enero		0.80	0.88	0.88	0.85
Febrero		0.86	0.87	0.86	0.86
Marzo		0.95	0.96	0.83	0.91
Abril		0.87	0.86	0.93	0.89
Mayo		0.89	1.03	0.97	0.96
Junio		0.99	1.08	1.06	1.05
Julio	0.96	1.15	1.30		1.13
Agosto	1.30	1.32	1.12		1.25
Septiembre	1.19	1.20	1.22		1.20
Octubre	0.91	0.92	0.91		0.91
Noviembre	1.01	0.87	0.91		0.93
Diciembre	1.11	0.91	1.02		1.01

Como el promedio móvil ajustado parte de julio de 2010 y termina en junio de 2013, en cada mes se calcularon tres variaciones, que al promediarse serán los factores estacionales.

Los factores estacionales muestran una mayor actividad en los meses de julio, agosto y septiembre, donde el índice es, respectivamente, 13%, 25% y 20% mayor al promedio. La menor actividad se registra en enero y febrero, donde los factores son 0.85 y 0.86.

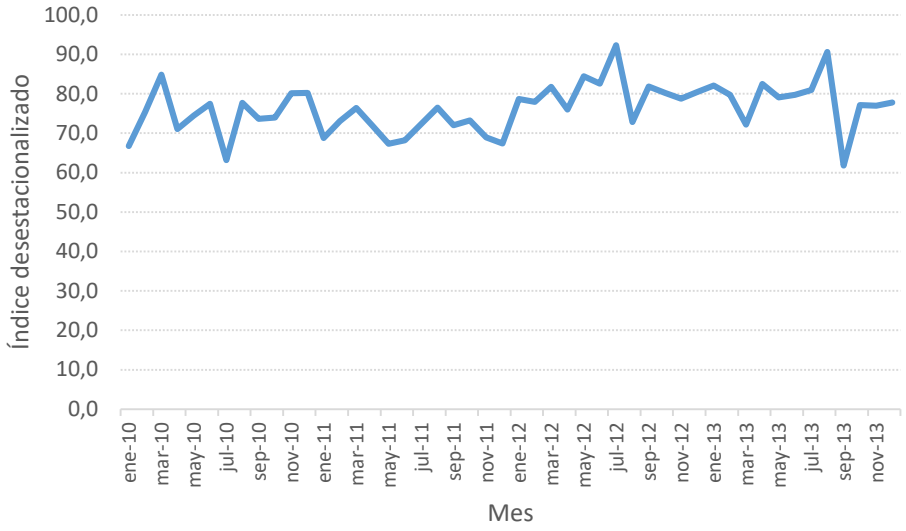
Una vez calculados los factores estacionales, sigue desestacionalizar los datos, dividiendo el valor original de la serie entre el factor que le corresponda.

Mes	Índice	Factor	Índice desestacionalizado
ene-10	57.0	0.85	66.7
feb-10	65.1	0.86	75.3
mar-10	77.6	0.91	84.8
abr-10	63.1	0.89	71.0
may-10	71.8	0.96	74.4
jun-10	81.1	1.05	77.5
jul-10	71.7	1.13	63.2
ago-10	96.9	1.25	77.7
sep-10	88.7	1.20	73.7
oct-10	67.7	0.91	73.9
nov-10	74.3	0.93	80.1
dic-10	81.2	1.01	80.2
ene-11	58.7	0.85	68.8
feb-11	63.2	0.86	73.1
mar-11	69.9	0.91	76.4
abr-11	63.9	0.89	71.9
may-11	64.9	0.96	67.3
jun-11	71.4	1.05	68.2
jul-11	82.0	1.13	72.3
ago-11	95.4	1.25	76.5
sep-11	86.7	1.20	72.0
oct-11	67.0	0.91	73.2
nov-11	63.9	0.93	68.9
dic-11	68.2	1.01	67.4
ene-12	67.2	0.85	78.6
feb-12	67.4	0.86	77.9
mar-12	74.8	0.91	81.8
abr-12	67.5	0.89	76.0
may-12	81.4	0.96	84.5
jun-12	86.5	1.05	82.6
jul-12	104.7	1.13	92.3
ago-12	90.9	1.25	72.9
sep-12	98.5	1.20	81.8
oct-12	73.4	0.91	80.2
nov-12	73.0	0.93	78.8
dic-12	81.4	1.01	80.5
ene-13	70.1	0.85	82.1
feb-13	68.9	0.86	79.7
mar-13	66.0	0.91	72.2
abr-13	73.3	0.89	82.5
may-13	76.2	0.96	79.1
jun-13	83.5	1.05	79.8
jul-13	91.8	1.13	80.9
ago-13	113.1	1.25	90.6
sep-13	74.4	1.20	61.8

oct-13	70.6	0.91	77.2
nov-13	71.4	0.93	77.0
dic-13	78.7	1.01	77.8

En la tabla anterior, los valores de la última columna son resultado de dividir el índice entre el factor.

La serie desestacionalizada queda así:



La gráfica anterior muestra los datos desestacionalizados, que no reflejan una tendencia aparente. Para confirmar lo anterior, se ajusta una regresión, la cual indica que sí existe una tendencia.

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	73.0707108	1.82817565	39.969196	8.4506E-37
Índice desestacionalizado	0.1381169	0.06427523	2.14883546	0.03706385

Entonces, la tendencia de los datos desestacionalizados es $\boldsymbol{y_t = 73.07 + 0.14t}$

La ecuación expresa que, por cada mes transcurrido, el índice desestacionalizado se incrementa en 0.14.

Supóngase que se desea realizar un pronóstico para los siguientes cinco meses, es decir, para las observaciones 49, 50, 51, 52 y 53. Primero, se sustituyen estos valores en el modelo de la tendencia:

t	73.07 + 0.14t
49	73.07 + (0.14)·(49) = 79.93
50	73.07 + (0.14)·(50) = 80.07
51	73.07 + (0.14)·(51) = 80.21
52	73.07 + (0.14)·(52) = 80.35
53	73.07 + (0.14)·(53) = 80.49

Los valores obtenidos se multiplican por el factor estacional:

t	Índice desestacionalizado	Factor estacional	Pronóstico
49	79.93	0.85	68.26
50	80.07	0.86	69.20
51	80.21	0.91	73.34
52	80.35	0.89	71.36
53	80.49	0.96	77.63

De esta manera, se alcanza el pronóstico.

7.5. Variaciones cíclicas

En la sección anterior, se trató cómo trabajar el componente estacional de una serie, el cual ofrece las variaciones que se presentan a lo largo de un año. Ahora, en este subtema se muestra el tratamiento de variaciones presentadas en periodos mayores a un año, los cuales son el componente de ciclicidad.

Un ciclo consiste en cambios ascendentes y descendentes en la serie respecto a su tendencia, con duración mayor a un año. Ejemplos de ello son los ciclos económicos caracterizados por un periodo de expansión y recesión, o el ciclo de vida de un producto.

Un componente cíclico tiene un comportamiento parecido al de la figura siguiente:

Fuente: elaboración propia.

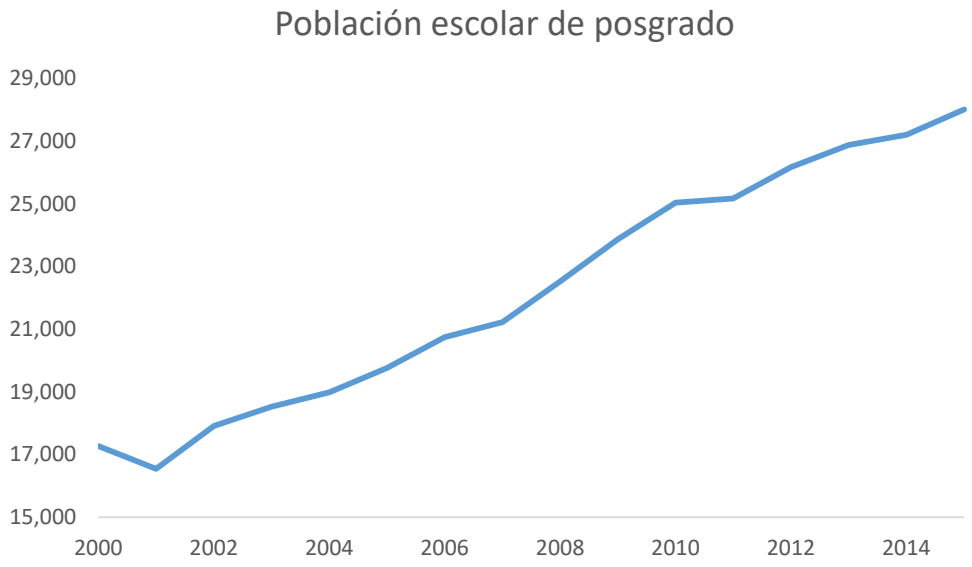
Como se muestra en la figura anterior, el ciclo se integra de dos partes: una expansiva, donde la serie aumenta de valor; y otra recesiva, donde disminuye el valor.

En cuanto al tratamiento que se dará a este componente, será bajo un enfoque aditivo. A continuación, se expone un ejemplo.

En la siguiente tabla, se muestra la población escolar de posgrado de cierta institución entre 2000 y 2015.

Año	Población escolar de posgrado
2000	17,270
2001	16,547
2002	17,910
2003	18,530
2004	18,987
2005	19,765
2006	20,747
2007	21,230
2008	22,527
2009	23,875
2010	25,036
2011	25,167
2012	26,169
2013	26,878
2014	27,210
2015	28,018

En el periodo de análisis, se advierte que la población creció de 17 270 en el año 2000 a 28 018 en 2015. Al graficar la serie, se observa el siguiente comportamiento:



La gráfica anterior manifiesta el crecimiento de la población de posgrado, caracterizado por una serie con tendencia positiva. A continuación se estimará la tendencia, con una regresión lineal simple a la serie, y se obtendrá la *y* estimada.

Consecutivo	Año	Población escolar de posgrado	Tendencia $\hat{y}$		
1	2000	17,270	16203		
2	2001	16,547	17008	15397.45	Intersección
3	2002	17,910	17813	805.197059	Pendiente
4	2003	18,530	18618		
5	2004	18,987	19423		
6	2005	19,765	20229		
7	2006	20,747	21034		
8	2007	21,230	21839		
9	2008	22,527	22644		
10	2009	23,875	23449		
11	2010	25,036	24255		
12	2011	25,167	25060		
13	2012	26,169	25865		
14	2013	26,878	26670		
15	2014	27,210	27475		
16	2015	28,018	28281		

Fuente: elaboración propia con Microsoft Excel (2013).

La tendencia de la población de posgrado se estima con la siguiente ecuación:

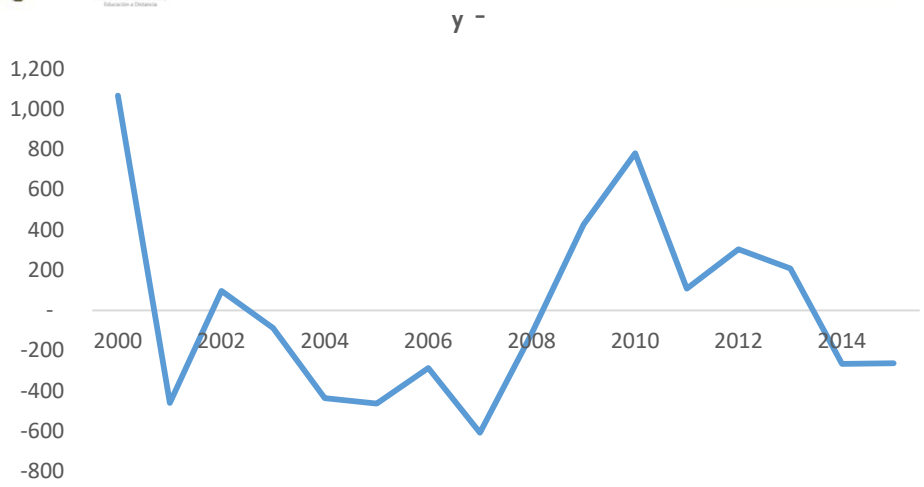
Población escolar de posgrado = 15,397 + 805 año

Enseguida, se elimina el componente de tendencia a la serie original. Para hacerlo, se resta la tendencia a los valores originales:

Consecutivo	Año	Población escolar de posgrado	Tendencia $\hat{y}$	Sin tendencia $y - \hat{y}$
1	2000	17,270	16,203	1,067
2	2001	16,547	17,008	- 461
3	2002	17,910	17,813	97
4	2003	18,530	18,618	- 88
5	2004	18,987	19,423	- 436
6	2005	19,765	20,229	- 464
7	2006	20,747	21,034	- 287
8	2007	21,230	21,839	- 609
9	2008	22,527	22,644	- 117
10	2009	23,875	23,449	426
11	2010	25,036	24,255	781
12	2011	25,167	25,060	107
13	2012	26,169	25,865	304
14	2013	26,878	26,670	208
15	2014	27,210	27,475	- 265
16	2015	28,018	28,281	- 263

El cuadro anterior presenta la serie original y la tendencia calculada con el modelo de regresión. La última columna es la serie sin tendencia, resultado de restar la tendencia de la serie original.

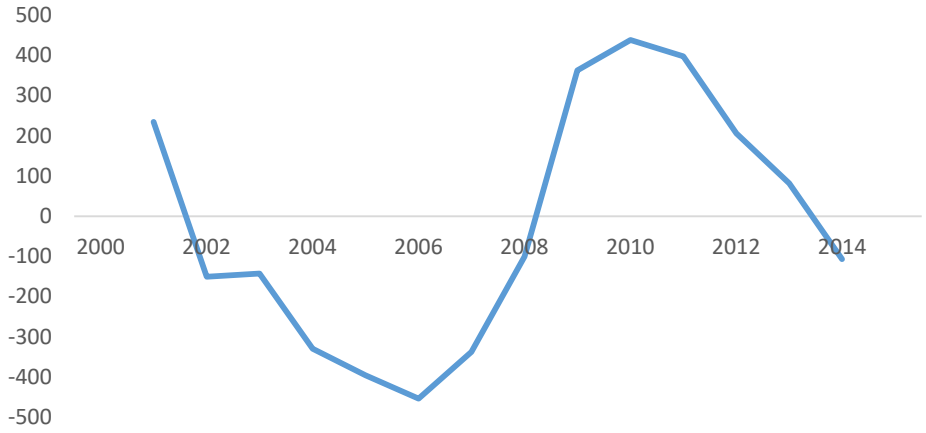
Ahora, la serie sin tendencia luce así:



La gráfica representa una serie con un comportamiento que se acerca a un ciclo. Obsérvese que aproximadamente cada tres años se cumple el ciclo, por lo que se utilizará un promedio móvil de orden 3 para obtener el componente cíclico (véase la siguiente tabla).

Año	Sin tendencia $y - \hat{y}$	Ciclo PM3
2000	1,067	
2001	- 461	234
2002	97	- 151
2003	- 88	- 143
2004	- 436	- 329
2005	- 464	- 396
2006	- 287	- 453
2007	- 609	- 338
2008	- 117	- 100
2009	426	363
2010	781	438
2011	107	398
2012	304	206
2013	208	82
2014	- 265	- 107
2015	- 263	

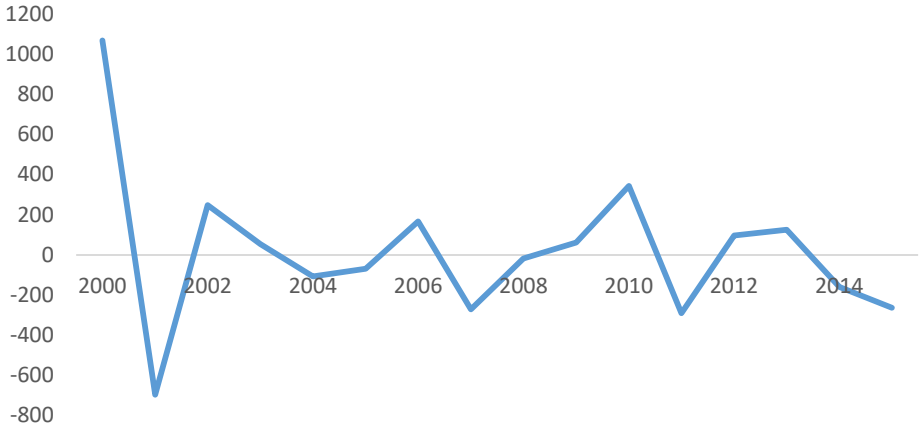
La tabla anterior expresa la nueva serie obtenida con el promedio móvil de orden 3, que al graficarse muestra el componente cíclico:



Para quitar el ciclo, se resta el componente a la serie sin tendencia:

Año	Población escolar de posgrado	Sin tendencia $y - \hat{y}$	Ciclo PM3	Aleatorio $(y - \hat{y}) - \text{PM3}$
2000	17,270	1,067		1067
2001	16,547	- 461	234	- 695
2002	17,910	97	-151	248
2003	18,530	- 88	-143	54
2004	18,987	- 436	-329	- 107
2005	19,765	- 464	-396	-68
2006	20,747	- 287	-453	166
2007	21,230	- 609	-338	- 271
2008	22,527	- 117	-100	- 17
2009	23,875	426	363	62
2010	25,036	781	438	343
2011	25,167	107	398	- 290
2012	26,169	304	206	98
2013	26,878	208	82	126
2014	27,210	- 265	-107	- 159
2015	28,018	- 263		- 263

El resultado es una serie irregular o aleatoria:



Supóngase que se necesita realizar un pronóstico de alumnos de posgrado del 2016 al 2019. Para hacerlo, se darán los siguientes pasos.

1. Pronosticar la tendencia en los periodos futuros con la ecuación lineal.

Población escolar de posgrado = 15,397 + 805 año

	Año	Población escolar de posgrado	Tendencia	Sin tendencia $y - \hat{y}$	Ciclo PM3	Aleatorio $(y - \hat{y}) - \text{PM3}$
1	2000	17,270	16,203	1,067		1067
2	2001	16,547	17,008	- 461	234	- 695
3	2002	17,910	17,813	97	-151	248
4	2003	18,530	18,618	- 88	-143	54
5	2004	18,987	19,423	- 436	-329	- 107
6	2005	19,765	20,229	- 464	-396	- 68
7	2006	20,747	21,034	- 287	-453	166
8	2007	21,230	21,839	- 609	-338	- 271
9	2008	22,527	22,644	- 117	-100	- 17
10	2009	23,875	23,449	426	363	62
11	2010	25,036	24,255	781	438	343
12	2011	25,167	25,060	107	398	- 290
13	2012	26,169	25,865	304	206	98
14	2013	26,878	26,670	208	82	126
15	2014	27,210	27,475	- 265	-107	- 159
16	2015	28,018	28,281	- 263		- 263

17	**2016**	29,086		
18	**2017**	29,891		
19	**2018**	30,696		
20	**2019**	31,501		
21	**2020**	32,307		

2. Para estimar el ciclo, se recurre al procedimiento de promedio móvil: se copia el último valor (–107) de la serie Ciclo PM3 en la columna Sin tendencia, debajo del valor –263, y en ambas columnas se replican las fórmulas ya trabajadas en cada una de ellas:

	Año	Población escolar de posgrado	Tendencia	Sin tendencia $y - \hat{y}$	Ciclo PM3	Aleatorio $(y - \hat{y}) - \text{PM3}$
1	2000	17,270	16,203	1,067		1067
2	2001	16,547	17,008	- 461	234	- 695
3	2002	17,910	17,813	97	-151	248
4	2003	18,530	18,618	- 88	-143	54
5	2004	18,987	19,423	- 436	-329	- 107
6	2005	19,765	20,229	- 464	-396	- 68
7	2006	20,747	21,034	- 287	-453	166
8	2007	21,230	21,839	- 609	-338	- 271
9	2008	22,527	22,644	- 117	-100	- 17
10	2009	23,875	23,449	426	363	62
11	2010	25,036	24,255	781	438	343
12	2011	25,167	25,060	107	398	- 290
13	2012	26,169	25,865	304	206	98
14	2013	26,878	26,670	208	82	126
15	2014	27,210	27,475	- 265	-107	- 159
16	2015	28,018	28,281	- 263	- 212	- 51
17	**2016**		29,086	**- 107**	- 194	87
18	**2017**		29,891	- 212	- 171	
19	**2018**		30,696	- 194	- 192	
20	**2019**		31,501	- 171	- 185	
21	**2020**		32,307	- 192		

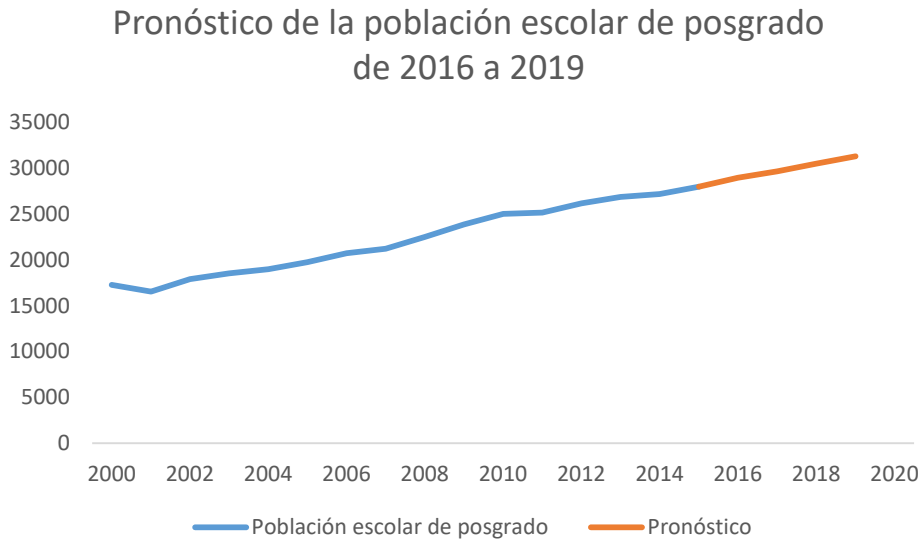
De igual manera, se replica la fórmula de la columna Aleatorio para los periodos a pronosticar:

	Año	Población escolar de posgrado	Tendencia	Sin tendencia $y - \hat{y}$	Ciclo PM3	Aleatorio $(y - \hat{y}) - \text{PM3}$
1	2000	17,270	16,203	1,067		1067
2	2001	16,547	17,008	- 461	234	- 695
3	2002	17,910	17,813	97	-151	248
4	2003	18,530	18,618	- 88	-143	54
5	2004	18,987	19,423	- 436	-329	- 107
6	2005	19,765	20,229	- 464	-396	- 68
7	2006	20,747	21,034	- 287	-453	166
8	2007	21,230	21,839	- 609	-338	- 271
9	2008	22,527	22,644	- 117	-100	- 17
10	2009	23,875	23,449	426	363	62
11	2010	25,036	24,255	781	438	343
12	2011	25,167	25,060	107	398	- 290
13	2012	26,169	25,865	304	206	98
14	2013	26,878	26,670	208	82	126
15	2014	27,210	27,475	- 265	-107	- 159
16	2015	28,018	28,281	- 263	- 212	- 51
17	**2016**		29,086	**- 107**	- 194	87
18	**2017**		29,891	- 212	- 171	- 41
19	**2018**		30,696	- 194	- 192	- 2
20	**2019**		31,501	- 171	- 185	15
21	**2020**		32,307	- 192		- 192

Se crea una nueva columna para realizar el pronóstico, sumando los valores de las columnas Tendencia, Ciclo y Aleatorio:

	Año	Población escolar de posgrado	Tendencia	Sin tendencia $y - \hat{y}$	Ciclo PM3	Aleatorio $(y - \hat{y}) - \text{PM3}$	Pronóstico
1	2000	17,270	16,203	1,067		1067	
2	2001	16,547	17,008	- 461	234	- 695	
3	2002	17,910	17,813	97	-151	248	
4	2003	18,530	18,618	- 88	-143	54	
5	2004	18,987	19,423	- 436	-329	- 107	
6	2005	19,765	20,229	- 464	-396	- 68	
7	2006	20,747	21,034	- 287	-453	166	
8	2007	21,230	21,839	- 609	-338	- 271	
9	2008	22,527	22,644	- 117	-100	- 17	
10	2009	23,875	23,449	426	363	62	
11	2010	25,036	24,255	781	438	343	
12	2011	25,167	25,060	107	398	- 290	
13	2012	26,169	25,865	304	206	98	
14	2013	26,878	26,670	208	82	126	
15	2014	27,210	27,475	- 265	-107	- 159	
16	2015	28,018	28,281	- 263	- 212	- 51	28,018
17	**2016**		29,086	**- 107**	- 194	87	28,979
18	**2017**		29,891	- 212	- 171	- 41	29,679
19	**2018**		30,696	- 194	- 192	- 2	30,503
20	**2019**		31,501	- 171	- 185	15	31,331
21	**2020**		32,307	- 192		- 192	

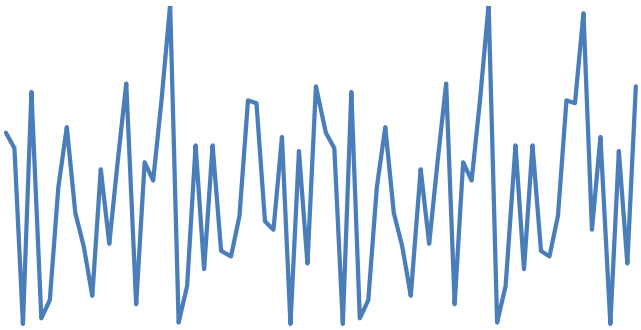
Por tanto, la población escolar aumentará de 28 018 a 31 331 alumnos entre 2016 y 2019.



7.6. Fluctuaciones irregulares

El último componente de una serie de tiempo es de fluctuaciones irregulares. Este componente se caracteriza por tener un comportamiento difícil de modelar, debido a que sus variaciones se deben a causas particulares que normalmente no son predecibles (por ejemplo, las variaciones en el tráfico a causa de un accidente o una manifestación).

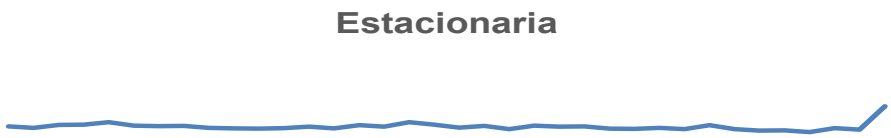
Una serie irregular se ejemplifica en la siguiente figura, donde no se aprecia un patrón:



Fuente: elaboración propia.

En el ejemplo de la sección anterior, después de quitar los componentes de tendencia y ciclicidad, se obtuvo como resultado una serie aleatoria con la cual ya no se hizo tratamiento adicional.

En el análisis de series de tiempo, luego de quitar los componentes, se busca trabajar con series estacionarias, las cuales tienen un comportamiento constante, donde su media y varianza se mantienen a lo largo del tiempo, como lo muestra la siguiente imagen:



Fuente: elaboración propia.

Para profundizar en este tema, se sugiere consultar Hanke, J. (2010).

7.7. Modelos autorregresivos de promedios móviles

El empleo de estos modelos se realiza con series estacionarias. Debido a que se requieren mayores bases de probabilidad y manejo de *software* estadístico como STATA, EVIEWS, SAS, entre otros, solamente se mencionarán las principales características de estos modelos.

Los procesos autorregresivos son aquellos que se modelan en función de sus observaciones pasadas:

$$Y_t = \rho_0 + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \cdots + \rho_p Y_{t-p} + Z_t$$

Donde ρ_k es la autocorrelación de rezago *k*.

Supóngase que se tiene la siguiente serie:

t	Y_t
1	5
2	2
3	2
4	5
5	4

La autocorrelación de rezago 1, ρ_1 , se calcula con los datos de la observación siguiente:

t	Y_t	Y_{t+1}
1	5	2
2	2	2
3	2	5
4	5	4
5	4	

En el cálculo no se considera el dato en gris.

Utilizando la fórmula COEF.DE.CORREL, de Excel, la autocorrelación de rezago 1 que se obtiene es –0.1924. El resultado indica que la observación actual tiene una correlación baja negativa con una observación anterior.

Otro proceso estacionario es el de medias móviles. En este proceso, la estimación de la observación actual se encuentra en función de los errores de las observaciones pasadas:

$$Y_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}$$

Un proceso integrado es aquel que puede convertirse en estacionario aplicando diferencias. En cuanto al orden de integración de un proceso, es el número de diferencias que debemos aplicarle para convertirlo en estacionario.

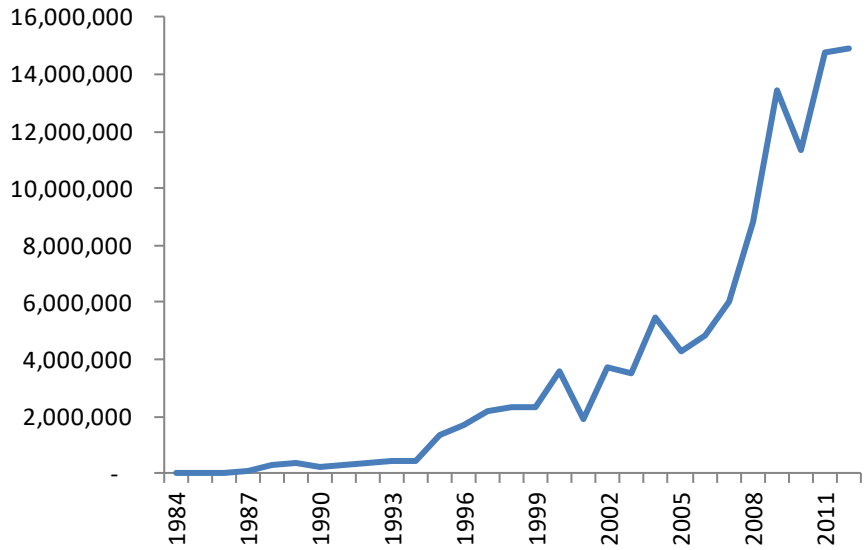
Estos modelos combinan procesos autorregresivos y de medias móviles a un proceso integrado.

Se denota ARIMA (p,d,q), donde *p* es el orden de la parte autorregresiva; *d*, el número de diferencias realizadas al modelo original para convertirla en estacionaria; y *q,* el orden de la parte de medias móviles.

Se denota ARIMA (p,d,q), donde:	*p* es el orden de la parte autorregresiva;
	d, el número de diferencias realizadas al modelo original para convertirla en estacionaria;
	q, el orden de la parte de medias móviles.

Se ejemplifica este modelo con las ventas registradas en el periodo 1984-2012 de una empresa (véase la gráfica correspondiente).

Ventas anuales 1984-2012

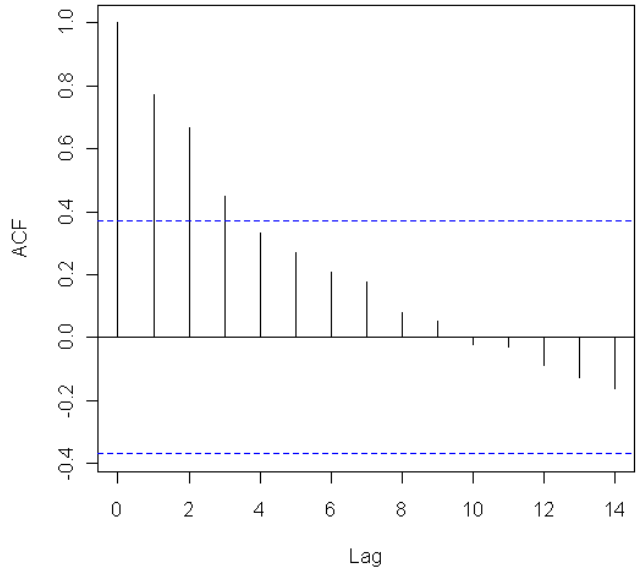


Fuente: elaboración propia.

A partir de 1994, se observa una tendencia creciente en las ventas, la cual se acentúa desde 2005.

Luego, se calculan las autocorrelaciones de la serie con diferentes rezagos y se grafica. A este gráfico se le conoce como *autocorrelograma.*

Autocorrelograma de la serie original

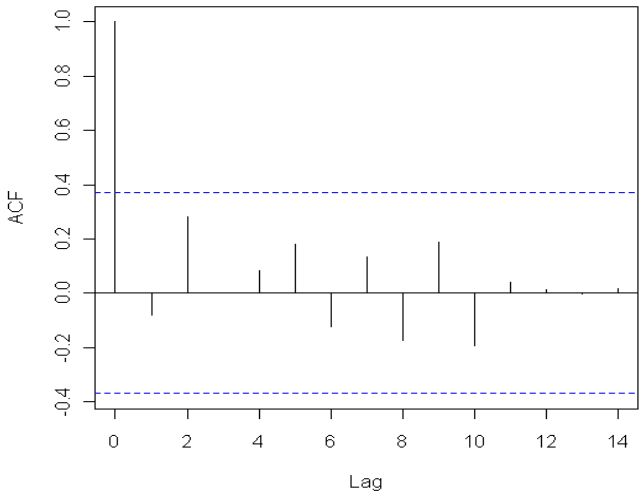


Fuente: elaboración propia. Datos procesados en el paquete estadístico R.

La gráfica anterior muestra que la observación actual está influenciada por una o dos observaciones anteriores. Después de ajustar varios modelos, se eligió un ARIMA (2, 2, 2), el que mejor se ajusta a la serie.

Para validar la calidad del modelo, se acostumbra realizar el autocorrelograma de los residuos.

Autocorrelograma de los residuos del modelo ARIMA (2, 2, 2)

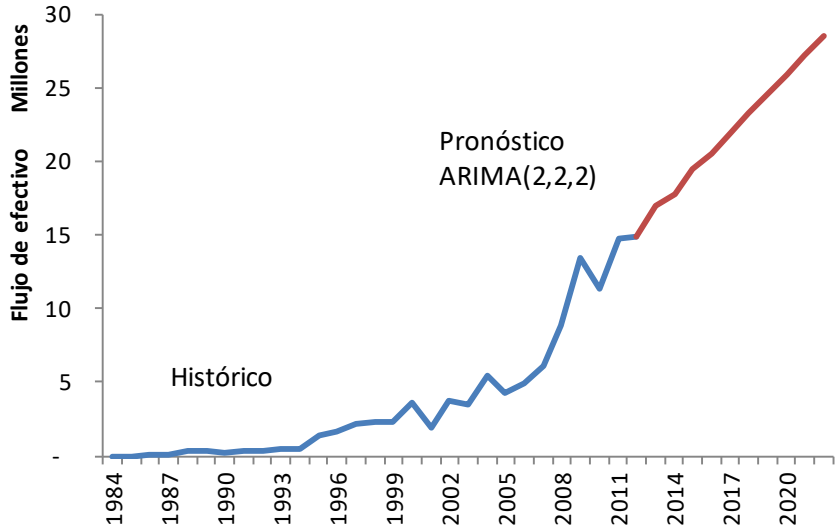


Fuente: elaboración propia. Datos procesados en el paquete estadístico R.

La gráfica anterior muestra que los residuos tienen un comportamiento de ruido blanco (aleatorio) porque se hallan dentro de la banda donde se espera caiga el 95% de las observaciones.

En la siguiente gráfica, se muestra la proyección de 10 observaciones de la serie con el empleo del modelo ARIMA (2, 2, 2).

Pronóstico de la serie con el modelo ARIMA (2,2,2)



Fuente: elaboración propia.

RESUMEN

Una serie de tiempo es una observación de los valores de una variable durante un periodo, y consta de cuatro componentes: tendencia, estacionalidad, ciclicidad y un elemento irregular o aleatorio.

Una serie puede tratarse bajo dos enfoques: el aditivo y multiplicativo. En el primero, la serie se considera que es resultado de la suma de sus componentes; mientras que, en el segundo los componentes se expresan como factores que alteran la tendencia.

Para estimar la tendencia, se utilizaron los métodos de regresión lineal y promedios móviles. Para trabajar la estacionalidad, se estimaron factores aplicados a la tendencia. Para manejar la ciclicidad, se construyó una serie cíclica que, al restarse de la serie original, da como resultado una serie irregular, la cual es deseable que sea estacionaria para poder aplicar modelos autorregresivos o de medias móviles.

Por último, se expusieron los términos *irregular* y *aleatorio*, y se mencionaron las características del modelo ARIMA, cuya aplicación se ejemplificó en una serie.

BIBLIOGRAFÍA

SUGERIDA

Autor	Capítulo	Páginas
Anderson, S.	18	785-852
Levin, R.	15	673-718
Lind, D.	16	604-647

Licenciatura: **Administración**

UNIDAD 8

Pruebas estadísticas no paramétricas

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 **2016** actualizado

OBJETIVO PARTICULAR

Al terminar la unidad, el alumno identificará las pruebas no paramétricas más utilizadas.

TEMARIO DETALLADO

(8 horas)

8. Pruebas estadísticas no paramétricas
   - 8.1. Diferencias entre los métodos estadísticos paramétricos y no paramétricos
   - 8.2. La prueba de rachas para aleatoriedad
   - 8.3. La prueba del signo
   - 8.4. La prueba de signos y rangos de Wilcoxon

INTRODUCCIÓN

En este material se ha estudiado que, para desarrollar una inferencia estadística, se debe contar con una población cuya distribución depende de un parámetro del cual se buscará inferir su valor a partir de una muestra. Asimismo, se han trabajado distribuciones muestrales que permiten realizar una estimación por intervalo o llevar a cabo una prueba, y se apoyan, en algunos casos, en supuestos como la normalidad de la población o que la muestra es considerablemente grande. Sin embargo, no siempre se puede garantizar que la población se apegue a los supuestos, por lo que es útil recurrir a pruebas no paramétricas.

Durante la quinta unidad, se utilizó la distribución χ^2 para realizar pruebas de bondad de ajuste e inferir sobre el comportamiento de una población. Ahora, esta última unidad se enfocará a los métodos no paramétricos de rachas, de signo y de signos y rangos de Wilcoxon.

Esta unidad debe tomarse como un curso introductorio a la estadística no paramétrica, en tanto brinda las bases para profundizar en el estudio de esta metodología.

8.1. Diferencias entre los métodos estadísticos paramétricos y no paramétricos

Hasta este momento, los métodos presentados tanto de estimación como de prueba de hipótesis son paramétricos, caracterizados por buscar inferir un parámetro de una población que determina la distribución de la población. Para aplicar la metodología, en ocasiones se parte de que la población sigue una distribución (frecuentemente es la normal). Sin embargo, no siempre es posible conocer o garantizar los supuestos de una distribución, por lo que se recurren a otras alternativas, las cuales no realizan restricciones acerca de la distribución de la población (a estas metodologías se les conoce como *no paramétricas*).

En estadística no paramétrica, se trabaja generalmente con datos cualitativos; a diferencia de los métodos paramétricos, donde se emplean datos cuantitativos. Cuando se manejan variables cuantitativas en estadística no paramétrica, la práctica es categorizarlas y realizar las pruebas correspondientes.

Ventajas de los métodos no paramétricos	No asumen una distribución asociada a la población.
	Su planteamiento es sencillo; y su cálculo, fácil.
	En ocasiones, los datos no requieren ser ordenados o clasificados.
	Se pueden usar en variables cualitativas.
Desventajas de los métodos no paramétricos	Ignoran información como resultado de utilizar ordenamientos en vez de los valores cuantitativos.
	Los métodos no paramétricos son menos potentes que los paramétricos.

8.2. La prueba de rachas para aleatoriedad

La primera prueba no paramétrica que se expone es la de rachas, utilizada para inferir si una muestra es aleatoria. Para aplicarla, normalmente se consideran dos resultados, como el género de una persona, el resultado del lanzamiento de una moneda, los valores por encima o debajo de la mediana, entre otros. Se enlistan los elementos de la muestra de acuerdo con el orden de aparición y se cuentan las rachas. Una racha es una secuencia de valores con una característica común precedida y seguida por valores que no presentan esa característica.

Para ilustrar una racha, supóngase que los resultados asociados a una muestra son dos: ganar (G) y perder (P). La información de una muestra de siete individuos se enlista según el orden de aparición:

$$\underbrace{G}_{R_1}\ \underbrace{G}_{R_2}\ \underbrace{P}_{R_3}\ \underbrace{G}_{R_4}\ \underbrace{P\ P}_{R_5}\ \underbrace{G}$$

En esta muestra, hay cinco rachas: la primera la forman los primeros dos individuos; la segunda y tercera, el tercer y cuarto individuos; la cuarta, el quinto y sexto individuos; y la quinta, el último individuo.

El número de rachas es un indicador de la aleatoriedad de la muestra. Si existen pocas rachas o son excesivas, entonces se está enfrentando a una muestra que no es aleatoria.

Para ilustrar lo anterior, supóngase que se hacen 10 lanzamientos de una moneda, cuyos resultados son águila (A) y sol (S), y se observan los siguientes resultados:

AAAAASSSSS

La secuencia anterior sugiere alguna carencia de independencia en los eventos.

Una secuencia como la siguiente presenta el mismo número de rachas que de lanzamientos, lo que hace pensar que los lanzamientos no se llevaron a cabo en condiciones normales:

ASASASASAS

Prueba de rachas

Planteamiento de la prueba de rachas:

$$H_o: La\ muestra\ es\ aleatoria$$
$$H_1: La\ muestra\ no\ es\ aleatoria$$

Esta prueba se centra en el número de rachas $\boldsymbol{R}$, cuya media es:

$$\mu_{R=\frac{2n_1n_2}{n_1+n_2}+1}$$

Y su desviación es:

$$\sigma_\theta = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1+n_2-1)}}$$

Donde:

n_1 = número de elementos con el primer resultado

n_2 = número de elementos con el segundo resultado

El estadístico de prueba es:

$$Z = \frac{R - \mu_R}{\sigma_R}$$

- Donde:

R = número de rachas
μ_R = media del número de rachas
σ_R = desviación del número de rachas

El estadístico de prueba se acerca a una distribución normal si n_1 o n_2 es mayor a 20. En caso de trabajar con muestras menores a 20, no es necesario calcular la media y desviación de *R*; es suficiente consultar si *R* se halla en zona de aceptación o rechazo en la tabla de valores críticos de *R*, en la prueba de rachas incluida en el apéndice.[12]

Para realizar esta prueba, conviene dar los siguientes pasos:

1. Enlistar los elementos de la muestra
2. Calcular n_1 y n_2
3. Calcular R
4. Realizar la prueba o consultar la tabla en caso de trabajar con n_1 y $n_2 < 20$.

A continuación, se plantean ejemplos.

[12] Tables for testing randomness of grouping in a sequence of alternatives. Annals of Mathematics Statistics. 4. 1943. pp. 83-86.

Ejemplo 1

Con la intención de justificar una nueva política de puntualidad en una PYME, se analizó una muestra de 20 días, donde se registra si todo el personal llegó a tiempo (P) o al menos un elemento de la organización llegó tarde (T).

La muestra deja los siguientes resultados:

Día	Resultado	Día	Resultado
1	P	11	P
2	P	12	P
3	P	13	T
4	P	14	T
5	T	15	P
6	P	16	P
7	T	17	P
8	T	18	P
9	T	19	T
10	T	20	P

Un empleado está en desacuerdo con la política que se desea implementar y dice que la muestra no fue extraída al azar. Con un nivel de significancia de 0.05, ¿se apoya lo dicho por este empleado?

Solución:

Este problema requiere de la realización de una prueba de rachas para validar la aleatoriedad de la muestra. Dando los pasos recomendados, la prueba se hace de la siguiente manera.

1. Determinar el número de veces que se registra el total de días en que todos los empleados llegaron puntuales (P) y el número de días en que al menos un empleado llegó tarde (T). Estos datos representarán el tamaño de las muestras $P = n_1$ y $T = n_2$:
$P = n_1 = 12$
$T = n_2 = 8$

2. Calcular el número de rachas (*R*)

Día	Resultado	Racha	Día	Resultado	Racha
1	P	1	11	P	5
2	P		12	P	
3	P		13	T	6
4	P		14	T	
5	T	2	15	P	7
6	P	3	16	P	
7	T	4	17	P	
8	T		18	P	
9	T		19	T	8
10	T		20	P	9

Rachas = = 9
Como n1 y n2 son menores a 20, es suficiente consultar la tabla de valores críticos de *R* en la prueba de rachas.

3. Realizar la prueba
Al consultar la tabla, se obtiene que la prueba se rechaza si $R \leq 6$ o $R \geq 16$.
Como R = 9, no hay elementos para rechazar la aleatoriedad de la muestra.

Ejemplo 2

Un médico clasifica a sus pacientes según si son recomendados por un seguro (S) o son de procedencia particular (P). Una muestra de pacientes atendidos por día se muestra a continuación:

Lunes	Martes	Miércoles	Jueves	Viernes
S	S	P	S	P
S	S	P	S	S
S	S	P	S	P
P	P	S	P	S
P	S	S	P	S
S	S	P	P	S
S	P	P	P	P
P	P	P	S	P
S	P	P	S	P
S	S	P	S	S
P	S	S	S	S
P	S	S	P	P
P	P	S	P	P
P	P	S	S	P

Con la intención de negociar incentivos con las aseguradoras, el médico envía esta información. El área técnica encargada de evaluar si la información que le manda el médico es válida, realiza una prueba de rachas. Con una significancia del 5% la muestra es válida.

Solución:

1. Determinar n_1 y n_2
n_1 = S = 35
n_2 = P = 35

2. Determinar el número de rachas

Lunes	Racha	Martes	Racha	Miércoles	Racha	Jueves	Racha	Viernes	Racha
S	1	S	7	P	...	S	...	P	20
S		S		P	12	S	15	S	21
S		S		P		S		P	22
P	2	P	8	S	13	P	16	S	23
P		S	9	S		P		S	
S	3	S		P	14	P		S	
S		P	10	P		P		P	24
P	4	P		P		S	17	P	
S	5	P		P		S		P	
S		S	11	P		S		S	25
P	6	S		S	...	S		S	
P		S		S		P	18	P	26
P		P	...	S		P		P	
P		P		S		S	19	P	

R = 26

3. Realizar la prueba
a) Calcular la media:

$$\mu_{R=}\frac{2n_1n_2}{n_1+n_2}+1$$
$$\mu_{R=}\frac{2\cdot35\cdot35}{35+35}+1$$
$$\mu_{R=}\frac{2{,}450}{70}+1$$
$$\mu_{R=}35+1$$
$$\mu_{R=}36$$

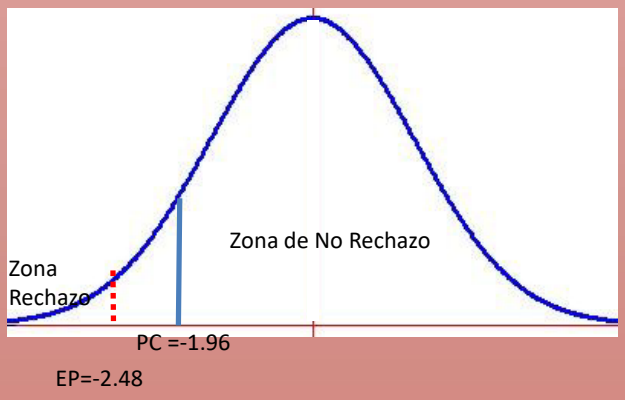
b) Calcular la desviación estándar:

$$\sigma_R=\sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2\,(n_1+n_2-1)}}$$
$$\sigma_R=\sqrt{\frac{2\cdot35\cdot35(2\cdot35\cdot35-35-35)}{(35+35)^2\,(35+35-1)}}$$
$$\sigma_R=\sqrt{\frac{2{,}480(2{,}380)}{(70)^2\,(69)}}$$
$$\sigma_R=\sqrt{\frac{2{,}480(2{,}380)}{4{,}900\,(69)}}$$
$$\sigma_R=\sqrt{\frac{5{,}831{,}000}{338{,}100}}$$
$$\sigma_R=\sqrt{17.24}$$
$$\sigma_R=4.15288$$

c) Calcular el estadístico de prueba:

$$Z=\frac{R-\mu_R}{\sigma_R}$$
$$Z=\frac{26-36}{4.15288}$$
$$Z=\frac{-10}{4.15288}$$
$$Z=-2.408$$

Calculamos el estadístico de prueba a través de una distribución normal:
DISTR.NORM.ESTAND.INV(0.05/2) = –1.9599



Como el estadístico de prueba cae en zona de rechazo, no existe evidencia estadística para apoyar la aleatoriedad de la muestra.

8.3. La prueba del signo

Esta prueba recibe el nombre “del signo” porque se basa en la dirección de la diferencia entre dos mediciones, expresada con un signo “+” o “–”, más que en los datos de donde proceden. Normalmente, se emplea para hacer pruebas relacionadas con la mediana de una población o comparar muestras apareadas.

La esencia de esta prueba es apoyar que la proporción de diferencias positivas es la misma que las negativas (p = 0.5). Obsérvese que en esta prueba hay dos resultados posibles: “+” o “– “, hay *n* observaciones independientes y una probabilidad *p* constante en cada ensayo, por lo que esta prueba está asociada a una distribución binomial cuyos parámetros son el número de observaciones (n) y la probabilidad de éxito (p). Como es sabido, si $n \cdot p > 5$ y $n \cdot (1-p) > 5$, la distribución binomial puede aproximarse a una normal.

Para este caso, el estadístico de prueba es:

$$Z = \frac{R^+ - 0.5n}{0.5\sqrt{n}}$$

- Donde:

Z = estadístico de prueba
R^+ = número de datos positivos
n = tamaño de la muestra

Para muestras pequeñas, se utiliza la distribución binomial.

A continuación, se presentan algunos ejemplos.

Ejemplo 1

El número de horas extras trabajadas al mes por una muestra de 20 empleados es la siguiente:

Empleado	Horas extras
1	22
2	27
3	25
4	12
5	14
6	11
7	16
8	24
9	13
10	20
11	12
12	13
13	21
14	17
15	18
16	11
17	27
18	18
19	14
20	17

Con un nivel de significancia del 5%, ¿se apoya la hipótesis de que la mediana es de 17 horas?

Solución

La prueba se plantea de la siguiente manera:

$$H_0: mediana = 17$$
$$H_0: mediana \neq 17$$

La prueba es de dos extremos.

A cada valor de la muestra se resta el valor de la mediana bajo la hipótesis nula y se anota el signo de la diferencia:

Empleado	Horas extras	Mediana	Diferencia el dato y la mediana		Signo
			Cálculo	Resultado	
1	22	17	22-17	5	+
2	27		27-17	10	+
3	25		25-17	8	+
4	12		12-17	-5	-
5	14		14-17	-3	-
6	11		11-17	-6	-
7	16		16-17	-1	-
8	24		24-17	7	+
9	13		13-17	-4	-
10	20		20-17	3	+
11	12		12-17	-5	-
12	13		13-17	-4	-
13	21		21-17	4	+
14	17		17-17	0	=
15	18		18-17	1	+
16	11		11-17	-6	-
17	27		27-17	10	+
18	18		18-17	1	+
19	14		14-17	-3	-
20	17		17-17	0	=

El siguiente paso es contar el número de signos positivos, negativos e iguales:

Signo	Frecuencia
+	9
-	9
-	2
Total	20

Como existen dos signos “=”, se restan al total de la muestra, por lo que disminuye su valor a 18 elementos:

$$n = 20 – 2 = 18$$

Como $18 \cdot 0.5 = 9$, se utilizará una aproximación normal.

Sustituyendo los valores en el estadístico de prueba, se obtiene lo siguiente:

R+ = número de signos positivos en la muestra = **9**
n = 18

$$Z = \frac{R^{+} - 0.5n}{0.5\sqrt{n}}$$

$$Z = \frac{9 - 0.5 \cdot 18}{0.5\sqrt{18}}$$

$$Z = \frac{9 - 9}{2.1213}$$

$$Z = \frac{0}{2.1213}$$

$$Z = 0$$

Por el planteamiento, se conoce que la prueba es de dos extremos con un nivel de significancia de α = 5% = 0.05. Para determinar los puntos críticos, se utiliza la fórmula de Excel:

DISTR.NORM.ESTAND.INV(1-α/2)
DISTR.NORM.ESTAND.INV(1-0.05/2) = 1.9599

Las zonas de rechazo son en valores menores o iguales a –1.96 y mayores o iguales a 1.96. El valor del estadístico de prueba es 0, por lo que no existe evidencia para rechazar la hipótesis nula.

En el siguiente ejemplo, se contrastarán los resultados de dos muestras.

Ejemplo 2

Una panadería quiere introducir un nuevo tipo de pan blanco, para ello hornea dos panes: uno de avena y otro de linaza. Dan a probar los panes a 10 clientes para conocer su nivel de aceptación y decidir qué sabor deben introducir en su producción, y se les pide calificar el sabor de cada pan en una escala del 1 al 10: 1 significa que el sabor es desagradable; y 10, muy agradable. Los resultados del ejercicio se muestran a continuación.

Cliente	Pan con avena	Pan con linaza
1	5	9
2	8	8
3	10	8
4	9	10
5	9	8
6	5	7
7	5	10
8	8	9
9	6	9
10	5	10

Con un nivel de significancia de 5%, ¿se apoya que los clientes prefieren más el pan de linaza que de avena?

Solución

El primer paso consiste en calcular las diferencias entre las calificaciones que los clientes dieron a los panes. Se asigna “+” cuando la calificación del pan de avena supere al de linaza, y “–” en caso contrario. Cuando la calificación es la misma, se asigna “=”. Las diferencias se muestran a continuación.

Cliente	Pan con avena	Pan con linaza	Diferencia		Signo
			Cálculo	Resultado	
1	5	9	5-9	-4	-
2	8	7	8-7	1	+
3	10	8	10-8	2	+
4	9	10	9-10	-1	-
5	9	8	9-8	1	+
6	5	7	5-7	-2	-
7	5	10	5-10	-5	-
8	8	9	8-9	-1	-
9	6	9	6-9	-3	-
10	5	10	5-10	-5	-

El conteo de los signos es el siguiente:

Signo	Frecuencia
+	3
-	7
-	0
Total	10

La prueba queda planteada de la siguiente manera:

$$H_0: p = 0.5$$

$$H_1: p > 0.5$$

La prueba es de un extremo (derecho).

Como n·p = 10 ·0.5 = 5, se empleará una distribución normal.

Por tanto, los datos que sustituiremos en la fórmula del estadístico de prueba serán los siguientes:

$\boldsymbol{R^-}$ = número de signos negativos = 7

n = 10

$$\boldsymbol{Z = \frac{R^- - 0.5n}{0.5\sqrt{n}}}$$

$$\boldsymbol{Z = \frac{7 - 0.5 \cdot 10}{0.5\sqrt{10}}}$$

$$\boldsymbol{Z = \frac{7 - 5}{1.5811}}$$

$$\boldsymbol{Z = Z = \frac{2}{1.5811}}$$

$$\boldsymbol{Z = 1.26}$$

Por el planteamiento de la prueba, se conoce que es de un extremo con un nivel de significancia de α = 5% = 0.05. Para determinar el punto crítico, se utiliza la fórmula de Excel:

DISTR.NORM.ESTAND.INV(1-α)
DISTR.NORM.ESTAND.INV(1-0.05) = 1.64

Las zonas de rechazo son en valores mayores o iguales a 1.64. Como el valor del estadístico de prueba es 1.26, se concluye que no existe evidencia para rechazar la hipótesis nula: no hay evidencia para afirmar que los clientes prefieran más el pan de linaza que el de avena.

8.4. La prueba de signos y rangos de Wilcoxon

En la sección anterior, se mostró la prueba de los signos, que se basa en la dirección de las diferencias de las mediciones, más que en su magnitud. Existe una prueba con mayor potencia, la cual, además de considerar la dirección de las desviaciones, toma en cuenta su magnitud, es la prueba de rangos asignados de Wilcoxon.

La prueba de Wilcoxon parte de las diferencias apareadas de puntuaciones entre las variables *X* y *Y*. Después, estas variables son ordenadas de manera ascendente conforme a su valor absoluto, y se les asigna un rango, al cual se le da el signo que corresponde a la diferencia. En caso de que se presenten diferencias con valor cero, se eliminan del análisis; y si existen diferencias con la misma magnitud, se les asignará un rango promedio.

La hipótesis nula es que las variables *X* y *Y* son equivalentes, con la misma mediana y la misma distribución continua. Es decir, si H_0 es cierta, se esperaría observar el mismo número de diferencias en favor de *X* que dé *Y*, o lo que es equivalente, que la suma de rangos positivos sea igual a la de negativos.

Para desarrollar esta prueba, se definen dos estadísticos:

T^+ = suma de los rangos de las diferencias positivas
T^- = suma de los rangos de las diferencias negativas

El estadístico en que se enfocará la prueba es en T^+, el cual tiene como media:

$$\mu_{T^+} = \frac{n(n+1)}{4}$$

Y desviación estándar:

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)+(2n+1)}{24}}$$

Para valores de *n* mayores o iguales a 10, tiene una distribución aproximadamente normal.

A continuación, se muestra un ejemplo del uso de esta prueba.

Ejemplo

Una compañía farmacéutica lanzó un medicamento para el estrés. Desea medir si el nivel de aceptación ha cambiado en los médicos que trabajan en el hospital Alta Tensión después de haber capacitado a la fuerza de ventas. Para ello selecciona una muestra de 15 médicos, de los que compara los resultados de dos reportes, uno anterior y otro posterior a la capacitación; en el reporte, cada médico asigna una calificación del 1 al 5 para evaluar al producto. En la siguiente tabla, se muestran las calificaciones de los médicos hacia el medicamento antes y después de la capacitación.

	Calificación del producto	
Médico	**Antes**	**Después**
1	5	3
2	5	5
3	2	2
4	4	4
5	5	1
6	4	3
7	5	3
8	2	3
9	3	2
10	1	3
11	4	1
12	4	5
13	4	5
14	5	2
15	4	2

Con un nivel de significancia del 5%, ¿se podría apoyar que la capacitación a la fuerza de ventas cambió la aceptación del producto?

Solución

A continuación, se muestra cada paso para realizar la prueba.

Calcular la diferencia de las calificaciones:

Calificación del producto			
Médico	Antes	Después	Diferencia
1	5	3	2
2	5	5	0
3	2	2	0
4	4	4	0
5	2	2	0
6	4	3	1
7	5	3	2
8	2	3	-1
9	3	2	1
10	1	3	-2
11	4	1	3
12	4	5	-1
13	4	5	-1
14	5	2	3
15	4	2	2

Calcular el valor absoluto de las diferencias:

Calificación del producto				Valor absoluto
Médico	Antes	Después	Diferencia	Diferencia
1	5	3	2	2
2	5	5	0	0
3	2	2	0	0
4	4	4	0	0
5	5	1	4	4
6	4	3	1	1
7	5	3	2	2
8	2	3	-1	1
9	3	2	1	1
10	1	3	-2	2
11	4	1	3	3
12	4	5	-1	1
13	4	5	-1	1
14	5	2	3	3
15	4	2	2	2

Eliminar las diferencias que son cero:

Médico	Calificación del producto		Diferencia	Valor absoluto diferencia	Rango
	Antes	Después			
1	5	3	2	2	
2	5	5	0	0	Se eliminan
3	2	2	0	0	
4	4	4	0	0	
5	5	1	4	4	
6	4	3	1	1	
7	5	3	2	2	
8	2	3	-1	1	
9	3	2	1	1	
10	1	3	-2	2	
11	4	1	3	3	
12	4	5	-1	1	
13	4	5	-1	1	
14	5	2	3	3	
15	4	2	2	2	

$$n = 15$$

Se eliminaron tres datos:

$$n = 15 – 3 = 12$$

Se determina el rango, es decir, se asigna a cada diferencia un número de menor a mayor. En caso de que se repita el valor de las diferencias, se calcula el promedio de los rangos que les corresponde; dicho resultado será el rango asignado a cada una de las diferencias involucradas en el cálculo de ese rango.

Después, se calcula el rango promedio para las diferencias de valor 1:

Médico	Calificación del producto Antes	Después	Diferencia	Valor absoluto diferencia	Rango	Promedio de los rangos	Rango definitivo
1	5	3	2	2	7		
5	5	1	4	4	12		
6	4	3	1	1	1	$\frac{1+2+3+4+5}{5}$	3
7	5	3	2	2	8		
8	2	3	-1	1	2	$\frac{15}{5}=3$	3
9	3	2	1	1	3		3
10	1	3	-2	2	6		
11	4	1	3	3	10		
12	4	5	-1	1	4		3
13	4	5	-1	1	5		3
14	5	2	3	3	11		
15	4	2	2	2	9		

El rango para las diferencias de valor 2 se muestra a continuación:

Médico	Calificación del producto Antes	Después	Diferencia	Valor absoluto diferencia	Rango	Promedio de los rangos	Rango definitivo
1	5	3	2	2	6	$\frac{6+7+8+9}{4}$	7.5
5	5	1	4	4	12		
6	4	3	1	1	1		3
7	5	3	2	2	7		7.5
8	2	3	-1	1	2		3
9	3	2	1	1	3		3
10	1	3	-2	2	8	$\frac{30}{4}=7.5$	7.5
11	4	1	3	3	10		
12	4	5	-1	1	4		3
13	4	5	-1	1	5		3
14	5	2	3	3	11		
15	4	2	2	2	9		7.5

El rango promedio para las diferencias de valor 3 es el siguiente:

Médico	Calificación del producto Antes	Después	Diferencia	Valor absoluto diferencia	Rango	Promedio de los rangos	Rango definitivo
1	5	3	2	2	6		7.5
5	5	1	4	4	12		
6	4	3	1	1	1		3
7	5	3	2	2	7		7.5
8	2	3	-1	1	2		3
9	3	2	1	1	3		3
10	1	3	-2	2	8		7.5
11	4	1	3	3	10	$\frac{10+11}{2}$	10.5
12	4	5	-1	1	4		3
13	4	5	-1	1	5		3
14	5	2	3	3	11		10.5
15	4	2	2	2	9		7.5

Para el valor máximo, como es único, su rango se mantiene:

Médico	Calificación del producto Antes	Después	Diferencia	Valor absoluto diferencia	Rango	Promedio de los rangos	Rango definitivo
1	5	3	2	2	6		7.5
5	5	1	4	4	12		12
6	4	3	1	1	1		3
7	5	3	2	2	7		7.5
8	2	3	-1	1	2		3
9	3	2	1	1	3		3
10	1	3	-2	2	8		7.5
11	4	1	3	3	10		10.5
12	4	5	-1	1	4		3
13	4	5	-1	1	5		3
14	5	2	3	3	11		10.5
15	4	2	2	2	9		7.5

A cada rango se le pone el signo de su diferencia original:

Médico	Calificación del producto Antes	Después	Diferencia	Valor absoluto diferencia	Rango	Rango definitivo signo original
1	5	3	2	2	6	7.5
2	5	5	0	0	Se eliminan	
3	2	2	0	0		
4	4	4	0	0		
5	5	1	4	4	12	12
6	4	3	1	1	1	3
7	5	3	2	2	7	7.5
8	2	3	-1	1	2	-3
9	3	2	1	1	3	3
10	1	3	-2	2	8	-7.5
11	4	1	3	3	10	10.5
12	4	5	-1	1	4	-3
13	4	5	-1	1	5	-3
14	5	2	3	3	11	10.5
15	4	2	2	2	9	7.5

Se suman los rangos positivos:

Médico	Calificación del producto Antes	Después	Diferencia	Valor absoluto diferencia	Rango	Rango definitivo signo original	Suma rangos positivos
1	5	3	2	2	6	7.5	7.5
2	5	5	0	0	Se eliminan		
3	2	2	0	0			
4	4	4	0	0			
5	5	1	4	4	12	12	12
6	4	3	1	1	1	3	3
7	5	3	2	2	7	7.5	7.5
8	2	3	-1	1	2	-3	
9	3	2	1	1	3	3	3
10	1	3	-2	2	8	-7.5	
11	4	1	3	3	10	10.5	10.5
12	4	5	-1	1	4	-3	
13	4	5	-1	1	5	-3	
14	5	2	3	3	11	10.5	10.5
15	4	2	2	2	9	7.5	7.5
						T^+ =	61.5

Entonces, la prueba se define de la siguiente manera:

H_0: *No hay diferencia en la calificación antes y después de la capacitación*
H_1: *Hay diferencia en la calificación antes y después de la capacitación*

La prueba es de dos colas.

Se realiza la prueba.

Se calculan el promedio y la desviación.

Donde n = 13

$$\mu_{r^+} = \frac{n(n+1)}{4}$$

$$\mu_{r^+} = \frac{12(12+1)}{4}$$

$$\mu_{r^+} = \frac{12(13)}{4}$$

$$\mu_{r^+} = \frac{156}{4}$$

$$\mu_{r^+} = 39$$

$$\sigma_{r^+} = \sqrt{\frac{n(n+1)+(2n+1)}{24}}$$

$$\sigma_{r^+} = \sqrt{\frac{12(12+1)+(2\cdot 12+1)}{24}}$$

$$\sigma_{r^+} = \sqrt{\frac{12(13)+(24+1)}{24}}$$

$$\sigma_{r^+} = \sqrt{\frac{156+25}{24}}$$

$$\sigma_{r^+} = \sqrt{\frac{181}{24}}$$

$$\sigma_{r^+} = \sqrt{7.5412}$$

$$\sigma_{r^+} = 2.7462$$

Calcular el estadístico de prueba:

$$z = \frac{T^+ - \mu_{r^+}}{\sigma_{r^+}}$$

$$z = \frac{61.5 - 39}{2.7462}$$

$$z = \frac{22.5}{2.7462}$$

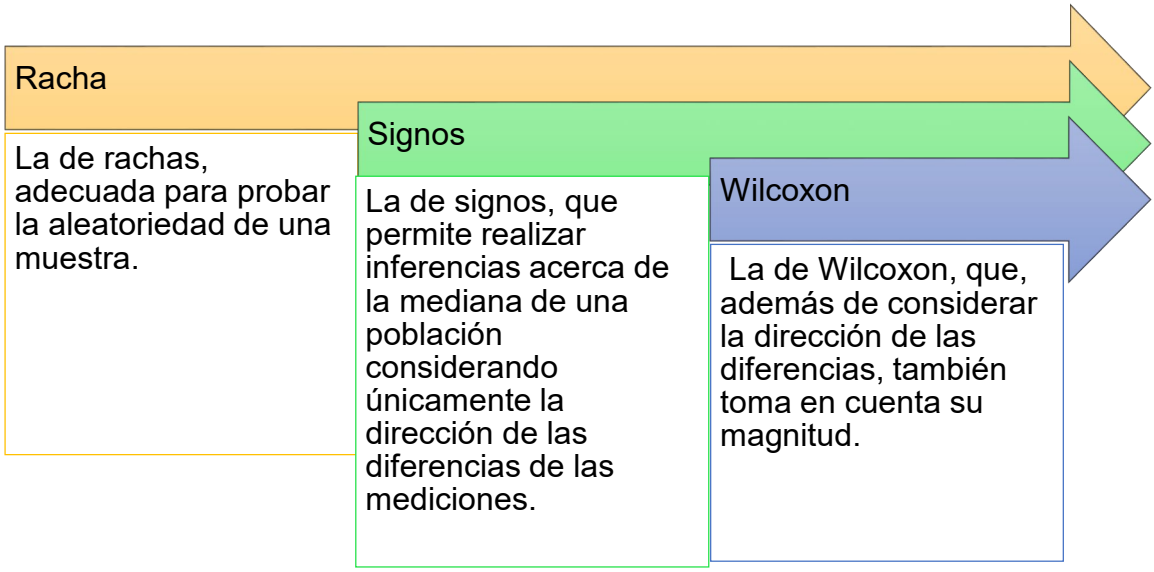
$$z = 8.1931$$

Como el estadístico de prueba es notablemente mayor a los puntos críticos (±1.96), no hay evidencia estadística para apoyar la hipótesis nula de que no hay diferencia en la calificación antes y después de la capacitación.

RESUMEN

En esta unidad, se presentó un primer acercamiento a la realización de pruebas con el empleo de métodos no paramétricos. Estos métodos tienen la ventaja de no asumir que la población sigue una distribución; sus pruebas son sencillas y entendibles, aunque no tienen la misma potencia que las pruebas paramétricas.

Se expusieron tres pruebas:



BIBLIOGRAFÍA

SUGERIDA

Autor	Capítulo	Páginas
Anderson, S.	19	855-904
Levin, R.	14	621-663
Lind, D.	18	680-719

APÉNDICE

Apéndice 1. Valores críticos de *R* en la prueba de rachas

Los diferentes valores críticos de R están proporcionados en las tablas para valores n1 y n2 menores o iguales a 20. Para la prueba de rachas de una muestra, cualquier valor observado de R que sea menor o igual al valor más pequeño, o que sea mayor o igual al valor más grande en un par, es significativo en el nivel α = 0.05.

	n2																		
n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2	2	2	2	2	2	2	2	2
											-	-	-	-	-	-	-	-	-
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
				9	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
			9	10	10	11	11	-	-	-	-	-	-	-	-	-	-	-	-
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
		-	9	10	11	12	12	13	13	13	13	-	-	-	-	-	-	-	-
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
		-	-	11	12	13	13	14	14	14	14	15	15	15	-	-	-	-	-
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
		-	-	11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
		-	-	-	13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
		-	-	-	13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
		-	-	-	13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
	-	-	-	-	13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
	-	-	-	-	-	15	16	17	18	19	19	20	20	21	21	22	22	23	23
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
	-	-	-	-	-	15	16	17	18	19	20	20	21	22	22	23	23	23	24
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
	-	-	-	-	-	15	16	18	18	19	20	21	22	22	23	23	24	24	25
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
	-	-	-	-	-	-	17	18	19	20	21	21	22	23	23	24	25	25	25
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
	-	-	-	-	-	-	17	18	19	20	21	22	23	23	24	25	25	26	26
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
	-	-	-	-	-	-	17	18	19	20	21	22	23	24	25	25	26	26	27
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
	-	-	-	-	-	-	17	18	20	21	22	23	23	24	25	26	26	27	27
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14
	-	-	-	-	-	-	17	18	20	21	22	23	24	25	25	26	27	27	28

Fuente: Siegel (1995, p. 369).

REFERENCIA BIBLIOGRÁFICA

BIBLIOGRAFÍA BÁSICA

Anderson, D. R. (2016). *Estadística para negocios y economía.* (12a ed.), México: Cengage Leraning.

Levine, D. M. (2014). *Estadística para administración.* (6 ed.), México: Pearson.

Lind, A. D. (2015). *Estadística aplicada a los negocios y a la economía.* (16a ed.), México: McGraw-Hill.

Mendenhall, W. (2015). *Introducción a la probabilidad y estadística.* (14a ed.), México: Cengage Learning.

Rodríguez, F. J. (2014). *Estadística aplicada II: estadística en administración para la toma de decisiones.* México: Grupo Editorial Patria.

Rodríguez, F. J. (2014). *Estadística para administración.* México: Grupo Editorial Patria.

Triola, M. F. (2013). *Estadística: actualización tecnológica.* (11a ed.), México: Pearson Educación.

BIBLIOGRAFÍA COMPLEMENTARIA

Alvarado, V. V. (2014). *Probabilidad y estadística.* México: Grupo Editorial Patria.

Domínguez, D. J. (2015). *Estadística para administración y economía.* México: Alfaomega.

Fontana, D. B. (2014). *Probabilidad y estadística.* México: UNAM Facultad de Ingeniería.

Funelabrada, D. T. (2014). *Probabilidad y estadística.* (4a ed.), México: McGraw-Hill.

Garza, O. B. (2014). *Estadística y probabilidad.* México: Pearson Educación.

Newbold, P. (2013). *Estadística para administración y economía.* (8a ed.), Madrid: Pearson.

Spiegel, M. R. (2013). *Probabilidad y estadística.* (4a ed.), New York: McGraw-Hill.

Plan 2012

2016

actualizado