Estadística aplicada

A las ciencias económicas y administrativas

Ciencias Sociales y Humanidades

Andrés Venereo Bravo

Dossier Académico ULEAM

EDITORIAL MAR ABIERTO

Estadística aplicada a las Ciencias Económicas y Administrativas

Andrés Venereo Bravo

EDITORIAL MAR ABIERTO

Este libro ha sido evaluado bajo el sistema de pares académicos y mediante la modalidad de doble ciego.

Estadística aplicada a las Ciencias Económicas y Administrativas

Universidad Laica Eloy Alfaro de Manabí (ULEAM)
Ciudadela universitaria vía circunvalación (Manta)
www.uleam.edu.ec

Departamento de Edición y Publicación Universitaria (DEPU)
Editorial Mar Abierto
Telef. 2 623 026 Ext. 255
www.marabierto.uleam.edu.ec
www.depuleam.blogspot.com
www.editorialmarabierto.blogspot.com

Cuidado de la edición: Alexis Cuzme
Diagramación y diseño de portada: José Márquez
Corrección: Patricio Lovato

ISBN: 978-9942-959-53-9

Primera edición: noviembre de 2016

Manta, Manabí, Ecuador.

Resumen

El libro **Estadística aplicada a las Ciencias Económicas y Administrativas** cuenta con un total de 17 capítulos mediante los cuales se ha intentado explicar de la forma más pedagógica posible, los contenidos principales de la Estadística no Paramétrica, la Estadística Paramétrica y otros métodos de esta importante rama de la Matemática.

En el Capítulo 1 se estudian los elementos básicos de la Estadística con el objetivo de que sirvan como una introducción al estudio de esta disciplina.

El agrupamiento de datos y la forma adecuada de elaborar tablas y gráficos es tratado en el Capítulo 2, y en el mismo se estudian las *Tablas de frecuencias* para datos cualitativos y las *Distribuciones de frecuencias* para datos cuantitativos.

En el Capítulo 3 se desarrolla un estudio detallado de las *Medidas de tendencia central y de variación* las cuales representan un elemento de singular importancia en el desarrollo del resto de los capítulos del libro.

Una *Introducción a la teoría de probabilidades* y el estudio de las *Distribuciones teóricas de probabilidad discretas y continuas* son abordados en los Capítulos 4 y 5, mientras que el importante tema del *Muestreo y las distribuciones de muestreo* es revisado en el Capítulo 6.

Los Capítulos 7,8 y 9 son dedicados al estudio de los temas relacionados con la teoría de la estimación estadística, y con ese objetivo, se abordan los contenidos relacionados con la *Estimación e intervalos de confianza*, la *Prueba de hipótesis para una sola muestra* y la *Prueba de hipótesis para dos muestras*.

Los elementos teóricos y prácticos vinculados con la técnica del *Análisis de varianza* son revisados en el Capítulo 10, en el cual se estudian los diseños Completamente al Azar y Bloques al Azar así como los arreglos factoriales y el no menos importante tema relativo al número de repeticiones en el diseño experimental.

Las técnicas relacionadas con el *Análisis de regresión simple y correlación* así como las de *Regresión múltiple* son tratadas en los Capítulos 11 y 12, mientras que *Los métodos no paramétricos*, específicamente las *Aplicaciones de la Ji – Cuadrada* y el *Análisis de datos ordenados* son estudiados en los Capítulos 13 y 14.

Por último en los Capítulos 15, 16 y 17 se abordan aspectos de vital importancia para los profesionales de la Economía y la Administración. Concretamente nos referimos a las *Series cronológicas*, los *Números índice* y una *Introducción a la teoría de decisiones*, con lo cual se completa el contenido planificado para este libro.

Palabras claves: Estadística Paramétrica, estadística no Paramétrica, métodos aplicados a la Economía y la Administración.

Índice de contenidos

Para Guadalupe, que además de mi esposa, siempre ha sido mi apoyo y sin lugar a dudas mi mejor amiga; para mis hijos Nubia, Andrés y Arian; para la memoria de mis padres que fueron siempre un ejemplo a imitar.

Introducción

La Estadística es una ciencia formal que tiene múltiples aplicaciones en casi todas las áreas del conocimiento, y de manera especial, es una herramienta de trabajo que nos permite llevar a vías de hecho y de manera eficiente el proceso relacionado con la investigación científica, lo cual la convierte en una ciencia universal.

Es transversal con relación a una gran cantidad de otras disciplinas, entre las cuales podríamos citar a modo de ejemplo las relacionadas con las ciencias naturales y sociales, la física, las ciencias médicas, y en particular, las ciencias económicas y administrativas.

Los profesionales de la economía al pretender comprender el comportamiento de un determinado sistema económico, lo hacen mediante el empleo de modelos económicos y matemáticos que son construidos mediante la aplicación de la Estadística. La construcción de los modelos a los que hemos hecho referencia es uno de los objetivos específicos de la Econometría.

La integración adecuada de la Estadística, la teoría económica y las matemáticas se convierte en un poderoso instrumento que permite estudiar el comportamiento de las relaciones económicas en la actualidad.

Por las razones explicadas anteriormente, la malla curricular de la carrera de Economía de la Universidad Laica "Eloy Alfaro" de Manabí, presenta tres niveles en los que se estudia la Estadística y dos en los que se explica la Econometría.

Por otra parte, en un mundo globalizado donde los cambios se producen de manera acelerada y de la noche a la mañana, toda empresa requiere hacer de la estadística una herramienta de uso cotidiano. Un administrador que se respete y desee mantener su puesto, requiere poder estimar con cierta confiabilidad los niveles de demanda de sus productos, conocer de forma inmediata los cambios que se producen en el mercado y saber los gastos en que ha incurrido su empresa y en qué rubros estos se han producido.

Según lo expresado por W. Edwards, el cual fue uno de los primeros en utilizar métodos de la Estadística en los procesos de control de calidad, el conocimiento de los métodos estadísticos por parte de los que administran una empresa resulta algo de vital importancia.

Sin el uso de la Estadística una empresa está imposibilitada de poder reconocer y mantener bajo control aquellas actividades que le generan ganancias y aquellas que por el contrario le producen pérdidas.

Considero que los argumentos que han sido expuestos en los párrafos anteriores justifican de alguna manera la elaboración del presente libro, y espero, que la utilización del mismo por parte de los profesionales de la economía, la administración y otras esferas del conocimiento, sirva de instrumento para la aplicación en sus esferas de actividad de la importante herramienta que representa la Estadística.

Capítulo 1
Una introducción

El problema

Un importante medio de comunicación escrita del Ecuador asegura que aproximadamente el 12% de los accidentes de tránsito en el país se encuentran vinculados con la ingesta de bebidas alcohólicas. Según este dato estadístico, ¿podemos llegar a la conclusión que no ingerir alcohol es la causa del 88% restante de los accidentes de tránsito, o en realidad esta afirmación sería una forma irresponsable y descabellada de interpretar un resultado estadístico?

1.1 Definición de Estadística

Desde el punto de vista etimológico, el término estadística proviene del latín *statiscum collegium*, que traducido al español significa *Consejo de Estado*, y de su derivado italiano *statista* que significa *hombre de Estado o político*. Por lo antes expuesto podemos concluir que los orígenes de la estadística están íntimamente vinculados con el gobierno y su aparato administrativo.

En el año 1749, el filósofo, historiador, economista, jurista y estadístico de origen alemán Gottfried Achenwall, introdujo el término *statistik* asociado con *el análisis de datos estatales*.

Wikipedia, la enciclopedia libre, al definir la estadística señala, y cito: *La estadística es una ciencia formal y una herramienta que estudia el uso y los análisis provenientes de una muestra representativa de datos, busca explicar las correlaciones y dependencias de un fenómeno físico o natural, de ocurrencia en forma aleatoria o condicional.* En otras palabras, podemos definir la estadística como *la ciencia que reúne, clasifica, organiza, analiza e interpreta conjuntos de datos con el objetivo de llegar a determinadas conclusiones que permitan la toma de decisiones más adecuada*.

1.2 Antecedentes históricos

El uso de métodos estadísticos sencillos e incipientes se remonta a los inicios de nuestra civilización cuando los hombres y mujeres de esa época utilizaban determinados gráficos y símbolos para contar personas, animales y cosas y que quedaron plasmados en pieles, rocas, pedazos de madera y en las paredes de algunas cuevas.

Con el objetivo de recopilar información con relación a la producción agropecuaria y a los artículos cambiados mediante trueque, los babilonios utilizaban tablillas de arcilla hacia el año 3000 antes de Cristo.

Alrededor del año 3050 antes de Cristo, los faraones del antiguo Egipto pudieron obtener importantes datos relacionados con la población y las riquezas del país. Según el historiador y geógrafo griego Heródoto de Halicarnaso (484 – 425 A.C.), lo anterior fue hecho con el objetivo de iniciar la construcción de las pirámides de Egipto.

En los libros bíblicos de Números y Crónicas aparecen algunos trabajos realizados con la aplicación incipiente de la estadística. En el primero de estos libros aparecen dos censos realizados en la población de Israel y en el segundo se puede apreciar información sobre el bienestar material de las tribus judías.

Con anterioridad al año 2000 antes de Cristo, en China existían registros muy similares a los encontrados en los libros bíblicos, y los antiguos griegos hacia el año 594 antes de Cristo realizaban censos periódicos que tenían un fin tributario. Se ha comprobado a través de investigaciones de carácter histórico, que se realizaron 69 censos para calcular los impuestos, establecer los derechos de voto y estimar la fuerza militar de la época.

En la época del Imperio Romano los funcionarios públicos tenían un registro que debían mantener actualizado de forma obligatoria y en el que anotaban los nacimientos, fallecimientos y matrimonios que se producían. Cada cinco años realizaban empadronamientos de la población.

En los años 758 y 762 respectivamente, Pipino el Breve y Carlomagno elaboraron informes sobre las tierras que eran propiedad de la iglesia.

En el año 1532 existían en Inglaterra registros de los fallecimientos producidos por la peste y en Francia resultaba una obligación que los clérigos registraran los bautismos, fallecimientos y matrimonios que se producían en la época.

En su obra *Observaciones políticas y naturales echas a partir de las Cuentas de Mortalidad*, el Capitán John Graunt (1620 – 1674), demógrafo inglés considerado el fundador de la bioestadística y el precursor de la epidemiología, realizó inferencias acerca del número de personas que morirían a causa de determinadas enfermedades.

A partir del siglo XVII y principios del siglo XVIII comienza el desarrollo de la teoría de probabilidades. Destacados matemáticos tales como Girolano Cardano, Blaise Pascal, Pierre de Fermat, Christian Huygens, Daniel Bernoulli, Joseph – Louis Lagrange, Pierre Simon de Laplace, Carl Friedrich Gauss, Simeón – Denis Poisson y Abraham de Moivre crearon las bases teóricas de la teoría de probabilidades.

El debut de la estadística moderna puede ser ubicado a partir de los trabajos de Francis Galton y Kurt Pearson. A Pearson se debe la prueba de Chi – Cuadrado y la publicación en 1892 del libro La Gramática de la Ciencia, el cual constituyó un gran aporte a la ciencia. Otra destacadísima figura de la estadística moderna lo constituye el científico, matemático, estadístico, biólogo evolutivo y genetista inglés Ronald Ayl-

mer Fisher (1890 – 1962). Fue el pionero en el estudio del diseño experimental y el análisis de varianza. Hizo aportes importantes en la aplicación de métodos estadísticos a la biología y la agronomía. Una mancha en la trayectoria científica de Fisher fue su apoyo a la teoría de la *eugenesia*, la cual estudia las vías para mejorar la raza humana y cómo lograr que las características que se consideran mejores se desarrollen a mayor velocidad que las consideradas inadecuadas. El gobierno alemán presidido por Hitler promulgó en 1933 la ley de *esterilización eugenésica* la que constituyó un preámbulo de los exterminios masivos realizados en los campos de concentración bajo una supuesta investigación de carácter médica.

1.3 Población y Muestra

Población es cualquier conjunto formado por elementos que pueden ser inequívocamente identificados y que además poseen uno o más atributos factibles de ser medidos**.** Por ejemplo, el conjunto de todas las familias de clase alta es una *población*, ya que los elementos de este conjunto son fácilmente identificables el uno del otro y además, poseen características medibles u observables, tales como, la cantidad de miembros, la edad promedio, sus ingresos, etc.

Sin embargo, resulta necesario dejar bien claro que en Estadística el concepto de población no debe ser confundido con el concepto de población humana, la cual está integrada por personas. En Estadística una población puede ser el conjunto de computadoras que pertenecen a la Universidad Laica Eloy Alfaro de Manabí, o el conjunto de contribuyentes de una determinada provincia, o el número de libros que integran una biblioteca nacional, etc. Es decir, una población en Estadística puede ser un conjunto integrado por personas, animales, cosas o cualquier otro tipo de identidad.

Pero estudiar una característica cualquiera usando para ello todos los elementos de una población resulta, en la inmensa mayoría de los casos, una tarea imposible, y es por ello que en la práctica, solo sea factible analizar una parte de la misma. A cualquier subconjunto de elementos extraídos de una determinada población se le llama *Muestra*.

El *tamaño* de una población o de una muestra viene dado por la cantidad de elementos que la conforman y se representan con las letras N y n respectivamente. El tamaño de la población puede ser finito o infinito. El tamaño de una muestra es siempre finito.

La extracción de una muestra de una determinada población tiene siempre un objetivo central: tomar decisiones acerca del comportamiento de un atributo de la población, conocido con el nombre de *parámetro*, tomando como base el comportamiento de este mismo atributo en la muestra extraída, conocido con el nombre de

estadígrafo.

Precisando, cualquier indicador que haya sido calculado utilizando la totalidad de las observaciones de la población recibe el nombre de *parámetro*. Cualquier indicador que haya sido calculado utilizando los datos de una muestra extraída de una población recibe el nombre de *estadígrafo*.

1.4 División de la estadística

Con el objetivo de facilitar su estudio la estadística se divide en dos grupos o categorías, a saber, la *estadística descriptiva* y la *estadística inferencial*.

1.4.1 Estadística descriptiva

La *estadística descriptiva* es la parte de la estadística que se ocupa de recolectar, ordenar, organizar, resumir y analizar un conjunto de datos con el objetivo de describir de manera informativa las principales características de éste. Para lograr su objetivo la *estadística descriptiva* se apoya en el cálculo de algunos *indicadores* y en el empleo de diferentes tipos de *gráficos*. En capítulos posteriores estudiaremos con más detalle esta importante parte de la estadística.

1.4.2 Estadística inferencial

La *estadística inferencial* también llamada *inferencia estadística*, se ocupa de estudiar los métodos y procedimientos que permiten *predecir o inferir* el valor de un parámetro poblacional a partir del conocimiento del valor del estadígrafo correspondiente. En capítulos posteriores estudiaremos con más detalle esta importante parte de la estadística.

1.5 Tipos de variables

En general existen dos tipos de variables, las *cualitativas* y las *cuantitativas*.

Las variables *cualitativas*, también llamadas *atributos*, son aquellas cuyo valor no puede ser expresado de forma numérica. Se pueden citar como ejemplos de variables cualitativas el color, el sabor, el género, la raza, el estado civil, etc.

Las variables *cuantitativas*, como su nombre lo indica, son aquellas cuyo valor puede ser expresado de forma numérica. Ejemplos de variables cuantitativas son la edad, el saldo de una cuenta bancaria, el impuesto pagado en un año fiscal, la inflación, el sueldo mensual, etc.

1.6 Escalas de medición

En el libro Handbook of Experimental Psychology (1951), Stevens señala que *medir es asignar números a los objetos según ciertas reglas* y considera cuatro dife-

rentes *escalas de medición: nominal, ordinal, de intervalo y de razón*.

Escala nominal

En este nivel los datos de una variable cualitativa solo se clasifican y se cuentan. No admite ninguna operación aritmética entre ellos pues el resultado no tendría ningún sentido. Por ejemplo, en una reunión de negocios hay 7 economistas de los cuales 4 son hombres y 3 son mujeres. Sería correcto asignarle a los hombres el código 1 y a las mujeres el código 2, pero ¿tendría sentido hacer lo siguiente

$$PROMEDIO = \frac{1+1+1+1+2+2+2}{7} = \frac{10}{7} = 1.43$$

Por supuesto que no tendría ningún sentido concluir que el género promedio de economistas en la reunión es igual a 1.43. Otros ejemplos son:

Estado civil (1 Soltero, 2 Casado, 3 Viudo, 4 Divorciado)

Preferencia (1 De acuerdo, 2 En desacuerdo)

Ubicación (1 Manabí, 2 El Oro, 3 Guayas, 4 Pichincha)

Calidad del agua (1 Potable, 2 No potable)

Nivel de posgrado (1 Especialista, 2 Magister, 3 Doctor en Ciencias)

Escala ordinal

La escala ordinal establece categorías o grupos que tienen un orden jerárquico. Establece posiciones relativas de los objetos bajo estudio con relación a una determinada característica sin reflejar distancias entre ellos.

Un ejemplo de datos ordinales se muestra en la tabla 1.1, donde se han encuestado a 300 clientes de un negocio con relación a la calidad del servicio que reciben.

Tenga en cuenta que la opción 1 de ninguna manera significa que *Muy satisfecho* es 4 veces mejor que *Muy insatisfecho*.

En la tabla 1.1 la opinión *Muy satisfecho* es mejor que *Satisfecho*, *Satisfecho* es mejor que *Insatisfecho* y éste mejor que *Muy insatisfecho*. Sin embargo no es posible establecer la magnitud de la diferencia entre los grupos.

TABLA 1.1 Frecuencias de la opinión de los clientes encuestados

Opinión	Frecuencia
1. Muy satisfecho	85
2. Satisfecho	125
3. Insatisfecho	50
4. Muy insatisfecho	40
Total	300

Escala de intervalo

Este nivel de medición posee todas las cualidades del nivel de medición ordinal, y adicionalmente, tiene la característica de que la distancia entre sus valores está claramente determinada y es siempre una cantidad constante. Un ejemplo claro de este tipo de medición viene dado por las escalas de temperatura. En este caso, la diferencia existente entre 17 y 18 grados centígrados es la misma que la existente entre 32 y 33 grados. Una limitante de este tipo de escala es que no es posible definir el cero absoluto, es decir, la ausencia total de la cualidad que se mide. De este modo, la temperatura de cero grados centígrados no significa en forma alguna que haya ausencia de temperatura.

Las escalas utilizadas para medir algunas pruebas sicológicas y de rendimiento pertenecen a este tipo de medición. La escala de intervalo no permite expresar equivalencias desde el punto de vista matemático, es decir, no es posible establecer que 7 grados centígrados es la mitad de temperatura que 14 grados centígrados.

Escala de razón

Este tipo de medición, también llamada *escala de cocientes*, posee todas las características de los casos anteriores y adicionalmente permite establecer un cero real, lo cual hace posible realizar algunas operaciones matemáticas entre las que se encuentran las proporciones y los cocientes.

En una escala de este tipo un valor de 30 es el doble que un valor de 15 y también las dos terceras partes de un valor igual a 45.

Ejemplos de escala de razón son aquellas empleadas en la física, tales como, las que miden la intensidad de la corriente, la masa, etc. Son la economía y la demografía, entre otras disciplinas, las que con mayor frecuencia utilizan la escala de razón.

1.7 Impugnaciones a la estadística

En el artículo de divulgación titulado *Mentiras, pecados y abusos estadísticos*, su autor, Bartolo Luque Serrano escribió, y cito:

No es de extrañar que, con las manipulaciones y los malos usos estadísticos, el público en general acabe navegando entre la fascinación y la repudia por las cifras. "Existen medias mentiras, mentiras y estadísticas", oímos con frecuencia decir satisfecho al tertuliano de turno. La frase correcta debería ser: "Existen medios mentirosos, mentirosos y estadísticos embaucadores".

Desde hace algún tiempo el uso de la estadística y las conclusiones a las que se arriba con su utilización, han sido blanco de impugnaciones alejadas totalmente de la

realidad. Una de las más conocida se le atribuye *supuestamente* al político, escritor y aristócrata británico Benjamín Disraeli, conocido también como Conde de Beaconsfield o Lord Beaconsfield, quien expresó que existen tres tipos de mentiras, *Mentiras, Malditas Mentiras y Estadísticas*. Esta famosa frase ha sido atribuida también al escritor y humorista estadounidense Samuel Langhorne Clemens, más conocido por su seudónimo de Mark Twain, y al político, académico y hombre de letras británico Leonard Henry Courtney.

Alguien que apoye tan *lamentablemente frase célebre* expresada en el párrafo anterior y que maneje la estadística de forma superficial, podría tener la osadía de llegar a la conclusión que al tener ésta tres posibles autores y también tres palabras diferentes, entonces a cada uno de ellos se le podría otorgar el 33.3% de la autoría.

En realidad las injustas y calumniosas críticas al uso de la estadística se deben, en el fondo, *al abuso estadístico del cual somos víctimas por nuestro analfabetismo numérico*, tal y como lo expresó el científico, filósofo y académico estadounidense Douglas Richard Hofstadter.

Como señaló el matemático e informático canadiense Alexander Keewatin Dewdney, y cito, *"Aquellos que abusan de las matemáticas también abusan de nosotros. Nos convertimos en presas de las triquiñuelas comerciales, las estafas financieras, la charlatanería médica y el terrorismo numérico de los grupos de presión, todo porque somos incapaces (o no estamos dispuestos) a pensar con claridad durante unos momentos".*

Capítulo 2
Agrupamiento de datos. Tablas y gráficos.

El problema

La Corporación Nacional de Telecomunicaciones de la provincia de Manabí, desea conocer el grado de satisfacción que tienen los usuarios del servicio de Internet que ellos ofrecen. Para ello escogió a 200 de estos usuarios y a través de una encuesta les preguntó cómo consideraban la calidad del servicio, si excelente, bien, regular o mala. ¿Tiene la corporación la posibilidad de resumir los resultados con la finalidad de poder mostrar la distribución de los mismos y en qué punto tienden a concentrarse?

2.1 Introducción

En el capítulo anterior comenzamos el estudio de la estadística descriptiva y vimos que una de sus principales funciones era la de organizar los datos con el objetivo de describir la forma en que éstos se distribuyen resumiendo los mismos mediante tablas y gráficos.

En el presente capítulo estudiaremos las *tablas y distribuciones de frecuencias*, las cuales son instrumentos básicos que nos permiten organizar y resumir un conjunto de datos. Las *tablas de frecuencias* corresponden a *variables de tipo discretas y las distribuciones de frecuencias a variables de tipo continuas*.

2.2 Tabla de frecuencias. Datos cualitativos

Si quisiéramos expresar una definición de lo que es una *tabla de frecuencias* podríamos decir que *es el resultado de agrupar en clases mutuamente excluyentes a datos que provienen de una variable cualitativa, incluyendo en dicho agrupamiento el número de observaciones que pertenecen a cada clase*.

Para construir una *tabla de frecuencias* a través de un ejemplo, consideremos que los resultados de la encuesta realizada a los clientes de la Corporación Nacional Telefónica, con el objetivo de medir el grado de satisfacción que tienen los mismos con relación al servicio de Internet que ofrece la empresa, fueron los que se muestran en la tabla 2.1.

TABLA 2.1 Resultados de la encuesta realizada a clientes de CNT

Clasificación	Número de usuarios
Excelente	85
Bien	70
Regular	30
Mal	15
Total	200

En una *tabla de frecuencias*, la clasificación en cuatro categorías hecha por CNT recibe el nombre de *clases* y el número de usuarios que seleccionaron cada una de esas categorías recibe el nombre de *frecuencia de clase o frecuencia absoluta*.

TABLA 2.2 Tabla de frecuencias

Clases	Frecuencias
Excelente	85
Bien	70
Regular	30
Mal	15
Total	200

También en una tabla de frecuencias se suelen mostrar las llamadas *frecuencias relativas*, que son cantidades que expresan la relación existente entre las *frecuencias absolutas* y el total de observaciones. En nuestro ejemplo, muestran la proporción de usuarios que eligieron una determinada opción con relación al total de usuarios.

Para calcular estas frecuencias relativas se dividen las frecuencias absolutas entre el total de usuarios. Los resultados obtenidos se muestran en la tabla 2.3. Observe que *la suma de las frecuencias relativas es igual a 1*, lo cual resulta ser *una regla*.

TABLA 2.3 Frecuencias absolutas y relativas

Clases	Frecuencias absolutas	Frecuencias relativas
Excelente	85	0.425
Bien	70	0.350
Regular	30	0.150
Mal	15	0.075
	200	1

2.2.1 Representación gráfica de una tabla de frecuencia

Existen dos tipos de gráficos que pueden ser utilizados para representar de una

manera adecuada el comportamiento de una tabla de frecuencias.

Estos son el *gráfico de columnas y el gráfico circular*.

a) Gráfico de columnas.

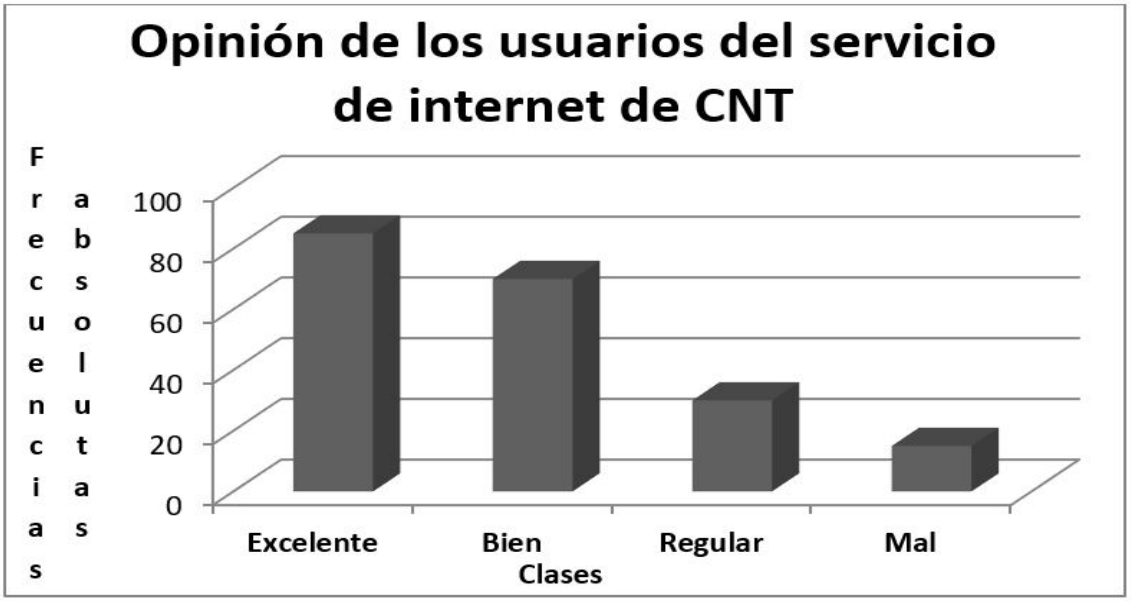
El eje horizontal de este tipo de gráfico muestra las *clases* y el eje vertical las *frecuencias absolutas*. Su característica fundamental es que al ser un gráfico de variable discreta, sus columnas no son adyacentes y en consecuencia muestran una separación entre ellas.

b) Gráfico circular.

El *gráfico circular* es también conocido como *gráfico de pastel, gráfico de torta o gráfico de 360 grados*.

En la figura 2.1 se muestra un gráfico de columnas con las frecuencias absolutas reportadas en la tabla 2.3. De igual forma, la figura 2.2 corresponde a un gráfico circular con las frecuencias relativas que aparecen en la mencionada tabla.

FIGURA 2.1 Gráfico de columnas



2.3 Distribuciones de frecuencias. Datos cuantitativos

Los datos que se muestran en la tabla 2.4 representan los ingresos mensuales de 100 familias pobres de barrios marginales, los cuales debido a que no presentan ningún tipo de *organización* y no están *resumidos*, reciben el nombre de *datos no agrupados*.

Veamos cómo elaborar una *distribución de frecuencias* haciendo uso de esta información.

FIGURA 2.2 Gráfico circular

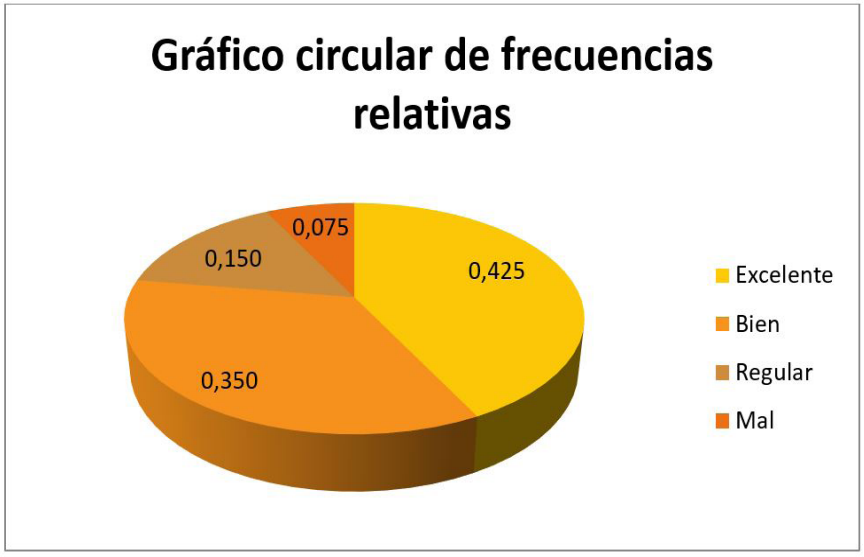


TABLA 2.4 Ingresos mensuales de 100 familias pobres

463	456	467	454	469	460	478	481	459	464
457	451	471	460	461	477	462	484	476	462
468	469	461	454	456	463	460	483	458	466
455	460	472	462	460	481	464	479	461	459
467	452	469	450	473	464	466	463	460	474
460	463	454	474	462	453	485	461	459	458
468	455	473	462	456	482	461	462	452	472
454	464	469	465	471	460	466	463	459	466
456	453	461	474	463	460	464	453	483	458
471	467	455	476	454	466	461	459	477	452

1.- El primer paso consiste en determinar el *rango o amplitud (R)* de las observaciones, el cual se obtiene restando el mayor y el menor de los ingresos mensuales. Si estudia la tabla anterior concluirá que el valor más bajo es *450* y el más alto *485*, por tanto, $R = 485 - 450 = 35$.

2.- El siguiente paso consiste en establecer el número de *clases o intervalos* que debemos utilizar en la distribución de frecuencias, de forma tal que permita visualizar de la mejor manera la tendencia en que se distribuyen los datos.

- Una vía para establecer el número de clases es mediante la *regla* 2^k donde k es la cantidad de clases. La regla consiste en seleccionar el menor valor de k para el cual $2^k > n$, donde n es el número de observaciones.
- En nuestro ejemplo n = 100, por tanto, el número de clases debe ser igual a 7 ya que $2^6 = 64 < 100$, pero $2^7 = 128 > 100$.
- Otra alternativa sería calcular el número de clases utilizando la *regla de Sturges* que consiste en determinar este valor mediante la expresión:

c = 1+3.322 log n,

donde n es el número de observaciones. Usando esta regla:

c = 1+3.322 log 100 = 1+3.322(2) = 7.6

Dada la cercanía entre ambos resultados (7 y 7.6), utilizaremos 7 como el número de clases.

3.- A continuación debemos proceder a determinar la *longitud de cada intervalo o ancho de clase*, de manera tal que las clases en su conjunto cubran todos los datos, es decir, desde el valor mínimo hasta el valor máximo. Esto se logra dividiendo el rango entre el número de clases o intervalos, es decir, $\frac{35}{7} = 5.$

De forma general, la *longitud de cada intervalo* se calcula redondeando el cociente anterior a la unidad (u) más pequeña inmediata superior en que se encuentran expresados los datos. Por ejemplo, si los datos están expresados en números enteros y el rango es igual a 5.7, entonces la longitud del intervalo debe ser igual a 6. En nuestro ejemplo no es necesario el redondeo por cuanto el rango es un número entero.

4.- El siguiente paso consiste en determinar los límites de clase, tanto el inferior como el superior, tomando en cuenta que ya establecimos que 5 debe ser el ancho de clase. Para la primera clase se coloca como límite inferior (Li) el valor más pequeño de los datos y como límite superior Ls = Li + (longitud del intervalo – 1), es decir, Li = 450 y Ls = 450 + (5 – 1) = 454. Para calcular Ls se le ha restado a 5 un 1 ya que éste último es la unidad más pequeña de los datos.

Para obtener el límite inferior de la segunda clase se le suma un entero al límite superior de la primera clase, esto es 454 + 1 (recuerde que 1 es la unidad más pequeña de los datos) y al resultado 455 se le suman (5 – 1) unidades para obtener el límite superior de la segunda clase, es decir, 455 + 4 = 459.

Este proceso se sigue repitiendo hasta completar los 7 intervalos. En la tabla 2.5 se muestran los intervalos y los resultados obtenidos.

TABLA 2.5 Intervalos de clase

Clases	Ingresos ($) Li - Ls
1	450 - 454
2	455 - 459
3	460 - 464
4	465 - 469
5	470 - 474
6	475 - 479
7	480 - 484

Como se puede apreciar, el límite superior de la última clase (484) es menor al

valor máximo de los datos (485), cuando en realidad, debe ser igual o mayor que él. Para intentar resolver la situación que se presenta en este caso o en cualquier otro caso similar, podemos disminuir el número de intervalos a 6, recalcular los nuevos límites de clase y ver si la situación queda resuelta.

Longitud de los intervalos $= \frac{Rango}{6} = \frac{35}{6} = 5.83 \approx 6$

Los nuevos intervalos de clase se muestran en la tabla 2.6.

TABLA 2.6 Intervalos de clase

Clases	**Ingresos ($) Li - Ls**
1	450 – 455
2	456 – 461
3	462 – 467
4	468 – 473
5	474 - 479
6	480 - 485

Como se puede apreciar en la tabla, la situación que se presentaba con el límite superior de la última clase ha quedado resuelta.

5.- A continuación debemos obtener los *límites reales de clase*, los cuales se obtienen restándole media unidad (u/2) a los límites inferiores de clase y sumándole esa misma cantidad a los límites superiores. En nuestro ejemplo u = 1, por tanto, la cantidad a sumar y restar es 0.5. Así, el *límite real inferior (Lri)* de la primera clase es *450 – 0.5 = 449.5 y el límite real superior (Lrs)* de esta clase es *455 + 0.5 = 455.5, el límite real inferior (Lri)* de la segunda clase es *456 – 0.5 = 455.5* y el *límite real superior (Lrs)* de esta clase es *461 + 0.5 = 461.5* y así para el resto de las clases.

TABLA 2.7 Incorporación de la columna de intervalos reales de clase

Clases	**Ingresos ($) Li - Ls**	**Ingresos ($) Lri - Lrs**
1	450 – 455	449.5 - 455.5
2	456 – 461	455.5 - 461.5
3	462 – 467	461.5 - 467.5
4	468 – 473	467.5 - 473.5
5	474 - 479	473.5 - 479.5
6	480 - 485	479.5 - 485.5

6.- El siguiente paso consiste en determinar la *marca de clase* o *punto medio* de cada uno de los intervalos, el cual se obtiene hallando la semisuma entre los dos límites del intervalo, es decir, para la primera clase (450 + 455)/2 = 452.5 para la segunda clase (456 + 461)/2 = 458.5 y así sucesivamente.

7.- Este paso consiste en determinar la cantidad de observaciones que perte-

necen a cada clase, es decir, *la frecuencia absoluta* de cada una de ellas. Las marcas de clase y las frecuencias absolutas se aprecian en las tablas 2.8 y 2.9.

TABLA 2.8 Incorporación de la columna marcas de clase

Clases	Ingresos ($) Li - Ls	**Ingresos ($) Lri - Lrs**	**Marcas de clase**
1	450 – 455	449.5 - 455.5	452.5
2	456 – 461	455.5 - 461.5	458.5
3	462 – 467	461.5 - 467.5	464.5
4	468 – 473	467.5 - 473.5	470.5
5	474 - 479	473.5 - 479.5	476.5
6	480 - 485	479.5 - 485.5	482.5

TABLA 2.9 Incorporación de las frecuencias absolutas de cada clase

Clases	**Ingresos ($) Li - Ls**	**Ingresos ($) Lri - Lrs**	**Marcas de clase**	**Frecuencias absolutas**
1	450 – 455	449.5 - 455.5	452.5	16
2	456 – 461	455.5 - 461.5	458.5	29
3	462 – 467	461.5 - 467.5	464.5	26
4	468 – 473	467.5 - 473.5	470.5	13
5	474 - 479	473.5 - 479.5	476.5	9
6	480 - 485	479.5 - 485.5	482.5	7
				100

8.- Al igual que para el caso de una variable discreta, podríamos presentar la ***distribución de frecuencias relativas*** de los ingresos mensuales de las 100 familias pobres con la finalidad de poner en evidencia la parte del total de las observaciones que pertenece a cada una de las clases.

TABLA 2.10 Incorporación de las frecuencias relativas de cada clase

Clases	**Ingresos ($) Li - Ls**	**Ingresos($) Lri - Lrs**	**Marcas de clase**	**Frecuencias absolutas**	**Frecuencias relativas**
1	450 – 455	449.5 - 455.5	452.5	16	0.16
2	456 – 461	455.5 - 461.5	458.5	29	0.29
3	462 – 467	461.5 - 467.5	464.5	26	0.26
4	468 – 473	467.5 - 473.5	470.5	13	0.13
5	474 - 479	473.5 - 479.5	476.5	9	0.09
6	480 - 485	479.5 - 485.5	482.5	7	0.07
				100	1

2.3.1 Representación gráfica de una distribución de frecuencia

Una manera eficiente de resumir aún más los datos y obtener de forma rápida información a partir de ellos es mediante un gráfico. Existen tres tipos de gráficos que

pueden ser utilizados para representar de una manera adecuada el comportamiento de una distribución de frecuencias. Estos son:

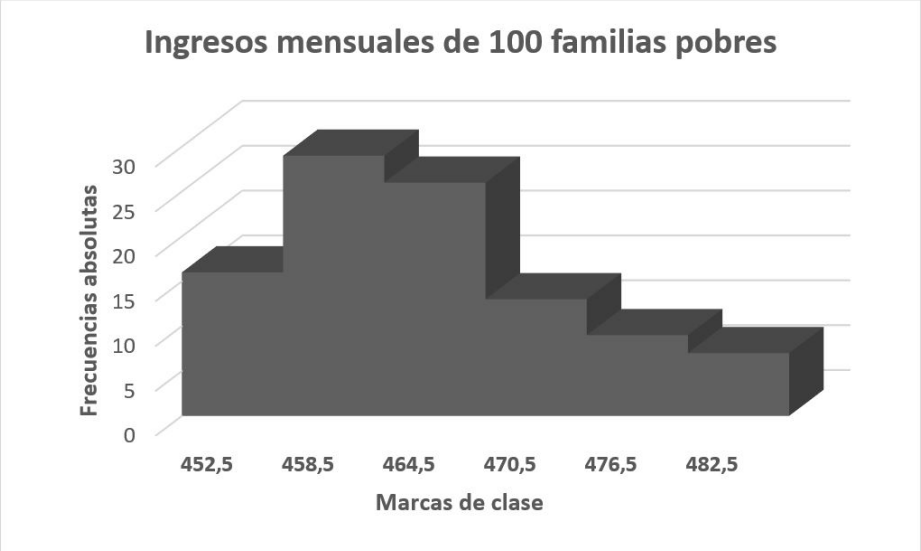
a. **El histograma.**
b. **El polígono de frecuencias.**
c. **El polígono de frecuencias relativas.**

En principio, un *histograma* es un gráfico de columnas muy similar al utilizado para representar a una tabla de frecuencias para variable discreta. Sin embargo, existe una diferencia fundamental entre ambos gráficos debido a la naturaleza de los datos originales. El *histograma* representa de forma gráfica a una distribución de variable continua, y por tanto, en el eje horizontal del gráfico las columnas son adyacentes con la finalidad de indicar la naturaleza continua de los datos. Al igual que el gráfico de columnas para variable discreta, las *marcas de clase* se ubican en el eje horizontal y las *frecuencias absolutas* en el eje vertical.

La figura 2.3 muestra el *histograma* correspondiente a la distribución de frecuencias de los ingresos mensuales de las 100 familias pobres.

Observe que hemos utilizado como etiquetas del eje horizontal los puntos medios de clase.

FIGURA 2.3 Histograma de los ingresos mensuales de 100 familias pobres

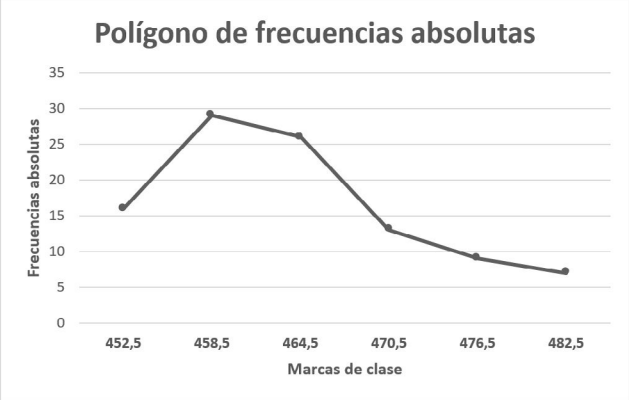


Al igual que el *histograma*, un *polígono de frecuencias* muestra el comportamiento de una distribución y consiste en segmentos de recta que unen los pares ordenados formados por las intersecciones de las marcas de clase con las correspondientes frecuencias.

La figura 2.4 muestra el *polígono de frecuencias* correspondiente al ingreso

mensual de las 100 familias pobres que hemos venido estudiando.

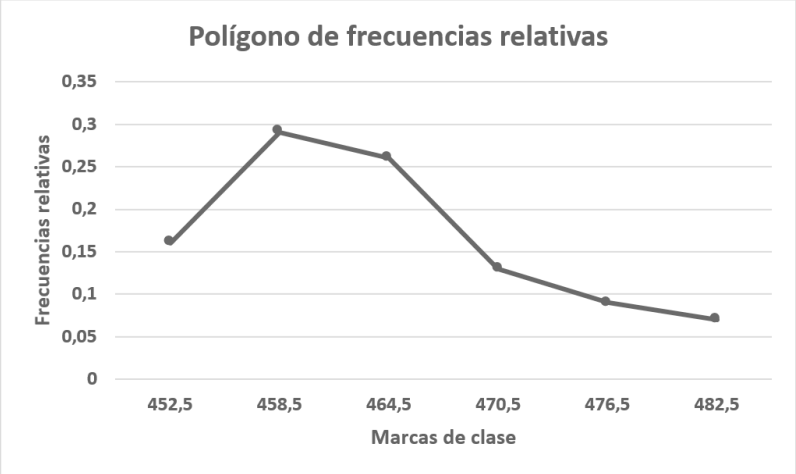
FIGURA 2.4 Polígono de frecuencias absolutas



Otra forma alternativa de presentar un polígono es utilizando las *frecuencias relativas*. Este tipo de representación gráfica recibe el nombre de *polígono de frecuencias relativas*. Por supuesto, los gráfico del *polígono de frecuencias* y del *polígono de frecuencias relativas* son exactamente iguales. La única diferencia consiste en que en el primero, el eje vertical está formado por las *frecuencias absolutas,* y en el segundo por las *frecuencias relativas***.**

Un gráfico de este tipo se muestra en la figura 2.5

FIGURA 2.5 Polígono de frecuencias relativas



2.4 Distribución de frecuencias acumuladas

Si tuviéramos un interés particular en conocer, por ejemplo, el número de familias pobres con ingresos mensuales inferiores a 468 dólares, nos veríamos obligados a sumar las frecuencias correspondientes a las clases *450 – 455, 456 – 461 y 462 – 467*, es decir, hallar la suma de *16+29+26 = 71*. Concluimos entonces que 71 familias tienen ingresos mensuales inferiores a 468 dólares.

Una alternativa más general y útil para situaciones similares a la anterior, consiste en calcular las *frecuencias absolutas acumuladas*.

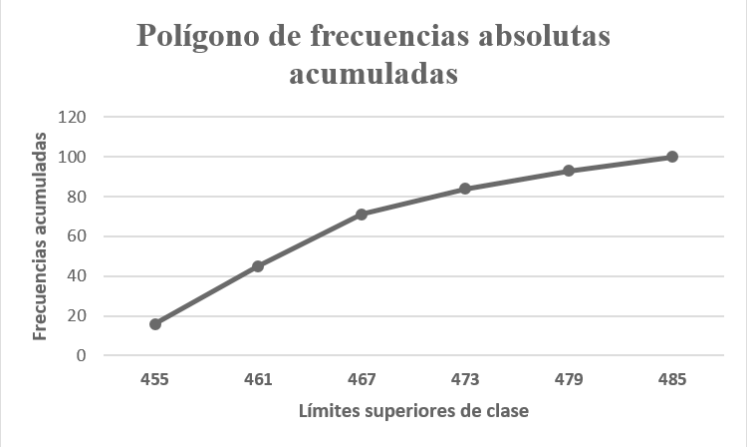
Para concluir con la construcción de la tabla de distribución de frecuencias debemos calcular de igual modo las *frecuencias relativas acumuladas*.

Si representamos por X_i las marcas de clase, por f_i las frecuencias absolutas, por h_i las frecuencias relativas, por F_i las frecuencias absolutas acumuladas y por H_i las frecuencias relativas acumuladas obtenemos la tabla 2.11.

TABLA 2.11 Columnas de frecuencias absolutas y relativas acumuladas

Clases	Ingresos ($) Li - Ls	Ingresos ($) Lri - Lrs	X_i	f_i	h_i	F_i	H_i
1	450 – 455	449.5 - 455.5	452.5	16	0.16	16	0.16
2	456 – 461	455.5 - 461.5	458.5	29	0.29	45	0.45
3	462 – 467	461.5 - 467.5	464.5	26	0.26	71	0.71
4	468 – 473	467.5 - 473.5	470.5	13	0.13	84	0.84
5	474 - 479	473.5 - 479.5	476.5	9	0.09	93	0.93
6	480 - 485	479.5 - 485.5	482.5	7	0.07	100	0.93
				100	1	100	1

FIGURA 2.6 Polígono de frecuencias absolutas acumuladas



Observe que los valores utilizados en el eje horizontal corresponden a los límites superiores de cada una de las clases.

Con el objetivo de precisar lo que hemos estudiado desarrollemos un ejemplo de la razón de precio – ganancia de una emisión de acciones. Los datos se muestran en la tabla 2.12.

TABLA 2.12 Razón precio – ganancia de una emisión de acciones

4.65	6.90	8.64	5.47	6.07	6.48	8.72	9.05	5.85	4.75
8.96	7.23	4.83	5.88	7.62	5.67	9.00	5.60	7.64	8.82
5.64	4.00	7.63	6.81	7.49	4.56	7.16	8.61	3.86	6.78
9.02	8.65	6.72	6.95	7.90	6.65	7.25	6.26	6.43	6.67
7.52	6.68	7.98	7.10	7.78	7.17	8.06	6.66	7.74	6.67
6.25	7.63	6.73	7.60	8.14	7.12	7.82	6.86	7.75	7.36

1.- Si estudia la tabla anterior concluirá que el valor más bajo es *3.86* y el más alto *9.05*, por tanto el rango es $R = 9.05 - 3.86 = 5.19$.

2.- Establezcamos el número de *clases o intervalos* que debemos utilizar.

- En nuestro ejemplo n = 60, y por tanto el número de clases debe ser igual a *6* ya que $2^5 = 32 < 60$, pero $2^6 = 64 > 60$.
- Utilizando la *regla de Sturges*:

c = 1+3.322 log n,

c = 1+3.322 log 60 = 1+3.322(1.78) = 6.91

Para ser conservadores debemos utilizar 7 clases.

3.- Determinemos la *longitud de cada intervalo o ancho de clase*.

$$\frac{5.19}{7} = 0.741$$

En nuestro ejemplo u = 0.01, por tanto la *longitud de cada intervalo* debe ser igual a 0.75.

4.- Para la primera clase se coloca como límite inferior (Li) el valor más pequeño de los datos y como límite superior Ls = Li + (longitud del intervalo – 0.01), es decir, Li = 3.86 y Ls = 3.86 + (0.75 – 0.01) = 4.60.

Para obtener el límite inferior de la segunda clase se le suma 0.01 al límite superior de la primera clase, esto es 4.60 + 0.01 (recuerde que 0.01 es la unidad más pequeña de los datos) y al resultado 4.61 se le suman (0.75 – 0.01) unidades para obtener el límite superior de la segunda clase, es decir, 4.61 + 0.74 = 5.35.

Este proceso se sigue repitiendo hasta completar los 7 intervalos requeridos.

Los intervalos se muestran en la tabla 2.13:

TABLA 2.13 Límites de clase

Clases	**Ingresos Li - Ls**
1	3.86 – 4.60
2	4.61 – 5.35
3	5.36 – 6.10
4	6.11 – 6.85
5	6.86 – 7.60
6	7.61 – 8.35
7	8.36 – 9.10

5.- Obtengamos los *límites reales de clase*. Recuerden que estos límites se obtie-

nen restándole media unidad (u/2) a los límites inferiores de clase y sumándole esa misma cantidad a los límites superiores.

En nuestro ejemplo u = 0.01, por tanto, el *límite real inferior (Lri)* de la primera clase es *3.86 – 0.005 = 3.855* y el *límite real superior (Lrs)* de esta clase es *4.60 + 0.005 = 4.605*.

Este procedimiento se continúa hasta agotar todas las clases, obteniendo los resultados que se muestran en la tabla 2.14.

TABLA 2.14 Límites de clase y límites reales de clase

Clases	**Ingresos ($)** **Li - Ls**	**Ingresos ($)** **Lri - Lrs**
1	3.86 – 4.60	3.855 – 4.605
2	4.61 – 5.35	4.605 – 5.355
3	5.36 – 6.10	5.355 – 6.105
4	6.11 – 6.85	6.105 – 6.855
5	6.86 – 7.60	6.855 – 7.605
6	7.61 – 8.35	7.605 – 8.355
7	8.36 – 9.10	8.355 – 9.105

6.- Determinemos la *marca de clase o punto medio* de cada uno de los intervalos. Los resultados se muestran en la tabla 2.15.

TABLA 2.15 Incorporación de la columna marcas de clase

Clases	**Ingresos ($) Li - Ls**	**Ingresos ($) Lri - Lrs**	**Marcas de clase**
1	3.86 – 4.60	3.855 – 4.605	4.23
2	4.61 – 5.35	4.605 – 5.355	4.98
3	5.36 – 6.10	5.355 – 6.105	5.73
4	6.11 – 6.85	6.105 – 6.855	6.48
5	6.86 – 7.60	6.855 – 7.605	7.23
6	7.61 – 8.35	7.605 – 8.355	7.98
7	8.36 – 9.10	8.355 – 9.105	8.73

7.- Obtengamos las *frecuencias absolutas* de cada clase. Los resultados se muestran en la tabla 2.16.

TABLA 2.16 Incorporación de la columna frecuencias absolutas

Clases	Ingresos ($) Li - Ls	Ingresos ($) Lri - Lrs	Marcas de clase	Frecuencias absolutas
1	3.86 – 4.60	3.855 – 4.605	4.23	3
2	4.61 – 5.35	4.605 – 5.355	4.98	3
3	5.36 – 6.10	5.355– 6.105	5.73	7
4	6.11 – 6.85	6.105 – 6.855	6.48	13
5	6.86 – 7.60	6.855 – 7.605	7.23	13
6	7.61 – 8.35	7.605 – 8.355	7.98	12
7	8.36 – 9.10	8.355 – 9.105	8.73	9
				60

8.- Calculemos las *frecuencias relativas*, las *frecuencias absolutas acumuladas* y las *frecuencias relativas acumuladas*. Los resultados se muestran en la tabla 2.17.

TABLA 2.17 Columnas frecuencias absolutas y relativas acumuladas

Clases	Ingresos Li - Ls	Ingresos Lri - Lrs	X_i	f_i	h_i	F_i	H_1
1	3.86 – 4.60	3.855 – 4.605	4.23	3	0.05	3	0.05
2	4.61 – 5.35	4.605 – 5.355	4.98	3	0.05	6	0.10
3	5.36 – 6.10	– 6.105	5.73	7	0.12	13	0.22
4	6.11 – 6.85	6.105 – 6.855	6.48	13	0.22	26	0.43
5	6.86 – 7.60	6.855 – 7.605	7.23	13	0.22	39	0.65
6	7.61 – 8.35	7.605 – 8.355	7.98	12	0.20	51	0.85
7	8.36 – 9.10	8.355 – 9.105	8.73	9	0.15	60	1
				60	1		

En las figuras 2.7 y 2.8 se muestran el *histograma* y el *polígono de frecuencias* correspondiente.

FIGURA 2.7 Histograma de frecuencias

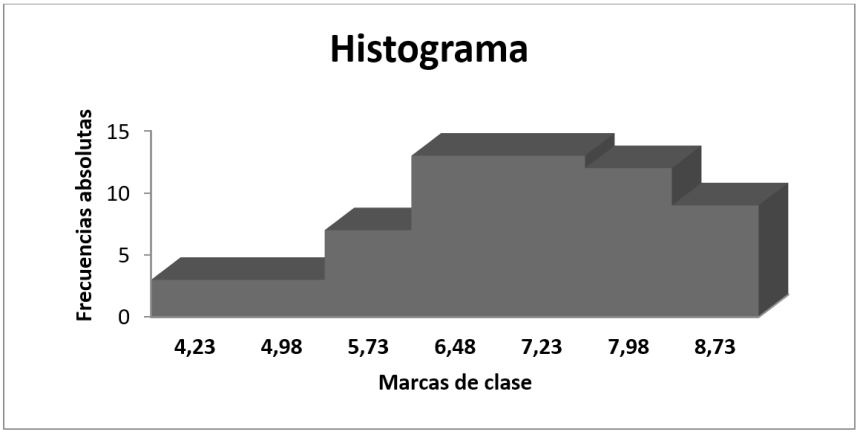
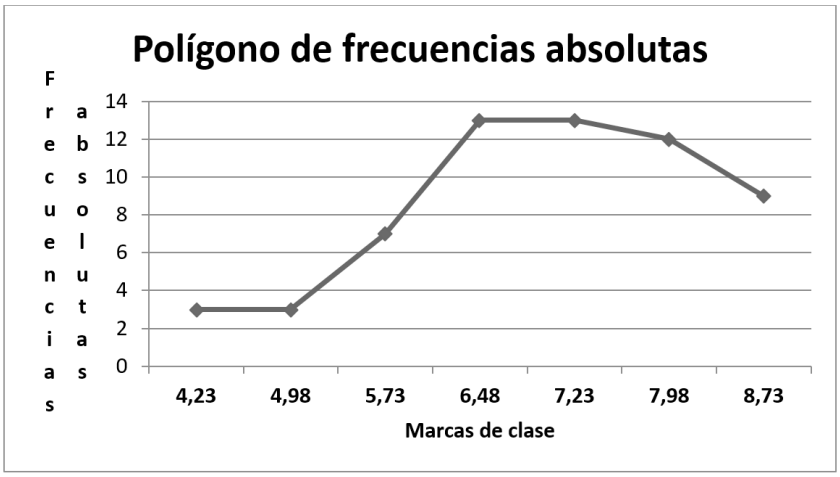


FIGURA 2.8 Polígono de frecuencias


Polígono de frecuencias absolutas
Frecuencias absolutas
14
12
10
8
6
4
2
0
4,23
4,98
5,73
6,48
7,23
7,98
8,73
Marcas de clase

Ejercicios del capítulo

2.1 Los datos que se muestran a continuación representan las calificaciones obtenidas por 50 estudiantes en un examen final de Estadística.

8.4	7.6	9.2	10	9.7	8.8	7.5	7.9	8.1	9.1
10	7.5	7.3	8.4	8.8	9.4	9.6	8.2	8.7	9
8	7.7	7.1	7	8.3	9	8.1	9.6	10	9
7.4	7.5	8.8	9.8	9.1	10	9	7.1	7.3	7
7.3	8.5	8.2	9	8	10	8.1	7.7	9	8.9

Haciendo uso de esta información, elabore una distribución de frecuencias con cuatro intervalos cumpliendo los siguientes pasos:

a. Determine el rango o amplitud de las observaciones.
b. Determine la longitud de cada intervalo o ancho de clase.
c. Obtenga el límite inferior y superior de cada clase.
d. Determine los límites reales de clase.
e. Obtenga la marca de clase o punto medio de cada uno de los intervalos.
f. Determine las frecuencias absolutas y relativas de cada clase.
g. Obtenga las frecuencias absolutas y relativas acumuladas de cada clase.

2.2 A continuación se muestra el número de minutos promedio diarios de utilización del servicio de internet de 60 clientes de la Corporación Nacional de Telecomunicaciones.

64	56	72	80	77	68	55	59	61	71
110	85	83	94	98	104	106	92	97	100
75	72	66	65	78	85	76	91	95	85
89	90	103	113	106	114	105	86	88	85
80	92	89	79	87	107	88	84	97	96
73	70	64	82	90	110	91	90	94	100

Elabore una distribución de frecuencias con cinco intervalos cumpliendo los siguientes pasos:

a. Determine el rango o amplitud de las observaciones.
b. Determine la longitud de cada intervalo o ancho de clase.
c. Obtenga el límite inferior y superior de cada clase.
d. Determine los límites reales de clase.
e. Obtenga la marca de clase o punto medio de cada uno de los intervalos.
f. Determine las frecuencias absolutas y relativas de cada clase.
g. Obtenga las frecuencias absolutas y relativas acumuladas de cada clase.

2.3 Con los datos del ejercicio 2.1 elabore los siguientes gráficos:

a) Histograma.
b) Polígono de frecuencias absolutas.
c) Polígono de frecuencias relativas.
d) Polígono de frecuencias absolutas acumuladas.

2.4 Con los datos del ejercicio 2.2 elabore los siguientes gráficos:

a) Histograma.
b) Polígono de frecuencias absolutas.
c) Polígono de frecuencias relativas.
d) Polígono de frecuencias absolutas acumuladas.

2.5 Los datos que se aprecian a continuación representan el porcentaje del sueldo de cien personas de clase baja con relación a la remuneración básica actual.

135.24	137.35	139.12	139.71	142.65	144.41	144.69	146.76	148.53	148.74
135.98	135.41	139.41	141.47	143.24	144.12	146.17	146.47	147.35	150.88
137.06	138.24	140.29	139.71	141.86	145.29	148.88	147.06	148.24	149.12
134.71	138.53	139.41	142.06	142.94	142.35	147.06	146.47	149.12	151.23
136.76	136.18	139.71	138.53	142.35	145.59	144.71	148.24	148.82	149.41
138.24	139.41	138.24	140.02	143.53	142.35	147.35	147.65	148.53	149.71
135.24	137.06	139.12	142.06	141.76	143.28	146.18	147.94	146.47	148.82
135.59	139.71	139.54	142.94	140.29	144.41	144.71	148.24	148.94	149.12
137.35	136.47	140.29	140.56	143.82	144.41	147.06	145.29	147.94	149.71
136.76	137.65	138.53	140.88	141.18	143.24	146.18	147.06	148.24	147.94

Elabore una distribución de frecuencias con seis intervalos cumpliendo los siguientes pasos:

a) Determine el rango o amplitud de las observaciones.
b) Determine la longitud de cada intervalo o ancho de clase.
c) Obtenga el límite inferior y superior de cada clase.
d) Determine los límites reales de clase.
e) Obtenga la marca de clase o punto medio de cada uno de los intervalos.
f) Determine las frecuencias absolutas y relativas de cada clase.
g) Obtenga las frecuencias absolutas y relativas acumuladas de cada clase.

2.6 Con los datos que se muestran a continuación elabore una tabla de distribución de frecuencias con cinco intervalos cumplimentando los pasos que se detallan:

1.07	1.63	1.25	1.33	1.28	1.13	0.92	0.98	1.02	1.18
1.83	1.42	1.38	1.57	1.63	1.73	1.77	1.53	1.62	1.67
1.25	1.21	1.17	1.08	1.36	1.42	1.27	1.52	1.58	1.42
1.48	1.56	1.72	1.88	1.77	1.91	1.75	1.43	1.47	1.42
1.33	1.53	1.48	1.32	1.45	1.78	1.47	1.42	1.62	1.61
1.22	1.17	1.07	1.37	1.53	1.83	1.52	1.51	1.57	1.67

a. Determine el rango o amplitud de las observaciones.
b. Determine la longitud de cada intervalo o ancho de clase.
c. Obtenga el límite inferior y superior de cada clase.
d. Determine los límites reales de clase.
e. Obtenga la marca de clase o punto medio de cada uno de los intervalos.
f. Determine las frecuencias absolutas y relativas de cada clase.
g. Obtenga las frecuencias absolutas y relativas acumuladas de cada clase.

2.7 Con los datos del ejercicio 2.5 elabore los siguientes gráficos:
a. Histograma.
b. Polígono de frecuencias absolutas.
c. Polígono de frecuencias relativas.
d. Polígono de frecuencias absolutas acumuladas.

2.8 Con los datos del ejercicio 2.6 elabore los siguientes gráficos:
a. Histograma.
b. Polígono de frecuencias absolutas.
c. Polígono de frecuencias relativas.
d. Polígono de frecuencias absolutas acumuladas.

Capítulo 3

Medidas de tendencia central y de variación

El problema

Se suele escuchar en los noticieros y también leer en la prensa que la temperatura promedio en una determinada ciudad es de 34 grados centígrados en época de verano y de 8 grados durante el invierno, y que además el nivel de humedad es altamente variable durante todo el año. ¿Existen métodos estadísticos que permiten calcular estos indicadores de temperatura promedio y de variabilidad de los niveles de humedad?

3.1 Introducción

En el capítulo anterior, y como parte del estudio de la estadística descriptiva, vimos la forma en que un conjunto de datos cuantitativos podían ser organizados, agrupados y graficados con el objetivo de describir el comportamiento de su distribución y poner al descubierto alrededor de qué valor se concentra la mayor parte de los mismos.

En el presente capítulo abordaremos el estudio de dos cantidades numéricas que nos permiten describir el comportamiento de un conjunto de datos cuantitativos. Estas cantidades numéricas a las cuales hacemos referencia son las *medidas de tendencia central* y las *medidas de variación*.

Durante el desarrollo del capítulo estudiaremos las siguientes *medidas de tendencia central y de variación*:

- **Medidas de tendencia central.**

5. La media aritmética.
6. La mediana.
7. La moda.

- **Medidas de variación.**

1. El alcance.
2. La varianza.
3. La desviación estándar.
4. El coeficiente de variación.

Las medidas de tendencia central son también conocidas como *medidas de posición* y las medidas de variación como *medidas de dispersión*.

Las medidas de posición y dispersión pueden ser calculadas para los casos de *datos no agrupados* y *agrupados* en una distribución de frecuencias. Para ejemplificar

estos cálculos, en las tablas 3.1 y 3.2 se reportan los datos no agrupados y agrupados de los ingresos de las 100 familias pobres, utilizados en el capítulo anterior.

TABLA 3.1 Datos no agrupados de los ingresos mensuales de 100 familias

463	456	467	454	469	460	478	481	459	464
457	451	471	460	461	477	462	484	476	462
468	469	461	454	456	463	460	483	458	466
455	460	472	462	460	481	464	479	461	459
467	452	469	450	473	464	466	463	460	474
460	463	454	474	462	453	485	461	459	458
468	455	473	462	456	482	461	462	452	472
454	464	469	465	471	460	466	463	459	466
456	453	461	474	463	460	464	453	483	458
471	467	455	476	454	466	461	459	477	452

TABLA 3.2 Distribución de frecuencias de los ingresos mensuales

Clases	**Ingresos ($) Li - Ls**	**Ingresos ($) Lri - Lrs**	**Marcas de clase**	**Frecuencias absolutas**
1	450 – 455	449.5 - 455.5	452.5	16
2	456 – 461	455.5 - 461.5	458.5	29
3	462 – 467	461.5 - 467.5	464.5	26
4	468 – 473	467.5 - 473.5	470.5	13
5	474 - 479	473.5 - 479.5	476.5	9
6	480 - 485	479.5 - 485.5	482.5	7
				100

3.2 Medidas de tendencia central

3.2.1 La media aritmética

Uno de los estadígrafos más utilizados en trabajos relacionados con la inferencia estadística es la *media aritmética*, también conocida como *promedio*. Su valor numérico se denota mediante la expresión $\overline{X}$ y se lee X barra. La *media aritmética* de un conjunto de datos expresa en qué punto se encuentra el valor central de dicho conjunto.

La *media aritmética* de una muestra puede ser, y de hecho es utilizada, como una estimación del valor de la media aritmética de una población *(μ)*, lo cual la convierte en un elemento muy importante al momento de establecer una inferencia sobre el comportamiento de los datos en una población cualquiera. A continuación una descripción numérica de cómo calcular este indicador en los casos de datos no agrupados y agrupados.

Datos no agrupados

Sea el conjunto $X=\{X_1, X_2, X_3;....., X_n\}$.

La media aritmética de este conjunto se obtiene de la siguiente forma:

$$\overline{X}=\frac{\sum_{1}^{n} X_i}{n}=\frac{X_1+X_2+X_3+.....+X_n}{n}$$

En nuestro ejemplo:

$$\overline{X}=\frac{\sum_{1}^{n} X_i}{n}=\frac{453+456+457+.....+452}{100}=\frac{46413}{100}=464.13$$

Datos agrupados

La media aritmética para datos agrupados viene dada por la expresión:

$\overline{X}=\frac{\sum_{1}^{k} f_k \times x_k}{n}$, donde k es el número de clases, f_k la frecuencia absoluta de la clase k-ésima y x_k la marca de clase de la clase k-ésima.

En nuestro ejemplo:

$$\overline{X}=\frac{16(452.5)+29(458.5)+.....+7(482.5)}{100}=\frac{46396}{100}=463.96$$

Observe que existe una diferencia de 0.17 dólares entre la media aritmética para datos no agrupados y la media aritmética para datos agrupados, lo cual se explica por la pérdida de información provocada por el agrupamiento de los datos.

La *media poblacional* viene dada por la expresión $\mu=\frac{\sum_{1}^{N} X_i}{N}$.

3.2.1.1 Propiedades de la media aritmética

Propiedad 1:

Si a cada valor de un conjunto de datos $X=\{X_1, X_2, X_3;....., X_n\}$ se le suma (o se le resta) una constante k, la media aritmética del nuevo conjunto de datos es igual a la media aritmética del conjunto original más (o menos) la constante k.

Demostración:

Si $Y_i = X_i \pm k$ entonces $\overline{Y}=\overline{X}\pm k$

$$\overline{Y} = \sum_{1}^{n} \frac{(X_i \pm k)}{n} = \sum_{1}^{n} \frac{X_i}{n} \pm \sum_{1}^{n} \frac{k}{n} = \overline{X} \pm \frac{nk}{n} = \overline{X} \pm k$$

Propiedad 2:

Si cada valor de un conjunto $X = \{X_1, X_2, X_3; ....., X_n\}$ se multiplica (o se divide) por una constante k, entonces la media aritmética del nuevo conjunto es igual a la media aritmética del conjunto original multiplicada (o dividida) por la constante k.

Demostración:

Si $Y_i = k\, X_i$ entonces $\overline{Y} = k\, \overline{X}$

$$\overline{Y} = \sum_{1}^{n} \frac{k\, X_i}{n} = k \sum_{1}^{n} \frac{X_i}{n} = k\, \overline{X}$$

Propiedad 3:

La suma de las desviaciones de cada valor de un conjunto de datos $X = \{X_1, X_2, X_3; ....., X_n\}$ con respecto a su media aritmética es igual a cero.

Demostración:

$\sum_{1}^{n} (X_i - \overline{X}) = \sum_{1}^{n} X_i - \sum_{1}^{n} \overline{X} = n\,\overline{X} - n\,\overline{X} = 0$, ya que $\sum_{1}^{n} X_i = n\,\overline{X}$

3.2.2 La mediana

Una desventaja de la media aritmética como medida de tendencia central es su susceptibilidad de distorsión a causa de valores extremos en el conjunto de datos. Es decir, si en el conjunto al que se le pretende calcular la media aritmética existe una o más observaciones cuyos valores numéricos son mucho más altos o mucho más bajos que el resto, entonces el valor de esta media *no indicará* el centro del conjunto, sino que habrá un sesgo a la izquierda o a la derecha en función de cuáles son los valores extremos.

La *mediana* es una medida de posición que muestra el valor central de un conjunto de datos independientemente de que existan o no valores extremos.

En un conjunto de datos ordenado de forma ascendente o descendente la *mediana* se calcula de la siguiente manera:

- Si el conjunto tiene un número impar de observaciones la *mediana* es el valor central del conjunto.
- Si el conjunto tiene un número par de observaciones la *mediana* es el pro-

medio de los dos valores centrales.

Datos no agrupados

Los datos no agrupados del ejemplo que estamos desarrollando ordenados de forma ascendente se muestran en la tabla 3.3.

TABLA 3.3 Ingresos mensuales ordenados de forma ascendente

450	454	457	460	461	462	464	467	472	477
451	454	458	460	461	463	464	468	472	478
452	454	458	460	461	463	465	468	473	479
452	455	458	460	461	463	466	469	473	481
452	455	459	460	461	463	466	469	474	481
453	455	459	460	462	463	466	469	474	482
453	456	459	460	462	463	466	469	474	483
453	456	459	460	462	464	466	471	476	483
454	456	459	461	462	464	467	471	476	484
454	456	460	461	462	464	467	471	477	485

Por ser *n = 100,* un número par, la mediana será el promedio de las observaciones que ocupan las posiciones 50 y 51 en el ordenamiento, es decir, el promedio de *462 y 462*. Las posiciones 50 y 51 pueden ser obtenidas mediante la expresión:

$\frac{n+1}{2} = \frac{100+1}{2} = 50.5$, es decir, entre 50 y 51.

$Mediana = \frac{462+462}{2} = 462$

Datos agrupados

El primer paso para calcular la mediana en la *distribución de frecuencias* del ejemplo que nos ocupa, consiste en determinar en cuál de las seis clases se encuentra la mediana.

Como sabemos, la mediana está ubicada en la posición $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$.

Si estudia la tabla 3.4 podrá percatarse que la posición 50.5 (entre 50 y 51) se encuentra ubicada en la *tercera clase o intervalo*, la cual tiene una frecuencia acumulada igual a 71.

TABLA 3.4 Frecuencias acumuladas de los ingresos mensuales

Ingresos ($)	Frecuencias	Frecuencias acumuladas
450 – 455	16	16
456 – 461	29	45
462 – 467	26	71
468 – 473	13	84
474 - 479	9	93
480 - 485	7	100
	100	

Seguidamente determinemos a que observación de la clase mediana le corresponde la posición 50 y a cuál la posición 51. Observe que la suma de las frecuencias de las clases que están *por encima* de la clase mediana es 45, por tanto a la posición 50 de la clase mediana le corresponde la observación *50 – 45 = 5* y en consecuencia a la posición 51 le corresponde la observación *6* de la clase mediana.

El *ancho* de cada una de las posiciones de la clase mediana es igual a:

$$\frac{467-462+1}{71}=0.084$$

por tanto la posición *5* sería *462 + (5-1) x 0.084 = 462.336*, y en consecuencia, la posición *6* es *462.336 + 0.084 = 462.420*.

$$Mediana = \frac{462.336+462.420}{2}=462.378\approx 462$$

Observe que en este caso el valor de la mediana para datos no agrupados coincide con el valor de la mediana para datos agrupados.

Otro método

Otra forma de calcular la mediana de un conjunto de datos es mediante la expresión: $M_e=\left(\frac{\frac{(n+1)}{2}-(F+1)}{f_{M_e}}\right)w+L_{M_e}$, donde

n es el total de observaciones = 100

F es la suma de las frecuencias *hasta*, pero *sin incluir*, la clase mediana, es decir, 45.

$\mathbf{f_{Me}}$ es la frecuencia de la clase mediana = 26

w es el ancho del intervalo de clase = 6

$\mathbf{L_{Me}}$ es el límite inferior del intervalo de la clase mediana = 462

$$M_e = \left(\frac{\frac{(100+1)}{2} - (45+1)}{26} \right) 6 + 462 = 1.04 + 462 = 463.04 \approx 463$$

3.2.3 La Moda

La *Moda* de un conjunto de datos es el valor que más se repite en dicho conjunto, es decir, el valor que aparece con una mayor frecuencia. Según lo expresado, un conjunto de datos puede no tener *Moda* o tener una *Moda* o más de una.

Datos no agrupados

En la tabla 3.5 se muestran los ingresos mensuales y sus frecuencias.

Como se puede apreciar en la tabla 3.5, el valor con mayor frecuencia de aparición es 460 (9 frecuencias), por lo tanto el valor de la Moda es 460.

TABLA 3.5 Ingresos mensuales y sus correspondientes frecuencias

Valor	f	Valor	f	Valor	f	Valor	f
450	1	459	5	468	2	477	2
451	1	460	9	469	4	478	1
452	3	461	7	470	0	479	1
453	3	462	6	471	3	480	0
454	5	463	6	472	2	481	2
455	3	464	5	473	2	482	1
456	4	465	1	474	3	483	2
457	1	466	5	475	0	484	1
458	3	467	3	476	2	485	1

Datos agrupados

Cuando los datos están agrupados en una distribución de frecuencias es razonable pensar que la *Moda* está ubicada en la clase con una mayor frecuencia. A partir de esta *clase modal* podemos calcular el valor de la *Moda* utilizando la siguiente expresión:

$$M_0 = L_{M_0} + \left(\frac{d_1}{d_1 + d_2} \right) w \text{ , donde}$$

$\mathbf{M_0}$ es la *Moda* y $\mathbf{L_{Mo}}$ es el límite inferior de la ***clase modal***.

$\mathbf{d_1}$ es la frecuencia de la *clase modal* menos la frecuencia de la clase que se encuentra inmediatamente por debajo de ésta.

d_2 es la frecuencia de la *clase modal* menos la frecuencia de la clase que se encuentra inmediatamente por encima de ésta.

w es el ancho del intervalo de la clase modal.

Según los datos que se observan en la tabla 3.6, la *clase modal* es la tercera y adicionalmente:

$L_{Mo} = 456$

$d_1 = 29 - 26 = 3$

$d_2 = 29 - 16 = 13$

$w = 6$

Por tanto, $M_0 = 456 + \left(\frac{3}{3+13}\right)6 = 456 + 1.12 = 457.12 \approx 457$

TABLA 3.6 Determinación de la frecuencia modal

Clases	Ingresos ($) Li - Ls	Ingresos ($) Lri - Lrs	Marcas de clase	Frecuencias absolutas
1	450 – 455	449.5 - 455.5	452.5	16
2	456 – 461	455.5 - 461.5	458.5	29
3	462 – 467	461.5 - 467.5	464.5	26
4	468 – 473	467.5 - 473.5	470.5	13
5	474 - 479	473.5 - 479.5	476.5	9
6	480 - 485	479.5 - 485.5	482.5	7

Observe que existe una diferencia de aproximadamente 3 dólares entre la Moda para datos no agrupados y la Moda para datos agrupados, lo cual se explica por la pérdida de información provocada por el agrupamiento de los datos.

3.3 Medidas de variación

A continuación comenzaremos el estudio de algunos indicadores que expresan el grado de *variación (dispersión o variabilidad)* que tiene un conjunto cualquiera de datos, los cuales tienen un valor inestimable en la aplicación de algunos métodos de la Estadística Inferencial, y que además, calculados a nivel de una muestra, son utilizados como una estimación del comportamiento de los mismos a nivel de toda la población.

Las principales *medidas de variación* son el *alcance*, la *varianza*, la *desviación estándar* y el *coeficiente de variación*.

3.3.1 El Alcance
Datos no agrupados

El *alcance* de un conjunto de datos $X = \{X_1, X_2, X_3; ....., X_n\}$ se define como

la *diferencia* entre el mayor y el menor de los datos del conjunto, es decir:

Alcance = valor mayor – valor menor
En nuestro ejemplo, *Alcance = 485 – 450 = 35.*

La utilidad práctica de esta *medida de variación* es limitada, no obstante, en un capítulo posterior estudiaremos una situación en la que el *Alcance* resultará de mucha importancia.

3.3.2 La Varianza
Datos no agrupados

La *varianza* S^2 de un conjunto de datos $X=\{X_1, X_2, X_3;....., X_n\}$ se define como:

$$S^2=\frac{\sum_1^n\left(X_i-\overline{X}\right)^2}{n-1}.$$

Observe que en la medida en que están más dispersos (o menos dispersos) los datos alrededor de su media aritmética, mayor (o menor) será el valor numérico de la varianza.

Demostremos a continuación, que otra forma de calcular la varianza es mediante la expresión:

$$S^2=\frac{\sum_1^n X_i^2-\frac{\left(\sum_1^n X_i\right)^2}{n}}{n-1}$$

$$S^2=\frac{\sum_1^n\left(X_i-\overline{X}\right)^2}{n-1}=\frac{\sum_1^n\left(X_i^2-2X_i\overline{X}+\overline{X}^2\right)}{n-1}=$$

$$\frac{\sum_1^n X_i^2-2\overline{X}\sum_1^n X_i+\sum_1^n\overline{X}^2}{n-1}=\frac{\sum_1^n X_i^2-2\overline{X}\left(n\overline{X}\right)+n\overline{X}^2}{n-1}=\frac{\sum_1^n X_i^2-n\overline{X}^2}{n-1}=$$

$$\frac{\sum_1^n X_i^2-n\left(\frac{\sum_1^n X_i}{n}\right)^2}{n-1}=\frac{\sum_1^n X_i^2-\frac{\left(\sum_1^n X_i\right)^2}{n}}{n-1}$$

En nuestro ejemplo:

$$S^2 = \frac{\left(463^2 + 456^2 + \ldots\ldots\ldots + 452^2\right) - \frac{\left(463 + 456 + \ldots\ldots\ldots + 452\right)^2}{100}}{9} =$$

$$\frac{21548771 - \frac{2154166569}{100}}{99} = \frac{7105.31}{99} = 71.77$$

La *varianza poblacional* se denota como σ^2 y viene dada por la expresión:

$$\sigma^2 = \sum_1^N \frac{\left(X_i - \overline{X}\right)^2}{N} = \frac{\sum_1^N X_i^2 - \frac{\left(\sum_1^N X_i\right)^2}{N}}{N}$$

Datos agrupados

La varianza muestral para datos agrupados en una distribución de frecuencias viene dada por la expresión:

$S^2 = \frac{\sum_1^n f_i \left(X_i - \overline{X}\right)^2}{n-1}$, donde f_i es la frecuencia de la clase i-ésima, X_i es el punto medio de la clase i-ésima y n el número de observaciones.

$$S^2 = \frac{16(452.5 - 464.13)^2 + 29(458.5 - 464.13)^2 + \ldots. + 7(482.5 - 464.13)^2}{99}$$

$$= \frac{7353.73}{99} = 74.28$$

De igual manera la *varianza poblacional* de datos agrupados viene dada por:

$$\sigma^2 = \frac{\sum_1^N f_i \left(X_i - \mu\right)^2}{N}$$

3.3.2.1 Propiedades de la Varianza

Propiedad 1:

La varianza de un conjunto de valores es siempre un número no negativo.

Esta propiedad significa que para cualquier conjunto de datos $S^2 \geq 0$, lo cual

puede ser fácilmente verificado si observamos que el numerador de la varianza es siempre un número no negativo, ya que es el resultado de una suma de cuadrados. Por supuesto, el denominador también es una cantidad no negativa.

La propiedad anterior implica también que $\sigma^2 \geq 0$, es decir, la propiedad es válida tanto para la varianza muestral como para la poblacional.

Propiedad 2:

La varianza de un conjunto de valores $X = \{X_1, X_2, X_3; ....., X_n\}$ todos iguales, es igual a cero.

Demostración:

Si cualquiera sea i, X_i = k entonces $\overline{X} = \frac{\sum_1^n X_i}{n} = \frac{n\,k}{n} = k$ y por tanto,

$$S^2 = \frac{\sum_1^n \left(X_i - \overline{X}\right)^2}{n-1} = \frac{\sum_1^n (k-k)^2}{n-1} = \frac{0}{n-1} = 0$$

Propiedad 3:

Si a cada valor de un conjunto $X = \{X_1, X_2, X_3; ....., X_n\}$ se le suma (o se le resta) una constante k > 0, la varianza del nuevo conjunto de valores es igual a la varianza del conjunto original.

Demostración:

La *propiedad 1* de la media aritmética establece que si $Y_i = X_i \pm k$ entonces $\overline{Y} = \overline{X} \pm k$ y por tanto:

$$S_Y^2 = \frac{\sum_1^n \left(Y_i - \overline{Y}\right)^2}{n-1} = \frac{\sum_1^n \left[X_i \pm k - \left(\overline{X} \pm k\right)\right]^2}{n-1} = \frac{\sum_1^n \left(X_i - \overline{X}\right)}{n-1} = S_X^2$$

Propiedad 4:

Si cada valor de un conjunto $X = \{X_1, X_2, X_3; ....., X_n\}$ se multiplica (o divide) por una constante k, la varianza del nuevo conjunto es igual a la varianza del conjunto original multiplicada (o dividida) por el cuadrado de la constante.

Demostración:

Si c = k, o $c = \frac{1}{k}$ y $Y_i = c\, X_i$, entonces:

Según la *propiedad 2* de la media, $\overline{Y} = c\,\overline{X}$ y

$$S_Y^2 = \frac{\sum_1^n \left(Y_i - \overline{Y}\right)}{n-1} = \frac{\sum_1^n \left(c\,X_i - c\,\overline{X}\right)^2}{n-1} = \frac{c^2 \sum_1^n \left(X_i - \overline{X}\right)^2}{n-1} = c^2\, S_X^2$$

3.3.3 La desviación estándar

La *desviación estándar* de un conjunto $X = \{X_1, X_2, X_3; ....., X_n\}$ se define como la raíz cuadrada positiva de la varianza.

Es decir, $S = \sqrt{S^2}$

En el ejemplo que estamos desarrollando:

$S = \sqrt{71.77} = 8.47$

De igual manera la *desviación estándar poblacional* es $\sigma = \sqrt{\sigma^2}$.

3.3.4 El coeficiente de variación

Se define el *coeficiente de variación* de un conjunto de observaciones $X = \{X_1, X_2, X_3; ....., X_n\}$ como el cociente entre la desviación estándar y la media de dicho conjunto. Es usual que el *coeficiente de variación* se exprese en términos porcentuales, para lo cual debe ser multiplicado por 100, es decir,

$$C.V.(X) = \frac{S}{\overline{X}}\, x\ 100.$$

En nuestro ejemplo $C.V.(X) = \frac{8.47}{464.45} x100 = 1.82\%.$

3.4 Ejercicio resumen

Calculemos las medidas de tendencia central y de variación para los datos no agrupados y agrupados de la razón de precio – ganancia de una emisión de acciones reportados en el capítulo anterior.

TABLA 3.7 Razón precio – ganancia de una emisión de acciones (datos no agrupados)

4.65	6.90	8.64	5.47	6.07	6.48	8.72	9.05	5.85	4.75
8.96	7.23	4.83	5.88	7.62	5.67	9.00	5.60	7.64	8.82
5.64	4.00	7.63	6.81	7.49	4.56	7.16	8.61	3.86	6.78
9.02	8.65	6.72	6.95	7.90	6.65	7.25	6.26	6.43	6.67
7.52	6.68	7.98	7.10	7.78	7.17	8.06	6.66	7.74	6.67
6.25	7.63	6.73	7.60	8.14	7.12	7.82	6.86	7.75	7.36

1.- Media aritmética.

$$\overline{X} = \frac{4.65 + 6.90 + \ldots\ldots\ldots + 7.36}{60} = \frac{419.47}{60} = 6.99$$

2.- Mediana.

- Número par de observaciones.
- Si ordena los datos podrá comprobar que los valores centrales son 7.10 y 7.12.

- $M_e = \frac{7.10 + 7.12}{2} = 7.11$.

3.- Moda.

En la tabla 3.8 aparecen en orden creciente los 60 datos correspondientes a la razón de precio – ganancia con sus correspondientes frecuencias de aparición.

Como se puede apreciar en la tabla el conjunto tiene dos modas, 6.67 y 7.63, ambos valores con frecuencia 2.

4.- Alcance.

Alcance = 9.05 – 3.86 = 5.19

TABLA 3.8 Razón precio – ganancia ordenadas de forma ascendente

Valor	f	Valor	f	Valor	f	Valor	f	Valor	f
3.86	1	6.07	1	6.81	1	7.52	1	8.14	1
4.00	1	6.25	1	6.84	1	7.60	1	8.61	1
4.56	1	6.26	1	6.90	1	7.62	1	8.64	1
4.65	1	6.43	1	6.95	1	7.63	2	8.65	1
4.75	1	6.48	1	7.10	1	7.64	1	8.72	1
4.83	1	6.65	1	7.12	1	7.74	1	8.82	1
5.47	1	6.66	1	7.16	1	7.75	1	8.96	1
5.60	1	6.67	2	7.17	1	7.78	1	9.00	1
5.64	1	6.68	1	7.23	1	7.82	1	9.02	1
5.67	1	6.72	1	7.25	1	7.90	1	9.05	1
5,85	1	6.73	1	7.36	1	7.98	1		
5.88	1	6.78	1	7.49	1	8.06	1		

5.- Varianza.

$$S^2 = \frac{(4.65)^2 + (6.90)^2 + \ldots\ldots\ldots + (7.36)^2 - \dfrac{(4.65 + 6.90 + \ldots\ldots\ldots + 7.36)^2}{60}}{59}$$

$$= \frac{3026.63 - \dfrac{(419.47)^2}{60}}{59} = 1.59$$

6.- Desviación estándar.

$$S = \sqrt{1.59} = 1.26$$

7.- Coeficiente de variación.

$$CV = \frac{1.26}{6.99} x100 = 18.03\,\%$$

TABLA 3.9 Distribución de frecuencias de la razón precio-ganancia (datos agrupados)

Clases	Ingresos Li - Ls	Ingresos Lri - Lrs	X_i	f_i	h_i	F_i	H_1
1	3.86 – 4.60	3.855 – 4.605	4.23	3	0.05	3	0.05
2	4.61 – 5.35	4.605 – 5.355	4.98	3	0.05	6	0.10
3	5.36 – 6.10	– 6.105	5.73	7	0.12	13	0.22
4	6.11 – 6.85	6.105 – 6.855	6.48	13	0.22	26	0.43
5	6.86 – 7.60	6.855 – 7.605	7.23	13	0.22	39	0.65
6	7.61 – 8.35	7.605 – 8.355	7.98	12	0.20	51	0.85
7	8.36 – 9.10	8.355 – 9.105	8.73	9	0.15	60	1
				60	1		

1.- Media aritmética.

$$\overline{X} = \frac{\sum_{1}^{k} f_k \times x_k}{n} = \frac{3(4.23)+3(4.98)+.....+9(8.73)}{60} = \frac{420.3}{60} = 7.01$$

2.- Mediana.

La clase mediana es la número 5.

$$M_e = \left(\frac{\frac{(n+1)}{2} - (F+1)}{f_{M_e}} \right) w + L_{M_e}$$

, donde

$$M_e = \left(\frac{\frac{(60+1)}{2} - (26+1)}{13} \right) 0.75 + 6.86 = 0.20 + 6.86 = 7.06$$

3.- Moda.

$$M_0 = L_{M_0} + \left(\frac{d_1}{d_1 + d_2} \right) w$$

, donde

Hay dos clases modal, la número 4 y la número 5.

Para la clase modal número 4: $M_0 = 6.11 + \left(\frac{0}{0+6} \right) 0.75 = 6.11$

Para la clase modal 5: $M_0 = 6.11 + \left(\frac{1}{1+0} \right) 0.75 = 6.11 + 0.75 = 6.86$

4.- La varianza.

$$S^2 = \frac{3(4.23-7.01)^2 + 3(4.98-7.01)^2 + ...+9(8.73-7.01)^2}{59} = \frac{89.2}{59} = 1.51$$

5.- La desviación estándar.

$$S = \sqrt{1.51} = 1.23$$

6.- Coeficiente de variación.

$$CV = \frac{1.23}{7.01} x100 = 17.5\%$$

3.5 La importancia de la varianza

Al ser la varianza una medida que expresa el grado de variabilidad de los datos de un conjunto alrededor de su media aritmética, su valor numérico nos da información de la confiabilidad de una medida de tendencia central. Si los datos se encuentran muy dispersos, la posición central que calculemos será mucho menos representativa de estos datos que cuando se agrupan más cercanamente alrededor de su media. Para evidenciar lo que acabamos de señalar, calculemos la media y la varianza de los conjuntos que se muestran en la tabla 3.10.

TABLA 3.10 Datos de dos conjuntos A y B

A	14.3	15.8	14.6	14.3	15.5	16.1	16.0	15.7	14.9
B	10.3	20.8	10.6	14.3	21.5	10.1	20.0	10.7	18.9

Resultará fácil para el lector comprobar que la media y la varianza del Conjunto A es 15.24 y 0.53 respectivamente, y la del Conjunto B 15.24 y 25, es decir, ambos conjuntos tienen la misma media aritmética pero el segundo muestra una varianza mucho mayor que la del primero. Adicionalmente, compare ambos conjuntos y podrá apreciar que la media 15.24 es mucho más representativa de los valores del primer conjunto que de los valores del segundo, es decir, a menor variabilidad mayor representatividad de la media aritmética calculada.

3.6 La importancia del Coeficiente de Variación

El valor numérico de la varianza de un conjunto de datos depende de la unidad en que éstos hayan sido medidos. Según la *Propiedad 4* de la varianza, si X representa el peso de un grupo de encomiendas transportadas por Servientrega medido en kilogramos y la variable Y este mismo peso medido en libras, entonces entre las varianzas de ambos conjuntos existe la relación $S_Y^2 = (2.2)^2 \, S_X^2$, ya que un kilogramo es igual a 2.2 libras, es decir, la varianza del peso de las encomiendas medido en libras es $(2.2)^2 = 4.84$ veces mayor que la varianza de los pesos de estas mismas encomiendas medido en kilogramos, siendo ambos sin embargo el mismo conjunto.

Al ser el Coeficiente de Variación una cantidad relativa, su valor es el mismo independientemente de las unidades en que fue medido el conjunto, es decir,

$$CV(X) = \frac{S(Kg)}{\overline{X}(Kg)} = \frac{S(Lb)}{\overline{X}(Lb)} = \frac{S(Gr)}{\overline{X}(Gr)} = \frac{S(Ton)}{\overline{X}(Ton)}$$

ya que las unidades del numerador y el denominador se cancelan entre sí.

Ejercicios del capítulo

3.1 Los datos que se muestran a continuación representan las calificaciones obtenidas por 50 estudiantes en un examen final de estadística.

8.4	7.6	9.2	10	9.7	8.8	7.5	7.9	8.1	9.1
10	7.5	7.3	8.4	8.8	9.4	9.6	8.2	8.7	9
8	7.7	7.1	7	8.3	9	8.1	9.6	10	9
7.4	7.5	8.8	9.8	9.1	10	9	7.1	7.3	7
7.3	8.5	8.2	9	8	10	8.1	7.7	9	8.9

Con estos datos no agrupados, obtenga las medidas de tendencia central y de variación correspondientes.

3.2 A continuación se muestra el número de minutos promedio diarios de utilización del servicio de internet de 60 clientes de la Corporación Nacional de Telecomunicaciones.

64	56	72	80	77	68	55	59	61	71
110	85	83	94	98	104	106	92	97	100
75	72	66	65	78	85	76	91	95	85
89	90	103	113	106	114	105	86	88	85
80	92	89	79	87	107	88	84	97	96
73	70	64	82	90	110	91	90	94	100

Con estos datos no agrupados, obtenga las medidas de tendencia central y de variación correspondientes.

3.3 Los datos que se aprecian a continuación representan el porcentaje del sueldo de cien personas de clase baja con relación a la remuneración básica actual.

135.24	137.35	139.12	139.71	142.65	144.41	144.69	146.76	148.53	148.74
135.98	135.41	139.41	141.47	143.24	144.12	146.17	146.47	147.35	150.88
137.06	138.24	140.29	139.71	141.86	145.29	148.88	147.06	148.24	149.12
134.71	138.53	139.41	142.06	142.94	142.35	147.06	146.47	149.12	151.23
136.76	136.18	139.71	138.53	142.35	145.59	144.71	148.24	148.82	149.41
138.24	139.41	138.24	140.02	143.53	142.35	147.35	147.65	148.53	149.71
135.24	137.06	139.12	142.06	141.76	143.28	146.18	147.94	146.47	148.82
135.59	139.71	139.54	142.94	140.29	144.41	144.71	148.24	148.94	149.12
137.35	136.47	140.29	140.56	143.82	144.41	147.06	145.29	147.94	149.71
136.76	137.65	138.53	140.88	141.18	143.24	146.18	147.06	148.24	147.94

Con estos datos no agrupados, obtenga las medidas de tendencia central y de variación correspondientes.

3.4 Para el siguiente conjunto de datos:

1.07	1.63	1.25	1.33	1.28	1.13	0.92	0.98	1.02	1.18
1.83	1.42	1.38	1.57	1.63	1.73	1.77	1.53	1.62	1.67
1.25	1.21	1.17	1.08	1.36	1.42	1.27	1.52	1.58	1.42
1.48	1.56	1.72	1.88	1.77	1.91	1.75	1.43	1.47	1.42
1.33	1.53	1.48	1.32	1.45	1.78	1.47	1.42	1.62	1.61
1.22	1.17	1.07	1.37	1.53	1.83	1.52	1.51	1.57	1.67

Obtenga las medidas de tendencia central y de variación correspondientes.

3.5 Utilizando la tabla de distribución de frecuencias que se muestra a continuación, calcule las medidas de tendencia central y de variación correspondientes a estos datos agrupados.

Clases	**Sueldos**		**Sueldos**		**Marca de clase**	f_i	F_i
	Li	**Ls**	**Lri**	**Lrs**			
1	134.71	137.46	134.705	137.465	136.085	14	14
2	137.47	140.22	137.465	140.225	138.845	18	32
3	140.23	142.98	140.225	142.985	141.605	17	49
4	142.99	145.74	142.985	145.745	144.365	15	64
5	145.75	148.5	145.745	148.505	147.125	21	85
6	148.51	151.26	148.505	151.265	149.885	15	100
						100	

3.6 Calcule las medidas de tendencia central y de variación correspondientes a los datos agrupados que aparecen a continuación.

Clases					**Marca de clase**	f_i	F_i
	Li	**Ls**	**Lri**	**Lrs**			
1	0.92	1.11	0.915	1.115	1.015	6	6
2	1.12	1.31	1.115	1.315	1.215	10	16
3	1.32	1.51	1.315	1.515	1.415	18	34
4	1.52	1.71	1.515	1.715	1.615	16	50
5	1.72	1.91	1.715	1.915	1.815	10	60
						60	

3.7 Utilizando la tabla de distribución de frecuencias que se muestra a continuación, calcule las medidas de tendencia central y de variación correspondientes a estos datos agrupados.

Clases	Calificaciones		Calificaciones		Marca de clase	f_i	F_i
	Li	Ls	Lri	Lrs			
1	7.0	7.7	6.95	7.75	7.35	14	14
2	7.8	8.5	7.75	8.55	8.15	12	26
3	8.6	9.3	8.55	9.35	8.95	14	40
4	9.4	10.1	9.35	10.15	9.75	10	50
						50	

3.8 Obtenga las medidas de tendencia central y de variación correspondientes a los datos agrupados que se muestran en la siguiente tabla.

Clases	Minutos		Minutos		Marca de clase	f_i	F_i
	Li	Ls	Lri	Lrs			
1	55	64	54.5	64.5	59.5	6	6
2	65	74	64.5	74.5	69.5	8	14
3	75	84	74.5	84.5	79.5	10	24
4	85	94	84.5	94.5	89.5	19	43
5	95	104	94.5	104.5	99.5	9	52
6	105	114	104.5	114.5	109.5	8	60
						60	

Capítulo 4

Introducción a la teoría de probabilidades

El problema

El último censo de población y vivienda realizado en el Ecuador en el 2010 reveló que alrededor del 26% de los hogares en el país cuentan con una computadora en casa. ¿Podemos conocer cuál es la probabilidad que al seleccionar al azar un hogar ecuatoriano éste posea una computadora?

4.1 Introducción

El inicio de la *teoría de la probabilidad* se remonta al siglo XVII cuando Blaise Pascal (matemático, físico, filósofo cristiano y escritor francés) y Pierre de Fermat (jurista y matemático francés) intercambiaron correspondencia sobre el tema. Posteriormente, Christiaan Huygens, (astrónomo, físico y matemático holandés) estudió la probabilidad con un carácter científico y fueron Jacob Bernoulli (matemático y científico suizo) y Abraham de Moivre (matemático francés famoso por predecir el día de su muerte mediante un cálculo matemático) los que ubicaron la materia como una rama de la matemática.

Los métodos de la *estadística matemática* surgen a partir de la *teoría de probabilidades*. Uno de los aspectos medulares en los que se basa la *Estadística Inferencial* consiste en conocer cuál es la *probabilidad de que algo ocurra* en un momento posterior al presente. Al tomar una decisión cualquiera sin tener a la mano toda la información requerida para ello, la persona o entidad que lo hace siempre corre el riesgo de equivocarse, y en consecuencia, actuar de forma errónea. Por esa razón, resulta muy importante conocer de antemano los riesgos que se corren al tomar una decisión con la finalidad de prever una posible alternativa.

La *teoría de probabilidades* es una herramienta estadística de suma utilidad al momento de evaluar los riesgos que se corren y la magnitud de los posibles errores que pueden cometerse al tomar una decisión sin poseer toda la información necesaria para ello.

En este capítulo estudiaremos los conceptos básicos relacionados con *la teoría de probabilidades* y las reglas que permiten su determinación numérica.

4.2 Conceptos básicos

4.2.1 Experimento Aleatorio

La actividad que da lugar a la ocurrencia de un determinado resultado se llama

*experimento***.** Un experimento se considera *aleatorio* cuando su resultado no se puede predecir con exactitud.

Ejemplos de experimentos aleatorios son:

1.- El lanzamiento de una moneda.

2.- El lanzamiento de un dado.

3.- El estudio de la precipitación ocurrida en una etapa.

4.- El estudio del rendimiento de un cultivo.

5.- El estudio del comportamiento de las declaraciones de impuestos en un año determinado, etc.

Para los fines que persiguen los métodos estadísticos, los experimentos aleatorios deben tener la característica de poderlos repetir, como es el caso de observar el comportamiento de un determinado fenómeno.

La experiencia ha demostrado que si se repite un experimento aleatorio un gran número de veces, la frecuencia de un resultado (o algunos resultados) tiende a ser constante, y decimos que el resultado (o los resultados) muestran una *regularidad estadística*.

4.2.2 Espacio muestral

El *espacio muestral de un experimento aleatorio* es el conjunto integrado por todos los posibles resultados de dicho experimento. Los espacios muestrales pueden ser representados en forma de conjuntos o a veces en forma gráfica.

Para los dos primeros ejemplos de experimentos aleatorios dados anteriormente, se tiene que los espacios muestrales correspondientes son:

1.- {C, S} C = Cara S = Sello

2.- {1, 2, 3, 4, 5, 6}

El espacio muestral se denota por **S** y será finito o infinito en dependencia de que el conjunto tenga un número finito o infinito de elementos.

4.2.3 Suceso o Evento

En la teoría de probabilidades, un *suceso* o *evento* es uno o más de los posibles resultados de un experimento aleatorio, por tanto un *suceso* o *evento* es un subconjunto del espacio muestral. Como ejemplos de sucesos o eventos podemos citar los siguientes:

- Que *"salga cara"* al lanzar una moneda al aire. Este *suceso* o *evento* se representa por el conjunto A = {C} el cual es un subconjunto del espacio muestral

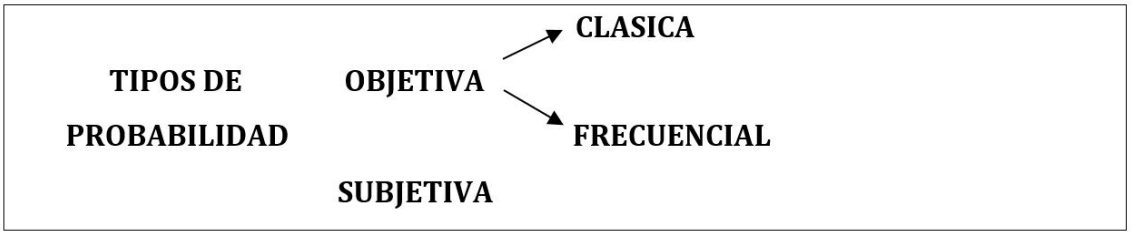
S = {C, S}.

- Que *"salga un número par"* al lanzar un dado. Este evento se representa por el conjunto A = {2, 4, 6} el cual es un subconjunto del espacio muestral S = {1, 2, 3, 4, 5, 6}.

De acuerdo con el número de elementos que integran un evento, este puede ser *elemental o simple* si consta de un solo elemento, o *no elemental o compuesto* si consta de dos o más elementos. El espacio muestral y cualquier evento asociado a un experimento aleatorio se puede representar gráficamente mediante un *diagrama de Venn*, el cual es un esquema muy utilizado en la teoría de conjuntos.

4.3 Tipos de probabilidad

Existen dos *tipos de probabilidad*, la *probabilidad objetiva* y la *probabilidad subjetiva*. La *probabilidad objetiva* a su vez se divide en *probabilidad clásica y probabilidad frecuencial.*



4.3.1 Definición clásica de probabilidad

Se llama *probabilidad* de ocurrencia de un evento al cociente del número de resultados favorables al evento y el número total de resultados posibles del experimento.

De aquí que la probabilidad de que ocurra un evento *A* se expresa como:

$P(A) = \frac{N_A}{N}$, donde N_A es el número de resultados favorables al evento *A* y *N* es el número de todos los resultados posibles del experimento.

Es necesario destacar que la definición de probabilidad que hemos expresado solo es aplicable a espacios muestrales finitos.

Si *A* fuera un *evento seguro* entonces *P(A)=1*

Si *A* fuera un *evento imposible* entonces *P(A)=0*

Probemos que la probabilidad de ocurrencia de un evento es siempre una cantidad entre 0 y 1.

$0 \leq N_A \leq N$ y dividiendo para N,

$\frac{0}{N} \leq \frac{N_A}{N} \leq \frac{N}{N}$, de donde

$0 \leq P(A) \leq 1$

EJEMPLOS

1.- Si se lanzan dos monedas, calcule la probabilidad de ocurrencia del evento *"salga una cara"*.

El espacio muestral es S = {CC, CS, SC, SS}

El evento es A = {CS, SC}

La probabilidad será entonces P(A)=2/4=1/2=0.5

2.- Si se lanza un dado, calcule la probabilidad de ocurrencia del evento *"salga un número par"*.

El espacio muestral es S = {1, 2, 3, 4, 5, 6}

El evento es A = {2, 4, 6}

La probabilidad será entonces P(A)=3/6=1/2=0.5

3.- Si se lanzan dos dados, cual es la probabilidad de que las caras que caigan hacia arriba:

Sumen 7

El espacio muestral está compuesto por 36 elementos, que son los siguientes:

(1,1) (2,1) (3,1) (4,1) (5,1) (6,1)
(1,2) (2,2) (3,2) (4,2) (5,2) (6,2)
(1,3) (2,3) (3,3) (4,3) (5,3) (6,3)
(1,4) (2,4) (3,4) (4,4) (5,4) (6,4)
(1,5) (2,5) (3,5) (4,5) (5,5) (6,5)
(1,6) (2,6) (3,6) (4,6) (5,6) (6,6)

El evento es A = {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}

La probabilidad será entonces P(A)=6/36=1/6=0.17

Sean dobles

El espacio muestral tiene 36 elementos.

El evento es A = {(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)}

La probabilidad será entonces P(A)=6/36=1/6=0.17

Sumen 8 y la diferencia entre el valor mayor y el menor sea 4

El espacio muestral tiene 36 elementos.

El evento es A = {(2,6), (6,2)}

La probabilidad será entonces P(A)=2/36=1/18=0.06

4.3.2 Probabilidad frecuencial

La *probabilidad frecuencial* de ocurrencia de un suceso se basa en el principio de la llamada *frecuencia relativa de presentación* de un evento, mediante el cual se define la probabilidad como *la frecuencia relativa observada de un evento* en un gran número de repeticiones del experimento.

Es decir, si f_A es la frecuencia con la que se presenta un evento o suceso A, y N es el número de elementos del espacio muestral, entonces:

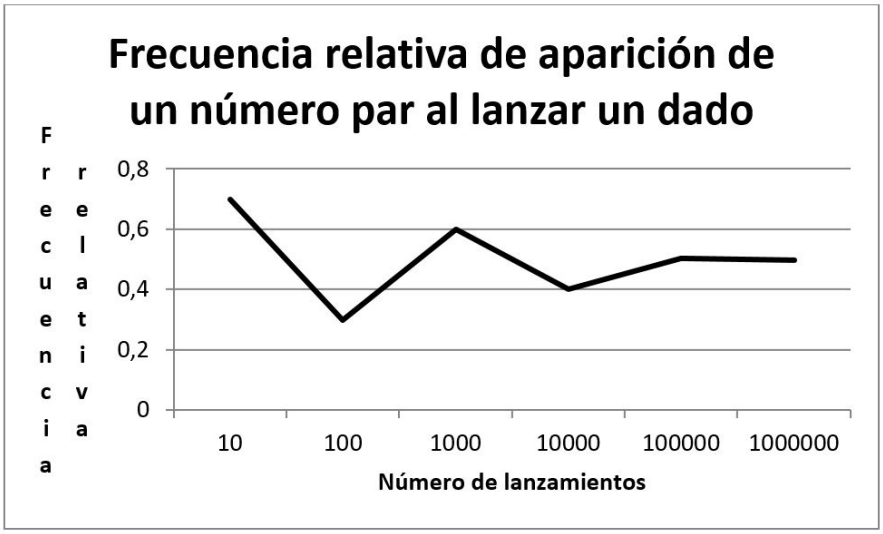
$$P(A) = \frac{f_A}{N}$$

Cuando aplicamos el principio de frecuencia relativa de presentación para el cálculo de una probabilidad, el valor que obtengamos alcanzará una mayor precisión en la medida en que las repeticiones del experimento aleatorio vayan aumentando.

Así por ejemplo, si lanzamos un dado en diez ocasiones, un resultado probable pudiera ser que obtengamos como resultado que en 7 ocasiones salga un número par y solo en 3 un número impar. Sin embargo, si lanzamos un dado cien veces ya no sería lógico esperar un resultado de 70 veces un número par y 30 veces un número impar, y mucho menos si lo hacemos un millón de veces obtener un resultado par en 700000 ocasiones y uno impar en 300000 de las veces.

La *regularidad estadística* debe comportarse aproximadamente como se muestra en la figura 4.1.

FIGURA 4.1 Aparición de número par al lanzar un dado



La *tendencia* a largo plazo deberá ser que la frecuencia relativa de ocurrencia del evento *"salga un número par"* sea igual a 0.5.

OTRO EJEMPLO

Los datos que se muestran a continuación representan los gastos semanales en comisariato de 80 profesionales.

TABLA 4.1 Gastos semanales en comisariato

157	151	158	160	161	159	162	157	155	162
158	159	161	154	156	163	160	159	158	156
155	160	158	162	160	153	164	157	161	159
157	152	159	150	158	164	156	163	160	157
160	163	154	155	162	153	165	161	159	158
158	155	157	162	156	158	161	162	152	155
154	164	159	165	151	160	156	163	159	156
156	153	161	157	163	160	164	153	157	158

Calcule la probabilidad de que al seleccionar al azar uno de estos profesionales su gasto semanal en comisariato sea:

1.- igual a 156 dólares.

Evento A: gasto semanal en comisariato igual a 156 dólares.

Cantidad de profesionales que gastan 156 dólares semanales en comisariato = 7 = f_A.

Cantidad total de profesionales = 80 = N.

De donde:

$$P(A) = \frac{f_A}{N} = \frac{7}{80} = 0.0875$$

es decir, la probabilidad de que al seleccionar al azar un profesional éste tenga un gasto semanal en comisariato igual a 156 dólares es igual a 0.0875, o lo que es lo mismo, 8.75%.

2.- menor a 154 dólares.

Evento A: gasto semanal en comisariato menor a 154 dólares.

Cantidad de profesionales que gastan menos de 154 dólares semanales en comisariato = 9 = f_A.

Cantidad total de profesionales = 80 = N.

De donde:

$$P(A) = \frac{f_A}{N} = \frac{9}{80} = 0.1125$$

es decir, la probabilidad de que al seleccionar al azar un profesional éste tenga un gasto semanal en comisariato menor a 154 dólares es igual a 0.1125, o lo

que es lo mismo, 11.25%.

3.- mayor a 161 dólares.

Evento A: gasto semanal en comisariato mayor a 161 dólares.

Cantidad de profesionales que gastan más de 161 dólares semanales en comisariato = 17 = f_A.

Cantidad total de profesionales = 80 = N.

De donde:

$$P(A) = \frac{f_A}{N} = \frac{17}{80} = 0.2125$$

es decir, la probabilidad de que al seleccionar al azar un profesional éste tenga un gasto semanal en comisariato mayor a 161 dólares es igual a 0.2125, o lo que es lo mismo, 21.25%.

4.3.3 Probabilidad subjetiva

Podemos definir este tipo de probabilidad, diciendo que es aquella que calcula la probabilidad de ocurrencia de un evento basándose en experiencias anteriores o puntos de vista y opiniones de las personas que la calculan.

La utilización del elemento subjetivo en el cálculo de una probabilidad fue introducida en el año 1926 por el matemático y filósofo inglés Frank Plumpton Ramsey.

4.4 Eventos mutuamente excluyentes

Antes de comenzar a trabajar con algunas reglas que norman el cálculo de una probabilidad, es necesario establecer el concepto de *eventos mutuamente excluyentes*.

Dos o más eventos se dice que son *mutuamente excluyentes* cuando uno y solo uno de ellos pueden ocurrir a un mismo tiempo.

Al lanzar una moneda se puede obtener como resultado cara o sello, pero no ambos, por lo cual estos dos eventos son *mutuamente excluyentes*.

De igual manera, en un partido de futbol un jugador puede desempeñarse como portero, defensa, volante o delantero. Solo uno de estos cuatro resultados es posible, por tanto, son posiciones *mutuamente excluyentes*.

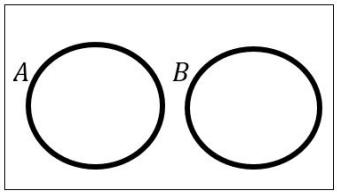
Cuando dos o más eventos pueden ocurrir al mismo tiempo se dice que *no son mutuamente excluyentes*.

Existe una forma práctica de representar un espacio muestral y sus respectivos eventos mediante una representación gráfica conocida con el nombre de *Diagramas de Venn*, en honor al matemático y lógico británico del siglo XIX John Venn Sykes, el cual se destacó por sus investigaciones en el campo de la lógica inductiva.

En estos diagramas el espacio muestral se representa mediante un rectángulo y los eventos en forma de círculos o elipses.

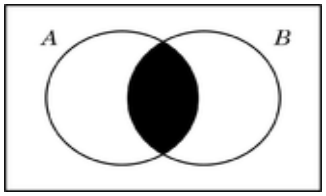
En la figura 4.2 se puede apreciar un *diagrama de Venn* para dos *eventos mutuamente excluyentes*. Observe que los círculos que representan a los eventos A y B no se interceptan entre sí.

FIGURA 4.2 Diagrama de Venn. Eventos mutuamente excluyentes



En el *diagrama de Venn* para eventos *no mutuamente excluyentes* de la figura 4.3, se aprecia que los círculos que representan a ambos sucesos se interceptan entre sí.

FIGURA 4.3 Diagrama de Venn. Eventos no mutuamente excluyentes



4.4.1 Regla de adición para eventos mutuamente excluyentes

A menudo resulta necesario calcular la probabilidad de ocurrencia de dos o más eventos mutuamente excluyentes. En estos casos podemos obtener esta probabilidad utilizando la *regla de adición* para eventos mutuamente excluyentes, la cual viene dada por la expresión P (A o B) = P (A) + P (B) para el caso de dos eventos, P (A o B o C) = P (A) + P (B) + P (C) para el caso de tres eventos, y así sucesivamente.

Precisemos mediante un ejemplo el cálculo de la probabilidad de ocurrencia de eventos mutuamente excluyentes, y con este propósito, consideremos los resultados de una muestra de lápices producidos por una empresa cuyas longitudes están por debajo de la norma establecida, son iguales a dicha norma o están por encima de ella. Estos datos se muestran en la tabla 4.2.

TABLA 4.2 Longitud de los lápices producidos por una empresa

Longitud de los lápices	Evento	Número de lápices
Por debajo de la norma	A	180
Igual a la norma	B	4400
Por encima de la norma	C	420
		5000

Calcule la probabilidad que al seleccionar al azar uno de estos lápices su longitud:

a. Esté por debajo de la norma.
b. Sea igual a la norma.
c. Esté por encima de la norma.
d. Esté por debajo de la norma o igual a ella.
e. Esté por debajo de la norma o por encima de ella.
f. Sea igual a la norma o esté por encima de ella.

Según la definición frecuencial de probabilidad,

a) $P(A)=\frac{180}{5000}=0.036$

b) $P(B)=\frac{4400}{5000}=0.88$

c) $P(B)=\frac{420}{5000}=0.084$

Aplicando la regla de la adición,

d) $P(A\,o\,B)=P(A)+P(B)=0.036+0.88=0.916$.

e) $P(A\,o\,C)=P(A)+P(C)=0.036+0.084=0.12$.

f) $P(B\,o\,C)=P(B)+P(C)=0.88+0.084=0.964$.

Adicionalmente a lo anterior, ¿cuál es la probabilidad que al seleccionar al azar uno de los lápices producidos por la empresa su longitud *no esté por debajo de la norma*? Sin lugar a dudas, la probabilidad de ocurrencia de este evento, denominado **no A** y que se denota **~A**, viene dada por P (~A) = 1 - P (A) = 1 – 0.036 = 0.964.

Tomando en cuenta lo anterior podemos formular la *regla del complemento*, la cual establece que cualquiera sea el evento A:

P (A) + P (~A) = 1

4.4.2 Regla de adición para eventos no mutuamente excluyentes

Consideremos que el Banco del Pichincha extrajo una muestra de 500 de sus

clientes, la cual dio como resultado que 232 de ellos poseían cuentas de ahorro y 295 cuentas corrientes. Según estos datos, la probabilidad de que un cliente del banco del Pichincha escogido de forma aleatoria tenga *una cuenta de ahorro o una cuenta corriente* viene dada por:

$P(A\,o\,B) = P(A) + P(B) = \frac{232}{500} + \frac{295}{500} = \frac{527}{500} = 1.054$, lo cual resulta una contradicción por cuanto una probabilidad no puede ser nunca mayor a 1. La explicación del por qué se obtuvo este resultado es que algunos de los 500 clientes del banco poseen tanto cuenta de ahorro como cuenta corriente, y por tanto, están siendo considerados dos veces. Es decir, los eventos poseen una cuenta de ahorro y poseer una cuenta corriente *no son* mutuamente excluyentes, y en consecuencia, la *regla de adición* para eventos mutuamente excluyentes no puede ser utilizada en este caso.

En situaciones similares a ésta, la regla de la adición viene dada por la expresión:
P (A o B) = P (A) + P (B) – P (A y B)

En nuestro ejemplo, un ajuste de los resultados reportados por el banco evidenció que 80 de los 500 clientes poseían ambos tipos de cuenta y por tanto si el evento A consiste en poseer cuenta de ahorro y B en poseer cuenta corriente, entonces,

$$P(A\,o\,B) = \frac{232}{500} + \frac{295}{500} - \frac{80}{500} = \frac{447}{500} = 0.894$$

La probabilidad de ocurrencia de dos eventos al mismo tiempo recibe el nombre de *probabilidad conjunta*. En este caso, $\frac{80}{500} = 0.16$ es un ejemplo de una *probabilidad conjunta.*

Estudiemos un segundo ejemplo. Consideremos que una caja contiene 20 fichas numeradas del 1 al 20 y deseamos conocer cuál es la probabilidad que al extraer una ficha de la caja ésta sea divisible por 3 o por 4.

Evento A: divisible por 3.

Evento B: divisible por 4.

S={1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20}

S_A = {3, 6, 9, 12, 15, 18}

S_B = {4, 8, 12, 16, 20}

S_{AB} = {12}

$$P(A\,o\,B) = \frac{6}{20} + \frac{5}{20} - \frac{1}{20} = \frac{10}{20} = \frac{1}{2} = 0.5$$

4.5 Probabilidad condicional

Muchas veces necesitamos calcular la probabilidad de ocurrencia de un evento A después de haber ocurrido un evento B. En este caso el evento B sirve como un es-

pacio muestral nuevo, según la definición clásica de probabilidad. A esta probabilidad se le llama *probabilidad condicional de A dado B* y se define como:

$$P(A/B) = \frac{N_{AB}}{N_B}$$

donde N_{AB} representa la cantidad de elementos del espacio muestral que pertenecen al evento *AB* y N_B la cantidad de elementos que pertenecen al evento *B*.

Puede demostrarse que:

$$P(A/B) = \frac{P(A\,y\,B)}{P(B)} \; ya\ que\ P(A\,y\,B) = \frac{N_{AB}}{N}\ y\ P(B) = \frac{N_B}{N}$$

Veamos el siguiente ejemplo. Si se lanzan dos dados y A representa el evento de que ambos dados muestren al menos 4 puntos cada uno y B es el evento de que ambos dados muestren una suma igual a 8 puntos, calcule:

a) La probabilidad condicional de A dado B.

El espacio muestral S del experimento aleatorio está formado por los 36 elementos que se muestran a continuación:

(1,1) (2,1) (3,1) (4,1) (5,1) (6,1)
(1,2) (2,2) (3,2) (4,2) (5,2) (6,2)
(1,3) (2,3) (3,3) (4,3) (5,3) (6,3)
(1,4) (2,4) (3,4) (4,4) (5,4) (6,4)
(1,5) (2,5) (3,5) (4,5) (5,5) (6,5)
(1,6) (2,6) (3,6) (4,6) (5,6) (6,6)

S_{AB} = {(4,4)} y N_{AB} = 1
S_B = {(2,6), (3,5), (4,4), (5,3), (6,2)} y N_B= 5

$$P(A/B) = \frac{N_{AB}}{N_B} = \frac{1}{5} = 0.2$$

b) La probabilidad condicional de B dado A.

S_{AB} = {(4,4)} y N_{AB}= 1
S_A = {(4,4), (4,5), (4,6), (5,4), (5,5), (5,6), (6,4), (6,5), (6,6)} y $N_{A=}$ 9

$$P(B/A) = \frac{N_{AB}}{N_A} = \frac{1}{9} = 0.11$$

4.6 Eventos independientes

Dos eventos A y B son *independientes***,** cuando la probabilidad de ocurrencia de uno de ellos no depende de la ocurrencia o no del otro.

Cuando dos eventos A y B son independientes, entonces:

P (A/B)=P (A) y P (B/A)=P (B)

4.6.1 Regla del producto o de la multiplicación

Si tenemos dos eventos A y B de un espacio muestral S, la probabilidad de ocurrencia de A y B se expresa como:

P (A y B) = P (A) P (B/A) = P (B) P (A/B)

Si A y B son *eventos independientes*, entonces se cumple que P (A/B) = P (A) y P (B/A) = P (B) y la expresión anterior se reduce a:

P (A y B) = P (A) P (B)

Veamos el siguiente ejemplo. Si lanzamos dos monedas, cual es la probabilidad que:

a) Aparezca cara en ambas monedas.

Ambos eventos son independientes, por tanto:

P(C y C) = P (C) P (C) = 1/2x1/2 = 1/4 = 0.25

b) Aparezca cara en una moneda y sello en la otra.

P (C y S o S y C) = P (C y S) + P (S y C) por ser eventos independientes.

P (C y S) = 1/2x1/2 = 1/4 P (S y C) = 1/2x1/2 = 1/4

P (C y S o S y C)=1/4+1/4 = 2/4 = 1/2 = 0.5

A modo de resumen de los aspectos estudiados en el capítulo, veamos los siguientes ejemplos.

1.- Un juego consiste en sacar de una caja que contiene 20 bolas numeradas del 1 al 20 una con número par y divisible por 3. ¿Cuál es la probabilidad de ganar el juego?

S = {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20}

S_A = {6, 12, 18}

$$P(A) = \frac{3}{20} = 0.15$$

2.- Si se lanzan dos dados, cual es la probabilidad que las caras que queden hacia arriba sumen 8, dado que su diferencia en valor absoluto es 4.

A = sumen 8.

B = diferencia 4.

P (A/B) = P (AB) / P (B)

S_{AB} = {(2,6), (6,2)}

S_B = {((1,5), (2,6), (5,1), (6,2)}

P(A/B) = 2/4 = 1/2 = 0.5

3.- La tabla 4.3 muestra los datos correspondientes al tipo de producción y al número de trabajadores de 35 empresas de un Complejo Industrial.

TABLA 4.3 Cantidad de trabajadores de un Complejo Industrial

Tipo de	**Trabajadores**			**Subtotal**
producción	Más de 200	Menos de 100	Entre 100 y 200	
Básica	8	6	6	20
Ligera	6	4	5	15
Subtotal	14	10	11	

Si se selecciona al azar una empresa, obtenga la probabilidad que la empresa escogida:

a) sea de producción ligera.

Sea A el evento *"empresa de producción ligera".*

El espacio muestral tiene 35 elementos.

El número de resultados favorables al evento es 15, por tanto:

$$P(A) = \frac{15}{35} = 0.43$$

b) tenga por lo menos 100 trabajadores.

Sea A el evento *"tenga por lo menos 100 trabajadores".*

El espacio muestral tiene 35 elementos. El número de resultados favorables al evento es 8 + 6 + 6 + 5 = 25, por tanto:

$$P(A) = \frac{25}{35} = 0.71$$

c) sea una empresa básica y tenga más de 200 trabajadores.

Sea A el evento *"empresa básica con más de 200 trabajadores".*

El espacio muestral tiene 35 elementos.

El número de resultados favorables es 8, por tanto:

$$P(A) = \frac{8}{35} = 0.23$$

Ejercicios del capítulo

Capítulo 4

4.1 Los datos que se muestran a continuación representan las calificaciones obtenidas por 50 estudiantes en un examen final de estadística.

8.4	7.6	9.2	10	9.7	8.8	7.5	7.9	8.1	9.1
10	7.5	7.3	8.4	8.8	9.4	9.6	8.2	8.7	9
8	7.7	7.1	7	8.3	9	8.1	9.6	10	9
7.4	7.5	8.8	9.8	9.1	10	9	7.1	7.3	7
7.3	8.5	8.2	9	8	10	8.1	7.7	9	8.9

Determine la probabilidad que al seleccionar al azar uno de estos estudiantes su calificación sea:

1.- Igual a 7.3.

2.- Menor a 9.

3.- Mayor a 8.5.

4.- Entre 8.2 y 9.1.

4.2 A continuación se muestra el número de minutos promedio diarios de utilización del servicio de internet de 60 clientes de la Corporación Nacional de Telecomunicaciones.

64	56	72	80	77	68	55	59	61	71
110	85	83	94	98	104	106	92	97	100
75	72	66	65	78	85	76	91	95	85
89	90	103	113	106	114	105	86	88	85
80	92	89	79	87	107	88	84	97	96
73	70	64	82	90	110	91	90	94	100

Determine la probabilidad que al seleccionar al azar uno de estos clientes, el número de minutos promedio diarios sea:

1.- Igual a 85.

2.- Menor a 100.

3.- Mayor a 80.

4.- Entre 76 y 90.

4.3 Los datos que se aprecian a continuación representan el porcentaje del sueldo de cien personas de clase baja con relación a la remuneración básica actual.

135.24	137.35	139.12	139.71	142.65	144.41	144.69	146.76	148.53	148.74
135.98	135.41	139.41	141.47	143.24	144.12	146.17	146.47	147.35	150.88
137.06	138.24	140.29	139.71	141.86	145.29	148.88	147.06	148.24	149.12
134.71	138.53	139.41	142.06	142.94	142.35	147.06	146.47	149.12	151.23
136.76	136.18	139.71	138.53	142.35	145.59	144.71	148.24	148.82	149.41
138.24	139.41	138.24	140.02	143.53	142.35	147.35	147.65	148.53	149.71
135.24	137.06	139.12	142.06	141.76	143.28	146.18	147.94	146.47	148.82
135.59	139.71	139.54	142.94	140.29	144.41	144.71	148.24	148.94	149.12
137.35	136.47	140.29	140.56	143.82	144.41	147.06	145.29	147.94	149.71
136.76	137.65	138.53	140.88	141.18	143.24	146.18	147.06	148.24	147.94

Determine la probabilidad que al seleccionar al azar una de estas personas, el porcentaje de su sueldo sea:

1.- Igual a 144.41.

2.- Menor a 136.76.

3.- Mayor a 140.

4.- Entre 145 y 150.

4.4 Dado los datos que se muestran a continuación:

1.07	1.63	1.25	1.33	1.28	1.13	0.92	0.98	1.02	1.18
1.83	1.42	1.38	1.57	1.63	1.73	1.77	1.53	1.62	1.67
1.25	1.21	1.17	1.08	1.36	1.42	1.27	1.52	1.58	1.42
1.48	1.56	1.72	1.88	1.77	1.91	1.75	1.43	1.47	1.42
1.33	1.53	1.48	1.32	1.45	1.78	1.47	1.42	1.62	1.61
1.22	1.17	1.07	1.37	1.53	1.83	1.52	1.51	1.57	1.67

Determine la probabilidad que al seleccionar al azar un elemento del anterior conjunto, su valor sea:

1.- Igual a 1.33.

2.- Menor a 1.88.

3.- Mayor a 1.5.

4.- Entre 1.2 y 1.7.

4.5 En una muestra de 2000 equipajes pesados en un importante aeropuerto del país se obtuvo los resultados que se aprecian a continuación:

Peso de los equipajes	**Evento**	**Cantidad de equipajes**
Inferior al establecido	A	420
Igual al establecido	B	1350
Superior al establecido	C	230
		2000

Calcule la probabilidad que al seleccionar al azar uno de estos equipajes su peso:

1.- Sea inferior a lo establecido.

2.- Sea igual a lo establecido.

3.- Sea superior a lo establecido.

4.- Sea inferior o igual a lo establecido.

5.- Sea inferior o superior a lo establecido.

6.- Sea igual o superior a lo establecido.

4.6 En un evento desarrollado en el paraninfo de la ULEAM al cual asistieron un total de 490 personas, se anotaron aquellas que llegaron antes de la hora prevista, a la hora prevista y después de dicha hora. Los resultados alcanzados se aprecian a continuación:

Llegada a la reunión	**Evento**	**Cantidad de personas**
Antes de la hora prevista	A	60
A la hora prevista	B	130
Después de la hora prevista	C	300
		490

Calcule la probabilidad que al seleccionar al azar una de estas personas, la misma haya llegado:

1.- Antes de la hora prevista.

2.- A la hora prevista.

3.- Después de la hora prevista.

4.- Antes o a la hora prevista.

5.- Antes o después de la hora prevista.

6.- A la hora o después de la hora prevista.

4.7 Si al evento desarrollado en el paraninfo de la ULEAM asistieron un total de 270 economistas, 200 ingenieros comerciales y 20 que poseían ambos títulos, calcule la probabilidad que al seleccionar al azar uno de estos profesionales el mismo sea:

1.- Economista.

2.- Ingeniero Comercial.

3. De ambas profesiones.

4.8 Una ruleta tiene 36 sectores numerados del 1 al 36 de los cuales los 12 primeros son de color rojo, los 12 siguientes de color azul y los 12 últimos de color negro. Calcule la probabilidad que al hacer girar la ruleta salga un número par o un número de color azul.

Capítulo 5

Distribuciones teóricas de probabilidad discretas y continuas

El problema

En el mes de Marzo la ciudad de Guayaquil suele soportar temperaturas de 30 grados centígrados. ¿Hay algún método estadístico que permita, con un poco más de información, determinar la probabilidad de que en un día de ese mes la temperatura esté entre 26 y 29 grados centígrados?

5.1 Introducción

El presente capítulo será dedicada al estudio de las *distribuciones de probabilidad*, las cuales pueden ser definidas diciendo que son *modelos teóricos que permiten calcular la probabilidad de ocurrencia de un suceso determinado, o lo que es lo mismo, una lista de todos los posibles resultados de un experimento y la probabilidad de su ocurrencia .*

Para iniciar este estudio, recordemos el concepto de *distribución de frecuencia* visto en el Capítulo 2.

En el capítulo de referencia señalamos que:

Una distribución de frecuencias no es más que un listado de las frecuencias observadas de todos los posibles resultados de un experimento, elaborada con posterioridad a la realización de dicho experimento.

Basados en esta definición podemos expresar que:

Una distribución de probabilidad es un listado de las probabilidades de ocurrencia de todos los posibles resultados que podrían ser obtenidos en el caso de que este experimento sea llevado a cabo.

Lo anteriormente expuesto es la esencia de la diferencia entre distribuciones de frecuencias y distribuciones de probabilidad.

Para ejemplificar la construcción de una distribución de probabilidad consideremos el experimento aleatorio consistente en medir la cantidad de victorias obtenidas por un equipo de futbol en 4 partidos realizados. La tabla 5.1 muestra los 16 posibles resultados en juegos ganados y perdidos que puede tener el equipo al realizar los 4 partidos.

TABLA 5.1 Combinación de juegos ganados y perdidos en 4 partidos

Resultados posibles	PARTIDOS				Número de victorias
	Primero	Segundo	Tercero	Cuarto	
1	G	G	G	G	4
2	G	G	G	P	3
3	G	G	P	G	3
4	G	P	G	G	3
5	P	G	G	G	3
6	G	G	P	P	2
7	G	P	G	P	2
8	P	G	G	P	2
9	G	P	P	G	2
10	P	G	P	G	2
11	P	P	G	G	2
12	G	P	P	P	1
13	P	G	P	P	1
14	P	P	G	P	1
15	P	P	P	G	1
16	P	P	P	P	0

Observe en la tabla 5.1 lo siguiente:

1.- Existe una sola forma de obtener 4 victorias (1 frecuencia).

2.- Existen cuatro formas diferentes de obtener 3 victorias (4 frecuencias).

3.- Existen seis formas diferentes de obtener 2 victorias (6 frecuencias).

4.- Existen cuatro formas diferentes de obtener 1 victoria (4 frecuencias).

5.- Existe una sola forma de no obtener victorias (1 frecuencia).

La tabla 5.2 muestra la *distribución de probabilidad* correspondiente al ejemplo que estamos considerando. La última columna de la tabla indica la forma de calcular la probabilidad.

TABLA 5.2 Distribución de probabilidad

Número de victorias	Frecuencias	Probabilidad	Cálculo
0	1	0.0625	1/16
1	4	0.25	4/16
2	6	0.375	6/16
3	4	0.25	4/16
4	1	0.0625	1/16
Suma	16		

La figura 5.1 muestra el gráfico correspondiente al número de victorias y su correspondiente probabilidad.

FIGURA 5.1 Probabilidad de obtener entre 0 y 4 victorias en 4 partidos

5.2 Variables aleatorias

Podríamos formular el concepto intuitivo de lo que es una *variable aleatoria* diciendo que es una variable estadística cuyos valores provienen de un experimento aleatorio. Es un valor que se encuentra afectado por el azar. A las variables aleatorias se les conoce también con el nombre de *variables estocásticas*.

Una definición más formal de *variable aleatoria* sería la siguiente:

Una variable aleatoria X es una función real que toma valores en un espacio muestral Ω el cual está asociado a un experimento aleatorio, es decir, $X : \Omega \Rightarrow \Re$, donde $\Re$ representa el conjunto de los números reales.

Ejemplos de variables aleatorias son:

1. Si lanzamos dos monedas al aire, el número de caras resultante del experimento depende del azar y por esta razón el número de caras es una variable aleatoria.
2. El número de llamadas telefónicas recibidas por mi esposa en el Día de la Madre es una variable aleatoria ya que ese número puede ser 0, 1, 2, 3……..
3. La cantidad de lápices defectuosos producidos por una empresa en una jornada laboral.
4. El número de vuelos de Copa Airlines que llegan retrasados a un determinado destino en el lapso de una semana.

5.2.1 Rango de una variable aleatoria

Se llama *recorrido o rango* de una variable aleatoria y se denota como R_X a la imagen de la función X, es decir, al conjunto de valores reales que esta función puede tomar.

Expresado en forma matemática $R_X = \{x \in \Re \mid \exists w \in \Omega : X(w) = x\}$

Para esclarecer lo planteado, calculemos el recorrido o rango que se produce al lanzar dos monedas al aire. Sea X la variable aleatoria "salga cara".

Si se lanzan dos monedas el espacio muestral Ω está conformado por los elementos:

(Cara y Cara) – (Cara y Cruz) – (Cruz y Cara) – (Cruz y Cruz)

Variable aleatoria $X : \Omega \Rightarrow \Re$

Ω		ℜ	
Cara y Cara	⟶	**2**	
Cara y Cruz	⟶	**1**	
Cruz y Cara	⟶	**1**	
Cruz y Cruz	⟶	**0**	Y entonces $\mathbf{R_X = \{0, 1, 2\}}$

5.2.2 Tipos de variables aleatorias

Existen dos tipos de variables aleatorias, las *discretas* y las *continuas*. Veamos la diferencia entre ellas.

Una variable aleatoria se dice que es *discreta* cuando su rango o recorrido es un *conjunto discreto*, es decir, un conjunto formado por elementos en los que se puede reconocer la existencia de un primer elemento, un segundo elemento, un tercer elemento y así sucesivamente. En este caso se dice que el rango o recorrido es un conjunto numerable. En una variable discreta el recorrido está integrado por elementos que tienen entre si una determinada separación.

Ejemplos de variables discretas son:

1. El número de hijos de una pareja. Una pareja puede tener 0, 1, 2, 3,……. hijos, pero nunca 3.7 hijos.
2. El número de viajes al exterior de una persona. Situación similar al ejemplo anterior.
3. El número de calzado. Puede ser 6, 6 y 1/2, 7, 7 y 1/2……., pero nunca 7.38.

Una variable aleatoria se dice que es *continua* cuando su recorrido no es un conjunto numerable, es decir, cuando la variable dentro de un rango admisible puede tomar cualquier valor real.

Ejemplos de variables continuas son:

1. Los ingresos de los empleados de una empresa.
2. El costo del consumo de energía eléctrica en una vivienda.
3. La vida útil de los focos producidos por una empresa.

Resulta necesario precisar la diferencia entre una variable aleatoria y una distribución de probabilidad. La diferencia radica en que la primera se refiere al resultado particular de un experimento, mientras la segunda representa todos los resultados posibles y su correspondiente probabilidad.

5.2.3 Valor esperado (esperanza matemática) y varianza de una distribución de probabilidad discreta

Para una variable aleatoria discreta que toma los valores $x_1, x_2, \ldots, x_n$ con probabilidades $p(x_i)$, la *esperanza matemática, esperanza* o *valor esperado* se calcula mediante la expresión:

$$E[X] = \sum_{1}^{n} x_i \; p(x_i)$$

Si todos los sucesos tienen la misma probabilidad de ocurrencia, entonces el *valor esperado* es la media aritmética.

Se suele escribir $\mu = E[X]$

Para una variable aleatoria discreta que toma los valores $x_1, x_2, \ldots, x_n$ con probabilidades $p(x_i)$, la varianza se calcula mediante la expresión:

$$V(X) = \sum_{1}^{n} p_i (x_i - \mu)^2 \text{ , donde } \mu = E[X] = \sum_{1}^{n} x_i \; p(x_i)$$

5.3 Distribución Binomial

Antes de comenzar a explicar la distribución binomial resulta aconsejable estudiar los llamados *Ensayos de Bernoulli*. Un ensayo de Bernoulli es un experimento aleatorio en el cual solo se pueden obtener dos resultados, los cuales se suelen identificar como éxito **y** *fracaso*, aunque debemos estar claros que es una forma de caracterizar el resultado y por tanto no debe usarse en el sentido literal de la palabra. Este tipo de ensayo recibe este nombre en honor al ilustre matemático y científico suizo *Jakob Bernoulli***.** Estos ensayos son variables aleatorias que solo permiten dos resultados 0 y 1, considerándose a 1 como éxito y a 0 como fracaso.

En un ensayo de Bernoulli si π es la probabilidad de éxito, entonces el valor esperado de la variable aleatoria es π y su varianza π (1-π).

Existen con bastante frecuencia experimentos de carácter repetitivo en los cuales nos interesa solamente registrar la ocurrencia o no de un suceso, por ejemplo, aparición o no de cara al lanzar una moneda, asistencia o no de un empleado a su jornada de trabajo, clientes de un banco que tienen o no tienen cuentas de acumulación, plantas de banano atacadas o no por una plaga, etc.

Estamos al momento en condiciones de comenzar a estudiar la *distribución binomial* expresando que es una distribución de probabilidad discreta que contabiliza el número de éxitos que se obtienen en una serie de n ensayos de Bernoulli, caracterizados por ser independientes entre sí y que tienen una probabilidad de ocurrencia de éxito entre ensayos igual a un valor fijo π.

A modo de resumen, un experimento de probabilidad binomial debe reunir las siguientes características:

1. Los resultados de un experimento binomial son excluyentes y solamente dos, éxito o fracaso.
2. Los experimentos binomiales son independientes, lo cual significa que el resultado de uno no influye en el resultado del otro.
3. La variable aleatoria contabiliza el número de éxitos en una cantidad fija de ensayos.
4. La probabilidad de éxito (π) y la probabilidad de fracaso (1-π) son la misma en cada ensayo.

Si realizamos n repeticiones independientes de un experimento con característica binomial y representamos por x el número de éxitos obtenidos en las n repeticiones del mismo, entonces la probabilidad de obtener x éxitos en n repeticiones viene dada por:

$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$, donde π es la probabilidad de éxito.

${}_nC_x = \frac{n!}{x!(n-x)!}$, representa la combinación de n números tomados x a x.

Con el objetivo de precisar los aspectos estudiados en este numeral, desarrollemos el siguiente ejemplo ilustrativo:

Si la probabilidad que un auto se dañe antes de cumplir el período de garantía es 0.15, determinemos la probabilidad que de 9 autos nuevos que están en garantía:

1. Ningún auto se dañe.
2. 1 auto se dañe.
3. 2 autos se dañen.
4. 3 autos se dañen.
5. 4 autos se dañen.

n = 9 π = 0.15 1 - π = 0.85

- **x = 0**

$$P(x = 0) = \frac{9!}{0!9!}\ (0.15)^0\ (0.85)^9 = \frac{(362880)(1)(0.2316)}{362880} = 0.2316$$

La probabilidad que de 9 autos nuevos ninguno se dañe antes de concluir la etapa de garantía es igual a 0.2316.

- **x = 1**

$$P(x = 1) = \frac{9!}{1!\ 8!}\ (0.15)^1\ (0.85)^8 = \frac{(362880)(0.15)(0.2724)}{40320} = 0.3679$$

La probabilidad que de 9 autos nuevos uno se dañe antes de concluir la etapa de garantía es igual a 0.3679.

- **x = 2**

$$P(x = 2) = \frac{9!}{2!7!}\ (0.15)^2\ (0.85)^7 = \frac{(362880)(0.0225)(0.3206)}{10080} = 0.2597$$

La probabilidad que de 9 autos nuevos dos se dañen antes de concluir la etapa de garantía es igual a 0.2597.

- **x = 3**

$$P(x = 3) = \frac{9!}{3!6!}\ (0.15)^3\ (0.85)^6 = \frac{(362880)(0.0034)(0.3771)}{4380} = 0.1069$$

La probabilidad que de 9 autos nuevos tres se dañen antes de concluir la etapa de garantía es igual a 0.1069.

- **x = 4**

$$P(x = 4) = \frac{9!}{4!5!}\ (0.15)^4\ (0.85)^5 = \frac{(362880)(0.0005)(0.4437)}{2880} = 0.0283$$

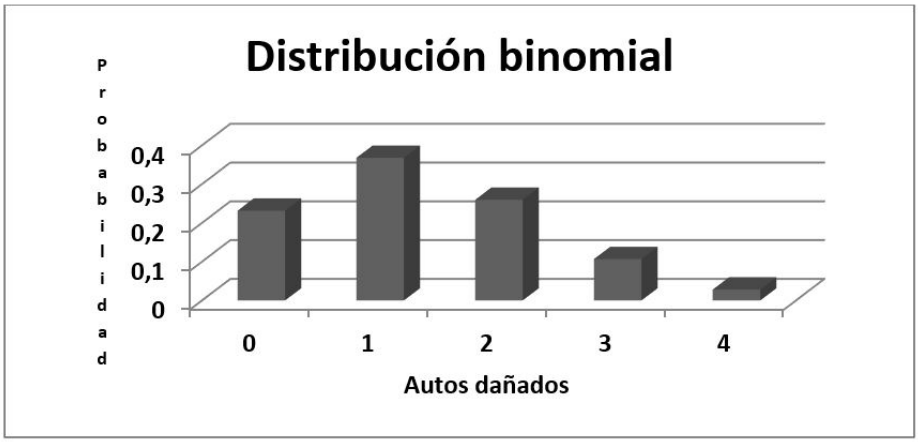
La probabilidad que de 9 autos nuevos cuatro se dañen antes de concluir la etapa de garantía es igual a 0.0283.

La tabla 5.3 muestra la distribución de probabilidad binomial obtenida.

TABLA 5.3 Distribución de probabilidad binomial obtenida

Autos dañados	Probabilidad
0	0.2316
1	0.3679
2	0.2597
3	0.1069
4	0.0283

FIGURA 5.2 Gráfico de la distribución binomial



Ya señalamos que en un ensayo de Bernoulli si π es la probabilidad de éxito, entonces la media o valor esperado de la variable aleatoria es π y su varianza π (1-π).

Como en una distribución binomial se producen n repeticiones independientes, entonces:

$$\mu = n\pi = 9(0.15) = 1.35 \quad \sigma^2 = n\pi(1-\pi) = 9(0.15)(0.85) = 1.15$$

Cuando el valor de n es grande, el cálculo de las probabilidades de una distribución binomial se convierte en una actividad larga y engorrosa. En estos casos, y en general cualquiera sea el valor de n, el cálculo de las probabilidades de la binomial se pueden realizar utilizando la **Tabla T.6** que aparece en el Anexo A, la cual ha sido calculada para diferentes valores de n y π.

A continuación se muestra parte de la tabla de referencia para n = 9 y diferentes valores de π:

	π									
x	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
0	0,630	0,387	0,232	0,134	0,075	0,040	0,021	0,010	0,005	0,002
1	0,299	0,387	0,368	0,302	0,225	0,156	0,100	0,060	0,034	0,018
2	0,063	0,172	0,260	0,302	0,300	0,267	0,216	0,161	0,111	0,070
3	0,008	0,045	0,107	0,176	0,234	0,267	0,272	0,251	0,212	0,164
4	0,001	0,007	0,028	0,066	0,117	0,172	0,219	0,251	0,260	0,246
5	0,000	0,001	0,005	0,017	0,039	0,074	0,118	0,167	0,213	0,246
6	0,000	0,000	0,001	0,003	0,009	0,021	0,042	0,074	0,116	0,164
7	0,000	0,000	0,000	0,000	0,001	0,004	0,010	0,021	0,041	0,070
8	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,004	0,008	0,018
9	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,002

En el ejercicio que acabamos de concluir π = 0.15.

Observe que en la tabla los valores de las probabilidades coinciden (con tres

decimales) con las calculadas en el ejercicio.

5.4 Distribución de Poisson

Otra alternativa cuando el número de repeticiones del experimento (n) es grande, y además π es pequeña, consiste en *aproximar* la Binomial mediante la conocida *distribución de probabilidad de Poisson* la cual viene dada por la expresión:

$$P(x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$$

donde x = 0, 1, 2, ……. e = constante Euler = 2.7183 y $\lambda = n\,\pi$.

Esta distribución se debe al físico y matemático francés Simeón Denis Poisson y en esencia es una distribución de probabilidad discreta que contabiliza la cantidad de veces que un evento se manifiesta durante un determinado intervalo. Se ha demostrado que la aproximación de la distribución binomial mediante la distribución de Poisson es buena cuando $n \geq 20$ o $\pi < 0.05$. Veamos un ejemplo. La probabilidad que un televisor Lcd se dañe al cabo de un período de tiempo es de 0.002. Si se seleccionan al azar 1000 televisores con ese período o más de trabajo, cual es la probabilidad que de ellos 4 se dañen.

Evidentemente estamos en presencia de una distribución binomial en la que el número de repeticiones es bastante grande.

$n = 1000 \geq 20$, $\pi = 0.002 < 0.05$, por tanto, se puede proceder a aproximar la binomial mediante la distribución de Poisson.

$$P(x)=\frac{e^{-2}(2)^{4}}{4!}=\frac{(0.1353)(16)}{24}=0.0902$$

Por tanto, en las condiciones del problema, la probabilidad que cuatro televisores Lcd se dañen es igual a 0.0902.

Al igual que con la distribución binomial, el cálculo de las probabilidades de la distribución de Poisson se pueden realizar utilizando la **Tabla T.7** que aparece en el Anexo A, la cual ha sido calculada para diferentes valores de λ y x. A continuación se muestra parte de la tabla de referencia para diferentes valores de **λ** y x:

	λ								
x	1	2	3	4	5	6	7	8	9
0	0.3679	0.1353	0.0498	0.0183	0.0067	0.0025	0.0009	0.0003	0.0001
1	0.3679	0.2707	0.1494	0.0733	0.0337	0.0149	0.0064	0.0027	0.0011
2	0.1839	0.2707	0.2240	0.1465	0.0842	0.0446	0.0223	0.0107	0.0050
3	0.0613	0.1804	0.2240	0.1954	0.1404	0.0892	0.0521	0.0286	0.0150
4	0.0153	0.0902	0.1680	0.1954	0.1755	0.1339	0.0912	0.0573	0.0337
5	0.0031	0.0361	0.1008	0.1563	0.1755	0.1606	0.1277	0.0916	0.0607
6	0.0005	0.0120	0.0504	0.1042	0.1462	0.1606	0.1490	0.1221	0.0911
7	0.0001	0.0034	0.0216	0.0595	0.1044	0.1377	0.1490	0.1396	0.1171
8	0.0000	0.0009	0.0081	0.0298	0.0653	0.1033	0.1304	0.1396	0.1318

En el ejercicio que acabamos de concluir λ = 2 y x = 4. Observe que en la tabla el valor de la probabilidad coincide con la calculada en el ejercicio.

5.5 La Distribución Normal

La *Distribución Normal* es una *distribución de probabilidad continua*, la cual fue introducida por Abraham de Moivre en el año 1733. El nombre del astrónomo, matemático y físico del siglo XIX Johann Carl Friedrich Gauss, se ha vinculado con la distribución normal porque la utilizó de una forma profunda a raíz de sus trabajos relacionados con la teoría de los errores de medidas físicas, y en especial, al analizar datos astronómicos. Incluso algunos historiadores en el ámbito de la ciencia le otorgan a Gauss un descubrimiento de la distribución independiente del hecho por Moivre.

Debido a la inestimable contribución de Gauss en el desarrollo de la teoría de la distribución normal, y también por la forma física de su gráfico, ésta también es conocida como *distribución Gaussiana* o *Campana de Gauss*.

Una gran cantidad de métodos estadísticos utilizan la distribución normal, razón por la cual ocupa un lugar destacado dentro de la estadística.

Solo con el objetivo de poder brindar una definición formal de la distribución normal, y sin el ánimo de atiborrar al lector con términos matemáticos, pasemos a definir qué se entiende por *función de distribución*.

Una *función de distribución* es aquella función *F* para la que se cumple que

$F(x) = P(X \leq x)$, es decir, dicho *quizás* de una forma *más clara*, es una función a la que llamamos ***F*** **y** que le asigna a cada valor real x la probabilidad de que la variable aleatoria *X* sea menor o igual a **x**. Dicho esto podemos decir que:

Una variable aleatoria continua X sigue una distribución normal si la misma tiene la siguiente *función de distribución*:

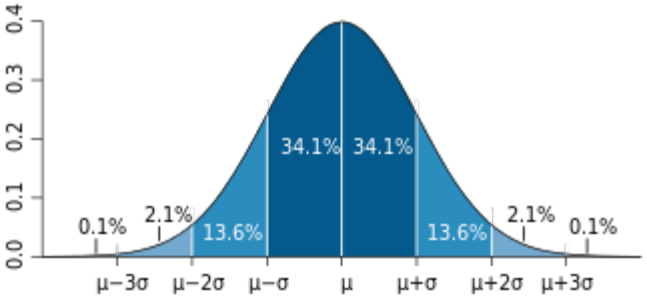
$$f(x)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

donde μ y σ representan la media y la desviación estándar poblacional respectivamente.

La forma resumida X ~ N (μ , σ^2) significa que la variable X sigue una distribución normal con media μ y varianza σ^2.

El gráfico de la distribución de probabilidad normal con media μ y varianza σ^2 se muestra en la figura 5.3.

FIGURA 5.3 Distribución de probabilidad normal



Fuente: es.wikipedia.org

5.5.1 Características de la distribución normal

Observando el gráfico podemos apreciar que algunas propiedades de la distribución normal son:

- Alcanza su valor máximo en la media de la población, la cual se encuentra en el centro de la distribución.
- Es simétrica con relación a la vertical que pasa por su media.
- Los puntos de inflexión de la curva se presentan en $x=\mu-\sigma$ y $x=\mu+\sigma$.
- Es asintótica con relación al eje horizontal. Cuando el valor de X aumenta o disminuye, la curva de la distribución se acerca al eje X sin llegar a tocarlo jamás.
- En los intervalos:
  - $[\mu-\sigma,\mu+\sigma]$, se encuentra aproximadamente el 68.26% de la distribución.
  - $[\mu-2\sigma,\mu+2\sigma]$, se encuentra aproximadamente el 95.44% de la distribución.
  - $[\mu-3\sigma,\mu+3\sigma]$, se encuentra aproximadamente el 99.74% de la distribución.

5.5.2 Distribución de probabilidad normal estándar

Debido a que para definir una distribución normal es necesario establecer el valor de dos parámetros, existe una gran familia de distribuciones normales cada una de ellas definida por una media (μ) y una desviación estándar (σ) específica, lo que hace imposible poder suministrar tablas para el cálculo de probabilidades para esta enorme cantidad de distribuciones normales.

Para suerte de todos nosotros, cualquier distribución de probabilidad normal se puede transformar en otra cuya media siempre es cero y cuya varianza es también siempre 1, y en consecuencia, el cálculo de una probabilidad se reduce al uso de una sola tabla. La distribución de la que hablamos recibe el nombre de *distribución de probabilidad normal estándar*, y se obtiene de la siguiente forma:

Si $X \sim N(\mu, \sigma^2)$ entonces $Z = \frac{X - \mu}{\sigma} \sim N(1,0)$

Es decir, cualquier miembro de la familia de las distribuciones normales puede convertirse en una *distribución de probabilidad normal estándar* restando el valor de la media y el resultado dividiéndolo para la desviación estándar.

Con el objetivo de precisar los aspectos estudiados, veamos algunos ejemplos del cálculo de la probabilidad de ocurrencia de un suceso utilizando la distribución normal estándar.

Supongamos que la edad (X) con la que los estudiantes egresan de la Universidad Laica Eloy Alfaro de Manabí sigue una distribución normal con media poblacional igual a 24 años y una varianza poblacional igual a 0.12 años, es decir, $X \sim N(26, 0.12)$. Calcule la probabilidad que al seleccionar al azar un estudiante recién egresado de esta institución educativa, su edad sea:

- **Menor a 24.5 años.**

$$P(X < 24.5) = P\left(Z < \frac{24.5 - 24}{\sqrt{0.12}}\right) = P\left(Z < \frac{0.5}{0.35}\right) = P(Z < 1.43) = 0.9236$$

A continuación le mostramos un segmento de la **TABLA T.1** del Anexo A**,** en la cual se aprecian las probabilidades de la distribución normal estándar.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147
1.4	0.9192	0.9207	0.9222	**0.9236**	0.9251	0.9265	0.9279	0.9292
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616

- **Mayor a 24.3 años.**

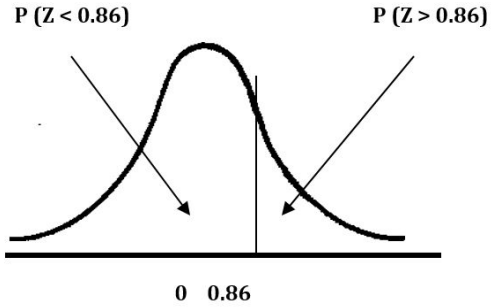
Debemos calcular P (X > 24.3). Tipificando:

$Z = \frac{24.3 - 24}{0.35} = \frac{0.3}{0.35} = 0.86$ es decir, debemos calcular P (Z > 0.86), la cual gráficamente implica obtener el área bajo la curva normal que se encuentra a la derecha del percentil 0.86.

Las áreas a la derecha de un percentil no se encuentran tabuladas, razón por la cual haremos uso de las propiedades gráficas de la curva para encontrar la probabilidad deseada.

Veamos gráficamente la situación:

FIGURA 5.4 Áreas bajo la curva normal



Como el área total bajo la curva es igual a 1, entonces:

P (Z < 0.86) + P (Z > 0.86) = 1 o lo que es lo mismo:

P (Z > 0.86) = 1 – P (Z < 0.86)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	**0.8051**	0.8078

y como P (Z < 0.86) = 0.8051 entonces:

P (X > 24.3) = P (Z > 0.86) = 1 – 0.8051 = 0.1949

De manera general, si a es un percentil no negativo (>=0) de una distribución normal, entonces:

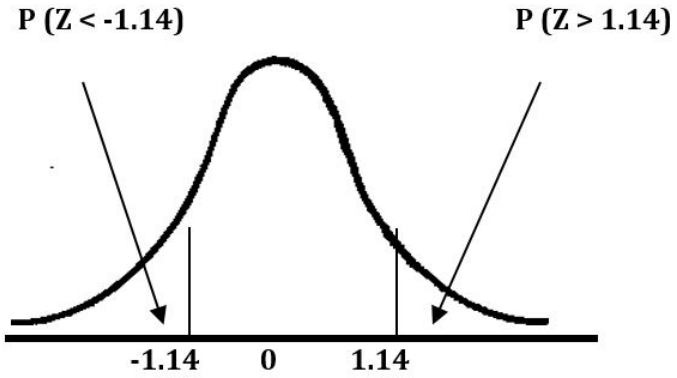
P (Z > a) = 1 – P (Z < a)

- **Menor a 23.6 años.**

Debemos calcular P (X < 23.6). Tipificando:

$Z = \frac{23.6 - 24}{0.35} = -1.14$ es decir, debemos calcular P (Z < -1.14), la cual gráficamente implica obtener el área bajo la curva normal que se encuentra a la izquierda del percentil –1.14. Las áreas correspondientes a percentiles negativos no se encuentran tabuladas, razón por la cual debemos nuevamente hacer uso de las propiedades gráficas de la curva. Veamos gráficamente la situación:

FIGURA 5.5 Áreas bajo la curva normal



En el gráfico, P (Z < -1.14) es exactamente igual a P (Z > 1.14) debido a la simetría de la curva con relación a 0, es decir, P (Z < -1.14) = P (Z > 1.14)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
1.1	0.8643	0.8665	0.8686	0.8708	**0.8729**	0.8749	0.8770	0.8790

P (Z > 1.14) = 1- P(Z < 1.14) y como P (Z < 1.14) = 0.8729 entonces:

P (X < 23.6) = P (Z < -1.14) = 1 – 0.8729 = 0.1271

De manera general, si **a** es un percentil no negativo (>=0) de una distribución normal, entonces:

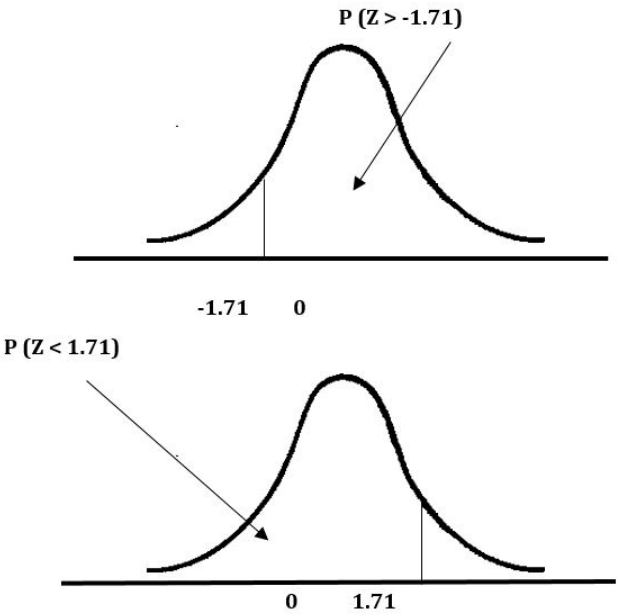
P (Z < -a) = 1 – P (Z < a)

- **Mayor a 23.4 años.**

Debemos calcular P (X > 23.4). Tipificando $Z = \frac{23.4 - 24}{0.35} = -1.71$, es decir, debemos calcular P (Z > -1.71), la cual gráficamente implica obtener el área bajo la curva normal que se encuentra a la derecha del percentil –1.71.

Esta área no se encuentra tabulada, razón por la cual debemos nuevamente hacer uso de las propiedades gráficas de la curva.

FIGURA 5.6 Áreas bajo la curva normal



En los dos gráficos anteriores se puede apreciar que P (Z > -1.71) es exactamente igual a P (Z < -1.71) debido a la simetría de la curva, es decir:

P (Z > -1.71) = P (Z < 1.71)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
1.7	0.9554	**0.9564**	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616

y como P (Z < 1.71) = 0.9564 entonces:

P (X > 23.4) = P (Z > -1.71) = 0.9564

De manera general, si **a** es un percentil no negativo (>=0) de una distribución normal, entonces:

P (Z > -a) = P (Z < a)

- **Sea menor a 24.3 y mayor a 23.5.**

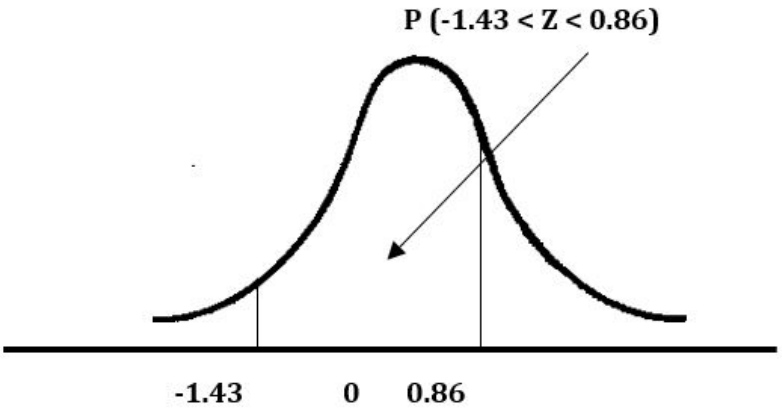
Debemos calcular P (23.5 < X < 24.3). Tipificando:

$$\frac{23.5-24}{0.35} < Z < \frac{24.3-24}{0.35} \qquad -1.43 < Z < 0.86$$ es decir, debemos calcular

P (-1.43 < Z < 0.86).

Gráficamente puede apreciarse que:

FIGURA 5.7 Áreas bajo la curva normal



P (-1.43 < Z < 0.86 = P (Z < 0.86) – P (Z < -1.43) =

P (Z < 0.86) – [1 – P (Z < 1.43)] =

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	**0.8051**	0.8078

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
1.4	0.9192	0.9207	0.9222	**0.9236**	0.9251	0.9265	0.9279	0.9292

P (Z < 0.86) – 1 + P (Z < 1.43) = 0.8051 – 1 + 0.9236 = 0.7287

de donde P (23.5 < X < 24.3) = P (-1.43 < Z < 0.86) = 0.7287

De manera general, si **a** y **b** son percentiles no negativos (>=0) de una distribución normal, entonces:

P (b < Z < a) = P (Z < a) – P (Z < b)

5.5.3 Aproximación de la normal a la binomial

En el numeral **5.3** estudiamos que *"cuando el valor de n es grande, el cálculo de las probabilidades de una distribución binomial se convierte en una actividad larga y engorrosa".* Vimos que en estos casos una primera alternativa era hacer uso de la **TABLA T.6** que aparece en el Anexo A y una segunda opción consistía en aproximar la binomial a través de la distribución de Poisson.

A continuación describiremos una tercera posibilidad la cual consiste en hacer una *aproximación de la normal a la binomial*.

Observe en los gráficos que se muestran en las figuras 5.8, 5.9 y 5.10, cómo para

un valor fijo de π = 0.35, la forma de la distribución binomial se va acercando a una distribución normal en la medida en que n aumenta. Los gráficos que se muestran en dichas figuras corresponden a valores de n iguales a 2, 5 y 10.

FIGURA 5.8 Distribución binomial para π = 0.35 y n = 2

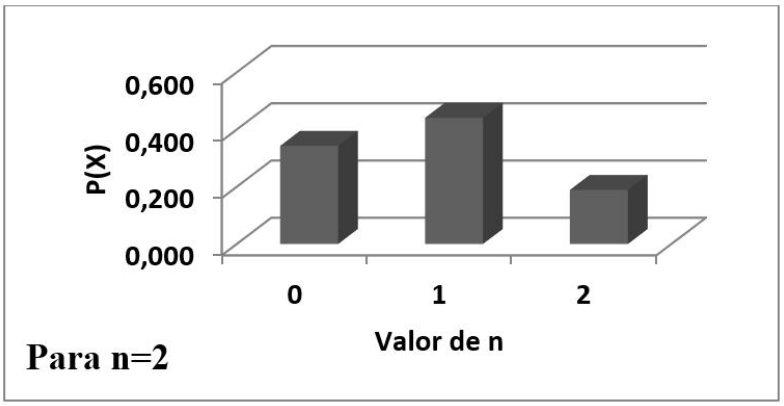


FIGURA 5.9 Distribución binomial para π = 0.35 y n = 5

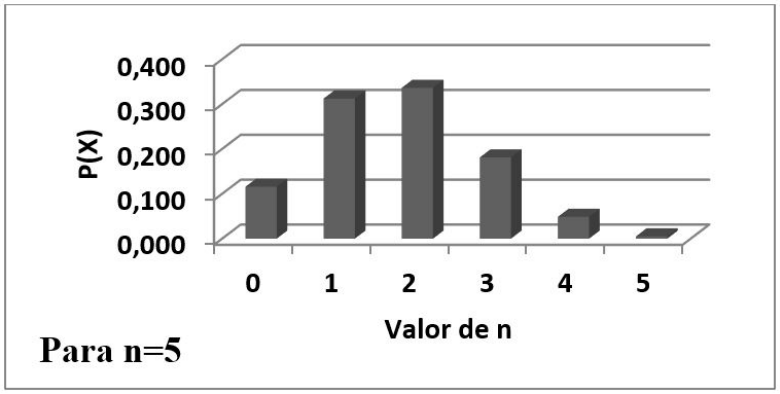
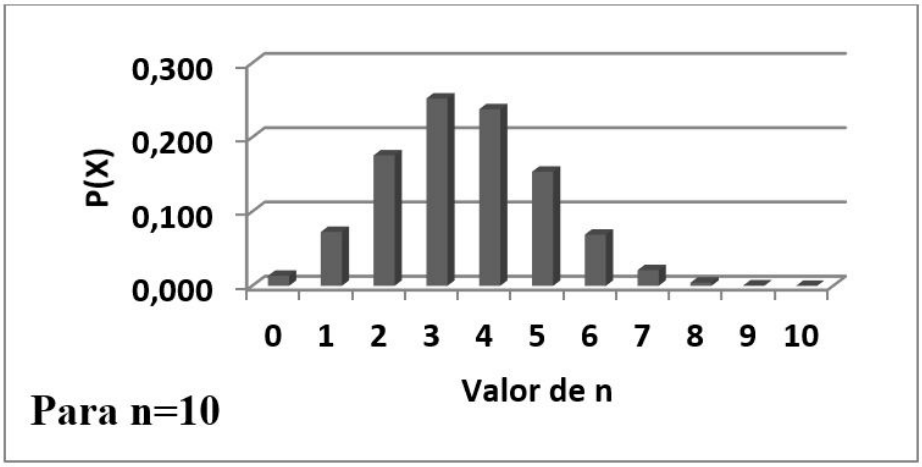


FIGURA 5.10 Distribución binomial para π = 0.35 y n = 10



Por lo señalado anteriormente parece lógico que podamos aproximar la normal a la binomial. Algunos autores sugieren que esta aproximación es adecuada cuando tanto nπ como n (1-π) son mayores o iguales a 5.

Antes de iniciar el cálculo de una probabilidad utilizando la aproximación normal en lugar de la binomial, es necesario tomar en cuenta que la primera es una distribución de probabilidad continua y la segunda es una distribución de probabilidad *discreta.*

Frank Yates, estadístico inglés del siglo XX, propuso un *factor de corrección*, el cual lleva su nombre, y consiste en sumar o restar, según sea el caso, 0.5 unidades a la variable al momento de hacer la aproximación.

La regla es la siguiente:

Probabilidad binomial		**Probabilidad normal**
$P(X = a)$	=	$P(a - 0.5 < X' < a + 0.5)$
$P(X \le a)$	=	$P(X' \le a + 0.5)$
$P(X < a)$	=	$P(X' \le a - 0.5)$
$P(X \ge a)$	=	$P(X' \ge a - 0.5)$
$P(X > a)$	=	$P(X' > a + 0.5)$

Y entonces $Z = \dfrac{(X \pm 0.5) - \mu}{\sigma}$

Precisemos lo estudiado con un ejemplo:

El 3% de los pernos producidos por un equipo resultan defectuosos. Calcule la probabilidad que de un lote de 3000 pernos nuevos,

a) 95 resulten defectuosos.

Según los datos, μ = nπ = 3000 (0.03) = 90 y

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{3000(0.03)(0.97)} = 9.34$$

$$P(X = 95) = P(95 - 0.5 < X' < 95 + 0.5) = P(94.5 < X' < 95.5)$$

Estandarizando la variable X':

$$P\left(\frac{94.5-90}{9.34} < Z < \frac{95.5+90}{9.34}\right) = P(0.48 < Z < 0.59)$$

$$= P(Z < 0.59) - P(Z < 0.48) = 0.7224 - 0.6844 = 0.038.$$

b) 93 o menos resulten defectuosos.

$$P(X \le 93) = P(X' \le 93 + 0.5) = P(X' \le 93.5)$$

Estandarizando la variable X':

$$P\left(Z \le \frac{93.5-90}{9.34}\right) = P(Z \le 0.37) = 0.6443.$$

c) Menos de 98 resulten defectuosos.

$$P(X<98)=P(X'<98-0.5)=P(X'<97.5)$$

Estandarizando la variable X':

$$P\left(Z<\frac{97.5-90}{9.34}\right)=P(Z<0.80)=0.7881.$$

d) 92 o más resulten defectuosos.

$$P(X\geq 92)=P(X'\geq 92-0.5)=P(X'\geq 91.5)$$

Estandarizando la variable X':

$$P\left(Z\geq\frac{91.5-90}{9.34}\right)=P(Z\geq 0.16)=1-P(Z<0.16)=1-0.5636=0.4364$$

e) Más de 88 resulten defectuosas.

$$P(X>88)=P(X'>88+0.5)=P(X'>88.5)$$

Estandarizando la variable X':

$$P\left(Z>\frac{88.5-90}{9.34}\right)=P(Z>-0.16)=P(Z<0.16)=0.5636.$$

f) Entre 84 y 97 resulten defectuosas.

$$P(84<X<97)=P(X<97)-P(X<84)=P(X'<97-0.5)-P(X'<84-0.5)$$

Estandarizando la variable X':

$$P\left(Z<\frac{96.5-90}{9.34}\right)-P\left(Z<\frac{83.5-90}{9.34}\right)=P(Z<0.70)-P(Z<-0.70)$$

$$=P(Z<0.70)-1+P(Z<0.70)$$

$$=0.7580-1+0.7580=0.5160$$

5.6 La distribución *F*

La *distribución F* es una distribución de probabilidad continua conocida también con el nombre de distribución F de Fisher, en honor a *Ronald Aylmer Fisher*, destacado científico inglés de los siglos XIX y XX. Algunos autores también le llaman *distribución de Fisher - Snedecor* en reconocimiento al matemático y estadístico estadounidense de los siglos XIX y XX *George Waddel Snedecor*. En el año 1924, Ronald A. Fisher presentó esta distribución, la cual es de inestimable utilidad para establecer procedimientos

de inferencia usando la razón entre dos varianzas muestrales.

En esencia la naturaleza de la distribución *F* de Fisher es la siguiente:

Si S_1^2 y S_2^2 son varianzas calculadas a partir de muestras aleatorias independientes de tamaño n_1 y n_2, extraídas de poblaciones distribuidas normalmente con varianzas σ_1^2 y σ_2^2 respectivamente, entonces la variable aleatoria

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$$

sigue una distribución *F* de Fisher con $n_1 - 1$ y $n_2 - 1$ grados de libertad en el numerador y denominador respectivamente.

Cuando las varianzas de ambas poblaciones son iguales entonces:

$$F = \frac{S_1^2}{S_2^2}$$

La distribución *F* posee las siguientes características:

1. **Es no negativa**. El menor valor que toma es 0.
2. **Es asintótica con relación al eje horizontal**. Cuando el valor de X aumenta, la curva de la distribución se acerca al eje X sin llegar a tocarlo jamás.
3. **Tiene sesgo positivo**. La curva tiene una cola larga hacia la derecha. Cuando el número de grados de libertad del numerador y del denominador aumentan la distribución se aproxima a la normal.
4. **La *F* es una familia de distribuciones.** Un miembro en particular de la familia queda definido por el número de grados de libertad del numerador y del denominador.

La distribución *F* tiene dos aplicaciones principales.

1. Permite establecer si dos muestras provienen de poblaciones que tienen la misma varianza. Este método estadístico se denomina *prueba de hipótesis* y será tratado en un capítulo posterior.
2. Someter a una prueba de hipótesis la igualdad de un conjunto de medias poblacionales. Este método estadístico se denomina *análisis de varianza* y será tratado en un capítulo posterior.

En la **TABLA T.3** del Anexo A aparecen los valores críticos de la distribución *F*, la cual estudiaremos con profundidad en un capítulo posterior.

Ejercicios del capítulo

5.1 La probabilidad de que un conductor no lleve puesto el cinturón de seguridad es igual a 0.12.

a) Si un agente de tránsito detiene a 10 conductores al azar, determine la probabilidad que de los conductores detenidos:

1. Ninguno tenga puesto el cinturón.
2. Uno tenga puesto el cinturón.
3. Dos tengan puesto el cinturón.
4. Tres tengan puesto el cinturón.
5. Cuatro tengan puesto el cinturón.

b) Obtenga la distribución binomial correspondiente.

c) Elabore el gráfico de la distribución binomial obtenida.

5.2 La probabilidad de que un equipo en un almacén de maquinarias se encuentre defectuoso es igual a 0.3.

a) Si se extraen 12 equipos del almacén, determine la probabilidad que de los equipos extraídos:

1. Ninguno esté defectuoso.
2. Uno esté defectuoso.
3. Dos estén defectuosos.
4. Tres estén defectuosos.
5. Cuatro estén defectuosos.

b) Obtenga la distribución binomial correspondiente.

c) Elabore el gráfico de la distribución binomial obtenida.

5.3 Una agencia de seguros le vende pólizas a 5 propietarios de vehículos nuevos. La agencia conoce que la probabilidad de que uno de estos autos tenga un siniestro en los próximos 10 años es igual a 0.17. Determine la probabilidad que después de transcurridos los 10 años,

1. Al menos tres vehículos hayan sufrido un siniestro.
2. Como máximo dos vehículos hayan sufrido un siniestro.
3. Al menos un vehículo haya sufrido un siniestro.

5.4 El 57% de los profesionales que laboran en una entidad bancaria son economistas. Si se eligen al azar 8 profesionales de esa entidad, determine la probabilidad de que:

1. Al menos cinco sean economistas.
2. Como máximo cuatro sean economistas.
3. Al menos uno sea economista.

5.5 La probabilidad de que una prueba de alcoholemia realizada a un conductor dé positiva es igual a 0.008. Si un agente de tránsito detiene a 1000 conductores al azar, determine la probabilidad que de los conductores detenidos:

1. Ninguno resulte con la prueba de alcoholemia positiva.
2. Uno resulte con la prueba de alcoholemia positiva.
3. Dos resulten con la prueba de alcoholemia positiva.
4. Tres resulten con la prueba de alcoholemia positiva.
5. Cuatro resulten con la prueba de alcoholemia positiva.

5.6 La probabilidad de morir a lo largo de la vida a causa de un accidente laboral es igual a 0.012. Calcule la probabilidad de que en una empresa con 500 trabajadores:

1. Ninguno tenga un accidente laboral a lo largo de su vida.
2. Uno tenga un accidente laboral a lo largo de su vida.
3. Dos tengan un accidente laboral a lo largo de su vida.
4. Tres tengan un accidente laboral a lo largo de su vida.
5. Cuatro tengan un accidente laboral a lo largo de su vida.

5.7 Considere que las calificaciones (X) obtenidas por los estudiantes de una Facultad de Ciencias Económicas en un examen final de estadística sigue una distribución normal con una media poblacional igual a 8.5 y una varianza poblacional igual a 0.85. Calcule la probabilidad que al seleccionar al azar a un estudiante de dicha facultad, la calificación obtenida en el examen sea:

1. Menor a 8.8.
2. Mayor a 8.6.
3. Menor a 8.3.
4. Mayor a 8.2.
5. Menor a 8.7 y mayor a 8.4.

5.8 El número de minutos promedio diarios de utilización del servicio de internet de los clientes de la Corporación Nacional de Telecomunicaciones sigue una distribución normal con media poblacional igual a 86 y varianza poblacional igual a 222. Calcule la probabilidad que al seleccionar al azar a un cliente de CNT, el número de minutos promedio diarios de utilización del servicio de internet sea:

1. Menor a 89.
2. Mayor a 87.
3. Menor a 83.
4. Mayor a 82.
5. Menor a 88 y mayor a 85.

5.9 La probabilidad de que una máquina envasadora de leche en polvo se exceda (X) en la cantidad de gramos de este producto que debe depositar en cada envase es igual a 0.04. Calcule la probabilidad que de un lote de 4000 envases:

1. 165 resulten con exceso de peso.
2. 163 o menos resulten con exceso de peso.
3. Menos de 167 resulten con exceso de peso.
4. 162 o más resulten con exceso de peso.
5. Más de 155 resulten con exceso de peso.
6. Entre 153 y 164 resulten con exceso de peso.

5.10 La probabilidad de que un equipo de cómputo se dañe antes de que se venza su garantía es de 0.03. Calcule la probabilidad que de una muestra de 3000 equipos de cómputo:

1. 140 se dañen antes de que se venza su garantía.
2. 135 o menos se dañen antes de que se venza su garantía.
3. Menos de 143 se dañen antes de que se venza su garantía.
4. 136 o más se dañen antes de que se venza su garantía.
5. Más de 130 se dañen antes de que se venza su garantía.
6. Entre 128 y 142 se dañen antes de que se venza su garantía.

Capítulo 6

Muestreo y distribuciones de muestreo

El problema

El Gobierno ecuatoriano ha decidido sustituir las actuales cocinas a gas existentes en el país por cocinas de inducción, y en este sentido está interesado en conocer el grado de aceptación que tendrían dichas cocinas de inducción por parte de todos los hogares del país. La interrogante es:

¿En qué forma debe ser seleccionada la muestra para que el objetivo planteado por el Gobierno pueda ser alcanzado?

6.1 Introducción

Cualquier proceso investigativo que desarrollemos tiene un objetivo central, *conocer el comportamiento de una característica a nivel poblacional*. Pero en la mayor parte de las ocasiones este objetivo no puede ser alcanzado a menos que estimemos su valor mediante una muestra extraída de esa población. De manera general, existen cinco razones que en la práctica no permiten estudiar a todos los elementos de una población, y aconsejan solo estudiar los elementos de una muestra extraída de ella. Esas cinco razones son las siguientes:

1. Poder acceder a todos los elementos de la población pudiera requerir un tiempo demasiado grande.
2. En la mayoría de los casos estudiar todos los elementos de la población resulta altamente costoso.
3. En algunas ocasiones medirle un indicador a un elemento de la población puede tener un carácter destructivo.
4. Hay poblaciones sumamente grandes y algunas inclusive de tamaño infinito.
5. En muchas ocasiones, aunque sea posible, estudiar todos los elementos de la población puede resultar excesivo, y una muestra pudiera ser representativa de ella.

En este capítulo iniciaremos el estudio del *muestreo*, el cual es una herramienta estadística que nos permitirá inferir un valor poblacional sobre la base de la muestra extraída.

De forma general, el muestreo puede ser dividido en dos grandes e importantes partes:

1. El tipo de muestreo.

Es la parte del muestreo que nos permite determinar ante una situación concreta

la forma más adecuada de extraer los elementos de la muestra.

2. **El tamaño de la muestra.**

Parte del muestreo que nos facilita determinar la cantidad de elementos de la muestra, de forma tal, que la inferencia realizada tenga una determinada confiabilidad.

En el presente capítulo solo estudiaremos *los tipos de muestreo* más utilizados y algunas *distribuciones muestrales*, dejando *el tamaño de la muestra* para capítulos posteriores.

6.2 Métodos de muestreo

Existen cuatro tipos de muestreo fundamentales:

1. **El muestreo aleatorio simple.**
2. **El muestreo aleatorio sistemático.**
3. **El muestreo aleatorio estratificado.**
4. **El muestreo por conglomerados.**

Pasemos a estudiar cada uno de estos métodos por separado.

6.2.1 Muestreo aleatorio simple

Como su nombre lo indica es el más simple de todos los tipos de muestreo, y en esencia, consiste en seleccionar los elementos de la muestra de forma tal que cada individuo de la población tenga la misma probabilidad de ser elegido.

A modo de ejemplo, supongamos una población conformada por 600 clientes de una empresa y de la cual deseamos extraer una muestra de 50 clientes mediante un muestreo aleatorio simple. Un método satisfactorio para extraer esta muestra es identificando a cada uno de los clientes con un número escrito en un pequeño papel y depositar estos papeles en una caja u otro recipiente cualquiera. Después de mezclar adecuadamente todos los papeles se procede a extraerlos uno a uno sin mirar el interior del recipiente, siguiendo el proceso hasta que se hayan extraído los 50 clientes que integran la totalidad de la muestra.

Otro método *"quizás"* un poco más técnico para extraer esta muestra es identificando a cada cliente de la forma que explicamos anteriormente y utilizando ***"una" tabla de números aleatorios*** como la que se muestra en la **TABLA T.8** del Anexo A, la cual permite que todos los elementos de la población tengan la misma probabilidad de ser seleccionados.

En el párrafo anterior hemos escrito *una tabla* y no *la tabla* puesto que no existe una tabla *"única"* de números aleatorios, ya que, como tal, la tabla en sí, también es aleatoria.

Con el objetivo de ejemplificar cómo seleccionar una muestra aleatoria simple

utilizando una tabla de números aleatorios, a continuación presentamos un segmento de una de ellas.

55133	02146	47361	38334	68477	16419	45059
76479	43969	60921	96963	43709	72625	00267
52707	23409	65911	42673	55278	91093	89919
89487	08474	77111	06716	36269	87105	19261
16880	57646	17957	94212	22987	75512	25590
10382	81608	57990	69473	16748	68865	56894
15227	26288	81202	63471	15311	41805	34311
36023	16559	08389	90665	42430	60165	72330
19080	80699	24311	87541	90053	35686	67122
15697	11793	04840	64444	04844	35912	35256
49717	60306	77889	41285	01900	23574	65326
12449	84770	79756	80698	89599	99857	26690

El paso inicial para seleccionar los clientes que integrarán la muestra es determinar *un punto de partida* en la tabla. Lo más sencillo para escoger este punto consiste en cerrar los ojos y con la punta de un lápiz o algo similar señalar un elemento de la tabla.

Supongamos que procediendo de esta forma seleccionamos como punto de partida al número *06716.* El primer cliente elegido para la muestra es el *67*, pues al ser el tamaño de la muestra igual a 600 solo necesitamos los tres primeros dígitos del número aleatorio. Para elegir al segundo cliente que integrará la muestra se puede continuar en cualquier sentido, es decir de abajo hacia arriba o a la inversa, de izquierda a derecha o viceversa. Supongamos que lo hacemos de izquierda a derecha y por tanto el próximo cliente elegido es el identificado con el número *362*, el siguiente cliente el *192* y así sucesivamente hasta escoger a todos los elementos de la muestra. Observe que en la tabla omitimos el número aleatorio *87105*. La razón fue que al ser el tamaño de la población igual a 600 el cliente identificado con el número *871* no existe.

6.2.2 Muestreo aleatorio sistemático

En ocasiones el muestreo aleatorio simple resulta poco práctico. Por ejemplo, si en una Empresa Bananera desean sacar una muestra de tamaño 200 del número de manos por racimo en una hectárea que tiene una densidad de 2400 plantas, la utilización del muestreo aleatorio simple requeriría que se numere cada planta antes de hacer uso de la tabla de números aleatorios, lo cual implicaría mucho tiempo y esfuerzo. En situaciones como ésta es aconsejable utilizar el muestreo aleatorio sistemático.

Este muestreo consiste en calcular en primer término un valor *k* que sería el resultado de dividir el tamaño de la población (2400) entre el tamaño de la muestra (200), es decir:

$k = \frac{2400}{200} = 12$. Si el valor de k no es un número entero entonces será necesario redondearlo.

Hecho esto seleccionamos la primera planta tal y como procedimos en el muestreo aleatorio simple, es decir, seleccionamos un número en la tabla de números aleatorios entre 1 y k, en este caso, entre 1 y 12, supongamos que resultó ser 8. Entonces a partir de la planta 8, se seleccionan las plantas 20 = (8+12), 32 = (20+12), 44 = (32+12) y de esa forma hasta muestrear las 200 plantas, es decir, de forma *sistemática* cada 8 plantas.

En este caso concreto el muestreo podría hacerse comenzando por el primer surco hasta concluir en el último de ellos.

6.2.3 Muestreo aleatorio estratificado

Una encuesta realizada por el Instituto Nacional de Estadísticas y Censos (INEC) en Quito, Guayaquil, Cuenca, Ambato y Machala reveló que de 9.744 hogares analizados el 11,2% representa al estrato medio alto; el 22,8% al medio típico; y el 49,3% al medio bajo. En tanto, el 1,9% de los hogares estudiados pertenece a la clase alta; y el 14,9% a la baja.

En esta encuesta reportada por el Telégrafo, el INEC divide a las clases sociales en grupos llamados *estratos.*

Consideremos que un almacén de electrodomésticos en Ambato, desea encuestar a 500 hogares para conocer su interés en la adquisición de una refrigeradora que tienen en promoción. Conocedores de que no todas las personas en Ambato tienen la misma capacidad adquisitiva, dividió a la población en **estratos**, garantizando de esta manera que cada grupo social se encuentre representado en la muestra. Supongamos que el número de hogares a muestrear es el que se indica en la tabla 6.1.

TABLA 6.1 Hogares a muestrear por tipo de estrato de la muestra de 500

Clase	Hogares	Proporción	Muestra
Alta	779	0,02	10
Media alta	4592	0,11	55
Media típica	9348	0,23	115
Media baja	20213	0,49	245
Baja	6109	0,15	75

Observe que de haberse utilizado un muestreo aleatorio simple pudo haber ocurrido que ningún hogar de clase alta hubiera sido seleccionado (2% de probabilidad), mientras que con toda seguridad la mayor parte de los hogares seleccionados en la muestra serían de clase media baja (49% de probabilidad). El muestreo aleatorio estratificado no permite que esto ocurra pues garantiza que al menos un hogar por

estrato será seleccionado en la muestra.

Una vez definidos los estratos la selección de la muestra en cada grupo se hace aplicando el muestreo aleatorio simple o el muestreo aleatorio sistemático, en dependencia del caso.

6.2.4 Muestreo por conglomerados

Este método de muestreo es usualmente utilizado cuando se desea extraer una muestra en un área geográfica que tiene la característica de ser muy dispersa.

Por ejemplo, supongamos que se desea encuestar a los habitantes de la provincia de Manabí con relación a la opinión que tienen acerca de la gestión de la Prefectura de la provincia. Poder encuestar de forma individual a cada uno de los habitantes de la región sería en exceso laborioso y costoso, por tanto podríamos aplicar el muestreo por conglomerados y dividir a la provincia en *unidades primarias* (podrían ser los cantones).

De esta forma concentraríamos la atención en seleccionar de forma aleatoria un grupo de estos cantones, y dentro de ellos, seleccionar una muestra aleatoria de sus habitantes los cuales serían encuestados con la finalidad que se persigue.

6.3 Distribuciones muestrales

Una distribución muestral es una distribución de probabilidad de un estadígrafo muestral, el cual es calculado tomando como base todas las muestras posibles de tamaño n, seleccionadas al azar de una población determinada. Dicho de otra manera, *la distribución muestral es lo que resulta de considerar todas las muestras posibles de una población. Su estudio permite calcular la probabilidad que se tiene, dada una sola muestra, de acercarse al parámetro de la población (es.wikipedia.org).*

6.3.1 Distribución muestral de la media

Consideremos que obtenemos todas las posibles muestras de tamaño n de una determinada población y que calculamos la media de cada una de estas muestras. El conjunto de los números obtenidos de esta forma pueden ser considerados como los valores de una variable aleatoria. A la distribución de esta variable aleatoria se le llama *distribución muestral de la media.*

Demostremos que si $\bar{x}$ es la media de una muestra aleatoria de tamaño n de una determina población, entonces la media y la varianza de la distribución muestral de $\bar{x}$ son μ y $\frac{\sigma^2}{n}$ donde μ y σ^2 son la media y la varianza de la población de donde provienen las muestras.

Si representamos por M(X) a la media de la población y por V(X) a su respectiva varianza, entonces M(X) = μ y V(X) = σ^2 y aplicando las propiedades de la media y la varianza:

$$M(\bar{x}) = M\left(\sum_1^n \frac{x_i}{n}\right) = \sum_1^n \frac{M(x_i)}{n} = \frac{n\mu}{n} = \mu$$

$$V(\bar{x}) = V\left(\sum_1^n \frac{x_i}{n}\right) = \sum_1^n \frac{V(x_i)}{n} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Si las muestras provienen de una población con distribución normal o la población no es normal pero n es lo suficientemente grande, entonces: $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Es decir, $\bar{x}$ sigue una distribución normal con media μ y varianza $\frac{\sigma^2}{n}$. Este resultado se conoce como *teorema del límite central* y es de gran importancia en la estadística. Por tanto, la variable $\bar{x}$ tipificada, $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ sigue una distribución normal con media cero y varianza 1.

Los percentiles correspondientes a la distribución normal tipificada pueden ser encontrados en la **TABLA T.1** del Anexo A.

En el siguiente ejemplo se describe de forma sencilla la construcción de una distribución muestral de la media. Desarrollaremos el ejemplo para un tamaño de muestra igual a 2, sin embargo, el procedimiento es válido para cualquier otro tamaño de muestra que decidamos. La tabla 6.2 muestra las calificaciones en Estadística I de 8 estudiantes de la Facultad de Ciencias Económicas de la Universidad Laica Eloy Alfaro de Manabí.

TABLA 6.2 Calificaciones de 8 estudiantes en la materia de Estadística I

Nombres	Notas
Juan	8
Raúl	10
Pedro	7
Ada	8
Eva	9
Luis	10
Carlos	8
María	10

1.- Calculemos la media de la población.

$$\mu = \frac{8+10+7+8+9+10+8+10}{8} = 8.75$$

2.- Consideremos todas las medias posibles de tamaño 2.

$${}_8C_2 = \frac{8!}{2!6!} = \frac{8x7x6x5x4x3x2}{2x6x5x4x3x2} = 28\,muestras$$

Las medias muestrales de todas las posibles muestras de tamaño 2 se pueden apreciar en la tabla 6.3

TABLA 6.3 Medias muestrales de todas las posibles muestras de tamaño 2

Nombres	Notas	Media		Nombres	Notas	Media
Juan, Raúl	8+10	9		Pedro, Eva	7+9	8
Juan, Pedro	8+7	7.5		Pedro, Luis	7+10	8.5
Juan, Ada	8+8	8		Pedro, Carlos	7+8	7.5
Juan, Eva	8+9	8.5		Pedro, María	7+10	8.5
Juan, Luis	8+10	9		Ada, Eva	8+9	8.5
Juan, Carlos	8+8	8		Ada, Luis	8+10	9
Juan, María	8+10	9		Ada, Carlos	8+8	8
Raúl, Pedro	10+7	8.5		Ada, María	8+10	9
Raúl, Ada	10+8	9		Eva, Luis	9+10	9.5
Raúl, Eva	10+9	9.5		Eva, Carlos	9+8	8.5
Raúl, Luis	10+10	10		Eva, María	9+10	9.5
Raúl, Carlos	10+8	9		Luis, Carlos	10+8	9
Raúl, María	10+10	10		Luis, María	10+10	10
Pedro, Ada	7+8	7.5		Carlos, María	8+10	9

La distribución muestral de la media para n = 2 se muestra en la tabla 6.4.

TABLA 6.4 Distribución muestral de la media para n = 2

Media muestral	Número de medias	Probabilidad
7.5	3	0.1071
8	4	0.1429
8.5	6	0.2143
9	9	0.3214
9.5	3	0.1071
10	3	0.1071

3.- Calculemos la media de la distribución muestral de la media.

Si denotamos por $\mu_{\bar{X}}$ a la media de la distribución muestral de la media

tenemos: $\mu_{\bar{X}} = \frac{9+7.5+8+.....+9+10+9}{28} = \frac{245}{28} = 8.75$

Observe que la media de la distribución muestral de la media, es exactamente igual a la media poblacional, es decir, $\mu_{\bar{X}} = \mu = 8.75.$

A modo de resumen de lo realizado podemos señalar que:

La media de la distribución muestral de la media, es exactamente igual a la media de la población.

La dispersión de la distribución muestral de la media, es menor que la de la población.

La distribución muestral de la media tiende a mostrar una forma de campana, aproximándose a una distribución normal.

Lo expresado en los tres numerales anteriores ya lo habíamos considerado cuando estudiamos el **teorema del límite central**.

FIGURA 6.1 Gráfico de barras de la distribución poblacional

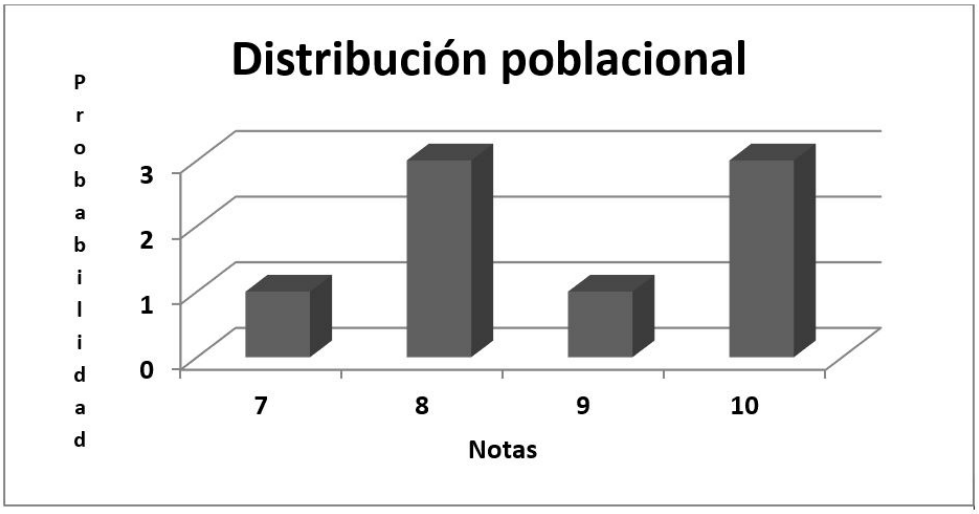
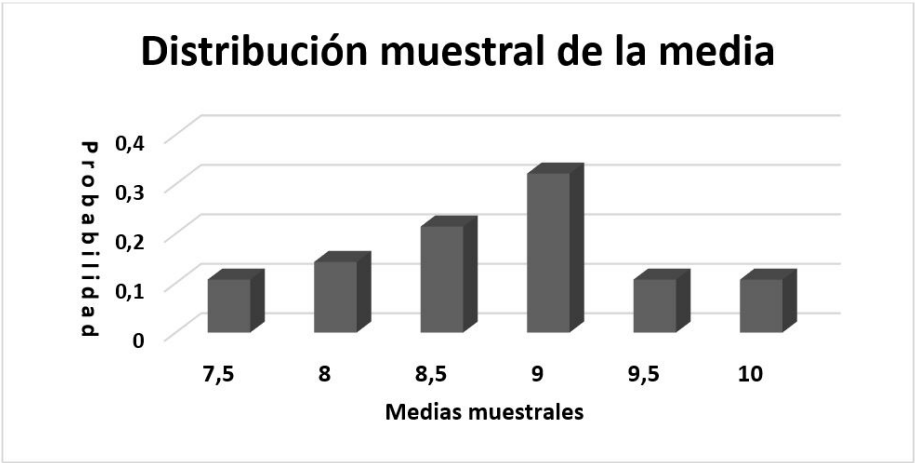


FIGURA 6.2 Gráfico de barras de la distribución muestral



6.3.2 Distribución t (t de Student)

En el numeral anterior vimos que: $Z = \dfrac{\overline{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0,1)$

Pero por regla general, el valor de la desviación estándar poblacional no es conocido, y en consecuencia se hace necesario utilizar la desviación estándar muestral, dando lugar a la nueva variable:

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

La distribución de t es conocida como *distribución t de Student* o simplemente *distribución t.* La distribución t fue introducida por el estadístico William Sealy Gosset el cual nació en Nueva Zelandia en el siglo XIX. Gosset realizó publicaciones bajo el seudónimo de Student en el año 1908.

Entre la distribución t y la distribución normal existe una determinada relación. Esta relación viene dada por los siguientes aspectos:

1. Ambas son simétricas con respecto a la media y por regla general la distribución t es más plana que la normal.
2. La distribución t alcanza una menor altura en la media que la normal, y por el contrario, la primera es mayor en los extremos que la segunda.
3. Hay una distribución t diferente para cada tamaño posible de muestra y cuando ésta es mayor a 30, se vuelve aproximadamente igual a la normal.
4. Por lo expuesto en el numeral anterior, para tamaños de muestra mayores a 30 se suele utilizar la distribución normal en lugar de la t de Student.

6.3.2.1 Grados de libertad

Hemos señalado que para cada tamaño posible de muestra existe una distribución t diferente, o dicho en términos estadísticos, existe una distribución t distinta para cada uno de los posibles *grados de libertad*. Este término estadístico fue introducido por Ronald Fisher y reviste una gran importancia dentro de la inferencia estadística.

Pasemos a precisar este concepto, y para ello, supongamos un conjunto formado por tres elementos de los cuales conocemos que su media es igual a 36. Sean a, b, y c estos elementos.

$$\frac{a+b+c}{3} = 36 \text{ o lo que es lo mismo a + b + c =108}$$

Si seleccionamos a = 45 y b = 38, entonces el valor de c no puede ser escogido

arbitrariamente sino que *tiene* que tomar el valor:

$$a+b+c=108 \quad c=108-45-38 \quad c=25$$

En este conjunto formado por tres datos podremos asignarles valores arbitrarios a dos de ellos, el valor del tercer dato queda establecido.

Se dice entonces que el conjunto formado por estos tres elementos tiene 2 *grados de libertad* ya que somos libres de establecer el valor de solamente dos de estos elementos, quedando el valor del tercer elemento inequívocamente establecido.

En general un conjunto formado por n elementos tiene *n – 1 grados de libertad*.

Para la determinación del percentil de una distribución t de Student específica, podemos utilizar la **TABLA T.2** del Anexo A, la cual estudiaremos en un capítulo posterior.

6.3.2 Aplicación de la distribución muestral de la media

La compañía Nestlé, y en especial Nescafé, envasan el café soluble en polvo de la marca DOLCA en tachos que deben contener un peso neto de 170 gramos. Sin embargo, en ocasiones la cantidad real de café en los tachos puede presentar una diferencia con relación al peso neto que esta debe tener, y en dependencia de la magnitud de esa diferencia, la compañía puede perder confiabilidad por parte de sus clientes si es que el peso está por debajo de lo establecido, o perder utilidades, si es que el peso está por encima.

Por tal motivo, un especialista de control de la calidad extrajo una muestra del peso neto de 20 tachos de café la cual se muestra en la tabla 6.5.

TABLA 6.5 Peso neto de 20 tachos de café

170.4	170.3	170	169.8	170.2	169.4	170.4	170.1	169.6	170.3
169.8	170.2	170.8	170	170.6	171	170.4	170.6	170	170.1

Por otra parte, el especialista tiene datos históricos que le permiten estimar que el peso neto de los tachos de café de la marca DOLCA sigue una distribución normal con una desviación estándar poblacional igual a 0.46 gramos.

Con esta información, ¿podemos llegar a la conclusión de que es probable que se esté envasando el café DOLCA con un peso neto superior al que tiene establecido la compañía, el cual es de 170 gramos?

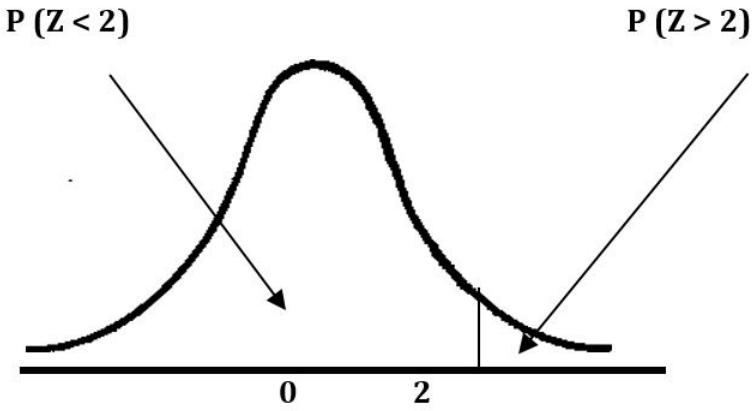
Sea X la variable peso neto del tacho de café DOLCA:}

$$\overline{X}=\frac{3404.00}{20}=170.2 \quad \mu=170 \quad \sigma=0.46 \quad n=20$$

$$P(\bar{X} > 170.2) = P\left(Z > \frac{170.2 - 170}{\frac{0.46}{\sqrt{20}}} \right) = P\left(Z > \frac{0.2}{0.1} \right) = P(Z > 2)$$

Gráficamente estamos ante la siguiente situación:

FIGURA 6.3 Áreas bajo la curva normal



Como el área bajo la curva es igual a 1, entonces:

P (Z > 2) = 1 – P (Z < 2)

En la **TABLA T1** del Anexo A podemos encontrar que:

P (Z < 2) = 0.9772 de donde P (Z > 2) = 0.0228 o 2.28%

Esta probabilidad del 2.28% significa que es bastante poco probable que se pueda seleccionar una muestra de 20 observaciones con una media igual a 170.2 gramos, de una población normal con media igual a 170 gramos y desviación estándar igual a 0.46 gramos.

Debemos entonces concluir que los resultados del control de calidad efectuado indican que los tachos de café DOLCA están siendo envasados con un peso neto superior a 170 gramos, lo cual está provocando en consecuencia una pérdida de utilidades de la compañía.

6.3.3 Distribución de la diferencia entre dos medias muestrales de poblaciones normalmente distribuidas con varianzas σ_1^2 y σ_2^2 conocidas

Partiendo del mismo razonamiento utilizado en el numeral relacionado con la distribución muestral de la media, podemos establecer que la distribución muestral de la diferencia entre dos medias muestrales, calculadas a partir de muestras aleatorias independientes de tamaño n_1 y $n_{2,}$ extraídas de dos poblaciones distribuidas normalmente con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 conocidas, estará también distribuida

normalmente con media μ_1 - μ_2 y varianza $\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$.

Si n_1 y n_2 son grandes, entonces la distribución muestral de la diferencia entre las dos medias será aproximadamente normal con la media y la varianza señalada, con independencia de la forma funcional de las poblaciones originales.

6.3.4 Distribución de la diferencia entre dos medias muestrales de poblaciones normalmente distribuidas con varianzas σ_1^2 y σ_2^2 desconocidas e iguales

Si $\bar{x}_1$ y s_1^2 son respectivamente la media y la varianza de una muestra de tamaño n_1 extraída de una población distribuida normalmente con media μ_1 y varianza σ_1^2 y si $\bar{x}_2$ y s_2^2 son respectivamente la media y la varianza de una muestra de tamaño n_2 extraída de otra población distribuida normalmente con media μ_2 y varianza σ_2^2, igual a σ_1^2, entonces la expresión:

$$\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_c^2}{n_1}+\frac{s_c^2}{n_2}}}=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{s_c^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

está distribuida según una t de Student, con $n_1 + n_2 - 2$ grados de libertad, donde:

$s_c^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ se denomina varianza muestral combinada.

6.3.5 Distribución de la diferencia entre dos medias muestrales de poblaciones normalmente distribuidas con varianzas σ_1^2 y σ_2^2 desconocidas y desiguales

Si $\bar{x}_1$ y s_1^2 son respectivamente la media y la varianza de una muestra de tamaño n_1 extraída de una población distribuida normalmente con media μ_1 y varianza σ_1^2, y si $\bar{x}_2$ y s_2^2 son respectivamente la media y la varianza de una muestra de tamaño n_2 extraída de otra población distribuida normalmente con media μ_2 y varianza σ_2^2, distinta de σ_1^2, entonces la expresión:

$\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ está distribuida según una t de Student con los grados de libertad

calculados mediante la expresión:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

6.3.6 Distribución de una proporción muestral

La distribución muestral de una proporción p calculada a partir de muestras aleatorias de tamaño n extraídas de una población en la que la proporción poblacional es π, tiene una media igual a p y una varianza igual a $\frac{p(1-p)}{n}$

Si n es lo suficientemente grande, la distribución muestral de p se aproxima a una distribución normal.

6.3.7 Distribución de la diferencia entre dos proporciones muestrales

Si p_1 es una proporción muestral calculada a partir de todas las muestras aleatorias de tamaño n_1 que se pueden extraer de una población con parámetro π_1 y p_2 es una proporción muestral calculada a partir de todas las muestras aleatorias de tamaño n_2 que se pueden extraer de una población con parámetro π_2, entonces la distribución muestral de $p_1 - p_2$ tiene una media igual a $\pi_1 - \pi_2$ y una varianza igual a $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

Si n_1 y n_2 son lo suficientemente grandes, entonces la distribución muestral de $p_1 - p_2$ se aproxima a una distribución normal.

Ejercicios del capítulo

6.1 Los datos que se muestran a continuación representan el precio de venta al público de un detergente en siete diferentes establecimientos de una ciudad.

ESTABLECIMIENTO	P.V.P.
A	3.37
B	3.33
C	3.35
D	3.39
E	3.31
F	3.35
G	3.33

Construya la distribución muestral de la media para un tamaño de muestra igual a 2.

6.2 Con los mismos datos del ejercicio anterior, construya la distribución muestral de la media para un tamaño de muestra igual a 3.

6.3 La tabla que aparece a continuación muestra las calificaciones obtenidas por 5 estudiantes de una Facultad de Administración de Empresas en una evaluación de Estadística Aplicada.

Nombres	Notas
Juan	8
Raúl	10
Pedro	7
Ada	8
Eva	9

Construya la distribución muestral de la media para un tamaño de muestra igual a 3.

6.4 Con los mismos datos del ejercicio anterior, construya la distribución muestral de la media para un tamaño de muestra igual a 3.

6.5 Una variable aleatoria X sigue una distribución normal con media 340 y varianza poblacional igual a 20. Calcule la probabilidad que al seleccionar al azar una muestra de tamaño 24, su media sea:

1.- Menor a 342.

2.- Menor a 339.

3.- Mayor a 341.

4.- Mayor a 338.

6.6 Una variable aleatoria X sigue una distribución normal con media 143 y varianza poblacional igual a 18. Calcule la probabilidad que al seleccionar al azar una muestra de tamaño 24, su media sea:

1.- Menor a 145.
2.- Menor a 142.
3.- Mayor a 144.
4.- Mayor a 141.

Capítulo 7
Estimación e intervalos de confianza

El problema

Al gerente del centro comercial Paseo Shopping de la ciudad ecuatoriana de Manta, le interesa estimar el gasto promedio poblacional en compra de calzado de los clientes que visitan el local Payless. ¿cuál es la forma más adecuada de hacer esta estimación teniendo la garantía de hacerlo con una confiabilidad aceptable?

7.1 Introducción.

En el capítulo anterior, el cual fue dedicado al muestreo y a las distribuciones de muestreo, estudiamos cinco razones que en la práctica no permiten estudiar a todos los elementos de una población, y aconsejan solo estudiar los elementos de una muestra extraída de ella. Esas cinco razones fueron las siguientes:

1. Poder acceder a todos los elementos de la población pudiera requerir un tiempo demasiado grande.
2. En la mayoría de los casos estudiar todos los elementos de la población resulta altamente costoso.
3. En algunas ocasiones medirle un indicador a un elemento de la población puede tener un carácter destructivo.
4. Hay poblaciones sumamente grandes y algunas inclusive de tamaño infinito.
5. En muchas ocasiones, aunque sea posible, estudiar todos los elementos de la población puede resultar excesivo, y una muestra pudiera ser representativa de ella.

A menudo se necesita conocer el valor de un parámetro y resulta de suma complejidad su obtención, debido a cualquiera de las cinco razones señaladas anteriormente. Por ejemplo, un investigador podría estar interesado en conocer el ingreso promedio mensual de todas las mujeres que residen en la ciudad de Guayaquil.

Resultaría prácticamente imposible contactar a todas las mujeres de esta ciudad para hallar el valor promedio de sus ingresos. Una solución pudiera ser determinar un tamaño de muestra adecuado y promediar los ingresos de las mujeres seleccionadas en la muestra, considerando entonces el resultado obtenido como representativo del valor poblacional. En este caso se dice que hemos hecho una *estimación puntual* del valor promedio poblacional de los ingresos.

En la mayoría de las veces una estimación puntual de un parámetro resulta

insuficiente porque al hacerse solo puede ocurrir una de dos situaciones, o la estimación puntual realizada es correcta o no lo es, pero además si no es correcta nunca sabremos en que magnitud nos hemos equivocado al hacer la estimación.

Por otra parte, al hacer una estimación puntual nunca se tiene la certeza de la confiabilidad de dicha estimación y resulta una incógnita el posible éxito o no que hayamos obtenido. Podríamos entonces concluir que una estimación puntual sería realmente utilizable siempre y cuando se nos dé acompañada por una estimación del posible error que podríamos estar cometiendo.

La solución a este problema conduce a la utilización de otro método de estimación llamado *estimación por intervalo o intervalo de confianza*, mediante el cual podemos obtener un intervalo cuyos extremos son funciones de la muestra, es decir, variables aleatorias entre las cuales con determinada probabilidad se halla el verdadero valor del parámetro estimado.

Si designamos con la letra griega θ al parámetro que deseamos estimar, entonces el procedimiento consiste en tomar una muestra aleatoria de la correspondiente población y hallar un intervalo aleatorio [I , D] donde I representa el extremo izquierdo del intervalo y D representa el extremo derecho, de tal forma que la probabilidad de que el verdadero valor del parámetro esté dentro del intervalo calculado sea lo *suficientemente grande*, es decir,

$P \{ \theta \varepsilon [I , D] \} = 1 - \alpha$

donde α es un valor lo suficientemente pequeño y que definiremos con una mayor precisión en el próximo capítulo. A α *se le denomina nivel de significación* y a $1 - \alpha$ *se le conoce como nivel de confianza* de la estimación por intervalo. Se dice entonces que [I , D] es un intervalo de confianza del $100(1 - \alpha)\%$, o lo que es lo mismo, una estimación por intervalo de θ con un nivel de confianza de $1 - \alpha$.

Los valores de $1 - \alpha$ deben ser lo más cercano posible a 1. Usualmente se escogen los valores de $1 - \alpha$ iguales a 0.95, 0.99 o 0.999, o lo que es lo mismo, α igual a 0.05, 0.01 o 0.001. Por supuesto que otros valores de α pueden ser escogidos según los intereses del investigador.

Al describir el procedimiento para construir un intervalo de confianza resulta necesario considerar dos posibles casos:

- Se conoce el valor de la varianza poblacional
- Se desconoce el valor de la varianza poblacional

7.2 Intervalo de confianza para la media μ de una población distribuida normalmente con varianza σ^2 conocida.

Supongamos que disponemos de una muestra aleatoria de tamaño n de una

población con distribución normal con media poblacional μ y varianza poblacional σ^2 conocida, entonces:

$$I = \left[\bar{x} - \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \right]$$ es un intervalo de confianza para estimar μ con un nivel de confianza de 1 - α , donde $Z_{1-\frac{\alpha}{2}}$ es el percentil de orden $1-\frac{\alpha}{2}$ de la distribución normal estándar.

Para demostrarlo basta con probar que la probabilidad de que μ pertenezca al intervalo I es igual a 1 - α, es decir,

$$P\{\mu \ \varepsilon \ I\} = P\left[\bar{x} - \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \le \mu \le \bar{x} + \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \right] =$$

$$P\left[- \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \le \mu - \bar{x} \le \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \right] =$$

$$P\left[- Z_{1-\frac{\alpha}{2}} \le \frac{\mu - \bar{x}}{\frac{\sigma}{\sqrt{n}}} \le Z_{1-\frac{\alpha}{2}} \right] = N\left(Z_{1-\frac{\alpha}{2}} \right) - N\left(- Z_{1-\frac{\alpha}{2}} \right) =$$

$$N\left(Z_{1-\frac{\alpha}{2}} \right) - \left(1 - N\left(Z_{1-\frac{\alpha}{2}} \right) \right) = N\left(Z_{1-\frac{\alpha}{2}} \right) - 1 + N\left(Z_{1-\frac{\alpha}{2}} \right) =$$

$$1 - \frac{\alpha}{2} - 1 + 1 - \frac{\alpha}{2} = 1 - \alpha$$

El proceso anterior nos permite demostrar lo que nos habíamos propuesto.

Veamos un ejemplo. En una investigación desarrollada con el objetivo de estimar el gasto promedio mensual por persona en consumo de bebidas alcohólicas en una importante ciudad del Ecuador, se obtuvo una muestra de tamaño 25 que dio como resultado una media de 60 dólares. Suponiendo que este indicador tiene una varianza poblacional de 6 dólares, obtenga un intervalo de confianza del 95% para la media poblacional del gasto mensual en consumo de alcohol en la mencionada ciudad.

$$\bar{x} = 60 \quad n = 25 \quad \sigma = \sqrt{6} = 2.45 \quad 1 - \alpha = 0.95 \quad \alpha = 1 - 0.95 = 0.05$$

$$1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 1 - 0.025 = 0.975$$

Haciendo uso de la **TABLA T1** del Anexo A encontramos que el percentil de la distribución normal estándar con un área a su izquierda igual a 0.975 es 1.96. El

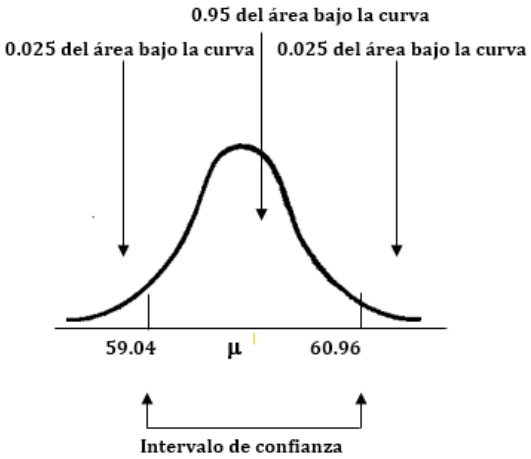
intervalo de confianza viene dado por:

$$\left[60-\frac{2.45}{\sqrt{25}}(1.96)\ ,\ 60+\frac{2.45}{\sqrt{25}}(1.96)\right] = [60-0.96\ ,\ 60+0.96] = [59.04,\ 60.96]$$

lo cual significa que con un 95% de confiabilidad podemos asegurar que el gasto promedio a nivel poblacional en consumo de bebidas alcohólicas en la ciudad estudiada, está entre 59.04 y 60.96 dólares, o lo que es lo mismo, $59.04 \leq \mu \leq 60.96$ con un nivel de confianza del 95 %.

En la figura 7.1 se muestra de forma gráfica el intervalo de confianza obtenido:

FIGURA 7.1 Intervalo de confianza



Intervalo de confianza

Observe como el nivel de significación 0.05 se encuentra equitativamente *distribuido* en ambos extremos de la curva, es decir, $\frac{\alpha}{2}=\frac{0.05}{2}=0.025$ del área en cada una de las *colas* de la curva.

7.3 Intervalo de confianza para la media μ de una población distribuida normalmente con varianza σ^2 desconocida.

Cuando la varianza de la población es desconocida, se hace necesario estimar el valor de la misma a través de la varianza de una muestra extraída de esta población, y en ese caso, una estimación por intervalo de confianza de la media μ de dicha población cuya distribución es normal y viene dada por:

$$\left[\bar{x}\ -\ \frac{s}{\sqrt{n}}\ t_{\alpha}^{(n-1)},\ \bar{x}\ +\frac{s}{\sqrt{n}}\ t_{\alpha}^{(n-1)}\right]$$

donde s es la desviación estándar de la muestra, n el tamaño de la misma y

$t_{\alpha}^{(n-1)}$ el percentil para una prueba de dos colas de la distribución t de Student con n-1 grados de libertad y un nivel de significación α.

Veamos un ejemplo en el que se presenta esta situación. Una empresa que produce focos incandescentes desea investigar la durabilidad en meses de los mismos, y para ello, extrajo una muestra de la durabilidad de 20 focos en la que obtuvo los resultados que se muestran en la tabla 7.1.

TABLA 7.1 Muestra de la durabilidad de 20 focos incandescentes

11.7	12.3	12.1	11.6	12.2	12.3	11.7	11.6	12.3	11.8
12.3	12.5	12.1	11.6	11.7	12.1	12.6	11.4	11.7	12.1

Construya un intervalo de confianza del 99% para la media poblacional de la durabilidad de los focos producidos por la empresa.

$$1-\alpha = 0.99 \quad \alpha = 1-0.99 = 0.01$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{239.7}{20} = 11.98 \qquad s^2 = \frac{\sum \left(x_i - \bar{x} \right)^2}{n-1} = \frac{2.29}{19} = 0.12$$

$$s = \sqrt{0.12} = 0.35$$

En la **TABLA T.2** del Anexo A se observa que el percentil de la distribución t de Student para una prueba de dos colas, $\alpha = 0.01$ y 19 grados de libertad es igual a 2.861.

La expresión del intervalo de confianza viene dado por:

$$\left[11.98 - \frac{0.35}{\sqrt{20}}(2.861) \, , \, 11.98 + \frac{0.35}{\sqrt{20}}(2.861) \right]$$

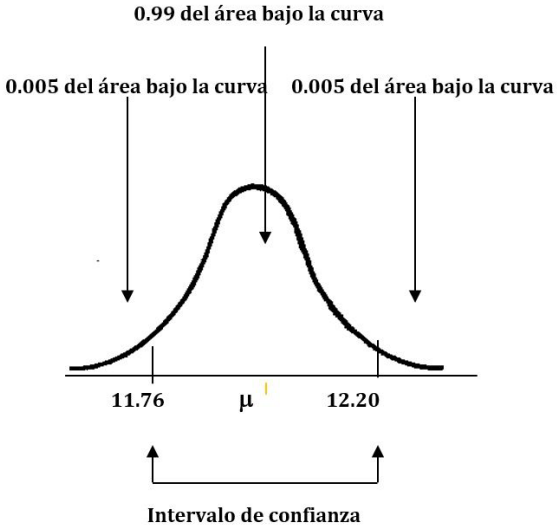
$$\left[11.98 - 0.22 \, , \, 11.98 + 0.22 \right]$$

$$[11.76, 12.20]$$

El resultado obtenido nos permite concluir que con un 99% de confiabilidad podemos asegurar que la durabilidad promedio a nivel poblacional de los focos producidos por la empresa, está entre 11.76 y 12.20 horas, o lo que es lo mismo:
$11.76 \leq \mu \leq 12.20$ con un nivel de confianza del 99 %.

Como ya conocemos, para n suficientemente grande ($n \geq 30$), la distribución t puede ser aproximada por la distribución normal, por lo que en lugar del percentil t podría haberse utilizado, *de haber sido el caso*, el percentil correspondiente de la normal. En la figura 7.2 se muestra de forma gráfica el intervalo de confianza obtenido.

FIGURA 7.2 Intervalo de confianza



7.4 Intervalo de confianza para una proporción.

El contenido estudiado hasta el momento en el presente capítulo ha sido presentado utilizando siempre variables de tipo cuantitativas, es decir, variables que representan características que pueden ser cuantificadas o expresadas mediante un valor numérico, por ejemplo, los gastos en dólares del consumo de bebidas alcohólicas y la durabilidad en horas de focos incandescentes.

Sin embargo, resulta en muchas ocasiones necesario estimar un valor poblacional de una variable de tipo cualitativa nominal, es decir, una variable que se refiere a atributos que no pueden ser representados con números. Ejemplos de este tipo son los siguientes:

- Aproximadamente el 35% (35 de cada 100) de los graduados de la Facultad de Ciencias Económicas de la Universidad Laica Eloy Alfaro de Manabí poseen al menos un título de 4to nivel.
- El 87% (87 de cada 100) de las viviendas de la ciudad de Portoviejo posee alcantarillado público.
- Una empresa que produce focos incandescentes afirma que la proporción de este producto con una durabilidad mayor a un año es igual a 0.63 (63 de cada 100).
- La proporción de personas de la 3era edad con bajos recursos económicos en un sector del país es de 0.28 (28 de cada 100).

En este numeral trabajaremos con variables cualitativas nominales cuyas observaciones pueden ser clasificadas en dos grupos que son mutuamente excluyentes. Por ejemplo,

- Un graduado universitario tiene título de 4to nivel o no.

- Una vivienda posee alcantarillado público o no.
- Un foco incandescente dura más de un año o no.
- Una persona de la 3era edad es de bajos recursos o no.

Observe como en un párrafo reciente nos hemos referido a cifras que han sido dos de ellas expresadas en porcentajes y las otras dos en términos de *proporción*. La razón es que resulta equivalente hablar de:

- 35% de graduados o de una proporción de graduados igual a 0.35.
- 87% de viviendas o una proporción de viviendas igual a 0.87.
- Una proporción de durabilidad igual a 0.63 o un 63% de durabilidad.
- Una proporción de la 3era edad igual a 0.28 o un 28% de personas de la 3era edad.

Una *proporción* es un valor que señala la parte de la muestra de una población que tiene un rasgo distintivo que la identifica, es decir, $p = \frac{X}{n}$ donde X es el número de elementos de la muestra que poseen el rasgo distintivo y n el tamaño de la muestra.

Al igual que a las variables cuantitativas, a las variables cualitativas nominales se les puede estimar el valor de la proporción poblacional a través de un intervalo de confianza.

Una estimación por intervalo de confianza para la proporción poblacional π de elementos con cierta característica en una población, viene dada por:

$p - \sqrt{\frac{p(1-p)}{n}}\ Z_{1-\frac{\alpha}{2}}$, $p + \sqrt{\frac{p(1-p)}{n}}\ Z_{1-\frac{\alpha}{2}}$ donde p es la proporción correspondiente a una muestra de tamaño n.

Desarrollemos un ejemplo. Supongamos que un director provincial de salud desea estimar la proporción poblacional de fármacos antidiarreicos que se encuentran caducados en un almacén de la provincia. Debido a que el almacén es relativamente grande y tiene una gran cantidad de este tipo de fármaco, extrajo una muestra de 150 de ellos, de los cuales detectó que 21 habían caducado. Construya un intervalo de confianza del 99% para estimar la proporción poblacional de fármacos antidiarreicos caducados dentro del almacén.

$n = 150 \quad X = 21 \qquad p = \frac{21}{150} = 0.14$

$\alpha = 0.01 \qquad 1 - \frac{0.01}{2} = 0.995 \qquad Z_{0.995} = 2.58$

$$0.14 - \sqrt{\frac{(0.14)(1-0.14)}{150}}(2.58)\ ,\ 0.14 + \sqrt{\frac{(0.14)(1-0.14)}{150}}(2.58)$$

$[\ 0.14 - 0.07\ ,\ 0.14 + 0.07\ ]$

$[\ 0.07\ ,\ 0.21\ ]$ o sea, $0.07 \leq \pi \leq 0.21$

Podemos entonces asegurar con un nivel de confiabilidad del 99 % que el porcentaje de fármacos caducados en el almacén se encuentra entre un 7% y un 21%.

7.5 Factor de corrección para poblaciones finitas.

En el cálculo de los intervalos de confianza realizados hasta el momento hemos supuesto que la población ha sido grande o infinita. Sin embargo, en ocasiones podemos encontrarnos ante la situación de que el tamaño de la población es pequeño y debemos estimar una media o una proporción poblacional de dicha distribución. En casos como éste se hace necesario hacer ajustes en la determinación del error estándar tanto de las medias muestrales como de las proporciones muestrales. El ajuste consiste en multiplicar el error estándar por el *factor de corrección para una población finita*, el cual viene dado por la expresión:

$$\sqrt{\frac{N-n}{N-1}}$$

De esta forma, para poblaciones finitas el ajuste del error estándar queda de la siguiente forma:

- Para estimar una media poblacional

$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ para varianza poblacional conocida.

$\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ para varianza poblacional desconocida.

- Para estimar una proporción poblacional

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

Llegado a este punto cabe preguntarnos, ¿cuando la población es pequeña o finita es siempre necesario realizar el ajuste planteado? Para dar respuesta a esta

pregunta hagamos los cálculos que se muestran en la tabla 7.2 tomando como base una población de tamaño 800:

TABLA 7.2 Factor de corrección para diferentes valores de n y N = 800

n	$\frac{n}{N}$	$\sqrt{\frac{N-n}{N-1}}$
10	0.013	0.9944
20	0.025	0.9880
40	0.050	0.9753
80	0.100	0.9493
160	0.200	0.8950
320	0.400	0.7751
640	0.800	0.4475

Como puede apreciarse en la tabla, cuando la fracción $\frac{n}{N}$ es menor a 0.05, es decir, cuando el tamaño de la muestra es menor que el 5% del tamaño de la población, el efecto del factor de corrección no es significativo y la respuesta a la pregunta formulada sería que el ajuste es solo necesario cuando la razón $\frac{n}{N}$ es mayor a 0.05, o dicho de otra manera, cuando el tamaño de la muestra es mayor que el 5% del tamaño de la población.

Ejemplo 1:

Con el objetivo de estimar el ingreso promedio mensual de 300 familias de clase baja de un suburbio de Guayaquil, se extrajo una muestra de tamaño 20 la cual arrojó los resultados que se muestran en la tabla 7.3

TABLA 7.3 Muestra del ingreso promedio mensual de 300 familias

454	423	438	465	451	433	448	462	433	450
437	457	461	460	439	448	450	441	453	460

Con una confiabilidad del 99.9%, obtenga un intervalo de confianza para estimar el ingreso promedio mensual de todas las familias del suburbio estudiado.

Observe en primer lugar que la población es finita, N = 300.

$\frac{n}{N}=\frac{20}{300}=0.067$ es decir, el tamaño de la muestra representa el 6.7% del tamaño

de la población, por tanto, se requiere hacer el ajuste para población finita.

La expresión para calcular el intervalo de confianza quedaría como:

$$\left[\bar{x} - \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\ t_{\alpha}^{(n-1)},\ \bar{x} + \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\ t_{\alpha}^{(n-1)}\right]$$

$$1-\alpha = 0.999 \quad \alpha = 1-0.999 = 0.001$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{8963}{20} = 448.15 \qquad s^2 = \frac{\sum\left(x_i - \bar{x}\right)^2}{n-1} = \frac{2526.5}{19} = 132.97$$

$$s = \sqrt{132.97} = 11.53 \qquad \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{300-20}{300-1}} = 0.97$$

El percentil de la distribución t de Student para una prueba de dos colas, $\alpha = 0.001$ y 19 grados de libertad es igual a 3.883, lo cual puede comprobarse haciendo uso de la **TABLA T.2** del Anexo A.

La expresión del intervalo de confianza viene dado por:

$$\left[448.15 - \frac{11.53}{\sqrt{20}}(0.97)(3.883)\ ,\ 448.15 + \frac{11.53}{\sqrt{20}}(0.97)(3.883)\right]$$

$$\left[448.15 - 9.71\ ,\ 448.15 + 9.71\right]$$

[438.44 , 457.86]

El resultado obtenido nos permite concluir que con un 99.9% de confiabilidad podemos asegurar que el promedio de ingreso mensual a nivel poblacional de las familias que residen en el suburbio estudiado, está entre 438.44 y 457.86 dólares, o lo que es lo mismo, 438.44 ≤ μ ≤ 457.86 con un nivel de confianza del 99.9 %.

Ejemplo 2:

Para ingresar a laborar en una importante empresa de productos de belleza resulta necesario aprobar un examen de habilidades manuales. En la última convocatoria hecha por la empresa para someterse a éste tipo de evaluación, se presentaron un total de 280 personas de las cuales 12 no sobrepasaron la evaluación. Construya un intervalo de confianza del 95% para la proporción poblacional de solicitantes a laborar en la empresa que no aprueban el examen aplicado.

$$\frac{n}{N}=\frac{12}{280}=0.04$$

Observe que a pesar que la población bajo estudio es finita, no resulta necesario aplicar el factor de corrección ya que el tamaño de la muestra es solo el 4% del tamaño de la población, lo cual implica que el procedimiento para calcular el intervalo de confianza es:

$$n = 280 \quad X = 12 \qquad p=\frac{12}{280}=0.04$$

$$\alpha = 0.05 \qquad 1-\frac{0.05}{2}=0.975 \qquad Z_{0.975} = 1.96$$

$$0.04 - \sqrt{\frac{(0.04)(1-0.04)}{280}}(1.96)\ ,\ 0.04 + \sqrt{\frac{(0.04)(1-0.04)}{180}}(1.96)$$

$$[\ 0.04 - 0.02\ ,\ 0.04 + 0.02\ ]$$

$$[0.02\ ,\ 0.06\ ] \text{ o sea, } 0.02 \leq \pi \leq 0.06$$

Podemos asegurar con un nivel de confiabilidad del 95 % que el porcentaje de solicitantes que no aprueban las evaluaciones realizadas por la empresa se encuentra entre un 2% y un 6%.

7.6 Tamaño de muestra.

Uno de los aspectos de mayor importancia en cualquier proceso estadístico inferencial, consiste en determinar con la mayor exactitud posible el tamaño de la muestra requerido para proceder a la estimación de un valor poblacional.

Si en un estudio que involucre la estadística utilizamos un tamaño de muestra innecesariamente grande estaremos dilapidando recursos sin una razón objetiva, y por el contrario, si el número de elementos de la muestra es muy pequeño perderemos precisión en la estimación.

Debido a que en ningún método estadístico estudiaremos a la población completa, siempre al seleccionar en su lugar una muestra, perderemos *una parte* de información útil con respecto a la población, y en consecuencia, la estimación que hagamos con ésta última estará sujeta a un error.

Este error que se comete al estimar un valor poblacional haciendo uso de los datos de la muestra recibe el nombre de *error de muestreo* y por regla general se representa con la letra *E*.

7.6.1 Tamaño de muestra para estimar una media poblacional.

- **Población infinita**

Analicemos la expresión para calcular un intervalo de confianza.

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \right]$$

Si estudiamos con detenimiento la expresión anterior, podemos percatarnos de que el término dentro de ella que determina que ambos extremos del intervalo estén cercanos o alejados de μ, es decir, que tengamos un *alto nivel de precisión* en la estimación de μ o un *bajo nivel de precisión* en esta estimación, es el término:

$$\frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}$$

Esta expresión sería entonces el error de muestreo al que hicimos referencia en párrafos anteriores y por tanto:

$$E = \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}$$

Bastaría entonces despejar el valor de n en la expresión matemática anterior para obtener una fórmula que nos permita calcular el tamaño de la muestra. Hagamos este proceso:

$$E^2 = \frac{\sigma^2}{n} Z^2_{1-\frac{\alpha}{2}}$$

$$n = \left(\frac{\sigma Z_{1-\frac{\alpha}{2}}}{E} \right)^2$$

- **Población finita**

Siguiendo la misma línea de pensamiento utilizada en el epígrafe anterior:

$$E = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} Z_{1-\frac{\alpha}{2}}$$

$$E^2 = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right) Z^2_{1-\frac{\alpha}{2}}$$

$$n\,(N-1)E^2 = (N-n)\,\sigma^2\, Z^2_{1-\frac{\alpha}{2}}$$

$$n\,(N-1)E^2 + n\,\sigma^2\, Z^2_{1-\frac{\alpha}{2}} = N\sigma^2\, Z^2_{1-\frac{\alpha}{2}}$$

$$n=\frac{N\sigma^2 Z^2_{1-\frac{\alpha}{2}}}{\sigma^2 Z^2_{1-\frac{\alpha}{2}}+(N-1)E^2}$$

Para determinar el tamaño de la muestra en los dos casos anteriormente estudiados, se necesita establecer, el nivel de confianza deseado, el error de muestreo que el investigador está dispuesto a cometer y el nivel de variabilidad de la población bajo estudio. Por su importancia analicemos brevemente estos tres factores que intervienen en la determinación del tamaño de la muestra para estimar una media poblacional.

a) El nivel de confianza o nivel de confiablidad con el cual el investigador desea hacer una estimación, es un valor entre 0% y 100% elegido bajo su entera responsabilidad, es decir, el propio investigador decide por su cuenta y riesgo el nivel de confiabilidad de la estimación. Por regla general los niveles de confiabilidad más utilizados son el 95%, el 99% y el 99.9%, lo cual no quiere decir que otros valores no puedan ser utilizados. Debe tomarse en cuenta que a mayor nivel de confiabilidad de la estimación mayor será el tamaño de la muestra requerido, pero nunca deberá trabajarse con un nivel de confianza por debajo de lo que se desea, solo con el objetivo de reducir el tamaño de la muestra.

b) El error de muestreo es como su nombre lo indica *un error* por exceso o por defecto al estimar una media poblacional. Por ejemplo, si al estimar el ingreso promedio mensual poblacional de los ingenieros comerciales de una provincia, obtengo que estos ganan 1200 dólares cuando en realidad el ingreso es de 1340 dólares, hemos incurrido en un error de muestreo de 140 dólares. Por supuesto que hubiéramos incurrido también en un error si en realidad el ingreso promedio poblacional hubiera sido de 1060 dólares.

En general, si μ es una media poblacional y $\hat{\mu}$ una estimación de la misma, entonces el error de muestreo sería:

$$E=\left|\mu-\hat{\mu}\right|$$

La responsabilidad de determinar la magnitud numérica del error de muestreo recae en la persona que desarrolla el proceso de estimación, y se debe tomar en cuenta que un error de muestreo pequeño implica un tamaño de muestra grande, y a la inversa, un error de muestreo grande determina un tamaño de muestrea pequeño.

Sería una decisión totalmente errónea seleccionar un error de muestreo grande solo con el objetivo de reducir el tamaño de la muestra.

La determinación de la magnitud numérica del error de muestro es una decisión

de carácter técnico y debe ser asumida con mucha responsabilidad.

c) El tercer elemento que interviene en la determinación del tamaño de la muestra es la varianza poblacional, y un inconveniente que por regla general se presenta es que su valor numérico es desconocido.

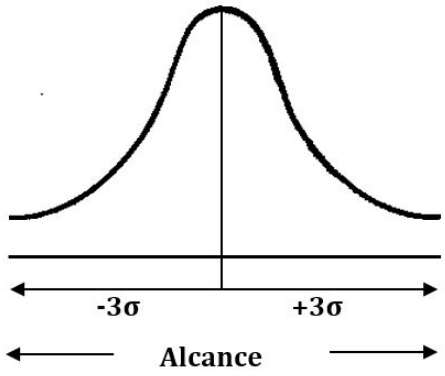
Cuando esto ocurre se hace necesario utilizar un estimador de su valor, y para ello, se suele proceder de una de las tres formas siguientes:

1.- Si tenemos conocimiento de un estudio realizado anteriormente con características similares al nuestro y disponemos de un valor confiable de la varianza obtenida en el mismo, podemos utilizar éste para obtener el tamaño de la muestra que estamos intentando calcular.

Este método no resulta del todo aplicable ya que es poco probable que dispongamos de un estudio anterior con características parecidas al nuestro, y en caso de que exista, nuestro estudio dejaría de tener un carácter original.

2.- En un capítulo anterior establecimos que en una distribución normal, más menos tres desviaciones estándar incluyen el 99.7% del área total bajo la curva, es decir, más tres desviaciones estándar y menos tres desviaciones estándar de la media incluyen a casi toda el área de la distribución, tal y como se muestra en la figura 7.3.

FIGURA 7.3 Alcance de una distribución normal



Valor máximo – Valor mínimo

Esto quiere decir, que si tenemos una idea más o menos clara del *alcance* de la población, es decir, la diferencia entre su valor máximo y su valor mínimo, entonces podemos emplearlo para obtener una estimación aproximada pero utilizable del valor de la desviación estándar de la población, y en consecuencia, del valor de la varianza.

Por tanto, una estimación aproximada de la desviación estándar de la población vendría dada por:

$$6\hat{\sigma} = Alcance \qquad \hat{\sigma} = \frac{Alcance}{6}$$

La estimación de la desviación estándar de la población mediante este método

no es del todo precisa, pero puede marcar la diferencia entre poder calcular un tamaño de muestra utilizable y no poder hacerlo.

3.- El método más comúnmente utilizado es el conocido con el nombre de *estudio piloto*, el cual consiste en obtener una pequeña muestra del indicador bajo estudio y con estos datos calcular el valor de la varianza, el cual nos permite determinar el tamaño de la muestra al ser utilizado en la fórmula correspondiente. Por supuesto que la muestra obtenida en el estudio piloto puede ser utilizada como parte de la muestra necesaria para desarrollar el estudio.

Veamos un ejemplo del cálculo del tamaño de una muestra. Un investigador del área de la salud desea estimar el valor promedio poblacional con el cual se vende un determinado medicamento en una determinada ciudad en la que existen 736 farmacias.

Para lograr este objetivo pretende utilizar un nivel de confiabilidad del 99% y está satisfecho si comete en la estimación un error de muestreo de ± 0.25 dólares. ¿Cuantas farmacias deberá visitar para conocer en cada una de ellas el precio de venta del medicamento?

La población es finita por tanto:}

$$n=\frac{N\sigma^2 Z^2_{1-\frac{\alpha}{2}}}{\sigma^2 Z^2_{1-\frac{\alpha}{2}}+(N-1)E^2}$$

N = 736 $\quad 1-\frac{0.01}{2}=0.995 \quad Z_{0.995} = 2.58 \quad Z^2_{0.995}=6.66 \quad E^2 = 0.06$

Debemos obtener un estimador de la varianza poblacional σ^2, y para ello, supongamos que desarrollamos un estudio piloto en 10 farmacias de la ciudad en el que obtuvimos los precios de venta del medicamento que se muestran en la tabla 7.4.

TABLA 7.4 Resultados del estudio piloto sobre precios del medicamento

7.36	8.23	7.14	7.95	7.12	8.14	8.05	7.23	8.12	7.84

$$\bar{x}=\frac{\sum x_i}{n}=\frac{77.18}{10}=7.72 \qquad s^2=\frac{\sum\left(x_i-\bar{x}\right)^2}{n-1}=\frac{1.84}{9}=0.20$$

Sutituyendo valores:

$$n=\frac{736(0.2)(6.66)}{0.2(6.66)+(736-1)(0.06)}=\frac{980.35}{44.23}=21.16$$

7.6.2 Tamaño de muestra para estimar una proporción poblacional.

- **Población infinita**

El error de muestreo es $E=\sqrt{\frac{p(1-p)}{n}}Z_{1-\frac{\alpha}{2}}$ de donde:

$$E^2=\frac{p(1-p)}{n}Z^{2}_{1-\frac{\alpha}{2}}$$

$$n=\frac{p(1-p)Z^{2}_{1-\frac{\alpha}{2}}}{E^2}$$

- **Población finita**

$$E=\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}Z_{1-\frac{\alpha}{2}}$$

$$E^2=\left(\frac{p(1-p)}{n}\right)\left(\frac{N-n}{N-1}\right)Z^{2}_{1-\frac{\alpha}{2}}$$

$$n(N-1)E^2=Z^{2}_{1-\frac{\alpha}{2}}\,p(1-p)N-Z^{2}_{1-\frac{\alpha}{2}}\,p(1-p)n$$

$$nZ^{2}_{1-\frac{\alpha}{2}}\,p(1-p)+n(N-1)E^2=Z^{2}_{1-\frac{\alpha}{2}}\,p(1-p)N$$

$$n=\frac{Z^{2}_{1-\frac{\alpha}{2}}\,p(1-p)N}{Z^{2}_{1-\frac{\alpha}{2}}\,p(1-p)+(N-1)E^2}$$

Ya sea la población finta o infinita, la determinación del tamaño de muestra para estimar una proporción poblacional requiere que sean establecidos por parte del investigador el nivel de confiabilidad, el error de muestreo y el valor de p. De los dos primeros factores hablamos en el numeral anterior, por tanto, concentremos nuestra atención en la magnitud numérica de p.

Para mayor claridad precisemos el término p mediante un ejemplo. Supongamos que el autor de este libro desea desarrollar una investigación con el objetivo de conocer el grado de aceptación del mismo por parte de los estudiantes de la Universidad Laica Eloy Alfaro de Manabí, y en especial, los relacionados con carreras de las esferas económicas y administrativas, los cuales suman un total de 2700 estudiantes. Si el

autor decide utilizar un nivel de confiabilidad del 99.9% y un error de muestreo de ± 0.04, es decir, ± %4 en términos de porcentaje, ¿Cuantos estudiantes deberán ser entrevistados para conocer el grado de aceptación que tiene el libro?

La población es finita, por tanto, el tamaño de la muestra viene dado por:

$$n = \frac{Z^2_{1-\frac{\alpha}{2}}\, p(1-p)N}{Z^2_{1-\frac{\alpha}{2}}\, p(1-p) + (N-1)E^2}$$

El error de muestreo se toma siempre en términos de proporción, nunca expresado como un porcentaje, es decir, E = 0.04 y por tanto, $E^2 = 0.002$.

$$N = 2700 \quad 1 - \frac{0.001}{2} = 0.9995 \quad Z_{0.9995} = 3.27 \quad Z^2_{0.9995} = 10.69$$

y sustituyendo valores, el tamaño de la muestra sin precisar aún el valor de p sería:

$$n = \frac{28863(p)(1-p)}{10.69(p)(1-p) + 5.4}$$

Si creemos conocer con una buena aproximación a la realidad cual será en la universidad el nivel de aceptación que tenga el libro, entonces podemos utilizar este valor como el mejor estimador del valor de p para sustituirlo en la ecuación anterior.

Lamentablemente, por regla general el investigador no tiene ni siquiera la menor idea del valor de p y se hace entonces necesario escogerlo de forma tal que el tamaño de la muestra resulte lo *suficientemente grande* como para garantizarnos la precisión que necesitamos sin preocuparnos del verdadero valor de p. La tabla 7.5 muestra un ensayo con diferentes valores de p y su incidencia en el tamaño de la muestra.

TABLA 7.5 Valores de p y su incidencia en el tamaño de la muestra

P	**n**
0.1	408
0.2	649
0.3	793
0.4	870
0.5	894 ←
0.6	870
0.7	793
0.8	649
0.9	408

El valor de p con el cual se obtiene un tamaño de muestra *suficientemente grande* para garantizar la precisión de la estimación sin importarnos su verdadero valor es *0.5*.

En conclusión, siempre que necesitemos calcular el tamaño de muestra necesario para estimar una proporción poblacional y no dispongamos de una buena estimación de p, lo recomendable es asignarle el valor 0.5.

No obstante, la sugerencia es tratar de estimar el valor de p por todos los medios que tengamos a nuestro alcance, ya que en muchas ocasiones no resulta una tarea tan crítica como parece, y si lo logramos, obtendremos el beneficio de trabajar con una muestra menor.

Observe que si estamos seguros que la proporción de estudiantes que harán una buena aceptación del libro es de 0.8, entonces el número de estudiantes a entrevistarse serian 649 y no 894 que son los necesarios si tomamos p igual a 0.5.

Ejercicios del capítulo

7.1 Una empresa produce clavos de dos pulgadas y los envasa en cajas que contienen 50 clavos cada una. Control de calidad detecta que al parecer la máquina que realiza el trabajo de envasado tiene un desperfecto, y para comprobarlo, extrajo una muestra de 36 cajas las cuales contenían el número de clavos que se muestran a continuación:

50	51	50	50	51	52	51	48	49
50	50	48	47	53	50	50	51	50
49	48	50	50	50	49	51	52	48
50	50	50	49	49	51	50	50	49

Con un nivel de significación del 5%, obtenga un intervalo de confianza que permita estimar el número de clavos poblacional promedio que la máquina está ubicando en cada caja.

a. ¿Considera necesario utilizar el factor de corrección para población finita?
b. Suponga una varianza poblacional igual a 1.55
c. Considere que la varianza poblacional es desconocida

7.2 Los datos que se muestran a continuación representan la demanda diaria de un producto registrado por una empresa:

34	42	39	36	44	40	39	41	44
36	45	41	38	35	37	43	40	42

Con un nivel de significación del 0.1%, obtenga un intervalo de confianza para estimar la demanda diaria poblacional del producto de referencia.

a. ¿Considera necesario utilizar el factor de corrección para población finita?
b. Suponga una varianza poblacional igual a 11.5
c. Considere que la varianza poblacional es desconocida

7.3 Una empresa agropecuaria desea estimar la producción diaria de leche de vacas que están siendo suplementadas con un balanceado de alto nivel proteico. Para ello extrajo una muestra de tamaño 24 de un lote de 400 vacas, la cual dio los resultados de producción de leche que se observan a continuación:

12.5	13.6	12.8	12.9	13.4	13.8	12.5	12.7	13.1	13.5	11.9	12.9
13.6	13.8	12.7	12.5	12.1	13.8	13.5	12.4	13.1	13.6	12.2	12.4

Con un nivel de significación del 1%, obtenga un intervalo de confianza para estimar la producción diaria de leche poblacional de vacas suplementadas con el balanceado.

a. ¿Considera necesario utilizar el factor de corrección para población finita?

b. Suponga una varianza poblacional igual a 11.5
c. Considere que la varianza poblacional es desconocida

7.4 Se seleccionó una muestra de 30 contribuyentes de un total de 600 con el objetivo de estimar el impuesto a la renta pagado por ellos en el año 2013. La media de la muestra dio como resultado 70 dólares. Con un nivel de significación del 5%, obtenga un intervalo de confianza para estimar la media poblacional del impuesto pagado en el año 2013.

a. ¿Considera necesario utilizar el factor de corrección para población finita?
a. Suponga una varianza poblacional igual a 6.2
a. Considere que la varianza poblacional es desconocida

7.5 El Servicio de Rentas Internas de la provincia de Manabí clausuró 50 restaurantes de 250 que fueron inspeccionados por no emitir factura con el IVA correspondiente. Con un nivel de significación del 5%, obtenga un intervalo de confianza para estimar la proporción poblacional de restaurantes que no emiten factura en la provincia de Manabí. Determine si es requerido ajustar el error estándar a través del factor de corrección para población finita.

7.6 Una muestra al azar de 25 pasajeros que arribaron a Quito provenientes de España en un Boeing 737 – 900 con 180 ocupantes, dio como resultado que 6 de ellos eran de la tercera edad. Con un nivel de significación del 1%, obtenga un intervalo de confianza para estimar la proporción poblacional de pasajeros de la tercera edad que arriban a Quito provenientes de España.

7.7 Un investigador del área de la salud desea estimar el nivel de colesterol en un determinado sector en el que viven 700 personas. Para su investigación necesita determinar el tamaño de la muestra, y en tal sentido, decide utilizar un nivel de confiabilidad del 99% y está dispuesto a aceptar un error de muestreo de $\pm 5mg./dL$..

Como no conoce el valor de la varianza poblacional del colesterol en el sector, desarrolló un estudio piloto en el cual obtuvo los siguientes niveles de colesterol:

205	195	210	225	215	194	200	210	224	205	200	220

Determine el tamaño de muestra requerido para desarrollar la investigación con las premisas planteadas.

7.8 El departamento de control de la calidad de una empresa desea estimar la proporción poblacional de pilas de níquel – cadmio que salen con defecto durante el proceso de producción, utilizando para ello un nivel de confiabilidad del 95%. El departamento no conoce con una buena aproximación a la realidad cuál es la proporción

de pilas defectuosas y considera que un error de muestreo de un 4% sería razonable. Determine el tamaño de la muestra requerido para desarrollar la investigación con los antecedentes planteados.

Capítulo 8

Prueba de Hipótesis para una sola muestra

El problema

El peso neto de una lata de lomito de atún enlatado en aceite de una determinada marca debe ser igual a 160 gramos. Una muestra aleatoria del peso neto de un grupo de estas latas dio como resultado un valor promedio igual a 158 gramos.

Con una confiabilidad lo suficientemente aceptable, ¿sugieren los datos de la muestra una diferencia entre el peso neto reglamentario de la lata de atún y el valor promedio obtenido en la muestra?

8.1 Introducción.

El procedimiento estadístico que da una adecuada respuesta a la problemática planteada en el párrafo anterior se conoce como *PRUEBA DE HIPÓTESIS*, la cual se inicia con una *suposición* que se establece con respecto al valor de un parámetro poblacional. Esta suposición recibe el nombre de hipótesis estadística. En este capítulo estudiaremos el procedimiento que nos permitirá comprobar la validez o no de un enunciado establecido con relación a un parámetro poblacional.

Algunos ejemplos de enunciados cuya validez podría ser sometida a prueba son los siguientes:

- En el Ecuador los conductores de clase media tienen un consumo anual de gasolina súper de 350 galones.
- En la provincia de Manabí solo el 30% de sus habitantes tiene una cuenta de ahorro.
- El 36% de hogares ecuatorianos están asentados en casa propia.
- El consumo de alcohol por habitante en el Ecuador alcanza los 9,38 litros por año.

Por regla general, los problemas que involucran una toma de decisión pueden reducirse a un procedimiento que implique el rechazo o la aceptación de una hipótesis o suposición sobre el valor de la media poblacional de una distribución.

Antes de continuar dejemos establecida las dos siguientes definiciones:

HIPÓTESIS: Suposición que se establece con relación a un parámetro poblacional factible de ser verificada.

PRUEBA DE HIPÓTESIS: Procedimiento estadístico que basado en los resultados de una muestra permite con un determinado nivel de confiabilidad establecer si la hi-

pótesis es o no razonable.

A continuación pasaremos a ilustrar de la forma más sencilla posible el procedimiento para someter a prueba una hipótesis estadística, y para ello, consideremos tomar una decisión sobre el valor de una media poblacional de la cual conocemos que solo puede tomar los valores 14 o 19.

Resulta razonable entonces formular el problema mediante las siguientes hipótesis:

$H_{0:}\ \mu = 14$ y $H_1: \mu = 19$

En términos de una Prueba de Hipótesis, a H_0 (H subcero) se le conoce como la *hipótesis nula* y a H_1 (H subuno) como la *hipótesis alternativa*. La hipótesis nula es la suposición que deseamos probar mientras que la hipótesis alternativa es la afirmación que debemos aceptar cuando la hipótesis nula es rechazada.

El término "hipótesis nula" surgió cuando en las primeras aplicaciones estadísticas a las ciencias agrícolas y médicas la hipótesis que se probaba era que un nuevo fertilizante o un nuevo medicamento *"no tenía efecto".*

Decidir si la media poblacional es igual a 14 o igual a 19, se reduce a tomar una de las dos siguientes decisiones:

- Rechazar H_0, es decir, la hipótesis nula es falsa y por tanto se concluye que $\mu = 19$
- No se rechaza H_0, lo cual determina que $\mu = 14$

Pero cualquiera sea la decisión que tomemos, en el procedimiento podemos cometer dos tipos de errores:

- Si rechazamos H_0 y realmente esta hipótesis es verdadera, cometemos el error conocido como *Error de Tipo I*
- Si no rechazamos H_0 y en realidad esta hipótesis es falsa, cometemos el error conocido como *Error de Tipo II*

La tabla 8.1 muestra el conjunto de decisiones posibles a tomar en una prueba de hipótesis y sus posibles resultados.

TABLA 8.1 Conjunto posible de decisiones en una prueba de hipótesis

Prueba de hipótesis	No se rechaza H_0	Se rechaza H_0
H_0 es verdadera	Decisión correcta	Error de Tipo I
H_0 es falsa	Error de tipo II	Decisión correcta

La probabilidad de cometer el Error de Tipo I se denomina *nivel de significación* y se le representa mediante la letra α. La probabilidad de cometer el Error de Tipo II se designa con la letra β.

En el procedimiento estadístico y con el objetivo de optimizar el método, las hipótesis se formulan de tal manera que el Error de Tipo I sea el de consecuencias más graves, fijando el nivel de significación con un valor lo suficientemente pequeño y que sea aceptable para el investigador. Por regla general este valor de α oscila entre 0,001 y 0,05, o lo que es lo mismo, entre un 0,1% y 5%. Por supuesto que valores mayores o menores a los indicados pueden ser utilizados.

Un principio del cual debemos estar alertas es que no resulta adecuado rechazar o no una hipótesis relacionada con un parámetro de población, siguiendo un instinto personal o una simple apreciación, pues lo correcto sería basarnos en los datos de una muestra para decidir de forma objetiva si rechazamos o no la hipótesis planteada.

Según lo establecido en el párrafo anterior, *una regla de decisión razonable* sería no rechazar H_0 si la media de una muestra extraída de la población bajo estudio es menor o igual que un cierto *valor crítico z* comprendido entre 14 y 19, y rechazar H_0 si esta media es mayor que dicho valor. Supongamos entonces que la media de una muestra de tamaño 16 extraída de la población bajo estudio dio como resultado una media igual a 17, y que conocemos que la desviación estándar poblacional tiene un valor igual a 9. Si fijamos el valor de α, entonces la probabilidad de rechazar H_0 siendo verdadera, viene dada por la expresión:

$P\{\bar{x} > z \mid \mu = 14\} = \alpha$

Si consideramos $\alpha = 0.05$, entonces, $P\{\bar{x} > z \mid \mu = 14\} = 0.05$

$P\{\bar{x} > z \mid \mu = 14\} = 1 - P\{\bar{x} \leq z \mid \mu = 14\} = 0.05$

Tipificando: $1 - P\left\{\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \dfrac{z - 14}{9/4}\right\} = 0.05$

o lo que es lo mismo:

$$P\left\{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{z - 14}{9/4}\right\} = 1 - 0.05 = 0.95$$

$$N\left(\frac{z - 14}{\frac{9}{4}}\right) = 0.95, \text{ de donde } \frac{z - 14}{2.25} = Z_{0.95} = 1.64 \text{ y entonces } z = 17.69$$

La regla de decisión queda expresada como sigue:

No rechazar H_0 si $\bar{x} \leq 17.69$

Rechazar H_0 si $\bar{x} > 17.69$

En nuestro caso $\bar{x}$ = 17 por tanto no debemos rechazar $H_{0.}$
Calculemos a continuación los valores de β para diferentes valores de α.

Para α = 0.05 ya vimos que z = 17.69

$$\beta = P\{\bar{x} \leq 17.69 \mid \mu = 19\} = N\left(\frac{17.69 - 19}{\frac{9}{4}}\right) = N(-0.58)$$

N(-0.58) = 1- N(0.58) = 1- 0.7190 = 0.2810

Para α = 0.01

$$N\left(\frac{z - 14}{\frac{9}{4}}\right) = 0.99, \text{ de donde } \frac{z-14}{2.25} = Z_{0.99} = 2.33 \text{ y entonces z = 19.24}$$

$$\beta = P\{\bar{x} \leq 19.24 \mid \mu = 19\} = N\left(\frac{19.24 - 19}{\frac{9}{4}}\right) = N(0.11) = 0.5438$$

Para α = 0.001

$$N\left(\frac{z - 14}{\frac{9}{4}}\right) = 0.999, \text{ de donde } \frac{z-14}{2.25} = Z_{0.999} = 3.08 \text{ y entonces z = 20.93}$$

$$\beta = P\{\bar{x} \leq 20.93 \mid \mu = 19\} = N\left(\frac{20.93 - 19}{\frac{9}{4}}\right) = N(0.86) = 0.8051$$

En la tabla 8.2 aparece un resumen de los cálculos realizados.

TABLA 8.2 Valores de α y el correspondiente valor de β

α	β
0.05	0.2810
0.01	0.5438
0.001	0.8051

Observe como al disminuir la probabilidad de cometer el Error de Tipo I se incrementó la probabilidad de cometer el Error de Tipo II y viceversa.

8.2 Potencia de la prueba de hipótesis.

En párrafos anteriores señalamos que en una prueba de hipótesis existe la probabilidad de cometer el Error de Tipo II cuando no rechazamos la hipótesis nula siendo ésta falsa, y a esta probabilidad la identificamos con la letra β. Partiendo de esto la expresión 1- β representa entonces la probabilidad de rechazar la hipótesis nula cuando ésta es falsa y resultaría deseable entonces que en una prueba de hipótesis cualquiera éste valor fuese lo más cercano a 1 posible.

Al valor 1-β se le conoce como la *potencia de la prueba de hipótesis*. En la tabla 8.3 aparece un resumen de los resultados obtenidos.

TABLA 8.3 Resumen de los resultados

n	α	β	1 - β
16	0.05	0.2810	0.7190
16	0.01	0.5438	0.4562
16	0.001	0.8051	0.1949

En la tabla anterior se puede apreciar que para un tamaño de muestra fijo existe una relación inversa entre el nivel de significación utilizado y la potencia de la prueba de hipótesis, es decir, cuando se disminuye la probabilidad de rechazar H_0 siendo ésta verdadera (α) entonces se disminuye también la potencia de la prueba (1-β), lo cual resulta no deseable.

Veamos a continuación que aumentando el tamaño de la muestra podemos incrementar la potencia de la prueba de hipótesis con independencia del valor de α que haya sido escogido.

Calculemos la potencia de la prueba si incrementamos el tamaño de la muestra a 25.

Para α = 0.05 y n = 25

$$N\left(\frac{z-14}{\frac{9}{5}}\right) = 0.95, \text{ de donde } \frac{z-14}{1.8} = Z_{0.95} = 1.64 \text{ y entonces z = 16.95}$$

$$\beta = P\{\bar{x} \leq 16.95 \mid \mu = 19\} = N\left(\frac{16.95-19}{\frac{9}{5}}\right)$$

N(-1.14) = 1 – N (1.14) = 1 – 0.8729 = 0.1271 , y por tanto:

1 - β = 1 – 0.1271 = 0.8729

valor que al compararlo con la potencia de la prueba de hipótesis (0.7190) para n = 16 y α = 0.05, nos indica que *al aumentar el tamaño de la muestra podemos incrementar el valor de la potencia de la prueba de hipótesis*, lo cual representa un resultado de vital importancia dentro de la Estadística. El lector podrá comprobar sin mucha dificultad los resultados que se reportan en la tabla 8.4 y compararlos con los que aparecen al inicio de este numeral.

TABLA 8.4 Valores de α, β y 1 – β para n = 25

n	α	β	**1 -** β
25	0.05	0.1271	0.8729
25	0.01	0.3264	0.6736
25	0.001	0.3821	0.6179

8.3 Prueba de hipótesis para la media de una población con varianza poblacional conocida.

En toda prueba de hipótesis se hace necesario asignarle un valor hipotético al parámetro poblacional sobre el cual se desarrolla la investigación. Una prueba de hipótesis para la media de una población se hace necesaria cuando ocurre un evento que de alguna manera nos hace suponer un cambio en la media poblacional. El valor de la media poblacional antes de que tal evento ocurra, es el valor hipotético al que nos referíamos con anterioridad, y se le suele designar como μ_0.

Para una mejor comprensión del aspecto que estamos tratando, analicemos un ejemplo similar al que se formuló al iniciar este capítulo.

El peso neto de una lata de lomito de atún enlatado en aceite de una determinada marca debe ser igual a 160 gramos. Una muestra aleatoria del peso neto de un grupo de 36 de estas latas dio como resultado un valor promedio de 159.33 gramos.

Adicionalmente se conoce que el valor de la varianza poblacional del indicador

es igual a 4.5 gramos. Con un nivel de significación del 5%, ¿sugieren los resultados de la muestra que el peso neto de las latas de atún producidas por la empresa es menor a 160 gramos?

Observe que antes de que se produjera el evento relacionado con la determinación de la muestra, existía un valor establecido del peso neto de la lata de atún (160 gramos).

Por tanto, y sin lugar a dudas, podemos concluir que el valor hipotético de la media de la población es de 160 gramos, y en consecuencia:

$\mu_0 = 160$

La formulación de la hipótesis nula sería entonces:

$H_0: \mu = 160$

Tres hipótesis alternativas podrían entonces ser formuladas en correspondencia con el interés de la investigación. Estas diferentes hipótesis alternativas son:

- $H_1: \mu \neq 160$ si nos interesa conocer si el peso neto de la lata de atún ha cambiado.
- $H_1: \mu > 160$ si esperamos que el peso neto de la lata de atún se ha incrementado.
- $H_1: \mu < 160$ si suponemos que el peso neto de la lata de atún ha disminuido.
- Esta última hipótesis alternativa es la que se ajusta al ejemplo que estamos procesando.
- Un resumen de los datos de la prueba de hipótesis se muestra a continuación:

$\mu_0 = 160$ $\quad n = 36$ $\quad \bar{x} = 159.33$

$\sigma^2 = 4.5 \Rightarrow \sigma = 2.12$ $\quad \alpha = 0.05$

Con el objetivo de establecer las regiones de rechazo y aceptación para cada una de las tres situaciones planteadas, desarrollaremos de forma teórica las tres pruebas de hipótesis para con posterioridad referirnos específicamente al caso que nos ocupa:

Consideremos en primer lugar las hipótesis:

- $H_{0:}\ \mu = \mu_0$ $\quad H_{1:}\ \mu \neq \mu_0$

donde μ es la media de la distribución de una población normal con varianza σ^2 y μ_0 es un número real.

Si $\bar{x}$ es la media de una muestra aleatoria de esta población, resulta razonable rechazar H_0 si $\bar{x}$ difiere "lo suficiente" de μ, o sea, si

$|\bar{x} - \mu| > x$ para x suficientemente grande. Tipificando:

$\left|\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right| > \frac{x}{\frac{\sigma}{\sqrt{n}}}$ = z = valor crítico

Establecer la regla de decisión se reduce entonces a determinar el valor de z.

Partiendo de la relación:

P {rechazar H_0 | H_0 es cierta} = α

$\left[\left|\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right|\right] > z = \alpha$ ya que si H_0 es cierta entonces $\mu = \mu_0$

$P\left[\left|\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right| \le z\right] = 1-\alpha$ $P\left[-z \le \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \le z\right] = 1-\alpha$

$N(z) - N(-z) = 1 - \alpha$ $N(z) - 1 + N(z) = 1 - \alpha$

$2N(z) = 2 - \alpha$ $N(z) = 1 - \frac{\alpha}{2}$

y entonces z sería igual a $Z_{1-\frac{\alpha}{2}}$ que es el percentil de orden $1-\frac{\alpha}{2}$ de la distribución normal con media 0 y varianza 1.

La regla de decisión quedaría entonces:

Rechazar H_0 si $\left|\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right| > Z_{1-\frac{\alpha}{2}}$

No rechazar H_0 si $\left|\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right| \le Z_{1-\frac{\alpha}{2}}$

Apliquemos la regla de decisión en el caso que estamos estudiando.

- $H_{0:}\ \mu = 160$ $H_{1:}\ \mu \neq 160$

$$\left|\frac{159.33-160}{\frac{2.12}{\sqrt{36}}}\right|=\left|\frac{-0.67}{0.35}\right|=1.91$$

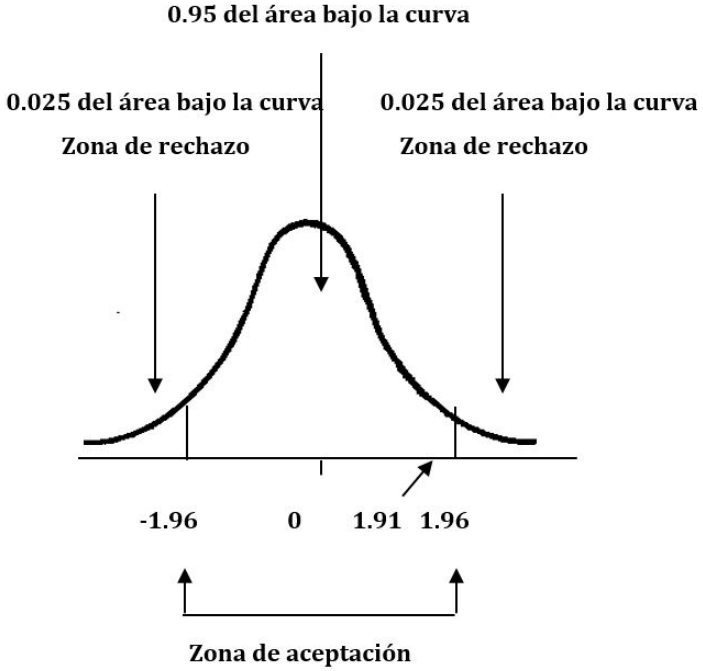
$$1-\frac{\alpha}{2}=1-\frac{0.05}{2}=1-0.025=0.975$$

$Z_{0.975}$ = 1.96 = valor crítico

y como 1.91 < 1.96, no rechazamos H_0 para un nivel de significación del 5%, es decir, la muestra obtenida no nos permite asegurar que el peso neto de las latas de atún ha cambiado.

Las zonas de rechazo y de aceptación de esta prueba de hipótesis se muestran en la figura 8.1. Aprecie en el gráfico que la zona de rechazo está conformada por *dos colas* que se extienden a la derecha e izquierda de la curva de la distribución normal. Se dice que la prueba de hipótesis que hemos desarrollado es una *prueba de dos colas*.

FIGURA 8.1 Zona de aceptación y de rechazo de la prueba de hipótesis



Consideremos las hipótesis:

- $H_{0:}\ \mu = \mu_0$ $\quad$ $H_{1:}\ \mu > \mu_0$

Si $\bar{x}$ es la media de una muestra aleatoria de esta población, resulta razonable rechazar H_0 si $\bar{x}$ es "lo suficientemente" mayor que μ, o sea, si $\bar{x} - \mu > x$ para x suficientemente grande. Tipificando:

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} > \frac{x}{\frac{\sigma}{\sqrt{n}}} = z = \text{valor crítico}$$

$$P\left[\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right] > z = \alpha \quad \text{ya que si } H_0 \text{ es cierta entonces } \mu = \mu_0$$

$$P\left[\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right] \leq z = 1-\alpha \qquad N(z) = 1 - \alpha$$

y entonces z = $Z_{1-\alpha}$ que es el percentil de orden $1-\alpha$ de la distribución normal con media 0 y varianza 1.

La regla de decisión quedaría entonces como sigue:

Rechazar H_0 si $\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > Z_{1-\alpha}$

No rechazar H_0 si $\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \leq Z_{1-\alpha}$

En nuestro caso:

- $H_{0:}\ \mu = 160$ $H_{1:}\ \mu > 160$

$$\frac{159.33-160}{\frac{2.12}{\sqrt{36}}} = \frac{-0.67}{0.35} = -1.91$$

$$1-\alpha = 1-0.05 = 0.95$$

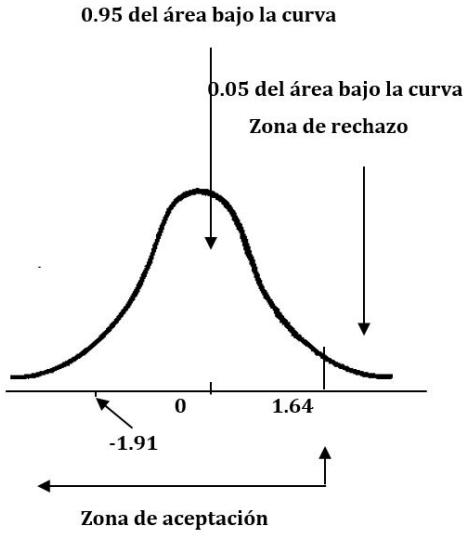
$Z_{0.95} = 1.64$ = valor crítico

y como -1.91 < 1.64, no rechazamos H_0 con un nivel de significación del 5%, es decir, la muestra obtenida no nos permite asegurar que el peso neto de las latas de atún es mayor a lo establecido.

Las zonas de rechazo y de aceptación de esta prueba de hipótesis se muestran

en la figura 8.2.

FIGURA 8.2 Zona de aceptación y de rechazo de la prueba de hipótesis



La prueba de hipótesis desarrollada es una *prueba de una cola.*

Sean las hipótesis:

- $H_{0:}\ \mu = \mu_0 \qquad H_{1:}\ \mu < \mu_0$

donde μ el la media de la distribución de una población normal con varianza σ^2 y μ_0 es un número real.

Si $\bar{x}$ es la media de una muestra aleatoria de esta población, resulta razonable rechazar H_0 si $\bar{x}$ es lo suficientemente menor que μ, o sea, si $\bar{x} - \mu < -x$ para x suficientemente grande, de donde:

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < \frac{-x}{\frac{\sigma}{\sqrt{n}}} = -z = \text{valor crítico}$$

$$\left[\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right] < -z = \alpha \quad \text{ya que si } H_0 \text{ es cierta entonces } \mu = \mu_0$$

$$1 - P\left[\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right] < z = \alpha \qquad P\left[\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right] < z = 1-\alpha$$

N (z) = 1 - α de donde z = $Z_{1-\alpha}$.

La regla de decisión quedaría entonces:

Rechazar H_0 si $\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} < -Z_{1-\alpha}$

No rechazar H_0 si $\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \geq -Z_{1-\alpha}$

Particularizando:

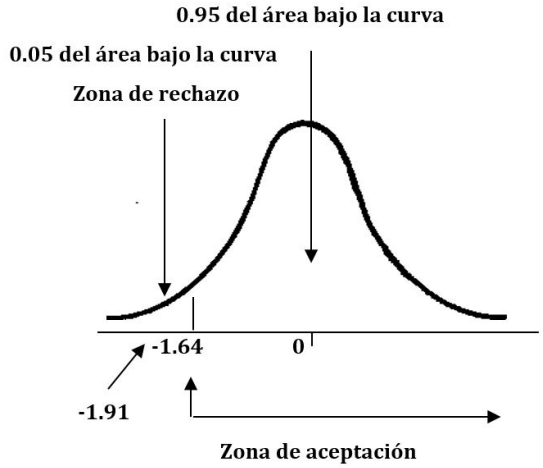
- $H_{0:}\ \mu = 160$ $H_{1:}\ \mu < 160$

$$\frac{159.33-160}{\frac{2.12}{\sqrt{36}}} = \frac{-0.67}{0.35} = -1.91$$

$1-\alpha = 1-0.05 = 0.95$ $-Z_{0.95} = -1.64$ = valor crítico

y como -1.91 < -1.64, rechazamos H_0 con un nivel de significación del 5%, es decir, la muestra obtenida nos permite concluir que el peso neto de las latas de atún es menor a lo establecido. Las zonas de rechazo y de aceptación de esta prueba de hipótesis se muestran en la figura 8.3.

FIGURA 8.3 Zona de aceptación y de rechazo de la prueba de hipótesis



La prueba de hipótesis realizada es una *prueba de una cola*.

8.4 Prueba de hipótesis para la media de una población con varianza poblacional desconocida.

En el Capítulo 7 al estudiar el tema relacionado con Intervalos de Confianza precisamos que diferenciar entre muestras grandes o pequeñas es solo importante cuando la varianza poblacional es desconocida, y por tanto, es necesario estimarla a través de la varianza de la muestra.

En estos casos el estadígrafo a utilizar es el siguiente:

$T = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}}$ el cual sigue una distribución t de Student con n-1 grados de libertad.

En consecuencia, las reglas de decisión para las tres pruebas de hipótesis desarrolladas en el numeral anterior, en el caso que la varianza poblacional sea desconocida, quedarían como sigue:

- $H_{0:}\ \mu = \mu_0 \qquad H_{1:}\ \mu \neq \mu_0$

Rechazar H_0 si $\left| \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}} \right| > t_{\alpha}^{(n-1)}$

No rechazar H_0 si $\left| \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}} \right| \leq t_{\alpha}^{(n-1)}$

Esta prueba es de dos colas.

- $H_{0:}\ \mu = \mu_0 \qquad H_{1:}\ \mu > \mu_0$

Rechazar H_0 si $\dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}} > t_{\alpha}^{(n-1)}$

No rechazar H_0 si $\dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}} \leq t_{\alpha}^{(n-1)}$

Esta prueba es de una cola.

- $H_{0:}\ \mu = \mu_0 \qquad H_{1:}\ \mu < \mu_0$

Rechazar H_0 si $\dfrac{\bar{x}-\mu_0}{\dfrac{s}{\sqrt{n}}} < -t_{\alpha}^{(n-1)}$

No rechazar H_0 si $\dfrac{\bar{x}-\mu_0}{\dfrac{s}{\sqrt{n}}} \geq -t_{\alpha}^{(n-1)}$

Esta prueba es de una cola.

Veamos un ejemplo. Datos históricos indican que el consumo mensual de energía eléctrica de familias de clase media en la ciudad de Manta es de 254 kilowatts. La Corporación Nacional de Electricidad de la provincia a la cual pertenece la ciudad desarrolló una investigación con el objetivo de establecer el consumo de energía eléctrica en el periodo navideño, y para ello, extrajo una muestra de consumo en 30 hogares de clase media en la cual obtuvo los resultados que se muestran en la tabla 8.5

TABLA 8.5 Consumo de energía eléctrica en periodo navideño

261	258	253	252	257	255	261	260	254	255
252	257	258	253	254	254	261	258	260	253
255	254	261	254	253	257	263	261	255	260

Con un nivel de significación del 1% procedamos a desarrollar las tres pruebas de hipótesis antes señaladas:

En primer lugar, calculemos y resumamos los datos necesarios:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{7699}{30} = 256.63 \qquad s^2 = \frac{\sum \left(x_i - \bar{x} \right)^2}{n-1} = \frac{317}{29} = 10.93$$

$$s = \sqrt{10.93} = 3.31 \qquad \mu_0 = 254$$

entonces:

- $H_{0:}\ \mu = 254$ $\qquad$ $H_{1:}\ \mu \neq 254$

$$\left|\frac{256.63-254}{\frac{3.31}{\sqrt{30}}}\right| = \frac{2.63}{0.60} = 4.38$$

En la **TABLA T.2** del Anexo A podemos encontrar el valor del percentil de la t de Student para un nivel de significación del 1%, 29 grados de libertad y una prueba de dos colas. Este valor es 2.756. Al ser 4.38 > 2.756, rechazamos la hipótesis nula, es decir, en periodo navideño el consumo de energía eléctrica de familias de clase media en la ciudad de Manta es distinto del consumo histórico.

- $H_{0:}\ \mu = 254$ $H_{1:}\ \mu > 254$

$$\left|\frac{256.63-254}{\frac{3.31}{\sqrt{30}}}\right| = \frac{2.63}{0.60} = 4.38$$

El percentil de la t de Student para un nivel de significación del 1%, 29 grados de libertad y una prueba de una cola, tiene un valor igual a 2.462.

Por ser 4.38 > 2.462, rechazamos la hipótesis nula, es decir, en periodo navideño el consumo de energía eléctrica de familias de clase media en Manta es mayor al consumo histórico.

- $H_{0:}\ \mu = 254$ $H_{1:}\ \mu < 254$

$$\left|\frac{256.63-254}{\frac{3.31}{\sqrt{30}}}\right| = \frac{2.63}{0.60} = 4.38$$

Como 4.38 > -2.462, no rechazamos la hipótesis nula, es decir, los datos de la muestra no nos permite concluir que en periodo navideño el consumo de energía eléctrica de familias de clase media en la ciudad de Manta es menor al consumo histórico.

Cabe aclarar que de los tres casos estudiados, solo el segundo se justifica desde el punto de vista técnico, pues sin lugar a dudas, el interés de la Corporación Nacional de Electricidad fue investigar si en periodo navideño el consumo de energía eléctrica de familias de clase media de la ciudad de Manta era *mayor* al valor histórico.

8.5 Proceso de cinco pasos para una prueba de hipótesis.

Lo que hemos estudiado hasta el momento nos permite formular un procedimiento conformado por cinco pasos para desarrollar de forma eficiente una prueba de hipótesis.

Paso 1	Se formulan las hipótesis nula y alternativa
Paso 2	Se establece el nivel de significación
Paso 3	Se identifica la distribución a utilizar
Paso 4	En dependencia de las hipótesis se escoge la regla de decisión adecuada
Paso 5	En base a la muestra tomada se decide o no rechazar la hipótesis nula

Paso 1: Se formulan las hipótesis nula y alternativa

- La hipótesis nula es siempre la que deseamos probar, y por regla general, se establece partiendo de un valor hipotético o preestablecido de la media poblacional.
- Si los datos de la muestra extraída en el proceso no corroboran lo establecido en la hipótesis nula, entonces debemos aceptar lo establecido en la hipótesis alternativa.
- Si μ_0 es el valor hipotético o preestablecido de la media poblacional y $H_{0:}\ \mu = \mu_0$ es la hipótesis nula, entonces las tres hipótesis alternativas posibles serian:

1. $H_{1:}\ \mu \neq \mu_0$
2. $H_{1:}\ \mu > \mu_0$
3. $H_{1:}\ \mu < \mu_0$

Paso 2: Se establece el nivel de significación

- Por ser el nivel de significación la probabilidad de rechazar una hipótesis nula cuando ésta es verdadera, resulta razonable elegir éste valor lo más pequeño posible en correspondencia con la importancia del tema tratado en la prueba.
- Pero debemos tomar en cuenta que reducir el valor del nivel de significación incrementa la probabilidad de no rechazar una hipótesis nula cuando ésta es falsa.
- Debemos entonces considerar elegir el tamaño de la muestra de tal manera que nos permita trabajar con una potencia de la prueba de hipótesis aceptable, es decir, con una adecuada probabilidad de rechazar la hipótesis nula cuando es falsa.

Paso 3: Se identifica la distribución a utilizar

- Si el valor de la varianza poblacional es conocido, se utiliza entonces la distribución normal con media cero y varianza 1.

- En caso contrario, es decir, cuando el valor de la varianza poblacional es desconocido, se utiliza entonces la distribución t de Student.

Paso 4: En dependencia de la hipótesis alternativa, se escoge la regla de decisión adecuada

- La regla de decisión establece la condición para la cual la hipótesis nula es o no rechazada.
- Cuando la hipótesis alternativa es:

1. $H_{1:}\ \mu \neq \mu_0$ la prueba de hipótesis es de dos colas.
2. $H_{1:}\ \mu > \mu_0$ la prueba de hipótesis es de una cola.
3. $H_{1:}\ \mu < \mu_0$ la prueba de hipótesis es de una cola.

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula

- Este último paso consiste en obtener el valor del estadístico de prueba basándonos en los datos de la muestra, compararlo con el valor crítico correspondiente y tomar la decisión de rechazar o no la hipótesis nula.
- Es necesario enfatizar que en una prueba de hipótesis solo es posible una de dos decisiones: la hipótesis nula se rechaza o no se rechaza. Algunos autores prefieren utilizar el término *"se acepta la hipótesis nula"* en lugar de *"no se rechaza la hipótesis nula".*

8.6 Prueba de hipótesis para una proporción.

En el capítulo anterior desarrollamos el procedimiento para calcular un intervalo de confianza para una proporción e indicamos que una proporción es un valor que señala la parte de la muestra de una población que tiene un rasgo distintivo que la identifica, es decir, $p = \frac{X}{n}$ donde X es el número de elementos de la muestra que poseen el rasgo distintivo y n el tamaño de la muestra. En ese capítulo señalamos que en muchas ocasiones se hacía necesario estimar un valor poblacional de una variable de tipo cualitativa nominal, es decir, una variable que se refiere a atributos que no pueden ser representados con números, y pusimos como ejemplo los siguientes casos:

- Aproximadamente el 35% (35 de cada 100) de los graduados de la Facultad de Ciencias Económicas de la Universidad Laica de Manabí poseen al menos un título de 4to nivel.
- El 87% (87 de cada 100) de las viviendas de la ciudad de Portoviejo posee alcantarillado público.
- Una empresa que produce focos incandescentes afirma que la proporción

de este producto con una durabilidad mayor a un año es igual a 0.63 (63 de cada 100).

- La proporción de persona de la 3era edad con bajos recursos económicos en un sector del país es de 0.28 (28 de cada 100).

Veremos a continuación como en presencia de variables cualitativas nominales es posible desarrollar una prueba de hipótesis para una proporción obtenida de las mismas.

Un estudio realizado por la Organización Mundial de la Salud sobre el tabaquismo en adolescentes entre 13 y 15 años en la ciudad de Guayaquil dio como resultado una proporción de 0.12. El Estado ecuatoriano ha desarrollado una fuerte campaña en contra de este nocivo hábito, y en una muestra de 400 adolescentes entre 13 y 15 años de esta ciudad, se encontró que 35 de ellos eran fumadores. Con un nivel de confiabilidad del 95%, ¿los datos de la muestra indican que la proporción de adolescentes de Guayaquil que tiene el hábito de fumar ha disminuido?

Apliquemos el proceso de cinco pasos para una prueba de hipótesis, en este caso, para una proporción.

Paso 1: Se formulan las hipótesis nula y alternativa

En este caso y en forma general, las hipótesis nula y alternativa serian:

$H_0 : \pi = \pi_0 \qquad H_1 : \pi < \pi_0$

Como π_0 = 0.12 es la proporción hipotética entonces en particular:

$H_0 : \pi = 0.12 \quad H_1 : \pi < 0.12$

Paso 2: Se establece el nivel de significación

Si el nivel de confiabilidad deseado es del 95%, entonces el nivel de significación es $\alpha = 0.05$.

Paso 3: Se identifica la distribución a utilizar

La distribución a utilizar es la normal y el percentil a calcular es

$Z_{1-\alpha} = Z_{1-0.05} = Z_{0.95} = 1.64$

Paso 4: En dependencia de la hipótesis alternativa, se escoge la regla de decisión adecuada

Para este caso la regla de decisión quedaría como sigue:

Rechazar H_0 si $\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} < -Z_{1-\alpha}$

No rechazar H_0 si $\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} \geq -Z_{1-\alpha}$

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula

$$p=\frac{X}{n}=\frac{35}{400}=0.09$$

$$\frac{0.09-0.12}{\sqrt{\dfrac{0.12(1-0.12)}{400}}}=\frac{-0.03}{0.02}=-1.5$$

Como -1.5 es mayor que -1.64 no se rechaza la hipótesis nula, y en consecuencia, no existen evidencias de que la proporción de adolescentes de Guayaquil que tiene el hábito de fumar haya disminuido.

Las reglas de decisión para los dos casos restantes de prueba de hipótesis para una proporción son:

1) $H_0:\pi=\pi_0 \quad H_1:\pi\neq\pi_0$

Rechazar H_0 si $\left|\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}\right| > Z_{1-\frac{\alpha}{2}}$

No rechazar H_0 si $\left|\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}\right| \leq Z_{1-\frac{\alpha}{2}}$

2) $H_0:\pi=\pi_0 \quad H_1:\pi>\pi_0$

Rechazar H_0 si $\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} > Z_{1-\alpha}$

No rechazar H_0 si $\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} \leq Z_{1-\alpha}$

Ejercicios del capítulo

8.1 Un investigador especialista en hábitos de comportamiento afirma que los niños con una edad entre 3 y 5 años dedican 20 horas semanales a ver la televisión. Una empresa cuya función es también el estudio de los hábitos de comportamiento no está de acuerdo con esta afirmación, y para comprobarlo, escogió una muestra aleatoria de tamaño 24 de niños con la edad ya señalada la cual dio como resultado los siguientes tiempos de dedicación semanal a la televisión:

21	20	23	22	20	20	24	21	20	21	22	24
20	20	22	24	21	23	22	20	23	22	20	21

Con un nivel de significación del 1%, ¿podemos llegar a la conclusión que el tiempo semanal dedicado por los niños a ver la televisión es diferente al afirmado por el investigador?

8.2 Con los mismos datos del ejercicio anterior y para un nivel de significación del 5%, ¿podemos llegar a la conclusión que el tiempo semanal dedicado por los niños a ver la televisión es mayor al afirmado por el investigador? Considere una varianza poblacional igual a 2.8.

8.3 Un metodólogo del Ministerio de Educación tiene la hipótesis de que los estudiantes de enseñanza media dedican un tiempo semanal de 10 horas al estudio y cumplimiento de sus deberes escolares, y para comprobarlo, extrajo una muestra del tiempo de dedicación de 16 estudiantes la cual dio los resultados que se muestran a continuación:

12	10	9	8	10	11	8	10
11	12	10	10	9	8	11	10

Con un nivel de significación del 5%, ¿podemos afirmar que el tiempo semanal dedicado al estudio por los estudiantes de enseñanza media es menor al afirmado por el metodólogo? Considere una varianza poblacional igual a 1.94.

8.4 Con los mismos datos del ejercicio anterior y utilizando un nivel de significación del 0.1%, ¿podemos afirmar que el tiempo semanal dedicado al estudio por los estudiantes de enseñanza media es diferente al afirmado por el metodólogo?

8.5 Un estudio realizado en el año 2008 en la Escuela Superior Politécnica de Chimborazo señala que la demanda por parte de los estudiantes para seguir una carrera en el área de las Ciencias Económicas y Administrativas es aproximadamente igual al 14%. En una investigación actual desarrollada en otra universidad se obtuvo que de 500 estudiantes encuestados 65 expresaron su interés en estudiar una carrera

en el área de las Ciencias Económicas y Administrativas. Con una confiabilidad del 95%, ¿hay evidencias que permitan asegurar que el interés de los estudiantes por estudiar carreras del área señalada ha disminuido?

8.6 Datos suministrados por el Ministerio de Inclusión Económica y Social señalan que a finales del año 2011 aproximadamente el 14% de los ecuatorianos eran beneficiarios del Bono de Desarrollo Humano. Una encuesta realizada recientemente arrojó que de 1000 personas entrevistadas 165 eran beneficiarios del bono. Con un nivel de significación del 1%, ¿podemos llegar a la conclusión que la proporción de beneficiarios del bono se ha incrementado?

Capítulo 9

Prueba de Hipótesis para dos muestras

El problema

El gerente general de una compañía estatal productora de autos livianos desea comparar el número de kilómetros por litro que recorren sus automóviles con relación a los que produce una compañía privada y cuyos autos tienen el mismo cilindraje.

Para ello extrajo una muestra de 50 autos producidos por su compañía la cual arrojó un promedio de 11.3 km/l con una varianza poblacional de 4.3 km/l.

Adicionalmente obtuvo una muestra de 45 autos producidos por la compañía privada la cual dio por resultado un promedio de 11.9 km/l con una varianza poblacional igual a 4.6 km/l.

Con un nivel de significación del 5%, ¿hay motivos que nos permitan llegar a la conclusión que los autos producidos por la compañía estatal hacen una menor cantidad de km/l que la compañía privada?

9.1 Introducción.

En el capítulo anterior comenzamos el estudio de las pruebas de hipótesis, las cuales son procedimientos que nos permiten comprobar la validez o no de un enunciado establecido con relación a un parámetro poblacional.

Pero en ese capítulo nos limitamos a estudiar el procedimiento estadístico para una sola población, es decir, se extrajo una sola muestra aleatoria para con ella decidir si era o no razonable *la suposición* establecida con relación a un valor poblacional.

En este capítulo ampliaremos el estudio de las pruebas de hipótesis para dos poblaciones y a través de la selección de dos muestras diferentes decidiremos si las medias o proporciones poblacionales son iguales o no.

Iniciaremos la descripción del procedimiento estadístico resolviendo el *problema* planteado al inicio del capítulo.

9.2 Prueba de hipótesis para las medias de dos muestras independientes con varianzas σ_1^2 y σ_2^2 conocidas.

En primer lugar observe que las dos muestras extraídas son *independientes* ya que los kilómetros por litro que recorre un auto de la compañía estatal no está relacionado en lo absoluto con lo que recorre un auto de la compañía privada. Según los datos del problema y denotando con **e** a la compañía estatal y con ***p*** a la privada

tenemos:

$n_e = 50 \quad n_p = 45 \quad \bar{x}_e = 11.3 \quad \bar{x}_p = 11.9 \quad \sigma_e^2 = 4.3 \quad \sigma_p^2 = 4.6$

En el Capítulo 6 estudiamos que la distribución muestral de la diferencia entre dos medias muestrales, calculadas a partir de muestras aleatorias independientes de tamaño n_1 y n_2 extraídas de dos poblaciones distribuidas normalmente con varianzas σ_1^2 y σ_2^2 conocidas, estará también distribuida normalmente con media $\mu_1 - \mu_2$ y varianza $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

Si n_1 y n_2 son grandes, entonces la distribución muestral de la diferencia entre las dos medias será aproximadamente normal con la media y la varianza señalada, con independencia de la forma funcional de las poblaciones originales.

Para desarrollar la prueba de hipótesis sigamos el procedimiento de los *cinco pasos* propuesto en el capítulo anterior.

Paso 1: Se formulan las hipótesis nula y alternativa

La hipótesis nula es que no existen diferencias entre la cantidad de kilómetros por litro que recorren los autos producidos por la compañía estatal y los producidos por la compañía privada, mientras que la alternativa es que la cantidad de kilómetros por litro que recorren los autos producidos por la compañía estatal es menor que los recorridos por la empresa privada, es decir, $H_{0:}\ \mu_e = \mu_p \qquad H_{1:}\ \mu_e < \mu_p$

Paso 2: Se establece el nivel de significación

El nivel de significación elegido por el gerente es del 5%.

Paso 3: Se identifica la distribución a utilizar

Como el valor de las varianzas de ambas poblaciones es conocido, se utiliza la distribución normal con media cero y varianza 1.

Paso 4: En dependencia de la hipótesis alternativa, se escoge la regla de decisión adecuada

Rechazar H_0 si $\dfrac{\bar{x}_e - \bar{x}_p}{\sqrt{\dfrac{\sigma_e^2}{n_e} + \dfrac{\sigma_p^2}{n_p}}} < -Z_{1-\alpha}$

No rechazar H_0 si $\dfrac{\bar{x}_e - \bar{x}_p}{\sqrt{\dfrac{\sigma_e^2}{n_e} + \dfrac{\sigma_p^2}{n_p}}} \geq -Z_{1-\alpha}$

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula

$Z_{1-\alpha} = Z_{1-0.05} = Z_{0.95} = 1.64$

$$\frac{11.3-11.9}{\sqrt{\frac{4.3}{50}+\frac{4.6}{45}}}=\frac{-0.6}{0.43}=-1.4$$

Como -1.4 > -1.64 no se rechaza la hipótesis nula, y por tanto, los datos de la muestra no permiten llegar a la conclusión que los autos producidos por la compañía estatal hacen una menor cantidad de kilómetros por litro que la compañía privada.

Las reglas de decisión para los dos casos restantes de prueba de hipótesis son:

1) $H_{0:}\ \mu_1 = \mu_2 \qquad H_{1:}\ \mu_1 \neq \mu_2$

$$\text{Rechazar } H_0 \text{ si } \left|\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{\sigma_1}{n_1}+\frac{\sigma_2}{n_2}}}\right| > Z_{1-\frac{\alpha}{2}}$$

$$\text{No rechazar } H_0 \text{ si } \left|\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{\sigma_1}{n_1}+\frac{\sigma_2}{n_2}}}\right| \leq Z_{1-\frac{\alpha}{2}}$$

2) $H_{0:}\ \mu_1 = \mu_2 \qquad H_{1:}\ \mu_1 > \mu_2$

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} > Z_{1-\alpha}$$

$$\text{No rechazar } H_0 \text{ si } \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leq Z_{1-\alpha}$$

9.3 Prueba de hipótesis para medias de dos muestras independientes con varianzas σ_1^2 y σ_2^2 desconocidas e iguales.

Consideremos el caso que más que una excepción es una regla, en el que las

varianzas de ambas poblacionales son desconocidas pero iguales.

En el Capítulo 6 vimos que si $\bar{x}_1$ y s_1^2 son respectivamente la media y la varianza de una muestra de tamaño n_1 extraída de una población distribuida normalmente con media μ_1 y varianza σ_1^2, y si $\bar{x}_2$ y s_2^2 son respectivamente la media y la varianza de una muestra de tamaño n_2 extraída de otra población distribuida normalmente con media μ_2 y varianza σ_2^2 igual a σ_1^2, entonces el valor

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_c^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

está distribuido según una t de Student, con $n_1 + n_2 - 2$ grados de libertad, donde:

$$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

se denomina varianza muestral combinada.

Recordado este aspecto desarrollemos el siguiente ejemplo. Nestlé S.A., la compañía agroalimentaria más grande del mundo, desea conocer si existe diferencia en el volumen de venta de la marca de leche La Lechera cuando ésta viene envasada en fundas plásticas o en envases de cartón. Para ello, se extrajeron muestras del volumen de venta medidos en miles de unidades para cada uno de los dos tipos de envase.

Los resultados se muestran en las tablas 9.1 y 9.2.

TABLA 9.1 Volumen de venta de leche La Lechera en envase de cartón

1.78	1.66	1.84	1.81	1.69	1.73	1.68	1.81	1.83	1.77
1.75	1.68	1.82	1.78	1.77	1.83	1.71	1.73	1.88	1.69

TABLA 9.2 Volumen de venta de leche La Lechera en funda plástica

1.32	1.18	1.41	1.25	1.36	1.39	1.42	1.38	1.17	1.21
1.43	1.36	1.32	1.33	1.28	1.24	1.15	1.42	1.37	1.31

Partiendo del supuesto que las dos muestras extraídas son independientes y provienen de poblaciones normales con varianzas desconocidas e iguales, ¿podemos asegurar con un nivel de significación del 1% que el volumen de venta de leche envasada en cartón es mayor que la que se envasa en funda?

Paso 1: Se formulan las hipótesis nula y alternativa

$H_{0:}\ \mu_1 = \mu_2$ $\quad$ $H_{1:}\ \mu_1 > \mu_2$

Paso 2: Se establece el nivel de significación

$\alpha = 0.01$

Paso 3: Se identifica la distribución a utilizar

La distribución es la t de Student con 20+20-2 = 38 grados de libertad.

Paso 4: En dependencia de la hipótesis alternativa, se escoge la regla de decisión adecuada

Por ser $\mu_1 = \mu_2$, es decir, $\mu_1 - \mu_2 = 0$, entonces la regla de decisión queda:

$$\text{Rechazar } H_0 \text{ si } \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_c^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} > t_{0.01}^{(38)}$$

$$\text{Rechazar } H_0 \text{ si } \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_c^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \leq t_{0.01}^{(38)}$$

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula

$$n_1 = 20 \quad n_2 = 20 \quad \bar{x}_1 = \frac{\sum x_{1i}}{n_1} = \frac{35.24}{20} = 1.76 \quad \bar{x}_2 = \frac{\sum x_{2i}}{n_2} = \frac{26.3}{20} = 1.32$$

$$s_1^2 = \frac{\sum (x_{1i} - \bar{x}_1)^2}{n-1} = \frac{0.08}{19} = 0.004 \quad s_2^2 = \frac{\sum (x_{2i} - \bar{x}_2)^2}{n-1} = \frac{0.15}{19} = 0.01$$

$$s_c^2 = \frac{(20-1)(0.004) + (20-1)(0.01)}{20+20-2} = \frac{0.27}{38} = 0.007$$

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_c^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{1.76-1.32}{\sqrt{0.007\left(\frac{1}{20}+\frac{1}{20}\right)}} = \frac{0.44}{0.03} = 14.67$$

El percentil de la t de Student para un nivel de significación del 1% y una prueba de una cola con 30 grados de libertad tiene un valor igual a 2.457, y con 40 grados de libertad 2.423, es decir, cuando los grados de libertad aumentan 10 unidades (de 30 a 40) el percentil disminuye 0.034 (de 2.457 a 2.423). ¿Cuánto disminuye el percentil (X) cuando los grados de libertad aumentan en 8 unidades (de 30 a 38)? Apliquemos una regla de tres simple:

10 0.034
8 X

$X = \frac{8x\,0.034}{10} = 0.027$ y entonces $t_{0.01}^{(38)} = 2.457 - 0.027 = 2.43$

Como 14.67 es mayor que 2.43 rechazamos la hipótesis nula y concluimos que efectivamente el volumen de venta de la leche envasada en cartón es mayor que el que se obtiene con la leche envasada en funda.

Las reglas de decisión para los dos casos restantes de prueba de hipótesis para dos muestras independientes con varianzas poblacionales desconocidas iguales son:

1) $H_{0:}\ \mu_1 = \mu_2$ $H_{1:}\ \mu_1 \neq \mu_2$

Rechazar H_0 si $\left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_c^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| > t_\alpha^{(n_1+n_2-2)}$

No rechazar H_0 si $\left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_c^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \leq t_\alpha^{(n_1+n_2-2)}$

2) $H_{0:}\ \mu_1 = \mu_2$ $H_{1:}\ \mu_1 < \mu_2$

Rechazar H_0 si $\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_c^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < -t_\alpha^{(n_1+n_2-2)}$

No rechazar H_0 si $\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_c^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \geq -t_\alpha^{(n_1+n_2-2)}$

9.4 Prueba de hipótesis para medias de dos muestras independientes con varianzas σ_1^2 y σ_2^2 desconocidas y desiguales.

En el numeral anterior se consideró que las varianzas poblacionales además de desconocidas eran iguales, sin embargo, en muchas ocasiones esta suposición puede

no ser razonable y en consecuencia las varianzas podrían resultar desiguales.

En un caso como éste el procedimiento para desarrollar una prueba de hipótesis presenta algunas modificaciones.

Para ilustrar esta situación utilicemos el siguiente ejemplo. El gerente de una importante línea aérea desea comparar el peso del equipaje de los pasajeros adultos que pertenecen al sexo femenino con los que son de sexo masculino, y para ello obtuvo una muestra aleatoria del peso en kilogramos del equipaje de 36 pasajeros adultos de los cuales 18 eran mujeres y otra cifra igual eran hombres. Los resultados de ambas muestras se aprecian en las tablas 9.3 y 9.4.

TABLA 9.3 Peso del equipaje de pasajeros adultos mujeres

32.3	31.5	33.4	32.8	30.6	31.3	33.8	34.2	33.6
31.7	32.5	30.4	31.2	32.4	30.8	30.5	31.2	32.6

TABLA 9.4 Peso del equipaje de pasajeros adultos hombres

31.2	30.4	32.1	30.6	30.2	31.3	32.6	32.9	33.4
30.4	31.6	30.1	30.8	31.3	30.4	30.2	31.1	32.3

Con un nivel de significación del 0.1%, ¿los datos de las muestras obtenidas ofrecen evidencias que permitan asegurar que el peso del equipaje transportado por mujeres y el transportado por hombres son diferentes?

Paso 1: Se formulan las hipótesis nula y alternativa

$H_{0:}\ \mu_1 = \mu_2$ $\quad$ $H_{1:}\ \mu_1 \neq \mu_2$

Paso 2: Se establece el nivel de significación

$\alpha = 0.001$

Paso 3: Se identifica la distribución a utilizar

Por ser las varianzas poblacionales desconocidas, la distribución que debemos utilizar es la t de Student para una prueba de dos colas.

Como señalamos en el Capítulo 6 los grados de libertad del percentil de la t de Student vienen dados por la expresión:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Paso 4: En dependencia de la hipótesis alternativa, se escoge la regla de decisión adecuada

En este caso, la regla de decisión queda como se muestra a continuación:

Rechazar H_0 si $\left|\frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}}\right| > t_\alpha^{(gl)}$

No rechazar H_0 si $\left|\frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}}\right| \leq t_\alpha^{(gl)}$

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula

$$n_1=18 \quad n_2=18 \quad \bar{x}_1=\frac{\sum x_{1i}}{n_1}=\frac{576.8}{18}=32.04 \quad \bar{x}_2=\frac{\sum x_{2i}}{n_2}=\frac{562.9}{18}=31.27$$

$$s_1^2=\frac{\sum(x_{1i}-\bar{x}_1)^2}{n-1}=\frac{24.38}{17}=1.43 \quad s_2^2=\frac{\sum(x_{2i}-\bar{x}_2)^2}{n-1}=\frac{17.46}{17}=1.03$$

$$\left|\frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}}\right|=\left|\frac{32.04-31.27}{\sqrt{\left(\frac{1.43}{18}+\frac{1.03}{18}\right)}}\right|=\left|\frac{0.77}{0.37}\right|=2.08$$

$$gl=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}=\frac{\left(\frac{1.43}{18}+\frac{1.03}{18}\right)^2}{\frac{\left(\frac{1.43}{18}\right)^2}{18-1}+\frac{\left(\frac{1.03}{18}\right)^2}{18-1}}=\frac{0.02}{0.0006}=33.33\approx 33$$

El percentil de la t de Student para un nivel de significación del 0.1% y una prueba de dos colas con 30 grados de libertad tiene un valor igual a 3.646, y con 40 grados de libertad 3.551, es decir, cuando los grados de libertad aumentan 10 unidades (de 30 a 40) el percentil disminuye 0.095 (3.646 a 3.551). ¿Cuánto disminuye el percentil (X) cuando los grados de libertad aumentan en 3 unidades (de 30 a 33)? Apliquemos

una regla de tres simple:

10	0.095
3	X

$X = \frac{3\,x\,0.095}{10} = 0.028$ y entonces $t_{0.001}^{(3)} = 3.646 - 0.028 = 3.618$

Como 2.08 es menor que 3.618 no rechazamos la hipótesis nula, y en consecuencia, no podemos asegurar que el peso del equipaje transportado por mujeres y el transportado por hombres sean diferentes.

Las reglas de decisión para los dos casos restantes de prueba de hipótesis para medias de dos muestras independientes con varianzas poblacionales desconocidas y desiguales son:

1) $H_{0:}\ \mu_1 = \mu_2$ $H_{1:}\ \mu_1 > \mu_2$

Rechazar H_0 si $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} > t_\alpha^{(gl)}$

No rechazar H_0 si $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \leq t_\alpha^{(gl)}$

2) $H_{0:}\ \mu_1 = \mu_2$ $H_{1:}\ \mu_1 < \mu_2$

Rechazar H_0 si $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} < -t_\alpha^{(gl)}$

No rechazar H_0 si $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \geq -t_\alpha^{(gl)}$

9.5 Prueba de hipótesis para proporciones de dos muestras independientes.

En el Capítulo 6 estudiamos que si p_1 es una proporción muestral calculada a partir de todas las muestras aleatorias de tamaño n_1 que se pueden extraer de una

población con parámetro π_1 y p_2 es una proporción muestral calculada a partir de todas las muestras aleatorias de tamaño n_2 que se pueden extraer de una población con parámetro π_2, entonces la distribución muestral de $p_1 - p_2$ tiene una media igual a $\pi_1 - \pi_2$ y una varianza igual a $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

Si n_1 y n_2 son lo suficientemente grandes, entonces la distribución muestral de $p_1 - p_2$ se aproxima a una distribución normal. Veamos el siguiente ejemplo.

Una compañía de calzado para hombres ha puesto recientemente a la venta un modelo que trata de satisfacer tanto el gusto de personas jóvenes como también de personas mayores, y desea conocer si realmente el nuevo modelo de calzado ha logrado este objetivo.

Para ello extrajo una muestra de 100 hombres jóvenes en la que se obtuvo como resultado que 31 expresaron interés por comprar el calzado, y de forma análoga extrajo una muestra de 100 hombres mayores de los cuales 58 expresaron su disposición por comprar el producto.

Con un nivel de confiabilidad del 95%, ¿considera que existe una diferencia en la proporción de hombres jóvenes y mayores interesados en comprar el nuevo modelo de calzado?

Paso 1: Se formulan las hipótesis nula y alternativa

$$H_0 : \pi_j = \pi_m \qquad H_1 : \pi_j \neq \pi_m$$

Paso 2: Se establece el nivel de significación

α = 0.01

Paso 3: Se identifica la distribución a utilizar

Se utiliza la distribución normal en una prueba de dos colas.

Paso 4: En dependencia de la hipótesis alternativa, se escoge la regla de decisión adecuada

Rechazar H_0 si $\left| \frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} \right| > Z_{1-\frac{\alpha}{2}}$

No rechazar H_0 si $\left| \frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} \right| \leq Z_{1-\frac{\alpha}{2}}$

donde la expresión $p_c = \frac{X_1 + X_2}{n_1 + n_2}$ es conocida con el nombre de *proporción conjunta*.

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula

$$Z_{1-\frac{\alpha}{2}} = Z_{1-\frac{0.01}{2}} = Z_{0.995} = 2.58$$

$$p_1 = \frac{X_1}{n_1} = \frac{31}{100} = 0.31 \quad p_2 = \frac{X_2}{n_2} = \frac{58}{100} = 0.58$$

$$p_c = \frac{X_1 + X_2}{n_1 + n_2} = \frac{31 + 58}{100 + 100} = \frac{89}{200} = 0.44$$

$$\left| \frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} \right| = \left| \frac{0.31 - 0.58}{\sqrt{\frac{0.44(1-0.44)}{100} + \frac{0.44(1-0.44)}{100}}} \right| =$$

$$\left| \frac{-0.27}{0.07} \right| = 3.86$$

Por ser 3.86 mayor que 2.58 rechazamos la hipótesis nula, y en consecuencia, existen diferencias entre la proporción de hombres jóvenes y mayores interesados en comprar el nuevo modelo de calzado.

Las reglas de decisión para los dos casos restantes de prueba de hipótesis para proporciones de dos muestras independientes son:

1) $H_{0:}\ \pi_1 = \pi_2 \qquad H_{1:}\ \pi_1 > \pi_2$

Rechazar H_0 si $$\frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} > Z_{1-\alpha}$$

No rechazar H_0 si $$\frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} \leq Z_{1-\alpha}$$

2) $H_{0:}\ \pi_1 = \pi_2 \qquad H_{1:}\ \pi_1 < \pi_2$

Rechazar H_0 si $$\frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} < -Z_{1-\alpha}$$

No rechazar H_0 si $$\frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} \geq -Z_{1-\alpha}$$

9.6 Prueba de hipótesis de dos muestras dependientes. Muestra apareada.

Hasta el momento, todas las pruebas de hipótesis desarrolladas en este capítulo se hicieron sobre la base de que las dos muestras obtenidas eran *independientes*, es decir, no estaban *relacionadas*.

Sin embargo, existen en la práctica casos en los cuales las muestras *no son independientes*, o dicho de otra manera, las muestras *están relacionadas*.

Una situación como esta se presenta cuando a una misma unidad experimental se le miden dos características diferentes con el ánimo de compararlas posteriormente, por ejemplo, cuando le medimos a una misma persona el nivel de aprendizaje de un idioma utilizando dos procedimientos metodológicos diferentes. Veamos uno de estos casos.

Existe el criterio compartido por algunos conductores y rechazados por otros, que conducir con el aire acondicionado en funcionamiento no provoca realmente un mayor gasto de combustible pues a pesar que el uso de este equipo determina un consumo adicional del mismo, también es cierto que en esa situación existe menos resistencia del aire sobre el auto pues se conduce con las ventanillas cerradas, provocando un ahorro de combustible por esa vía.

Una compañía productora de equipos de aire acondicionado para automóviles desea comprobar si existen diferencias en la cantidad de kilómetros recorridos por litro cuando se conduce el auto sin aire acondicionado o con él, y para ello, midió el kilometraje por litro de 15 autos que recorrieron una distancia determinada sin el aire acondicionado y con las ventanillas abiertas, midiendo posteriormente dicho kilometraje con *los mismos autos* y *los mismos conductores* pero con el aire acondicionado en funcionamiento.

Los resultados se muestran en la tabla 9.5.

TABLA 9.5 Kilómetros por litro recorridos por 15 autos

Auto	Sin aire acondicionado km/l	Con aire acondicionado km/l
1	12.7	12.4
2	12.5	12.3
3	12.3	12.5
4	12.9	12.5

Auto	Sin aire acondicionado km/l	Con aire acondicionado km/l
5	12.7	12.5
6	12.5	12.3
7	12.6	12.6
8	12.7	12.9
9	12.9	12.7
10	12.4	12.7
11	12.5	12.5
12	12.6	12.3
13	12.9	12.6
14	12.7	12.2
15	12.5	12.3

Observe en la tabla 9.5 que el recorrido en kilómetros por litro de los autos conducidos sin aire acondicionado y con él, se le midió *al mismo auto conducido por el mismo conductor,* y por tanto, ambas observaciones no son independientes. A una muestra con las características señaladas anteriormente se le conoce con el nombre de *muestra apareada*.

En este tipo de situación realmente estamos trabajando *con una muestra* pues lo que nos interesa es conocer si la distribución de las diferencias entre las distancias recorridas por litro es igual a 0. Por tanto la muestra en realidad son las diferencias entre estos recorridos y no los recorridos individuales.

Emplearemos el símbolo μ_d para referirnos a la media poblacional de la distribución de las diferencias.

Como ya hemos reiterado en múltiples ocasiones el primer paso consiste en formular la hipótesis nula y la hipótesis alternativa. Debido a que el interés de la compañía es conocer si existen *diferencias* en el recorrido de los autos al conducir sin aire acondicionado o con él, resulta razonable establecer una prueba de dos colas, es decir,

$$H_0 : \mu_d = 0 \qquad H_1 : \mu_d \neq 0$$

TABLA 9.6 Diferencia en el gasto de combustible de cada auto

Auto	*Diferencia*(d)	$d-\bar{d}$	$(d-\bar{d})^2$
1	+0.3	+0.16	0.0256
2	+0.2	+0.06	0.0036
3	-0.2	-0.34	0.1156
4	+0.4	+0.26	0.0676
5	+0.2	+0.06	0.0036
6	+0.2	+0.06	0.0036
7	0	-0.14	0.0196
8	-0.2	-0.34	0.1156
9	+0.2	+0.06	0.0036
10	-0.3	-0.44	0.1936
11	0	-0.14	0.0196
12	+0.3	+0.16	0.0256
13	+0.3	+0.16	0.0256
14	+0.5	+0.36	0.1296
15	+0.2	+0.06	0.0036
	2.1		0.7560

$$\bar{d}=\frac{\sum d}{n}=\frac{2.1}{15}=0.14$$

$$s_d=\sqrt{\frac{\sum(d-\bar{d})^2}{n-1}}=\sqrt{\frac{0.756}{14}}=0.23$$

Si decidimos utilizar un nivel de significación del 0.1%, $t_{0.001}^{14}=4.14$

La regla de decisión queda entonces:

Rechazar H_0 si $\left|\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}\right| > t_\alpha^{(n-1)}$

No rechazar H_0 si $\left|\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}\right| \le t_\alpha^{(n-1)}$

para una prueba de dos colas.

$$\left|\frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}\right| = \left|\frac{0.14}{\frac{0.23}{\sqrt{15}}}\right| = \left|\frac{0.14}{0.06}\right| = 2.33$$

y como 2.33 < 4.14 no rechazamos la hipótesis nula y concluimos que no existen diferencias en consumo de combustible entre conducir sin aire acondicionado y con él en funcionamiento.

Las reglas de decisión para los dos casos restantes de prueba de hipótesis (pruebas de una sola cola) para dos muestras dependientes, o lo que es lo mismo, para una *muestra apareada*, se aprecian a continuación:

1) $H_0 : \mu_d = 0 \qquad H_1 : \mu_d > 0$

Rechazar H_0 si $\dfrac{\bar{d}}{\frac{s_d}{\sqrt{n}}} > t_\alpha^{(n-1)}$ No rechazar H_0 si $\dfrac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \leq t_\alpha^{(n-1)}$

2) $H_0 : \mu_d = 0 \qquad H_1 : \mu_d < 0$

Rechazar H_0 si $\dfrac{\bar{d}}{\frac{s_d}{\sqrt{n}}} < -t_\alpha^{(n-1)}$ No rechazar H_0 si $\dfrac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \geq -t_\alpha^{(n-1)}$

9.7 Prueba de hipótesis para las varianzas de dos muestras.

En el Capítulo 5 vimos que si S_1^2 y S_2^2 son varianzas calculadas a partir de muestras aleatorias independientes de tamaño n_1 y n_2, extraídas de poblaciones distribuidas normalmente con varianzas σ_1^2 y σ_2^2 respectivamente, entonces la variable aleatoria:

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$$

sigue una distribución *F* de Fisher con $n_1 - 1$ y $n_2 - 1$ grados de libertad en el numerador y denominador respectivamente.

Cuando las varianzas de ambas poblaciones son iguales entonces $F = \frac{S_1^2}{S_2^2}$

Precisamos también que una de las aplicaciones de la distribución *F* es que nos permite establecer si dos muestras provienen de poblaciones que tienen la misma varianza.

Veamos un ejemplo de comparación de dos varianzas. Los datos que se observan en la tabla 9.7 son muestras aleatorias tomadas en dos empresas A y B correspondiente al sueldo mensual de sus empleados administrativos.

TABLA 9.7 Sueldo mensual de empleados administrativos

A	758	743	766	735	778	761	749	756	732	744
B	728	736	744	731	726	749	752	730	744	732

Con un nivel de confiabilidad del 95%, ¿hay evidencias que nos permitan concluir que las varianzas de ambas poblaciones son diferentes?

Para desarrollar la prueba de hipótesis sigamos el procedimiento de los *cinco pasos* propuesto en el Capítulo 8.

Paso 1: Se formulan las hipótesis nula y alternativa

Las hipótesis nula y alternativa son:

$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$

Paso 2: Se establece el nivel de significación

El nivel de significación elegido es del 5%.

Paso 3: Se identifica la distribución a utilizar

La distribución a utilizar es la *F* de Fisher con un nivel de significación del 5% con 9 grados de libertad en el numerador (hay 10 observaciones en la muestra A) y 9 grados de libertad en el denominador (hay 10 observaciones en la muestra B), es decir, $F_{5\%}(9,9)$.

En la **TABLA T.3** del Anexo A aparecen los valores críticos de la distribución F de Fisher. A continuación se muestra un segmento de dicha tabla.

G.L.		G.L. DEL NUMERADOR								
ERROR	α	1	2	3	4	5	6	7	8	9
7	0.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	0.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	0.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	0.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51
	0.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33
8	0.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	0.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	0.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	0.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34
	0.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77
9	0.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	0.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	0.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	0.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54
	0.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11

Como se observa en la tabla $F_{5\%}(9,9) = 3.18$.

Paso 4: En dependencia de la hipótesis alternativa, se escoge la regla de decisión adecuada

Rechazar H_0 si $F > F_{5\%}(9,9)$ No rechazar H_0 si $F \leq F_{5\%}(9,9)$ donde

$$F = \frac{S_1^2}{S_2^2}$$

$$S_A^2 = (758)^2 + (743)^2 + ... + (744)^2 - \frac{(758 + 743 + ... + 744)^2}{10}$$

$$S_1^2 = S_A^2 = 5659896 - 5658048.4 = 1847.6$$

$$S_B^2 = (728)^2 + (736)^2 + ... + (732)^2 - \frac{(728 + 736 + ... + 732)^2}{10}$$

$$S_2^2 = S_B^2 = 5435418 - 5434638.4 = 779.6$$

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula

Como $F = \frac{1847.6}{779.6} = 2.37$ es menor que 3.18, no rechazamos la hipótesis nula

y en consecuencia, no podemos asegurar con un nivel de significación del 5% que existan diferencias entre las dos varianzas.

Ejercicios del capítulo

9.1 Una empresa productora de pintura con fines domésticos desea comparar dos mezclas diferentes (A y B) de este producto con el objetivo de determinar si existen diferencias entre ellas con relación al tiempo de secado de las mismas. Para ello, escogió muestras independientes de ambas mezclas y les midió el tiempo de secado en minutos, obteniendo los resultados que se muestran a continuación:

Mezclas	**Tiempo de secado en minutos**											
A	30	32	35	28	33	30	30	28	29	33	30	31
B	28	33	32	29	30	30	27	28	30	32	31	30

Con un nivel de confiabilidad del 99.9%, ¿existen elementos que nos permitan concluir que el tiempo de secado de ambas mezclas de pintura es diferente? Considere una varianza poblacional de la mezcla A igual a 5.2 y una varianza poblacional de la mezcla B igual a 3.5.

9.2 La compañía Coca – Cola está interesada en conocer si existe diferencia en el volumen de venta entre la Coca Cola Light y la Coca Cola Zero que ella produce**, y** para ello, extrajo muestras independientes del volumen de venta de ambos productos medido en miles de unidades. Los resultados obtenidos fueron los siguientes:

Producto	**Volúmenes de venta**									
Coca Cola	1.75	1.69	1.72	1.63	1.67	1.77	1.73	1.81	1.79	1.66
Light	1.81	1.73	1.64	1.69	1.72	1.77	1.66	1.67	1.74	1.82
Coca Cola	1.63	1.65	1.67	1.59	1.58	1.66	1.65	1.72	1.71	1.61
Zero	1.74	1.69	1.63	1.59	1.57	1.65	1.62	1.67	1.63	1.58

Con un nivel de confiabilidad del 99%, ¿existen elementos que nos permitan concluir que el volumen de venta de la Coca Cola Light es diferente al de la Coca Cola Zero? Considere una varianza poblacional de la Coca Cola Light igual a 0.01 y una varianza poblacional de la Coca Cola Zero igual a 0.03.

9.3 Utilizando los tiempos de secado en minutos de las dos mezclas de pintura del ejercicio 9.1 y con un nivel de significación del 1%, ¿existen elementos que nos permitan concluir que el tiempo de secado de ambas mezclas de pintura es diferente? Considere que las varianzas poblacionales son desconocidas e iguales.

9.4 Utilizando los volúmenes de venta de la Coca Cola Light y la Coca Cola Zero del ejercicio 9.2 y con un nivel de significación del 5%, ¿existen evidencias que nos permitan llegar a la conclusión que el volumen de venta de la Coca Cola Light es diferente al de la Coca Cola Zero? Considere que las varianzas poblacionales son desconocidas e iguales.

9.5 Utilizando los tiempos de secado en minutos de las dos mezclas de pintura del ejercicio 9.1 y con un nivel de significación del 5%, ¿existen elementos que nos

permitan concluir que el tiempo de secado de ambas mezclas de pintura es diferente? Considere que las varianzas poblacionales son desconocidas y desiguales.

9.6 Utilizando los volúmenes de venta de la Coca Cola Light y la Coca Cola Zero del ejercicio 9.2 y con un nivel de significación del 0.1%, ¿existen evidencias que nos permitan llegar a la conclusión que el volumen de venta de la Coca Cola Light es diferente al de la Coca Cola Zero? Considere que las varianzas poblacionales son desconocidas y desiguales.

9.7 Una compañía de seguros está interesada en conocer si existen diferencias entre la proporción de hombres y mujeres que tienen su auto asegurado, y para conseguirlo, extrajo muestras independientes de 100 hombres y 100 mujeres de la población la cual dio como resultado que 62 hombres y 57 mujeres tenían su auto asegurado. Con un nivel de significación del 1%, ¿es la proporción poblacional de hombres que tiene su auto asegurado mayor a la de las mujeres?

9.8 Si se extraen dos muestras independientes de 200 profesores cada una en una universidad en proceso de elecciones de sus autoridades, y en la primera de ellas 160 manifiestan que votarán para rector por el candidato A, mientras que en la segunda 125 tienen la intención de hacerlo por el candidato B, ¿podemos estar seguros con una confiabilidad del 95% que las elecciones serán ganadas por el candidato A?

9.9 El gerente de un prestigioso banco tiene la hipótesis de que el continuo e intenso trabajo al cual se ven sometidas las personas que laboran en el balcón de servicios, pueden provocar un incremento de la tensión arterial diastólica de las mismas. Para comprobarlo midió la presión diastólica de 15 personas que laboran en esta área antes de comenzar la jornada laboral y al concluir la misma. Los resultados alcanzados en la investigación desarrollada son los que se muestran a continuación:

	Presión diastólica														
Antes	72	75	76	73	74	77	76	74	78	75	80	79	71	76	77
Después	83	85	75	84	83	76	83	81	77	82	83	78	80	80	76

Con un nivel de significación del 5%, ¿es razonable pensar que según los resultados obtenidos, la presión diastólica de las personas que laboran en los balcones de servicio es menor al iniciar la jornada laboral que al finalizarla?

9.10 Se dice que el efecto de la fuerza de gravedad que soportamos durante el día hace que nuestra estatura al levantarnos en la mañana sea mayor que al acostarnos en la noche. Un investigador que desea comprobar lo expresado midió la estatura en cm de 12 personas al momento de levantarse y cuando se disponían a acostarse, obteniendo los siguientes resultados:

	Estatura											
Al levantarse	172	166	183	175	161	166	173	181	184	177	163	161
Al acostarse	170	164	182	173	160	165	172	179	182	175	162	159

Con un nivel de significación del 1%, ¿podemos concluir que según los resultados obtenidos, la estatura de las personas al levantarse es mayor que al acostarse?

Capítulo 10
Análisis de varianza

El problema

El gerente general de la empresa Conservas Isabel, la cual en la actualidad es una de las plantas de procesamiento de atún más moderna del continente americano, desea conocer si existen diferencias en los volúmenes de venta de atún enlatado solo con agua o con aceites de Soya, Girasol y Oliva. ¿Qué método estadístico deberá emplear para cumplir su objetivo y que procedimiento conlleva la utilización de ese método?

10.1 Introducción.

La respuesta a la pregunta formulada en el párrafo anterior, es que el método estadístico que nos permite realizar una prueba de hipótesis para comprobar si existen o no diferencias en las medias poblacionales de las cuatro poblaciones sometidas a prueba, es el conocido con el nombre de Análisis de Varianza. En los Capítulos 8 y 9 estudiamos la teoría general de las pruebas de hipótesis para medias y proporciones en los casos de una muestra y dos muestras independientes o no. El presente capítulo lo dedicaremos a ampliar lo relacionado con las pruebas de hipótesis cuando deseamos comparar varias medias para poder determinar si provienen de poblaciones iguales.

10.2 El modelo lineal general.

Siempre que pretendamos realizar un análisis de resultados obtenidos en una investigación, se hará casi imprescindible que cada observación obtenida sea representada mediante un modelo matemático que exprese los diferentes factores que de una manera u otra han sido responsables de esa respuesta.

Si estamos interesados, por ejemplo, en estudiar las pérdidas que se han producido en una empresa, el valor numérico de una de estas pérdidas podría ser el efecto de un número más o menos grande de diferentes causas o factores que pueden ser expresados mediante un modelo matemático que permita conocer cuál de ellos son los realmente importantes, y poder estimar su grado de contribución a las pérdidas alcanzadas.

Factores tales como los costos de las materias primas, el costo de la mano de obra, la disponibilidad semanal de tiempo, y otros más, son factores que pueden haber influido en las pérdidas y que por tanto el modelo debe considerar y estimar. Veamos de forma general, la expresión matemática del modelo al cual estamos haciendo referencia.

Supongamos que $y_1, y_2, \ldots\ldots, y_n$ es un conjunto formado por n datos u observaciones provenientes de una investigación y sobre las cuales pueden estar influyendo o no el efecto de p factores $\beta_1, \beta_2, \ldots\ldots\ldots\ldots, \beta_p$.

El modelo lineal general para la i – ésima observación del modelo viene dado entonces por la expresión:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots\ldots\ldots\ldots + x_{pi}\beta_p + e_i$$

donde i = 1, 2, …….. , n y e_i son los *errores experimentales*, es decir, la diferencia entre el verdadero valor de y_i y el estimado mediante el modelo, o sea:

$$e_i = y_i - \hat{y}_i$$

Los coeficientes x_{ji} conocidos con el nombre de variables indicadoras, son utilizados para *indicar* en un caso específico cuando un parámetro determinado es considerado o no en el modelo, siendo el coeficiente x_{ji} igual a 1 o 0 respectivamente. El modelo en forma matricial queda expresado de la siguiente forma:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11}\ x_{21}\ x_{31} \ldots\ldots\ldots\ldots x_{p1} \\ x_{12}\ x_{22}\ x_{32} \ldots\ldots\ldots\ldots x_{p2} \\ x_{13}\ x_{23}\ x_{33} \ldots\ldots\ldots\ldots x_{p3} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_{1n}\ x_{2n}\ x_{3n} \ldots\ldots\ldots\ldots x_{pn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

o lo que es lo mismo $\mathbf{Y = X'\beta + e}$

10.2.1 Clasificación de los modelos lineales.

Los modelos lineales se clasifican atendiendo a dos aspectos diferentes:

a) Según los valores de las variables x_{ji}.

a.1) Se dice que el modelo es de Análisis de Varianza cuando todas las variables x_{ji} son indicadoras, es decir, solo toman los valores 0 o 1.

a.2) Si por el contrario, todas las variables x_{ji}, excepto una que tiene valor 1, son variables aleatorias, entonces se dice que el modelo es de Análisis de Regresión.

a.3) Cuando algunas de las variables x_{ji} tienen valor 0 o 1 y otras son variables aleatorias, decimos que el modelo es de Análisis de Covarianza.

Según el valor de los parámetros β_j.

b.1) Cuando todos los β_j son valores constantes, decimos que el modelo es de Efectos Fijos.

b.2) Si por el contrario, todos los β_j son variables aleatorias, entonces se dice que el modelo es de Efectos Aleatorios

b.3) Cuando algunos β_j son valores constantes y otros variables aleatorias se dice que el modelo es de Efectos Mixtos.

Durante el desarrollo del presente capítulo estudiaremos los modelos de Análisis de Varianza y de Regresión solo de efectos fijos. Los modelos de efectos aleatorios y mixtos son solo utilizados en casos muy especiales, y por ello no se justifica su inclusión en este libro.

10.2.2 Hipótesis de base.

El desarrollo teórico de los modelos lineales, requiere que se cumplan un grupo de suposiciones iniciales que reciben el nombre de *hipótesis de base*.

Estas hipótesis son las siguientes:

1. Los errores experimentales e_i son independientes.
2. Los errores siguen una distribución normal con media cero y varianza σ^2. De forma más resumida $e_i \sim N(0, \sigma^2)$.
3. La varianza de los errores es homogénea.
4. La homogeneidad de las varianzas significa que las variables aleatorias e_i tienen la misma varianza.

10.2.3 Estimación de los parámetros en el modelo lineal.

Para estimar el valor numérico de los parámetros β_j utilizaremos el *método de los mínimos – cuadrados*, el cual consiste en seleccionar los estimadores de los β_j de forma tal que hagan mínima la suma de los cuadrados de los errores experimentales, es decir, los mejores estimadores β_j son los que hacen mínima la expresión:

$$\phi = \sum e_i^2 = \sum \left(y_i - x_{1i}\beta_1 - x_{2i}\beta_2 - \ldots\ldots\ldots x_{pi}\beta_p \right)^2 \quad \phi = \mathbf{e'e = (Y\text{-}X'\beta)'(Y\text{-}X'\beta)}$$

Con el objetivo de clarificar lo planteado veamos el siguiente ejemplo. Supongamos que estamos interesados en desarrollar una investigación con el objetivo de estudiar el efecto de dos tipos de campañas publicitarias sobre los volúmenes de venta de un determinado artículo expresados en miles de unidades. Cada una de las campañas fue puesta en ejecución en tres diferentes provincias, y en cada una de estas provincias se seleccionó un supermercado para medir las ventas del artículo. El número de artículos vendidos por campaña publicitaria en cada una de las provincias respectivas se muestra en la tabla 10.1

TABLA 10.1 Número de artículos vendidos en cada campaña publicitaria

CAMPAÑAS	ARTÍCULOS VENDIDOS		
A	18	25	24
B	15	18	19

En términos del modelo:

$18 = 1\beta_1+1\beta_2+0\beta_3+e_1$

$25 = 1\beta_1+1\beta_2+0\beta_3+e_2$

$24 = 1\beta_1+1\beta_2+0\beta_3+e_3$

$15 = 1\beta_1+0\beta_2+1\beta_3+e_4$

$18 = 1\beta_1+0\beta_2+1\beta_3+e_5$

$19 = 1\beta_1+0\beta_2+1\beta_3+e_6$

En forma matricial:

$$\mathbf{Y} = \mathbf{X'}\ \beta + \mathbf{e}$$

$$\begin{bmatrix}18\\25\\24\\15\\18\\19\end{bmatrix} = \begin{bmatrix}1&1&0\\1&1&0\\1&1&0\\1&0&1\\1&0&1\\1&0&1\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\\\beta_3\end{bmatrix} + \begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\\e_6\end{bmatrix}$$

donde:

β_1 = es una constante común a todas las observaciones la cual está presente en todos los modelos

β_2 = es el efecto de la campaña A sobre el número de artículos vendidos

β_3 = es el efecto de la campaña B sobre el número de artículos vendidos

e_i = son errores experimentales normalmente distribuidos con media cero y varianza homogénea σ^2.

Sin que esto implique que por hacerlo perderemos generalidad, consideremos que la suma de los efectos de las campañas es igual a cero, es decir, que $\beta_2 + \beta_3 = 0$, lo cual implica que $\beta_3 = -\beta_2$, o expresado de forma más explícita:

$1\beta_3 = -1\beta_2$

El importantísimo resultado expresado en la igualdad anterior puede ser *traducido* diciendo que es lo *mismo* colocar un -1 como coeficiente de β_2 que un 1 como coeficiente de β_3, es decir, puede ser expresado como se muestra a continuación:

$$\begin{bmatrix}18\\25\\24\\15\\18\\19\end{bmatrix}=\begin{bmatrix}1&1\\1&1\\1&1\\1&-1\\1&-1\\1&-1\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\end{bmatrix}+\begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\\e_6\end{bmatrix}$$

de donde ϕ **= e'e = (Y-X'**β**)'(Y-X'**β**)** viene dado entonces por :

$$\phi=(18-\beta_1-\beta_2)^2+(25-\beta_1-\beta_2)^2+(24-\beta_1-\beta_2)^2+(15-\beta_1+\beta_2)^2+(18-\beta_1+\beta_2)^2$$
$$+(19-\beta_1+\beta_2)^2$$

Igualando a cero las respectivas derivadas parciales de ϕ con relación a β_1 y β_2 obtenemos los parámetros que hace mínima la función ϕ:

$$\frac{\partial\varphi}{\partial\beta_1}=2(18-\beta_1-\beta_2)(-1)+2(25-\beta_1-\beta_2)(-1)+2(24-\beta_1-\beta_2)(-1)+2(15-\beta_1+\beta_2)(-1)$$

$$+2(18-\beta_1+\beta_2)(-1)+2(19-\beta_1+\beta_2)(-1)=0$$

$$6\beta_1=18+25+24+15+18+19\quad\beta_1=\frac{119}{6}=19.83$$

Procediendo de idéntica forma para β_2 tenemos:

$$\frac{\partial\phi}{\partial\beta_2}=2(18-\beta_1-\beta_2)(-1)+2(25-\beta_1-\beta_2)(-1)+2(24-\beta_1-\beta_2)(-1)+2(15-\beta_1+\beta_2)(1)$$
$$+2(18-\beta_1+\beta_2)(1)+2(19-\beta_1+\beta_2)(1)=0$$

$$6\beta_2=18+25+24-15-18-19\quad\beta_2=\frac{15}{6}=2.50$$

y como $\beta_3=-\beta_2\quad\beta_3=-2.50$

Hallemos los errores experimentales y la suma de cuadrados de los mismos:

$e_1 = 18 - (\beta_1+\beta_2) = 18 - (19.83 + 2.5) = -4.33$
$e_2 = 25 - (\beta_1+\beta_2) = 25 - (19.83 + 2.5) = 2.67$
$e_3 = 24 - (\beta_1+\beta_2) = 24 - (19.83 + 2.5) = 1.67$
$e_4 = 15 - (\beta_1+\beta_3) = 15 - (19.83 - 2.5) = -2.33$
$e_5 = 18 - (\beta_1+\beta_3) = 18 - (19.83 - 2.5) = 0.67$
$e_6 = 19 - (\beta_1+\beta_3) = 19 - (19.83 - 2.5) = 1.67$

$$\phi = \sum e_i^2 = (-4.33)^2 + (2.67)^2 + (1.67)^2 + (-2.33)^2 + (0.67)^2 + (1.67)^2$$

$$\phi = \sum e_i^2 = 37.33$$

Para cualquier otro valor de β que escojamos, la *suma de cuadrados del error* va a ser mayor a 37.33.

Observe también que el valor de β_1 = 19.83 coincide con el valor de la media general de todos los datos y que la media correspondiente a los volúmenes de venta de la campaña A es igual a $\beta_1 + \beta_2$ = 19.83 + 2.50, es decir, 22.33.

De igual forma la media correspondiente a los volúmenes de venta de la campaña B es igual a $\beta_1 + \beta_3$ = 19.83 - 2.50, es decir, 17.33.

Esto que acabamos de ver es una regla.

10.3 Análisis de Varianza de datos provenientes de un diseño Completamente al Azar.

Para el desarrollo del presente numeral utilizaremos *el problema* del inicio de este capítulo en el que el gerente general de Conservas Isabel desea establecer posibles diferencias en los volúmenes de venta de atún enlatado con agua sola o con aceites de Soya, Girasol y Oliva. Con el objetivo de generalizar para cualquier investigación, se dice, en este caso, que el gerente desea comparar el efecto de cuatro *tratamientos*. En general se entiende por *tratamientos experimentales*, o simplemente *tratamientos,* a los procesos cuyos efectos se pretenden estimar y comparar en una investigación.

Los *tratamientos* en cuestión serían los siguientes:

A: Atún enlatado con agua

B: Atún enlatado con aceite de Soya

C: Atún enlatado con aceite de Girasol

D: Atún enlatado con aceite de Oliva

Si representamos como:

μ_A la media poblacional del tratamiento A

μ_B la media poblacional del tratamiento B

μ_C la media poblacional del tratamiento C

μ_D la media poblacional del tratamiento D, y entoces:

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

H_1 : No todas las medias poblacionales son iguales

Aunque al final de este capítulo precisaremos este aspecto, consideraremos

que se utilizaron ocho réplicas o repeticiones (supermercados) para cada uno de los tratamientos bajo estudio, es decir, los volúmenes de venta de cada tratamiento se midieron en ocho supermercados diferentes. Supongamos además que los 32 supermercados seleccionados para desarrollar la investigación presentaban características similares u homogéneas, lo cual no provocaba preferencias por parte del cliente para realizar su compra en uno u otro de los establecimientos. Esto determina, que sin correr el riesgo de favorecer a un tratamiento sobre el otro, podamos asignarle los tratamientos a los supermercados de forma totalmente aleatoria, tal y como se observa en la siguiente *aleatorización* (Tabla 10.2).

TABLA 10.2 Tratamientos asignados al azar a los supermercados

B	C	A	D	D	B	A	C
D	B	A	B	C	D	C	B
A	B	D	B	A	D	A	A
D	C	D	C	A	C	B	C

Cuando las *unidades experimentales* (supermercados) son asignados de forma totalmente aleatoria a los tratamientos, o viceversa, cuando los tratamientos son asignados de esta manera a las unidades experimentales, se dice que hemos usado *un diseño Completamente al Azar*. Supongamos que al medir los volúmenes de venta para cada tratamiento y en cada supermercado, se obtuvo los resultados que se muestran en la tabla 10.3.

TABLA 10.3 Volúmenes de venta

16.7	16.2	14.2	17.5	17.6	16.9	14.7	16.4
17.8	16.8	14.3	16.5	16.1	17.8	16.1	16.7
14.9	16.4	17.9	16.3	14.5	17.4	14.3	14.6
17.6	16.5	17.4	16.1	14.8	16.7	16.2	16.5

los cuales organizados por tratamientos quedarían:

TABLA 10.4 Volúmenes de venta organizados por tratamientos

									TOTAL	MEDIAS
A	14.9	14.2	14.3	14.5	14.8	14.7	14.3	14.6	116.3	14.54
B	16.7	16.8	16.4	16.5	16.3	16.9	16.2	16.7	132.5	16.56
C	16.2	16.5	16.1	16.1	16.7	16.1	16.4	16.5	130.6	16.33
D	17.8	17.6	17.9	17.4	17.5	17.6	17.8	17.4	141	17.63
									520.4	

El primer paso para iniciar el análisis estadístico de los datos es establecer un modelo matemático para representar cada una de las observaciones del experimento. Un estudio de las observaciones nos hace concluir que las mismas se encuentran

afectadas por tres tipos de componentes:

1) una constante general supuestamente alrededor de la cual fluctúan los valores de las observaciones, 2) una componente que representa el efecto directo del tratamiento, y 3) un efecto residual o error experimental, el cual incluye todos los demás factores que pueden influir en el comportamiento de las observaciones y que no fueron considerados en el diseño experimental.

Si representamos con μ la constante general, con α_i (i=1, 2, 3, 4) el efecto del tratamiento i-ésimo y con e_{ij} (j=1, 2,……,8) el error experimental, entonces el modelo lineal queda expresado como:

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

donde:

$e_i \sim N(0,\sigma^2)$ e_i independientes y σ^2 homogénea.

Si observamos con detenimiento el modelo lineal al que hemos arribado, podremos concluir que el mismo está compuesto por tres *fuentes de variación*:

- La fuente de variación debida a la variable **y**, es decir, debida a los 32 volúmenes de venta presentes en la investigación. Esta fuente de variación recibe el nombre de *TOTAL.*
- Una fuente de variación debida a los tratamientos, y
- Una fuente de variación debida a los errores experimentales.

Con estos elementos podemos comenzar a construir la tabla 10.5 del *Análisis de Varianza***.** Esta tabla, como su nombre lo indica, es un análisis de las varianzas involucradas en la investigación, y que a la postre, nos permitirá rechazar o no la hipótesis nula planteada.

TABLA 10.5 Análisis de varianza

FUENTES DE VARIACION TOTAL	G.L.	S.C.	C.M.	F	SIGN.
TRATAMIENTOS					
ERROR					

Calculemos la varianza Total, la varianza debida a tratamientos y la varianza del error por el método que vimos en un capítulo anterior:

$$S^2_{TOTAL} = \frac{SC_{TOTAL}}{gl} = \frac{\sum (y_{ij} - \bar{y})^2}{n-1} = \frac{\sum y^2_{ij} - \frac{\left(\sum y_{ij}\right)^2}{n}}{n-1}$$

donde y_{ij} son los volúmenes de venta y n-1 sus grados de libertad. Al término $\frac{\left(\sum y_{ij}\right)^2}{n}$ se le conoce con el nombre de *Factor de Corrección*.

$$S^2_{TOTAL} = \frac{SC_{TOTAL}}{gl} = \frac{8503.94 - \frac{(520.4)^2}{32}}{31} = \frac{8503.94 - 8463.00}{31}$$

$$S^2_{TOTAL} = \frac{40.94}{31} = 1.32|$$

$$S^2_{TRAT.} = \frac{SC_{TRAT}}{gl} = \frac{\sum \frac{T_i^2}{m_i} - \frac{\left(\sum y_{ij}\right)^2}{n}}{t-1}$$

donde T_i es el total de venta del tratamiento i-ésimo, t el número de tratamientos y m_i es la cantidad de observaciones por tratamiento.

$$S^2_{TRAT.} = \frac{\frac{(116.3)^2 + (132.5)^2 + (130.6)^2 + (141)^2}{8} - \frac{(520.4)^2}{32}}{3}$$

$$S^2_{TRAT.} = \frac{8502.41 - 8463.00}{3} = \frac{39.41}{3} = 13.14$$

A esta varianza también se le conoce como *Cuadrado medio de tratamientos* o simplemente como CM_{TRAT}.

Como ya vimos, el error en el modelo puede ser calculado por diferencia, por tanto,

$$gl_{ERROR} = gl_{TOTAL} - gl_{TRAT} = 31 - 3 = 28$$

$$SC_{ERROR} = SC_{TOTAL} - SC_{TRAT} = 40.94 - 39.41 = 1.53 \text{ y } S^2_{ERROR} = \frac{1.53}{28} = 0.05$$

A esta varianza se le conoce también como *Cuadrado medio del error* o simplemente CM_{ERROR}.

Continuando la construcción de la tabla de Análisis de Varianza tenemos:

TABLA 10.6 Análisis de varianza

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	31	40.94			
TRATAMIENTOS	3	39.41	13.14		
ERROR	28	1.53	0.05		

En este punto ya estamos en condiciones de decidir el rechazo o no de la hipótesis nula planteada, y para ello, debemos proceder a determinar el valor de la llamada *F calculada*, la cual se obtiene dividiendo el cuadrado medio de tratamientos (varianza debida a tratamientos) entre el cuadrado medio del error (varianza debida al error), es decir:

$$F = \frac{CM_{TRAT.}}{CM_{ERROR}} = \frac{13.14}{0.05} = 262.8$$

Con relación a este resultado podemos establecer que:

- Si existe una diferencia significativa entre los tratamientos, es razonable pensar que el valor de la varianza debida a esta componente debe ser grande al igual que la F calculada.
- Por el contrario, si esta diferencia no existe entonces el valor de dicha varianza debe ser pequeño, y en consecuencia, la F calculada también lo será.
- En el capítulo anterior vimos que el cociente F sigue una distribución *F* de Fisher, y por tanto, una regla de decisión para el rechazo o no de la hipótesis nula viene dada de la siguiente manera:

Rechazar H_0 si $F > F_{\alpha}\left(GI_{TRAT.}, GI_{Error}\right)$

No rechazar H_0 si $F \leq F_{\alpha}\left(GL_{TRAT.}, GL_{Error}\right)$

En la **TABLA T.3** del Anexo A podemos encontrar los valores críticos de la distribución ***F*** de Fisher.

Estos valores críticos para 3 grados de libertad en la fuente de variación debida a tratamientos y 28 grados de libertad en la fuente de variación debida al error son:

$\boldsymbol{F_{5\%}}$ **(3,28) = 2.95**

$\boldsymbol{F_{1\%}}$ **(3,28) = 4.57**

$\boldsymbol{F_{0.1\%}}$ **(3,28) = 7.19**

Y como 262.8 > 7.19, rechazamos la hipótesis nula con un nivel de significación del 0.1% y por tanto *no todas las medias poblacionales son iguales*.

Concluyendo la construcción de la tabla de Análisis de Varianza tenemos:

TABLA 10.7 Análisis de varianza

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	31	40.94			
TRATAMIENTOS	3	39.41	13.14	262.8	P<0.001
ERROR	28	1.53	0.05		

Observe que en la columna *SIGN. (Significación)* de la tabla del Análisis de Varianza hemos indicado el rechazo de la hipótesis nula con un nivel de significación del 0.1%, escribiendo P<0.001.

Una manera equivalente de hacerlo es escribiendo ***. Más adelante nos referiremos a este aspecto.

Varios casos podían haberse presentado al hacer la comparación de la F calculada con el percentil de la *F* de Fisher. A continuación procedemos a describir cuatro situaciones que se pueden presentar en la práctica.

1. La F calculada es menor o igual que la *F* de Fisher al 5% (0.05).

   En esta situación concluimos no rechazar la hipótesis nula de la igualdad de las medias de tratamientos, y lo indicamos escribiendo *NS (no significativo)* en la columna *SIGN.* correspondiente a la fila *Tratamientos* de la tabla de análisis de varianza.

2. La F calculada es mayor que la *F* de Fisher al 5% (0.05) pero menor o igual que la *F de Fisher* al 1% (0.01).

   En este caso se concluye rechazar la hipótesis nula al 5% (0.05) y se escribe p<0.05 en la columna *SIGN.*

3. La F calculada es mayor que la *F* de Fisher al 1% (0.01) pero menor o igual que la *F de Fisher* al 0.1% (0.001).

   Se rechaza la hipótesis nula al 1% (0.01) y se escribe p<0.01 en la columna *SIGN.*

4. La F calculada es mayor que la *F* de Fisher al 0.1% (0.001).

   Se rechaza la hipótesis nula al 0.1% (0.001) y se escribe p<0.001 en la columna *SIGN.*

A continuación ejemplificamos numéricamente los cuatro casos estudiados anteriormente con supuestas tablas de análisis de varianza.

La tabla 10.8, 10.9, 10.10 y 10.11 muestran los cuatro casos posibles de diferencias entre tratamientos.

TABLA 10.8 Los tratamientos no difieren significativamente

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	31	40.94			
TRATAMIENTOS	3	2.91	0.97	0.71	NS
ERROR	28	38.03	1.36		

El percentil $F_{0.05}(3,28) = 2.95$. Los tratamientos no difieren significativamente ya que el valor 0.71 < 2.95.

TABLA 10.9 Los tratamientos difieren significativamente al 5%

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	31	40.94			
TRATAMIENTOS	3	10.45	3.48	3.19	P < 0.05
ERROR	28	30.49	1.09		

El percentil $F_{0.01}(3,28) = 4.57$. Los tratamientos solo difieren significativamente al 5% por cuanto 3.19 > 2.95, pero no es mayor que 4.57.

TABLA 10.10 Los tratamientos difieren significativamente al 1%

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	31	40.94			
TRATAMIENTOS	3	17.45	5.82	6.93	P < 0.01
ERROR	28	23.49	0.84		

El percentil $F_{0.001}(3,28) = 7.19$. Los tratamientos solo difieren significativamente al 1% por cuanto el valor 6.93 es mayor que 4.57 pero no es mayor que 7.19.

TABLA 10.11 Los tratamientos difieren significativamente al 0.1%

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	31	40.94			
TRATAMIENTOS	3	19.58	6.53	8.59	P < 0.001
ERROR	28	21.36	0.76		

Los tratamientos difieren significativamente al 0.1% por cuanto 8.59 > 7.19.

En el ejemplo que estamos desarrollando, hemos llegado a la conclusión que la hipótesis nula es falsa, y en consecuencia la hemos rechazado, lo cual implica que las cuatro medias poblacionales de los tratamientos no son iguales.

Pero este rechazo de la hipótesis nula no da información acerca de cuál o cuáles de las medias de tratamientos no son iguales entre sí, y por tanto, han sido las res-

ponsables del rechazo de la hipótesis. En resumen, resulta necesario comparar cada una de las medias de tratamiento con todas las demás, con el objetivo de determinar cuáles fueron las causantes del rechazo de la hipótesis nula.

A este tipo de comparaciones se les suele conocer como *prueba de comparación múltiple*.

A continuación estudiaremos dos de las pruebas de comparación múltiple más populares, aunque solo propondremos una de ellas.

10.4 Pruebas de Comparación Múltiple.

10.4.1 Comparación múltiple por t de Student.

Supongamos que en el ejemplo que estamos desarrollando deseamos establecer si la diferencia entre la media del tratamiento A y el tratamiento B es o no significativamente diferente de cero.

En el Capítulo 6 estudiamos que la distribución muestral de la diferencia entre dos medias muestrales, calculadas a partir de muestras aleatorias independientes de tamaño n_1 y n_2 extraídas de dos poblaciones distribuidas normalmente con varianza homogénea σ^2 conocida, está distribuida normalmente con media $\mu_1 - \mu_2$ y con varianza $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$ y que en el caso en que la varianza poblacional es desconocida, entonces la expresión:

$$\frac{\overline{T}_A - \overline{T}_B}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

sigue una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad, donde $\overline{T}_A$ es la media muestral del tratamiento A, $\overline{T}_B$ es la media muestral del tratamiento B y S^2 puede ser estimada mediante el cuadrado medio del error del análisis de varianza.

Desarrollemos la prueba de comparación múltiple por t de Student con los datos de nuestro ejemplo usando un nivel de significación del 5%.

$$\overline{T}_A = \frac{116.3}{8} = 14.54$$

$$\overline{T}_B = \frac{132.5}{8} = 16.56$$

$$\overline{T}_C = \frac{130.6}{8} = 16.33$$

$$\overline{T}_D = \frac{141}{8} = 17.63$$

$S^2 = CM_{ERROR} = 0.05$ $\quad n_1 = n_2 = 8$ $\quad t_{5\%}^{(14)} = 2.145$

1. Tratamiento A vs Tratamiento B

$$t = \left| \frac{14.54 - 16.56}{\sqrt{0.05\left(\frac{1}{8} + \frac{1}{8}\right)}} \right| = \left| \frac{-2.02}{0.11} \right| = 18.36$$

Como 18.36 > 2.145, ambos tratamientos difieren significativamente con un nivel de significación del 5%.

2. Tratamiento A vs Tratamiento C

$$t = \left| \frac{14.54 - 16.33}{\sqrt{0.05\left(\frac{1}{8} + \frac{1}{8}\right)}} \right| = \left| \frac{-1.79}{0.11} \right| = 16.27$$

Como 16.27 > 2.145, ambos tratamientos difieren significativamente con un nivel de significación del 5%.

3. Tratamiento A vs Tratamiento D

$$t = \left| \frac{14.54 - 17.63}{\sqrt{0.05\left(\frac{1}{8} + \frac{1}{8}\right)}} \right| = \left| \frac{-3.09}{0.11} \right| = 28.09$$

Como 28.09 > 2.145, ambos tratamientos difieren significativamente con un nivel de significación del 5%.

4. Tratamiento B vs Tratamiento C

$$t = \left| \frac{16.56 - 16.33}{\sqrt{0.05\left(\frac{1}{8} + \frac{1}{8}\right)}} \right| = \left| \frac{0.23}{0.11} \right| = 2.09$$

Como 2.09 < 2.145, ambos tratamientos no difieren significativamente con un nivel de significación del 5%.

5. Tratamiento B vs Tratamiento D

$$t = \left| \frac{16.56 - 17.63}{\sqrt{0.05\left(\frac{1}{8} + \frac{1}{8}\right)}} \right| = \left| \frac{-1.07}{0.11} \right| = 9.73$$

Como 9.73 > 2.145, ambos tratamientos difieren significativamente con un nivel de significación del 5%.

6. Tratamiento C vs Tratamiento D

$$t = \left| \frac{16.32 - 17.63}{\sqrt{0.05\left(\frac{1}{8} + \frac{1}{8}\right)}} \right| = \left| \frac{-1.31}{0.11} \right| = 11.91$$

Como 11.91 > 2.145, ambos tratamientos difieren significativamente con un nivel de significación del 5%.

Los resultados que hemos obtenido con las seis comparaciones realizadas para los diferentes pares de medias deben ser resumidos de alguna manera. Una manera de hacerlo es colocando un mismo súper índice a las medias que no difieren entre sí, y diferentes a las que sí difieren. Un procedimiento que nos permite con facilidad realizar lo antes expuesto consiste en ordenar las medias de mayor a menor (o de menor a mayor), y unir con una misma línea las medias que no difieren significativamente y colocar el mismo súper índice a las medias unidas por la misma línea y diferentes al resto, es decir:

17.63 **16.56** **16.33** **14.54**

A	B	C	D
14.54[b]	**16.56[a]**	**16.33[a]**	**17.63[c]**

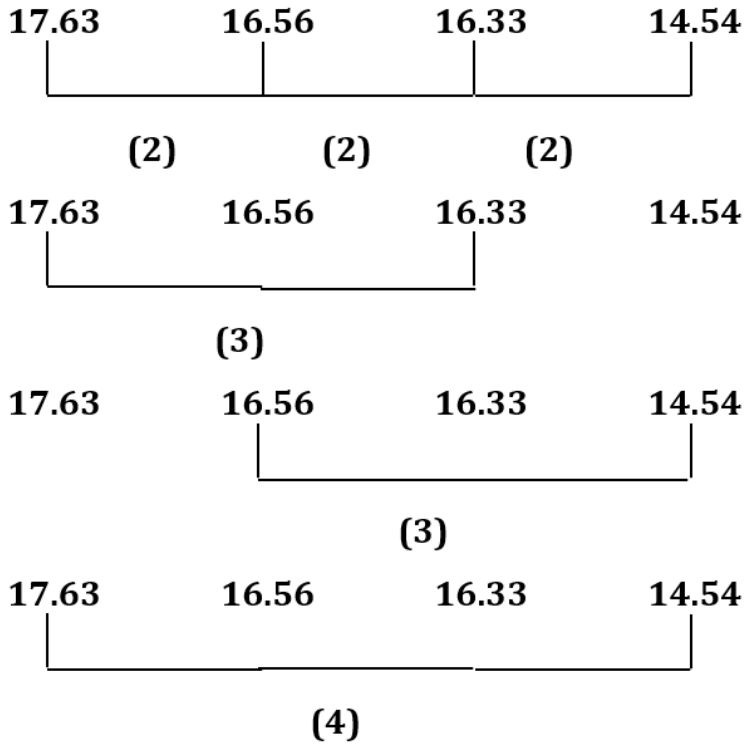
10.4.2 Prueba de rango múltiple de Duncan.

Algunos autores, entre ellos el de este libro, prefieren utilizar para la comparación entre medias de tratamientos un procedimiento desarrollado por *Duncan (1955)* el cual consiste en esencia en establecer un arreglo de las medias de los tratamientos

por orden de magnitud, y utilizar para cada comparación entre dos medias un percentil que depende de los grados de libertad del error en la tabla del Análisis de Varianza y del *rango* entre ambas medias.

La razón de preferir *esta prueba de comparación múltiple* en lugar de la t de Student consiste en que con esta última la probabilidad de que rechacemos la igualdad de dos medias cuando estas realmente son iguales, crece según aumenta el número de comparaciones realizadas.

Precisemos con más detalle lo que se entiende por *rango* entre dos medias utilizando el *arreglo* de mayor a menor de las medias de tratamientos utilizado en el numeral anterior:



Las comparaciones entre las siguientes medias tienen los rangos que se muestran a continuación:

Rango 2

17.63 vs 16.56

16.56 vs 16.33

16.33 vs 14.54

pues entre ambas medias en el arreglo existen 2 elementos.

Rango 3

17.63 vs 16.33

16.56 vs 14.54

pues entre ambas medias en el arreglo existen 3 elementos.

Rango 4

17.63 vs 14.54

pues entre ambas medias en el arreglo existen 4 elementos.

La prueba de Duncan para comparar dos medias consiste en esencia en obtener la cantidad:

$$D = \frac{\overline{T}_i - \overline{T}_j}{\sqrt{\frac{CM_{ERROR}}{m}}}$$

donde m es el número de observaciones por tratamiento.

Diremos entonces que las dos medias comparadas $\overline{T}_i$ y $\overline{T}_j$ difieren significativamente si el valor D resulta ser mayor que el percentil de Duncan correspondiente al rango entre las medias comparadas y para los grados de libertad del error de la tabla del Análisis de Varianza, indicando este resultado mediante un *****. En caso contrario, las medias no difieren entre si y lo indicaremos escribiendo **NS**. Los valores críticos para la prueba de Duncan pueden ser encontrados en la **TABLA T.4** del Anexo A.

En este libro, al igual que lo hicimos anteriormente para la comparación múltiple a través de la t de Student, solo usaremos percentiles de Duncan para un nivel de significación del 5%, con el objetivo de evitar de esta manera que la zona de rechazo de esta prueba supere la de la prueba F, con lo cual se obtendrían resultados contradictorios.

Para desarrollar la prueba de Duncan debemos establecer para una mayor facilidad en los cálculos, una tabla de doble entrada donde la primera columna sea el arreglo de las medias de tratamientos por orden creciente de magnitud sin necesidad de incluir la mayor de ellas, y la primera fila el arreglo por orden decreciente de magnitud de las medias de tratamientos sin necesidad de incluir la menor.

	17.63	**16.56**	**16.33**
14.54	**4**	**3**	**2**
16.33	**3**	**2**	
16.56	**2**		

Establecer la anterior tabla de doble entrada, permite de una forma ágil determinar el rango a utilizar en la comparación de dos medias cualesquiera, ya que las que se encuentran en la misma diagonal (señaladas con el mismo número) utilizan el mismo rango.

De arriba hacia abajo la comparación en la primera diagonal utiliza rango 4, las dos de la segunda diagonal rango 3 y las tres de la última diagonal rango 2.

Resulta también conveniente para los cálculos, incluir en la tabla de doble en-

trada una primera fila con los percentiles de Duncan correspondientes.

Los grados de libertad del error en la tabla del Análisis de Varianza son 28, por tanto los percentiles de Duncan para estos grados de libertad y los correspondientes rangos son:

Rango 4: **3.13**
Rango 3: **3.04**
Rango 2: **2.90**

quedando entonces la tabla de doble entrada como sigue:

3.13	**3.04**	**2.90**
17.63	**16.56**	**16.33**
14.54		
16.33		
16.56		

$$ET(\bar{T}) = \sqrt{\frac{CM_{ERROR}}{m}} = \sqrt{\frac{0.05}{8}} = 0.08$$

a) 17.63 vs 14.54

$$D = \frac{17.63 - 14.54}{0.08} = \frac{3.09}{0.08} = 38.62$$

38.65 > 3.13, por tanto ambas medias difieren y lo expresamos:

	3.13	**3.04**	**2.90**
	17.63	**16.56**	**16.33**
14.54	*		
16.33			
16.56			

b) 17.63 vs 16.33

$$D = \frac{17.63 - 16.33}{0.08} = \frac{1.30}{0.08} = 16.25$$

16.25 > 3.04, por tanto ambas medias difieren significativamente, y lo expresamos:

	3.13	**3.04**	**2.90**
	17.63	**16.56**	**16.33**
14.54	*		
16.33	*		
16.56			

c) 16.56 vs 14.54

$$D = \frac{16.56 - 14.54}{0.08} = \frac{2.02}{0.08} = 25.25$$

25.25 > 3.04, por tanto ambas medias difieren significativamente, y lo expresamos:

	3.13	**3.04**	**2.90**
	17.63	**16.56**	**16.33**
14.54	*	*	
16.33	*		
16.56			

d) 17.63 vs 16.56

$$D = \frac{17.63 - 16.56}{0.08} = \frac{1.07}{0.08} = 13.38$$

13.38 > 2.90, por tanto ambas medias difieren significativamente, y lo expresamos:

	3.13	**3.04**	**2.90**
	17.63	**16.56**	**16.33**
14.54	*	*	
16.33	*		
16.56	*		

e) 16.56 vs 16.33

$$D = \frac{16.56 - 16.33}{0.08} = \frac{0.23}{0.08} = 2.88$$

2.88 < 2.90, por tanto ambas medias no difieren, y lo expresamos:

	3.13	**3.04**	**2.90**
	17.63	**16.56**	**16.33**
14.54	*	*	
16.33	*	**NS**	
16.56	*		

f) 16.33 vs 14.54

$$D = \frac{16.33 - 14.54}{0.08} = \frac{1.79}{0.08} = 22.38$$

22.38 > 2.90, por tanto ambas medias difieren significativamente, y lo expresamos:

	3.13	3.04	2.90
	17.63	16.56	16.33
14.54	*	*	*
16.33	*	NS	
16.56	*		

Los resultados finales del ejercicio que hemos desarrollado se resumen en la tabla 10.12.

TABLA 10.12 Resumen de los resultados de la investigación

TRATAMIENTOS	A	B	C	D	SIGN.	E.E.
VENTAS	14.54^{b}	16.56^{a}	16.33^{a}	17.63^{c}	p<0.001	±0.08

En el procedimiento estadístico que acabamos de concluir utilizamos 32 datos provenientes de 4 tratamientos y 8 repeticiones. Sin embargo, ocurre con relativa frecuencia que al medir la respuesta de la variable que estamos estudiando en una investigación, por alguna razón, este valor no puede ser obtenido y en consecuencia se nos presenta la situación que debemos trabajar con uno o más *valores faltantes***.**

En casos como el expuesto anteriormente, el procedimiento para obtener la tabla que resume los resultados de la investigación presenta algunos cambios que procedemos a exponer de inmediato. Para ejemplificar lo antes expuesto supongamos que en el ejemplo que hemos estado desarrollando no fue posible obtener los datos correspondientes a la sexta repetición del tratamiento A y tercera del tratamiento C.

Además, con la finalidad de mostrar una vez más el método de cálculo para este tipo de diseño, eliminaremos de la investigación el tratamiento D (aceite de soya) y las repeticiones 4 y 5, es decir, que los datos quedan como se muestra en la tabla 10.13.

TABLA 10.13 Procesamiento estadístico con datos faltantes

							TOTAL	MEDIAS
A	14.9	14.2	14.3	-	14.3	14.6	72.3	14.46
B	16.7	16.8	16.4	16.9	16.2	16.7	99.7	16.62
C	16.2	16.5	-	16.1	16.4	16.5	81.7	16.34
							253.7	

El factor de corrección y la suma de cuadrados corregida total se calculan de la forma acostumbrada:

$$FC = \frac{(253.7)^2}{16} = 4022.73$$

Observe que el valor del denominador del FC es 16, ya que es el número de ob-

servaciones existentes.

Suma de cuadrados corregida total:

$$SC_{TOTAL} = (14.9)^2 + (14.2)^2 + \ldots\ldots\ldots\ldots + (16.5)^2 - FC$$

$$SC_{TOTAL} = 4037.93 - 4022.73 = 15.20$$

La suma de cuadrados corregida de tratamientos debe ser calculada tomando en cuenta el número real de observaciones en cada tratamiento, es decir:

Suma de cuadrados corregida de tratamientos:

$$SC_{TRAT.} = \frac{(72.3)^2}{5} + \frac{(99.7)^2}{6} + \frac{(81.7)^2}{5} - FC$$

$$SC_{TRAT.} = 4037.12 - 4022.73 = 14.39$$

Suma de cuadrados corregida del error:

$$SC_{ERROR} = 15.20 - 14.39 = 0.81$$

Observe en la tabla 10.4 del Análisis de Varianza que los grados de libertad para la fuente de variación total es igual a 15, es decir, uno menos la cantidad real de observaciones en la investigación. Por el contrario, observe que la fuente de variación debida a los tratamientos tiene 2 grados de libertad (uno menos la cantidad de tratamientos), es decir, que no se ven afectados por el número real de observaciones de la investigación.

No será difícil para el lector comprobar que las medias de tratamientos difieren significativamente al 0.1%.

TABLA 10.14 Análisis de varianza

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	15	15.20			
TRATAMIENTOS	2	14.39	7.20	120	P < 0.001
ERROR	13	0.81	0.06		

El error estándar de las medias de tratamientos viene dado por:

Para los tratamientos A y C

$$EE(\overline{T}) = \sqrt{\frac{CM_{ERROR}}{r}} = \sqrt{\frac{0.06}{5}} = \pm 0.11$$

Para el tratamiento B

$$EE(\overline{T}) = \sqrt{\frac{CM_{ERROR}}{r}} = \sqrt{\frac{0.06}{6}} = \pm 0.10$$

Para desarrollar la prueba de Duncan correspondiente a este ejemplo, debemos ajustar la forma en que obtenemos el error estándar entre tratamientos debido a que el número de observaciones entre las medias comparadas no siempre es el mismo. Es decir, la expresión para calcular el valor D en la prueba de Duncan quedaría como sigue:

$$D = \frac{\overline{T}_i - \overline{T}_j}{\sqrt{\frac{2CM_{ERROR}}{r_i + r_j}}} \quad cuando\, r_i \neq r_j$$

$$D = \frac{\overline{T}_i - \overline{T}_j}{\sqrt{\frac{2CM_{ERROR}}{r + r}}} = \frac{\overline{T}_i - \overline{T}_j}{\sqrt{\frac{CM_{ERROR}}{r}}} \quad cuando\, r_i = r_j = r$$

De esta forma, cuando comparamos las medias de los tratamientos con 5 observaciones (14.46 y 16.34) con la del tratamiento que tiene 6 (16.62), el valor de D es:

$$D = \frac{16.62 - 14.46}{\sqrt{\frac{2(0.06)}{5+6}}} = \frac{2.16}{0.10} = 21.6$$

$$D = \frac{16.62 - 16.34}{\sqrt{\frac{2(0.06)}{5+6}}} = \frac{0.28}{0.10} = 2.8$$

Cuando comparamos las medias de los dos tratamientos que tienen 5 observaciones (14.46 y 16.34):

$$D = \frac{16.34 - 14.46}{\sqrt{\frac{(0.06)}{5}}} = \frac{1.88}{0.11} = 17.09$$

Con estas indicaciones no le será difícil al lector completar la prueba de Duncan correspondiente.

La tabla 10.15 resume la investigación desarrollada.

TABLA 10.15 Resumen de los resultados de la investigación

TRATAMIENTOS	A	B	C	SIGN.
VENTAS	14.46[a]	16.62[b]	16.34[b]	p<0.001
	0.11±	0.10±	0.11±	

10.5 Análisis de Varianza de datos provenientes de un diseño en Bloques al Azar.

Para el desarrollo del presente numeral utilizaremos un ejemplo expuesto por el autor en su libro Diseño y Análisis de Experimentos Agropecuarios (2007), el cual consiste en la aplicación de tres dosis distintas de potasa (K_2O) y un Testigo (sin aplicación de potasa) sobre la producción de un determinado cultivo en una empresa agropecuaria. Sean estas dosis las siguientes:

A: 0 Kg de K_2O /Ha. (Testigo)

B: 40 Kg de K_2O /Ha.

C: 80 Kg de K_2O /Ha.

D: 120 Kg de K_2O /Ha.

Para el desarrollo de la investigación se utilizaron ocho réplicas por tratamiento, es decir, que a cada grupo de ocho parcelas experimentales sembradas con el cultivo se le aplicó una dosis de potasa diferente.

Consideremos que al seleccionar el área del terreno escogido para la investigación se detectaron diferencias en la fertilidad del suelo en sentido horizontal y de izquierda a derecha, lo que determinó una marcada heterogeneidad entre las unidades experimentales usadas en la investigación.

Supongamos, solo de manera provisional, que utilizamos un diseño Completamente al Azar, y en consecuencia, asignamos los tratamientos de forma totalmente aleatoria a las parcelas experimentales, quedando la aleatorización de la forma que se muestra en la tabla 10.16.

TABLA 10.16 Aleatorización completamente al azar de cuatro tratamientos

Mayor fertilidad → Menor fertilidad

A	C	C	C	B	B	D	D
A	A	B	B	C	D	C	B
B	A	A	B	B	D	D	D
A	A	A	C	C	C	D	D

Como se puede apreciar en la tabla anterior, el hecho de haber asignado **completamente al azar** los tratamientos a las parcelas experimentales, provocó que el tratamiento D correspondiente a 120 Kg de K_2O /Ha fuese aplicado en un área donde el suelo tenía una menor fertilidad que en la zona donde no se aplicó fertilizante

(Tratamiento A 0 Kg de K_2O/Ha.).

En conclusión, el haber asignado los tratamientos a las parcelas experimentales de forma *completamente aleatoria* determinó un *efecto confundido* de los tratamientos, por cuanto en el diseño utilizado y de manera inconsciente se *favoreció* al tratamiento A y se *perjudicó* al tratamiento D.

La solución al inconveniente expuesto en el párrafo anterior consiste en *agrupar* los tratamientos en *bloques*, de manera tal que dentro de cada bloque el material experimental sea homogéneo y se encuentren representados todos los tratamientos.

La asignación de los tratamientos a las unidades experimentales en cada bloque se realiza de forma aleatoria.

Una posible aleatorización podía ser la que se muestra en la tabla 10.17.

TABLA 10.17 Asignación por bloques de los tratamientos a las unidades

B	C	A	D	D	B	A	C
D	A	B	A	C	A	C	B
A	B	D	B	B	D	D	A
C	D	C	C	A	C	B	D

Cuando procedemos de la manera indicada en el párrafo anterior se dice que hemos utilizado un *diseño en Bloques al Azar*.

Consideremos que los resultados de la investigación fueron los que se muestran en la tabla 10.18.

TABLA 10.18 Producción de las parcelas experimentales

B L O Q U E S							
I	II	III	IV	V	VI	VII	VIII
B	C	A	D	D	B	A	C
3.4	4.8	2.7	4.3	4.1	3.6	2.5	4.5
D	A	B	A	C	A	C	B
4.6	2.4	3.4	2.9	4.4	2.5	4.3	3.3
A	B	D	B	B	D	D	A
2.4	3.6	4.4	3.3	3.7	4.7	4.3	2.2
C	D	C	C	A	C	B	D
4.5	4.9	4.4	4.1	2.7	4.1	3.3	4.5
14.9	15.7	14.9	14.6	14.9	14.9	14.4	14.5

donde la última fila son los totales por cada uno de los bloques.

Desarrollemos los cálculos necesarios para obtener la tabla de Análisis de Varianza correspondiente:

$$SC_{TOTAL} = \sum y_{ij} - \frac{\left(\sum y_{ij}\right)^2}{n}$$

$$SC_{TOTAL} = 465.28 - \frac{(118.8)^2}{32}$$

$$SC_{TOTAL} = 21.24$$

Los totales por tratamiento son:

T_A = 2.4+2.4+2.7+2.9+2.7+2.5+2.5+2.2= 20.3

T_B = 3.4+3.6+3.4+3.3+3.7+3.6+3.3+3.3= 27.6

T_C = 4.5+4.8+4.4+4.1+4.4+4.1+4.3+4.5 = 35.1

T_D = 4.6+4.9+4.4+4.3+4.1+4.7+4.3+4.5= 35.8

$$SC_{TRAT.} = \sum \frac{T_i^2}{b} - \frac{\left(\sum y_{ij}\right)^2}{n}$$

donde b es la cantidad de bloques.

$$SC_{TRAT.} = \frac{(20.3)^2 + (27.6)^2 + (35.1)^2 + (35.8)^2}{8} - \frac{(118.8)^2}{32}$$

$$SC_{TRAT.} = 460.94 - 441.04 = 19.90$$

Si B_j es el total del bloque j-ésimo, entonces:

$$SC_{BLOQUES} = (14.9)^2 + (15.7)^2 + \ldots\ldots\ldots + (14.4)^2 + (14.5)^2 - 441.04$$

$$SC_{BLOQUES} = 441.32 - 441.04 = 0.28$$

$$SC_{ERROR} = SC_{TOTAL} - SC_{TRAT} - SC_{BLOQUES} = 21.24 - 19.90 - 0.28 = 1.06$$

$$CM_{TRAT.} = \frac{19.90}{3} = 6.63$$

$$CM_{ERROR} = \frac{1.06}{21} = 0.05$$

$F = \frac{6.63}{0.05} = 132.6$ la cual es significativa al 0.1%.

Los resultados obtenidos se aprecian en la tabla 10.19.

TABLA 10.19 Análisis de varianza

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	31	21.24			
TRATAMIENTOS	3	19.90	6.63	132.6	P < 0.001
BLOQUES	7	0.28			
ERROR	21	1.06	0.05		

Procedamos a desarrollar la prueba de comparación múltiple de Duncan para lo cual debemos calcular el error estándar de las medias de tratamientos y obtener los percentiles de Duncan para 21 grados de libertad haciendo uso de la **TABLA T.4** que se encuentra en el Anexo A.

$$ET(\bar{T}) = \sqrt{\frac{CM_{ERROR}}{b}} = \sqrt{\frac{0.05}{8}} = 0.08$$

En la tabla de valores críticos de Duncan no aparecen los correspondientes a 21 grados de libertad lo que implica que debemos interpolar:

- Para 20 grados de libertad y un nivel de significación del 5% los percentiles de Duncan son:

Rango 4 = 3.18
Rango 3 = 3.10
Rango 2 = 2.95

- Para 22 grados de libertad y un nivel de significación del 5% los percentiles de Duncan son:

Rango 4 = 3.17
Rango 3 = 3.08
Rango 2 = 2.93

Como 21 grados de libertad está exactamente entre 20 y 22 podemos entonces fácilmente determinar que los percentiles requeridos son:

$$Rango\ 4 = \frac{3.18 + 3.17}{2} = 3.175$$

$$Rango\ 3 = \frac{3.10 + 3.08}{2} = 3.09$$

$$Rango\ 2 = \frac{2.95 + 2.93}{2} = 2.94$$

3.175 **3.09** **2.94**

	4.48	**4.39**	**3.45**
2.54			
3.45			
4.39			

$$D = \frac{4.48 - 2.54}{0.08} = \frac{1.94}{0.08} = 24.25$$

24.25 > 3.175, por tanto ambas medias difieren significativamente, y lo expresamos:

	3.175	**3.09**	**2.94**
	4.48	**4.39**	**3.45**
2.54	*		
3.45			
4.39			

$$D = \frac{4.48 - 3.45}{0.08} = \frac{1.03}{0.08} = 12.88$$

12.88 > 3.09, por tanto ambas medias difieren significativamente, y lo expresamos:

	3.175	**3.09**	**2.94**
	4.48	**4.39**	**3.45**
2.54	*		
3.45	*		
4.39			

$$D = \frac{4.39 - 3.45}{0.08} = \frac{0.94}{0.08} = 11.75$$

11.75 > 3.09, por tanto ambas medias difieren significativamente, y lo expresamos:

	3.175	**3.09**	**2.94**
	4.48	**4.39**	**3.45**
2.54	*	*	
3.45	*		
4.39			

d) 4.48 vs 4.39

$$D = \frac{4.48 - 4.39}{0.08} = \frac{0.09}{0.08} = 1.13$$

1.13 < 2.94, por tanto ambas medias no difieren, y lo expresamos:

	3.175	**3.09**	**2.94**
	4.48	**4.39**	**3.45**
2.54	*	*	
3.45	*		
4.39	**NS**		

e) 4.39 vs 3.45

$$D = \frac{4.39 - 3.45}{0.08} = \frac{0.94}{0.08} = 11.75$$

11.75 > 2.94, por tanto ambas medias difieren significativamente, y lo expresamos:

	3.175	**3.09**	**2.94**
	4.48	**4.39**	**3.45**
2.54	*	*	
3.45	*	*	
4.39	**NS**		

f) 3.45 vs 2.54

$$D = \frac{3.45 - 2.54}{0.08} = \frac{0.91}{0.08} = 11.38$$

11.38 > 2.94, por tanto ambas medias difieren significativamente, y lo expresamos:

	3.175	**3.09**	**2.94**
	4.48	**4.39**	**3.45**
2.54	*	*	*
3.45	*	*	
4.39	**NS**		

Los resultados finales se resumen en la tabla 10.20.

TABLA 10.20 Resumen de los resultados de la investigación

TRATAMIENTOS	A	B	C	D	SIGN.	E.E.
RENDIMIENTO	2.54^{a}	3.45^{b}	4.39^{c}	4.48^{c}	p<0.001	±0.08

A diferencia del diseño Completamente al Azar, en el diseño en Bloques al Azar no es posible desarrollar el Análisis de Varianza si se tiene un número desigual de observaciones por tratamiento como ocurre cuando hay *valores faltantes*. En casos como éste, se hace necesario *establecer* el valor numérico del dato faltante siguiendo un procedimiento que explicaremos en los siguientes párrafos.

Consideremos una investigación diseñada en Bloques al Azar con t tratamientos y b bloques. Denotemos por X al valor faltante y calculemos las sumas de cuadrados requeridas para el análisis de varianza.

Factor de corrección:

$$FC = \frac{(X+G)^2}{tb}$$

donde G es la suma total de todos los resultados experimentales existentes.

Suma de cuadrados total:

$$SC_{TOTAL} = X^2 + K1 - \frac{(X+G)^2}{tb}$$

donde K1 es la suma de los cuadrados de todos los resultados experimentales.

Suma de cuadrados de bloques (réplicas):

$$SC_{BLOQUES} = \frac{(X+B)^2 + K2}{t} - \frac{(X+G)^2}{tb}$$

donde B es la suma de los resultados experimentales del bloque donde se encuentra el valor faltante, y K2 es la suma de cuadrados de los totales de los bloques donde no se encuentra el valor faltante.

Suma de cuadrados de tratamientos:

$$SC_{TRAT.} = \frac{(X+T)^2 + K3}{b} - \frac{(X+G)^2}{tb}$$

donde T representa la suma de los resultados experimentales del tratamiento donde se encuentra el valor faltante, y K3 es la suma de cuadrados de los totales de los tratamientos donde no se encuentra el valor faltante.

La suma de cuadrados corregida del error será entonces:

$$SC_{ERROR} = X^2 + K1 - \frac{(X+G)^2}{tb} - \left[\frac{(X+B)^2 + K2}{t} - \frac{(X+G)^2}{tb}\right]$$

$$- \left[\frac{(X+T)^2 + K3}{b} - \frac{(X+G)^2}{tb}\right]$$

$$SC_{ERROR} = X^2 + K1 - \frac{(X+B)^2 + K2}{t} - \frac{(X+T)^2 + K3}{b} + \frac{(X+G)^2}{tb}$$

El valor de X que *asumiremos*, será aquel que hace mínima la suma de cuadrados del error anteriormente calculada:

$$\frac{d(SC_{ERROR})}{dX} = 2X - \frac{2(X+B)}{t} - \frac{2(X+T)}{b} + \frac{2(X+G)}{tb} = 0$$

$$\frac{2(tbX - bX - bB - tX - tT + X + G)}{tb} = 0$$

$$X(tb - b - t + 1) = bB + tT - G$$

y como tb – b – t + 1 = (b-1) (t-1) entonces:

$X = \frac{bB + tT - G}{(b-1)(t-1)}$, expresión que nos permite calcular un valor faltante en un diseño en Bloques al Azar.

A modo de ejemplo consideremos un estudio sobre el tiempo empleado para envasar un determinado producto utilizando cinco métodos diferentes, el cual fue desarrollado según un diseño en Bloques al Azar con cuatro réplicas. Supongamos que el valor numérico del tiempo empleado en el tercer bloque del tratamiento C no pudo ser medido. La aleatorización y los resultados experimentales se muestran en la tabla 10.21 y los tiempos empleados ordenados por tratamientos y bloques en la tabla 10.22.

TABLA 10.21 Aleatorización y resultados experimentales

BLOQUES			
I	**II**	**III**	**IV**
B 22.6	A 26.2	E 18.8	B 18.4
A 37.4	B 17.5	A 38.3	A 28.4
D 39.2	E 28.9	D 38.0	D 27.8
C 24.1	D 28.6	B 23.1	C 19.0
E 28.4	C 18.2	-	E 27.9

TABLA 10.22 Tiempos empleados ordenados por tratamientos y bloques

	BLOQUES				
	I	**II**	**III**	**IV**	**TOTAL**
A	37.4	26.2	38.3	28.4	130.3
B	22.6	17.5	23.1	18.4	81.6
C	24.1	18.2	-	19	61.3
D	39.2	28.6	38	27.8	133.6
E	28.4	28.9	18.8	27.9	104
TOTAL	151.7	119.4	118.2	121.5	510.8

Observe que B = 118.2, T = 61.3 **y** G = 510.8

El valor del dato faltante es:

$$X = \frac{(4)(118.2)+(5)(61.3)-510.8}{(4-1)(5-1)} = \frac{268.5}{12} = 22.4$$

Es necesario dejar totalmente claro que 22.4 es el valor que al ser incluido en lugar del valor faltante hace mínima la suma de cuadrados del error, y que de ninguna manera debe ser interpretado como una *estimación* del valor real.

Los datos para el cálculo del análisis de varianza se muestran en la tabla 10.23.

El lector podrá comprobar que la tabla del análisis de varianza es la que se muestra en la tabla 10.24.

Observe que los grados de libertad del Total son 18 debido a que solo existen 19 resultados experimentales independientes, por cuanto el valor faltante no lo es.

TABLA 10.23 Datos para el análisis de varianza

	BLOQUES				
	I	**II**	**III**	**IV**	**TOTAL**
A	37.4	26.2	38.3	28.4	130.3
B	22.6	17.5	23.1	18.4	81.6
C	24.1	18.2	22.4	19.0	83.7
D	39.2	28.6	38.0	27.8	133.6
E	28.4	28.9	18.8	27.9	104.0
TOTAL	151.7	119.4	140.6	121.5	533.2

TABLA 10.24 Análisis de varianza

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	18	952.59			
BLOQUES	3	144.86			
TRATAMIENTOS	4	611.71	152.93	8.58	P < 0.01
ERROR	11	196.01	17.82		

El error estándar para las medias de tratamientos sin valor faltante se calcula de la forma usual.

$$EE(\overline{T}) = \sqrt{\frac{CM_{ERROR}}{b}} = \sqrt{\frac{17.82}{4}} = \pm 2.11$$

En el caso del tratamiento al cual pertenece el valor faltante, el error estándar viene dado por la expresión:

$$EE(\overline{T}) = \sqrt{CM_{ERROR}\left[\frac{1}{b} + \frac{t}{b(b-1)(t-1)}\right]}$$

por tanto, para el tratamiento C:

$$EE(\overline{T}) = \sqrt{17.82\left[\frac{1}{4} + \frac{5}{4(4-1)(5-1)}\right]} = \sqrt{17.82\left[\frac{1}{4} + \frac{5}{48}\right]} = \pm 2.51$$

Debido a que en el análisis de varianza realizado se produjo el rechazo de la hipótesis nula de la igualdad entre las medias de tratamientos, resulta necesario proceder a desarrollar la prueba de comparación múltiple de Duncan, la cual con toda seguridad el lector está en condiciones de hacer, al igual que el resumen de los resultados de la investigación que aparece en la tabla 10.25.

TABLA 10.25 Resumen de los resultados de la investigación

TRATAMIENTOS	A	B	C	D	E	SIGN.
VENTAS	32.58^a	20.40^b	20.93^a	33.40^a	26.00^b	P<0.01
	± 2.1	± 2.1	± 2.5	± 2.1	± 2.1	

10.6 Partición de la suma de cuadrados de tratamientos.

Para estudiar este tema haremos referencia al ejercicio desarrollado al inicio del numeral 10.5 el cual consistía en la aplicación de tres dosis distintas de potasa (K_2O) y un Testigo (sin aplicación de potasa) sobre la producción de un determinado cultivo en una empresa agropecuaria.

Estos tratamientos y sus respectivos totales eran los siguientes:

A: 0 Kg de K_2O /Ha. $T_A = 20.3$

B: 40 Kg de K_2O /Ha. $T_B = 27.6$

C: 80 Kg de K_2O /Ha. $T_C = 35.1$

D: 120 Kg de K_2O /Ha. $T_D = 35.8$

Al profesional que desarrolló esta investigación le pudo haber interesado establecer si existían diferencias entre el tratamiento Testigo (0 Kg/Ha.) y el resto de los tratamientos, pues con ello podía llegar a la conclusión si hubo o no una respuesta favorable a la aplicación del fertilizante, o también establecer una posible diferencia entre la dosis más baja (40 Kg/Ha.) y las más altas (80 Kg/Ha. y 120 Kg/Ha.) o entre las dos dosis mayores (80 Kg/Ha. y 120 Kg/Ha.).

Las comparaciones anteriores pueden ser expresadas de la siguiente manera:

BCD vs A

$$\frac{T_B + T_C + T_D}{3} - T_A \quad o \quad 1T_B + 1T_C + 1T_D - 3T_A$$

CD vs B

$$\frac{T_C + T_D}{2} - T_B \qquad o \quad 1T_C + 1T_D - 2T_B$$

D vs C

$$1T_D - 1T_C$$

Para someter a prueba estas comparaciones o cualquier otra, podemos subdividir la suma de cuadrados de tratamientos y sus grados de libertad en un número determinado de componentes, cada uno de ellos correspondiente a una determinada hipótesis, para a través de la prueba F establecer su significación. Las comparaciones o cantidades anteriores se llaman *funciones lineales de las T*. Una condición que debe cumplir cualquier *función lineal de las T* es que la suma de los coeficientes de los totales de tratamientos sea igual a cero.

Cualquier función lineal de la forma:

$$L = \sum l_i T_i$$

es una *comparación entre las T* si se cumple que $\sum l_i = 0$

Por otra parte, si L es una comparación cualquiera entre las T_i entonces la can-

tidad $\frac{L^2}{Q}$ es una componente de la suma de cuadrados de tratamientos con 1 grado de libertad, donde $Q = r\sum l_i^2$ y r el número de repeticiones de cada tratamiento.

Considerando los totales de tratamientos del ejemplo y los aspectos señalados en los párrafos precedentes, tenemos que las suma de cuadrados para cada comparación puede ser calculada como se muestra en la tabla 10.26.

TABLA 10.26 Cálculo de las sumas de cuadrados para cada comparación

	A	B	C	D	L	Q	S.C.
Total	20.3	27.6	35.1	35.8			
BCD vs A	-3	1	1	1	37.6	96	14.73
CD vs B	0	-2	1	1	15.7	48	5.14
D vs C	0	0	-1	1	0.7	16	0.03
						Total	19.90

Las sumas de cuadrados de cada comparación fueron obtenidas de la siguiente manera:

BCD vs A

L = (-3) (20.3) + (1) (27.6) + (1) (35.1) + (1) (35.8) = 37.6

$$Q = 8\left[(-3)^2 + (1)^2 + (1)^2 + (1)^2\right] = 8(12) = 96$$

$$S.C. = \frac{(37.6)^2}{96} = 14.73$$

CD vs B

L = (0) (20.3) + (-2) (27.6) + (1) (35.1) + (1) (35.8) = 15.7

$$Q = 8\left[(0)^2 + (-2)^2 + (1)^2 + (1)^2\right] = 8(6) = 48$$

$$S.C. = \frac{(15.7)^2}{48} = 5.14$$

D vs C

L = (0) (20.3) + (0) (27.6) + (-1) (35.1) + (1) (35.8) = 0.7

$$Q = 8\left[(0)^2 + (0)^2 + (-1)^2 + (1)^2\right] = 8(2) = 16$$

$$S.C. = \frac{(0.7)^2}{16} = 0.03$$

En los cálculos anteriores se pueden observar los siguientes aspectos:

- La suma de cuadrados de tratamientos del análisis de varianza es igual a la suma de las sumas de cuadrados de todas las comparaciones.
- Las sumas de los coeficientes de cada comparación es igual a cero.
- La suma de los productos de los coeficientes de cualquier par de comparaciones es igual a cero.
- En este caso se dice que las comparaciones son *mutuamente ortogonales*, es decir, *ortogonales dos a dos*.

En la tabla 10.27 se muestra el análisis de varianza que se obtuvo al desarrollar el ejercicio al que hemos hecho referencia.

A esta tabla le hemos agregado la descomposición de la suma de cuadrados de tratamientos en sus respectivas componentes.

TABLA 10.27 Análisis de varianza

FUENTES DE VARIACIÓN	G.L	S.C.	C.M.	F	SIGN.
TOTAL	31	21.24			
TRATAMIENTOS	3	19.90	6.63	132.6	P<0.001
BCD vs A	1	14.73	14.73	294.6	P<0.001
CD vs B	1	5.14	5.14	102.8	P<0.001
D vs C	1	0.03	0.03	0.60	NS
BLOQUES	7	0.28			
ERROR	21	1.06	0.05		

Como conclusión, la tabla del análisis de varianza muestra que existe una respuesta altamente significativa a la aplicación de la potasa sobre el rendimiento del cultivo.

Por otra parte no existen diferencias entre las dosis altas del fertilizante, y éstas en promedio, sí difieren de la dosis más baja.

10.7 Arreglos Factoriales.

Consideremos que Nestlé S.A., la compañía agroalimentaria más grande del mundo, está interesada en estudiar el efecto de cuatro tipos de leche y tres formas de envase sobre los volúmenes de venta de este producto.

Si representamos por $\mathbf{T_1, T_2, T_3, y\ T_4}$ los cuatro tipos de leche y por $\mathbf{F_1, F_2\ y\ F_3}$ las tres formas de envase, entonces en la investigación se someterán a prueba un total de 12 tratamientos, los cuales se describen a continuación:

$T_1 F_1$	$T_1 F_2$	$T_1 F_3$
$T_2 F_1$	$T_2 F_2$	$T_2 F_3$
$T_3 F_1$	$T_3 F_2$	$T_3 F_3$
$T_4 F_1$	$T_4 F_2$	$T_1 F_3$

Para el desarrollo de la investigación la compañía decidió utilizar un diseño en Bloques al Azar con 5 réplicas, razón por la cual fue necesario medir los volúmenes de venta del producto en 60 diferentes supermercados. Por otra parte, para cumplir los requisitos de un diseño en bloques al azar, la compañía garantizó que los 12 supermercados que conformaban una réplica tuvieran características homogéneas, y que en caso de existir heterogeneidad entre los mismos, éstas se presentaran entre réplicas.

Supongamos que los volúmenes de venta medidos en cientos de unidades, son los que se muestran en la tabla 10.28.

TABLA 10.28 Volúmenes de venta en cientos de unidades

FACTORES							**TOTAL TRAT.**
TIPO DE LECHE	**FORMA DE ENVASE**	**B L O Q U E S**					
		I	**II**	**III**	**IV**	**V**	
	F_1	58	60	64	61	58	301
T_1	F_2	72	71	73	70	71	357
	F_3	68	67	71	68	69	343
	F_1	70	69	73	70	71	353
T2	F_2	61	63	66	62	60	312
	F_3	66	67	71	68	67	339
	F_1	55	58	62	60	56	291
T3	F_2	63	61	64	59	64	311
	F_3	72	71	73	70	70	356
	F_1	86	87	91	90	85	439
T4	F_2	74	72	75	78	74	373
	F_3	67	65	68	67	66	333
TOTAL RÉPLICAS		812	811	851	823	811	4108

Hagamos los cálculos necesarios para construir la tabla del análisis de varianza:

$$FC = \frac{(4108)^2}{60} = 281261.07$$

$$SC_{TOTAL} = (58)^2 + (60)^2 + (64)^2 + \ldots\ldots\ldots\ldots + (66)^2 - 281261.07$$

$$SC_{TOTAL} = 284884 - 281261.07 = 3622.93$$

$$SC_{TRAT.} = \frac{(301)^2 + (357)^2 + (343)^2 + \ldots\ldots\ldots + (333)^2}{5} - 281261.07$$

$$SC_{TRAT.} = 284690 - 281261.07 = 3428.93$$

$$SC_{BLOQUES.} = \frac{(812)^2 + (811)^2 + (851)^2 + (823)^2 + (811)^2}{12} - 281261.07$$

$$SC_{BLOQUES} = 281359.67 - 281261.07 = 98.6$$

$$SC_{ERROR} = 3622.93 - 3428.93 - 98.6 = 95.4$$

$$GL_{ERROR} = GL_{TOTAL} - GL_{BLOQUES} - GL_{TRAT.} = 59 - 4 - 11 = 44$$

En la **TABLA T.3** del Anexo A no están reportados los percentiles de la *F* de Fisher para 11 grados de libertad de Tratamientos y 44 del Error, por tanto debemos interpolar:

a) Obtengamos los percentiles ***F*** con 11 y 40 grados de libertad:

$$F_{5\%}(10,40) = 2.08 \qquad F_{1\%}(10,40) = 2.80 \qquad F_{0.1\%}(10,40) = 3.87$$

$$F_{5\%}(12,40) = 2.00 \qquad F_{1\%}(12,40) = 2.66 \qquad F_{0.1\%}(12,40) = 3.64$$

$$F_{5\%}(11,40) = \frac{2.08 + 2.00}{2} = 2.04 \qquad F_{1\%}(11,40) = \frac{2.80 + 2.66}{2} = 2.73$$

$$F_{0.1\%}(11,40) = \frac{3.87 + 3.64}{2} = 3.76$$

b) Obtengamos los percentiles ***F*** con 11 y 60 grados de libertad:

$$F_{5\%}(10,60) = 1.99 \qquad F_{1\%}(10,60) = 2.63 \qquad F_{0.1\%}(10,60) = 3.54$$

$$F_{5\%}(12,60) = 1.92 \qquad F_{1\%}(12,60) = 2.50 \qquad F_{0.1\%}(12,60) = 3.32$$

$$F_{5\%}(11,60) = \frac{1.99 + 1.92}{2} = 1.96 \qquad F_{1\%}(11,60) = \frac{2.63 + 2.50}{2} = 2.56$$

$$F_{0.1\%}(11,60) = \frac{3.54 + 3.32}{2} = 3.43$$

c) Obtengamos los percentiles ***F*** con 11 y 44 grados de libertad:

$$F_{5\%}(11,40)=\frac{2.08+2.00}{2}=2.04 \qquad F_{5\%}(11,60)=\frac{1.99+1.92}{2}=1.96$$

A un incremento de 20 unidades (40 a 60) en los grados de libertad del Error le corresponde un decrecimiento de 0.08 unidades (2.04 a 1.96) en el percentil, por tanto, a un incremento de 4 unidades en el Error corresponderá un decrecimiento de X unidades:

20	0.08
4	X

$$X=\frac{(4)(0.08)}{20}=\frac{0.32}{20}=0.02 \text{ y } F_{5\%}(11,44)=2.04-0.02=2.02$$

Procediendo de la misma manera para los dos percentiles restantes tenemos:

$$F_{1\%}(11,40)=\frac{2.80+2.66}{2}=2.73 \qquad F_{1\%}(11,60)=\frac{2.63+2.50}{2}=2.56$$

20	0.17
4	X

$$X=\frac{(4)(0.17)}{20}=\frac{0.68}{20}=0.03 \text{ y } F_{1\%}(11,44)=2.73-0.03=2.70$$

$$F_{0.1\%}(11,40)=\frac{3.87+3.64}{2}=3.76 \qquad F_{0.1\%}(11,60)=\frac{3.54+3.32}{2}=3.43$$

20	0.33
4	X

$$X=\frac{(4)(0.33)}{20}=\frac{1.32}{20}=0.07 \text{ y } F_{0.1\%}(11,44)=3.76-0.07=3.69$$

Por ser la F calculada 143.65 mayor que 2.02, mayor que 2.70 y también mayor que 3.69, se rechaza la hipótesis nula de la igualdad entre las medias de tratamientos con un nivel de significación del 0.1%. La tabla 10.29 muestra el análisis de varianza obtenido.

TABLA 10.29 Análisis de varianza

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	59	3622.93			
TRATAMIENTOS	11	3428.93	311.72	143.65	P < 0.001
BLOQUES	4	98.6			
ERROR	44	95.4	2.17		

Después de llegar a este punto se hace necesario poner en evidencia una situación.

Durante el desarrollo del presente ejemplo hemos considerado que estamos trabajando con 12 tratamientos, pero sin tomar aún en cuenta, que en realidad el objetivo de la investigación es estudiar el efecto de *dos factores* (tipo de leche y forma de envase), y que al tener el factor tipo de leche *cuatro niveles* (T_1 , T_2 , T_3 y T_4) y el factor forma de envase *tres niveles* (F_1 , F_2 y F_3), se produce al combinarlos la existencia de los 12 tratamientos.

Una investigación como la que estamos tratando se dice que ha sido desarrollada mediante un *arreglo factorial 4 x 3 en Bloques al Azar*.

Tomando en cuenta lo antes expuesto, con toda seguridad el interés de la compañía agroalimentaria Nestlé S.A. es hacer el estudio *por separado* de los *efectos principales* de los factores Tipo de Leche y Forma de Envase.

La posibilidad de que esto pueda hacerse depende de la *interacción* entre ambos factores, la cual se denota como A x B, y cuyo concepto tiene una gran connotación en cualquier arreglo factorial.

10.7.1 El concepto de interacción.

Una forma de definir el concepto de interacción es diciendo que la misma expresa el grado en que difiere el efecto de un factor sobre la variable bajo estudio según el valor que toma el otro factor, en nuestro caso, como difiere el efecto del tipo de leche sobre los volúmenes de venta del producto en dependencia de que se trate de uno u otro tipo de envase.

Por esta razón se dice con frecuencia que la interacción expresa la magnitud del *efecto cruzado* entre dos factores.

Con el objetivo de precisar con más detalle este importante concepto de *interacción,* en la tabla 10.30 se muestran las ventas totales por tratamiento y por efectos principales del factor Tipo de Leche y el factor Forma de Envase.

Observe en la tabla 10.30 lo siguiente:

- El tipo de leche 1 alcanza una mayor venta cuando es envasado con la forma de envase 2
- El tipo de leche 2 con la forma de envase 1

- El tipo de leche 3 con la forma de Envase 3
- El tipo de leche 4 con la forma de Envase 1

Lo mismo ocurre con las formas de envase:

- La forma de envase 1 alcanza su mayor volumen de venta con el tipo de leche 4
- La forma de envase 2 con el tipo de leche 4
- La forma de envase 3 con el tipo de leche 3

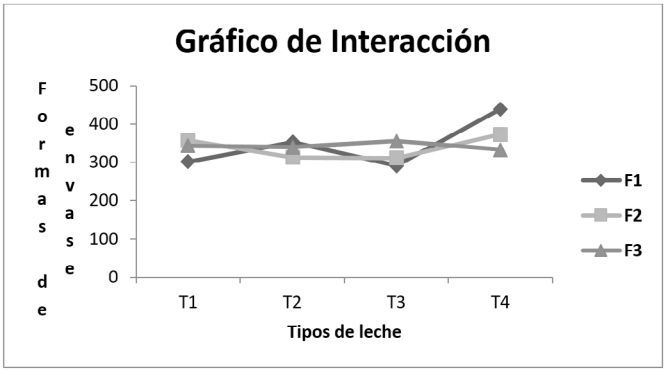
TABLA 10.30 Ventas totales por tratamiento y por efectos principales

Tipo de leche	**Forma de envase**			**Efecto Principal**
	F_1	**F_2**	**F_3**	
T_1	301	357	343	1001
T_2	353	312	339	1004
T_3	291	311	356	958
T_4	439	373	333	1145
Efecto Principal	1384	1353	1371	4108

El análisis efectuado anteriormente *sugiere* que entre ambos factores existe una interacción que nos obliga a encontrar, en caso de que ésta exista, la combinación óptima de los niveles de ambos factores, y en consecuencia, estudiar los efectos principales de los factores de forma separada. Por supuesto que tendremos que probar estadísticamente que lo anteriormente expresado es cierto, y que el *efecto cruzado o la interacción es significativa*.

El gráfico que se muestra en la figura 10.1 corrobora claramente el análisis que hemos efectuado.

FIGURA 10.1 Gráfico de efecto cruzado o interacción



10.7.2 Cálculo de la suma de cuadrados debida a la interacción.

Por lo visto hasta el momento, el modelo lineal para un arreglo factorial con dos factores según un diseño en Bloques al Azar es:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_k + e_{ijk}$$

donde:

μ = constante general común a todas las observaciones

α_i = efecto principal del factor A

β_j = efecto principal del factor B

$(\alpha\beta)_{ij}$ = efecto de la interacción entre los dos factores

δ_k = efecto del bloque

$\mathbf{e_{ijk}}$ = error aleatorio normalmente distribuido son media cero y varianza homogénea σ^2.

Es decir, el efecto de tratamiento puede ser expresado como la suma de los efectos principales de los factores más el efecto de la interacción, por tanto:

$SC_{TRAT.} = SC_T + SC_F + SC_{TXF}$ o lo que es lo mismo:

$SC_{TxF} = SC_{TRAT.} - SC_T - SC_F$

Calculemos las sumas de cuadrados requeridas y para ello tome en cuenta que cada nivel de Tipo de Leche tiene 15 observaciones y cada nivel de Forma de Envase tiene 20 observaciones:

$$SC_T = \frac{(1001)^2 + (1004)^2 + (958)^2 + (1145)^2}{15} - 281261.07$$

$$SC_T = 282587.07 - 281261.07 = 1326$$

$$SC_F = \frac{(1384)^2 + (1353)^2 + (1371)^2}{20} - 281261.07$$

$$SC_F = 281285.3 - 281261.07 = 24.23$$

$$SC_{TxF} = 3428.93 - 1326 - 24.23 = 2078.70$$

La determinación de los grados de libertad se realiza con un razonamiento análogo:

Por ser cuatro los tipos de leche, $GL_T = 3$

Por ser tres las formas de envase, $GL_F = 2$

$GL_{TxF} = GL_{TRAT.} - GL_T - GL_F = 11\text{-}3\text{-}2 = 6$

TABLA 10.31 Análisis de varianza

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	59	3622.93			
TRATAMIENTOS	11	3428.93	311.72		
TIPO DE LECHE	3	1326.00	442.00	203.69	P < 0.001
FORMA DE ENVASE	2	24.23	12.12	5.58	P < 0.01
T x F	6	2078.70	346.45	159.65	P < 0.001
BLOQUES	4	98.6			
ERROR	44	95.4	2.17		

Comprobemos la columna correspondiente a la significación (SIGN.).

En la **TABLA T.3** del Anexo A no están reportados los percentiles de la *F* de Fisher para 44 del Error, por tanto debemos interpolar según lo ya establecido:

a) Para el efecto principal Tipo de Leche

$$F_{5\%}(3,40)=2.84 \qquad F_{1\%}(3,40)=4.31 \qquad F_{0.1\%}(3,40)=6.59$$

$$F_{5\%}(3,60)=2.76 \qquad F_{1\%}(3,60)=4.13 \qquad F_{0.1\%}(3,60)=6.17$$

$$\begin{array}{ll} 20 & 0.08 \\ 4 & X \end{array} \qquad X=\frac{(4)(0.08)}{20}=0.02$$

$$F_{5\%}(3,44)=2.84-0.02=2.82$$

$$\begin{array}{ll} 20 & 0.18 \\ 4 & X \end{array} \qquad X=\frac{(4)(0.18)}{20}=0.04$$

$$F_{1\%}(3,44)=4.31-0.04=4.27$$

$$\begin{array}{ll} 20 & 0.42 \\ 4 & X \end{array} \qquad X=\frac{(4)(0.42)}{20}=0.08$$

$$F_{0.1\%}(3,44)=6.59-0.08=6.51$$

Como el valor de la F calculada para el efecto principal Tipo de Leche 203.69 es mayor que 2.82, 4.27 y 6.51 entonces las medias de efectos principales de este factor difieren significativamente al 0.1%.

b) Para el efecto principal Forma de Envase

$F_{5\%}(2,40)=3.23 \qquad F_{1\%}(2,40)=5.18 \qquad F_{0.1\%}(2,40)=8.25$

$F_{5\%}(2,60)=3.15 \qquad F_{1\%}(2,60)=4.98 \qquad F_{0.1\%}(2,60)=7.77$

20 0.08 $X=\dfrac{(4)(0.08)}{20}=0.02$

4 X

$F_{5\%}(2,44)=3.23-0.02=3.21$

20 0.20 $X=\dfrac{(4)(0.20)}{20}=0.04$

4 X

$F_{1\%}(2,44)=5.18-0.04=5.14$

20 0.48 $X=\dfrac{(4)(0.48)}{20}=0.10$

4 X

$F_{0.1\%}(2,44)=8.25-0.10=8.15$

Como el valor de la F calculada para el efecto principal Forma de Envase 5.58 es mayor que 3.21 y 5.14 pero es menor que 8.15, entonces las medias de efectos principales de este factor difieren significativamente al 1%.

c) Para la interacción

$F_{5\%}(6,40)=2.34 \qquad F_{1\%}(6,40)=3.29 \qquad F_{0.1\%}(6,40)=4.73$

$F_{5\%}(6,60)=2.25 \qquad F_{1\%}(6,60)=3.12 \qquad F_{0.1\%}(6,60)=4.37$

20 0.09 $X=\dfrac{(4)(0.09)}{20}=0.02$

4 X

$F_{5\%}(6,44)=2.34-0.02=2.32$

20 0.17 $X=\dfrac{(4)(0.17)}{20}=0.03$

4 X

$F_{1\%}(6,44)=3.29-0.03=3.26$

$$X = \frac{(4)(0.36)}{20} = 0.07$$

20	0.36
4	X

$$F_{0.1\%}(6,44) = 4.73 - 0.07 = 4.66$$

Como el valor de la F calculada para la interacción 159.65 es mayor que 2.32, 3.26 y 4.66 entonces ésta es significativa al 0.1%. Este último resultado indica que no es posible estudiar por separado los efectos principales de Tipo de Leche ni tampoco las Formas de Envase, y por tanto, deberemos centrar nuestra atención en los 12 tratamientos que resultan de combinar los 4 niveles de Tipo de Leche con los 3 niveles de Forma de Envase, para lo cual tendremos que desarrollar una prueba de comparación múltiple de Duncan para las medias de tratamientos.

Los valores críticos de rango múltiple de Duncan al 5% para rango 11 y 44 grados de libertad del error no aparecen reportados, como puede apreciarse en la **TABLA T.4** del Anexo A.

Por tal razón, debemos proceder a realizar una doble interpolación (horizontal y vertical) en dicha tabla.

Para ello procedemos de la siguiente manera:

A) Interpolación horizontal

a) Obtengamos el percentil de rango 11 para 40 grados de libertad:

El percentil de rango 10 es 3.35 y el de rango 12 es 3.39.

Por tanto, el percentil de rango 11 es:

$$\frac{3.35 + 3.39}{2} = 3.37$$

b) Obtengamos el percentil de rango 11 para 60 grados de libertad:

El percentil de rango 10 es 3.33 y el de rango 12 es 3.37.

Por tanto, el percentil de rango 11 es:

$$\frac{3.33 + 3.37}{2} = 3.35$$

B) Interpolación vertical

Calculemos entonces los 11 percentiles requeridos para 44 grados de libertad:

Para 40 grados de libertad:

2.86 3.01 3.10 3.17 3.22 3.27 3.30 3.33 3.35 3.37 3.39

Para 60 grados de libertad:

2.83 2.98 3.08 3.14 3.20 3.24 3.28 3.31 3.33 3.35 3.37

Aplicando la regla de tres simple para el primer valor tenemos:

20 0.03

4 X

$$X = \frac{(4)(0.03)}{20} = 0.01$$

Por tanto, el percentil de rango 2 con 44 grados de libertad en el Error es 2.86 – 0.01 = 2.85. Procediendo de la misma forma pueden ser calculados el resto de los percentiles. El lector podrá comprobar que finalmente los 11 percentiles para 44 grados de libertad son los siguientes:

(2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12)

2.85 3.00 3.10 3.16 3.22 3.26 3.30 3.33 3.35 3.37 3.39

La tabla 10.32 contiene las medias de tratamientos y efectos principales.

TABLA 10.32 Medias de tratamientos y de efectos principales

Tipo de leche	**Forma de envase**			
	F_1	F_2	F_3	Efecto Principal
T_1	60.2	71.4	68.6	66.7
T_2	70.6	62.4	67.8	66.9
T_3	58.2	62.2	71.2	63.9
T_4	87.8	74.6	66.6	76.3
Efecto Principal	69.2	67.6	68.6	

Los errores estándar para tratamientos y los efectos principales vienen dados por:

$$E.E._T = \sqrt{\frac{2.17}{15}} = \pm 0.38 \qquad E.E._F = \sqrt{\frac{2.17}{20}} = \pm 0.33 \qquad E.E._{TRAT.} = \sqrt{\frac{2.17}{5}} = \pm 0.66$$

La prueba de Duncan tiene la forma que se muestra a continuación:

	3.39	3.37	3.35	3.33	3.30	3.26	3.22	3.16	3.10	3.00	2.85
	87.8	74.6	71.4	71.2	70.6	68.6	67.8	66.6	62.4	62.2	60.2
58.2	*	*	*	*	*	*	*	*	*	*	*
60.2	*	*	*	*	*	*	*	*	*	*	
62.2	*	*	*	*	*	*	*	*	NS		
62.4	*	*	*	*	*	*	*	*			
66.6	*	*	*	*	*	*	NS				
67.8	*	*	*	*	*	NS					
68.6	*	*	*	*	*						
70.6	*	*	NS	NS							
71.2	*	*	NS								
71.4	*	*									
74.6	*										

El resumen de resultados se muestra en la tabla 10.33.

TABLA 10.33 Análisis de varianza

TIPO DE LECHE	**FORMA DE ENVASE**				
	F_1	**F_2**	**F_3**	**Efecto Principal**	
T_1	60.2^{e}	71.4^{d}	68.6^{c}	66.7	P<0.001 ±0.38
T_2	70.6^{d}	62.4^{a}	67.8^{bc}	66.9	
T_3	58.2^{f}	62.2^{a}	71.2^{d}	63.9	
T_4	87.8^{g}	74.6^{h}	66.6^{b}	76.3	
Efecto Principal	69.2	67.6	68.6	P<0.001 ±0.66	
	P<0.01 ± 0.33				

Al ser la interacción significativa la prueba de comparación múltiple de Duncan se le hizo a las medias de tratamientos y no a la de los efectos principales.

La interacción AxB entre dos factores, estudiada en el presente numeral, recibe el nombre de *interacción de primer orden*.

10.8 Análisis de varianza para proporciones.

En el Capítulo 7 estudiamos que una *proporción* es un valor que señala la parte de la muestra de una población que tiene un rasgo distintivo que la identifica, es decir,

una proporción se calcula como $p = \frac{X}{n}$, donde X es el número de elementos de la muestra que poseen el rasgo distintivo y n el tamaño de la muestra.

Con bastante frecuencia ocurre que nos interesa someter a prueba si existen diferencias significativas entre t tratamientos cuyos resultados experimentales están expresados en términos de *proporción*. En este caso, el procedimiento estadístico que hemos estado estudiando durante todo el presente capítulo puede ser utilizado con estos fines. Desarrollemos un ejemplo al respecto con el fin de exponer de forma detallada los métodos de cálculo correspondientes.

Una compañía de seguros está interesada en estudiar si existen diferencias entre cinco grupos de edades de conductores de autos con respecto a los accidentes de tránsito. Los tratamientos fueron:

A: Entre 20 y 29 años

B: Entre 30 y 39 años

C: Entre 40 y 49 años

D: Entre 50 y 59 años

E: Más de 60 años

La hipótesis nula y la hipótesis alternativa de esta prueba son:

$H_0 : \pi_A = \pi_B = \pi_C = \pi_D = \pi_E$

H_1 : No todas las proporciones poblacionales son iguales

Para lograr su objetivo la compañía extrajo una muestra de 400 conductores por grupo de edades en la cual obtuvo los resultados que se aprecian en la tabla 10.34.

TABLA 10.34 Conductores clasificados por edades y con al menos un accidente

GRUPOS DE EDADES	CON AL MENOS UN ACCIDENTE	TAMAÑO DE MUESTRA
Entre 20 y 29	132	400
Entre 30 y 39	125	400
Entre 40 y 49	129	400
Entre 50 y 59	98	400
Más de 60	75	400

El primer paso para desarrollar el análisis de varianza consiste en calcular las cifras que aparecen en la tabla 10.35.

TABLA 10.35 Cálculos requeridos para el análisis de varianza

TRATAMIENTOS	m	n	m/n	m^2/n
Entre 20 y 29	132	400	0.33	43.56
Entre 30 y 39	125	400	0.31	39.063
Entre 40 y 49	129	400	0.32	41.603
Entre 50 y 59	98	400	0.24	24.01
Más de 60	75	400	0.19	14.063
TOTAL	559	2000		156.24

Partiendo de la tabla 10.35 realicemos el análisis de varianza para este caso.

$$FC = \frac{\left(\sum m_i\right)^2}{\sum n_i} = \frac{M^2}{N} = \frac{(559)^2}{2000} = 156.24$$

$$= \left(\frac{132^2}{400} + \frac{125^2}{400} + \frac{129^2}{400} + \frac{98^2}{400} + \frac{75^2}{400}\right) - 156.24 = 162.30 - 156.24 = 6.06$$

$$CM_{ERROR} = P(1-P) = \frac{M}{N}\left(1 - \frac{M}{N}\right)$$

$$CM_{ERROR} = \frac{559}{2000}\left(1 - \frac{559}{2000}\right) = 0.28\,(0.72) = 0.20$$

el cual tiene infinitos grados de libertad.

Los percentiles de la distribución F de Fisher son:

$$F_{5\%}(4,\infty) = 2.37 \qquad F_{1\%}(4,\infty) = 3.32 \qquad F_{0.1\%}(4,\infty) = 4.62$$

Como la F calculada 7.58 es mayor que 2.37, 3.32 y 4.62 hay una diferencia significativa entre las cinco proporciones poblacionales con un nivel de significación del 0.1%.

El análisis de varianza correspondiente se muestra en la tabla 10.36.

TABLA 10.36 Análisis de varianza

FUENTES DE VARIACION	G.L.	S.C.	C.M.	F	SIGN.
TRATAMIENTOS	4	6.06	1.52	7.58	P < 0.001
ERROR	∞		0.20		

$$E.E.(P) = \sqrt{\frac{CM_{ERROR}}{n}} = \sqrt{\frac{0.20}{400}} = \pm 0.02$$

El lector podrá comprobar que la prueba de comparación múltiple de Duncan queda de la siguiente manera:

	3.09	3.02	2.92	2.77
	0.33	0.32	0.31	0.24
0.19	*	*	*	NS
0.24	*	*	*	
0.31	NS	NS		
0.32	NS			

La tabla 10.37 resume los resultados de la investigación.

TABLA 10.37 Resumen de los resultados de la investigación

GRUPO DE EDADES	PROPORCIONES	SIGN. Y E.E.
Entre 20 y 29	0.33^{b}	$P < 0.001 \pm 0.02$
Entre 30 y 39	0.31^{b}	
Entre 40 y 49	0.32^{b}	
Entre 50 y 59	0.24^{a}	
Más de 60	0.19^{a}	

10.9 Número de repeticiones en el diseño experimental.

Un aspecto de vital importancia en el diseño de una investigación es la determinación del número de réplicas o repeticiones necesarias para detectar diferencias entre medias de una determinada magnitud. En este sentido Cochran y Cox (1957), al referirse a los métodos para incrementar la exactitud de los experimentos señalaron que *"Cualquiera sea la fuente de los errores experimentales, la repetición del experimento disminuye constantemente el error asociado a la diferencia entre los resultados medios de dos tratamientos, siempre y cuando algunas precauciones (tales como la aleatorización) se hayan tenido...."*

Existen variados métodos que nos permiten determinar el número de réplicas en el diseño de experimentos. En este libro abordaremos el que en nuestra opinión resulta ser el más adecuado, nos referimos al método según la prueba F-Fisher, el cual está basado en la potencia de la prueba F (Scheffé, 1959), es decir, en el cálculo de la probabilidad:

$P = \left(F'_{\upsilon_1,\upsilon_2.\delta} > F_{\alpha;\upsilon_1;\upsilon_2}\right) = \beta$ donde $F'_{\upsilon_1,\upsilon_2,\delta}$ es la llamada *F no central*, ν_1 y ν_2 son los grados de libertad del numerador y denominador, δ es el llamado *parámetro de no centralidad*, α el nivel de significación y β la probabilidad de rechazar la hipótesis nula cuando ésta es falsa, es decir, la potencia de la prueba.

Patnaik (1949) propuso una aproximación a la F no central la cual consiste en la relación:

$$F'_{\upsilon_1,\upsilon_2;\delta} \cong c\upsilon_1^{-1}\upsilon_1' F\upsilon_1', \upsilon_2$$

donde c y ν_1 se determinan mediante las expresiones :

$$c\upsilon_1' = \upsilon_1 + \delta^2 \qquad c^2\upsilon_1' = \upsilon_1 + 2\delta^2$$

Aplicando la aproximación descrita:

$$P\left[F\upsilon_1,\upsilon_2 > \upsilon_1(c\upsilon_1')F_{\alpha;\upsilon_1,\upsilon_2}\right] = \beta$$

y el cálculo del número de réplicas se limita a determinar el valor de r para el cual la expresión anterior alcanza la magnitud deseada.

Si representamos con D la diferencia que se desea detectar entre dos medias poblacionales de tratamientos, entonces la expresión:

$$\nabla = \frac{D}{\sigma}$$

es el *rango estandarizado entre dos medias de tratamientos*.

Siguiendo el método utilizado por Menchaca (1974,1975) se calculó el número de réplicas requerido en diseños Completamente al Azar y Bloques al Azar para diferente número de tratamientos, niveles de significación del 5, 1 y 0.1 %, potencias de la prueba F del 70, 80 y 90 % y diferentes niveles del rango estandarizado entre medias de tratamientos. Los resultados se reportan a partir de la siguiente página.

El número de réplicas o repeticiones reportadas en las tablas con:

$\nabla = 1$ permite detectar diferencias entre tratamientos igual a σ.

$\nabla = 1.5$ permite detectar diferencias entre tratamientos igual a 1.5σ.

$\nabla = 2$ permite detectar diferencias entre tratamientos igual a 2σ.

$\nabla = 2.5$ permite detectar diferencias entre tratamientos igual a 2.5σ.

$\nabla = 3$ permite detectar diferencias entre tratamientos igual a 3σ.

TABLA 10.38 Determinación del número de réplicas en el diseño Completamente al Azar

∇= 1									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	6	7	9	8	9	12	12	14	17
4	8	10	13	10	12	16	15	18	21
6	10	12	15	12	15	18	17	20	24
8	11	13	17	13	16	19	19	21	26
10	12	14	18	14	17	21	20	23	28
12	12	15	19	15	18	23	21	25	29
14	13	16	20	16	19	24	22	26	31
16	14	17	21	17	20	25	24	27	32
18	15	18	22	18	21	26	25	28	33
20	15	19	23	19	22	27	25	29	35

TABLA 10.39 Determinación del número de réplicas en el diseño Completamente al Azar

∇= 1.5									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	4	4	5	4	6	6	7	8	9
4	5	5	6	6	6	8	8	9	10
6	5	6	7	6	7	9	9	10	11
8	6	6	8	7	8	10	9	11	12
10	6	6	9	7	8	10	10	11	13
12	6	6	9	8	9	11	10	12	14
14	7	7	10	8	9	11	11	12	14
16	7	7	10	8	10	12	11	13	15
18	7	7	10	9	10	12	12	13	16
20	7	7	11	9	10	13	12	14	16

TABLA 10.40 Determinación del número de réplicas en el diseño Completamente al Azar

∇= 2									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	3	3	4	3	4	4	5	5	6
4	3	4	4	4	4	5	5	6	7
6	4	4	5	4	5	6	6	6	7
8	4	4	5	4	5	6	6	7	8
10	4	5	5	5	5	6	6	7	8
12	4	5	6	5	6	7	7	7	9
14	4	5	6	5	6	7	7	8	9
16	4	5	6	5	6	7	7	8	9
18	5	5	6	5	6	7	7	8	9
20	5	6	7	6	6	8	7	8	10

TABLA 10.41 Determinación del número de réplicas en el diseño Completamente al Azar

∇= 2.5									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	2	3	3	3	3	4	4	4	5
4	3	3	3	3	3	4	4	5	5
6	3	3	4	3	4	4	4	5	5
8	3	3	4	3	4	4	4	5	6
10	3	3	4	4	4	5	5	5	6
12	3	4	4	4	4	5	5	5	6
14	3	4	4	4	4	5	5	5	6
16	3	4	4	4	4	5	5	6	6
18	3	4	5	4	4	5	5	6	7
20	3	4	5	4	5	5	5	6	7

TABLA 10.42 Determinación del número de réplicas en el diseño Completamente al Azar

∇= 3									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	2	2	3	3	3	3	4	4	4
4	2	3	3	3	3	3	4	4	4
6	2	3	3	3	3	3	4	4	4
8	2	3	3	3	3	4	4	4	4
10	3	3	3	3	3	4	4	4	5
12	3	3	3	3	3	4	4	4	5
14	3	3	3	3	3	4	4	4	5
16	3	3	4	3	3	4	4	4	5
18	3	3	4	3	4	4	4	4	5
20	3	3	4	3	4	4	4	5	5

TABLA 10.43 Determinación del número de réplicas en el diseño Bloques al Azar

∇= 1									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	7	8	11	9	11	13	14	16	19
4	9	10	13	11	13	16	16	18	22
6	10	12	15	12	15	18	17	20	24
8	11	13	17	13	16	20	19	22	26
10	12	14	18	14	17	21	20	23	28
12	12	15	19	15	18	23	21	25	29
14	13	16	20	16	19	24	22	26	31
16	14	17	21	17	20	25	24	27	32
18	15	18	22	18	21	26	25	28	33
20	15	18	23	19	22	27	25	29	35

TABLA 10.44 Determinación del número de réplicas en el diseño Bloques al Azar

∇= 1.5									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	4	5	6	5	6	7	8	9	10
4	5	6	7	6	7	8	8	9	11
6	5	6	8	6	7	9	9	10	12
8	6	7	8	7	8	10	9	11	13
10	6	7	9	7	8	10	10	11	13
12	6	8	9	8	9	11	10	12	14
14	7	8	10	8	9	11	11	12	15
16	7	8	10	8	10	12	11	13	15
18	7	9	11	9	10	12	12	13	16
20	7	9	11	9	11	13	12	14	16

TABLA 10.45 Determinación del número de réplicas en el diseño Bloques al Azar

∇= 2									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	3	4	4	4	5	5	6	7	8
4	3	4	5	4	5	5	6	6	7
6	4	4	5	4	5	6	6	7	8
8	4	4	5	5	5	6	6	7	8
10	4	5	6	5	5	6	6	7	8
12	4	5	6	5	6	7	7	7	9
14	4	5	6	5	6	7	7	8	9
16	4	5	6	5	6	7	7	8	9
18	5	5	7	5	6	8	7	8	10
20	5	6	7	6	6	8	7	8	10

TABLA 10.46 Determinación del número de réplicas en el diseño Bloques al Azar

∇= 2.5									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	3	3	4	4	4	4	5	6	6
4	3	3	4	3	4	4	5	5	6
6	3	3	4	3	4	4	4	5	6
8	3	3	4	4	4	5	5	5	6
10	3	3	4	4	4	5	5	5	6
12	3	4	4	4	4	5	5	5	6
14	3	4	4	4	4	5	5	6	6
16	3	4	5	4	4	5	5	6	6
18	3	4	5	4	5	5	5	6	7
20	4	4	5	4	5	5	5	6	7

TABLA 10.47 Determinación del número de réplicas en el diseño Bloques al Azar

∇= 3									
α	0.10			0.05			0.01		
β	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
T									
2	3	3	3	3	4	4	5	5	6
4	3	3	3	3	3	4	4	4	5
6	3	3	3	3	3	4	4	4	5
8	3	3	3	3	3	4	4	4	5
10	3	3	3	3	3	4	4	4	5
12	3	3	3	3	3	4	4	4	5
14	3	3	3	3	3	4	4	4	5
16	3	3	4	3	4	4	4	4	5
18	3	3	4	3	4	4	4	4	5
20	3	3	4	3	4	4	4	5	5

Ejercicios del capítulo

10.1 Los datos que se muestran a continuación corresponden a los resultados de una investigación según un diseño completamente aleatorizado con 5 tratamientos y 7 repeticiones:

	Repeticiones						
Tratamientos	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**
A	2.36	2.45	2.49	2.12	2.28	2.22	2.47
B	3.48	3.16	3.45	3.67	3.23	3.17	3.54
C	2.26	2.12	2.14	2.23	2.34	2.13	2.34
D	5.18	5.25	5.17	5.29	5.34	5.62	5.24
E	4.56	4.23	4.27	4.39	4.12	4.29	4.41

a. Desarrolle la tabla del análisis de varianza correspondiente.
b. Calcule las medias de tratamientos.
c. Calcule el error estándar de tratamientos.
d. Desarrolle la prueba de comparación múltiple de Duncan en caso de que sea necesario.
e. Obtenga la tabla que resume los resultados del análisis efectuado.

10.2 En una investigación desarrollada según un diseño completamente aleatorizado con 4 tratamientos y 6 repeticiones se obtuvieron los siguientes resultados:

	Repeticiones					
Tratamientos	I	II	III	IV	V	VI
A	12.47	12.89	12.04	12.47	12.36	12.68
B	14.18	14.56	14.78	14.09	14.23	14.27
C	11.18	11.14	11.69	11.57	11.46	11.36
D	14.29	14.89	14.67	14.16	14.36	14.37

a. Desarrolle la tabla del análisis de varianza correspondiente.
b. Calcule las medias de tratamientos.
c. Calcule el error estándar de tratamientos.
d. Desarrolle la prueba de comparación múltiple de Duncan en caso de que sea necesario.
e. Obtenga la tabla que resume los resultados del análisis efectuado.

10.3 Considere que los resultados del ejercicio anterior fueron los siguientes:

	Repeticiones					
Tratamientos	**I**	**II**	**III**	**IV**	**V**	**VI**
A	12.47	12.89	12.04	12.47	12.36	12.68
B	14.18	14.56		14.09		14.27
C	11.18	11.14	11.69	11.57	11.46	11.36
D	14.29	14.89	14.67	14.16		14.37

a. Desarrolle la tabla del análisis de varianza correspondiente.
b. Calcule las medias de tratamientos.
c. Calcule el error estándar de cada tratamiento.
d. Desarrolle la prueba de comparación múltiple de Duncan en caso de que sea necesario.
e. Obtenga la tabla que resume los resultados del análisis efectuado.

10.4 Los resultados que se muestran a continuación corresponden a una investigación según un diseño completamente aleatorizado con 4 tratamientos y 8 repeticiones:

	Repeticiones							
Tratamientos	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**
A	23.8	23.9	23.1	23.7	23.4	23.3	23.8	23.2
B	29.6	29.7	29.1	29.6	29.4	29.3	29.4	29.1
C	27.4		27.9	27.6	27.4	27.3	27.1	27.9
D	25.1	25.9	25.4	25.7		25.3	25.5	25.1

a. Desarrolle la tabla del análisis de varianza correspondiente.
b. Calcule las medias de tratamientos.
c. Calcule el error estándar de cada tratamiento.
d. De ser necesario, desarrolle la prueba de comparación múltiple de Duncan.
e. Obtenga la tabla que resume los resultados del análisis efectuado.

10.5 En un trabajo de investigación diseñado en bloques al azar con 4 tratamientos y 8 repeticiones se obtuvo los siguientes resultados:

	Repeticiones							
Tratamientos	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**
A	43.7	43.9	43.2	43.6	43.2	43.2	43.7	43.1
B	49.5	49.8	49.2	49.4	49.2	49.1	49.2	49.3
C	47.2	47.9	47.7	47.5	47.4	47.1	47.3	47.5
D	45.3	45.7	45.1	45.6	45.5	45.3	45.2	45.4

a. Desarrolle la tabla del análisis de varianza correspondiente.
b. Calcule las medias de tratamientos.

c. Calcule el error estándar de tratamientos.
d. Desarrolle la prueba de comparación múltiple de Duncan en caso de que sea necesario.
e. Obtenga la tabla que resume los resultados del análisis efectuado.

10.6 Se sometió a prueba el efecto de 6 tratamientos experimentales usando para ello un diseño en bloques al azar con 7 repeticiones. Los resultados de esta investigación se muestran a continuación:

	Repeticiones						
Tratamientos	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**
A	16.2	16.9	16.6	16.1	16.7	16.5	16.3
B	16.7	16.3	16.4	16.9	16.1	16.5	16.7
C	16.4	16.1	16.6	16.5	16.4	16.9	16.3
D	16.2	16.6	16.4	16.1	16.7	16.9	16.3
E	16.6	16.7	16.3	16.2	16.5	16.2	16.1
F	16.1	16.5	16.6	16.4	16.9	16.7	16.2

a. Desarrolle la tabla del análisis de varianza correspondiente.
b. Calcule las medias de tratamientos.
c. Calcule el error estándar de tratamientos.
d. Desarrolle la prueba de comparación múltiple de Duncan en caso de que sea necesario.
e. Obtenga la tabla que resume los resultados del análisis efectuado.

10.7 Considere que en el ejercicio anterior el dato correspondiente al tratamiento C réplica V no pudo ser obtenido, quedando la tabla de resultados como sigue:

	Repeticiones						
Tratamientos	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**
A	16.2	16.9	16.6	16.1	16.7	16.5	16.3
B	16.7	16.3	16.4	16.9	16.1	16.5	16.7
C	16.4	16.1	16.6	16.5		16.9	16.3
D	16.2	16.6	16.4	16.1	16.7	16.9	16.3
E	16.6	16.7	16.3	16.2	16.5	16.2	16.1
F	16.1	16.5	16.6	16.4	16.9	16.7	16.2

a. Obtenga la estimación del valor faltante.
b. Desarrolle la tabla del análisis de varianza correspondiente.
c. Calcule las medias de tratamientos.
d. Calcule el error estándar de cada tratamiento.
e. Desarrolle la prueba de comparación múltiple de Duncan en caso de que sea necesario.

f. Obtenga la tabla que resume los resultados del análisis efectuado.

10.8 Un experimento diseñado en bloques al azar con 5 tratamientos y 6 réplicas arrojó los resultados que se muestran a continuación:

	Repeticiones					
Tratamientos	**I**	**II**	**III**	**IV**	**V**	**VI**
A	17.5	17.9	17.4	17.1	17.3	17.8
B	16.9	16.8	16.5	16.7	16.9	16.8
C	17.3	17.1	17.5	17.2	17.1	17.4
D	16.5		16.4	16.6	16.8	16.5
E	17.5	17.4	17.8	17.3	17.1	17.2

a. Obtenga la estimación del valor faltante.
b. Desarrolle la tabla del análisis de varianza correspondiente.
c. Calcule las medias de tratamientos.
d. Calcule el error estándar de cada tratamiento.
e. Desarrolle la prueba de comparación múltiple de Duncan en caso de que sea necesario.
f. Obtenga la tabla que resume los resultados del análisis efectuado.

10.9 En un experimento factorial 4 x 2 completamente aleatorizado con 5 repeticiones se obtuvo los siguientes resultados:

Factores		**Repeticiones**				
A	**B**	**I**	**II**	**III**	**IV**	**V**
A_1	B_1	56.2	58.4	62.3	59.8	56.7
	B_2	70.2	69.8	71.4	68.3	69.7
A_2	B_1	68.2	67.9	71.4	68.2	69.6
	B_2	59.4	61.3	64.7	60.7	58.3
A_3	B_1	53.4	56.7	60.4	58.7	54.2
	B_2	61.3	59.7	62.2	57.5	62.8
A_4	B_1	84.2	85.6	79.8	88.4	83.2
	B_2	72.8	70.3	73.7	76.5	72.4

a. Desarrolle la tabla del análisis de varianza.
b. Calcule las medias de efectos principales y de tratamientos.
c. Calcule los errores estándar de las medias de efectos principales y de tratamientos.
d. Desarrolle las pruebas de comparación múltiple de Duncan en caso de que sean necesarias.

e. Obtenga la tabla que resume los resultados del análisis realizado.

10.10 Se desarrolló una investigación con el objetivo de estudiar el efecto de 5 niveles de un factor A y 3 niveles de un factor B en el comportamiento de una determinada variable. Para ello se utilizó un diseño en bloques al azar con 6 réplicas. Los resultados de la investigación se muestran a continuación:

Factores		**Repeticiones**					
A	**B**	**I**	**II**	**III**	**IV**	**V**	**VI**
A_1	B_1	7.66	7.26	7.48	7.39	7.74	7.21
	B_2	7.89	7.97	7.78	7.54	7.81	7.44
	B_3	8.01	8.05	8.12	8.16	8.14	8.22
A_2	B_1	8.65	8.36	8.54	8.37	8.71	8.29
	B_2	8.93	8.99	8.84	8.62	8.92	8.66
	B_3	9.12	9.11	9.06	9.15	9.21	9.17
A_3	B_1	9.24	9.36	9.48	9.24	9.44	9.22
	B_2	9.87	9.68	9.58	9.87	9.66	9.67
	B_3	10.12	10.15	10.27	10.11	10.22	10.31
A_4	B_1	9.26	9.44	9.56	9.32	9.55	9.36
	B_2	9.95	9.78	9.66	9.89	9.75	9.81
	B_3	10.24	10.26	10.38	10.27	10.25	10.41
A_5	B_1	9.35	9.57	9.61	9.44	9.63	9.47
	B_2	10.01	10.09	10.15	10.21	10.13	10.08
	B_3	10.31	10.39	10.44	10.34	10.33	10.61

a. Desarrolle la tabla del análisis de varianza correspondiente.
b. Calcule las medias de efectos principales y de tratamientos.
c. Calcule los errores estándar de las medias de efectos principales y de tratamientos.
d. Desarrolle las pruebas de comparación múltiple de Duncan en caso de que sean necesarias.
e. Obtenga la tabla que resume los resultados del análisis realizado.

10.11 Con el objetivo de conocer la opinión de los pacientes del hospital del IESS de la ciudad de Manta, acerca de la entrega de medicamentos por dicha institución, el hospital seleccionó una muestra de 500 pacientes dividida en 5 grupos de 100 y a cada uno de estos grupos le preguntó su opinión acerca del servicio de farmacia. Las respuestas de los pacientes se muestran a continuación:

Respuestas	m	n
Excelente	12	100
Muy bien	17	100
Bien	62	100
Regular	34	100
Mal	3	100

a. Obtenga la tabla del análisis de varianza.
b. Calcule las proporciones por tratamiento.
c. Calcule el error estándar de tratamientos.
d. Desarrolle la prueba de comparación múltiple de Duncan en caso de que sea necesario.
e. Obtenga la tabla que resume los resultados del análisis realizado.

10.12 Los datos que se muestran a continuación representan las respuestas de 800 personas que fueron encuestadas en cuanto a su preferencia entre 6 diferentes tipos de turismo.

Tratamientos	m	n
Turismo de salud	34	150
Turismo religioso	16	150
Turismo cultural	42	100
Turismo de diversión	64	100
Turismo de naturaleza	55	150
Turismo rural	25	100

a. Obtenga la tabla del análisis de varianza.
b. Calcule las proporciones por tratamiento.
c. Calcule el error estándar de tratamientos.
d. Desarrolle la prueba de comparación múltiple de Duncan en caso de que sea necesario.
e. Obtenga la tabla que resume los resultados del análisis realizado.

Capítulo 11

Regresión simple y correlación

El problema

El Banco Central de un país vecino dispone de 50 pares de observaciones correspondientes al producto interno bruto (PIB) de un año en específico y el correspondiente producto interno bruto per cápita (PIBPC) para ese mismo año. ¿Puede el Banco Central establecer con una determinada confiabilidad si existe o no una relación entre ambas variables, y consecuentemente, medir en forma numérica el grado de relación que existe entre las mismas?

11.1 Introducción.

Uno de los primeros trabajos de los que se tiene conocimiento en que se hizo uso de los conceptos iniciales relacionados con la regresión y la correlación, se remonta al siglo XIX cuando el ilustre *Sir Francis Galton (1822-1917)* se dio a la tarea de estudiar todo lo relacionado con la herencia y los modelos matemáticos relacionados con ella. Puede decirse que fue *el pionero* en establecer la relación existente entre dos variables y asignar un valor numérico para expresar de forma objetiva el *grado de relación* entre ellas.

Si quisiéramos dar un concepto de lo que se entiende por *análisis de regresión*, lo expresaríamos diciendo que es una *técnica estadística* que nos permite *modelar* la relación existente entre un grupo de variables *independientes* o también llamadas variables *explicativas* $\left(X_1, X_2, X_3, \ldots\ldots\ldots X_n\right)$ y una variable *dependiente* o también llamada variable *respuesta* (Y).

La regresión de una variable dependiente (Y) sobre una o más variables independientes ($X_1, X_2, \ldots\ldots\ldots, X_n$) expresa la variación que sufre la primera como consecuencia de la variación de las segundas.

La teoría de la regresión permite establecer un modelo matemático que permita hacer una estimación más o menos acertada de la variable dependiente al asignarle valores a las variables dependientes.

Veremos más adelante que la teoría de la correlación nos permitirá obtener de forma numérica el grado o intensidad de la dependencia entre dos o más variables, permitiendo establecer sin ambigüedades si esta relación entre las variables es o no significativa.

11.2 La ecuación de una línea recta.

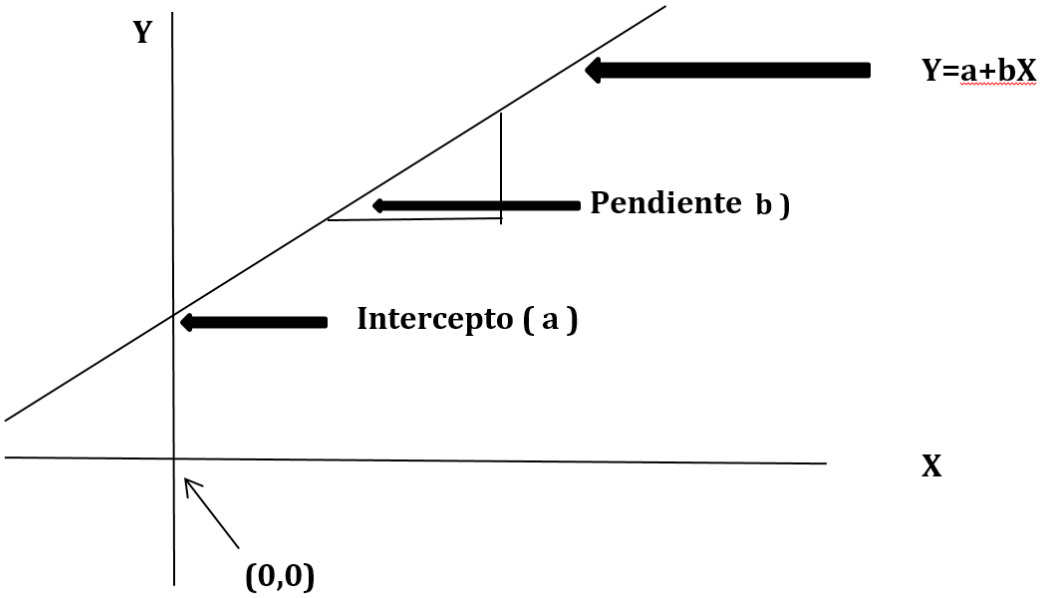
Dedicaremos unos pocos párrafos a *recordar* algunos aspectos relacionados con *la línea recta*, los cuales nos servirán de bastante ayuda en el resto del capítulo.

En primer lugar, la ecuación general de una línea recta viene dada por la expresión:

$Y = a + bX$

donde en un sistema de ejes Cartesianos o Coordenados, *a* es el *intercepto* o *intersección* de la línea recta con el eje vertical y *b* es la *gradiente o pendiente* de la línea recta, tal y como se muestra en la figura 11.1.

FIGURA 11.1 Gráfico y elementos de una línea recta



- Cuando **a > 0**, la línea recta corta el eje Y por encima de 0, mientras que cuando **a < 0** la línea recta corta el eje por debajo de 0.
- Cuando **a = 0** la línea recta pasa por el origen de coordenadas, es decir, por **(0,0)**.
- Cuando **b** > 0, la línea recta es *creciente*, es decir, a incrementos de la variable X corresponden incrementos de la variable Y.
- Cuando **b < 0**, la línea recta es decreciente, es decir, a incrementos de la variable

X corresponden decrecimientos de la variable Y.

- Cuando **b = 0** la línea recta es paralela al eje **X**.

11.3 Regresión lineal simple.

En el Capítulo 10 estudiamos que el modelo lineal general viene dado por la

expresión: $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots\ldots\ldots\ldots + x_{p}\beta_p + e_i$

El modelo anterior es de Análisis de Regresión cuando todas las variables x_{ji} excepto x_{1i} que es igual a 1, son variables aleatorias.

Para el caso de la regresión lineal simple, el modelo lineal general se reduce a la expresión:

$y_i = \beta_1 + x_i\beta_2 + e_i$ o expresado en forma análoga $y_i = \beta_1 + \beta_2 x_i + e_i$

11.3.1 Estimación de los parámetros del modelo.

En el Capítulo 10 vimos también que para estimar el valor numérico de los parámetros β_1 y β_2 se utiliza el *método de los mínimos cuadrados*, el cual consiste en seleccionar los estimadores de β_1 y β_2 de forma tal que hagan mínima la suma de los cuadrados de los errores experimentales, es decir, los mejores estimadores son los que hacen mínima la expresión:

$$\sum e_i^{\,2} = \sum (y_i - a - bx_i)^2$$

donde **a** y **b** son las estimaciones mínimo cuadráticas de β_1 y β_2 respectivamente.

Para obtener el valor de **a** debemos calcular:

$$\frac{\partial \sum e_i^2}{\partial a} = \sum 2(y_i - a - bx_i)\frac{\partial (y_i - a - bx_i)}{\partial a} =$$

$$\sum -2(y_i - a - bx_i) = 0 \Rightarrow \sum (a + bx_i) = \sum y_i \Rightarrow na = \sum y_i - b\sum x_i \Rightarrow a = \bar{y} - b\bar{x}$$

Para obtener el valor de **b** calculamos:

$$\frac{\partial \sum e_i^2}{\partial b} = \sum 2\left(y_i - a - bx_i\right)\frac{\partial \left(y_i - a - bx_i\right)}{\partial b}$$

$$\sum -2\left(y_i - a - b\,x_i\right)\left(x_i\right) = 0 \Rightarrow -\sum x_i y_i + a\sum x_i + b\sum x_i^2 = 0$$

Y como $a = \dfrac{\sum y_i}{n} - b\dfrac{\sum x_i}{n}$

$-\sum x_i y_i + \dfrac{\sum x_i \sum y_i}{n} - b\dfrac{\left(\sum x_i\right)^2}{n} + b\sum x_i^2 = 0$ y entonces:

$$b = \frac{\sum x_i y_i - \dfrac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}} = \frac{SPC_X}{SCC_X}$$

donde SPC_{XY} se lee Suma de productos corregida de XY y SCC_X es la Suma de cuadrados corregida de X.

Al valor **b** se le conoce como *coeficiente de regresión*, y su valor numérico tiene la siguiente interpretación:

a. Cuando **b > 0**, a un incremento de *una unidad* de la variable independiente X corresponde un incremento de *b unidades* de la variable dependiente Y. En este caso, es posible obtener una estimación del valor de Y sustituyendo en la ecuación un valor determinado de X.
b. Cuando **b < 0**, a un incremento de *una unidad* de la variable independiente X corresponde un decrecimiento de *b unidades* de la variable dependiente Y. También en este caso es posible predecir el valor de Y para un determinado valor de X.
c. Cuando **b = 0**, la ecuación de regresión asume la forma *Y = a* y por tanto es paralela al eje de las X. En esta situación resulta imposible predecir un valor de Y.
d. Veamos un ejemplo. La tabla 11.1 muestra los ingresos anuales (X) expresados en miles de dólares y la cuantía del ahorro anual (Y) también expresado en miles de dólares de 11 familias ecuatorianas.

TABLA 11.1 Ingresos y ahorro anual de 11 familias ecuatorianas

X	20.5	20.8	21.2	21.7	22.1	22.3	22.2	22.6	23.1	23.5	22.4
Y	1.9	1.8	2.1	2.1	1.9	2.2	2.2	2.3	2.7	3.1	2.2

Calculemos las sumas de cuadrados y productos necesarios para estimar los parámetros de la ecuación de regresión lineal simple:

$$\sum x_i = 20.5 + 20.8 + 21.2 + \ldots\ldots\ldots + 23.1 + 23.5 + 22.4 = 242.4$$

$$\sum y_i = 1.9 + 1.8 + 2.1 + \ldots\ldots\ldots + 2.7 + 3.1 + 2.2 = 24.5$$

$$\sum x_i y_i = (20.5)(1.9) + (20.8)(1.8) + \ldots\ldots\ldots + (22.4)(2.2) = 542.85$$

$$\sum x_i^2 = (20.5)^2 + (20.8)^2 + \ldots\ldots + (23.5)^2 + (22.4)^2 = 5350.14$$

y para ser utilizada con posterioridad:

$$\sum y_i^2 = (1.9)^2 + (1.8)^2 + \ldots\ldots + (3.1)^2 + (2.2)^2 = 55.99$$

$$b = \frac{542.85 - \frac{(242.4)(24.5)}{11}}{5350.14 - \frac{(242.4)^2}{11}} = \frac{2.96}{8.53} = 0.35$$

$$a = \frac{24.5}{11} - (0.35)\left(\frac{242.4}{11}\right) = 2.23 - 7.71 = -5.48$$

Observe que fue necesario obtener primero el valor de *b* para posteriormente calcular el valor de *a*.

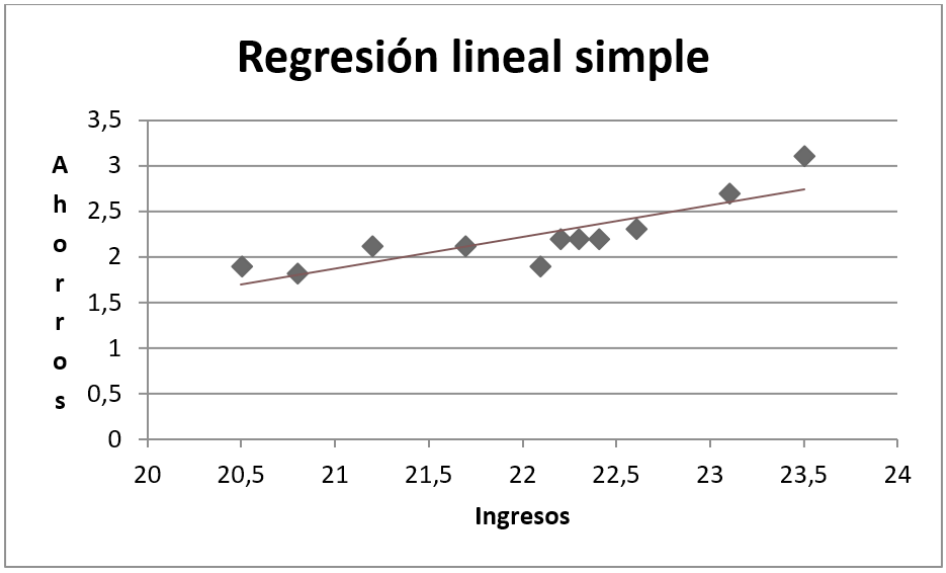
La razón es que el valor de *a* es *dependiente* del valor de *b*, o dicho de otra manera, el parámetro *a* no es *independiente*.

Finalmente, la ecuación de *regresión lineal simple* obtenida es:

Y = -5.48 + 0.35X

El *diagrama de dispersión* de los datos procesados así como el de la línea de regresión obtenida se muestra en la figura 11.2.

FIGURA 11.2 Diagrama de dispersión de la línea de regresión obtenida



Antes de abordar un aspecto diferente debemos precisar que al realizar una regresión las series de valores utilizados deben estar apareados siguiendo una determinada lógica de organización, la cual puede tener un carácter biológico, físico, natural o de otra índole, pero que justifique lo acertado del apareamiento.

En un párrafo anterior vimos que cuando el coeficiente de regresión *b* es igual a cero, la ecuación de regresión asume la forma *Y = a* y por tanto al ser paralela al eje de las X, resulta imposible predecir un valor de la variable Y, o dicho de otra manera, la ecuación de regresión lineal simple *no se ajusta a los datos*. En el caso que estamos tratando, el coeficiente de regresión tiene un valor igual a 0.35 y entonces cabe la

pregunta, ¿es este valor significativamente distinto de cero?

Para hallar la respuesta sobre *la bondad del ajuste* de la ecuación debemos realizar la siguiente prueba de hipótesis:

H_0: $\beta_2 = 0$

H_1: $\beta_2 \neq 0$

Como vimos en el Capítulo 10, una vía para tomar una decisión acerca de la veracidad o no de la hipótesis nula anterior es mediante la técnica de Análisis de Varianza, más conocida en este caso como *Análisis de Regresión*.

Comencemos por definir las fuentes de variación de la tabla del Análisis de Regresión y para ello analicemos las fuentes que causan variabilidad en el modelo matemático $y_i = \beta_1 + \beta_2 x_i + e_i$, las cuales son la *TOTAL* (debida a la variable Y), la debida a la *REGRESION* ($\beta_1 + \beta_2 x_i$) y la del *ERROR* (e_i).

El inicio de la tabla del análisis de regresión se muestra en la tabla 11.2.

TABLA 11.2 Análisis de regresión

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	10				
REGRESIÓN	1				
ERROR	9				

La razón por la cual la fuente de variación debida a la REGRESIÓN tiene un grado de libertad es porque de los *dos* parámetros del modelo solo β_2 es independiente ya que β_1 es dependiente de éste.

Calculemos las sumas de cuadrados requeridas para la tabla de análisis de regresión:

$$SCC_{TOTAL} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 55.99 - \frac{(24.5)^2}{11} = 1.42$$

Probemos a continuación que:

$$SCC_{REG.} = \frac{(SPC_{XY})^2}{SCC_X}$$

$$SCC_{REG.} = SCC_{TOTAL} - SCC_{ERROR} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} - \sum (y_i - a - bx_i)^2 \,|$$

$$= \sum y_i^2 - \frac{(\sum y_i)^2}{n} - \sum (y_i^2 - ay_i - bx_i y_i - ay_i + a^2 + abx_i - bx_i y_i + abx_i + b^2 x_i^2)$$

$$= \sum y_i^2 - \frac{(\sum y_i)^2}{n} - \sum y_i^2 + a\sum y_i + b\sum x_i y_i + a\sum y_i - na^2 - ab\sum x_i +$$

$$b\sum x_i y_i - ab\sum x_i - b^2 \sum x_i^2$$

$$= -\frac{(\sum y_i)^2}{n} + 2a\sum y_i + 2b\sum x_i y_i - na^2 - 2ab\sum x_i - b^2 \sum x_i^2$$

y como $a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n}$ *entonces*:

$$-\frac{(\sum y_i)^2}{n} + 2\left(\frac{\sum y_i}{n} - b\frac{\sum x_i}{n}\right)\sum y_i + 2b\sum x_i y_i - n\left(\frac{\sum y_i}{n} - b\frac{\sum x_i}{n}\right)^2$$

$$-2b\left(\frac{\sum y_i}{n} - b\frac{\sum x_i}{n}\right)\sum x_i - b^2 \sum x_i^2$$

$$= -\frac{(\sum y_i)^2}{n} + 2\frac{(\sum y_i)^2}{n} - 2b\frac{\sum x_i \sum y_i}{n} + 2b\sum x_i y_i - \frac{(\sum y_i)^2}{n}$$

$$+2b\frac{\sum y_i \sum x_i}{n} - b^2\frac{(\sum x_i)^2}{n} - 2b\frac{\sum x_i \sum y_i}{n} + 2b^2\frac{(\sum x_i)^2}{n} - b^2\sum x_i^2$$

$$= 2b\sum x_i y_i - 2b\frac{\sum x_i \sum y_i}{n} + b^2\frac{(\sum x_i)^2}{n} - b^2\sum x_i^2$$

$$2b\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right) - b^2\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) = 2bSPC_X - b^2 SCC_X$$

Y como $b = \frac{SPC_{XY}}{SCC_X}$, *entonces*:

$$SCC_{REG.} = 2\frac{(SPC_{XY})^2}{SCC_X} - \frac{(SPC_{XY})^2}{(SCC_X)^2} SCC_X = \frac{(SPC_{XY})^2}{SCC_X} \quad L.Q.Q.D.$$

Por tanto:

$$SCC_{REG.} = \frac{(SPC_{XY})^2}{SCC_X} = \frac{(2.96)^2}{8.53} = 1.03 \quad y\, entonces:$$

$$SCC_{ERROR} = 1.42 - 1.03 = 0.39$$

La tabla del análisis de regresión toma la forma que se observa en la tabla 11.3.

TABLA 11.3 Análisis de regresión

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	10	1.42			
REGRESIÓN	1	1.03	1.03	25.75	
ERROR	9	0.39	0.04		

Los percentiles de la distribución *F* de Fisher son:

$$F_{5\%}(1,9) = 5.12 \quad F_{1\%}(1,9) = 10.56 \quad F_{0.1\%}(1,9) = 22.86$$

Al ser el valor de la F calculada mayor a 22.86, rechazamos la hipótesis nula y en consecuencia, el coeficiente de regresión es significativamente distinto de cero, por tanto, la ecuación de regresión lineal simple obtenida se ajusta a los datos con un nivel de significación del 0.1%.

La tabla final del análisis de regresión es la que se muestra en la tabla 11.4.

TABLA 11.4 Análisis de regresión

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	10	1.42			
REGRESIÓN	1	1.03	1.03	25.75	P<0.001
ERROR	9	0.39	0.04		

11.3.2 Error estándar de estimación.

La estimación de un valor de la variable dependiente Y a través de la ecuación de regresión es, por regla general, un valor aproximado que puede estar más o menos cerca de su valor verdadero.

El error en la estimación da lugar a los *residuos*, es decir, la diferencia existente entre el verdadero valor de Y y su estimación mediante la ecuación de regresión.

Por lo antes expuesto sería conveniente disponer de un valor numérico que nos indique el grado de exactitud de estas estimaciones. Este valor es el denominado *error estándar de la estimación*, el cual viene dado por la expresión:

$$S_{Y.X} = \sqrt{\frac{SC_{ERROR}}{n-2}} = \sqrt{CM_{ERROR}}$$

En el ejemplo que estamos desarrollando:

$$S_{Y.X} = \sqrt{0.04} = \pm 0.2$$

11.3.3 Error estándar del coeficiente de regresión.

El valor del coeficiente de regresión *b* que hemos obtenido es una *estimación puntual de* β_2 basada en una muestra de los ingresos y ahorros anuales de 11 familias ecuatorianas, por lo que resulta aconsejable calcular un estadígrafo de dispersión para *b* con la finalidad de establecer el grado de precisión conque el mismo ha sido estimado.

El estadígrafo al cual hacemos referencia es el *error estándar del coeficiente de regresión.*

Al estudiar la teoría general de los modelos lineales, Scheffé (1959) reportó que la varianza del coeficiente de regresión lineal simple viene dado por la expresión:

$$V(b) = \frac{\sigma^2}{m_{ZZ;\Omega}} \quad donde\ \sigma^2 = CM_{ERROR}\ y\ m_{ZZ;\Omega} = SCC_X\ \ es\ decir,$$

$$V(b) = \frac{CM_{ERROR}}{SCC_X} \quad y\ entonces:$$

$$E.E.(b) = \sqrt{\frac{CM_{ERROR}}{SCC_X}}$$

En nuestro ejemplo $SCC_X = 8.53\ y\ CM_{ERROR} = 0.04,\ de\ donde:$

$$E.E.(b) = \sqrt{\frac{0.04}{8.53}} = \pm 0.07$$

11.3.4 Intervalo de confianza del coeficiente de regresión.

La expresión matemática que permite calcular el intervalo de confianza del coeficiente de regresión viene dada por:

$$b - t_{\alpha}^{GL_{ERROR}}\, E.E.(b) < \beta_2 < b + t_{\alpha}^{GL_{ERROR}}\, E.E.(b)$$

El percentil de la t de Student para una prueba de dos colas, con 9 grados de libertad y nivel de significación del 5% es 2.262, por tanto, un intervalo de confianza para β_2 con un nivel de confiabilidad del 95% viene dado por:

$$0.35-(2.262)(0.07)<\beta_2<0.35+(2.262)(0.07)$$

$0.19 < \beta_2 < 0.51$

11.3.5 Método alternativo para calcular la suma de cuadrados del error.

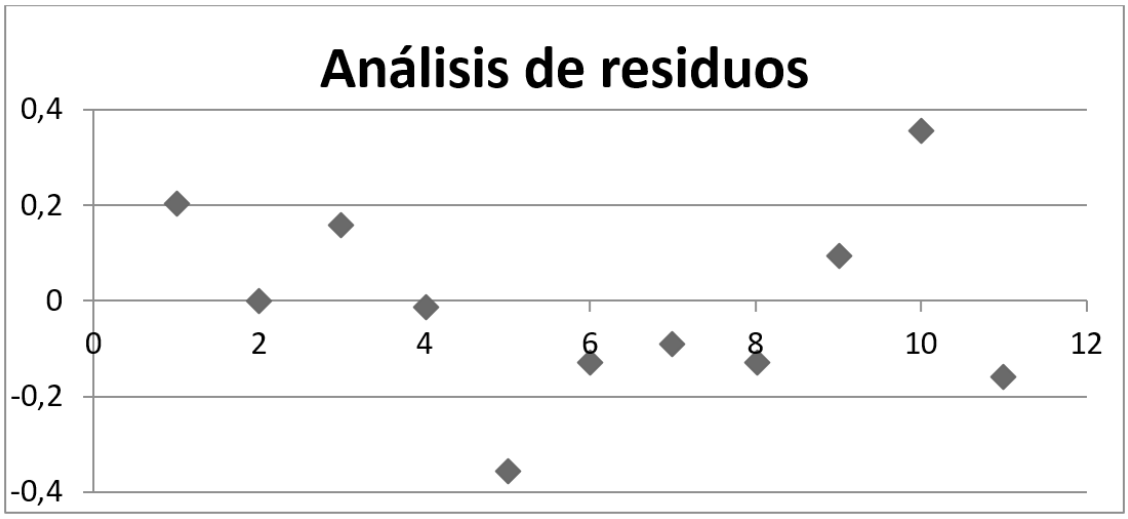
A continuación se presenta la tabla 11.5 en la que la tercera columna son las estimaciones de los valores de la variable dependiente (Y) calculada a partir de la ecuación de regresión, la cuarta columna son los errores en las estimaciones o *residuos* y la última columna los cuadrados de dichos errores. Observe en la tabla como la suma de los cuadrados de los errores o residuos (0.39) la cual ha sido redondeada a dos decimales, es igual a la SCC_{ERROR} de la tabla del análisis de regresión.

Adicionalmente la figura 11.3 muestra un diagrama de dispersión correspondiente a los residuos.

TABLA 11.5 Cálculo de la suma de cuadrados de los errores

X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
20.5	1.9	1.70	0.20	0.0400
20.8	1.8	1.80	0.00	0.0000
21.2	2.1	1.94	0.16	0.0256
21.7	2.1	2.12	- 0.02	0.004
22.1	1.9	2.26	-0.36	0.1296
22.3	2.2	2.33	-0.13	0.0169
22.2	2.2	2.29	-0.09	0.0081
22.6	2.3	2.43	-0.13	0.0169
23.1	2.7	2.61	0.09	0.0081
23.5	3.1	2.75	0.35	0.1225
22.4	2.2	2.36	-0.16	0.0256
			SUMA	0.3937

FIGURA 11.3 Diagrama de dispersión de los residuos



11.3.6 Prueba de hipótesis del coeficiente de regresión.

Un método alternativo para determinar la bondad del ajuste de la ecuación de regresión lineal simple sin tener que desarrollar el análisis de regresión, es apoyándonos en la distribución t de Student.

En el numeral 11.3.3 vimos que la expresión del error estándar del coeficiente de regresión viene dado por:

$$E.E.(b) = \sqrt{\frac{CM_{ERROR}}{SCC_X}} \quad y \ como:$$

$$CM_{ERROR} = \frac{\sum (Y_i - \bar{Y})^2}{n-2} \quad entonces \quad E.E.(b) = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{(n-2)SCC_X}}$$

resultado que permite calcular el error estándar del coeficiente de regresión sin tener que desarrollar el análisis de regresión correspondiente.

Formulemos nuevamente las hipótesis nula y alternativa:

$H_0: \beta_2 = 0 \quad H_1: \beta_2 \neq 0$

Coma ya conocemos

$$T = \frac{b}{E.E.(b)} = \frac{b}{\sqrt{\frac{\sum \left(Y_i - \hat{Y}\right)^2}{(n-2)SCC_X}}}$$

sigue una distribución t de Student con n-2 grados de libertad.

La regla de decisión queda entonces como sigue:

Rechazar H_0 si: $|T| > t_{\alpha}^{(n-2)}$

No rechazar H_0 si: $|T| \leq t_{\alpha}^{(n-2)}$

En el ejemplo:

$$\left|\frac{b}{E.E.(b)}\right| = \left|\frac{0.35}{\sqrt{\frac{0.39}{(11-2)(8.53)}}}\right| = \frac{0.35}{\sqrt{0.005}} = \frac{0.35}{0.07} = 5$$

Los percentiles de la t de Student para 9 grados y una prueba de dos colas son:

2.262 para un nivel de significación del 5%

3.250 para un nivel de significación del 1%

4.781 para un nivel de significación del 0.1%

Como 5 es mayor que 4.781, rechazamos la hipótesis nula con un nivel de significación del 0.1%, es decir, la ecuación de regresión se ajusta satisfactoriamente a los datos.

Observe que la significación obtenida en la prueba de hipótesis usando la distribución t de Student, es la misma que la alcanzada en el análisis de regresión.

11.4 Coeficiente de correlación lineal simple.

En el numeral anterior, concretamos en forma matemática la posible relación existente entre dos variables o series de valores entre las cuales la dependencia podía ser expresada mediante una línea recta.

A menudo resulta muy favorable poder expresar en términos numéricos el grado de dependencia que eventualmente pueda existir entre dos variables o entre las series de datos que la representan. Este valor matemáticamente calculado recibe el nombre de coeficiente de correlación. Para obtener la expresión matemática de este coeficiente partamos de la igualdad que se muestra a continuación:

$$\sum X_i Y_i = \sum \left[\overline{X} + (X_i - \overline{X})\right]\left[\overline{Y} + (Y_i - \overline{Y})\right]$$

$$\sum X_i Y_i = \sum \overline{X}\,\overline{Y} + \sum \overline{X}(Y_i - \overline{Y}) + \sum \overline{Y}(X_i - \overline{X}) + \sum (X_i - \overline{X})(Y_i - \overline{Y}) =$$

$$n\,\overline{X}\,\overline{Y} + \overline{X}\sum (Y_i - \overline{Y}) + \overline{Y}\sum (X_i - \overline{X}) + \sum (X_i - \overline{X})(Y_i - \overline{Y})$$

pero $\sum (X_i - \overline{X}) = 0$ y $\sum (Y_i - \overline{Y}) = 0$

y entonces dividiendo todo para n $\dfrac{\sum X_i Y_i}{n} = \overline{X}\,\overline{Y} + \dfrac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n}$

Partiendo de este último resultado y haciendo uso de algunas técnicas de la teoría de probabilidades, podemos probar matemáticamente que si X e Y son dos variables independientes, entonces se cumple que:

$$\frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n} = 0 \qquad (1)$$

Si por el contrario, X e Y no son independientes, entonces:

$$\frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n} \neq 0$$

Por lo antes expuesto podemos utilizar la expresión (1) para calcular de forma numérica *el grado de dependencia* entre dos variables en concordancia con el hecho de que dicha expresión se acerque o esté alejada significativamente de 0. Si en (1), expresamos los valores de las desviaciones con relación a su media de cada una de las variables en unidades de desviación estándar, tenemos:

$$\frac{\sum \frac{(X_i - \overline{X})}{\sigma_X}\frac{(Y_i - \overline{Y})}{\sigma_Y}}{n} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n\sigma_X\sigma_Y} \qquad (2)$$

expresión que recibe el nombre de *coeficiente de correlación* y se denota con la letra **r**.

Si escribimos la expresión (2) en términos de desviaciones, obtenemos:

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n\sqrt{\frac{\sum (X_i - \overline{X})^2}{n}}\sqrt{\frac{\sum (Y_i - \overline{Y})^2}{n}}}$$

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})}}$$

o lo que es lo mismo:

$$r = \frac{SPC_{XY}}{\sqrt{SCC_X . SCC_Y}}$$

Puede demostrarse matemáticamente que el coeficiente de correlación siempre es una cantidad que oscila entre 1 y -1, es decir, $\mathbf{-1 \leq r \leq 1}$

Cuando $\mathbf{r > 0}$, la correlación es directa, y a valores crecientes de una de las variables corresponden valores crecientes de la otra.

Cuando $\mathbf{r < 0}$, la correlación es inversa, y a valores crecientes de una de las variables corresponden valores decrecientes de la otra.

Cuando $\mathbf{r = 1}$ o $\mathbf{r = -1}$, la correlación es perfecta.

En la medida en que el valor del coeficiente de correlación se acerca a la unidad

por valores positivos o negativos, mayor es la relación lineal entre las dos variables estudiadas. En cambio, valores próximos a cero indican, en general, ausencia de relación lineal entre ambas variables.

Calculemos el coeficiente de correlación en el ejemplo que hemos estado desarrollando.

$$r = \frac{SPC_{XY}}{\sqrt{SCC_X . SCC_Y}} = \frac{2.96}{\sqrt{(8.53)(1.42)}} = 0.85$$

A menudo se suele reportar el valor del cuadrado del coeficiente de correlación en lugar de éste. Este valor recibe el nombre de *coeficiente de determinación* y se denota como $\mathbf{R^2}$. En nuestro caso, $\mathbf{R^2} = 0.72$

Si como hemos dicho, el valor de r oscila entre –1 y 1, entonces el valor de $\mathbf{R^2}$ es siempre menor o igual a 1 y mayor o igual a 0, es decir, $\mathbf{0 \leq R^2 \leq 1}$.

Veamos un método alternativo para calcular el coeficiente de determinación utilizando las sumas de cuadrados del análisis de regresión:

$$R^2 = \frac{SPC_{XY}^2}{SCC_X . SCC_Y} = \frac{SPC_{XY}^2}{SCC_X} \frac{1}{SCC_Y} = \frac{SCC_{REG.}}{SCC_{TOTAL}}$$

El coeficiente de determinación puede definirse entonces como *la cantidad de la variación en Y que es explicada por la regresión.*

En nuestro ejemplo $R^2 = \frac{1.03}{1.42} = 0.72$ valor que coincide exactamente con el calculado anteriormente. Podemos decir entonces que el 72% de la variación de la variable dependiente Y está explicada a través de su relación lineal con la variable independiente X.

11.4.1 Error estándar del coeficiente de correlación.

Con el objetivo de conocer el grado de precisión con el cual el coeficiente de correlación poblacional es estimado, podría resultar conveniente obtener un estadígrafo de dispersión para el coeficiente de correlación muestral calculado.

El error estándar de este coeficiente se calcula a través de las siguientes expresiones:

$E.E.(r) = \frac{1 - r^2}{\sqrt{n}}$ para el caso de muestras pequeñas, y

$E.E.(r) = \frac{1 - r^2}{\sqrt{n - 1}}$ para el caso de muestras grandes.

Fisher propuso que para muestras pequeñas, lo cual es común en la mayor parte de las investigaciones, se utilice el error estándar $E.E.(r)=\frac{1-r^2}{\sqrt{n-2}}$, expresión que utilizaremos en el desarrollo de este capítulo. En nuestro ejemplo:

$$E.E.(r)=\frac{1-(0.85)^2}{\sqrt{11-2}}=\frac{0.28}{3}=\pm 0.09$$

11.4.2 Prueba de hipótesis del coeficiente de correlación.

Sometamos a prueba las hipótesis que se muestran a continuación, donde ρ representa el coeficiente de correlación poblacional:

$H_0:\rho=0$

$H_1:\rho\neq 0$

Si mediante el procedimiento estadístico que explicaremos en los siguientes párrafos rechazamos la hipótesis nula, entonces podremos concluir que la ecuación de regresión lineal simple obtenida se ajusta eficientemente a los datos, de lo contrario, no podremos asegurar que esto ocurre.

$$t=\left|\frac{r}{E.E.(r)}\right|=\left|\frac{r}{\frac{1-r^2}{\sqrt{n-2}}}\right|=\left|\frac{r\sqrt{n-2}}{1-r^2}\right|=\left|\frac{0.85\sqrt{9}}{1-(0.85)^2}\right|=\frac{2.55}{0.28}=9.11$$

Como vimos en párrafos anteriores, los percentiles de la t de Student para 9 grados y una prueba de dos colas son:

2.262 para un nivel de significación del 5%

3.250 para un nivel de significación del 1%

4.781 para un nivel de significación del 0.1%

Como 9.11 es mayor que 4.781, rechazamos la hipótesis nula con un nivel de significación del 0.1%, es decir, la ecuación de regresión se *ajusta satisfactoriamente* a los datos.

Un método alternativo para desarrollar esta prueba de hipótesis consiste en comparar el valor obtenido del coeficiente de correlación (en valor absoluto) con el valor crítico reportado en la **TABLA T.10** del Anexo A. Para n = 11 y una prueba de dos colas, los valores críticos son 0.602 para el 5% y 0.735 para el 1%.

Como el valor de r es 0.85 es mayor que 0.735, entonces se rechaza la hipótesis nula para un nivel de significación del 1%.

Observe que la significación del coeficiente de correlación obtenida mediante la

t de Student coincide con la significación hallada usando la F de Fisher en el análisis de regresión y con la obtenida mediante la prueba de hipótesis del coeficiente de regresión, lo cual significa que estos métodos son vías alternativas para comprobar el ajuste de la ecuación de regresión lineal simple.

11.5 Regresión exponencial simple.

Los datos que se muestran en la tabla 11.6 representan el porcentaje (X) de llantas de cierta marca que aún pueden seguirse utilizando después de recorrer una determinada cantidad de miles de millas (Y).

TABLA 11.6 Porcentaje de llantas (X) y miles de millas de recorrido (Y)

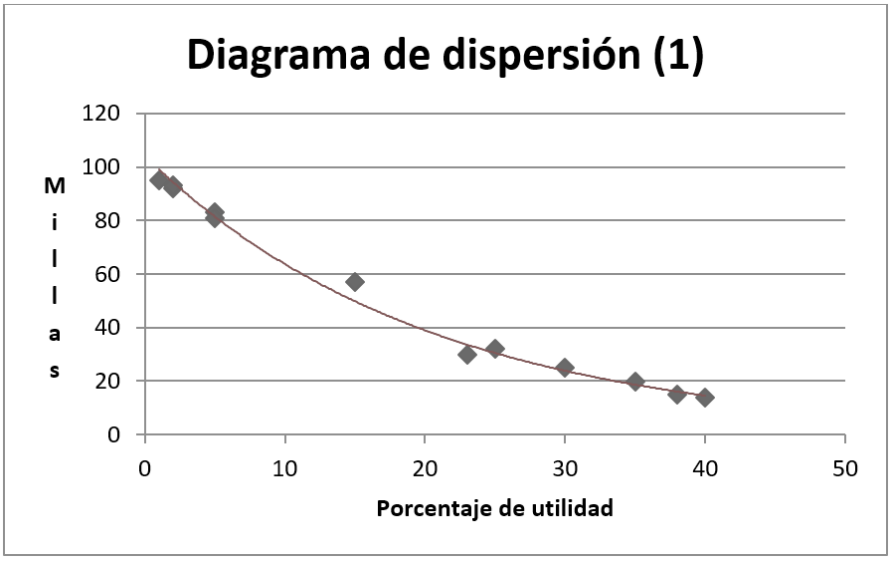
X	5	25	35	2	1	15	30	40	38	2	5	23
Y	83	32	20	92	95	57	25	14	15	93	81	30

Con los datos de la tabla se obtuvieron los diagramas de dispersión correspondiente a una ecuación de tipo exponencial $\left(Y = \beta_1 e^{\beta_2 X}\right)$ y a una ecuación de tipo lineal, la cual es ya conocida por nosotros $\left(Y = \beta_1 + \beta_2 X\right)$.

Los diagramas de dispersión correspondientes se muestran en la figura 11.4 y 11.5 respectivamente.

Observe comparando ambos diagramas que resulta muy difícil, sino imposible, determinar a *simple vista* cuál de las dos ecuaciones se ajusta *mejor* a los datos, razón que nos obliga a realizar el análisis de regresión para ambos modelos matemáticos.

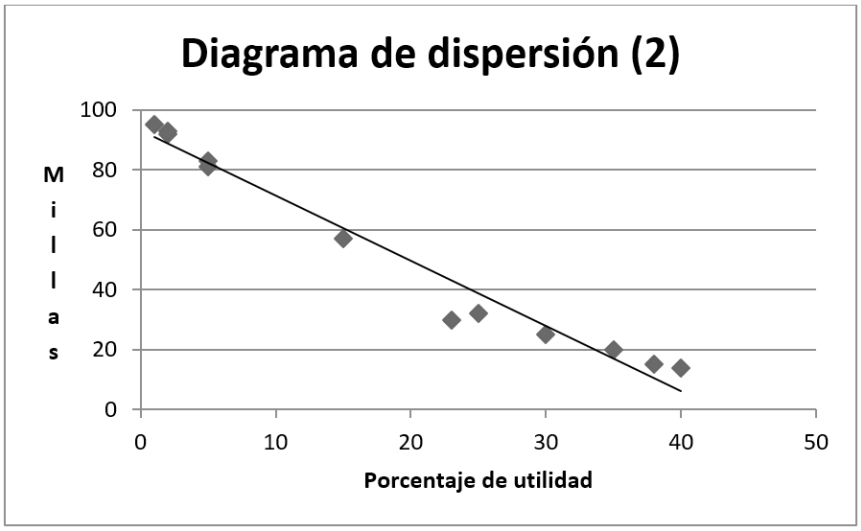
FIGURA 11.4 Diagrama de dispersión para regresión exponencial



Comencemos por realizar el ajuste de la ecuación exponencial, cuyo modelo matemático viene dado por $Y_i = \beta_1 e^{\beta_2 X_i} + e_i$ (3) donde e es la llamada *constante Euler*

(base del logaritmo natural con un valor igual a 2.7183), β_1 y β_2 son los parámetros del modelo y e_i son errores aleatorios que siguen una distribución normal con media cero y varianza σ^2.

FIGURA 11.5 Diagrama de dispersión para regresión lineal



Aplicando logaritmo natural a la expresión (3) y tomando en cuenta algunas propiedades de los logaritmos tenemos:

$\ln Y_i = \ln\left(a\, e^{bX_i}\right)$ donde a y b son las estimaciones mínimo cuadráticas de β_1 y β_2 respectivamente.

$$\ln Y_i = \ln\left(a\, e^{bX_i}\right) = \ln a + \ln e^{bX_i} = \ln a + bX_i \left(\ln e\right) = \ln a + bX_i \;\; pues \;\; \ln e = 1$$

Hemos convertido la ecuación exponencial en una ecuación lineal simple de la forma $\ln Y_i = \ln a + bX_i$, donde la variable dependiente es $\ln Y_i$ y la independiente X_i.

Hagamos el ajuste de ambas ecuaciones utilizando los resultados de la tabla 11.7.

TABLA 11.7 Datos y cálculos necesarios para ajustar ambas ecuaciones

X	Y	XY	X^2	Y^2	ln Y	X lnY	$(\ln Y)^2$
5	83	415	25	6889	4.42	22.1	19.54
25	32	800	625	1024	3.47	86.75	12.04
35	20	700	1225	400	3	105	9
2	92	184	4	8464	4.52	9.04	20.43
1	95	95	1	9025	4.55	4.55	20.7
15	57	855	225	3249	4.04	60.6	16.32
30	25	750	900	625	3.22	96.6	10.37
40	14	560	1600	196	2.64	105.6	6.97
38	15	570	1444	225	2.71	102.98	7.34
2	93	186	4	8649	4.53	9.06	20.52

5	81	405	25	6561	4.39	21.95	19.27
23	30	690	529	900	3.4	78.2	11.56
221	637	6210	6607	46207	44.89	702.43	174.06

Ecuación de regresión exponencial simple:

$$b = \frac{SPC_{X \ln Y}}{SCC_X} = \frac{702.43 - \frac{(221)(44.89)}{12}}{6607 - \frac{(221)^2}{12}} = \frac{-124.29}{2536.92} = -0.05$$

$$\ln a = \frac{44.89}{12} + 0.05\left(\frac{221}{12}\right) = 3.74 + 0.92 = 4.66 \Rightarrow a = e^{4.66} = 105.64$$

La ecuación de regresión exponencial mínimo cuadrática es:

$$Y = 105.64e^{-0.05X}$$

$$SCC_{TOTAL} = \sum (\ln Y_i)^2 - \frac{\left(\sum \ln Y_i\right)^2}{n} = 174.06 - \frac{(44.89)^2}{12} = 6.13$$

$$SCC_{REG.} = \frac{(SPC_{X \ln Y})^2}{SCC_X} = \frac{\left(702.43 - \frac{(221)(44.89)}{12}\right)^2}{6607 - \frac{(221)^2}{12}} = \frac{15449.04}{2536.92} = 6.09$$

$$SCC_{ERROR} = 6.13 - 6.09 = 0.04$$

El lector podrá comprobar que la tabla del análisis de regresión es la que se muestra en la tabla 11.8 :

TABLA 11.8 Análisis de regresión exponencial

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	11	6.13			
REGRESIÓN	1	6.09	6.09	1522.5	P<0.001
ERROR	10	0.04	0.004		

$$E.E.(b) = \sqrt{\frac{CM_{ERROR}}{SCC_X}} = \sqrt{\frac{0.004}{2536.92}} = \pm 0.0013$$

$$r = \frac{SPC_{XY}}{\sqrt{SCC_X\ SCC_Y}} = \frac{-124.29}{\sqrt{(2536.92)\ (6.13)}} = \frac{-124.29}{124.70} = -0.997 \approx -1$$

Ecuación de regresión lineal simple:

$$b = \frac{SPC_{XY}}{SCC_X} = \frac{6210 - \frac{(221)(637)}{12}}{6607 - \frac{(221)^2}{12}} = \frac{-5521.42}{2536.92} = -2.18$$

$$a = \frac{637}{12} + 2.18\left(\frac{221}{12}\right) = 53.08 + 40.15 = 93.23$$

La ecuación de regresión lineal simple mínimo cuadrática es:

$$Y = 93.23 - 2.18X$$

$$SCC_{TOTAL} = 46207 - \frac{(637)^2}{12} = 12392.92$$

$$SCC_{REG.} = \frac{(SPC_{XY})^2}{SCC_X} = \frac{(-5521.42)^2}{2536.92} = \frac{30486078.82}{2536.92} = 12016.96$$

$$SCC_{ERROR} = 12392.92 - 12016.96 = 375.96$$

El lector podrá comprobar que la tabla del análisis de regresión es la que se muestra en la tabla 11.9.

TABLA 11.9 Análisis de regresión lineal

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	11	12392.92			
REGRESIÓN	1	12016.96	12016.96	319.6	P<0.001
ERROR	10	375.96	37.60		

$$E.E.(b) = \sqrt{\frac{37.6}{2536.92}} = \pm 0.12$$

$$r = \frac{-5521.42}{\sqrt{(2536.92)\ (12392.92)}} = \frac{-5521.42}{5607.12} = -0.98$$

Resumiendo los resultados obtenidos en el ajuste de la ecuación lineal y la exponencial tenemos:

$Y = 93.25 - 2.18X$	$Y = 105.64e^{-0.05X}$
P < 0.001	P < 0.001
r = -0.98	r = -1
E.E. (b) = ± 0.12	E.E. (b) = ±0.0013

Como se puede apreciar en el resumen anterior, ambas ecuaciones de regresión se ajustan adecuadamente a los datos, sin embargo, la ecuación exponencial muestra un mejor coeficiente de correlación y un menor error estándar del coeficiente de regresión.

Debemos concluir entonces que la ecuación exponencial es la de *mejor ajuste.*

Para la ecuación de regresión exponencial obtenida, un intervalo de confianza para β_2 con un nivel de confiabilidad del 95% viene dado por:

$$-0.05 - (2.262)(0.0013) < \beta_2 < -0.05 + (2.262)(0.0013)$$

$$-0.053 < \beta_2 < -0.047$$

Ejercicios del capítulo

11.1 Los datos que se muestran a continuación representan los gastos de producción de una empresa y la cantidad de artículos producidos por la misma.

Gastos	525	643	748	513	523	588	716	603	576	612	801
Artículos	80	95	112	76	84	91	105	92	88	90	120

a. Determine la variable dependiente y la independiente.
b. Obtenga la ecuación de regresión lineal simple.
c. Desarrolle el análisis de regresión correspondiente.
d. Determine el error estándar de la estimación.

11.2 Los datos que aparecen a continuación representan el tamaño del núcleo familiar y los gastos mensuales en dólares de consumo de energía eléctrica.

Gastos	31.18	72.48	54.89	42.36	85.39	70.54	40.17	50.47	73.21
Tamaño	2	5	4	3	6	5	3	4	5

a. Determine la variable dependiente y la independiente.
b. Obtenga la ecuación de regresión lineal simple.
c. Desarrolle el análisis de regresión correspondiente.
d. Determine el error estándar de la estimación.

11.3 Con los datos del ejercicio 11.1:

a. Calcule el error estándar del coeficiente de regresión.
b. Obtenga el intervalo de confianza del coeficiente de regresión para $\alpha=0.05$.
c. Determine el valor del coeficiente de correlación.
d. Calcule el error estándar del coeficiente de correlación.

11.4 Con los datos del ejercicio 11.2:

a. Calcule el error estándar del coeficiente de regresión.
b. Obtenga el intervalo de confianza del coeficiente de regresión para $\alpha=0.01$.
c. Determine el valor del coeficiente de correlación.
d. Calcule el error estándar del coeficiente de correlación.

11.5 A continuación se describe el comportamiento de dos variables que al parecer tienen entre si una relación exponencial.

X	1	2	3	4	5	6	7	8	9	10
Y	0.23	8.36	9.97	38.54	35.47	90.14	84.17	105.67	236.14	221.33

a. Obtenga la ecuación de regresión exponencial simple entre estas dos variables.

b. Desarrolle el análisis de regresión correspondiente.
c. Determine el error estándar de la estimación.

11.6 A continuación se describe el comportamiento de dos variables experimentales.

X	1	2	3	4	5	6	7	8
Y	2.13	8.47	5.32	12.48	9.17	32.14	21.17	68.2

a. Obtenga la ecuación de regresión exponencial simple entre estas dos variables.
b. Desarrolle el análisis de regresión correspondiente.
c. Determine el error estándar de la estimación.

11.7 Con los datos del ejercicio 11.5:
a. Calcule el error estándar del coeficiente de regresión.
b. Obtenga el intervalo de confianza del coeficiente de regresión para $\alpha=0.01$.
c. Determine el valor del coeficiente de correlación.
d. Calcule el error estándar del coeficiente de correlación.

11.8 Con los datos del ejercicio 11.6:
a. Calcule el error estándar del coeficiente de regresión.
b. Obtenga el intervalo de confianza del coeficiente de regresión para $\alpha=0.05$.
c. Determine el valor del coeficiente de correlación.
d. Calcule el error estándar del coeficiente de correlación.

11.9 En una empresa metalúrgica dedicada a la producción de herramientas, se desarrolló una investigación con el objetivo de encontrar una posible relación entre la deformación del acero (X) expresada en milímetros y la dureza del mismo (Y) expresada en kg/mm^2. Los resultados de la investigación se muestran a continuación:

X	8	10	12	14	21	25	30	32	37
Y	91	75	71	58	42	38	40	33	34

a. Ajuste una ecuación de regresión lineal simple.
b. Ajuste una ecuación de regresión exponencial simple.
c. Determine de las dos ecuaciones cual se ajusta mejor a los datos.

11.10 Los datos que se muestran a continuación representan la concentración de alcohol en la sangre de una persona (X) y el riesgo de que la misma tenga un accidente automovilístico (Y).

X	0,05	0,07	0,09	0,11	0,13	0,15	0,17	0,19	0,21	0,23
Y	6,63	6,95	7,34	7,65	7,94	8,21	8,56	8,84	9,28	9,63

a. Ajuste una ecuación de regresión lineal simple.
b. Ajuste una ecuación de regresión exponencial simple.
c. Determine de las dos ecuaciones, cuál se ajusta mejor a los datos.

Capítulo 12
Regresión múltiple

El problema

El gerente de un establecimiento dedicado a las ofertas de entretenimiento al aire libre, desarrolló una investigación dentro de la cual estableció la relación que existía entre la asistencia de personas al establecimiento (X) y el consumo (Y) de una bebida caliente hecha a base de chocolate de muy alta calidad. En el análisis de regresión realizado se obtuvo como resultado que tanto el coeficiente de regresión como el coeficiente de correlación entre ambas variables resultaron negativos, evidenciando que a una mayor asistencia de público al establecimiento le correspondía una menor venta de la bebida de referencia. ¿Son estos resultados razonables o la técnica de regresión aplicada adolece de alguna deficiencia?

12.1 Introducción.

En el capítulo anterior, estudiamos los modelos matemáticos que nos permitían establecer la posible relación lineal existente entre una variable dependiente (Y) y otra variable independiente (X). En este capítulo ampliaremos el estudio de la regresión y la correlación a situaciones en las cuales existe una influencia de dos o más variables independientes sobre la variable dependiente.

Las técnicas de regresión y correlación múltiple se ocupan de estas situaciones.

12.2 Regresión lineal múltiple para el caso de dos variables independientes.

La forma general del modelo de regresión lineal múltiple para dos variables independientes X_1 y X_2 viene dado por la expresión:

$$Y_i = \beta_1 + \beta_2 X1_i + \beta_3 X2_i + e_i$$

Si **a** es la estimación mínimo cuadrática de β_1, **b** la estimación mínimo cuadrática de β_2 y **c** la estimación mínimo cuadrática de β_3 entonces $Y = a + bX1 + cX2$ es la ecuación de regresión lineal múltiple, donde X1 y X2 son las variables independientes, a es el término independiente, y b**,** c son los coeficientes de regresión de la línea recta.

Para obtener la fórmula matemática para calcular a, b y c debemos minimizar la suma de los cuadrados de los errores, es decir, la expresión:

$$\varphi = \sum e_i^2 = \sum (Y_i - a - bX1_i - cX2_i)^2$$

Estimación mínimo cuadrática de a

$$\frac{\partial \varphi}{\partial a} = -2\sum\left(Y_i - a - bX1_i - cX2_i\right) = 0$$

$$\sum Y_i - \sum a - \sum bX1_i - \sum cX2_i = 0$$

$$\sum Y_i = na + b\sum X1_i + c\sum X2_i \quad (1)$$

conocida como **primera ecuación normal**.

Estimación mínimo cuadrática de b

$$\frac{\partial \varphi}{\partial b} = -2\sum\left[\left(Y_i - a - bX1_i - cX2_i\right)X1_i\right] = 0$$

$$\sum X1_i Y_i - \sum aX1_i - \sum bX1_i^2 - \sum cX1_i X2_i = 0$$

$$\sum X1_i Y_i = a\sum X1_i + b\sum X1_i^2 + c\sum X1_i X2_i \quad (2)$$

conocida como **segunda ecuación normal**.

Estimación mínimo cuadrática de c

$$\frac{\partial \varphi}{\partial c} = -2\sum\left[\left(Y_i - a - bX1_i - cX2_i\right)X2_i\right] = 0$$

$$\sum X2_i Y_i - \sum aX2_i - \sum bX1_i X2_i - \sum cX2_i^2 = 0$$

$$\sum X2_i Y_i = a\sum X2_i + b\sum X1_i X2_i + c\sum X2_i^2 \quad (3)$$

conocida como **tercera ecuación normal**.

En resumen, para obtener las estimaciones mínimo cuadráticas de **a**, **b** y **c** debemos resolver el sistema de las tres ecuaciones normales con tres incógnitas (1), (2) y (3), es decir,

$$na + b\sum X1_i + c\sum X2_i = \sum Y_i$$

$$a\sum X1_i + b\sum X1_i^2 + c\sum X1_i X2_i = \sum X1_i Y_i$$

$$a\sum X2_i + b\sum X1_i X2_i + c\sum X2_i^2 = \sum X2_i Y_i$$

12.3 Ejemplo numérico.

Los datos que se muestran en la tabla 12.1 representan el consumo de agua (Y) expresado en litros por habitante, el precio del metro cúbico según el consumo realizado (X1) expresado en dólares y el número promedio de habitantes por domicilio (X2) en 9 cantones de la provincia de Manabí. Ajustemos a los datos una ecuación de regresión lineal múltiple.

TABLA 12.1 Datos para el ajuste de una ecuación lineal múltiple

Y	128	112	84	93	117	78	124	95	74
X1	106	93	63	69.8	97.1	58.5	102.9	71.3	55.5
X2	6	5	3	4	5	2	6	4	2

La tabla con las sumas, sumas de productos y sumas de cuadrados requeridas se muestran en la tabla 12.2.

TABLA 12.2 Sumas, sumas de productos y sumas de cuadrados

Y	**X1**	**X2**	**X1Y**	**X2Y**	**X1X2**	**$X1^2$**	**$X2^2$**	**Y^2**
128	106	6	13568.00	768	636	11236.00	36	16384
112	93	5	10416.00	560	465	8649.00	25	12544
84	63	3	5292.00	252	189	3969.00	9	7056
93	69.8	4	6491.40	372	279.2	4872.04	16	8649
117	97.1	5	11360.70	585	485.5	9428.41	25	13689
78	58.5	2	4563.00	156	117	3422.25	4	6084
124	102.9	6	12759.60	744	617.4	10588.41	36	15376
95	71.3	4	6773.50	380	285.2	5083.69	16	9025
74	55.5	2	4107.00	148	111	3080.25	4	5476
905	717.1	37	75331.2	3965	3185.3	60329.05	171	94283

Sustituyendo los totales obtenidos en las ecuaciones normales (1), (2) y (3) respectivamente tendremos el siguiente sistema de ecuaciones lineales:

$$9a + 717.1b + 37c = 905 \quad (1)$$
$$717.1a + 60329.05b + 3185.3c = 75331.2 \quad (2)$$
$$37a + 3185.3b + 171c = 3965 \quad (3)$$

Eliminemos "a" de las ecuaciones (1) y (2):

$$9a + 717.1b + 37c = 905 \quad \times -717.1$$
$$717.1a + 60329.05b + 3185.3c = 75331.2 \quad \times 9$$
$$28729.04b + 2135c = 29005.3 \quad (4)$$

Eliminemos "a" de las ecuaciones (1) y (3):

$$9a + 717.1b + 37c = 905 \quad \times -37$$
$$37a + 3185.3b + 171c = 3965 \quad \times 9$$
$$2135b + 170c = 2200 \quad (5)$$

Eliminando "b" entre (4) y (5):

$$28729.04b + 2135c = 29005.3 \quad \times -2135$$

2135b + 170c = 2200 x 28729.04

325711.8c = 1277572.5

$c = \frac{1277572.5}{325711.8}$ de donde **c = 3.92**

Sustituyendo c=3.92 en (5):

2135b = 2200 - 170(3.92) de donde **b = 0.72**

Sustituyendo b=0.72 y c=3.92 en (1):

9a = 905 – 717.1(0.72) – 37(3.92) de donde **a = 27.07**

La ecuación de regresión lineal múltiple queda entonces:

Y = 27.07 + 0.72X1 + 3.92X2

Procedamos a continuación a realizar la *Prueba de Hipótesis Global* de los coeficientes de regresión. Las hipótesis nula y alternativa son:

H_0: $\beta_2 = \beta_3 = 0$

H_1: Los parámetros no son iguales a cero

Para tomar la decisión de rechazar o no rechazar la hipótesis nula procedamos a realizar el análisis de regresión.

$$SCC_{TOTAL} = (128)^2 + (112)^2 + ... + (74)^2 - \frac{(128+112+.....+74)^2}{9}$$

$$SCC_{TOTAL} = 94283 - 91002.78 = 3280.22$$

Calculemos la SCC_{ERROR} a través del cálculo de los errores en la estimación y su correspondiente suma de cuadrados como se aprecia en la tabla 12.3.

TABLA 12.3 Cálculo de la suma de cuadrados del error

Y	X1	X2	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
128	106	6	126.910	1.090	1.1881
112	93	5	113.630	-1.630	2.6569
84	**63**	3	84.190	-0.190	0.0361
93	69.8	4	93.006	-0.006	0
117	97.1	5	116.582	0.418	0.1747
78	58.5	2	77.030	0.970	0.9409
124	102.9	6	124.678	-0.678	0.4597
95	71.3	4	94.086	0.914	0.8354
74	55.5	2	74.870	-0.870	0.7569
905	717.1	37	904.982	0.018	7.0487

De la tabla anterior, redondeando a dos decimales, $SCC_{ERROR} = 7.05$ por tanto:

$SCC_{REG.} = 3280.22 - 7.05 = 3273.17$

Por los motivos ya explicados en el capítulo anterior, los grados de libertad de la regresión son dos. La tabla 12.4 muestra el análisis de regresión.

TABLA 12.4 Análisis de regresión

FUENTES DE VARIACION	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	8	3280.22			
REGRESION	2	3273.17	1636.59	1386.94	P<0.001
ERROR	6	7.05	1.18		

En este caso, *el error estándar múltiple de la estimación* será igual a:

$$S_{Y.X} = \sqrt{1.18} = \pm 1.09$$

Los coeficientes de determinación y de correlación múltiples son:

$$R^2 = \frac{3273.17}{3280.22} = 0.998 \approx 1$$

$$r = \sqrt{0.998} = 0.999 \approx 1$$

12.4 Pruebas de hipótesis individuales para los coeficientes de regresión.

En el numeral anterior hicimos la *Prueba de Hipótesis Global* de los coeficientes de regresión mediante las hipótesis:

$H_0: \beta_2 = \beta_3 = 0$

H_1: Los parámetros no son iguales a cero

En ese numeral mediante la técnica estadística de análisis de regresión rechazamos la hipótesis nula. Sin embargo, al ser la prueba de tipo global, no tenemos la certeza si ambos coeficientes son realmente diferentes de cero o si uno solo de ellos fue el causante del rechazo de la hipótesis nula.

Scheffé (1959) reportó que para una ecuación de regresión lineal múltiple con dos variables independientes, las varianzas de los parámetros del modelo viene dada por:

$$V(b) = \frac{m_{WW.\Omega}\sigma^2}{M}$$

$$V(c) = \frac{m_{ZZ.\Omega}\sigma^2}{M}$$

$$donde: M = \left(m_{ZZ;\Omega}\right)\left(m_{WW;\Omega}\right) - \left(m_{ZW;\Omega}\right)^2 \quad y$$

En consecuencia, los errores estándar de los coeficientes de regresión se calculan mediante las expresiones:

$$E.E.(b) = \sqrt{\frac{SCC_{X2} CM_{ERROR}}{SCC_{X1} SCC_{X2} - SPC_{X1X2}^2}}$$

$$E.E.(c) = \sqrt{\frac{SCC_{X1} CM_{ERROR}}{SCC_{X1} SCC_{X2} - SPC_{X1X2}^2}}$$

En el ejemplo que estamos desarrollando:

$$\sum X1 = 717.1 \quad \sum X2 = 37 \quad \sum (X1)^2 = 60329.05 \quad \sum (X2)^2 = 171$$

$$\sum X1X2 = 3185.3$$

$$SCC_{X1} = 60329.05 - \frac{(717.1)^2}{9} = 3192.12$$

$$SCC_{X2} = 171 - \frac{(37)^2}{9} = 18.89$$

$$SPC_{X1X2} = 3185.3 - \frac{(717.1)(37)}{9} = 237.22$$

$$M = (3192.12)(18.89) - (237.22)^2 = 4025.82$$

$CM_{ERROR} = 1.18$

$$V(b) = \frac{(18.89)(1.18)}{4025.82} = 0.006$$

$$E.E.(b) = \sqrt{V(b)} = \sqrt{0.006} = \pm 0.08$$

$$V(c) = \frac{(3192.12)(1.18)}{4025.82} = 0.94$$

$$E.E.(c) = \sqrt{V(c)} = \sqrt{0.94} = \pm 0.97$$

Procedamos a realizar las pruebas de hipótesis individuales.

Para una prueba de dos colas, los valores de los percentiles t de Student para 6

grados de libertad son:

2.447 para un nivel de significación del 5%

3.707 para un nivel de significación del 1%

5.959 para un nivel de significación del 0.1%

Para el parámetro β_2

H_0: $\beta_2 = 0$

H_1: $\beta_2 \neq 0$

$$|T| = \left|\frac{b}{E.E.(b)}\right| = \left|\frac{0.72}{0.08}\right| = 9$$

y como 9 > 5.959 se rechaza la hipótesis nula con un nivel de significación del 0.1%, y por tanto, $\beta_2 \neq 0$.

Para el parámetro β_3

H_0: $\beta_3 = 0$

H_1: $\beta_3 \neq 0$

$$|T| = \left|\frac{c}{E.E.(c)}\right| = \left|\frac{3.92}{0.97}\right| = 4.04$$

y como 4.04 > 3.707 pero no mayor que 5.959, se rechaza la hipótesis nula con un nivel de significación del 1%.

Ambos coeficientes de regresión son significativamente diferentes de cero, lo cual corrobora la validez de la ecuación de regresión lineal múltiple obtenida.

12.5 Intervalos de confianza para β_2 y β_3.

Para el parámetro β_2

Un intervalo de confianza para β_2 con un nivel de confiabilidad del 95% viene dado por:

$$0.72 - (2.447)(0.08) < \beta_2 < 0.72 + (2.447)(0.08)$$

$0.52 < \beta_2 < 0.92$

Para el parámetro β_3

$$3.92 - (2.447)(0.97) < \beta_3 < 3.92 + (2.447)(0.97)$$

$1.55 < \beta_3 < 6.29$

$$m_{ZZ;\Omega} = SCC_{X1}\ ,\ m_{WW;\Omega} = SCC_{X2}\ ,\ m_{ZW;\Omega} = SPC_{X1X2}\ y\ \sigma^2 = CM_{ERROR}$$

12.6 Coeficiente de correlación parcial.

Al iniciar el presente capítulo hicimos referencia a una hipotética investigación en la cual el gerente de un establecimiento al aire libre estableció una regresión entre la asistencia de personas a su establecimiento (X) y el consumo (Y) de una bebida caliente a base de chocolate de alta calidad. Señalamos que en dicha regresión el gerente había obtenido resultados que mostraban un coeficiente de correlación negativo, lo cual evidenciaba una contradictoria relación que establecía que a mayor asistencia de público, al establecimiento correspondía un menor consumo de dicha bebida.

Pero en realidad el resultado contradictorio obtenido por el gerente tiene una explicación sencilla: *Al planificar la investigación, olvidó incluir la temperatura como segunda variable independiente*. Evidentemente cuando hay temperaturas altas, la asistencia de público al establecimiento también lo es, pero el consumo de una bebida caliente se ve disminuido. Por el contrario, cuando la temperatura es baja la asistencia de público también es baja, pero los pocos que asisten tienen preferencia por consumir dicha bebida. Es decir, las variables que debieron ser tomadas en cuenta en la regresión eran el consumo de la bebida caliente (Y), la asistencia de público al establecimiento (X1) y *la temperatura (X2).* Consideremos que los datos correspondientes a la investigación fueron los siguientes:

TABLA 12.5 Consumo de bebida, asistencia de público y temperatura

Y	12	30	78	10	70	31	13	82	75	12	34	20
X1	180	35	90	210	82	30	185	95	90	205	39	23
X2	27	11	14	30	13	10	28	15	17	29	11	9

donde Y son las unidades de bebida caliente vendidas, X1 la cantidad de personas asistentes al establecimiento y X2 la temperatura medida en grados centígrados.

Calculemos las sumas, sumas de cuadrados y sumas de productos requeridas para desarrollar el análisis que nos interesa. Estos cálculos se muestran en la tabla 12.6.

TABLA 12.6 Sumas, sumas de cuadrados y de productos

Y	X1	X2	X1Y	X2Y	X1X2	$X1^2$	$X2^2$	Y^2
12	180	27	2160	324	4860	32400	729	144
30	35	11	1050	330	385	1225	121	900
78	90	14	7020	1092	1260	8100	196	6084
10	210	30	2100	300	6300	44100	900	100
70	82	13	5740	910	1066	6724	169	4900
31	30	10	930	310	300	900	100	961
13	185	28	2405	364	5180	34225	784	169
82	95	15	7790	1230	1425	9025	225	6724
75	90	17	6750	1275	1530	8100	289	5625
12	205	29	2460	348	5945	42025	841	144
34	39	11	1326	374	429	1521	121	1156
20	23	9	460	180	207	529	81	400
467	1264	214	40191	7037	28887	188874	4556	27307

Con los resultados de la tabla 12.6, calculemos en primer término el coeficiente de correlación lineal entre la asistencia de público al establecimiento (X1) y el consumo de bebida caliente (Y):

$$SPC_{X1Y} = 40191 - \frac{(1264)(467)}{12} = -8999.67$$

$$SCC_{X1} = 188874 - \frac{(1264)^2}{12} = 55732.67$$

$$SCC_{Y} = 27307 - \frac{(467)^2}{12} = 9132.92$$

$$r = \frac{-8999.67}{\sqrt{(55732.67)(9132.92)}} = -0.40$$

En efecto, como podemos apreciar el coeficiente de correlación es *negativo*, debido a que no se ha tomado en cuenta el efecto de la temperatura sobre las ventas de bebida caliente.

Para resolver esta aparente contradicción calculemos el llamado *coeficiente de correlación parcial* entre Y y X1, el cual mide la relación entre X1 y Y eliminando el efecto lineal entre X2 y Y.

El procedimiento para calcular este coeficiente es el siguiente:

1. Determinamos la ecuación de regresión lineal entre la variable X2 y la va-

riable Y obteniendo los residuos correspondientes (Y^*). De esta manera se elimina la influencia de X2 sobre Y.

2. Obtenemos la regresión entre las variables X2 y X1 tomando a X1 como variable dependiente. A la ecuación resultante se le determinan sus residuos ($X1^*$). De esta manera se elimina la influencia de X2 sobre X1.
3. Calculamos el coeficiente de correlación entre los residuos Y^* y $X1^*$ el cual representaremos como $r_{YX1.X2}$ y llamaremos *coeficiente de correlación parcial* entre la variable X1 y la variable Y.

Regresión lineal de X2 vs Y

$$b = \frac{SPC_{X2Y}}{SCC_{X2}} = \frac{7037 - \frac{(214)(467)}{12}}{4556 - \frac{(214)^2}{12}} = \frac{-1291.17}{739.67} = -1.75$$

$$a = \overline{Y} - b\overline{X}2 = \frac{467}{12} - (-1.75)\frac{214}{12} = 70.13$$

La ecuación de regresión lineal simple obtenida es Y = 70.13-1.75X2.

Regresión lineal de X2 vs X1

$$b = \frac{SPC_{X2X1}}{SCC_{X2}} = \frac{28887 - \frac{(214)(1264)}{12}}{4556 - \frac{(214)^2}{12}} = \frac{6345.67}{739.67} = 8.58$$

$$a = \overline{X}1 - b\overline{X}2 = \frac{1264}{12} - 8.58\frac{214}{12} = -47.68$$

La ecuación de regresión lineal simple obtenida es X1 = -47.68 +8.58 X2.

Las tablas 12.7 y 12.8 muestran los residuos de ambas ecuaciones. Calculemos el coeficiente de correlación entre los residuos Y^* y $X1^*$**.** Usted podrá comprobar que:

$$SPC_{X1^*Y^*} = 2077.36$$
$$SCC_{X1^*} = 1292.62$$
$$SCC_{Y^*} = 6879.06$$

$$r_{YX1.X2} = \frac{2077.36}{\sqrt{(1292.62)(6879.06)}} = \frac{2077.36}{2981.95} = 0.70$$

TABLA 12.7 Cálculo de residuos. Regresión X2 vs Y

Y	X2	$\hat{Y}$	$Y^* = Y - \hat{Y}$
12	27	22,88	-10,88
30	11	50,88	-20,88
78	14	45,63	32,37
10	30	17,63	-7,63
70	13	47,38	22,62
31	10	52,63	-21,63
13	28	21,13	-8,13
82	15	43,88	38,12
75	17	40,38	34,62
12	29	19,38	-7,38
34	11	50,88	-16,88
20	9	54,38	-34,38
		SUMA	-0.06

TABLA 12.8 Cálculo de residuos. Regresión X2 vs X1

X1	X2	$\hat{X}1$	$X1^* = X1 - \hat{X}1$
180	27	183,98	-3,98
35	11	46,7	-11,7
90	14	72,44	17,56
210	30	209,72	0,28
82	13	63,86	18,14
30	10	38,12	-8,12
185	28	192,56	-7,56
95	15	81,02	13,98
90	17	98,18	-8,18
205	29	201,14	3,86
39	11	46,7	-7,7
23	9	29,54	-6,54
		SUMA	0.04

Observe dos importantes diferencias entre el coeficiente de correlación parcial recién calculado (0.70) y el obtenido sin tomar en cuenta el efecto que produce la temperatura sobre las ventas de bebida caliente (-0.40).

1. Como era de esperarse el coeficiente de correlación parcial es positivo, es decir, a mayor asistencia de público al establecimiento se produce un mayor consumo de bebida caliente, siempre que eliminemos el efecto lineal de la temperatura sobre las ventas de la dicha bebida.
2. Se evidencia un mayor grado de relación entre las dos variables, pues el

coeficiente de correlación parcial se incrementó de -0.40 a 0.70.

12.7 Regresión cuadrática simple.

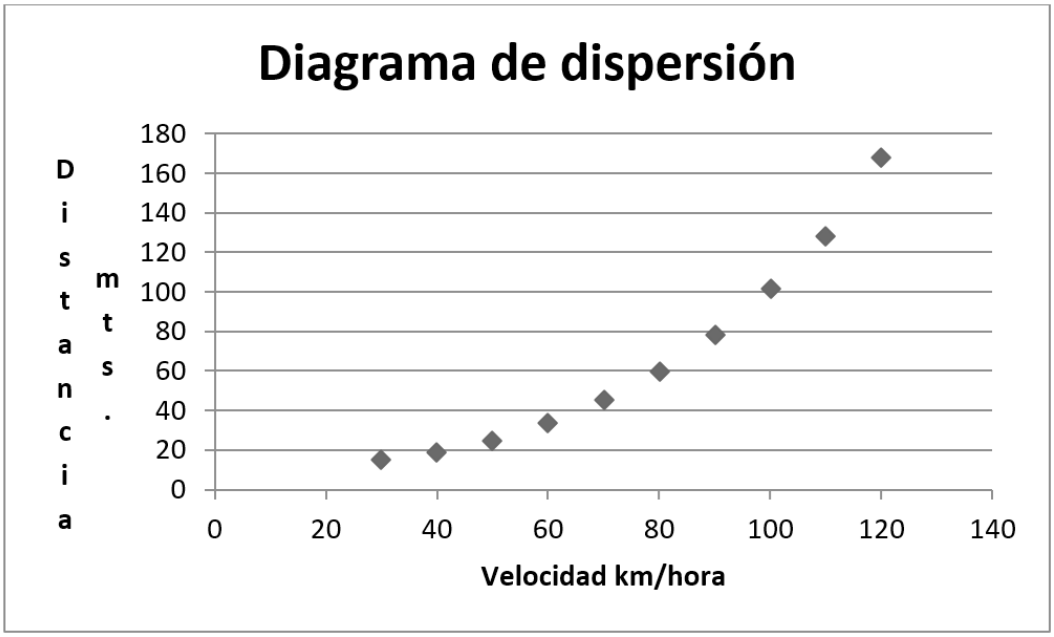
Se realizó un estudio con el objetivo de encontrar la relación existente entre la distancia en metros de frenado de un automóvil (Y) y la velocidad (X) en km/hora a la cual viajaba dicho vehículo.

Los resultados de la investigación se muestran en la tabla 12.9:

TABLA 12.9 Resultados de la investigación

X	30	40	50	60	70	80	90	100	110	120
Y	15	19	25	34	46	60	78	101	128	168

FIGURA 12.1 Diagrama de dispersión



El diagrama de dispersión muestra una tendencia que puede ser representada mediante una ecuación exponencial de la forma $Y = \beta_1 e^{\beta_2 X} + e_i$ o también mediante una ecuación cuadrática del tipo $Y = \beta_1 + \beta_2 X + \beta_3 X^2 + e_i$.

Ya estudiamos en párrafos anteriores el ajuste de una ecuación de regresión exponencial por lo que a continuación nos dedicaremos a realizar el ajuste de la regresión cuadrática.

Si comparamos el modelo cuadrático con el correspondiente a una regresión lineal múltiple $Y = \beta_1 + \beta_2 X1 + \beta_3 X2 + e_i$, podemos concluir que ambos modelos son iguales

si consideramos que $X1 = X \;\; y \;\; X2 = X^2$

De esta manera, las ecuaciones normales para ajustar una ecuación de regresión cuadrática son las siguientes:

$$\sum Y_i = na + b\sum X_i + c\sum X_i^2$$

$$\sum X_i Y_i = a\sum X_i + b\sum X_i^2 + c\sum X_i^3$$

$$\sum X_i^2 Y_i = a\sum X_i^2 + b\sum X_i^3 + c\sum X_i^4$$

Calculemos las sumas, sumas de productos y sumas de cuadrados necesarias para realizar el ajuste de la ecuación cuadrática tal como se muestra en la tabla 12.10.

TABLA 12.10 Cálculos requeridos para el ajuste de la ecuación cuadrática

X	Y	X^2	XY	X^2Y	X^3	X^4	Y^2
30	15	900	450	13500	27000	810000	225
40	19	1600	760	30400	64000	2560000	361
50	25	2500	1250	62500	125000	6250000	625
60	34	3600	2040	122400	216000	12960000	1156
70	46	4900	3220	225400	343000	24010000	2116
80	60	6400	4800	384000	512000	40960000	3600
90	78	8100	7020	631800	729000	65610000	6084
100	101	10000	10100	1010000	1000000	100000000	10201
110	128	12100	14080	1548800	1331000	146410000	16384
120	168	14400	20160	2419200	1728000	207360000	28224
750	674	64500	63880	6448000	6075000	606930000	68976

Sustituyendo los totales de la tabla anterior en las ecuaciones normales tenemos:

$$10a + 750b + 64500c = 674 \quad (1)$$

$$750a + 64500b + 6075000c = 63880 \quad (2)$$

$$64500a + 6075000b + 606930000c = 6448000 \quad (3)$$

Eliminemos "a" de las ecuaciones (1) y (2):

$$10a + 750b + 64500c = 674 \quad \text{x (-75)}$$

$$750a + 64500b + 6075000c = 63880$$

$$8250b + 1237500c = 13330 \quad (4)$$

Eliminemos "a" de las ecuaciones (1) y (3):

$$10a + 750b + 64500c = 674 \quad \text{x (-6450)}$$

$$64500a + 6075000b + 606930000c = 6448000$$

1237500b + 190905000c = 2100700 (5)

Eliminando "b" entre (4) y (5):

8250b + 1237500c = 13330 x (-150)

1237500b + 190905000c = 2100700

5280000c = 101200

$$c = \frac{101200}{5280000} = 0.02$$

c = 0.019

Sustituyendo c = 0.019 en (5):

1237500b + 190905000(0.019) = 2100700

1237500b = -1526495 $b = \frac{-1526495}{1237500} = -1.23$ **b = -1.23**

Sustituyendo b = -1.23 y c = 0.019 en (1):

10a + 750b + 64500c = 674 10a = 674 – 750(-1.23) – 64500(0.019)

a = 37.1

La ecuación de regresión cuadrática queda entonces:

$Y = 37.1 - 1.23X + 0.019X^2$

A continuación realicemos la *Prueba de Hipótesis Global* de los coeficientes de regresión:

H_0: $\beta_2 = \beta_3 = 0$

H_1: Los parámetros no son iguales a cero

Para ello se requiere, entre otros cálculos, determinar la suma de los cuadrados de los residuos, es decir,

$$SCC_{ERROR} = \sum (Y_i - \hat{Y})^2$$

En la tabla 12.11 puede observar los cálculos requeridos para obtener esta suma de cuadrados.

TABLA 12.11 Cálculos requeridos para obtener la SCC_{ERROR}

X	Y	$\hat{Y}$	Y- $\hat{Y}$	(Y- $\hat{Y}$)^2
30	15	17.30	-2.30	5.29
40	19	18.30	0.70	0.49
50	25	23.10	1.90	3.61
60	34	31.70	2.30	5.29
70	46	44.10	1.90	3.61
80	60	60.30	-0.30	0.09
90	78	80.30	-2.30	5.29
100	101	104.10	-3.10	9.61
110	128	131.70	-3.70	13.69
120	168	163.10	4.90	24.01
				70.98

Por tanto, $SCC_{ERROR} = 70.98$

$$SCC_{TOTALl} = 68976 - \frac{(674)^2}{10} = 23548.40$$

$$SCC_{REG.} = 23548.40 - 70.98 = 23477.42$$

La tabla 12.12 muestra el análisis de regresión obtenido.

TABLA 12.12 Análisis de regresión

FUENTES DE VARIACIÓN	**G.L.**	**S.C.**	**C.M.**	**F**	**SIGN.**
TOTAL	9	23548.40			
REGRESIÓN	2	23477.42	11738.71	1157.66	P<0.001
ERROR	7	70.98	10.14		

El error estándar múltiple de la estimación será igual a:

$$S_{Y.X} = \sqrt{10.14} = \pm 3.18$$

y los coeficientes de determinación y de correlación:

$$R^2 = \frac{23477.42}{23548.40} = 0.997$$

$$r = \sqrt{0.997} = 0.998$$

Es decir, el 99.7% de la variación en la variable dependiente Y está determinado

por su relación cuadrática simple con las variables independientes X y X^2.

Procedamos a realizar las pruebas de hipótesis individuales de cada coeficiente de regresión, para lo cual debemos calcular el error estándar de cada uno de dichos coeficientes:

$$SCC_X = 64500 - \frac{(750)^2}{10} = 8250$$

$$SCC_{X^2} = 606930000 - \frac{(64500)^2}{10} = 190905000$$

$$SPC_{XX^2} = 6075000 - \frac{(750)(64500)}{10} = 1237500$$

M = (8250) (190905000) – $(1237500)^2$ = 43560000000

$CM_{ERROR} = 10.14$ por tanto:

$$V(b) = \frac{(190905000)(10.14)}{43560000000} = 0.04$$

$$E.E.(b) = \sqrt{V(b)} = \sqrt{0.04} = \pm 0.2$$

$$V(c) = \frac{(8250)(10.14)}{43560000000} = 0.00000192$$

$$E.E.(c) = \sqrt{V(c)} = \sqrt{0.00000192} = \pm 0.0014$$

Para una prueba de dos colas, los valores de los percentiles t de Student para 7 grados de libertad son:

2.365 para un nivel de significación del 5%

3.499 para un nivel de significación del 1%

5.405 para un nivel de significación del 0.1%

Para el parámetro β_2

H_0: $\beta_2 = 0$

H_1: $\beta_2 \neq 0$

$$|T| = \left|\frac{-1.23}{0.2}\right| = 6.15$$

y como 6.15 > 5.405 se rechaza la hipótesis nula con un nivel de significación

del 0.1%.

Para el parámetro β_3

H_0: $\beta_3 = 0$

H_1: $\beta_3 \neq 0$

$$|T| = \left|\frac{0.019}{0.0014}\right| = 13.57$$

y como 13.57 > 5.405 se rechaza la hipótesis nula con un nivel de significación del 0.1%.

Los resultados anteriores corroboran la validez de la ecuación de regresión obtenida ya que ambos coeficientes de regresión son significativamente diferentes de cero.

Obtengamos los intervalos de confianza para β_2 y β_3.

Para un nivel de significación del 5%, el percentil de la t de Student con 7 grados de libertad y una prueba de dos colas es 2.365.

Para el parámetro β_2

$$-1.23 - (2.365)(0.2) < \beta_2 < -1.23 + (2.365)(0.2)$$

$-1.7 < \beta_2 < -0.76$

Para el parámetro β_3

$$0.019 - (2.365)(0.0014) < \beta_2 < 0.019 + (2.365)(0.0014)$$

$0.016 < \beta_3 < 0.022$

Ejercicios del capítulo

12.1 La filial de la empresa Servientrega ubicada en la ciudad de Manta está interesada en conocer si existe una relación lineal múltiple entre las variables número de kilómetros recorridos, cantidad de encomiendas entregadas y tiempo de recorrido en horas. La siguiente tabla muestra los resultados del estudio realizado por la empresa.

Recorrido Km	215	120	195	205	100	187	155	121	185	174	180
No. encomiendas	4	4	3	2	2	2	4	3	3	2	3
Tiempo de recorrido	10.2	5.1	9.3	6.4	5.2	6.1	8.3	6.5	7.5	6.3	6.5

a. Determine la variable dependiente y las independientes.
b. Obtenga la ecuación de regresión lineal múltiple.
c. Desarrolle el análisis de regresión correspondiente.
d. Determine el error estándar de la estimación.

12.2 Los datos que aparecen a continuación representan los gastos en publicidad televisiva (en miles de dólares), en publicidad escrita (en miles de dólares) y los ingresos (en miles de dólares) de una determinada compañía.

P. Televisiva	5.2	2.3	4.3	2.7	2.7	3.9	2.6	3.4	3.1
P. Escrita	1.7	1.9	1.5	2.3	3.7	2.4	4.1	2.8	2.7
Ingresos	100	87	98	90	99	96	93	91	90

a. Determine la variable dependiente y las independientes.
b. Obtenga la ecuación de regresión lineal múltiple.
c. Desarrolle el análisis de regresión correspondiente.
d. Determine el error estándar de la estimación.

12.3 Con los datos del ejercicio 12.1:
a. Calcule los errores estándar de los coeficientes de regresión.
b. Determine el valor del coeficiente de determinación.
c. Obtenga el valor del coeficiente de correlación.

12.4 Con los datos del ejercicio 12.2:
a. a) Calcule los errores estándar de los coeficientes de regresión.
b. b) Determine el valor del coeficiente de determinación.
c. c) Obtenga el valor del coeficiente de correlación.

12.5 A continuación se describe el comportamiento de dos variables entre las que al parecer existe una relación cuadrática.

X	1	2	3	4	5	6	7	8	9	10
Y	0.23	8.36	9.97	11.54	15.47	12.21	10.48	9.17	7.43	1.06

a. Obtenga la ecuación de regresión cuadrática entre estas dos variables.
b. Desarrolle el análisis de regresión correspondiente.
c. Determine el error estándar de la estimación.

12.6 A continuación se describe el comportamiento de dos variables experimentales.

X	1	2	3	4	5	6	7	8
Y	4.36	6.48	7.93	10.25	8.12	6.19	5.36	3.18

a. Obtenga la ecuación de regresión cuadrática entre estas dos variables.
b. Desarrolle el análisis de regresión correspondiente.
c. Determine el error estándar de la estimación.

12.7 Con los datos del ejercicio 12.5:

a. Calcule los errores estándar de los coeficientes de regresión.
b. Determine el valor del coeficiente de determinación.
c. Obtenga el valor del coeficiente de correlación.

12.8 Con los datos del ejercicio 12.6:

a. Calcule los errores estándar de los coeficientes de regresión.
b. Determine el valor del coeficiente de determinación.
c. Obtenga el valor del coeficiente de correlación.

12.9 Se desarrolló una investigación con el objetivo de establecer la relación existente entre las calificaciones (Y) obtenidas en una asignatura por 11 estudiantes y su coeficiente intelectual (X1).

Con el objetivo de eliminar de la relación planteada el efecto de las horas dedicadas al estudio por parte de los estudiantes, ésta variable (X2) también fue incluida en el estudio. Los resultados alcanzados fueron los siguientes:

X1	85	120	104	100	125	132	140	118	93
X2	11	9	20	22	12	15	17	22	17
Y	2.5	4.5	5.5	5.9	5.7	9	9.5	4.1	3.8

Obtenga el coeficiente de correlación parcial entre la variable calificaciones y la variable coeficiente intelectual eliminando el efecto de las horas dedicadas al estudio.

12.10 La siguiente tabla muestra las calificaciones (Y) obtenidas por los 11 estudiantes del ejercicio anterior y el tiempo dedicado al estudio (X1) por los mismos. Obtenga el coeficiente de correlación parcial entre la variable calificaciones y la variable tiempo dedicado al estudio (X1) eliminando el efecto del coeficiente intelectual (X2) sobre las calificaciones.

X1	11	9	20	22	12	15	17	22	17
X2	85	120	104	100	125	132	140	118	93
Y	2.5	4.5	5.5	5.9	5.7	9	9.5	4.1	3.8

Capítulo 13

Métodos no paramétricos.
Aplicaciones de la Ji-Cuadrada

El problema

El propietario de una dulcería está interesado en conocer si entre sus clientes existen preferencias en cuanto al sabor de los pasteles que el produce. ¿Existe algún método estadístico que le permita al dueño de la dulcería satisfacer el interés que tiene?

13.1 Introducción.

En los Capítulos 7, 8 y 9 en los que estudiamos los temas correspondientes a estimación e intervalos de confianza, prueba de hipótesis para una sola muestra, y prueba de hipótesis para dos muestras, fue necesario suponer cómo una hipótesis de base que las poblaciones involucradas seguían una distribución normal. Sin embargo, hay pruebas estadísticas conocidas como *no paramétricas* en las cuales no resulta necesario hacer suposiciones acerca de la forma en que se distribuye la población.

Por otra parte, cuando los datos que deben ser procesados están expresados en escala de medición nominal, las pruebas estadísticas correspondientes tienen un carácter particular y utilizan un estadístico de prueba no estudiado hasta el momento, el cual es conocido como *Ji – Cuadrada*. El presente capítulo está dedicado al estudio de la importantísima distribución *Ji – Cuadrada* la cual es ampliamente utilizada en pruebas estadísticas que involucran a datos medidos en escala nominal.

13.2 Distribución Ji – Cuadrada.

Si se elige una muestra de tamaño n con varianza muestral S^2 de una población normal con varianza poblacional σ^2, entonces el estadístico $\frac{(n-1)S^2}{\sigma^2}$ tiene una distribución muestral conocida como *Distribución Ji – Cuadrada con n-1 grados de libertad*, la cual se denota como $\chi^2(n-1)$.

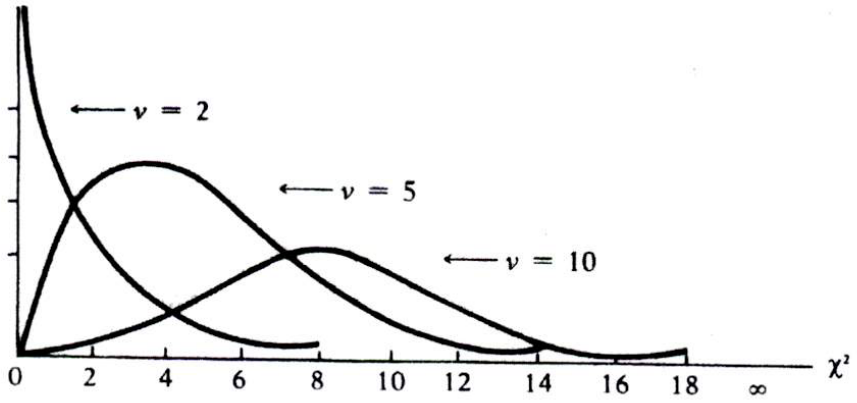
13.2.1 Propiedades de la distribución Ji – Cuadrada.

- Los valores de χ^2 son siempre no negativos.
- El área bajo una curva Ji – Cuadrada y sobre el eje horizontal es igual a 1.
- La distribución χ^2 está sesgada a la derecha.

- Cuando n > 2, la media de una distribución χ^2 es n-1 y su varianza 2(n-1).
- La moda de una distribución χ^2 es n-3.
- La forma de una distribución χ^2 depende de sus grados de libertad, por tanto, existe una familia de estas distribuciones.

Una característica importante de la distribución Ji – Cuadrada radica en que a medida que el número de grados de libertad aumenta, la distribución se va aproximando a la distribución normal, tal como se observa en el siguiente gráfico, donde ν representa los grados de libertad:

FIGURA 13.1 Gráfico de la distribución χ^2 para diferentes valores de ν



13.3 Prueba de bondad de ajuste.

Se entiende por *bondad de ajuste* a la posible discrepancia existente entre una distribución observada y la correspondiente distribución teórica.

La *prueba de bondad de ajuste* tiene por objetivo indicar en qué medida las diferencias entre ambas son significativas o simplemente son debidas al azar.

A través de dos ejemplos procederemos a describir la prueba de hipótesis de bondad de ajuste, el primero cuando las frecuencias esperadas son iguales y el segundo cuando éstas son desiguales.

13.3.1 Frecuencias esperadas iguales.

Una dulcería produce pasteles de diferentes sabores, y el propietario de la misma desea conocer si existe por parte de sus clientes diferencias en cuanto a la preferencia de dichos sabores con el objetivo de producir una mayor cantidad del pastel que más gusta.

Para ello, durante una jornada de trabajo ofreció pasteles de cinco sabores diferentes. Al concluir el día de trabajo la dulcería había vendido 370 pasteles. El volumen de venta de cada uno de los sabores se muestra a continuación:

TABLA 13.1 Volumen de venta de pasteles de distintos sabores

Sabores	**Ventas**
Chocolate blanco	85
Chocolate negro	93
Fresa	66
Coco	54
Vainilla	72
	370

Con un nivel de significación del 5%, ¿podemos concluir que existe una diferencia significativa en la preferencia de sabores por parte de los clientes de la dulcería?

Observe en primer lugar que los volúmenes de venta son medidos mediante una *escala nominal*, ya que entre los diferentes sabores de los pasteles no existe un orden natural. Por otra parte, por ser 370 el número de pasteles, resulta razonable pensar que cada sabor tenga un volumen de venta *esperado* igual a $\frac{370}{5} = 74.$ De esta forma, las frecuencias *observadas y esperadas* de las ventas de pasteles son las que se muestran en la tabla 13.2.

TABLA 13.2 Frecuencias observadas y esperadas de ventas de pasteles

Sabores	**Frecuencias observadas f_0**	**Frecuencias esperadas fe**
Chocolate blanco	85	74
Chocolate negro	93	74
Fresa	66	74
Coco	54	74
Vainilla	72	74
	370	370

Para desarrollar la prueba de hipótesis de la bondad de ajuste, apliquemos el procedimiento de *los 5 pasos* establecido en el Capítulo 8, el cual se resume de la siguiente forma:

Paso 1	Se formulan las hipótesis nula y alternativa
Paso 2	Se establece el nivel de significación
Paso 3	Se identifica la distribución a utilizar
Paso 4	En dependencia de las hipótesis se escoge la regla de decisión adecuada
Paso 5	En base a la muestra tomada se decide o no rechazar la hipótesis nula

Paso 1: Se formulan las hipótesis nula y alternativa.

H_0: No hay diferencias entre las frecuencias observadas y esperadas.

H_1: Hay diferencias entre las frecuencias observadas y esperadas.

Paso 2: Se establece el nivel de significación.

En el ejercicio se establece que el nivel de significación es del 5%.

Paso 3: Se identifica la distribución a utilizar.

La distribución es la Ji – Cuadrada con k-1 grados de libertad $\chi^2(k-1)$ donde k es el número de *categorías,* en este caso sabores.

El estadístico de prueba es $\sum\left[\frac{(f_0-f_e)^2}{f_e}\right]$

Paso 4: En dependencia de las hipótesis se escoge la regla de decisión adecuada.

En este caso, la regla de decisión queda como sigue:

Rechazar H_0 si $\sum\left[\frac{(f_0-f_e)^2}{f_e}\right] > \chi^2(k-1)$

No rechazar H_0 si $\sum\left[\frac{(f_0-f_e)^2}{f_e}\right] \leq \chi^2(k-1)$

Obtengamos el valor crítico de la distribución Ji – Cuadrada con 4 grados de libertad (5 sabores - 1) y el valor del estadístico de prueba correspondiente.

- **Valor crítico de la distribución Ji – Cuadrada.**

En la siguiente página se muestra un segmento de la **TABLA T.5** del Anexo donde se pueden apreciar los valores críticos de la distribución Ji – Cuadrada.

En esta tabla se ha marcado en color rojo el valor crítico de la distribución para 4 grados de libertad y un nivel de significación del 5% (9.49), el cual se obtiene interceptando la fila 4 con la columna correspondiente al valor 0.05.

G.L.	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	0.001	0.001	0.001	0.004	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.51
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59

- **Estadístico de prueba.**

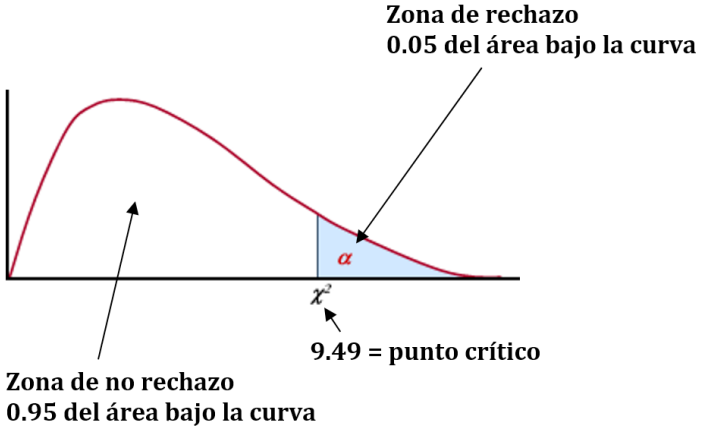
TABLA 13.3 Cálculos requeridos para hallar el estadístico de prueba

Sabores	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Chocolate blanco	85	74	11	121	1.64
Chocolate negro	93	74	19	361	4.88
Fresa	66	74	-8	64	0.86
Coco	54	74	-20	400	5.41
Vainilla	72	74	-2	4	0.05
	370	370	0		12.84

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula.

En la figura 13.2 se puede apreciar que como 12.84 es mayor que 9.49 se rechaza la hipótesis nula, y por tanto, con un nivel de confiabilidad del 95%, existen diferencias significativas entre las frecuencias observadas y esperadas, es decir, existen preferencias por parte de los clientes al momento de elegir el sabor de los pasteles.

FIGURA 13.2 Ubicación de las zonas de aceptación y rechazo



13.3.2 Frecuencias esperadas desiguales.

El Instituto Nacional de Estadísticas y Censo (INEC) en su censo de población y vivienda del 28 de Noviembre del 2010, estableció que los grupos étnicos existentes en el Ecuador y el porcentaje de personas que pertenecen a los mismos son los siguientes:

TABLA 13.4 Grupos étnicos y su composición

Grupos étnicos	**Porcentajes**
Mestizos	71.9
Blancos	6.1
Indígenas	7
Montubios	7.4
Afrodescendientes	7.2
Otros	0.4

Con la finalidad de estudiar si en la ciudad de Manta los grupos étnicos tienen una estructura semejante a la del país, se extrajo una muestra de 1000 personas en la que se obtuvo los resultados que se muestran en la siguiente tabla:

TABLA 13.5 Frecuencia observada por grupo étnico en una muestra

Grupos étnicos	**Cantidad, f_o**
Mestizos	658
Blancos	168
Indígenas	35
Montubios	60
Afrodescendientes	72
Otros	7

Con un nivel de significación igual a 0,01, ¿podemos llegar a la conclusión que en la ciudad de Manta los grupos étnicos tienen una estructura diferente a la del país?

Paso 1: Se formulan las hipótesis nula y alternativa.

H_0: No hay diferencias entre la estructura étnica de Manta y la del país.

H_1: Hay diferencias entre la estructura étnica de Manta y la del país.

Paso 2: Se establece el nivel de significación.

En el ejercicio se establece que el nivel de significación es del 1%.

Paso 3: Se identifica la distribución a utilizar.

La distribución es la Ji – Cuadrada con 6-1 grados de libertad $\chi^2(5)$.

El estadístico de prueba es $\sum\left[\frac{(f_0 - f_e)^2}{f_e}\right]$

Paso 4: En dependencia de las hipótesis se escoge la regla de decisión adecuada.

En este caso, la regla de decisión queda como sigue:

Rechazar H_0 si $\sum\left[\frac{(f_0 - f_e)^2}{f_e}\right] > \chi^2(5)$

No rechazar H_0 si $\sum\left[\frac{(f_0 - f_e)^2}{f_e}\right] \leq \chi^2(5)$

Obtengamos el valor crítico de la distribución Ji – Cuadrada con 5 grados de libertad y el valor del estadístico de prueba correspondiente.

- **Valor crítico de la distribución Ji – Cuadrada.**

Consultando la **TABLA T.5** del Anexo A encontramos que el valor crítico de la distribución para 5 grados de libertad y un nivel de significación del 1% es 15.09.

- **Estadístico de prueba.**

TABLA 13.6 Cálculos requeridos para hallar el estadístico de prueba

Grupos étnicos	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Mestizos	658	719	-61	3721	5.18
Blancos	168	61	107	11449	187.69
Indígenas	35	70	-35	1225	17.50
Montubios	60	74	-14	196	2.65
Afrodescendentes	72	72	0	0	0.00
Otros	7	4	3	9	2.25
					215.27

Las frecuencias esperadas para cada uno de los grupos étnicos fueron calculadas según la siguiente fórmula:

$$f_e = \frac{tamaño\ de\ muestra\ x\ porcentaje\ del\ grupo\ obtenido\ en\ el\ censo}{100}$$

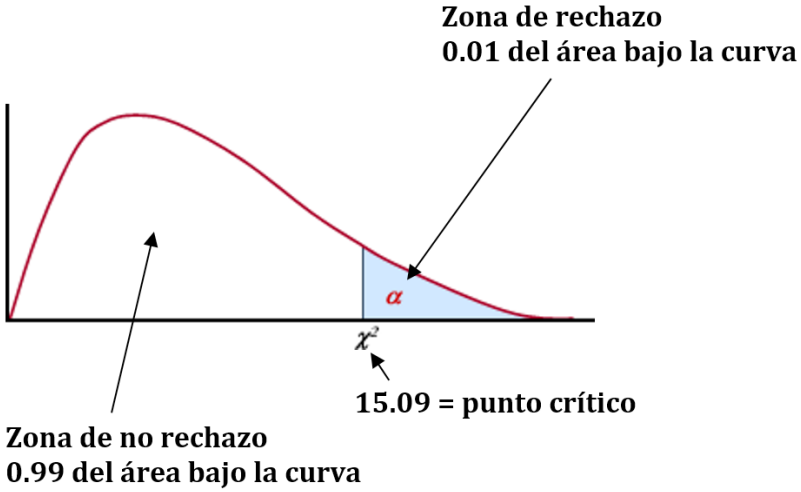
Por ejemplo, en el caso del grupo étnico Mestizos:

$$f_e = \frac{1000\,x\,71.9}{100} = \frac{71900}{100} = 719$$

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula.

En la figura 13.3 se puede apreciar que como 215.27 es mayor que 15.09 se rechaza la hipótesis nula, y por tanto, con un nivel de confiabilidad del 99% los grupos étnicos de Manta tienen una estructura diferente a la del país.

FIGURA 13.3 Ubicación de las zonas de aceptación y rechazo



13.4 Precaución al utilizar la prueba Ji – Cuadrada.

Cuando en una celda existe una frecuencia esperada con un valor muy pequeño, la aplicación de la Ji – Cuadrada puede conducirnos a un resultado equivocado, y por tanto, llegar a una conclusión también equivocada.

Esto se debe a que en el estadístico de prueba $\sum\left[\frac{(f_0 - f_e)^2}{f_e}\right]$ la frecuencia esperada se encuentra en el denominador, y por tanto, cuando f_e es pequeña el cociente resulta relativamente grande.

Existen dos reglas que han sido aceptadas de forma general. Estas dos reglas son:

1. Si en el análisis que estamos efectuando solo existen dos celdas, entonces la frecuencia esperada en cada una de ellas debe ser de al menos 5.

Por ejemplo, en una situación como la siguiente:

TABLA 13.7 Frecuencias observadas y esperadas en dos celdas

TURISTAS	f_o	f_e
Nacionales	148	146
Extranjeros	8	6

sería aplicable la utilización de la Ji – Cuadrada.

2. En los casos en que existen más de dos celdas no debemos utilizar la prueba Ji – Cuadrada si más del 20% de las mismas tienen frecuencias esperadas menores a 5.

Un ejemplo de este caso se muestra en la tabla 13.8, en la cual se puede apreciar con bastante claridad que no es recomendable el uso de la prueba ya que de ocho celdas en total, cuatro tienen frecuencias esperadas menores a 5, lo cual representa un 50%, evidentemente mayor al 20%.

TABLA 13.8 Frecuencias observadas y esperadas en más de dos celdas

CLASIFICACION I.M.C.	f_o	f_e
Peso Insuficiente	15	16
Normopeso	55	56
Sobrepeso Grado I	43	42
Sobrepeso Grado II	13	12
Obesidad I	5	2
Obesidad II	5	4
Obesidad III	4	1
Obesidad IV	2	1

Con el objetivo de evidenciar el porqué de la regla del 20%, calculemos el estadístico de prueba correspondiente a los datos de la tabla anterior y utilicemos un nivel de significación del 5% para la prueba de hipótesis.

TABLA 13.9 Cálculos requeridos para hallar el estadístico de prueba

CLASIFICACION I.M.C.	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Peso Insuficiente	15	16	-1	1	0.06
Normopeso	55	56	-1	1	0.02
Sobrepeso Grado I	43	42	1	1	0.02
Sobrepeso Grado II	13	12	1	1	0.08
Obesidad I	5	2	3	9	4.50
Obesidad II	5	4	1	1	0.25
Obesidad III	4	1	3	9	9.00
Obesidad IV	2	1	1	1	1.00
					14.93

La Ji – Cuadrada con 7 grados de libertad y un nivel de significación del 5%

es igual a 14.07. Como 14.93 es mayor que 14.07, se rechaza la hipótesis nula, y en consecuencia, existen evidencias de diferencias significativas entre las frecuencias observadas y esperadas.

Sin embargo, observe que el 98.79% del valor calculado del estadístico de prueba corresponde a las cuatro categorías de *obesidad*, es decir,

$$\left(\frac{4.50+0.25+9.00+1.00}{14.93}=\frac{14.75}{14.94}=0.9879\right) = 98.79\%$$

En casos similares a éste una solución podría ser, de ser esto posible, agrupar dos o más categorías en una sola de forma tal que se cumpla con la regla del 20%.

En el ejemplo que estamos desarrollando podríamos agrupar las cuatro categorías de *obesidad* en dos categorías.

Procediendo de esta manera los datos quedan tal y como se muestra en la tabla 13.10. En la tabla 13.11 se muestran los cálculos necesarios para determinar el valor del estadístico de prueba.

TABLA 13.10 Agrupamiento de categorías

CLASIFICACION I.M.C.	$\mathbf{f_o}$	$\mathbf{f_e}$
Peso Insuficiente	15	16
Normopeso	55	56
Sobrepeso Grado I	43	42
Sobrepeso Grado II	13	12
Obesidad I	10	6
Obesidad II	6	2

TABLA 13.11 Cálculos requeridos para hallar el estadístico de prueba

CLASIFICACION I.M.C.	$\mathbf{f_o}$	$\mathbf{f_e}$	$\mathbf{f_o - f_e}$	$\mathbf{(f_o - f_e)^2}$	$\mathbf{(f_o - f_e)^2/f_e}$
Peso Insuficiente	15	16	-1	1	0.06
Normopeso	55	56	-1	1	0.02
Sobrepeso Grado I	43	42	1	1	0.02
Sobrepeso Grado II	13	12	1	1	0.08
Obesidad I	10	6	4	16	2.67
Obesidad II	6	2	4	16	8.00
					10.85

La Ji – Cuadrada con 5 grados de libertad y un nivel de significación del 5% es igual a 11.07. Como 10.85 es menor que 11.07, no se rechaza la hipótesis nula, y en consecuencia, no existen evidencias de diferencias significativas entre las frecuencias observadas y esperadas.

Observe que el resultado que hemos obtenido al cumplir con la regla del 20%

es diferente al que obtuvimos cuando esta regla estaba siendo violada.

13.5 Tablas de contingencia.

Una muy importante aplicación de la distribución Ji – Cuadrada, es el caso en el cual estamos interesados en comprobar si dos variables medidas en una escala nominal son independientes o no. Una situación como ésta puede ser resuelta mediante la construcción de una *tabla de contingencia*, la cual en su forma más elemental es una tabla de doble entrada que tiene, en lo fundamental, dos objetivos básicos:

1. **Organizar la información proveniente de una variable medida en escala nominal y que tiene un carácter bidimensional, es decir, responde al efecto de dos factores distintos.**
2. **Evaluar si existe una relación de dependencia o no entre las dos variables bajo estudio.**

A continuación se muestra en la Tabla 13.12 la forma general de una tabla de contingencia.

TABLA 13.12 Forma general de una tabla de contingencia

		FACTOR A				
						TOTAL O
		A1	**A2**	**.**	**.**	**MARGINAL**
FACTOR B	**B1**	n_{11}	n_{12}	.	.	$n_{1.}$
	B2	n_{21}	n_{22}			$n_{2.}$
	.	.				.
	.	.				.
	TOTAL O MARGINAL	$n_{.1}$	$n_{.2}$	.	.	N

donde:

n_{ij} = número de observaciones que tienen el atributo i y j

$n_{i.}$ = número de observaciones que tienen el atributo i (marginal i)

$n._{j}$ = número de observaciones que tienen el atributo j (marginal j)

Para estudiar la posible independencia de los dos factores, Pearson propuso la utilización del ya conocido estadístico de prueba:

$$\frac{\sum_{1}^{h}\sum_{1}^{k}\left(n_{ij}-E_{ij}\right)^{2}}{E_{ij}}$$

donde $E_{ij}=\dfrac{n_{i.}x\ n_{.j}}{N}$

es la frecuencia esperada en condiciones de independencia.

El resultado de esta expresión es comparado con el valor de la distribución Ji –Cuadrada con (h-1) (k-1) grados de libertad y un determinado nivel de significación.

Si el estadístico de prueba es mayor que el valor crítico se rechaza la hipótesis nula y en caso contrario se no se rechaza, siendo tales hipótesis:

H_0: Los dos factores son independientes

H_1: Los dos factores no son independientes

Estudiemos un caso. Una empresa exportadora de latas de atún utiliza en el proceso de producción un total de cuatro máquinas en tres turnos diarios de trabajo. Los datos que se muestran a continuación representan el número de latas defectuosas en cada turno de trabajo y producida por cada una de las máquinas de la empresa.

TABLA 13.13 Número de latas defectuosas por máquina y por turno

TURNOS	**MÁQUINAS**				**TOTAL**
	M1	**M2**	**M3**	**M4**	
T1	5	3	1	2	11
T2	4	2	3	2	11
T3	3	4	3	5	15
TOTAL	12	9	7	9	37

Con un nivel de significación del 5%, ¿podemos asegurar que no existe una relación entre el turno de trabajo y la máquina utilizada con respecto a la producción de latas defectuosas?

Paso 1: Se formulan las hipótesis nula y alternativa.

H_0: No hay relación entre el turno de trabajo y la máquina utilizada.

H_1: Hay relación entre el turno de trabajo y la máquina utilizada.

Paso 2: Se establece el nivel de significación.

En el ejercicio se establece que el nivel de significación es del 5%.

Paso 3: Se identifica la distribución a utilizar.

La distribución es la Ji – Cuadrada con (4-1) (3-1) = 6 grados de libertad.

El estadístico de prueba es $\sum \frac{(f_0 - f_e)^2}{f_e}$

Paso 4: En dependencia de las hipótesis se escoge la regla de decisión adecuada.

En este caso, la regla de decisión queda como sigue:

Rechazar H_0 si $\sum \frac{(f_o - f_e)^2}{f_e} > \chi^2(6)$

No rechazar H_0 si $\sum \frac{(f_o - f_e)^2}{f_e} \leq \chi^2(6)$

Obtengamos el valor crítico de la distribución Ji – Cuadrada con 6 grados de libertad y el valor del estadístico de prueba correspondiente.

- **Valor crítico de la distribución Ji – Cuadrada.**

Consultando la **TABLA T.5** del Anexo A encontramos que el valor crítico de la distribución para 6 grados de libertad y un nivel de significación del 5% es 12.59.

- **Estadístico de prueba.**

TABLA 13.14 Cálculos requeridos para hallar el estadístico de prueba

TURNOS	MÁQUINAS								TOTAL	
	M1		M2		M3		M4			
	f_o	f_e	f_o	f_e	f_o	f_e	f_o	f_e	f_o	f_e
T1	5	3.568	3	2.676	1	2.081	2	2.676	11	11
T2	4	3.568	2	2.676	3	2.081	2	2.676	11	11
T3	3	4.865	4	3.649	3	2.838	5	3.649	15	15
TOTAL	12	12	9	9	7	7	9	9	37	37

Las frecuencias esperadas para cada combinación de máquina y turno fueron calculadas según la siguiente fórmula:

$$f_e = \frac{Total\ de\ la\ fila\ x\ Total\ de\ la\ columna}{Suma\ total}$$

Por ejemplo, la frecuencia esperada de la máquina 1 en el turno 1 es:

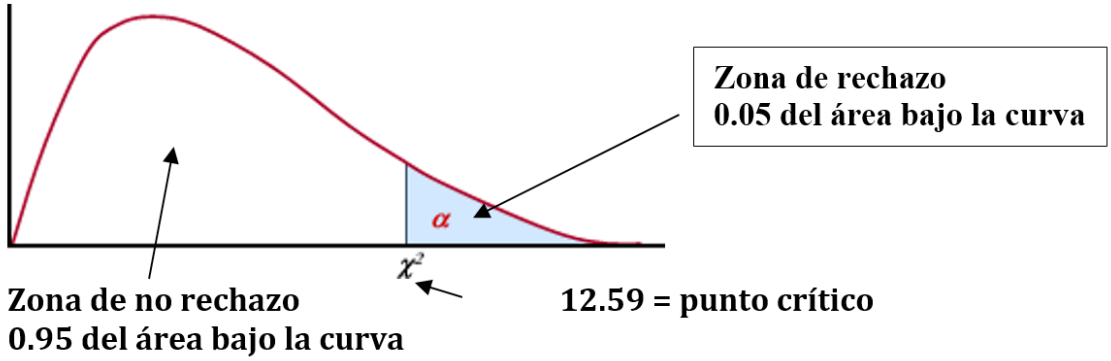
$$f_e = \frac{11x12}{37} = \frac{132}{37} = 3.568$$

$$\sum \frac{(f_o - f_e)^2}{f_e} = \frac{(5-3.568)^2}{3.568} + \frac{(3-2.676)^2}{2.676} + ... + \frac{(5-3.649)^2}{3.649} = 3.40$$

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula.

Como 3.40 es menor que 12.59 no se rechaza la hipótesis nula, y por tanto, con un nivel de confiabilidad del 95% no existe una relación entre el turno de trabajo y la máquina utilizada con respecto a la producción de latas defectuosas.

FIGURA 13.4 Ubicación de las zonas de aceptación y rechazo



Ejercicios del capítulo

13.1 Una encuesta realizada a un total de 2040 televidentes acerca de la preferencia por un determinado canal, dio los resultados que se muestran a continuación:

Canal	Frecuencias observadas
Ecuador Tv	425
RTS	298
TC	410
Manavisión	300
Gama	187
Ecuavisa	420

Considerando frecuencias esperadas iguales:

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el número de grados de libertad de la prueba
d. Determine el valor crítico correspondiente de la Ji – Cuadrada
e. Con un nivel de significación del 5%, ¿podemos concluir que existen diferencias significativas en cuanto a la preferencia de los televidentes por un determinado canal?

13.2 Un investigador observó durante un feriado largo el destino turístico elegido por 500 personas residentes en la ciudad de Quito. Los resultados del trabajo realizado fueron los siguientes:

Destino	Frecuencias observadas
Manta	120
Esmeraldas	100
Salinas	140
Galápagos	80
En la ciudad	60

Considerando frecuencias esperadas iguales:

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el número de grados de libertad de la prueba
d. Determine el valor crítico correspondiente de la Ji – Cuadrada
e. Con un nivel de significación del 1%, ¿hay evidencias que nos permitan concluir que existen diferencias significativas en cuanto a la preferencia de los turistas quiteños por un destino en particular?

13.3 La dosificación establecida para la elaboración de concreto con una determinada resistencia y por cada saco de cemento es de 6 baldes de arena, 8 baldes de ripio y 1.5 baldes de agua. El constructor de una obra reportó que en la preparación de 400 de estos concretos utilizó 160 baldes de arena, 203 baldes de ripio y 40 baldes de agua.

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el número de grados de libertad de la prueba
d. Determine el valor crítico correspondiente de la Ji – Cuadrada
e. Con un nivel de significación del 0.1%, ¿podemos llegar a la conclusión que el constructor no está preparando el concreto según la dosificación establecida?

13.4 El gerente de un supermercado asegura que las compras en su establecimiento son pagadas el 12% con cheques, el 20% con tarjetas de crédito, el 30% con tarjetas de débito y el 38% en efectivo. Una muestra de las vías utilizadas para hacer los pagos de sus compras de 600 clientes del supermercado, arrojó los resultados que se muestran en la siguiente tabla:

Forma de pago	**Frecuencias observadas**
Cheque	65
Tarjeta de Crédito	127
Tarjeta de Débito	190
Efectivo	218

a) Formule las hipótesis nula y alternativa
b) Obtenga el valor del estadístico de prueba
c) Determine el número de grados de libertad de la prueba
d) Determine el valor crítico correspondiente de la Ji – Cuadrada
e) Con un nivel de significación del 5%, ¿coincide la forma de pago de los clientes del supermercado con lo asegurado por su gerente?

13.5 El gerente del supermercado al que hicimos referencia en el ejercicio anterior asegura que la forma de pago de las compras realizadas por sus clientes está relacionada con su clase social, y para comprobarlo, extrajo una muestra de 1000 clientes la cual arrojó los siguientes resultados:

Clase social	**Forma de pago**			
	Cheque	**Tarjeta de crédito**	**Tarjeta de débito**	**Efectivo**
Alta	45	78	95	102
Media - Alta	14	66	114	145
Media	3	88	101	149

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el número de grados de libertad de la prueba
d. Determine el valor crítico correspondiente de la Ji – Cuadrada
e. Con un nivel de significación del 5%, ¿existe relación entre la clase social y la forma de pago de las compras realizadas en el supermercado?

13.6 La Universidad Laica Eloy Alfaro de Manabí está interesada en conocer si existe una relación entre la categoría docente de sus profesores y la opinión que tienen sus estudiantes con relación a la calidad de sus clases. Para ello realizó un estudio con 400 de sus profesores en el cual se obtuvo los siguientes resultados:

Opinión de los estudiantes	**Categoría Docente**		
	Principal	**Agregado**	**Auxiliar**
Excelente	45	38	23
Muy Bien	51	45	44
Bien	30	39	28
Regular	15	14	10
Mal	6	8	4

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el número de grados de libertad de la prueba
d. Determine el valor crítico correspondiente de la Ji – Cuadrada
e. Con un nivel de significación del 1%, ¿existe relación entre la categoría docente del profesor y la opinión que tienen sus estudiantes en cuanto a la calidad de sus clases?

Capítulo 14

Métodos no paramétricos. Análisis de datos ordenados

El problema

Los fabricantes del endulzante SPLENDA están interesados en estudiar si las personas que lo consumen detectan una diferencia en el sabor del café endulzado con este producto o con azúcar. Para ello seleccionó a 15 personas a las cuales les hizo beber una taza de café de una determinada marca endulzado con SPLENDA y una taza de café de la misma marca endulzado con azúcar, para posteriormente solicitarles expresaran su preferencia por el uno o por el otro. ¿Existe algún procedimiento estadístico que permita comprobar si hay diferencias en la preferencia de las personas al beber café endulzado con SPLENDA o con azúcar?

14.1 Introducción.

Los aspectos que trataremos en este capítulo representan una continuación de las pruebas de hipótesis aplicadas específicamente a datos *no paramétricos* y que fueron desarrolladas en el capítulo anterior, con la diferencia, que las respuestas deben estar medidas al menos en escala *ordinal,* es decir, las respuestas se clasifican de *alto a bajo*.

Durante el desarrollo del presente capítulo estudiaremos las pruebas no paramétricas más conocidas y utilizadas en la actualidad, aplicadas a datos ordenados.

Estas pruebas, entre otras, son:

1. **La prueba de los signos, utilizada en datos apareados.**
2. **Prueba de rangos con signo de Wilcoxon para muestras dependientes.**
3. **Prueba de Wilcoxon de la suma de rangos para muestras independientes.**
4. **Prueba U de Mann-Whitney y prueba de Kruskal-Wallis, las cuales generalizan el método de análisis de varianza permitiendo obviar la suposición de normalidad de las poblaciones.**
5. **Prueba de aleatoriedad de una muestra.**
6. **Correlación por rangos.**

14.2 La prueba de los signos.

La Facultad de Ciencias Económicas de una importante universidad del país diseñó un seminario de capacitación docente con el objetivo de intentar mejorar la

calidad de los sílabos elaborados por sus profesores en el año lectivo 2015 – 2016.

Para ello, le solicitó a la Comisión Académica de la Facultad calificara los sílabos entregados por los profesores en el año lectivo 2014 – 2015 como Excelente, Muy Bien, Bien y Regular. Después de impartido el seminario la misma comisión calificó de la misma forma los sílabos entregados en el año lectivo 2015 - 2016.

Los resultados obtenidos en el trabajo de la Comisión Académica se muestran en la tabla 14.1:

TABLA 14.1 Calificación de los sílabos

	AÑO LECTIVO	
PROFESORES	**2014 - 2015**	**2015 - 2016**
Antonio Delgado	Muy Bien	Excelente
Pedro Martínez	Regular	Muy Bien
Carmen Sánchez	Bien	Regular
Víctor González	Muy Bien	Muy Bien
Teresa Toala	Excelente	Muy Bien
Lidia Cedeño	Regular	Bien
Juan Salazar	Muy Bien	Excelente
Fausto Villamarín	Bien	Bien
César Gutiérrez	Excelente	Muy Bien
Martha Aguilar	Muy Bien	Excelente
Cecilia Díaz	Muy Bien	Excelente
Carlos Villamar	Regular	Muy Bien
Raúl Gómez	Excelente	Muy Bien
Saúl Vélez	Bien	Muy Bien
María de la Fuente	Muy Bien	Excelente

Recordamos que el interés de la Facultad es conocer si los profesores han mejorado la calidad de la elaboración de los sílabos a causa del seminario impartido.

En la tabla 14.2 se ha agregado los *signos de la diferencia* entre las calificaciones del año lectivo 2013 – 2014 y 2014 - 2015.

Un signo **+** indica una mejoría en la elaboración del sílabo, un signo **–** una pérdida de calidad en su elaboración y un 0 ni mejoría ni pérdida de calidad en la elaboración de dichos sílabos.

TABLA 14.2 Signos de la diferencia entre calificaciones

	AÑO LECTIVO		
PROFESORES	**2014 - 2015**	**2015 - 2016**	**SIGNO**
Antonio Delgado	Muy Bien	Excelente	+
Pedro Martínez	Regular	Muy Bien	+
Carmen Sánchez	Bien	Regular	-

	AÑO LECTIVO		
PROFESORES	**2014 - 2015**	**2015 - 2016**	**SIGNO**
Víctor González	Muy Bien	Muy Bien	0
Teresa Toala	Excelente	Muy Bien	-
Lidia Cedeño	Regular	Bien	+
Juan Salazar	Muy Bien	Excelente	+
Fausto Villamarín	Bien	Bien	0
César Gutiérrez	Excelente	Muy Bien	-
Martha Aguilar	Muy Bien	Excelente	+
Cecilia Díaz	Muy Bien	Excelente	+
Carlos Villamar	Regular	Muy Bien	+
Raúl Gómez	Excelente	Muy Bien	-
Saúl Vélez	Bien	Muy Bien	+
María de la Fuente	Muy Bien	Excelente	+

Para desarrollar la prueba de hipótesis, utilicemos el ya conocido método de los *cinco pasos*:

Paso 1: Se formulan las hipótesis nula y alternativa

Es lógico suponer que si el seminario no provocó una mejoría en la calidad de los sílabos, el porcentaje de profesores que lograron mejoras en dicha calidad fue menor o igual al 50%, por tanto las hipótesis nula y alternativa son:

$H_0 : \pi \leq 0.50$ No hay mejorías en la elaboración de los sílabos.

$H_1 : \pi > 0.50$ Sí hay mejorías en la elaboración de los sílabos.

Paso 2: Se establece el nivel de significación

Consideremos un nivel de confiabilidad del 95%, es decir, un nivel de significación igual a 0.05.

Paso 3: Se identifica la distribución a utilizar

La distribución a utilizar es la Binomial por las razones que se expresan a continuación:

a. Existen solamente dos posibles resultados, el profesor mejoró la calidad del sílabo entregado después del seminario o no.
b. La probabilidad de éxito y la probabilidad de fracaso son la misma en cada ensayo e igual a 0.5.
c. Cada ensayo es independiente, es decir, el resultado en la elaboración del sílabo de un profesor no se relaciona con el resultado obtenido por otro.
d. El número total de ensayos es fijo, quince en este caso.

Paso 4: En dependencia de la hipótesis alternativa, se escoge la regla de

decisión adecuada

Debido a que lo profesores Víctor González y Fausto Villamarín mantuvieron la misma calificación antes y después del seminario, son eliminados del estudio y por tanto el tamaño de la muestra se reduce a trece docentes.

La distribución de probabilidad binomial para n = 13 y π = 0.50 fue extraída de la **TABLA T.6** del Anexo A y se muestra en la tabla 14.3.

TABLA 14.3 Distribución binomial para n = 13 y π = 0.50

Exitos	Probabilidad de éxito	Probabilidad acumulada
0	0.000	1
1	0.002	1
2	0.010	0.998
3	0.035	0.988
4	0.087	0.953
5	0.157	0.866
6	0.209	0.709
7	0.209	0.5
8	0.157	0.291
9	0.087	0.134
10	0.035	0.047
11	0.010	0.012
12	0.002	0.002
13	0.000	0

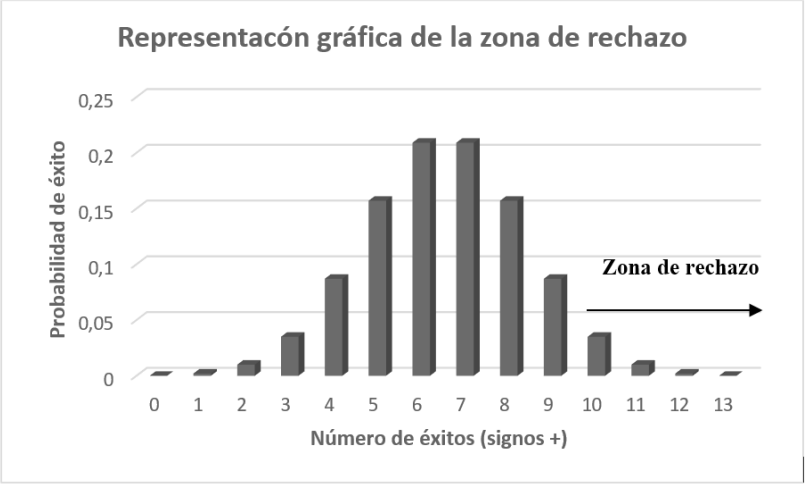
Por ser la prueba de hipótesis de una sola cola a la derecha $(H_1 : \pi > 0.50)$, las probabilidades acumuladas se calculan sumando las probabilidades de éxito de abajo hacia arriba, ya que la zona de rechazo está en la cola superior o derecha. Por ejemplo, para calcular la probabilidad acumulada de 10 o más éxitos se haya la suma 0.000 + 0.002 + 0.010 + 0.035 = 0.047.

La regla de decisión se establece buscando de abajo hacia arriba en la tabla anterior la *probabilidad acumulada más cercana, pero sin exceder*, el nivel de significación que se haya decidido utilizar, en este caso, 0.05.

Esta probabilidad es 0.047 la cual corresponde a un número de éxitos igual a 10. La regla de decisión queda entonces como sigue:

Si el número de signos + en la muestra es mayor o igual a 10, se rechaza la hipótesis nula, de lo contrario, no se rechaza.

FIGURA 14.1 Zona de rechazo para la prueba



Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula

Un total de 9 docentes de la Facultad de Economía mejoraron la calificación obtenida en el curso lectivo 2014 – 2015 (9 signos +) con relación a la calidad de sus sílabos, y como 9 < 10, no se rechaza la hipótesis nula, es decir, no hay mejorías de la calidad de los sílabos como consecuencia del seminario impartido a los profesores

- **Prueba de hipótesis de una sola cola a la izquierda.**

Si la prueba de hipótesis hubiese sido de una sola cola a la izquierda $(H_1 : \pi < 0.50)$, las probabilidades acumuladas se hubiesen calculado sumando las probabilidades de éxito de arriba hacia abajo, ya que la zona de rechazo estaría en la cola inferior o izquierda.

- **Prueba de hipótesis de dos colas.**

Si la prueba de hipótesis hubiese sido de dos colas $(H_1 : \pi \neq 0.50)$ entonces existirían dos zonas de rechazo, una en la cola inferior y la otra en la cola superior. Adicionalmente, el área en cada cola sería igual a $\alpha/2$ (en este caso 0.025).

Para establecer la cola inferior o izquierda buscaríamos de arriba hacia abajo la *probabilidad acumulada más cercana, pero sin exceder*, la mitad del nivel de significación que se haya decidido utilizar, en este caso 0.025, la cual corresponde a un número de éxitos igual a 2 puesto que 0.000 + 0.002 + 0.010 = 0.012.

Para la cola superior se procede de idéntica manera, pero buscando de abajo hacia arriba, lo cual corresponde a un número de éxitos igual a 11.

La regla de decisión quedaría entonces como sigue: ***rechazar la hipótesis nula si hubiese dos o menos signos + u once o más signos +.***

14.2.1 Prueba de los signos usando la distribución normal.

En el Capítulo 5 señalamos que cuando tanto nπ como n (1-π) son mayores o iguales a 5, podemos utilizar la distribución normal como una aproximación a la binomial, con una media μ igual a nπ y una desviación estándar σ igual a $\sqrt{n\pi(1-\pi)}$.

También expresamos en ese capítulo que Frank Yates, estadístico inglés del siglo XX, propuso un *factor de corrección de continuidad* que consiste en sumar o restar, según sea el caso, 0.5 unidades a la variable al momento de hacer la aproximación.

La tipificación de la variable quedaría entonces como se muestra a continuación:

$$z = \frac{(X \pm 0.5) - \mu}{\sigma}$$

En la prueba de los signos cuando existe la posibilidad de utilizar la normal como una aproximación a la binomial, el estadístico de prueba quedaría como sigue:

- **Cuando la cantidad de signos + es mayor que n/2**

$$z = \frac{(X - 0.5) - \mu}{\sigma} = \frac{(X - 0.50) - 0.50n}{0.50\sqrt{n}}$$

- **Cuando la cantidad de signos + es menor que n/2**

$$z = \frac{(X + 0.5) - \mu}{\sigma} = \frac{(X + 0.50) - 0.50n}{0.50\sqrt{n}}$$

En las fórmulas anteriores el valor +0.50 o -0.50 son necesarios como *factor de corrección de continuidad,* el cual fue tratado en el Capítulo 5.

En el ejemplo que hemos estado desarrollando n = 13 y π = 0.5, por tanto, nπ = 13 x 0.5 = 6.5 y n(1-π) = 13 (1-0.5) = 6.5, y en consecuencia, podemos utilizar la distribución normal. En el ejemplo X = 9.

Desarrollando el proceso de los *cinco pasos:*

Paso 1: Se formulan las hipótesis nula y alternativa

$H_0 : \pi \leq 0.50$ No hay mejorías en la elaboración de los sílabos.

$H_1 : \pi > 0.50$ Sí hay mejorías en la elaboración de los sílabos.

Paso 2: Se establece el nivel de significación

α = 0.05

Paso 3: Se identifica la distribución a utilizar

La distribución a utilizar es la Normal y el punto crítico:

$$Z_{1-\alpha} = Z_{1-0.05} = Z_{0.95} = 1.64$$

Paso 4: En dependencia de la hipótesis alternativa, se escoge la regla de decisión adecuada

Para este caso la regla de decisión quedaría como sigue:

Rechazar H_0 si $\dfrac{(X-0.50)-0.50n}{0.50\sqrt{n}} > Z_{1-\alpha}$

Aceptar H_0 si $\dfrac{(X-0.50)-0.50n}{0.50\sqrt{n}} \leq Z_{1-\alpha}$

Paso 5: En base a la muestra tomada se decide o no rechazar la hipótesis nula

$$\frac{(X-0.50)-0.50n}{0.50\sqrt{n}} = \frac{(9-0.50)-0.50(13)}{0.50\sqrt{13}} = \frac{2}{1.80} = 1.11$$

y como 1.11 < 1.64, no se rechaza la hipótesis nula, y en conclusión, según los datos de la muestra los docentes de la Facultad de Economía no mejoraron su calificación en la elaboración de los sílabos después del seminario impartido.

14.2.2 Prueba de hipótesis de una mediana.

Las pruebas de hipótesis que hemos estudiado hasta el momento han involucrado a la media o la proporción de una población. Sin embargo, en ocasiones resulta necesario someter a prueba el valor de una *mediana,* lo cual puede hacerse mediante *la prueba de los signos***.** Cuando se le realiza una prueba de hipótesis a una mediana, a cualquier valor en la muestra por encima de ella se le asigna un signo "+" y a un valor por debajo de dicha mediana un signo "_". Si el valor en la muestra es igual a la mediana este dato se elimina del análisis. Estudiemos un ejemplo.

El costo de una maestría en el Ecuador en los últimos años ha tenido una mediana igual a $6000. El Consejo de Educación Superior realizó un estudio actualizado sobre este tema y en una muestra de 80 maestrías obtuvo que 47 de ellas tenían un costo mayor a $6000, 30 un costo menor y 3 un costo igual. Con un nivel de significación del 1%, ¿podemos concluir que la mediana del costo de una maestría en Ecuador no es igual a $6000?

Las hipótesis nula y alternativa son:

H_0: Mediana = $6000

H_1: Mediana ≠ $6000

lo cual corresponde con una prueba de dos colas.

El estadístico de prueba es compatible con una distribución binomial ya que:

- El costo de una maestría es mayor o menor que la mediana, es decir, solo existen dos resultados posibles.
- La probabilidad de éxito y la probabilidad de fracaso son la misma en cada ensayo e igual a 0.5.
- Las maestrías seleccionadas en la muestra representan ensayos independientes.
- El número total de ensayos es fijo, 77 en este caso.

n = 77 y π = 0.5, por tanto, nπ = 77 x 0.5 = 38.5 y n (1-π) = 77 x 0.5 = 38.5, que al ser mayores que 5 permiten la utilización de la distribución normal para aproximar la binomial.

El valor crítico de z es igual a $Z_{1-\frac{\alpha}{2}} = Z_{1-\frac{0.01}{2}} = Z_{0.995} = 2.58$

Por ser la cantidad de *signos* **+** igual a 47 que es un valor mayor a 77/2 = 38.5, utilizamos la siguiente expresión para calcular el valor del estadístico de prueba z:

$$z = \frac{(X - 0.50) - 0.50\,n}{0.50\sqrt{n}} = \frac{(47 - 0.50) - 0.50(77)}{0.50\sqrt{77}} = 1.82$$

Como 1.82 < 2.58, no se rechaza la hipótesis nula y concluimos que no hay razones que nos permitan asegurar que la mediana del costo de una maestría en Ecuador no es igual a $6000.

14.3 Prueba de rangos de Wilcoxon.
1. Muestras dependientes.

En el Capítulo 9 estudiamos los aspectos relacionados con la prueba de hipótesis para dos muestras *dependientes*, también llamada muestra apareada, utilizando para ello la distribución *t de Student*.

Para describir el procedimiento correspondiente a esta prueba usamos un ejemplo en el cual "una compañía productora de equipos de aire acondicionado para automóviles deseaba comprobar si existían diferencias en la cantidad de kilómetros recorridos por litro cuando se conduce el auto sin aire acondicionado o con él, y para ello, midió el kilometraje por litro de 15 autos que recorrieron una distancia determinada sin el aire acondicionado y con las ventanillas abiertas, midiendo posteriormente dicho kilometraje con los mismos autos y los mismos conductores pero con el aire acondicionado en funcionamiento".

En este capítulo también señalamos que en este tipo de situación realmente lo que interesa es conocer si la distribución de las diferencias entre las distancias recorridas por litro es igual a 0, lo cual requiere *como requisito* que la distribución

de tales diferencias sea *la normal.*

Sin embargo, se presentan casos en los cuales no es posible suponer que las diferencias se aproximen a una distribución normal. Frank Wilcoxon, químico y estadístico estadounidense, propuso en el año 1945 una prueba no paramétrica llamada *prueba de rangos con signo de Wilcoxon*, la cual está soportada en las diferencias de muestras dependientes y que no requieren la suposición de normalidad. Desarrollemos un ejemplo para describir esta prueba.

Una compañía de taxis trata de decidir si el uso de llantas radiales en lugar de llantas regulares mejora la economía de combustible. Se equipan 16 automóviles con llantas radiales y se manejan por un recorrido de prueba establecido. Sin cambiar de conductores, se equipan los mismos autos con llantas regulares y se manejan una vez más por el recorrido de prueba. Se registra la cantidad de litros de gasolina consumidos en el recorrido de prueba, los cuales se muestran en la tabla 14.4.

TABLA 14.4 Litros de gasolina consumidos en el recorrido de prueba

Automóvil	Llantas radiales	Llantas regulares
1	4.2	4.1
2	4.7	4.9
3	6.6	6.2
4	7.0	6.9
5	6.7	6.8
6	4.5	4.4
7	5.7	5.7
8	6.0	5.8
9	7.4	6.9
10	4.9	4.9
11	6.1	6.0
12	5.2	4.9
13	5.7	5.3
14	6.9	6.5
15	6.8	7.1
16	4.9	4.8

www.itch.edu.mx

Con un nivel de significación del 5%, ¿podemos concluir que los automóviles equipados con llantas radiales consumen menos combustible que los equipados con llantas regulares?

Las hipótesis nula y alternativa son las siguientes:

H_0: El consumo de combustible entre los dos tipos de llantas es el mismo.

H_1: Las llantas radiales provocan un menor consumo de combustible que las

llantas regulares.

La prueba es de una sola cola.

Con relación al problema planteado podemos analizar que:

- Sin lugar a dudas las muestras son dependientes.
- El hecho de no haber definido una velocidad constante en la conducción de los automóviles, hace que no estemos seguros que la distribución de las diferencias entre los dos tipos de llantas siga una distribución normal.
- Las dos razones antes expuestas sugieren la utilización de la prueba de rangos con signo de Wilcoxon.

Para desarrollar esta prueba de rangos con signo de Wilcoxon se deben realizar los siguientes pasos (observe la tabla 14.5):

1. Se calcula la diferencia entre las llantas radiales y las llantas regulares en cuanto al consumo de combustible en litros. Por ejemplo, la diferencia para el automóvil 1 es 4.2 – 4.1 = 0.1
2. En el análisis solo se toman en cuenta las diferencias positivas y negativas. En los casos de diferencias iguales a 0 los automóviles que correspondan se eliminan del análisis y se reducen del tamaño de la muestra. Los automóviles 7 y 10 están en esta situación, por tanto, se eliminan del análisis y el tamaño de la muestra pasa de 16 a 14.
3. Se convierten las diferencias a valor absoluto, es decir, sin signo. Así la diferencia correspondiente al automóvil 2 pasa de ser -0.2 a 0.2.
4. Se ordenan las diferencias absolutas de menor a mayor. Este resultado recibe el nombre de *Rango* (R). Por ejemplo, después del ordenamiento, el automóvil 1 ocupa el 1er lugar, por tanto, se le asigna el rango 1. El automóvil 9 ocupa el 5to lugar, y por ello, se le asigna el rango 5.
5. En la columna R^+ se escriben los rangos de los automóviles cuya diferencia fue positiva y en R^- los rangos cuyas diferencias fueron negativas.
6. Se suman las columnas R^+ y R^-. La suma de la columna R^+ se designa como W^+, la suma de la columna R^- como W^- y al valor menor de estas dos sumas con la letra W.
7. El siguiente paso consiste en establecer la regla de decisión correspondiente, la cual en dependencia de las hipótesis nula y alternativa que se formulen quedaría de la siguiente manera:

1)

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Se rechaza la hipótesis nula si W es *menor o igual* que el valor crítico de la T de Wilcoxon para un tamaño de muestra n y un nivel de significación α .

2)

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 > \mu_2$

Se rechaza la hipótesis nula si W^- es *menor o igual* que el valor crítico de la T de Wilcoxon para un tamaño de muestra n y un nivel de significación α .

3)

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 < \mu_2$

Se rechaza la hipótesis nula si W^+ es *menor o igual* que el valor crítico de la T de Wilcoxon para un tamaño de muestra n y un nivel de significación α .

Este último caso es el correspondiente al ejemplo que hemos estado desarrollando.

La tabla 14.5 contiene los resultados obtenidos al cumplimentar los primeros seis pasos indicados anteriormente, y en la misma se puede apreciar que el estadístico de prueba que deberá ser utilizado es $W^+ = 27$.

La **TABLA T.9** del Anexo A contiene los valores críticos para la prueba de rangos con signo de Wilcoxon.

A continuación se muestra un segmento de dicha tabla. En ella se ha resaltado en color rojo el valor crítico de la T de Wilcoxon para un tamaño de muestra igual a 14 y un nivel de significación del 5%. Este valor crítico es el *mayor valor en la zona de rechazo.*

	Prueba de dos colas						
	0.15	**0.10**	**0.05**	**0.04**	**0.03**	**0.02**	**0.01**
	Prueba de una cola						
N	0.075	0.050	0.025	0.020	0.015	0.010	0.005
10	12	10	8	7	6	5	3
11	16	13	10	9	8	7	5
12	19	17	13	12	11	9	7
13	24	21	17	16	14	12	9
14	28	**25**	21	19	18	15	12
15	33	30	25	23	21	19	15

Como hemos explicado, la regla de decisión consiste en rechazar la hipótesis nula si el valor de W^+ es menor o igual a 25.

Como 27 > 25, no se rechaza la hipótesis nula, y en consecuencia, no podemos

asegurar que el consumo de combustible provocado por las llantas radiales sea menor que el determinado por las llantas regulares.

TABLA 14.5 Calculo de los rangos con signo

Autos	Llantas radiales	Llantas regulares	Diferencias	Diferencias absolutas	Rango con signo		
					R	R^+	R^-
1	4.2	4.1	0.1	0.1	1	1	
2	4.7	4.9	-0.2	0.2	2		2
3	6.6	6.2	0.4	0.4	4	4	
4	7.0	6.9	0.1	0.1	1	1	
5	6.7	6.8	-0.1	0.1	1		1
6	4.5	4.4	0.1	0.1	1	1	
7	5.7	5.7	0				
8	6.0	5.8	0.2	0.2	2	2	
9	7.4	6.9	0.5	0.5	5	5	
10	4.9	4.9	0				
11	6.1	6.0	0.1	0.1	1	1	
12	5.2	4.9	0.3	0.3	3	3	
13	5.7	5.3	0.4	0.4	4	4	
14	6.9	6.5	0.4	0.4	4	4	
15	6.8	7.1	-0.3	0.3	3		3
16	4.9	4.8	0.1	0.1	1	1	
						27	6

Si Ud. observa con detenimiento el planteamiento del ejercicio anterior, podrá apreciar que la binomial es la distribución de muestreo apropiada, y en consecuencia, cuando tanto np como nq son mayores a cinco podemos aproximarla mediante la distribución normal. Por ser el tamaño de la muestra igual a 14 y p = q = 0.5, podemos utilizar la distribución normal para desarrollar la prueba.

Si establecemos que:

$$\overline{p} = \frac{Cantidad\ de\ signos\ positivos}{Tamaño\ de\ la\ muestra}$$

$$\sigma_{\overline{p}} = \sqrt{\frac{0.25}{n}}$$

Las reglas de decisión para los tres casos de pruebas de hipótesis quedan como se muestra a continuación:

1)

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Rechazar H_0 si $\overline{p} < 0.5 - Z_{1-\frac{\alpha}{2}}\sigma_{\overline{p}}$ *o* $\overline{p} > 0.5 + Z_{1-\frac{\alpha}{2}}\sigma_{\overline{p}}$, No rechazar H_0 en caso contrario

2)

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 > \mu_2$

Rechazar H_0 si $\overline{p} > 0.5 + Z_{1-\alpha}\sigma_{\overline{p}}$, No rechazar H_0 en caso contrario

3)

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 < \mu_2$

Rechazar H_0 si $\overline{p} < 0.5 - Z_{1-\alpha}\sigma_{\overline{p}}$, No rechazar H_0 en caso contrario

En el caso concreto que estamos estudiando:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 < \mu_2$

$\alpha = 0.05$

$$\overline{p} = \frac{11}{14} = 0.79$$

$$Z_{1-\alpha} = Z_{1-0.05} = Z_{0.95} = 1.64$$

$$\sigma_{\overline{p}} = \sqrt{\frac{0.25}{14}} = \sqrt{0.0179} = 0.134$$

Límite de la región crítica: 0.5 - 1.64 (0.134) = 0.28

Como 0.79 > 0.28 no se rechaza la hipótesis nula y en consecuencia no podemos asegurar que el consumo de combustible provocado por las llantas radiales sea menor que el determinado por las llantas regulares.

2. Muestras independientes

En el Capítulo 9 estudiamos que cuando las muestras son independientes y las dos poblaciones siguen *distribuciones normales con varianzas iguales* resulta posible desarrollar una prueba de hipótesis para las medias de ambas poblaciones usando la distribución t de Student.

Sin embargo, existen ocasiones en las cuales una o ambas de estas dos condiciones no se satisfacen, y resulta necesario utilizar otro método estadístico para darle solución al problema planteado.

Un método específicamente concebido para este caso es la *prueba de Wilcoxon*

de la suma de rangos para muestras independientes.

Para el desarrollo de esta prueba, los datos de ambas muestras se clasifican como si constituyeran una sola. Si ambas muestran provienen de poblaciones equivalentes, los rangos tendrán entonces una distribución semejante entre ambas muestras, y en consecuencia, la suma de los rangos de dichas dos muestras serán aproximadamente iguales.

Cuando el tamaño de cada una de las muestras es al menos igual a ocho observaciones, entonces se utiliza la distribución normal estándar como estadístico de prueba, calculándose el valor de z de la siguiente manera:

$$z = \frac{W - \frac{n_1\left(n_1 + n_2 + 1\right)}{2}}{\sqrt{\frac{n_1 n_2\left(n_1 + n_2 + 1\right)}{12}}} \quad donde:$$

n_1 representa la cantidad de observaciones de la primera muestra.

n_2 representa la cantidad de observaciones de la segunda muestra.

W representa la suma de los rangos de la primera población.

Consideremos el siguiente ejemplo.

Un profesor de estadística que permite la entrada a su clase pasada la hora establecida para su comienzo, ha observado que los estudiantes de la noche son más impuntuales que los que estudian en la mañana, y para comprobarlo escogió a 9 estudiantes del curso matutino y a 9 del curso nocturno, a los cuales les midió los minutos de retraso al momento de ingresar al aula. Con un nivel de significación del 5%, ¿podemos aseverar que los estudiantes del curso nocturno son más impuntuales que los del curso matutino?

Los resultados de la investigación se muestran a continuación:

TABLA 14.6 Minutos de tardanza de los estudiantes

MINUTOS DE TARDANZA				
NOCTURNO	**MATUTINO**		**NOCTURNO**	**MATUTINO**
20	17		23	19
24	18		19	22
14	12		26	27
21	16		18	15
15	18			

Sin lugar a dudas las dos muestras anteriores *son independientes*, y en caso que sus poblaciones sigan una distribución normal con varianzas iguales, podríamos desarrollar una prueba de hipótesis para dos muestras utilizando la ya conocida

distribución t de Student.

Consideremos, sin embargo, que el profesor no está seguro de la igualdad de las varianzas y decide entonces utilizar la *prueba de Wilcoxon de la suma de rangos para muestras independientes.*

Esta prueba puede desarrollarse por cuanto cada muestra tiene un total de 9 observaciones.

Las hipótesis nula y alternativa son:

H_0: La distribución de la población de estudiantes impuntuales es la misma para el curso nocturno que para el curso diurno.

H_1: La distribución de la población de estudiantes impuntuales es mayor para el curso nocturno que para el curso diurno.

La asignación del rango a cada una de las observaciones correspondiente a los minutos de tardanza de los estudiantes se muestra en la tabla 14.7.

TABLA 14.7 Asignación de rangos a los minutos de tardanza

12	14	15	15	16	17	18	18	18	19	19	20	21	22	23	24	26	27
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Observe en la tabla las siguientes características:

1. Las observaciones de ambas muestras fueron ordenadas y los rangos asignados como si pertenecieran a una sola.
2. El tercer y cuarto estudiante cuyos rangos son tres y cuatro respectivamente, presentaron igual número de minutos de tardanza (15 minutos). En estos casos de igualdad se deben promediar los rangos inicialmente asignados, y el promedio resultante es el que se asigna. En consecuencia, a ambos tiempos de tardanza se le debe asignar el rango 3.5, el cual es el promedio entre 3 y 4. Lo mismo ocurre con los tiempos de tardanza correspondientes a 18 y 19 minutos. A continuación se muestran los rangos asignados a los minutos de tardanza de los estudiantes de los cursos nocturno y matutino.

TABLA 14.8 Rangos asignados a los minutos de tardanza

MINUTOS DE TARDANZA			
NOCTURNO	**RANGO**	**MATUTINO**	**RANGO**
20	12	17	6
24	16	18	8
14	2	12	1
21	13	16	5
15	3.5	18	8

MINUTOS DE TARDANZA			
NOCTURNO	**RANGO**	**MATUTINO**	**RANGO**
23	15	19	10.5
19	10.5	22	14
26	17	27	18
18	8	15	3.5
	97		74

El estadístico de prueba sigue una distribución normal y para un nivel de significación del 5% el valor crítico de z para una prueba de una cola es 1.64. Según esto, la hipótesis nula se rechaza si el estadístico de prueba z es mayor que 1.64.

Como la hipótesis alternativa es que las impuntuales son mayores para el curso nocturno que para el curso diurno, calculamos z con W = 97, es decir, la suma de los rangos del curso nocturno.

En el ejemplo, $n_1 = n_2 = 9$. Por supuesto que los valores de n_1 y n_2 no tienen necesariamente que ser iguales.

$$z = \frac{97 - \frac{9(9+9+1)}{2}}{\sqrt{\frac{(9)(9)(9+9+1)}{12}}} = \frac{11.5}{11.32} = 1.02$$

Como 1.02 < 1.64, no se rechaza la hipótesis nula, y en consecuencia, no hay suficientes evidencias que permitan asegurar que la distribución de la población de estudiantes impuntuales es mayor para el curso nocturno que para el curso diurno. Al parecer los estudiantes del curso diurno son tan impuntuales como los del curso nocturno.

14.4 Otras pruebas de suma de rangos para muestras independientes.

El Capítulo 10 fue dedicado al estudio del *Análisis de Varianza*, el cual es un método estadístico que nos permite realizar una prueba de hipótesis para comprobar si existen o no diferencias en las medias poblacionales de dos o más poblaciones. Estudiamos también que la aplicación del método requería asumir que las diferentes poblaciones estaban normalmente distribuidas con varianzas iguales. En ocasiones estas suposiciones no pueden ser asumidas, y en lugar del análisis de varianza se pueden utilizar dos pruebas no paramétricas ninguna de las cuales requiere suponer normalidad ni homogeneidad de varianzas.

Las pruebas a las que hacemos referencia son la *Prueba U de Mann-Whitney* y la *Prueba de Kruskal-Wallis*. Ambas pruebas se basan en la ya conocida *suma de rangos*. La primera de las dos pruebas es útil cuando trabajamos con solo dos poblaciones,

mientras que la segunda permite trabajar con más de dos de ellas.

14.4.1 Prueba U de Mann-Whitney.

La Comisión Académica de una Facultad de Ciencias Administrativas desea someter a prueba si el tiempo dedicado al estudio en época de exámenes a la materia de Estadística, es diferente al dedicado a la asignatura de Administración. Para ello, confeccionó dos grupos de 12 estudiantes cada uno, a los cuales les determinó mediante encuestas el número de horas dedicadas al estudio previas al examen de ambas asignaturas. Con un nivel de significación del 1%, ¿podemos asegurar que el tiempo dedicado al estudio en la materia de Estadística es diferente al dedicado a la asignatura de Administración?

Los resultados del tiempo dedicado al estudio previo al período de exámenes se muestran en la siguiente tabla:

TABLA 14.9 Tiempo dedicado al estudio previo al período de exámenes

Estadística	Administración		Estadística	Administración
12	14		11	7
8	9		15	15
14	12		16	14
9	10		7	10
13	11		10	12
10	12		9	11

Describamos en detalle los pasos a seguir para desarrollar esta prueba.

a. Se establecen las hipótesis nula y alternativa:

Las hipótesis nula y alternativa son:

$$H_0 : \mu_E = \mu_A$$

$$H_1 : \mu_E \neq \mu_A$$

donde μ_E representa la media poblacional del tiempo dedicado al estudio de la materia Estadística y μ_A la media poblacional del tiempo dedicado al estudio de la materia Administración.

b. Se fija el nivel de significación:

En el texto del ejercicio se define α = 0.01.

c. Se ordenan de menor a mayor los valores de las dos muestras de forma conjunta y se le asigna un rango a cada uno de estos valores:

En la tabla 14.10 se muestran las horas dedicadas al estudio ordenadas de menor a mayor y consignadas según el tipo de materia, es decir, E para Estadística y A para Administración.

TABLA 14.10 Rangos asignados a las horas dedicadas al estudio

HORAS	MATERIA	RANGO		HORAS	MATERIA	RANGO
7	E	1		11	A	13
7	A	2		12	E	14
8	E	3		12	A	15
9	E	4		12	A	16
9	E	5		12	A	17
9	A	6		13	E	18
10	E	7		14	E	19
10	E	8		14	A	20
10	A	9		14	A	21
10	A	10		15	E	22
11	E	11		15	A	23
11	A	12		16	E	24

Como se puede apreciar, en la tabla existen varios números de horas repetidos los cuales reciben el nombre de *ligaduras* y que debemos resolver de la forma que ya ha sido explicada.

Por ejemplo, para la observación correspondiente a 12 horas el rango que se le debe asignar se calcula de la siguiente forma:

$$\frac{14+15+16+17}{4}=\frac{62}{4}=15.5$$

Los rangos *"corregidos"* para el resto de las observaciones se muestran a continuación:

TABLA 14.11 Rangos asignados resolviendo las ligaduras

HORAS	MATERIA	RANGO		HORAS	MATERIA	RANGO
7	E	1.5		11	A	12
7	A	1.5		12	E	15.5
8	E	3		12	A	15.5
9	E	5		12	A	15.5
9	E	5		12	A	15.5
9	A	5		13	E	18
10	E	8.5		14	E	20
10	E	8.5		14	A	20
10	A	8.5		14	A	20
10	A	8.5		15	E	22.5
11	E	12		15	A	22.5
11	A	12		16	E	24

d. Se calculan R_1 y R_2 donde:

R_1 = Suma de los rangos de la primera muestra.

R_2 = Suma de los rangos de la segunda muestra.

Las sumas de los rangos a las que se hace referencia se muestran en la tabla 14.12.

TABLA 14.12 Cálculo de R_1 y R_2

ESTADÍSTICA			**ADMINISTRACIÓN**	
7	1.5		7	1.5
8	3		9	5
9	5		10	8.5
9	5		10	8.5
10	8.5		11	12
10	8.5		11	12
11	12		12	15.5
12	15.5		12	15.5
13	18		12	15.5
14	20		14	20
15	22.5		14	20
16	24		15	22.5
R_1	143.5		R_2	156.5

e. Se define y calcula el estadístico de prueba.

El estadístico de prueba es:

$$U = Mínimo\left(U_1, U_2\right) \; donde \left\{ \begin{array}{l} U_1 = n_1 n_2 + \dfrac{n_1\left(n_1+1\right)}{2} - R_1 \\ U_2 = n_1 n_2 + \dfrac{n_2\left(n_2+1\right)}{2} - R_2 \end{array} \right\}$$

$$U_1 = n_1 n_2 + \frac{n_1\left(n_1+1\right)}{2} - R_1 = (12)(12) + \frac{12\left(12+1\right)}{2} - 143.5 = 144 + 78 - 143.5 = 78.5$$

$$U_2 = n_1 n_2 + \frac{n_2\left(n_2+1\right)}{2} - R_2 = (12)(12) + \frac{12\left(12+1\right)}{2} - 156.5 = 144 + 78 - 156.5 = 65.5$$

Y por tanto, U = 65.5

f. Se decide la distribución de muestreo.

Cuando tanto n_1 como n_2 son mayores a 10, la distribución de muestreo del estadístico de prueba U puede aproximarse mediante la distribución normal en la cual:

La media poblacional y el error estándar son:

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(12)(12)}{2} = 72$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{3600}{12}} = 17.3$$

g. Se estandariza el estadístico de prueba U.

$$z = \frac{65.5 - 72}{17.3} = -0.38$$

h. Se determina el valor crítico de z.

El valor crítico de z para un nivel de significación del 1% y una prueba de dos colas es 2.58.

i. Conclusión de la prueba.

Por ser $|-0.38| = 0.38 < 2.58$ no se rechaza la hipótesis nula, y por tanto, no existen diferencias entre el tiempo de estudio dedicado a la asignatura de Estadística y el dedicado a la materia de Administración.

Cuando no se cumple que tanto n_1 como n_2 son mayores a 10, y en general, para cualesquiera sean los valores de n_1 y n_2, podemos comparar el estadístico de prueba U obtenido con los valores críticos que aparecen en la **TABLA T.14** de la prueba U de Mann–Whitney y de la cual presentamos a continuación un segmento:

		Una cola		Dos colas				Una cola		Dos colas				Una cola		Dos colas	
n_1	n_2	5%	1%	5%	1%	n_1	n_2	5%	1%	5%	1%	n_1	n_2	5%	1%	5%	1%
9	35	215	237	226	245	11	27	201	221	210	228	13	23	201	220	210	227
	36	221	244	232	252		28	208	229	218	236		24	209	229	218	237
	37	227	250	238	258		29	215	236	225	244		25	217	238	227	246
	38	233	257	244	265		30	222	244	232	252		26	225	247	236	255
	39	239	263	250	272		31	229	252	240	260		27	234	256	244	264
	40	245	270	257	279		32	236	260	247	268		28	242	265	253	273
10	10	73	81	77	84		33	243	267	255	276		29	250	274	261	283
	11	79	88	84	92		34	250	275	262	284		30	258	283	270	292
	12	86	96	91	99		35	257	283	269	292		31	267	292	278	301
	13	93	103	97	106		36	265	290	277	300		32	275	301	287	310
	14	99	110	104	114		37	272	298	284	308		33	283	310	296	319
	15	106	117	111	121		38	279	306	291	316		34	291	319	304	329
	16	112	124	118	129		39	286	314	299	323		35	299	328	312	338
	17	119	132	125	136		40	293	321	306	331		36	308	337	321	347
	18	125	139	132	143	12	12	102	113	107	117		37	316	346	330	356
	19	132	146	138	151		13	109	121	115	125		38	324	355	338	365
	20	138	153	145	158		14	117	130	123	134		39	332	363	347	374
	21	145	160	152	166		15	125	138	131	143		40	341	372	355	384
	22	152	167	159	173		16	132	146	139	151	14	14	135	149	141	154
	23	158	175	166	180		17	140	155	147	160		15	144	159	151	164
	24	165	182	173	188		18	148	163	155	169		16	153	168	160	174
	25	171	189	179	195		19	156	172	163	177		17	161	178	169	184
	26	178	196	186	202		20	163	180	171	186		18	170	187	178	194
	27	184	203	193	210		21	171	188	179	194		19	179	197	188	203
	28	191	210	200	217		22	179	197	187	203		20	188	207	197	213

En el ejemplo que hemos venido desarrollando $n_1 = 12$ y $n_2 = 12$. Como se puede apreciar en la tabla, el valor crítico de Mann – Whitney para una prueba de dos colas y una significación del 1% es 117, y como $65.5 < 117$ no se rechaza la hipótesis nula y por tanto, no existen diferencias entre el tiempo de estudio dedicado a la asignatura de Estadística y el dedicado a la materia de Administración.

Para aplicar este procedimiento los datos deben estar organizados de forma tal que n_1 sea menor o igual a n_2.

14.4.2 Prueba de Kruskal-Wallis.

Un investigador está interesado en conocer si existen diferencias entre las regiones Costa, Sierra y Amazonía en cuanto al gasto que sus familias realizan en actividades de turismo. Para ello encuestó a seis familias en cada región y les consultó que porcentaje de sus ingresos dedicaban anualmente a actividades turísticas.

Los resultados de las encuestas realizadas se muestran a continuación:

TABLA 14.13 Porcentaje de ingresos dedicados a actividades turísticas

Costa	Sierra	Amazonía
8.4	10.7	9.3
9.3	8.7	8.8
8.1	11.8	10.2
10.4	8.5	8.6
11.6	9.6	10.1
8.3	8.5	8.7

Con un nivel de significación del 5%, ¿existen diferencias entre las tres regiones en los gastos en turismo?

a. Se establecen las hipótesis nula y alternativa:

Las hipótesis nula y alternativa son:

H_0: No hay diferencias en gasto en turismo entre las tres regiones.

H_1: Al menos una región difiere de las otras dos en gasto en turismo.

b. Se fija el nivel de significación:

En el ejercicio quedó establecido en un 5% el nivel de significación.

c. Se ordenan de menor a mayor los valores de las tres muestras de forma conjunta y se le asigna un rango a cada uno de estos valores:

En la tabla 14.14 se muestran los porcentajes ordenados de menor a mayor y consignados según la región estudiada, es decir, C para Costa, S para Sierra y A para Amazonía. Adicionalmente aparecen los rangos para cada una de las observaciones.

TABLA 14.14 Rangos asignados a los porcentajes por región

%	REGIÓN	RANGO		%	REGIÓN	RANGO
8.1	C	1		9.3	A	10
8.3	C	2		9.3	C	11
8.4	C	3		9.6	S	12
8.5	S	4		10.1	A	13
8.5	S	5		10.2	A	14
8.6	A	6		10.4	C	15
8.7	S	7		10.7	S	16
8.7	A	8		11.6	C	17
8.8	A	9		11.8	S	18

Los rangos *"corregidos"* por concepto de *ligaduras* y el resto de los mismos se muestran a continuación:

TABLA 14.15 Rangos corregidos resolviendo ligaduras

%	REGIÓN	RANGO		%	REGIÓN	RANGO
8.1	C	1		9.3	A	10.5
8.3	C	2		9.3	C	10.5
8.4	C	3		9.6	S	12
8.5	S	4.5		10.1	A	13
8.5	S	4.5		10.2	A	14
8.6	A	6		10.4	C	15
8.7	S	7.5		10.7	S	16
8.7	A	7.5		11.6	C	17
8.8	A	9		11.8	S	18

d. Se calculan R_1, R_2 y R_3, es decir, la suma de los rangos para Costa, Sierra y Amazonía respectivamente:

TABLA 14.16 Cálculo de los rangos R_1, R_2 y R_3

COSTA	RANGO	SIERRA	RANGO	AMAZONÍA	RANGO
8.4	3	10.7	16	9.3	10.5
9.3	10.5	8.7	7.5	8.8	9
8.1	1	11.8	18	10.2	14
10.4	15	8.5	4.5	8.6	6
11.6	17	9.6	12	10.1	13
8.3	2	8.5	4.5	8.7	7.5
	48.5		62.5		60

e. Se define y calcula el estadístico de prueba.

El estadístico de prueba es:

$$H = \frac{12}{N(N+1)}\sum_{1=1}^{k}\frac{R_1^2}{n_i} - 3(N+1) \quad donde:$$

k = cantidad de muestras.

R_i = Suma de los rangos correspondientes a la muestra i-ésima.

n_i = número de observaciones de la muestra i-ésima.

$$N = \sum_{1}^{k} n_i$$

$$H = \frac{12}{18(18+1)}\left[\frac{(48.5)^2}{6} + \frac{(62.5)^2}{6} + \frac{(60)^2}{6}\right] - 3(18+1) = 0.0351(1643.08) - 57 = 0.67$$

f. Se decide la distribución de muestreo.

Cuando cada una de las muestras tiene al menos cinco observaciones, lo cual resulta completamente usual, la distribución del estadístico de prueba H es aproximadamente igual a la distribución Ji – Cuadrada con k-1 grados de libertad.

g. Se determina el valor crítico de H.

La Ji – Cuadrada con k – 1 = 3 – 1 = 2 grados de libertad y un nivel de significación del 5% tiene un valor igual a 5.99.

h. Conclusión de la prueba.

Como 0.67 es menor que 5.99, entonces no se rechaza la hipótesis nula, y en consecuencia, no existen evidencias que nos permitan concluir que existen diferencias en los gastos en turismo entre las tres regiones estudiadas.

14.5 Prueba de aleatoriedad de una muestra.

Hasta el momento hemos estado suponiendo que las muestras han sido extraídas de manera totalmente aleatoria, es decir, sin que mediara en el proceso ningún tipo de criterio de selección.

Sin embargo, pudiera ocurrir que tengamos que trabajar con una muestra seleccionada por otra persona y tengamos dudas si ésta fue realmente elegida en un *orden* aleatorio. Para comprobar una muestra en cuanto a la aleatoriedad del orden en que fue seleccionada, podemos hacer uso de la llamada *teoría de corridas o rachas*.

Una corrida o racha se define como una secuencia de ocurrencias idénticas las cuales pueden o no estar precedidas y seguidas de distintas ocurrencias.

Por ejemplo, supongamos que una compañía está desarrollando entrevistas a Economistas (E) e Ingenieros Comerciales (I) para seleccionar un Gerente Financiero, y que el orden declarado por la compañía en que los candidatos a la gerencia fueron entrevistados es el que se muestra en la tabla 14.17.

TABLA 14.17 Orden en que fueron entrevistados los candidatos

E	I	I	I	E	E	I	E	E	E	I	I
E	E	I	E	I	I	I	E	E	I	I	I
E	I	I	I	E	E	I	E	E	E	I	I
E	E	E	I	I	E	I	I	I	I	E	E

Con un nivel de significación del 5%, ¿podemos aceptar que la secuencia de entrevistas declarada por la compañía tiene un orden aleatorio?

Las hipótesis nula y alternativa son:

H_0: El orden en que los candidatos a la gerencia fueron entrevistados es aleatorio.

H_1: El orden en que los candidatos a la gerencia fueron entrevistados no es aleatorio.

En la siguiente tabla se muestran las *23 corridas o rachas* que contiene la secuencia de entrevistas realizadas:

TABLA 14.18 Corridas de la secuencia de entrevistas

E	I	I	I	E	E	I	E	E	E	I	I											
1	2			3		4	5			6												
E	E	I	E	I	I	I	E	E	I	I	I											
7		8	9	10			11		12													
E	I	I	I	E	E	I	E	E	E	I	I											
13	14			15		16	17			18												
E	E	E	I	I	E	I	I	I	I	E	E											
19			20		21	22				23												

Si denotamos por:

n_1 = número de ocurrencias del tipo 1

n_2 = número de ocurrencias del tipo 2

r = número de corridas

En nuestro ejemplo:

n_1 = número de Economistas entrevistados = 23

n_2 = número de Ingenieros Comerciales entrevistados = 25

r = número de corridas = 23

El número de corridas r (estadístico de prueba) tiene una distribución de muestreo con los siguientes parámetros:

$$\mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1 \quad \sigma_r = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

En nuestro ejemplo:

$$\mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1 = \frac{2(23)(25)}{23 + 25} + 1 = \frac{1150}{48} + 1 = 24.96$$

$$\sigma_r = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}} = \sqrt{\frac{2(23)(25)((2)(23)(25) - 23 - 25)}{(23 + 25)^2 (23 + 25 - 1)}}$$

$$\sigma_r = \sqrt{\frac{(1150)(1102)}{2304\, x\, 47}} = \sqrt{\frac{1267300}{108288}} = 3.42$$

Si n_1 o n_2 es mayor a 20, la distribución de muestreo de r se aproxima bastante a la normal. En nuestro ejemplo se cumple esta condición, por tanto, encontremos

el valor crítico de la distribución normal estándar para un área bajo la curva igual a $1-\frac{\alpha}{2}=1-\frac{0.05}{2}=0.975$ (prueba de dos colas). Este valor buscado en la **TABLA T.1** del Anexo es z = 1.96, y en consecuencia, los límites de la región crítica son:

Límite superior = 24.96 + 1.96 (3.42) = 31.66

Límite inferior = 24.96 – 1.96 (3.42) = 18.26

Al ser 18.26 < 23 < 31.66 no se rechaza la hipótesis nula, y por tanto, el orden declarado por la compañía en que los candidatos a la gerencia fueron entrevistados es aleatorio.

Cuando no se cumple la condición de que n_1 *o* n_2 *sea mayor a 20*, la distribución normal no puede ser utilizada.

Sin embargo, en este caso se puede hacer uso de las **TABLAS T.11a** y **T.11b** del Anexo A en las cuales se encuentran los valores críticos para la prueba de rachas para un nivel de significación del 5%.

Para ilustrar el uso de estas tablas supongamos que en el ejemplo anterior asistieron a las entrevistas 36 profesionales de los cuales 17 fueron Economistas y 19 Ingenieros Comerciales. La tabla 14.19 muestra la secuencia de llegadas:

TABLA 14.19 Secuencia de llegada a las entrevistas

E	I	I	I	E	E	I	E	E	E	I	I
E	E	I	E	I	I	I	E	E	I	I	I
E	I	I	I	E	E	I	E	E	E	I	I

En la siguiente tabla se muestran las *18 corridas o rachas* que contiene la secuencia de entrevistas realizadas:

TABLA 14.20 Corridas de la secuencia de entrevistas

E	I	I	I	E	E	I	E	E	E	I	I						
1	2			3		4	5			6							
E	E	I	E	I	I	I	E	E	I	I	I						
7		8	9	10			11		12								
E	I	I	I	E	E	I	E	E	E	I	I						
13	14			15		16	17			18							

n_1 = número de Economistas entrevistados = 17

n_2 = número de Ingenieros Comerciales entrevistados = 19

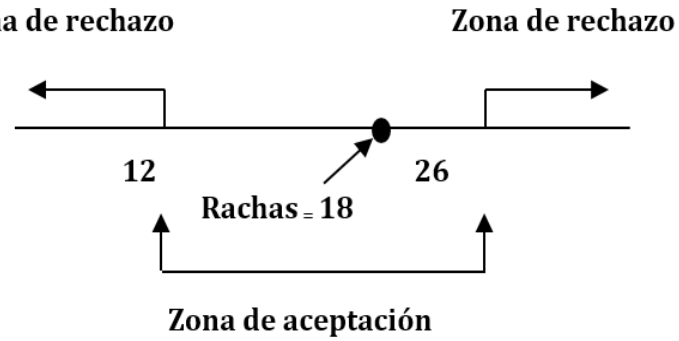
r = número de corridas = 18

Si en la **TABLA T.11a** interceptamos la fila 17 con la columna 19 encontramos

que el valor crítico de las rachas para α = 0.05 es 12. Si hacemos lo mismo con la **TABLA T.11b** hallamos que el valor crítico es 26.

En la figura 14.2 se muestran las zonas de aceptación y de rechazo.

FIGURA 14.2 Zonas de aceptación y de rechazo



Como el número de rachas se encuentra en la zona de aceptación, se acepta la hipótesis nula con un nivel de significación del 5% y se concluye que el orden declarado por la compañía en que los candidatos a la gerencia fueron entrevistados fue aleatorio.

14.6 Correlación por rango.

Al estudiar en los Capítulos 11 y 12 la teoría relacionada con la regresión simple y múltiple, expresamos el concepto de *correlación* entre dos variables medidas en una escala de intervalo o de razón, y señalamos que el cálculo numérico del mismo daba una indicación del grado en que ambas variables se encontraban relacionadas.

En ocasiones, al requerir conocer la correlación entre dos variables nos enfrentamos al problema de que están medidas en una escala nominal. En una situación como ésta podemos asignar rangos a cada una de las observaciones de las dos muestras y calcular un *coeficiente de correlación por rangos*, el cual fue propuesto por el inglés Charles Edward Spearman a principios del siglo XX.

Describamos mediante un ejemplo los pasos a seguir para desarrollar esta prueba.

Los datos que se muestran en la tabla 14.21 representan el orden de llegada a la meta (X) y la edad del participante (Y) de una muestra de 10 corredores en una carrera de 5000 metros.

Los organizadores del evento están interesados en conocer si con un nivel de significación del 1%, se puede concluir que existe una correlación entre estas dos variables.

TABLA 14.21 Orden de llegada a la meta (X) y edad del participante (Y)

X	1	2	3	4	5	6	7	8	9	10
Y	21	23	19	20	25	22	17	26	28	24

a. Se establecen las hipótesis nula y alternativa:

H_0: No existe correlación entre ambas variables

H_1: Existe correlación entre ambas variables

Si designamos como ρ_s al *coeficiente de correlación por rangos*, entonces las hipótesis nula y alternativa pueden quedar expresadas:

$$H_1 : \rho_s = 0$$

$$H_1 : \rho_s \neq 0$$

b. Se fija el nivel de significación:

El nivel de significación se estableció en el ejercicio en el 1%.

c. Se le asignan los rangos correspondientes a cada una de las variables, y adicionalmente, se calcula la diferencia (d) entre ellos, el cuadrado de esta diferencia (d^2) y la suma de éstas:

TABLA 14.22 Asignación de rangos a las variables y sumas requeridas

RANGO X	**RANGO Y**	**d**	**d^2**
1	4	-3	9
2	6	-4	16
3	2	1	1
4	3	1	1
5	8	-3	9
6	5	1	1
7	1	6	36
8	9	-1	1
9	10	-1	1
10	7	3	9
			84

d. Se define y calcula el estadístico de prueba.

El estadístico de prueba es el *coeficiente de correlación por rango*, el cual se calcula mediante la expresión:

$$r_s = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$ donde N es el número de pares de observaciones (10).

$$r_s = 1 - \frac{6(84)}{10(10^2 - 1)} = 1 - \frac{504}{990} = 0.49$$

e. Se decide la distribución de muestreo.

Para un valor de N menor o igual a 30, la distribución de r no es normal, y tampoco resulta posible utilizar la distribución t de Student para la prueba de la hipótesis. En una situación como ésta resulta necesario utilizar la **TABLA T.12** que se muestra en el Anexo A y que contiene los valores críticos del coeficiente de correlación de Spearman.

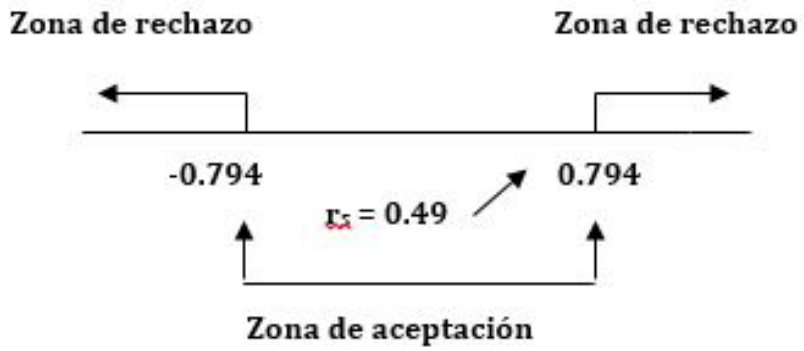
f. Se determina el valor crítico de r_s.

Para un nivel de significación del 1%, una prueba de dos colas y un valor de n igual a 10, el valor crítico de r_s es igual a 0.794.

g. Conclusión de la prueba.

En la figura 14.3 se aprecian gráficamente las ubicaciones de las zonas de aceptación y de rechazo.

FIGURA 14.3 Zonas de aceptación y de rechazo



Como $r_s = 0.49$ cae dentro de la zona de aceptación, entonces no se rechaza la hipótesis nula, y en consecuencia, no existe correlación entre ambas variables.

Cuando el número de pares de valores es mayor a 30 ya no resulta necesario utilizar la **TABLA T.12** de valores críticos, y cuando esto ocurre, la distribución de muestreo de r_s es aproximadamente normal con media 0 y un error estándar igual a $\sigma_{r_s} = \frac{1}{\sqrt{N-1}}$, y en consecuencia, podemos utilizar la **TABLA T.1** del Anexo A para realizar la prueba de hipótesis. Para ejemplificar lo anteriormente planteado, supongamos que los datos del ejemplo anterior son los que se muestran en la tabla 14.23, en la cual X representa el orden de llegada a la meta y Y la edad del participante de una muestra de 36 corredores en una carrera de 5000 metros, y que se mantiene el interés de los organizadores del evento en conocer si con un nivel de significación del

1%, se puede concluir que existe una correlación entre estas dos variables.

En la tabla 14.24 se observan los cálculos requeridos para lograr este objetivo.

TABLA 14.23 Orden de llegada a la meta (X) y edad del participante (Y)

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

TABLA 14.24 Asignación de rangos a las variables y sumas requeridas

RANGO X	**RANGO Y**	**d**	**d^2**		**RANGO X**	**RANGO Y**	**d**	**d^2**
1	7	-6	36		19	26,5	-7,5	56,25
2	9,5	-7,5	56,25		20	33,5	-13,5	182,25
3	4	-1	1		21	13	8	64
4	5,5	-1,5	2,25		22	21	1	1
5	13	-8	64		23	26,5	-3,5	12,25
6	8	-2	4		24	21	3	9
7	2	5	25		25	26,5	-1,5	2,25
8	15,5	-7,5	56,25		26	17,5	8,5	72,25
9	21	-12	144		27	35,5	-8,5	72,25
10	11	-1	1		28	31	-3	9
11	26,5	-15,5	240,25		29	33,5	-4,5	20,25
12	21	-9	81		30	13	17	289
13	3	10	100		31	17,5	13,5	182,25
14	9,5	4,5	20,25		32	26,5	5,5	30,25
15	5,5	9,5	90,25		33	15,5	17,5	306,25
16	31	-15	225		34	21	13	169
17	26,5	-9,5	90,25		35	35,5	-0,5	0,25
18	1	17	289		36	31	5	25
			1525,75					**1502,75**

$$\sum d^2 = 1525.75 + 1502.75 = 3028.5$$

$$r_s = 1 - \frac{6\sum d^2}{N(N^2-1)} = 1 - \frac{6(3028.5)}{36(36^2-1)} = 1 - \frac{18171}{46620} = 0.61$$

$$Z_{1-\frac{\alpha}{2}} = Z_{1-\frac{0.01}{2}} = Z_{0.995} = 2.58$$

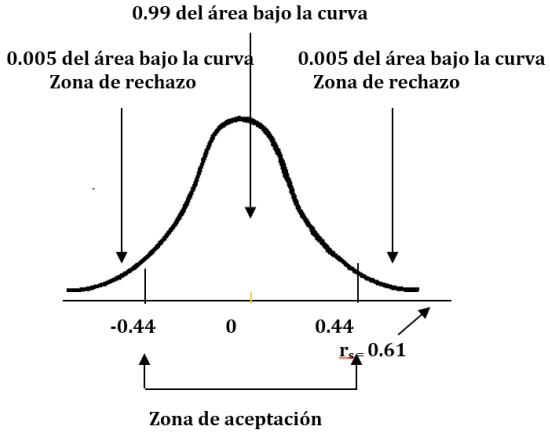
Los límites de las regiones de aceptación y de rechazo son:

$$\mu_{r_s} + Z_{1-\frac{\alpha}{2}}\left(\frac{1}{\sqrt{N-1}}\right) = 0 + 2.58\left(\frac{1}{\sqrt{36-1}}\right) = \frac{2.58}{5.92} = 0.44$$

$$\mu_{r_s} - Z_{1-\frac{\alpha}{2}}\left(\frac{1}{\sqrt{N-1}}\right) = 0 - 2.58\left(\frac{1}{\sqrt{36-1}}\right) = \frac{2.58}{5.92} = -0.44$$

En la figura 14.4 se muestran gráficamente las zonas de aceptación y de rechazo y la ubicación del coeficiente de correlación por rangos calculado.

FIGURA 14.4 Zonas de aceptación y de rechazo



Observe que el valor del coeficiente de correlación por rango r_s = 0.61 se encuentra dentro de la zona de rechazo y por tanto se rechaza la hipótesis nula, es decir, existe una correlación significativa entre las variables estudiadas.

Ejercicios del capítulo

14.1 El dueño de una panadería en la cual el pan se ha estado produciendo mediante la utilización de la tradicional harina de trigo, desea estudiar si a los consumidores de este producto les gustaría más el pan elaborado con la llamada harina de fuerza, la cual sin lugar a dudas, incrementa el valor proteico de dicho producto. Para ello, les dio a consumir a 13 personas pan elaborado con ambos tipos de harina, y posteriormente, les pidió indicaran cuál de los dos era el de su preferencia. Las respuestas de las personas consultadas fueron las siguientes:

Personas	**Preferencia**
1	Harina de fuerza
2	Harina de fuerza
3	Harina de trigo
4	Harina de fuerza
5	Harina de trigo
6	Harina de fuerza
7	Harina de fuerza
8	Harina de fuerza
9	Harina de trigo
10	Harina de trigo
11	Harina de fuerza
12	Harina de fuerza
13	Harina de trigo

Con un nivel de significación del 5%, ¿podemos arribar a la conclusión que existen diferencias significativas en cuanto a la preferencia de los consumidores por un determinado tipo de pan?

Utilice para desarrollar la prueba la distribución binomial.

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el o los valores críticos correspondientes de la distribución binomial
d. Decida el rechazo o no de la hipótesis nula

14.2 La tabla que aparece a continuación muestra la categoría docente de 15 profesores de una Facultad de Administración durante el año lectivo 2014 – 2015 y la correspondiente al año lectivo 2015 – 2016 después de haber sido sometido a un proceso escalafonario.

	Año lectivo	
Profesores	**2014 - 2015**	**2015 - 2016**
1	Auxiliar	Agregado
2	Agregado	Agregado
3	Auxiliar	Agregado
4	Agregado	Principal
5	Agregado	Principal
6	Agregado	Agregado
7	Auxiliar	Agregado
8	Agregado	Principal
9	Auxiliar	Agregado
10	Agregado	Agregado
11	Auxiliar	Agregado
12	Auxiliar	Agregado
13	Agregado	Principal
14	Agregado	Principal
15	Auxiliar	Auxiliar

Con un nivel de significación del 1%, ¿podemos arribar a la conclusión que existió una mejoría de la categoría docente de los profesores después del proceso escalafonario?

Utilice para desarrollar la prueba la distribución binomial.

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el o los valores críticos correspondientes de la distribución binomial
d. Decida el rechazo o no de la hipótesis nula

14.3 Con los mismos datos del ejercicio 14.1, y con un nivel de significación del 1%, ¿podemos arribar a la conclusión que existe una mayor preferencia de los consumidores por el pan elaborado con harina de fuerza?

a. Compruebe si para desarrollar la prueba puede ser utilizada la distribución normal, y en caso afirmativo, haga uso de ella.
b. Formule las hipótesis nula y alternativa
c. Obtenga el valor del estadístico de prueba
d. Determine el valor crítico correspondiente de la distribución normal
e. Decida el rechazo o no de la hipótesis nula

14.4 Con los mismos datos del ejercicio 14.2, y con un nivel de significación del 5%, ¿podemos llegar a la conclusión que hubo una mejoría en las categorías docentes

de los profesores después del proceso escalafonario?

a. Compruebe si para desarrollar la prueba puede ser utilizada la distribución normal, y en caso afirmativo, haga uso de ella.
b. Formule las hipótesis nula y alternativa
c. Obtenga el valor del estadístico de prueba
d. Determine el valor crítico correspondiente de la distribución normal
e. Decida el rechazo o no de la hipótesis nula

14.5 Una cooperativa de taxis turísticos de una importante ciudad de Ecuador tiene la intención que sus conductores mejoren sus niveles de información acerca de los principales sitios visitados por los turistas, con el objetivo de que les puedan brindar una mayor cantidad de datos sobre los mismos. Con esta finalidad, evaluó mediante un examen los conocimientos de 15 de sus conductores antes y después de que estos recibieran un curso de información turística acerca de la ciudad. El examen de referencia fue calificado sobre 10 y los resultados se muestran a continuación:

	Evaluación	
Conductores	**Antes**	**Después**
1	8.6	9.1
2	8.2	8.2
3	9.1	8.9
4	9.3	9.6
5	8.8	8.7
6	8.2	8.4
7	9.6	9.6
8	9.3	9.1
9	8.4	8.8
10	8.9	9.3
11	9.2	9.6
12	9.6	9.5
13	9.3	9.1
14	8.2	8.8
15	8.7	8.9

Con un nivel de significación del 1%, ¿existen evidencias que nos permitan llegar a la conclusión que el curso recibido por los conductores logró mejorar sus conocimientos acerca de los sitios de afluencia de turistas en la ciudad?

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el valor real del tamaño de la muestra
d. Determine el valor crítico de la prueba

e. Decida el rechazo o no de la hipótesis nula

14.6 Con el objetivo de medir el efecto de una campaña publicitaria sobre la cantidad de artículos vendidos de un determinado producto, se anotaron las ventas en miles de unidades en 12 supermercados antes de poner en práctica la campaña y dos semanas después de haberse realizado. Los resultados del estudio se muestran a continuación:

	Ventas	
Supermercados	**Antes**	**Después**
1	2.4	2.7
2	3.1	3.1
3	3.5	3.7
4	2.9	2.8
5	3.3	3.3
6	3.6	3.9
7	2.8	3.2
8	2.7	2.7
9	3.1	3.4
10	3.5	3.6
11	2.8	2.7
12	3.1	3.4

Con un nivel de significación del 5%, ¿podemos concluir que la campaña publicitaria logró incrementar las ventas del producto?

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el valor real del tamaño de la muestra
d. Determine el valor crítico de la prueba
e. Decida el rechazo o no de la hipótesis nula

14.7 La Policía Nacional desea comprobar si existe una diferencia significativa entre la cantidad de delitos que se cometen por día en las ciudades de Guayaquil y Manta, y para ello, recopiló durante 10 días consecutivos los delitos cometidos en ambas ciudades. Los resultados obtenidos se muestran a continuación:

Ciudad	**Delitos cometidos por día**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Manta	21	20	19	23	24	21	22	20	19	20
Guayaquil	23	21	18	26	23	19	26	24	20	18

Considerando que las varianzas de las dos poblaciones no son iguales y utilizando la prueba de Wilcoxon de la suma de rangos para muestras independientes, ¿existen razones que nos permitan asegurar con un nivel de significación del 5% que en la ciudad de Manta se cometen menos delitos que en la ciudad de Guayaquil?

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el valor crítico de la prueba
d. Decida el rechazo o no de la hipótesis nula

14.8 Se sometió a prueba el efecto que produce sobre los volúmenes de venta de un determinado artículo el hecho de que en la instalación donde éste se oferta, exista o no equipo de climatización. Para ello se midieron las ventas del artículo en 12 instalaciones sin aire acondicionado y en otras 12 con aire acondicionado. Los resultados de la investigación se muestran a continuación:

	Instalaciones											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Sin aire	34	41	32	40	38	37	33	39	42	35	38	34
Con aire	41	45	33	44	36	39	31	43	41	38	44	39

Considerando varianzas poblaciones desiguales utilice la prueba de Wilcoxon de la suma de rangos para muestras independientes, para establecer si existen diferencias significativas al 1% entre las ventas en instalaciones sin y con aire acondicionado.

a. Formule las hipótesis nula y alternativa
b. Obtenga el valor del estadístico de prueba
c. Determine el valor crítico de la prueba
d. Decida el rechazo o no de la hipótesis nula

14.9 Utilizando los datos del ejercicio 14.7, aplique la prueba U de Mann – Whitney para establecer con una significación del 5 % si existen diferencias significativas entre la ciudad de Manta y la ciudad de Guayaquil en cuanto al número diario de delitos cometidos.

14.10 Utilizando los datos del ejercicio 14.8, aplique la prueba U de Mann – Whitney para establecer con una significación del 1 % si existen diferencias significativas entre la ciudad de Manta y la ciudad de Guayaquil en cuanto al número diario de delitos cometidos.

14.11 Considere que los datos de la investigación planteada en el ejercicio 14.7 fueron los que se muestran a continuación:

Ciudad	Delitos cometidos por día											
	1	2	3	4	5	6	7	8	9	10	11	12
Manta	21	20	19	23	24	21	22	20	19	20	22	19
Guayaquil	23	21	18	26	23	19	26	24	20	18	21	24

Utilice la prueba U de Mann – Whitney para establecer con una significación del 1 % si existen diferencias significativas entre la ciudad de Manta y la ciudad de Guayaquil en cuanto al número diario de delitos cometidos. Aproxime el estadístico de prueba U mediante la distribución normal.

14.12 Un profesor universitario desea comparar los resultados obtenidos por estudiantes que cursan la materia de Estadística cuando son sometidos a un examen oral o cuando se les realiza un examen escrito. Para ello les aplicó un examen oral a 14 estudiantes de un paralelo y un examen escrito a otros 14 estudiantes del mismo grupo. Las calificaciones obtenidas fueron las siguientes:

Examen	Calificaciones													
Oral	7.8	8.4	9.2	7.9	8.8	9.4	7.6	7.4	8.2	9.1	9.5	8.1	8.6	7.7
Escrito	8.1	8.6	8.8	7.7	9.2	9.4	7.5	7.8	8.5	9.1	9.3	8.4	8.7	7.6

Utilice la prueba U de Mann – Whitney para establecer con una significación del 5 % si existen diferencias significativas entre las calificaciones obtenidas en ambos tipos de examen. Aproxime el estadístico de prueba U mediante la distribución normal.

14.13 Un investigador desea someter a prueba si existen diferencias significativas en cuanto al tiempo requerido para ensamblar un equipo industrial utilizando tres métodos diferentes. Para desarrollar el trabajo, formó 3 grupos de 8 trabajadores y a cada grupo se le asignó un método de ensamblaje específico. El tiempo en horas empleado por cada trabajador en cada grupo para ensamblar el equipo se muestra a continuación:

Método	Tiempo de ensamblaje							
A	14.4	14.1	14.6	14.3	14.8	14.1	14.5	14.3
B	15.1	14.8	14.7	15.2	15.3	14.9	14.7	15.2
C	13.9	13.6	14.1	14.3	13.8	13.7	14.1	14.2

Utilice la prueba de Kruskal – Wallis para establecer si existen diferencias significativas al 5% entre los tres métodos de ensamblaje. Emplee la distribución Ji – Cuadrada para obtener el valor crítico de la prueba.

14.14 En una empresa agropecuaria se sometió a prueba el efecto de 4 diferentes sistemas de alimentación sobre la producción de leche de vacas de mediano potencial. Los resultados de la investigación se muestran a continuación:

Sistema	Producción de leche									
A	13.8	13.6	13.9	13.5	13.4	13.7	13.3	13.5	13.1	13.3
B	14.3	14.2	14.6	14.1	14.8	14.4	14.2	14.8	14.6	14.2
C	13.6	13.4	13.9	13.4	13.6	13.5	13.1	13.4	13.8	13.5
D	15.2	15.3	15.1	15.7	15.4	15.5	15.2	15.3	15.4	15.7

Utilice la prueba de Kruskal – Wallis para establecer si existen diferencias significativas al 5% entre los cuatro sistemas de alimentación. Emplee la distribución Ji – Cuadrada para obtener el valor crítico de la prueba.

14.15 Una compañía está desarrollando entrevistas a profesionales del área económica con el objetivo de seleccionar a un hombre (H) o a una mujer (M) para un cargo financiero. El orden declarado por la compañía en que fueron entrevistados los profesionales fue el siguiente:

H	H	M	H	M	M	M	H	H	M	H	H
M	M	M	H	H	H	H	M	H	H	M	M
M	H	M	H	H	M	M	M	H	M	M	H
M	M	H	H	H	M	M	M	H	M	M	M

Con un nivel de significación del 5%, ¿podemos aceptar que la secuencia de entrevistas declarada por la compañía tiene un orden aleatorio? Aproxime la distribución de muestreo del número de corridas a una distribución normal.

14.16 Considere que en el ejercicio anterior el orden declarado por la compañía en que fueron entrevistados los profesionales fue el siguiente:

H	H	M	H	M	M	M	H	H	M	H	H
M	M	M	H	H	H	H	M	H	H	M	M
M	H	M	H	H	M	M	M	H	M	M	H

Con un nivel de significación del 1%, ¿podemos aceptar que la secuencia de entrevistas declarada por la compañía tiene un orden aleatorio? Utilice las tablas de valores críticos para la prueba de rachas.

14.17 Los datos que se muestran a continuación representan el orden de llegada a la meta (X) y el peso corporal (Y) de 36 corredores en una competencia de 5000 metros planos.

X	1	2	3	4	5	6	7	8	9	10	11	12
Y	60	62	58	63	61	65	63	68	70	68	71	71
X	13	14	15	16	17	18	19	20	21	22	23	24
Y	69	70	72	75	71	73	75	76	75	77	80	79
X	25	26	27	28	29	30	31	32	33	34	35	36
Y	78	79	80	78	78	82	80	81	84	82	81	82

Con un nivel de significación del 0.1%, ¿existe una correlación significativa entre ambas variables? Utilice la distribución normal como aproximación a la distribución de muestreo de r_s.

14.18 Considere que los datos del ejercicio anterior son los que se muestran a continuación:

X	1	2	3	4	5	6	7	8	9	10	11	12
Y	60	62	58	63	61	65	63	68	70	68	71	71

Con un nivel de significación del 1%, ¿existe una correlación significativa entre ambas variables? Utilice la tabla de valores críticos del coeficiente de correlación por rango de Spearman.

Capítulo 15
Series cronológicas

El problema

Con el objetivo de estudiar la evolución del número de sus usuarios con conexión a banda ancha en una importante ciudad del país, la Corporación Nacional de Telecomunicaciones dispone de un listado por semestre de los últimos 5 años de la cantidad de clientes que hacen uso de ese servicio. La corporación está interesada en pronosticar el número de usuarios con conexión a banda ancha en el segundo semestre del año 2015. ¿Resulta estadísticamente posible hacer este pronóstico?

15.1 Introducción.

En cualquier proceso de toma de decisiones en el que estemos inmersos, las predicciones o pronósticos se convierten en un instrumento de vital importancia para su correcta ejecución.

Sea para estimar el índice inflacionario para el año 2016, o cuales van a ser las necesidades de transportación pública o los requerimientos de equipos de línea blanca para ese año, o cualquier otro tipo de estimación del comportamiento de una característica a futuro, las predicciones o pronósticos se convierten en una herramienta fundamental para lograr estos propósitos.

Sin embargo, la precisión con la que estas predicciones a futuro puedan hacerse dependerá mucho de la calidad y cantidad de información de que dispongamos a través del tiempo.

El análisis de series cronológicas, series de tiempo o series temporales es un método estadístico que nos permite descubrir *patrones de comportamiento* de datos recolectados a intervalos regulares a través del tiempo, para con ello, proyectar estos patrones y obtener un *pronóstico de comportamiento a futuro*.

Resumiendo lo expresado hasta el momento, podemos concluir que ***una serie cronológica es cualquier conjunto de observaciones ordenadas según transcurre el tiempo, este último medido en intervalos regulares***.

El presente capítulo lo dedicaremos al estudio de datos que nos permitan proyectar el futuro, primero evaluando los componentes de una serie cronológica, segundo estudiando procedimientos para la valoración de los datos y por último, proyectando comportamientos futuros.

15.2 Componentes de una serie cronológica.

Existen cuatro componentes, tipos de cambio o variaciones inmersas en el análisis de una serie cronológica. Estas son:

1. **Tendencia secular**
2. **Fluctuación cíclica**
3. **Variación temporal o estacional**
4. **Variación irregular**

Estudiemos brevemente cada una de estas componentes o variaciones para posteriormente hacerlo con más detalle.

- **Tendencia secular**

Este tipo de variación se refiere a la dirección que sigue la serie cronológica en un largo periodo de tiempo, la cual se puede apreciar con bastante claridad si elaboramos un gráfico de líneas utilizando para ello los datos de la serie.

El estudio de la tendencia secular de una serie cronológica es lo que nos permitirá pronosticar el probable comportamiento de los datos en el futuro, el cual podrá estar sujeto a un sesgo en dependencia de la calidad de los datos y del método estadístico utilizado para el análisis de la tendencia.

Existen series cronológicas en las cuales sus valores muestran una tendencia ascendente, mientras que otras poseen una tendencia descendente. La proyección de una serie cronológica es una herramienta de extraordinaria importancia en los procesos de planificación a mediano y largo plazo.

La tendencia de una serie es factible de ser analizada utilizando para ello dos métodos en específico. Estos métodos son *el gráfico y el analítico*.

El método gráfico o empírico consiste en ajustar de manera visual una recta a los pares de observaciones de la serie que han sido ubicados en un gráfico.

Dado la subjetividad de este método al depender del criterio personal de la persona que hace el ajuste, hace que el mismo no sea muy aconsejable.

*El m*étodo *analítico* consiste en ajustar una línea de tendencia mediante el ya conocido método de los mínimos cuadrados. Este método es sin lugar a dudas el más utilizado.

- **Fluctuación cíclica**

El comportamiento común de cualquier ciclo de negocios está compuesto por etapas de prosperidad, a las cuales les siguen periodos de recesión, de depresión y posteriormente de recuperación. Es habitual observar como en etapas de prosperidad

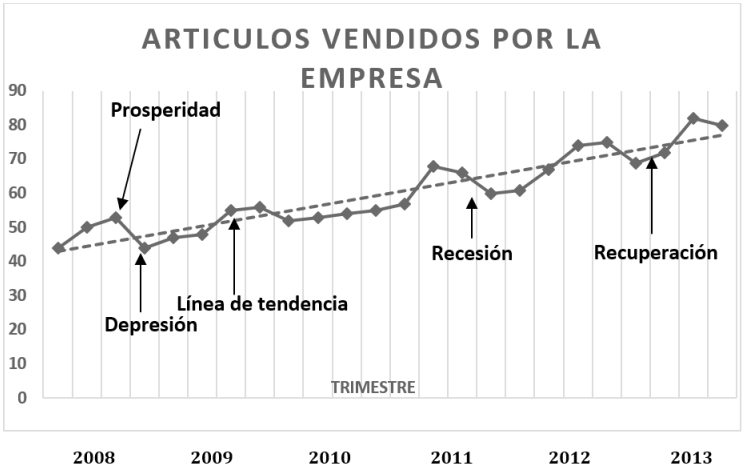
del ciclo de negocios los índices de empleo, por ejemplo, se encuentran por encima de la línea de tendencia a largo plazo, mientras en períodos de recesión se encuentran por debajo. La tabla 15.1 muestra la cantidad de artículos vendidos trimestralmente por una empresa entre los años 2008 y 2013.

TABLA 15.1 Cantidad de artículos vendidos por trimestre

	TRIMESTRE			
AÑO	**1**	**2**	**3**	**4**
2008	44	50	53	46
2009	47	48	55	56
2010	52	53	54	55
2011	57	66	64	60
2012	61	67	73	74
2013	69	72	77	78

Observe en la figura 15.1 las etapas que caracterizan a un ciclo de negocios, es decir, los períodos de prosperidad seguidos de recesión, depresión y recuperación.

FIGURA 15.1 Fluctuación cíclica



- **Variación temporal o estacional**

Es bien conocido por los especialistas en el tema que en dependencia de la temporada del año donde nos encontremos, hay productos y servicios que presentan una fluctuación en su demanda. Así por ejemplo, CopaAirlines tiene una buena demanda de pasajeros ecuatorianos para Cuba en los meses de Abril, Mayo y Junio por el inicio de la época de playa, incrementándose en Julio, Agosto y Septiembre por ser la etapa de mejor clima para la playa, decayendo de forma significativa en los meses de Octubre, Noviembre y Diciembre debido a la época ciclónica y las fiestas navideñas, para volver a incrementarse en los meses de Enero, Febrero y Marzo dado el clima favorable en la isla.

Un tipo de comportamiento similar al descrito anteriormente recibe el

nombre de *Variación temporal o estacional* el cual tiene la característica de ser un movimiento repetitivo y predecible en un periodo de un año o menos, es decir, que en el ejemplo anterior CopaAirlines enfrentará todos los años una situación similar a la descrita en sus vuelos de Ecuador a Cuba.

Representemos los trimestres como:

A-M-J (Abril, Mayo y Junio)

J-A-S (Julio, Agosto, Septiembre)

O-N-D (Octubre, Noviembre, Diciembre)

E-F-M (Enero, Febrero, Marzo)

El comportamiento en un determinado año de la demanda de pasajeros ecuatorianos a Cuba podría presentarse tal y como se muestra en la figura 15.2.

FIGURA 15.2 Demanda de pasajeros ecuatorianos a Cuba

Por supuesto, la variación temporal de la demanda de pasajeros a Cuba puede variar de un año a otro y comportarse de la manera que se muestra en la figura 15.3.

FIGURA 15.3 Demanda de pasajeros ecuatorianos a Cuba

- **Variación irregular.**

La variación irregular se debe a factores que se presentan a corto plazo y que tienen las características de ser impredecibles y no recurrentes. La variación irregular puede estar provocada por factores con características distintivas, por ejemplo, fenómenos naturales como huracanes y terremotos, revueltas sociales, epidemias, etc. También pueden estar provocadas por causas no identificables y que son aleatorias y producto de la casualidad.

Un ejemplo que permite describir estos comportamientos inusuales es la epidemia de ébola que se inició en marzo del 2014, la cual ha causado una impresionante cantidad de muertos y de personas infectadas con la enfermedad. Esta epidemia ha provocado fenómenos inusuales tales como la prohibición de venta y consumo de animales de caza en Guinea, el cierre de fronteras en Sierra Leona, los ministros de salud de muchos países de África Occidental acuerdan estrategias internacionales, y un sinnúmero más de acciones provocadas por un factor impredecible y de efectos a corto plazo.

A continuación pasaremos a estudiar de forma más profunda cada uno de los componentes de una serie cronológica, pero antes, debemos dejar claro que algunas de las suposiciones que haremos en los próximos párrafos no siempre están presentes en algunas series reales. Cuando estas suposiciones no se cumplen, es necesario hacer uso de métodos matemáticos de mayor complejidad que están por encima del nivel de los objetivos de este libro.

15.3 Análisis de tendencia.

En párrafos anteriores señalamos que la tendencia secular se refiere a la dirección que sigue la serie cronológica en un largo periodo de tiempo y que su estudio es lo que nos permite pronosticar el probable comportamiento de los datos en el futuro.

Vimos también que el método más adecuado para ajustar una tendencia es el método de los mínimos cuadrados.

- **Ajuste de una tendencia lineal**

Como ya conocemos la ecuación que representa una línea recta viene dada por la expresión:

$\hat{Y} = a + b\,X$ donde

$\hat{Y}$ **=** valor estimado de la variable dependiente

X = variable independiente (tiempo)

a = traza o intersección con el eje Y

b = pendiente de la línea de tendencia

En el Capítulo 11 al aplicar el método de los mínimos cuadrados obtuvimos que los valores de a y b de la ecuación de mejor ajuste vienen dados por las expresiones:

$$b = \frac{\sum X_i Y_i - \dfrac{\sum X_i \sum Y_i}{n}}{\sum X_i^2 - \dfrac{\left(\sum X_i\right)^2}{n}} = \frac{SPC_{XY}}{SCC_X}$$

$$a = \overline{Y} - b\,\overline{X}$$

Sin embargo, si tomamos en cuenta que en una serie cronológica la variable independiente es el tiempo, y que por regla general éste está medido en términos de semanas, meses o años, podríamos convertir estas unidades de tiempo de forma tal que los cálculos requeridos para obtener los valores de *a* y *b* se simplifiquen notablemente.

Supongamos para ejemplificar lo expuesto en el párrafo anterior, que en una serie cronológica en particular, la variable independiente X está medida en años y tiene los valores 2009, 2010, 2011, 2012 y 2013. Observe que la media de estos 5 años es igual a 2011, entonces podríamos *codificar* la variable X y convertirla en los valores que se muestran en la tabla 15.2.

TABLA 15.2 Valores de X codificados

X	$X - \overline{X}$	**x**
2009	**2009-2011**	**-2**
2010	**2010-2011**	**-1**
2011	**2011-2011**	0
2012	2012-2011	1
2013	2013-2011	2
	Suma	0

donde **x** (x minúscula) representa los valores de la nueva variable *X codificada*.

Observe que la nueva variable *X codificada* tiene una suma y por tanto una media igual a 0, y en consecuencia,

$$b = \frac{\sum x_i\, y_i - \frac{(0)\sum y_i}{n}}{\sum x_i^2 - \frac{(0)^2}{n}}$$

$$a = \overline{Y} - b(0)$$

$$b = \frac{\sum x_i y_i}{\sum x_i^2} \quad a = \overline{Y}$$

Las expresiones anteriores reducen de forma importante los cálculos a realizar, sobre todo, si tomamos en cuenta que los valores de x_i son -2, -1, 0, 1, y 2.

En el ejemplo anterior la cantidad de años de la serie cronológica era un número impar. Cuando este número de años es una cantidad par, entonces el procedimiento para la *codificación* de la variable independiente X varía un poco. Supongamos que los años son 2008, 2009, 2010, 2011, 2012 y 2013. En este caso la media es 2010.5 y la codificación quedaría como se muestra en la tabla 15.3.

TABLA 15.3 Valores de la variable X codificados

X	$X - \overline{X}$	**x**	**x por 2**
2008	**2008-2010.5**	**-2.5**	**-5**
2009	**2009-2010.5**	-1.5	-3
2010	2010-2010.5	-0.5	-1
2011	2011-2010.5	0.5	1
2012	2012-2010.5	1.5	3
2013	2013-2010.5	2.5	5
	Suma	0	0

Observe que con el objetivo de simplificar la codificación, los valores de x fueron multiplicados por 2, con lo que logramos también eliminar los valores decimales.

Calculemos a continuación una línea de tendencia mediante el método de los mínimos cuadrados y la codificación de la variable independiente.

En la tabla 15.4 se muestra el número de conductores de vehículos privados que pagaron peaje en una determinada carretera de la provincia del Guayas entre los años 2006 y 2013 medido en miles, y los cálculos necesarios para obtener la línea de tendencia correspondiente.

TABLA 15.4 Cantidad de conductores que pagaron peaje (miles)

X	Y	$X - \overline{X}$		x	xY	x^2
2006	3.5	2006-2009.5	-3.5	-7	-24.5	49
2007	5.1	2007-2009.5	-2.5	-5	-25.5	25
2008	4.6	2008-2009.5	-1.5	-3	-13.8	9
2009	4.8	2009-2009.5	-0.5	-1	-4.8	1
2010	6.1	2010-2009.5	0.5	1	6.1	1
2011	5.6	2011-2009.5	1.5	3	16.8	9
2012	6.8	2012-2009.5	2.5	5	34	25
2013	6.3	2013-2009.5	3.5	7	44.1	49
16076	42.8		0	0	32.4	168

$$b = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{32.4}{168} = 0.19$$

$$a = \overline{Y} = \frac{42.8}{8} = 5.35$$

La línea de tendencia de mejor ajuste viene dada por la ecuación:

$$\hat{Y} = 5.35 + 0.19x$$

Supongamos ahora que deseamos pronosticar el número de conductores de vehículos privados que pagarán peaje en la carretera de referencia en el año 2014. El primer paso consiste en codificar el año.

$$x = 2\left(X - \overline{X}\right) \qquad x = 2\left(2014 - 2009.5\right) = 9$$

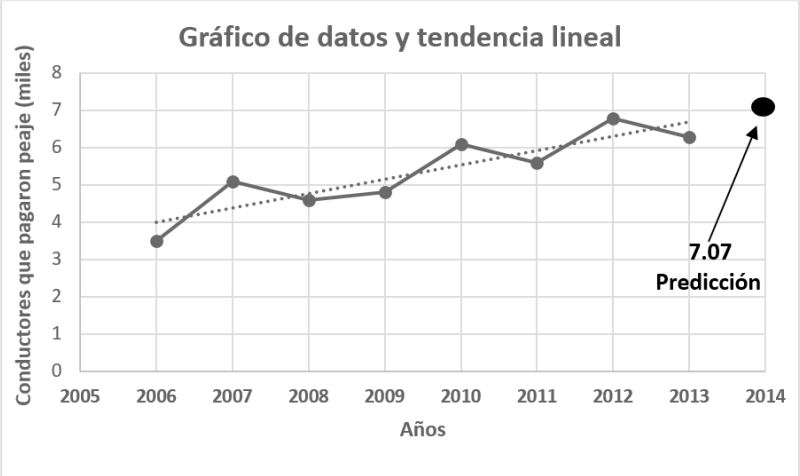
Sustituyendo este valor en la ecuación de la línea de tendencia obtenida, tenemos

$$\hat{Y} = 5.35 + 0.19x = 5.35 + 0.19\left(9\right) = 5.35 + 1.71 = 7.06$$

El número de conductores de vehículos privados que se espera paguen peaje en el año 2014 es de 7060.

En la figura 15.4 se muestra el gráfico de los datos y la tendencia lineal.

FIGURA 15.4 Gráfico de datos y tendencia lineal



- **Ajuste de una tendencia cuadrática**

En bastantes ocasiones sucede que la tendencia de una serie cronológica no puede ser descrita mediante una línea recta, y se hace entonces necesario ajustar una ecuación de segundo grado. Como ya conocemos, la expresión matemática de este tipo de ecuación viene dada por:

$$\hat{Y} = a + b\,X + c\,X^2$$

y utilizando valores codificados para la variable independiente:

$$\hat{Y} = a + b\,x + c\,x^2$$

En el Capítulo 12 obtuvimos el sistema de ecuaciones normales que nos permite obtener los valores de **a**, **b**, y **c** para una ecuación cuadrática.

Este sistema de ecuaciones con la variable independiente codificada viene dado por la expresión:

$$\sum Y_i = na + b\sum x_i + c\sum x_i^2$$

$$\sum x_i Y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3$$

$$\sum x_i^2 Y_i = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4$$

Al codificar la variable independiente X, tanto la sumatoria de las x_i como la sumatoria de las x_i^3 son iguales a 0, por tanto, el sistema de ecuaciones normales queda de la forma que se muestra a continuación:

$$\sum Y_i = na + c\sum x_i^2$$

$$\sum x_i Y_i = b\sum x_i^2 \longrightarrow b = \frac{\sum x_i Y_i}{\sum x_i^2}$$

$$\sum x_i^2 Y_i = a\sum x_i^2 + c\sum x_i^4$$

Veamos un ejemplo. En la tabla 15.5 se muestran los gastos en publicidad

(millones de dólares) entre los años 2007 y 2013 de una importante compañía trasnacional. La tabla 15.6 muestra los cálculos necesarios para realizar el ajuste de una tendencia cuadrática.

TABLA 15.5 Gastos en publicidad

2007	2008	2009	2010	2011	2012	2013
80.7	82.6	87.3	99.3	114.7	140.4	170.4

TABLA 15.6 Cálculos requeridos para un ajuste cuadrático

Y	**X**		**x**	**x^2**	**x^4**	**xY**	x^2Y
80.7	2007	2007 - 2010	-3	9	81	-242.1	726.3
82.6	2008	2008 - 2010	-2	4	16	-165.2	330.4
87.3	2009	2009 - 2010	-1	1	1	-87.3	87.3
99.3	2010	2010 - 2010	0	0	0	0	0
114.7	2011	2011 - 2010	1	1	1	114.7	114.7
140.4	2012	2012 - 2010	2	4	16	280.8	561.6
170.4	2013	2013 - 2010	3	9	81	511.2	1533.6
775.4			0	28	196	412.1	3353.9

Utilizando las ecuaciones normales, $b = \frac{\sum x_i Y_i}{\sum x_i^2} = \frac{412.1}{28} = 14.72$

$\sum Y_i = na + c\sum x_i^2 \qquad 775.4 = 7a + 28c$

$\sum x_i^2 Y_i = a\sum x_i^2 + c\sum x_i^4 \quad 3353.9 = 28a + 196c$

Resolvamos el anterior sistema de dos ecuaciones con dos incógnitas.

775.4 = 7a + 28 c x (-4)

3353.9 = 28a + 196 c

-3101.6 = -28a - 112 c

3353.9 = 28 a + 196 c

252.3 = 84 c $\quad c = \frac{252.3}{84} = 3$

Sustituyendo el valor de c:

775.4 = 7a +28 (3) 7a = 775.4 – 84 = 691.4 a = 98.77

La ecuación cuadrática que describe la tendencia de la serie cronológica cuyos datos fueron reportados en la tabla 15.5 es:

$$\hat{Y} = 98.77 + 14.72\,x + 3\,x^2$$

Si quisiéramos, por ejemplo, predecir los gastos en publicidad de la compañía para el año 2014, debemos en primer término codificar este año.

$$x = X - \overline{X} = 2014 - 2010 = 4$$

Sustituyendo este valor en la ecuación cuadrática:

$$\hat{Y} = 98.77 + 14.72(4) + 3(4)^2 = 98.77 + 58.88 + 48 = 205.65$$

Es decir, se espera que para el año 2014 la compañía gaste aproximadamente 205.65 millones de dólares en publicidad.

En el ejemplo anterior, y en cualquier otro en que utilicemos una ecuación de segundo grado para describir la tendencia de una serie cronológica, debemos tomar en cuenta que en la medida en que los años van pasando es posible que las necesidades de publicidad vayan disminuyendo, y en consecuencia, la tendencia no sería bien descrita a través de la ecuación cuadrática. Es decir, que al hacer predicciones debemos tener siempre presente que la tendencia de la serie cronológica puede cambiar.

Otro aspecto a tomar en cuenta cuando hacemos predicciones en una serie cronológica, es que en realidad la validez del ajuste de una ecuación de estimación es solo dentro del intervalo en que la muestra fue tomada. Claro que en una serie cronológica se requiere hacer estimaciones uno o dos periodos de tiempo más allá del intervalo de la muestra, pero en todo caso, no nos debemos exceder en hacer predicciones cuyos periodos de tiempo estén muy hacia el futuro.

En este numeral nos hemos limitado a estudiar dos tipos de tendencia, la lineal y la cuadrática. Sin embargo, de manera general, la tendencia de una serie cronológica puede tener otras formas diferentes que deberán ser tratadas de forma particular.

15.4 Fluctuación cíclica.

La fluctuación o variación cíclica, como ya expresamos, es la componente de una serie cronológica que muestra en periodos mayores a un año, un comportamiento oscilatorio por encima y por debajo de la línea de tendencia secular. El componente cíclico de una serie cronológica puede ser interpretado como el que persiste después de haber sido eliminado el efecto de la componente de tendencia. Cuando estamos trabajando con una serie cronológica cuya variable independiente son años, podemos expresar la fluctuación o variación cíclica como un porcentaje de la tendencia. Para esto, basta con dividir el verdadero valor de la variable dependiente Y entre el correspondiente valor estimado de tendencia $\hat{Y}$ multiplicando posteriormente por

100, es decir, $Porcentaje\ de\ tendencia = \frac{Y}{\hat{Y}} x100$

Calculemos el porcentaje de tendencia correspondiente a los datos de la tabla 15.4 (se observan en la tabla 15.7) en la que se mostró el número de conductores de vehículos privados que pagaron peaje en una determinada carretera de la provincia del Guayas entre los años 2006 y 2013 medido en miles.

TABLA 15.7 Cálculos requeridos para obtener los porcentajes de tendencia

X	**Y**	**x**	$\hat{Y}$	$Y/\hat{Y}\,x100$	$\left(Y/\hat{Y}\,x100\right)$-100
2006	3,5	-7	4,02	87,06	-12,94
2007	5,1	-5	4,4	115,91	15,91
2008	4,6	-3	4,78	96,23	-3,77
2009	4,8	-1	5,16	93,02	-6,98
2010	6,1	1	5,54	110,11	10,11
2011	5,6	3	5,92	94,59	-5,41
2012	6,8	5	6,3	107,94	7,94
2013	6,3	7	6,68	94,31	-5,69

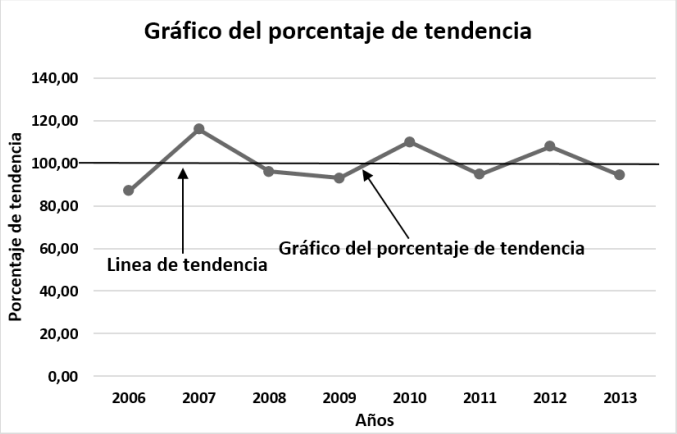
En la tabla 15.7 además del porcentaje de tendencia se obtuvo la expresión $\left(Y/\hat{Y}\,x100\right)$ **-** 100 la cual se conoce como *residuo cíclico relativo*. Observe que este valor se obtiene restando 100 del porcentaje de tendencia.

A modo de interpretación, se puede observar que en el año 2006 el porcentaje de tendencia indica que el número real de conductores que pagaron peaje fue el 87.06% de lo esperado, y sin embargo, en el año 2010 fue del 110.11%.

Lo dicho anteriormente es equivalente a expresar que en el año 2006 el residuo cíclico relativo indica que el número real de conductores que pagaron peaje estuvo un 12.94% por debajo de lo esperado, y sin embargo, en el año 2010 estuvo un 10.11% por encima.

En la figura 15.5 se muestra el gráfico de los porcentajes de tendencia obtenidos.

FIGURA 15.5 Gráfico del porcentaje de tendencia



Antes de concluir el estudio de las fluctuaciones cíclicas es necesario precisar con toda claridad que los métodos expuestos anteriormente solo son válidos para variaciones cíclicas pasadas, y de ninguna forma, para estimar variaciones futuras. Como ya expresamos en otros casos, la predicción de fluctuaciones cíclicas hace uso de métodos que no están dentro de los objetivos de este libro.

15.5 Variación temporal o estacional.

En párrafos anteriores estudiamos que la variación temporal o estacional tiene la característica de ser un movimiento repetitivo y predecible alrededor de la línea de tendencia en un periodo de un año o menos. El método por excelencia para medir la variación estacional está basado en el llamado *promedio móvil*, el cual permite además, observar la tendencia de una serie cronológica. Por esta razón, pasamos de inmediato a estudiar en qué consiste este método y para ello nos auxiliaremos de los datos que se muestran en la tabla 15.8 los cuales representan el número de artículos vendidos por una empresa en miles de unidades en los últimos 20 años.

TABLA 15.8 Número de artículos vendidos por una empresa

1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
4	6	7	6	5	5	7	8	7	6
2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
6	8	9	8	7	7	9	10	9	8

Observe en la tabla de referencia las siguientes regularidades:

a. El ciclo se repite cada cinco años.

El 2do ciclo (1999-2003) es igual al 1er ciclo (1994-1998) +1 (mil unidades)
El 3er ciclo (2004-2008) es igual al 2do ciclo (1999-2003) +1 (mil unidades)
El 4to ciclo (2009-2013) es igual al 3er ciclo (2004-2008) +1 (mil unidades)

b. La amplitud de cada ciclo es igual a 3.

Amplitud del 1er ciclo = Máximo – Mínimo = 7 – 4 = 3

Amplitud del 2do ciclo = Máximo – Mínimo = 8 – 5 = 3

Amplitud del 3er ciclo = Máximo – Mínimo = 9 – 6 = 3

Amplitud del 4to ciclo = Máximo – Mínimo = 10 – 7 = 3

Las regularidades señaladas anteriormente con relación a los ciclos y a la amplitud de los mismos se ilustran en la tabla 15.9.

TABLA 15.9 Ciclos y amplitud

	1er ciclo	2do ciclo	3er ciclo	4to ciclo
	4	5	6	7
	6	7	8	9
	7	8	9	10
	6	7	8	9
	5	6	7	8
Mínimo	4	5	6	7
Máximo	7	8	9	10
Amplitud	3	3	3	3

Cuando lo anterior ocurre, es decir, cuando la duración de los ciclos es una constante y la amplitud de los mismos es la misma, el promedio móvil elimina de la serie cronológica las componentes cíclicas e irregulares, quedando como resultado una tendencia lineal.

Procedamos a calcular el promedio móvil de cinco años con los datos de la serie cronológica que estamos estudiando.

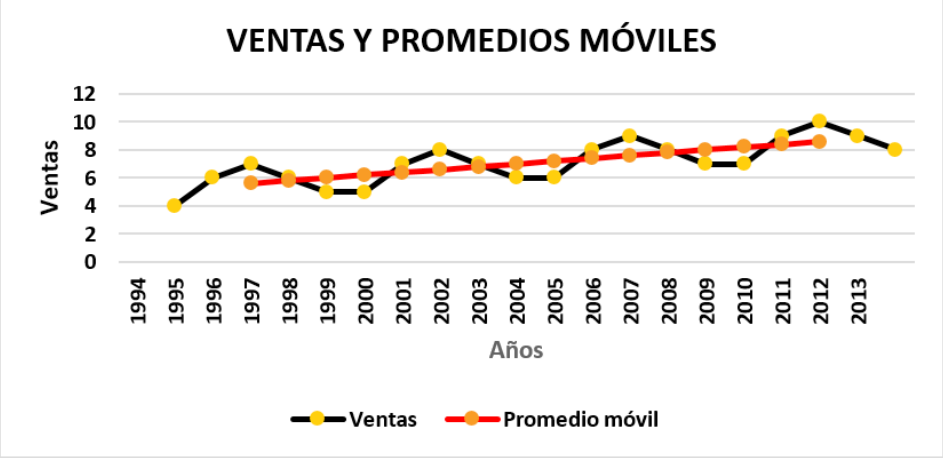
El primer paso consiste en obtener los *totales móviles* y los *promedios móviles* de cinco años. El primer total móvil se obtiene sumando los artículos vendidos en los primeros cinco años (1994 – 1998), es decir, 4+6+7+6+5 = 28 y el primer promedio móvil dividiendo este valor para 5, es decir, $\frac{28}{5} = 5.6.$

El total (28) y el promedio (5.6) se colocan en la fila correspondiente al año promedio de los cinco años utilizados, es decir, 1996. A continuación se determinan el total móvil y el promedio móvil de los siguientes cinco años (1995-1999), 6+7+6+5+5 = 29 y $\frac{29}{5} = 5.8.$ Procediendo de idéntica manera se calculan el resto de los totales y los promedios. Estos resultados se muestran en la tabla 15.10. El gráfico de las ventas y del promedio móvil se muestra en la figura 15.6.

TABLA 15.10 Promedio móvil de cinco años

Años	Ventas	Total móvil de cinco años	Promedio móvil de cinco años
1994	4		
1995	6		
1996	7	28	5.6
1997	6	29	5.8
1998	5	30	6.0
1999	5	31	6.2
2000	7	32	6.4
2001	8	33	6.6
2002	7	34	6.8
2003	6	35	7.0
2004	6	36	7.2
2005	8	37	7.4
2006	9	38	7.6
2007	8	39	7.8
2008	7	40	8.0
2009	7	41	8.2
2010	9	42	8.4
2011	10	43	8.6
2012	9		
2013	8		

FIGURA 15.6 Ventas y promedios móviles



Observe en la figura que la tendencia que describe el comportamiento de los

promedios móviles es una línea recta *perfecta*.

En la práctica, es poco probable que una serie cronológica tenga constante la duración de sus ciclos, y que además, la amplitud de estos ciclos sea la misma. En estos casos la tendencia pudiera no ser *exactamente* una línea recta.

Cuando el promedio móvil se calcula para un número par de años, al no existir un período central, los métodos de cálculo para estos promedios varían un poco con relación al procedimiento que se utiliza cuando el número de años es una cantidad impar.

Procedamos a calcular el promedio móvil de seis años con los datos que se muestran en la tabla 15.11.

TABLA 15.11 Datos de una **s**erie cronológica

2000	2001	2002	2003	2004	2005	2006
15	13	11	16	14	11	16
2007	2008	2009	2010	2011	2012	2013
12	12	10	11	13	14	16

Con los datos de la tabla 15.11 procedemos de la siguiente manera:

- El total de los primeros seis años 15+13+11+16+14+11 = 80 y su promedio $\frac{80}{6} = 13.33$ se colocan entre los años 2002 y 2003 (vendría a ser el punto medio entre los años 2000 y 2005).
- Se obtiene la suma de los siguientes seis años (81) y su promedio (13.50) y se ubican entre los años 2003 y 2004 y así sucesivamente.

- Por último se halla el promedio móvil centrado calculando la media entre 13.33 y 13.50 (13.42), entre 13.50 y 13.33 (13.42), entre 13.33 y 13.50 (13.42), entre 13.50 y 12.50 (13.00) y así sucesivamente.

En la tabla 15.12 se muestran los resultados obtenidos.

TABLA 15.12 Promedio móvil de seis años

Años	Ventas	Total móvil de seis años	Promedio móvil	Promedio móvil centrado
2000	15			
2001	13			
2002	11			

Años	Ventas	Total móvil de seis años	Promedio móvil	Promedio móvil centrado
		80	13.33	
2003	16			13.42
		81	13.50	
2004	14			13.42
		80	13.33	
2005	11			13.42
		81	13.50	
2006	16			13.00
		75	12.50	
2007	12			12.25
		72	12.00	
2008	12			12.17
		74	12.33	
2009	10			12.17
		72	12.00	
2010	11			12.33
		76	12.67	
2011	13			
2012	14			
2013	16			

15.6 Variación irregular.

En párrafos anteriores señalamos que *la variación irregular se debe a factores que se presentan a corto plazo y que tienen las características de ser impredecibles y no recurrentes. Esta variación puede estar provocada por causas no identificables y que son aleatorias y producto de la casualidad.*

Dadas las características de la variación irregular no podremos representarla de forma matemática y nos limitaremos en este libro solo a mencionar su existencia y describir su comportamiento.

15.7 Descripción integral de una serie cronológica.

A continuación desarrollaremos un ejemplo para describir de forma integral el comportamiento de una serie cronológica tomando en cuenta sus cuatro componentes.

Para ello realizaremos las etapas que se señalan a continuación:

Primera: Eliminaremos de la serie temporal los efectos de la estacionalidad haciendo uso de los llamados índices estacionales o temporales, los cuales permitirán describir la variación estacional. Este procedimiento se conoce como *desestacionaliza-*

ción o destemporalización de la serie cronológica. Para obtener los índices estacionales haremos uso del *método de razón de promedio móvil*.

Segunda: Una vez cumplida esta etapa calcularemos una línea de tendencia desestacionalizada la cual nos permitirá hacer predicciones hacia el futuro.

Tercera: Una vez identificadas las componentes estacional y de tendencia pasaremos a calcular la variación cíclica alrededor de la línea de tendencia.

La tabla 15.13 muestra la cantidad de motos acuáticas expresada en cientos de unidades que fueron rentadas por trimestre en un importante balneario entre los años 2009 y 2013.

Desarrollemos las tres fases señaladas anteriormente con el objetivo de poder predecir el número de motos acuáticas que serán rentadas en el segundo trimestre del año 2014.

TABLA 15.13 Número de motos acuáticas rentadas

	TRIMESTRES			
AÑOS	**I**	**II**	**III**	**IV**
2009	8	5	12	9
2010	7	6	10	10
2011	8	7	13	12
2012	8	5	14	11
2013	9	6	13	13

- **PRIMERA FASE**

El primer paso para obtener el índice temporal consiste en calcular el total móvil, el promedio móvil y el promedio móvil centrado de cuatro trimestres, tal y como ya fue explicado en el numeral 15.5. Los resultados de estos cálculos se ilustran en la tabla 15.14.

El segundo paso consiste en calcular los porcentajes del valor real de las motos rentadas para cada trimestre con respecto al correspondiente valor del promedio móvil centrado.

Por ejemplo, para el año 2009 III trimestre, este valor seria $\frac{12}{8.38} x100 = 143.20$

Para el año 2009 IV trimestre $\frac{9}{8.38} x100 = 107.40$

y así sucesivamente para el resto de los datos. Los resultados de estos cálculos se muestran en la última columna de la tabla 15.14.

TABLA 15.14 Porcentaje del valor real con respecto al promedio móvil

Años	Trimestres	Motos rentadas	Total móvil de 4 trimestres	Promedio móvil	Promedio móvil centrado	(Real / P. móvil) x 100
	I	8				
	II	5				
2009			34	8.50		
	III	12			8.38	143.20
			33	8.25		
	IV	9			8.38	107.40
			34	8.50		
	I	7			8.25	84.85
			32	8.00		
	II	6			8.13	73.80
2010			33	8.25		
	III	10			8.38	119.33
			34	8.50		
	IV	10			8.63	115.87
			35	8.75		
	I	8			9.13	87.62
			38	9.50		
	II	7			9.75	71.79
2011			40	10.00		
	III	13			10.00	130.00
			40	10.00		
	IV	12			9.75	123.08
			38	9.50		
	I	8			9.63	83.07
			39	9.75		
	II	5			9.63	51.92
2012			38	9.50		
	III	14			9.63	145.38
			39	9.75		
	IV	11			9.88	111.34
			40	10.00		
	I	9			9.88	91.09
			39	9.75		
	II	6			10.00	60.00
2013			41	10.25		
	III	13				
	IV	13				

Como tercer paso organizamos por trimestre los resultados de los porcentajes relacionados en la última columna de la tabla 15.14 y calculamos la *media modificada* para cada trimestre, la cual se obtiene eliminando los valores mínimos y máximos de cada trimestre y hallando la media de los valores restantes. Los porcentajes organi-

zados por trimestre se muestran en la tabla 15.15

TABLA 15.15 Porcentajes organizados por trimestre

Años	Trimestre I	Trimestre II	Trimestre III	Trimestre IV
2009			143.20	~~107.40~~
2010	84.45	~~73.80~~	~~119.33~~	115.87
2011	87.76	71.79	130.00	~~123.08~~
2012	~~83.07~~	~~51.92~~	~~145.38~~	111.34
2013	~~91.09~~	60.00		
Suma	172.21	131.79	273.20	227.21

Medias modificadas:

Trimestre I: $\frac{172.21}{2} = 86.10$

Trimestre II: $\frac{131.79}{2} = 65.90$

Trimestre III: $\frac{273.20}{2} = 136.60$

Trimestre IV: $\frac{227.21}{2} = 113.60$

Total de índices = 86.10 + 65.90 + 136.60 + 113.60 = 402.2

El cuarto paso de la primera fase consiste en ajustar el valor de las medias modificadas, ya que como se puede apreciar el total de índices es de 402.2 debiendo ser igual a 400 ya que la base de cada índice por trimestre es igual a 100.

El ajuste se logra multiplicando cada índice trimestral por un *factor de ajuste* el cual se obtiene dividiendo la suma correcta de los índices (400) entre la suma real obtenida (402.2). Este factor de ajuste es igual a 0.9945.

Los índices temporales por trimestre son:

1. Trimestre I: 86.10 x 0.9945 = 85.63
2. Trimestre II: 65.90 x 0.9945 = 65.54
3. Trimestre III: 136.60 x 0.9945 = 135.85
4. Trimestre IV: 113.60 x 0.9945 = 112.98

Observe que después del ajuste realizado, la suma total de los índices temporales es exactamente igual a 400 lo cual era el objetivo que estábamos persiguiendo con el ajuste realizado.

El quinto y último paso de la primera fase consiste en utilizar los índices tem-

porales para eliminar los efectos de la estacionalidad de la serie cronológica que estamos estudiando. Para desestacionalizar la serie cronológica dividimos cada valor real de la serie entre su correspondiente índice temporal expresado como una proporción, es decir, dividido para 100. Por ejemplo, en el I trimestre del año 2009 el valor real de las motos rentadas expresado en cientos de unidades fue igual a 8. El índice temporal para el I trimestre expresado como proporción es igual a $\frac{85.63}{100} = 0.8563$. El valor de las motos rentadas desestacionalizada en el I trimestre del año 2009 es igual a $\frac{8}{0.8563} = 9.34$.

Procediendo de esta manera obtenemos los valores de las rentas de motos desestacionalizadas los cuales se muestran en la tabla 15.16.

TABLA 15.16 Valores desestacionalizados de las rentas de motos

Años	Trimestres	Motos rentadas	Proporción del índice temporal	Rentas desestacionalizadas
2009	I	8	0.8563	9.34
	II	5	0.6554	7.63
	III	12	1.3585	8.83
	IV	9	1.1298	7.95
2010	I	7	0.8563	8.17
	II	6	0.6554	9.15
	III	10	1.3585	7.36
	IV	10	1.1298	8.84
2011	I	8	0.8563	9.34
	II	7	0.6554	10.68
	III	13	1.3585	9.57
	IV	12	1.1298	10.61
2012	I	8	0.8563	9.34
	II	5	0.6554	7.63
	III	14	1.3585	10.30
	IV	11	1.1298	9.72
2013	I	9	0.8563	10.51
	II	6	0.6554	9.15
	III	13	1.3585	9.57
	IV	13	1.1298	11.49

Una vez que ha sido concluida esta primera fase, las rentas de motos acuáticas desestacionalizadas solamente reflejan las componentes de tendencia secular, de

fluctuación cíclica y de variación irregular, por tanto, podemos pasar a la segunda fase la cual consiste en establecer una línea de tendencia que nos permita hacer predicciones hacia el futuro.

- **SEGUNDA FASE**

Con los datos de la tabla 15.16 de las rentas desestacionalizadas podemos determinar la ecuación de la línea de tendencia de nuestra serie cronológica, y para ello, en la tabla 15.17 se muestran los cálculos requeridos.

TABLA 15.17 Cálculos requeridos para la línea de tendencia

Año	Trimestres	Rentas desestacionalizadas	Codificación x/2	x	xY	x^2
2009	I	9.34	-9.5	-19	-177.46	361
	II	7.63	-8.5	-17	-129.71	289
	III	8.83	-7.5	-15	-132.46	225
	IV	7.95	-6.5	-13	-103.35	169
2010	I	8.17	-5.5	-11	-89.87	121
	II	9.15	-4.5	-9	-82.35	81
	III	7.36	-3.5	-7	-51.52	49
	IV	8.84	-2.5	-5	-44.20	25
2011	I	9.34	-1.5	-3	-28.02	9
	II	10.68	-0.5	-1	-10.68	1
	Trimestre medio		0			
	III	9.57	0.5	1	9.57	1
	IV	10.61	1.5	3	31.83	9
2012	I	9.34	2.5	5	46.70	25
	II	7.63	3.5	7	53.41	49
	III	10.30	4.5	9	92.70	81
	IV	9.72	5.5	11	106.92	121
2013	I	10.51	6.5	13	136.63	169
	II	9.15	7.5	15	137.25	225
	III	9.57	8.5	17	162.69	289
	IV	11.49	9.5	19	218.31	361
		185.18			146.40	2660

Para realizar la codificación tome en cuenta que la serie cronológica tiene un número par de periodos, es decir, 20 periodos (desde el trimestre 1 hasta el trimestre 20). El trimestre medio (10.5) estaría entre el segundo (10) y tercer (11) trimestre del año 2011. Procediendo a restar el valor 10.5 de los valores comprendidos entre

1 y 20, obtenemos los códigos que aparecen en la tabla 15.17.

Estamos en condiciones de poder calcular los parámetros de la línea de tendencia.

$$b = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{146.40}{2660} = 0.06 \quad a = \overline{Y} = \frac{185.18}{20} = 9.26$$

La línea de tendencia es $\hat{Y} = 9.26 + 0.06x$. En este punto ya hemos identificado la componente estacional y de tendencia, por tanto pasaremos a calcular la variación cíclica alrededor de la línea de tendencia.

- **TERCERA FASE**

Como señalamos al tratar el aspecto relacionado con la componente cíclica de una serie cronológica, la variación cíclica alrededor de la línea de tendencia puede ser identificada calculando el llamado $Porcentaje\ de\ tendencia = \frac{Y}{\hat{Y}} x100$, donde Y representa el valor de la variable dependiente desestacionalizada y $\hat{Y}$ su correspondiente valor estimado mediante la ecuación de tendencia. El porcentaje de tendencia para cada trimestre se muestra en la tabla 15.18.

TABLA 15.18 Porcentaje de tendencia

Años	**Trimestres**	**Rentas (Y) desestacionalizadas**	**x**	$\hat{Y}$	$\frac{Y}{\hat{Y}} x100$
2009	I	9.34	-19	8.12	115.02
	II	7.63	-17	8.24	92.60
	III	8.83	-15	8.36	105.62
	IV	7.95	-13	8.48	93.75
2010	I	8.17	-11	8.6	95.00
	II	9.15	-9	8.72	104.93
	III	7.36	-7	8.84	83.26
	IV	8.84	-5	8.96	98.66
2011	I	9.34	-3	9.08	102.86
	II	10.68	-1	9.2	116.09
	III	9.57	1	9.32	102.68
	IV	10.61	3	9.44	112.39
2012	I	9.34	5	9.56	97.70
	II	7.63	7	9.68	78.82
	III	10.30	9	9.8	105.10
	IV	9.72	11	9.92	97.98

2013	I	10.51	13	10.04	104.68
	II	9.15	15	10.16	90.06
	III	9.57	17	10.28	93.09
	IV	11.49	19	10.4	110.48

En un párrafo anterior señalamos que al concluir la tercera fase pasaríamos a predecir el número de motos acuáticas que serán rentadas en el segundo trimestre del año 2014.

El segundo trimestre del año 2014 se ubica dos trimestres posterior al cuarto trimestre del año 2013, el cual tiene un valor de x igual a 19 según se puede apreciar en la tabla anterior.

Por tanto, el valor de x para el segundo trimestre del año 2014 es igual a 23 = (19 + 2 + 2). Sustituyendo este valor en la ecuación de tendencia tenemos:

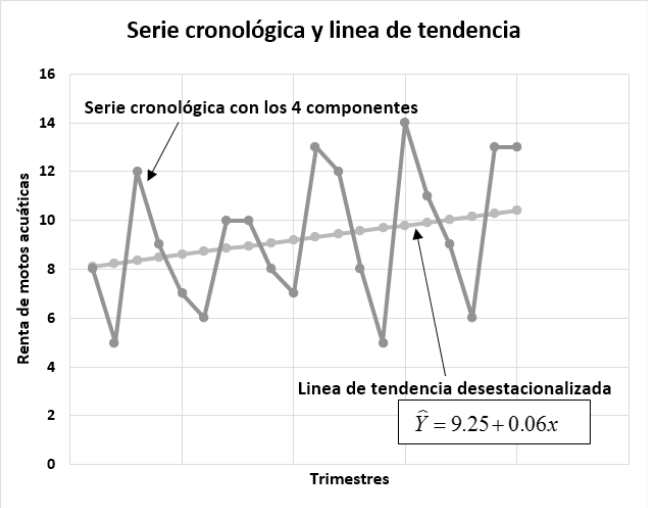
$$\hat{Y} = 9.25 + 0.06(23) = 1063$$

En el segundo trimestre del año 2014, según la línea de tendencia, el valor de $\widehat{Y}$ es igual a 1063, el cual se debe *estacionalizar* multiplicándola por el índice temporal correspondiente al segundo semestre expresado de forma proporcional mostrado en la tabla 15.16, es decir:

(1063) (0.6564) = 697.75 $\approx$ 698 motos acuáticas.

La figura 15.7 muestra el gráfico de la serie cronológica con sus cuatro componentes utilizando los datos tomados de la tabla 15.13 y la correspondiente línea de tendencia desestacionalizada obtenida.

FIGURA 15.7 Serie cronológica y línea de tendencia



15.8 Autocorrelación o correlación residual.

El análisis de series cronológicas y la consecuente predicción hacia el futuro deben ser tratadas por el economista y el administrador con el cuidado que el tema merece.

Para poner un ejemplo, e introducir este aspecto, describamos la línea de tendencia lineal apropiada para la serie cronológica que se muestra en la tabla 15.19 donde Y representa utilidades y X gastos.

TABLA 15.19 Gastos y utilidades medidos durante 18 meses

	Meses								
	1	2	3	4	5	6	7	8	9
X	20.5	20.8	21.2	21.7	22.1	22.3	22.2	22.6	23.1
Y	1.9	1.8	2.1	2.1	1.9	2.2	2.2	2.3	2.7
	Meses								
	10	11	12	13	14	15	16	17	18
X	20.5	20.8	22.2	22.7	23.1	23.3	23.2	23.6	24.1
Y	2.5	2.4	2.2	2.2	2	2.3	2.3	2.4	2.8

Calculemos las sumas de cuadrados y productos necesarios para estimar los parámetros de la ecuación de regresión lineal simple:

$$\sum x_i = 20.5 + 20.8 + 21.2 + \ldots\ldots\ldots + 23.2 + 23.6 + 24.1 = 400$$

$$\sum y_i = 1.9 + 1.8 + 2.1 + \ldots\ldots\ldots + 2.3 + 2.4 + 2.8 = 40.3$$

$$\sum x_i y_i = (20.5)(1.9) + (20.8)(1.8) + \ldots\ldots\ldots + (24.1)(2.8) = 897.94$$

$$\sum x_i^2 = (20.5)^2 + (20.8)^2 + \ldots\ldots + (23.6)^2 + (24.1)^2 = 8909.66$$

y para ser utilizada con posterioridad:

$$\sum y_i^2 = (1.9)^2 + (1.8)^2 + \ldots\ldots + (2.4)^2 + (2.8)^2 = 91.41$$

$$b = \frac{897.94 - \dfrac{(400)(40.3)}{18}}{8909.66 - \dfrac{(400)^2}{18}} = \frac{2.38}{20.77} = 0.1146$$

$$a = \frac{40.3}{18} - (0.1146)\left(\frac{400}{18}\right) = 2.24 - 2.55 = -0.31$$

Finalmente, la ecuación de *regresión lineal simple* obtenida es:

Y = -0.31 + 0.11X

Calculemos las sumas de cuadrados y cuadrados medios requeridas para la

tabla de análisis de regresión:

$$SCC_{TOTAL} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 91.41 - \frac{(40.3)^2}{18} = 1.18$$

$$SCC_{REG.} = \frac{\left(SPC_{XY}\right)^2}{SCC_X} = \frac{(2.38)^2}{20.77} = 0.27 \quad y\ entonces:$$

$$SCC_{ERROR} = 1.18 - 0.27 = 0.91$$

$$CM_{REG.} = \frac{0.27}{1} = 0.27$$

$$CM_{ERROR} = \frac{0.91}{16} = 0.06$$

$$F = \frac{0.27}{0.06} = 4.5$$

Los percentiles de la distribución *F* de Fisher son:

$$F_{5\%}(1,16) = 4.49 \quad F_{1\%}(1,16) = 8.53 \quad F_{0.1\%}(1,16) = 16.12$$

Por ser 4.5 > 4.49, la ecuación se ajusta a los datos para un nivel de significación del 5%.

La tabla del análisis de regresión se muestra en la tabla 15.20:

TABLA 15.20 Análisis de regresión

FUENTES DE VARIACIÓN	G.L.	S.C.	C.M.	F	SIGN.
TOTAL	17	1.18			
REGRESIÓN	1	0.27	0.27	4.5	P<0.05
ERROR	16	0.91	0.06		

A modo de resumen, la ecuación lineal obtenida Y = -0.31 + 0.11X *"puede ser utilizada"* para describir la tendencia secular de la serie cronológica que estamos estudiando.

En este punto resulta necesario hacer una precisión. Una de las hipótesis de base para obtener una ecuación de regresión y la correspondiente prueba de hipótesis de su coeficiente de regresión (análisis de regresión) es que los errores o residuos tienen que ser *independientes*. La característica de una serie cronológica consistente en que sus datos son medidos en periodos de tiempo sucesivos, y que un evento en uno de ellos puede influir en el evento del periodo siguiente, puede traer como consecuencia que los residuos estén *correlacionados*, y por tanto, el método de regresión no tenga *validez estadística*. Cuando esta situación se presenta decimos que existe una *Auto-correlación o Correlación Residual.*

El *estadístico de Durbin-Watson* es una prueba estadística que permite comprobar

si existe autocorrelación en los residuos de un análisis de regresión.

Para desarrollar la prueba de hipótesis de autocorrelación usando este procedimiento se plantean las siguientes hipótesis nula y alternativa:

H_0: No hay correlación residual

H_1: Hay correlación residual

o lo que es equivalente:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

El estadístico de prueba se calcula mediante la expresión:

$$d = \frac{\sum_{t=2}^{T} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{T} e_t}$$

donde T es el número de periodos de la serie cronológica y e_t el residuo de la ecuación de regresión correspondiente al periodo t, que como ya conocemos, se calcula mediante la expresión $e_t = Y_t - \hat{Y}_t$.

Por ejemplo, para t = 1 (Mes 1):

$$e_t = Y_t - \hat{Y}_t = Y_1 - \hat{Y}_1 = 1.9 - \left[-0.31 + 0.11(20.5)\right] = 2.21 - 2.255 = -0.045$$

$e_t - e_{t-1} = e_1 - e_0$ no se puede obtener ya que el residuo e_0 no existe.

Para t = 2 (Mes 2):

$$e_t = Y_t - \hat{Y}_t = Y_2 - \hat{Y}_2 = 1.8 - \left[-0.31 + 0.11(20.8)\right] = 2.11 - 2.288 = -0.178$$

$$e_t - e_{t-1} = e_2 - e_1 = -0.178 - (-0.045) = -0.133$$

Para t = 3 (Mes 3):

$$e_t = Y_t - \hat{Y}_t = Y_3 - \hat{Y}_3 = 2.1 - \left[-0.31 + 0.11(21.2)\right] = 2.41 - 2.288 = 0.078$$

$$e_t - e_{t-1} = e_2 - e_1 = 0.078 - (-0.178) = 0.256$$

Procediendo de forma similar para el resto de los meses obtenemos la tabla 15.21 donde aparecen los cálculos requeridos para determinar el valor del estadístico de Durbin-Watson (d).

TABLA 15.21 Cálculos requeridos para determinar el valor d

MES	X_t	Y_t	$\hat{Y}_t$	$e_t = Y_t - \hat{Y}_t$	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$	e_t^2
1	20,5	1,9	1,945	-0,045			0,0020
2	20,8	1,8	1,978	-0,178	-0,133	0,0177	0,0317
3	21,2	2,1	2,022	0,078	0,256	0,0655	0,0061
4	21,7	2,1	2,077	0,023	-0,055	0,0030	0,0005
5	22,1	1,9	2,121	-0,221	-0,244	0,0595	0,0488
6	22,3	2,2	2,143	0,057	0,278	0,0773	0,0032
7	22,2	2,2	2,132	0,068	0,011	0,0001	0,0046
8	22,6	2,3	2,176	0,124	0,056	0,0031	0,0154
9	23,1	2,7	2,231	0,469	0,345	0,1190	0,2200
10	20,5	2,5	1,945	0,555	0,086	0,0074	0,3080
11	20,8	2,4	1,978	0,422	-0,133	0,0177	0,1781
12	22,2	2,2	2,132	0,068	-0,354	0,1253	0,0046
13	22,7	2,2	2,187	0,013	-0,055	0,0030	0,0002
14	23,1	2	2,231	-0,231	-0,244	0,0595	0,0534
15	23,3	2,3	2,253	0,047	0,278	0,0773	0,0022
16	23,2	2,3	2,242	0,058	0,011	0,0001	0,0034
17	23,6	2,4	2,286	0,114	0,056	0,0031	0,0130
18	24,1	2,8	2,341	0,459	0,345	0,1190	0,2107
						0,7576	1,1059

$$d = \frac{\sum_{t=2}^{18}(e_t - e_{t-1})^2}{\sum_{t=1}^{18} e_t} = \frac{0.7576}{1.1059} = 0.69$$

Si decidimos utilizar un nivel de significación del 5% para realizar la prueba de hipótesis debemos encontrar entonces el valor crítico de Durbin-Watson, el cual aparece en la **TABLA T.13a** del Anexo A y de la cual a continuación mostramos un segmento:

α	Número de variables independientes									
0.05	k = 1		k = 2		k = 3		k = 4		k = 5	
n	d_I	d_S	d_I	d_S	d_I	d_S	d_I	d_S	d_I	d_S
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99

Como hay una sola variable independiente y el tamaño de la muestra es 18, elegimos los valores reportados en la columna k = 1 y en la fila 18, es decir, d_I =1.16 y d_S = 1.39 (I significa Inferior y S significa Superior).

La regla de decisión queda de la siguiente manera:

Se rechaza H_0 si $d < d_I$

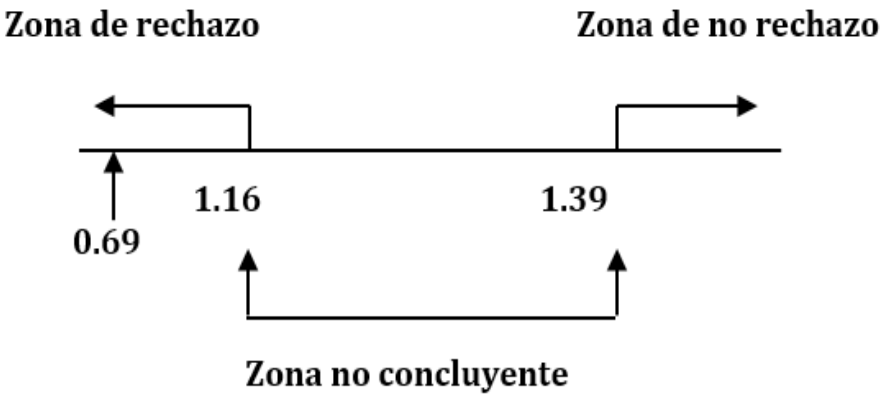
No se rechaza H_0 si $d > d_S$

Si $d_I \leq d \leq d_S$ la prueba no es concluyente y por tanto no se puede tomar una decisión al respecto.

Como d = 0.69 es menor que d_I = 1.16, rechazamos la hipótesis nula y se concluye que para un nivel de significación del 5% existe una correlación residual significativa.

Desde el punto de vista gráfico la situación expuesta anteriormente se muestra en la figura 15.8.

FIGURA 15.8 Zonas de rechazo, aceptación y no concluyente



Los resultados obtenidos muestran que en el procedimiento aplicado para determinar la expresión de la línea de regresión lineal, se violó la hipótesis de base relacionada con la independencia de los residuos, y por tanto, la ecuación y la prueba de hipótesis de su coeficiente de regresión adolecen de falta de *validez estadística.*

Los economistas y administradores que apliquen el procedimiento mecánico del análisis de series cronológicas con el objetivo de proyectar la tendencia y la variación estacional del pasado hacia el futuro, deberán acompañar estos procedimientos con el análisis de otros factores que le permitan alcanzar predicciones más exitosas. Recuerden que la matemática y la estadística no podrán sustituir nunca con toda exactitud a los fenómenos que se presentan a diario en la vida real.

Ejercicios del capítulo

15.1 La serie temporal que se muestra a continuación representa las ventas anuales de gasolina (Y) en la ciudad de Manta medida en millones de litros entre los años 2005 y 2013 (X).

X	2005	2006	2007	2008	2009	2010	2011	2012	2013
Y	350	425	500	525	600	675	750	775	850

Obtenga la tendencia de tipo lineal entre ambas variables mediante el método de codificación de la variable independiente. Pronostique las ventas de gasolina para los años 2014, 2015 y 2016.

15.2 La serie cronológica que se muestra a continuación representa las ventas anuales de gasolina (Y) en la ciudad de Manta medida en millones de litros entre los años 2006 y 2013 (X).

X	2006	2007	2008	2009	2010	2011	2012	2013
Y	425	500	575	650	725	800	875	950

Obtenga la tendencia de tipo lineal entre ambas variables mediante el método de codificación de la variable independiente. Pronostique las ventas de gasolina para los años 2014, 2015 y 2016.

15.3 Haciendo uso de los datos del ejercicio 15.1, obtenga la tendencia cuadrática entre ambas variables mediante el método de codificación de la variable independiente. Pronostique las ventas de gasolina para los años 2014, 2015 y 2016.

15.4 Haciendo uso de los datos del ejercicio 15.2, obtenga la tendencia cuadrática entre ambas variables mediante el método de codificación de la variable independiente. Pronostique las ventas de gasolina para los años 2014, 2015 y 2016.

15.5 Haciendo uso de los datos del ejercicio 15.1, obtenga el porcentaje de tendencia y el residuo cíclico relativo para cada uno de los años involucrados en la serie temporal. ¿Cuál fue el porcentaje real de venta de gasolina en el año 2005? ¿Y en el año 2011?

15.6 Haciendo uso de los datos del ejercicio 15.2, obtenga el porcentaje de tendencia y el residuo cíclico relativo para cada uno de los años involucrados en la serie temporal. ¿Cuál fue el porcentaje real de venta de gasolina en el año 2005? ¿Y en el año 2011?

15.7 A continuación se observa el número de equipos vendidos por una empresa en miles de unidades en los últimos 12 años.

2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
7	9	8	8	10	9	9	11	10	10	12	11

Obtenga los totales móviles y los promedios móviles de 3 años de la anterior serie cronológica.

15.8 Para la serie cronológica cuyos datos se muestran a continuación, obtenga los correspondientes totales y promedios móviles de 4 años.

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
4	6	5	5	5	7	6	6	6	8	7	7
2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
7	9	8	8	8	10	9	9	9	11	10	10

15.9 La siguiente tabla muestra las miles de visitas realizadas por trimestre a un centro de diversiones entre los años 2009 y 2013.

	Trimestres			
Años	**I**	**II**	**III**	**IV**
2009	11	10	14	14
2010	12	11	17	16
2011	12	9	18	15
2012	13	10	17	17
2013	12	9	14	14

a. Elimine de la serie temporal los efectos de la estacionalidad.
b. Calcule la línea de tendencia correspondiente.
c. Calcule la variación cíclica alrededor de la línea de tendencia
d. Haga una predicción del número de visitas que se producirán en el tercer trimestre del año 2014.

15.10 Con los datos de la siguiente serie cronológica:

	Trimestres			
Años	**I**	**II**	**III**	**IV**
2008	21	20	23	23
2009	20	19	23	24
2010	23	21	26	25
2011	22	18	28	25
2012	24	20	27	27
2013	23	18	26	27
2013	12	9	14	14

a. Elimine de la serie temporal los efectos de la estacionalidad.

b. Calcule la línea de tendencia correspondiente.
c. Calcule la variación cíclica alrededor de la línea de tendencia
d. Haga una predicción del número de visitas que se producirán en el primer trimestre del año 2015.

15.11 Los datos contenidos en la tabla que se observa a continuación representan el número promedio de horas diarias empleadas en la producción de un determinado artículo (X) y la cantidad que fueron elaborados medido en millones de unidades (Y).

MES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	9	8	10	8	8	9	11	9	10	11	12	10	8	11	9
Y	2.4	2.6	2.5	2.4	2.7	2.6	2.8	2.5	2.9	2.7	2.7	2.9	2.5	2.9	2.4

Con un nivel de significación del 1%, ¿existe una correlación residual significativa entre ambas variables? Utilice el método de Durbin – Watson.

15.12 Considere que los datos correspondientes al ejercicio anterior son los siguientes:

MES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X	10	9	11	9	9	10	12	10	11	12	13	11	9	12	10	9
Y	2.5	2.7	2.7	2.2	2.6	2.7	2.8	2.7	2.7	2.6	2.8	2.9	2.3	2.8	2.4	2.7

Con un nivel de significación del 5%, ¿existe una correlación residual significativa entre ambas variables? Utilice el método de Durbin – Watson.

Capítulo 16

Números índice

El problema

Las amas de casa del Ecuador se quejan constantemente de la subida de precio que han tenido los comestibles en general y que deben ajustar su presupuesto para enfrentar esta situación. Sin embargo, si se les pregunta en qué basan su afirmación no tienen argumentos de carácter técnico para demostrarlo. ¿Existe algún procedimiento que permita de manera científica demostrar si en realidad se ha presentado una subida de precio de los comestibles con relación a años anteriores?

16.1 Introducción.

De una manera sostenida y progresiva los *números índice* se han convertido en una herramienta básica de economistas y administradores, los cuales los han venido utilizando como *indicadores* de los cambios que se producen de manera constante en la actividad económica y de los negocios.

En muchas situaciones relacionadas con el área de la economía, resulta conveniente combinar varios índices para poder obtener un índice más global que permita estudiar la evolución de una característica difícil de medir.

El presente capítulo estará dedicado al estudio de los números índice, los cuales se constituyen en un sencillo pero importante instrumento de trabajo para los especialistas en economía y administración.

16.2 ¿Qué es un número índice?

Un número índice es una medida de carácter estadístico que *permite estudiar las variaciones de una o más variables en relación al tiempo o cualquier otra característica.*

Los índices más comunes son aquellos que realizan las comparaciones a través del tiempo, por lo que en realidad los números índices son series cronológicas.

Un número índice se calcula dividiendo el *valor actual* de la variable en estudio por un *valor base*, el cual es multiplicado por 100 con la finalidad de expresarlo en términos de porcentaje. A este resultado se le conoce con el nombre de *porcentaje relativo.*

De esta manera, el número índice para el punto base siempre tiene un valor igual a 100.

Veamos un ejemplo. En la tabla 16.1 se puede apreciar el número de personas

de la tercera edad entre los años 2009 y 2013 en una determinada ciudad del país. Si 2009 es el *año base*, entonces los números índice pueden ser calculados como se muestran en dicha tabla.

TABLA 16.1 Personas de la tercera edad entre los años 2009 y 2013

Años	Personas de la tercera edad	Cociente	Índice
2009	600	$\frac{600}{600}=1.00$	100
2010	675	$\frac{675}{600}=1.13$	113
2011	720	$\frac{720}{600}=1.20$	120
2012	845	$\frac{845}{600}=1.41$	141
2013	890	$\frac{890}{600}=1.48$	148

Como se puede apreciar en la tabla 16.1, el número de personas de la tercera edad en el año 2013 fue 148 % del número de personas de la tercera edad en el año base, es decir, en el año 2009.

Por regla general, un índice mide el cambio de una variable con relación al tiempo, sin embargo, en ocasiones es necesario medir este cambio entre diferentes regiones y lugares.

Por ejemplo, la tabla 16.2 muestra el precio promedio de un medicamento en cinco diferentes ciudades de Manabí.

Si Manta es la *ciudad base*, los números índice son calculados como se muestra en dicha tabla.

TABLA 16.2 Números índice del precio de un medicamento por ciudades

Ciudades	Precio promedio del medicamento ($)	Cociente	Índice
Manta	**15.45**	$\frac{15.45}{15.45}=1.00$	100

Ciudades	Precio promedio del medicamento ($)	Cociente	Índice
Portoviejo	16.38	$\frac{16.38}{15.45}=1.06$	106
Chone	15.93	$\frac{15.93}{15.45}=1.03$	103
El Carmen	15.69	$\frac{15.69}{15.45}=1.02$	102
Jipijapa	16.64	$\frac{16.64}{15.45}=1.08$	108

16.3 Tipos de números índice.

Existen tres tipos de índice: el índice de precios, el índice de cantidad y el índice de valores.

1. Índice de precios

De los tres tipos de índices éste es el más utilizado, y su objetivo es comparar niveles de precios de un período de tiempo a otro. Un famoso índice de precios es el conocido índice de precios al consumidor (IPC), el cual evalúa el costo de la vida a través de la medición de los cambios de precios que se producen en una gama de bienes de consumo y de servicios.

El IPC fue clasificado por la Oficina de Estadística Laboral de los Estados Unidos, pero en realidad cada país tiene su propio índice de precios al consumidor.

2. Índices de cantidad

Este índice mide la magnitud con que cambia el número o la cantidad de una variable con relación a periodos de tiempo.

El índice que fue calculado en la tabla 16.1 es un ejemplo al respecto.

3. Índices de valor

Este índice mide los cambios monetarios de una variable. El índice que fue calculado en la tabla 16.2 es un ejemplo al respecto.

16.4 Clasificación de los números índice.

Los números índice se clasifican en *simples* y *compuestos*.

- **Índices simples**

Miden una sola magnitud. Proporcionan la variación que ha sufrido esa magnitud en dos períodos distintos.

Si representamos por I_o^t a un número índice en el periodo t con relación al período base o, entonces:

$I_o^t = \frac{x_t}{x_o} x\ 100$ donde x_t representa el valor en el periodo t y x_o el valor en el periodo base.

- **Índices compuestos**

En ocasiones deseamos medir una magnitud compleja que está conformada por un conjunto de magnitudes simples.

Por ejemplo, el precio de las frutas no puede ser estudiado mediante un índice simple ya que esta magnitud está conformada por el precio de la naranja, la sandía, la manzana, etc.

En un caso como éste se hace necesario utilizar un índice compuesto, el cual se obtiene como una combinación de los índices simples de las magnitudes bajo estudio (naranja, sandía, manzana, etc.).

Los números índice compuestos se clasifican a su vez en *no ponderados* y ponderados.

- El índice compuesto no ponderado es aquel que mide la evolución de una magnitud compleja, pero donde las diferentes magnitudes simples que la conforman tienen la misma importancia.
- El índice compuesto ponderado es aquel que mide la evolución de una magnitud compleja, pero considerando los cambios de algunas variables mucho más importantes que los de otras.

16.5 Índices de agregados no ponderados.

En la tabla 16.3 se reportan los precios de mantenimiento de un vehículo en los años 2009 y 2013.

Con estos valores calculemos un *promedio simple de los índices de precios* para cada una de las acciones de mantenimiento tomando 2009 como año base y 2013 como año dado.

El índice simple fue calculado como ya se explicó en párrafos anteriores.

TABLA 16.3 Índice de precios de acciones de mantenimiento

Acciones de mantenimiento	Precio Año 2009	Precio Año 2013	Índice simple
ABC del motor	21.00	25.00	119.05
Limpieza de inyectores	23.15	26.79	115.72
Calibración de frenos	5.44	8.93	164.15
Cambio de empaque de válvulas	2.00	4.46	223.00
Cambio de aceite (galón)	15.21	17.86	117.42
Total	66.80	83.04	739.34

Con estos resultados podemos calcular el *promedio simple de los índices de precios (P) no ponderado* el cual viene dado por:

$P = \frac{\sum_{1}^{n} I_i}{n} = \frac{739.34}{5} = 147.87$ donde I_i representa el índice simple de cada una de las acciones de mantenimiento y n el número de estas acciones.

Otro tipo de índice factible de ser calculado con los datos de la tabla 16.3 es el índice *de agregados no ponderados* el cual consiste en determinar el índice utilizando para ello las sumas (de aquí la palabra *agregados*) de los precios de los dos periodos estudiados, es decir:

$$P = \frac{83.04}{66.80} x\ 100 = 124.31$$

En determinadas situaciones un índice no ponderado puede presentar algunas limitaciones prácticas. Para describir estas limitaciones, incluyamos en las acciones de mantenimiento de la tabla 16.3 el *cambio de kit de distribución*. La adición planteada se muestra en la tabla 16.4.

TABLA 16.4 Índice de precios de acciones de mantenimiento

Acciones de mantenimiento	Precio Año 2009	Precio Año 2013	Índice simple
ABC del motor	21	25	119.05
Limpieza de inyectores	23.15	26.79	115.72
Calibración de frenos	5.44	8.93	164.15
Cambio de empaque de válvulas	2	4.46	223.00
Cambio de aceite (galón)	15.21	17.86	117.42
Cambio de kit de distribución	41.64	44.64	117.20
Total	108.44	127.68	846.54

Si recalculamos el índice *de agregados no ponderados* con los nuevos datos de la tabla 16.4 obtenemos lo siguiente:

$$P = \frac{127.68}{108.44}\ x\ 100 = 117.74$$

Observe que este último índice es un 6.57% menor que el calculado con los datos de la tabla 16.3, y sin embargo, el aumento de precio del *cambio de kit de distribución* fue de 3 dólares, cifra muy similar de cambio del resto de los componentes de mantenimiento.

La razón que el índice de agregados no ponderados haya disminuido en lugar de mantenerse o aumentar, es que el peso en cuanto al precio del cambio de kit de distribución es mucho mayor que el peso del resto de las componentes, lo cual en estos casos es una limitante cuando se usa un índice no ponderado.

16.6 Índices de agregados ponderados.

Como acabamos de ver, en bastantes ocasiones al calcular un índice se hace necesario asignarle a los cambios que se producen en unas variables una mayor importancia que a los que se producen en otras.

El procedimiento radica en asignar ponderaciones a las diferentes variables en función de su importancia en el cálculo del índice, mejorando de esta forma la precisión de la estimación que pretendemos realizar.

La expresión para calcular un índice de precios de agregados ponderados es:

$$Indice\ de\ precios\ de\ agregados\ ponderados = \frac{\sum_{1}^{n} P_{1i}\, Q_i}{\sum_{1}^{n} P_{0i}\, Q_i}\ x\ 100$$

donde:

P_{1i} = precio del elemento i-ésimo en el año actual

P_{oi} = precio del elemento i-ésimo en el año base

Q_i = factor de ponderación escogido para el elemento i-ésimo

n = número de elementos del compuesto

Prestemos atención a la tabla 16.5 en la cual los elementos del compuesto están tomados de la tabla 16.4 y han sido ponderados en función de la cantidad de acciones de mantenimiento (Q) realizadas por tipo de elemento. Utilizando los resultados de la tabla 16.5, tenemos que el índice de precios de agregados ponderados es:

$$\frac{\sum_{1}^{n} P_{1i}\, Q_i}{\sum_{1}^{n} P_{0i}\, Q_i}\ x\ 100 = \frac{148215.6}{123115.2}\ x 100 = 120.39$$

TABLA 16.5 Cálculo del índice de agregados ponderados

Acciones de mantenimiento	Q	P_o	P_1	$P_o Q$	$P_1 Q$
ABC del motor	1248	21.00	25.00	26208.00	31200.00
Limpieza de inyectores	1872	23.15	26.79	43336.80	50150.88
Calibración de frenos	1560	5.44	8.93	8486.40	13930.80
Cambio de empaque de válvulas	312	2.00	4.46	624.00	1391.52
Cambio de aceite (galón)	2496	15.21	17.86	37964.16	44578.56
Cambio de Kit de distribución	156	41.64	44.64	6495.84	6963.84
Total				123115.2	148215.6

En este punto debemos responder una pregunta crucial, ¿qué valores de Q, o lo que es lo mismo, qué ponderaciones debemos utilizar al momento de calcular un índice de agregados ponderados?

Existen cuatro métodos de ponderar un índice. Estos son:

- **El método de Laspeyres**
- **El método de Paasche**
- **El método ideal de Fisher**
- **El método de agregados de ponderación fija**

Pasemos a estudiar con cierto detalle cada uno de estos métodos. Los métodos de Laspeyres y de Paasche solo difieren en que en el primero se utilizan las *ponderaciones en el periodo base* y en el segundo se utilizan las ponderaciones en el periodo actual. Por tal motivo, utilizaremos el mismo ejemplo para describir ambas formas de calcular el índice de precios ponderado. La tabla 16.6 muestra los precios y la cantidad de pernos de diferentes tamaños y formas vendidos por una ferretería local en el año 2009 y 2013.

Con estos datos calculemos los índices de precios de Laspeyres y de Paasche.

TABLA 16.6 Precios y ventas de pernos

Artículo	Precio base 2009 P_o	Precio actual 2013 P_1	Cantidad vendida 2009 Q_o
Perno Tipo A	0.12	0.14	34000
Perno Tipo B	0.15	0.16	44000
Perno Tipo C	0.18	0.18	38000
Perno Tipo D	0.23	0.26	42000
Perno Tipo E	0.31	0.32	32000

- **Índice de precios de Laspeyres**

La tabla 16.7 contiene los cálculos requeridos para obtener el índice de precios de Laspeyres.

TABLA 16.7 Cálculo del índice de Laspeyres

Artículo	Precio base 2009 P_0	Precio actual 2013 P_1	Cantidad vendida 2009 Q_0	$P_0 Q_0$	$P_1 Q_0$
Perno Tipo A	0.12	0.14	34000	4080	4760
Perno Tipo B	0.15	0.16	44000	6600	7040
Perno Tipo C	0.18	0.18	38000	6840	6840
Perno Tipo D	0.23	0.26	42000	9660	10920
Perno Tipo E	0.31	0.32	32000	9920	10240
				37100	39800

$$Indice\ de\ Laspeyres = \frac{\sum_{1}^{n} P_{1i} Q_i}{\sum_{1}^{n} Poi\, Q_i} x\ 100 = \frac{39800}{37100}\ x\ 100 = 107.28\ donde$$

P_{1i} = precio i-ésimo del periodo actual

P_{ϕ} = precio i-ésimo del periodo base

Q_{ϕ} = cantidad i-ésima vendida en el periodo base

n = número de elementos del compuesto

Partiendo del supuesto de que hemos seleccionado una muestra lo suficientemente representativa del compuesto que estamos estudiando, podemos concluir que los precios se han incrementado en un 7.28% tomando como base que el índice de 2009 es igual a 100.

El método de Laspeyres presenta una desventaja de importancia, la cual radica en el hecho de que no se toman en cuenta los cambios que se puedan producir en los patrones de consumo.

- **Índice de precios de Paasche**

La tabla 16.8 contiene los cálculos requeridos para obtener el índice de precios de Paasche.

TABLA 16.8 Cálculo del índice de Paasche

Artículo	Precio base 2009 P_o	Precio actual 2013 P_1	Cantidad vendida 2013 Q_1	$P_0 Q_1$	$P_1 Q_1$
Perno Tipo A	0.12	0.14	36000	4320	5040
Perno Tipo B	0.15	0.16	51000	7650	8160
Perno Tipo C	0.18	0.18	39000	7020	7020
Perno Tipo D	0.23	0.26	44000	10120	11440
Perno Tipo E	0.31	0.32	26000	8060	8320
				37170	39980

$$Indice\ de\ Paasche = \frac{\sum_{1}^{n} P_{1i}\, Q_{1i}}{\sum_{1}^{n} Poi\, Q_{1i}}\ x\ 100 = \frac{39980}{37170}\ x\ 100 = 107.56$$

donde

P_{1i} = precio i-ésimo del periodo actual

P_{oi} = precio i-ésimo del periodo base

Q_{1i} = cantidad i-ésima vendida en el periodo actual

n = número de elementos del compuesto

En este caso podemos concluir que los precios se han incrementado en 7.56% tomando como base que el índice de 2009 es de 100.

- **Índice de precios ideal de Fisher**

Con el objetivo de hacer un intento para compensar el hecho de que el método de Laspeyres suele ponderar en demasía los elementos cuyos precios aumentaron, y por el contrario, el método de Paasche tiende a ponderar en demasía los elementos cuyos precios disminuyeron, se utiliza el *método ideal de Fisher* el cual se calcula a través de la siguiente expresión:

$$Indice\ ideal\ de\ Fisher = \sqrt{\left(Indice\ de\ Laspeyres\right)\left(Indice\ de\ Paasche\right)}$$

$$Indice\ ideal\ de\ Fisher = \sqrt{\left(107.28\right)\left(107.56\right)} = 107.42$$

A este índice se le denomina *ideal* por el supuesto de que integra las características más notables de los métodos de Laspeyres y Paasche.

- **Índice de precios de agregados de ponderación fija**

Este método es similar a los métodos de Laspeyres y Paasche, pero en lugar de utilizar ponderaciones de periodo base o de periodo actual, utiliza ponderaciones tomadas de un periodo representativo cualquiera.

El índice de precios de agregados de ponderación fija viene dado por la expresión:

$$Indice\ de\ precios\ de\ agregados\ de\ ponderación\ fija = \frac{\sum_{1}^{n} P_{1i} Q_2}{\sum_{1}^{n} P_{oi} Q_2}\ x100$$

El cálculo del índice de precios de agregados de ponderación fija se muestra en la

tabla 16.9, en la cual las ponderaciones han sido tomadas del año 2011 (considerado representativo) y el 2009 tomado como año base.

TABLA 16.9 Cálculo del índice de agregados de ponderación fija

Artículo	**Precio base 2009 P_o**	**Precio actual 2013 P_1**	**Cantidad vendida 2011 Q_2**	**Po Q_2**	**P_1 Q_2**
Perno Tipo A	0.12	0.14	35000	4200	4900
Perno Tipo B	0.15	0.16	50000	7500	8000
Perno Tipo C	0.18	0.18	41000	7380	7380
Perno Tipo D	0.23	0.26	41000	9430	10660
Perno Tipo E	0.31	0.32	29000	8990	9280
				37500	40220

Utilizando los resultados de la tabla tenemos:

$$Indice\ de\ precios\ de\ agregados\ de\ ponderación\ fija = \frac{40220}{37500} x100 = 107.25$$

Llegado a este punto el lector se deberá estar preguntando:

- ¿de qué modo debo ponderar un índice?
- ¿de los cuatro métodos explicados cuál es el más aconsejable?

Responder de forma categórica esta pregunta no es fácil ni tampoco factible.

El autor de este libro piensa que lo más razonable resulta listar las ventajas y desventajas de cada método para que sirva de una guía en este sentido, y que sea el propio lector el que tome la decisión que considere más acertada.

A continuación ofrecemos una relación de las ventajas y desventajas de cada uno de los métodos explicados.

Laspeyres

Ventajas

Solo requiere datos de cantidades del periodo base, de manera que no son necesarios datos anuales de cantidades que hayan sido consumidas.

Los cambios en el precio se reflejan en los cambios en el índice.

Desventajas

No toma en cuenta los cambios que se pueden producir en los patrones de consumo.

Suele ponderar en demasía los artículos que tienden a aumentar de precio.

Paasche

Ventajas

Combina los cambios de patrones de consumo y de precios.

Muestra los actuales patrones de compra al
utilizar cantidades del periodo actual.

Desventajas

Suele ponderar en demasía los artículos que tienden a disminuir de precio.

Demanda que los precios sean recalculados anualmente.

Demanda información de cantidades para el año actual.

Ideal de Fisher

Ventajas

Teóricamente integra las características más notables de los métodos de Laspeyres y Paasche.

Desventajas

Aproximadamente las mismas que el método de Paasche.

Ponderación fija

Ventajas

Para determinar el periodo base y la ponderación fija, el método resulta extraordinariamente flexible.

Desventajas

Al ser similar a los métodos de Laspeyres y Paasche también tiene similares desventajas que ellos.

16.7 Métodos de promedio de relativos.

El método de promedio de relativos es una herramienta que nos permite calcular un índice como alternativa del método de agregados. Estudiaremos las dos formas de construir un índice utilizando este método las cuales son el *método de promedio no ponderado de relativos y el método de promedio ponderado de relativos.*

En realidad, sin hacer referencia a ello, ya habíamos utilizado una variante del método de promedio de relativos cuando calculamos el índice simple de la tabla 16.1 al inicio de este capítulo.

Iniciemos el estudio describiendo el procedimiento para el cálculo de un índice

utilizando el método de promedio *no ponderado* de relativos.

- **Método de promedio no ponderado de relativos**

La expresión para calcular el índice de promedio no ponderado de relativos es la que se muestra a continuación:

Índice de promedio no ponderado de relativos = $\frac{\sum_{1}^{n}\left(\frac{P_{1i}}{P_{oi}}x100\right)}{n}$ donde

P_{1i} = precio del elemento i-ésimo del periodo actual

P_{oi} = precio del elemento i-ésimo del periodo base

n = número de elementos del compuesto

La tabla 16.10 muestra 4 productos con sus respectivos precios en los años 2007 y 2013 así como los cálculos requeridos para determinar el índice de promedio no ponderado de relativos.

TABLA 16.10 Cálculo de un índice de promedio no ponderado de relativos

Producto	**Precio base 2007 P_o**	**Precio actual 2013 P_1**	$\frac{P_1}{P_o}x100$
Leche (litro)	0.65	0.75	115.38
Papa (quintal)	8.00	25.00	312.50
Arroz (kilo)	0.86	0.59	68.60
Tomate árbol (caja)	13.00	12.00	92.31
			588.79

Índice de promedio no ponderado de relativos = $\frac{588.80}{4} = 147.2$

Observe cómo el incremento desmesurado del precio del quintal de papa fue básicamente el causante del índice igual a 147.2, lo cual sugiere ponderar los datos para encontrar una mejor estimación.

- **Método de promedio ponderado de relativos**

En el ejemplo anterior se evidenció la necesidad de ponderar los elementos del compuesto con el objetivo de asignarle, de manera general, *una importancia* diferente a cada uno de ellos.

Existen diferentes formas de determinar las ponderaciones en el método de promedio ponderado de relativos. La expresión general para calcular un índice de

precios aplicando este método es:

$$\frac{\sum_{1}^{n}\left[\left(\frac{P_{1i}}{P_{oi}}x100\right)\left(P_{mi}Q_{mi}\right)\right]}{\sum_{1}^{n}P_{mi}Q_{mi}} \text{ donde}$$

P_{mi} y Q_{mi} = i-ésima cantidad e i-ésimo precio respectivamente que establecen las ponderaciones que serán utilizadas. En particular, m = 0 para el periodo base, m = 1 para el periodo actual y m = 2 para un periodo fijo.

$P_{mi}\ Q_{mi}$ = Valor i-ésimo (multiplicación de precio por cantidad)

P_{oi} = i-ésimo precio del periodo base

P_{1i} = i-ésimo precio del periodo actual

Cuando m = 0, es decir, cuando calculamos un índice de promedio pesado de relativos usando periodos de base, la expresión general se convierte en:

$$\frac{\sum_{1}^{n}\left[\left(\frac{P_{1i}}{P_{oi}}x100\right)\left(P_{oi}Q_{oi}\right)\right]}{\sum_{1}^{n}P_{oi}Q_{oi}}$$

Observe que si en la fórmula anterior, eliminamos del numerador el valor P_{oi}, la expresión se reduce a:

$$\frac{\sum_{1}^{n}\left[\left(P_{1i}\,x100\right)\left(Q_{oi}\right)\right]}{\sum_{1}^{n}P_{oi}Q_{oi}}=\frac{\sum_{1}^{n}P_{1i}Q_{oi}}{\sum_{1}^{n}P_{oi}Q_{oi}}x100$$

la cual coincide con la expresión utilizada por el método de Laspeyres. Es decir, cuando usamos para la ponderación el periodo base, el método de promedio ponderado de relativos coincide exactamente con el método de Laspeyres.

Los datos que se aprecian en la tabla 16.11 corresponden a la compra semanal promedio en comisariato de una familia con una composición estándar en los años 2007 y 2013. Obtengamos el índice de promedio ponderado de relativos usando valores base.

Los cálculos requeridos para ello se muestran en la misma tabla.

TABLA 16.11 Cálculo de un índice de promedio ponderado de relativos

Producto	**Precio base 2007 P_o**	**Precio actual 2013 P_1**	**Cantidad 2007 Q_o**	$\frac{P_1}{P_o}x100$ (1)	P_oQ_o (2)	$(1)\,x\,(2)$
Leche (litros)	0.65	0.75	6	115.38	3.90	449.98
Huevos (docenas)	0.92	1.2	2	130.43	1.84	239.99
Queso (paquetes)	1.12	2.06	1	183.93	1.12	206.00
Mantequilla (lb)	0.75	0.98	1	130.67	0.75	98.00
					7.61	993.97

El índice de promedio ponderado de relativos es igual a:

$$\frac{\sum_{1}^{n}\left[\left(\frac{P_{1i}}{P_{oi}}x100\right)\left(P_{oi}Q_{oi}\right)\right]}{\sum_{1}^{n}P_{oi}Q_{oi}}=\frac{993.97}{7.61}=130.61$$

Solo con el objetivo de comprobar si el valor del índice es igual, haremos el cálculo de éste utilizando el método de Laspeyres. Los cálculos requeridos se presentan en la tabla 16.12.

TABLA 16.12 Cálculo del índice de Laspeyres

Producto	**Precio base 2007 P_o**	**Precio actual 2013 P_1**	**Cantidad 2007 Q_o**	$P_{1i}Q_{1i}$	$P_{1i}Q_{1i}$
Leche (litros)	**0.65**	**0.75**	**6**	3.90	4.50
Huevos (docenas)	0.92	1.2	2	1.84	2.40
Queso (paquetes)	1.12	2.06	1	1.12	2.06
Mantequilla (lb)	0.75	0.98	1	0.75	0.98
				7.61	9.94

Índice de Laspeyres = $\frac{9.94}{7.61}=130.62$ el cual coincide aproximadamente con el calculado por el método de promedio ponderado de relativos.

16.8 Índices de cantidad.

Hasta el momento el estudio de los números índice se ha visto limitado a los índices de precios, sin embargo, también es factible utilizar otro tipo de índice el cual nos permite establecer cambios en las cantidades y en los valores asociadas a estos precios. Nos referimos a los índices de cantidad.

Cuando trabajamos con mercancías que tienen importantes fluctuaciones de precios se hace recomendable el uso de índices de cantidad.

En general, los métodos utilizados para calcular índices de precios pueden ser

utilizados para el cálculo de un índice de cantidad. La expresión general para calcular el índice de cantidad aplicando el método de promedio ponderado de relativos es:

$$\frac{\sum_{1}^{n}\left[\left(\frac{Q_{1i}}{Q_{oi}}x100\right)\left(Q_{mi}P_{mi}\right)\right]}{\sum_{1}^{n}Q_{mi}P_{mi}}$$ donde

Q_{1i} = cantidad i-ésima para el periodo actual

Q_{oi} = cantidad i-ésima para el periodo base

$Q_{mi}\ y\ P_{mi}$ = cantidad i-ésima y precio i-ésimo que determinan los valores que se utilizarán para la ponderación. Específicamente, m = 0 para el periodo base, m = 1 para el periodo actual y m = 2 para un periodo fijo.

A continuación calcularemos un índice de cantidad utilizando el método de promedio ponderado de relativos y 2007 como periodo base. Note que la diferencia con la aplicación del método hecha en párrafos anteriores, es que en este caso el valor se obtiene multiplicando cantidad por precio.

La tabla 16.13 muestra los cálculos necesarios para lograr este objetivo.

TABLA 16.13 Índice de cantidad de promedio ponderado de relativos

Producto	**Cantidad 2007 Q_o**	**Cantidad 2013 Q_1**	**Precio 2007 P_o**	$\frac{Q_1}{Q_o}x100$ (1)	Q_oP_o **(2)**	$(1)x(2)$
Leche (litros)	6	7	0.65	116.67	3.90	455.01
Huevos (docena)	2	2	0.92	100	1.84	184
Queso (paquetes)	1	2	1.12	200	1.12	224
Mantequilla (lb)	1	1	0.75	100	0.75	75
					7.61	938.01

El índice de cantidad es $\frac{938.01}{7.61}=123.26$

16.9 Índices de valores.

Un índice de valores mide los cambios en los precios y en las cantidades asociadas a ellos para un conjunto de elementos de un compuesto, y para su determinación, requiere los precios y las cantidades de los elementos para el año base así como los precios y las cantidades de los elementos para el año actual. La expresión para calcular un índice de valores es:

$$Indice\,de\,valores = \frac{\sum_{1}^{n} P_{1i}Q_{1i}}{\sum_{1}^{n} P_{oi}Q_{oi}} x100$$ donde

P_{1i} = precio i-ésimo para el periodo actual

Q_{1i} = cantidad i-ésima para el periodo actual

P_{oi} = precio i-ésimo para el periodo base

Q_{oi} = cantidad i-ésima para el periodo base

n = número de elementos del compuesto

Los datos de la tabla 16.14 representan el precio en dólares y la cantidad de docenas vendidas de cuatro diferentes artículos de bisutería en el mes de Abril del año 2005 y en ese mismo mes del año 2013. Obtengamos el índice de valores de Abril del 2013 considerando a Abril del 2005 como periodo base.

TABLA 16.14 Precios y cantidad de artículos de bisutería vendidos

	2005		2013			
Artículos	**Precio**	**Cantidad**	**Precio**	**Cantidad**	P_oQ_o	P_1Q_1
A	1.85	22	2.25	33	40.7	74.25
B	1.92	11	2.45	17	21.12	41.65
C	2.59	42	3.31	40	108.78	132.4
D	3.24	30	4.25	36	97.2	153
					267.8	401.3

Las dos últimas columnas de la tabla contienen las sumas requeridas para el cálculo del referido índice de valores.

$$Indice\,de\,valores = \frac{\sum_{1}^{n} P_{1i}Q_{1i}}{\sum_{1}^{n} P_{oi}Q_{oi}} x100 = \frac{401.3}{267.8} = 149.85$$

El resultado anterior significa que el valor (cantidad x precio) de las ventas de bisutería se incrementó en un 49.85% entre Abril del 2005 y Abril del año 2013.

16.10 Índice de Precios al Consumidor.

El Índice de Precios al Consumidor o como también se le conoce por sus siglas, el *IPC*, es sin lugar a dudas uno de los más relevantes índices de precios a nivel mundial.

Este índice le permite conocer al consumidor en qué medida el aumento de los

precios influye en su poder adquisitivo. De igual forma, es un excelente indicador económico de la tasa de inflación en cualquier país del mundo.

En la página web www.ecuadorencifras.gob.ec, se hace referencia al IPC de nuestro país diciendo, y cito:

El índice de Precios al Consumidor (IPC), es un indicador mensual, nacional y para ocho ciudades que mide los cambios en el tiempo del nivel general de los precios, correspondientes al consumo final de bienes y servicios de los hogares de estratos de ingreso: alto, medio y bajo, residentes en el área urbana del país. La variable principal que se investiga es el precio, para los 299 artículos de la canasta fija de investigación. El período base es el año 2004, donde los índices se igualan a 100.

En la figura 16.1 se puede apreciar un gráfico de columnas con los índices de precios al consumidor en el Ecuador según la metodología establecida entre el mes de Agosto del año 2013 y el mes de Julio del 2014.

FIGURA 16.1 Índice de precios al consumidor en Ecuador

Además de las aplicaciones del IPC señaladas anteriormente, existen dos cantidades que pueden ser calculadas haciendo uso del mismo. Estas cantidades son:

- **Ingreso real**
- **Poder de compra del dólar**

Veamos cómo se calculan e interpretan cada una de estas cantidades.

- **Ingreso real**

Supongamos que un ciudadano de la ciudad de Manta acumuló durante el año 2004 (año base) un ingreso de $17000. Si el ingreso actual de este señor es de $24408 en el año, ¿ha mejorado su estándar de vida con relación al periodo base?

El ingreso real se obtiene mediante la expresión:

$$Ingreso\,real = \frac{Ingreso\,monetario}{IPC} x100$$

Si tomamos en cuenta que el IPC actual es de 148.81 (ver figura 16.1) entonces el ingreso real durante el año en curso en el ejemplo que nos ocupa es de:

$$Ingreso\,real = \frac{Ingreso\,monetario}{IPC} x100 = \frac{24408}{148.81} x100 = 16402.12$$

Observe que a pesar de que los ingresos del ciudadano aumentaron de $17000 en el año 2004 a $24408 en el año 2014, su estándar de vida en realidad disminuyó ya que su *ingreso real* en el año 2014 fue de $16402.12.

- **Poder de compra del dólar**

El *poder de compra* del dólar puede ser determinado mediante la expresión:

$$Poder\,de\,compra\,del\,dólar = \frac{\$1}{IPC} x100$$

Según la fórmula anterior el **poder de compra** del dólar en el Ecuador en el momento actual (Julio del 2014) es:

$$Poder\,de\,compra\,del\,dólar = \frac{\$1}{IPC} x100 = \frac{\$1}{148.81} x100 = \$0.67$$

Es decir, un dólar del año 2004 es equivalente en la actualidad a 67 centavos.

Ejercicios del capítulo

16.1 Los valores que se observan en la siguiente tabla corresponden al número de contribuyentes entre los años 2008 y 2013 en una determinada ciudad del país.

Años	2008	2009	2010	2011	2012	2013
Contribuyentes	24380	25118	25782	26321	26788	27451

Determine los números índice correspondiente a cada año tomando como base el año 2013.

16.2 El consumo promedio mensual en kilowatts hora de los habitantes de diferentes ciudades de la provincia de Manabí se aprecia en la siguiente tabla:

Ciudad	Chone	Calceta	Manta	Jipijapa	El Carmen	Portoviejo	Rocafuerte
Consumo	182	165	269	188	200	275	158

Determine los números índice correspondiente a cada ciudad tomando como base a la ciudad de Chone.

16.3 Las ventas tarifa 0% efectuadas por 12 contribuyentes en los años 2012 y 2013 pueden ser apreciadas en la tabla que se muestra a continuación:

	VENTAS TARIFA 0%	
NOMBRES	**2012**	**2013**
JUAN CARLOS GONZÁLEZ	1500.36	1720.48
MARITZA GUTIÉRREZ	1000.78	938.43
VICTOR CASTRO	2500.36	2212.78
TERESA ÁVILA	2700.24	2923.45
JACINTO TERÁN	800.25	714.06
ROSA ALVIA	971.88	1029.89
PEDRO LUIS VITERI	1587.36	1323.65
MAGDALENA ANCHUNDIA	1324.48	1500.25
IGNACIO PÉREZ	548.36	647.12
CARMEN TOALA	368.47	448.93
ARTURO VÉLEZ	568.14	489.71
CRISTINA RODRÍGUEZ	454.89	621.36

a. Calcule el promedio simple de los índices no ponderados.
b. Determine el índice de agregados no ponderados.

16.4 Un grupo de personas estudiaron durante dos años la evolución en el precio de las entradas a cinco lugares de su preferencia. Los resultados obtenidos en el trabajo realizado se observan a continuación:

	Precios	
	Año 2012	**Año 2013**
Teatro	12	15
Cine	3	4
Ballet	25	30
Conciertos	20	25
Fútbol	6	8

a. Calcule el promedio simple de los índices no ponderados.

b. Determine el índice de agregados no ponderados.

16.5 Los datos de la tabla reportada en el ejercicio 16.3 ha sido modificada agregándole a la misma la cantidad de ventas efectuadas por cada uno de los contribuyentes. La nueva tabla quedó de la siguiente forma:

NOMBRES	**Cantidad de ventas**	**Ventas tarifa 0%**	
		2012	**2013**
JUAN CARLOS GONZÁLEZ	50	1500.36	1720.48
MARITZA GUTIÉRREZ	30	1000.78	938.43
VICTOR CASTRO	80	2500.36	2212.78
TERESA ÁVILA	85	2700.24	2923.45
JACINTO TERÁN	25	800.25	714.06
ROSA ALVIA	31	971.88	1029.89
PEDRO LUIS VITERI	55	1587.36	1323.65
MAGDALENA ANCHUNDIA	43	1324.48	1500.25
IGNACIO PÉREZ	15	548.36	647.12
CARMEN TOALA	12	368.47	448.93
ARTURO VÉLEZ	17	568.14	489.71
CRISTINA RODRÍGUEZ	20	454.89	621.36

Calcule el índice de ventas de agregados ponderados utilizando el número de ventas como factor de ponderación y 2012 como año base.

16.6 Los datos de la tabla reportada en el ejercicio 16.4 ha sido modificada agregándole a la misma la cantidad de visitas efectuadas por cada persona a los lugares de su preferencia.

La nueva tabla quedó de la siguiente forma:

	Número de visitas	**Precios**	
		Año 2012	**Año 2013**
Teatro	10	12	15
Cine	15	3	4

	Número de visitas	Precios	
		Año 2012	Año 2013
Ballet	8	25	30
Conciertos	9	20	25
Fútbol	23	6	8

Calcule el índice de precios de agregados ponderados utilizando el número de visitas como factor de ponderación y 2012 como año base.

16.7 Con los datos de la tabla que se muestra a continuación y utilizando 2012 como año base, calcule:

a) El índice de Laspeyres b) El índice de Paasche c) El índice ideal de Fisher

	Cantidad de ventas		Ventas tarifa 0%	
NOMBRES	2012	2013	2012	2013
JUAN CARLOS GONZÁLEZ	48	50	1500.36	1720.48
MARITZA GUTIÉRREZ	27	30	1000.78	938.43
VICTOR CASTRO	75	80	2500.36	2212.78
TERESA ÁVILA	79	85	2700.24	2923.45
JACINTO TERÁN	20	25	800.25	714.06
ROSA ALVIA	26	31	971.88	1029.89
PEDRO LUIS VITERI	51	55	1587.36	1323.65
MAGDALENA ANCHUNDIA	39	43	1324.48	1500.25
IGNACIO PÉREZ	12	15	548.36	647.12
CARMEN TOALA	9	12	368.47	448.93
ARTURO VÉLEZ	12	17	568.14	489.71
CRISTINA RODRÍGUEZ	16	20	454.89	621.36

16.8 Con los datos de la tabla que se muestra a continuación y utilizando 2012 como año base, calcule:

a) El índice de Laspeyres b) El índice de Paasche c) El índice ideal de Fisher

	Número de visitas		Precios	
	2012	2013	Año 2012	Año 2013
Teatro	8	10	12	15
Cine	11	15	3	4
Ballet	5	8	25	30
Conciertos	7	9	20	25
Fútbol	19	23	6	8

16.9 La siguiente tabla muestra el precio de seis diferentes artículos en el año 2008 y 2013 así como la cantidad de ellos que fueron vendidos en el año 2011.

Artículo	Precio		Cantidad vendida 2011
	2008	2013	
A	5.45	7.32	62
B	7.89	9.18	54
C	3.45	5.25	73
D	17.36	20.12	31
E	9.48	11.26	44

Utilizando 2008 como año base y 2011 como ponderación:

a. Calcule el índice de precios de agregados de ponderación fija.
b. Calcule el índice de promedio no ponderado de relativos.

16.10 Con los datos que aparecen en la siguiente tabla y utilizando 2007 como año base y 2010 como ponderación, calcule:

a. El índice de precios de agregados de ponderación fija.
b. El índice de promedio no ponderado de relativos.

Artículo	Precio		Cantidad vendida 2010
	2007	2013	
A	7.38	9,64	81
B	9.63	12.63	79
C	4.66	6.38	96
D	21.18	24.16	52

16.11 La tabla reportada en el ejercicio 16.9 ha sido modificada en su última columna quedando de la siguiente manera:

Artículo	Precio		Cantidad vendida 2008
	2008	2013	
A	5.45	7.32	69
B	7.89	9.18	88
C	3.45	5.25	96
D	17.36	20.12	54
E	9.48	11.26	72

Utilizando 2008 como año base, calcule el índice de promedio ponderado de relativos usando valores base.

16.12 La tabla reportada en el ejercicio 16.10 ha sido modificada en su última columna quedando de la siguiente manera:

Artículo	Precio		Cantidad vendida 2007
	2007	2013	
A	7.38	9.64	94
B	9.63	12.63	91

Artículo	Precio		Cantidad vendida 2007
	2007	2013	
C	4.66	6.38	107
D	21.18	24.16	76

Utilizando 2007 como año base, calcule el índice de promedio ponderado de relativos usando valores base.

16.13 Los precios y las cantidades vendidas de cinco líneas de artículos en los años 2008 y 2013 se muestran a continuación:

Artículo	Precio		Cantidad vendida 2008	Cantidad vendida 2013
	2008	2013		
A	5.45	7.32	77	62
B	7.89	9.18	67	54
C	3.45	5.25	89	73
D	17.36	20.12	46	31
E	9.48	11.26	59	44

Calcule el índice de cantidad y el índice de valores utilizando 2008 como año base.

16.14 Los precios y las cantidades vendidas de cuatro líneas de artículos en los años 2007 y 2013 se muestran a continuación:

Artículo	Precio		Cantidad vendida 2007	Cantidad vendida 2013
	2008	2013		
A	7.38	9.64	77	95
B	9.63	12.63	67	89
C	4.66	6.38	89	105
D	21.18	24.16	46	71

Calcule el índice de cantidad y el índice de valores utilizando 2007 como año base.

Capítulo 17

Introducción a la teoría de decisiones

El problema

Un vendedor minorista de frambuesas, la cual es una fruta con una permanencia en buen estado sumamente corta, quiere conocer cuántas cajas de esta fruta debe tener diariamente en existencia para no sufrir pérdidas ni por obsolescencia, ni por oportunidad, lo cual se produce al tener demasiadas frutas en existencia o por carecer de ellas. ¿Existe algún método estadístico que permita conocer cuántas cajas debe ordenar por día el vendedor para maximizar las ganancias?

17.1 Introducción.

La respuesta al problema planteado en el párrafo anterior la podemos encontrar a través del estudio de la *teoría de decisiones*. Esta área de la estadística también es conocida con el nombre de *teoría de decisiones bayesianas* en honor al ministro presbiteriano y matemático inglés Thomas Bayer (1702–1761), el cual estableció la base matemática para la inferencia probabilística.

La teoría de la decisión es un área en la que participan varias ramas de la ciencia, en particular, la Economía y la Administración. Tiene un carácter normativo y presupone que la persona encargada de tomar la decisión es capaz de hacerlo con precisión y total racionalidad, por supuesto con toda la información disponible.

17.2 Elementos que intervienen en una decisión.

Existen tres elementos que intervienen de forma directa al momento de tomar una decisión cualquiera. Estos elementos son:

1. **Las opciones que están disponibles**
2. **Los estados de la naturaleza**
3. **Los pagos**

Para describir los elementos que intervienen en una decisión, consideremos la situación en la que una compañía de taxis está estudiando la posibilidad de comprar un terreno para construir sus oficinas y un parqueadero con área de servicios para sus autos, en un área cercana al lugar donde se construirá una terminal de ómnibus, cuya ubicación se decidirá el próximo año, pero que con toda seguridad, será en uno de dos sitios diferentes A y B que ya fueron elegidos.

Las opciones que están disponibles

Las *opciones, acciones* o *casos* que están disponibles para la compañía de taxis son:

- Comprar el terreno en el sitio A
- Comprar el terreno en el sitio B
- Comprar el terreno en ambos sitios
- No comprar ningún terreno

Los estados de la naturaleza

- La terminal de ómnibus se construye en el sitio A
- La terminal de ómnibus se construye en el sitio B

Los estados de la naturaleza son eventos que sucederán en el futuro y que no están bajo el control del que toma la decisión. El gerente de la compañía de taxis no conoce a ciencia cierta cuál va a ser la decisión final en cuanto a la elección del sitio donde se construirá la terminal de ómnibus y por tanto está fuera de su control.

Los pagos

Es necesario un pago para poder establecer comparaciones entre las combinaciones que se deriven de las opciones y los estados de la naturaleza, y que nos pueda conducir a una decisión.

Observe los datos que se muestran en la tabla 17.1 los cuales están expresados en miles de dólares. En ella se detallan el precio de compra del terreno en cada ubicación, los beneficios estimados que recibirá la compañía de taxis en cada ubicación, si es que el terminal se construye en ella y el precio de venta del terreno, si el terminal no se construye en ese lugar.

La tabla 17.1 contiene la información necesaria que permite construir la llamada *tabla de pagos*.

TABLA 17.1 Información sobre costos y beneficios

	Posible ubicación	
	Sitio A	Sitio B
Precio de compra del terreno	38	24
Beneficio estimado de la instalación	60	43
Precio de venta del terreno	12	7

Según los datos de la tabla, si la compañía de taxis decide comprar el terreno en el sitio A y el terminal se construye ahí, obtendrá como rendimiento 60 – 38 = 22, que

corresponde a la diferencia entre el beneficio esperado de la instalación y el precio de compra del terreno.

Si por el contrario, el terminal se construye en el sitio B entonces la compañía tendrá que vender el terreno comprado en A con lo que obtiene un rendimiento de 12 al que se le deberá restar la inversión inicial en la compra de dicho terreno, es decir, 12 – 38 = - 26.

Si la compañía de taxis decide comprar el terreno en el sitio B y el terminal se construye ahí, obtendrá como rendimiento 43 – 24 = 19.

Si por el contrario, el terminal se construye en el sitio A entonces la compañía tendrá que vender el terreno comprado en B con lo que obtiene un rendimiento de 7 al que se le deberá restar la inversión inicial en la compra de dicho terreno, es decir, 7 – 24 = - 17.

Si la compañía de taxis decide comprar el terreno en ambos sitios, y el terminal se construye en el sitio A, obtendrá como rendimiento 60 - 38 = 22. Deberá además vender el terreno comprado en el sitio B con lo que el rendimiento se incrementa a 22 + 7 = 29, pero a esta cifra habrá que restarle la inversión inicial al comprar el terreno en el sitio B, es decir, 29 – 24 = 5.

Si por el contrario, el terminal se construye en el sitio B entonces la compañía tendrá un rendimiento 43 - 24 = 19. Deberá además vender el terreno que fue comprado en el sitio A con lo que el rendimiento se incrementa a 19 + 12 = 31, pero a esta cifra habrá que restarle la inversión inicial al comprar el terreno en el sitio A, es decir, 31 – 38 = -7.

Por supuesto que si la compañía decide no comprar ningún terreno, sea la terminal construida en el sitio A o sea construida en el sitio B, el valor de esta alternativa será 0.

La tabla 17.2 muestra la tabla de pagos

TABLA 17.2 Tabla de pagos

	Estados de la naturaleza	
Alternativas	Terminal en A	Terminal en B
Comprar el terreno en el sitio A	22	-26
Comprar el terreno en el sitio B	-17	19
Comprar el terreno en ambos sitios	5	-7
No comprar ningún terreno	0	0

17.3 Decisión en condiciones de incertidumbre. Caso oferta y demanda.

Un cardiólogo que posee un consultorio privado, requiere tener en existencia

un grupo de equipos Holter de tensión arterial que le permita monitorear de forma continua durante 24 horas la presión arterial de los pacientes que sufren de esta dolencia. De momento el cardiólogo no está en condiciones económicas que le permita comprar los equipos que requiere, y en su lugar, renta los equipos a un laboratorio médico a un costo de $50 y le cobra al paciente $70 por utilizarlo, es decir, con ganancia de $20. Si un paciente necesita el Holter y el cardiólogo no lo tiene en existencia deja de ganar $20 por la utilización, y por el contrario, si tiene en el consultorio equipos en existencia que no son requeridos por ningún paciente pierde $50 por cada uno de ellos por concepto de renta.

El problema que se presenta es conocer cuántos equipos debe tener en existencia el consultorio.

A continuación describiremos los pasos a seguir para tomar una decisión al respecto, pero solo tomando en cuenta los elementos básicos que nos permita llegar a la toma de decisión.

- **Primer paso: Cálculo de las ganancias condicionales**

La *tabla de ganancias condicionales* se aprecia en la tabla 17.3.

TABLA 17.3 Tabla de ganancias condicionales

Equipos ofertados	Equipos demandados				
	1	**2**	**3**	**4**	**5**
1	20	20	20	20	20
2	-30	40	40	40	40
3	-80	-10	60	60	60
4	-130	-60	10	80	80
5	-180	-110	-40	30	100

La tabla de ganancias condicionales puede ser considerada como una tabla de pagos en la cual los estados de la naturaleza son los equipos requeridos por los pacientes y las opciones o acciones el número de equipos ofertados o rentados.

Analicemos cómo se calculó esta tabla. Tome en cuenta que el cardiólogo, por supuesto, no puede utilizar más equipos de los que tenga en existencia.

Primera columna en la cual *un solo paciente* requirió el equipo.

- Un equipo rentado: (1 x $70) – (1 x $50) es igual a $70 - $20 = $20 de ganancia.
- Dos equipos rentados: (1 x $70) – (2 x $50) es igual a $70 – $100 = $30 de pérdida.
- Tres equipos rentados: (1 x $70) – (3 x $50) es igual a $70 – $150 = $80 de pérdida.

- Cuatro equipos rentados: (1 x $70) – (4 x $50) es igual a $70 – $200 = $130 de pérdida.
- Cinco equipos rentados: (1 x $70) – (5 x $50) es igual a $70 – $250 = $180 de pérdida.

Segunda columna en la cual *dos pacientes* requirieron el equipo.

- Un equipo rentado: (1 x $70) – (1 x $50) es igual a $70 - $50 = $20 de ganancia.
- Dos equipos rentados: (2 x $70) – (2 x $50) es igual a $140 – $100 = $40 de ganancia.
- Tres equipos rentados: (2 x $70) – (3 x $50) es igual a $140 – $150 = $10 de pérdida.
- Cuatro equipos rentados: (2 x $70) – (4 x $50) es igual a $140 – $200 = $60 de pérdida.
- Cinco equipos rentados: (2 x $70) – (5 x $50) es igual a $140 – $250 = $110 de pérdida.

Con toda seguridad el lector podrá comprobar la validez de los resultados de las tres columnas restantes de la tabla de ganancias condicionales.

Segundo paso: Cálculo de las ganancias esperadas

El cardiólogo, por suerte para él, tiene un registro del número de equipos demandados por sus pacientes en los últimos 90 días, el cual se muestra en la tabla 17.4.

TABLA 17.4 Distribución de frecuencias de equipos rentados

Número de equipos demandados	Número de días	Probabilidad	
		Cálculo	Valor
1	15	15/90	0.17
2	22	22/90	0.24
3	30	30/90	0.33
4	13	13/90	0.14
5	10	10/90	0.11
	90		0.99 ≈ 1.00

La ganancia esperada se pude obtener multiplicando la ganancia condicional por la probabilidad de este tal como se muestra en las tablas siguientes:

TABLA 17.5 Ganancia esperada por tener un Holter en existencia

Equipos demandados	Ganancia condicional	Probabilidad	Ganancia esperada
1	20	0.17	3.40
2	20	0.24	4.80
3	20	0.33	6.60
4	20	0.14	2.80
5	20	0.11	2.20
			19.80

TABLA 17.6 Ganancia esperada por tener dos Holter en existencia

Equipos demandados	Ganancia condicional	Probabilidad	Ganancia esperada
1	-30	0.17	-5.10
2	40	0.24	9.60
3	40	0.33	13.20
4	40	0.14	5.60
5	40	0.11	4.40
			27.70

TABLA 17.7 Ganancia esperada por tener tres Holter en existencia

Equipos demandados	Ganancia condicional	Probabilidad	Ganancia esperada
1	-80	0.17	-13.60
2	-10	0.24	-2.40
3	60	0.33	19.80
4	60	0.14	8.40
5	60	0.11	6.60
			18.80

TABLA 17.8 Ganancia esperada por tener cuatro Holter en existencia

Equipos demandados	Ganancia condicional	Probabilidad	Ganancia esperada
1	-130	0.17	-22.10
2	-60	0.24	-14.40
3	10	0.33	3.30
4	80	0.14	11.20
5	80	0.11	8.80
			-13.20

TABLA 17.9 Ganancia esperada por tener cinco Holter en existencia

Equipos demandados	Ganancia condicional	Probabilidad	Ganancia esperada
1	-180	0.17	-30.60
2	-110	0.24	-26.40
3	-40	0.33	-13.20
4	30	0.14	4.20
5	100	0.11	11.00
			-55.00

Un resumen de las ganancias esperadas se muestra en la tabla 17.10.

TABLA 17.10 Resumen de ganancias esperadas

Holter en existencia	Ganancia esperada
1	19.80
2	27.70
3	18.80
4	-13.20
5	-55.00

En la tabla 17.10 se observa que la mayor ganancia esperada ($27.70) se obtiene con dos equipos Holter en existencia.

Otro método, equivalente al que acabamos de estudiar, es tomar la decisión de acuerdo a la ganancia que se perdería a causa del desconocimiento del estado de la naturaleza.

Esta pérdida se le conoce como *pérdida de oportunidad* o *pérdida relativa*, la cual definiremos posteriormente.

Primer paso: Cálculo de las pérdidas de oportunidad

La tabla de *pérdidas de oportunidad* se muestra en la tabla 17.11.

TABLA 17.11 Tabla de pérdidas de oportunidad

Equipos ofertados	Equipos demandados				
	1	**2**	**3**	**4**	**5**
1	0	20	40	60	80
2	50	0	20	40	60
3	100	50	0	20	40
4	150	100	50	0	20
5	200	150	100	50	0

Analicemos cómo se calculó esta tabla, para lo cual debemos remitirnos a la tabla 17.3 de ganancias condiciones la cual repetimos a continuación para mayor facilidad del lector.

TABLA 17.3 Tabla de ganancias condicionales

Equipos ofertados	Equipos demandados				
	1	**2**	**3**	**4**	**5**
1	20	20	20	20	20
2	-30	40	40	40	40
3	-80	-10	60	60	60
4	-130	-60	10	80	80
5	-180	-110	-40	30	100

Buscamos en la primera columna de la tabla 17.3 de ganancias condicionales la mayor de las ganancias, en este caso 20, y a este valor se le restan todos los valores de esa primera columna, dando esta operación como resultado la primera columna de la tabla 17.11, es decir,

$20 - 20 = 0$

$20 - (-30) = 20 + 30 = 50$

$20 - (-80) = 20 + 80 = 100$

$20 - (-130) = 20 + 130 = 150$

$20 - (-180) = 20 + 180 = 200$

Para la segunda columna de la tabla 17.11 hacemos el mismo procedimiento.

El mayor valor de la segunda columna de la tabla 17.3 es 40, por tanto,

$40 - 20 = 20$

$40 - 40 = 0$

$40 - (-10) = 40 + 10 = 50$ etc.

Con toda seguridad el lector podrá comprobar la validez de los resultados restantes de la tabla 17.11.

- **Segundo paso: Cálculo de las pérdidas de oportunidad esperadas**

Calculemos las ganancias esperadas para cada valor en existencia de los equipos Holter en el consultorio.

TABLA 17.12 Pérdida esperada por tener un Holter en existencia

Equipos demandados	Pérdida de oportunidad	Probabilidad	Pérdida esperada
1	0	0.17	0.00
2	20	0.24	4.80
3	40	0.33	13.20
4	60	0.14	8.40
5	80	0.11	8.80
			35.20

TABLA 17.13 Pérdida esperada por tener dos Holter en existencia

Equipos demandados	Pérdida de oportunidad	Probabilidad	Pérdida esperada
1	50	0.17	8.50
2	0	0.24	0.00
3	20	0.33	6.60
4	40	0.14	5.60
5	60	0.11	6.60
			27.30

TABLA 17.14 Pérdida esperada por tener tres Holter en existencia

Equipos demandados	Pérdida de oportunidad	Probabilidad	Pérdida esperada
1	100	0.17	17.00
2	50	0.24	12.00
3	0	0.33	0.00
4	20	0.14	2.80
5	40	0.11	4.40
			36.20

TABLA 17.15 Pérdida esperada por tener cuatro Holter en existencia

Equipos demandados	Pérdida de oportunidad	Probabilidad	Pérdida esperada
1	150	0.17	25.50
2	100	0.24	24.00
3	50	0.33	16.50
4	0	0.14	0.00
5	20	0.11	2.20
			68.20

TABLA 17.16 Pérdida esperada por tener cinco Holter en existencia

Equipos demandados	Pérdida de oportunidad	Probabilidad	Pérdida esperada
1	200	0.17	34.00
2	150	0.24	36.00
3	100	0.33	33.00
4	50	0.14	7.00
5	0	0.11	0.00
			110.00

Un resumen de las perdidas esperadas se muestra en la tabla 17.17.

TABLA 17.17 Resumen de pérdidas de oportunidad esperadas

Holter en existencia	Pérdida esperada
1	35.2
2	27.3
3	36.2
4	68.2
5	110.0

En la tabla 17.17 se observa que la menor pérdida de oportunidad esperada ($27.3) se obtiene con dos equipos Holter en existencia, lo cual era de esperarse dada la equivalencia de este método con el de las ganancias esperadas.

17.4 Otros criterios de decisión.

En los párrafos siguientes nos dedicaremos a describir otros criterios de decisión en condiciones de incertidumbre para lo cual continuaremos utilizando el ejemplo de la compañía de taxis y su interés de comprar un terreno en un lugar cercano al sitio de construcción de una terminal de ómnibus.

Los criterios a que hemos hecho referencia son:

1. **Criterio de Wald**
2. **Criterio Maximax**
3. **Criterio de Hurwicz**
4. **Criterio de Savage**
5. **Criterio de Laplace**

Antes de iniciar la explicación de cada uno de estos criterios, y precisamente con este objetivo, esquematicemos lo que se entiende por tabla de decisión o tabla de pago de la siguiente manera:

Tabla de decisión o tabla de pago

n = número de alternativas		**Estados de la naturaleza**			
m = número de estados		e_1	e_2	. . .	e_m
Alternativas	a_1	x_{11}	x_{12}	. . .	x_{1m}
	a_2	x_{21}	x_{22}	. . .	x_{2m}
	. . .	. . .	. . .	. . .	. . .
	a_n	x_{n1}	x_{n2}	. . .	x_{nm}

Criterio de Wald

El criterio de Wald se debe al matemático rumano Abraham Wald (1902-1950) quien hizo importantes contribuciones a la teoría de la decisión, la geometría, la eco-

nomía y fue fundador del análisis secuencial.

El criterio de decisión consiste en elegir la alternativa a_i de manera tal que si $s_i = \min_{j=1}^{j=m} x_j$, entonces:

$a_i = \max_{i=1}^{i=n} s_i = \max_{i=1}^{i=n} \min_{j=1}^{j=m} x_j$ donde $s_i = \min_{j=1}^{j=m} x_j$ recibe el nombre de nivel de seguridad. Este criterio recibe también el nombre de *maximin*.

Veamos la aplicación del criterio de Wald en el ejemplo de la compañía de taxis.

En la tabla 17.18 se ha calculado para cada alternativa a_i el nivel de seguridad $s_i = \min_{j=1}^{j=m} x_j$.

TABLA 17.18 Niveles de seguridad

	Estados de la naturaleza		**s_i**
Alternativas	**Terminal en A**	**Terminal en B**	
Comprar el terreno en el sitio A	22	-26	-26
Comprar el terreno en el sitio B	-17	19	-17
Comprar el terreno en ambos sitios	7	-7	-7
No comprar ningún terreno	0	0	0

Según el criterio la mejor alternativa a_i es no comprar ningún terreno, ya que es la opción con el mayor nivel de seguridad.

Criterio Maximax

Este criterio de decisión consiste en elegir la alternativa a_i de manera tal que si $o_i = \max_{j=1}^{j=m} x_j$, entonces:

$a_i = \max_{i=1}^{i=n} o_i = \max_{i=1}^{i=n} \max_{j=1}^{j=m} x_j$

donde $o_i = \max_{j=1}^{j=m} x_j$ recibe el nombre de nivel de optimismo.

Veamos la aplicación del criterio Maximax en el ejemplo de la compañía de taxis.

En la tabla 17.19 se ha calculado para cada alternativa a_i el nivel de optimismo $o_i = \max_{j=1}^{j=m} x_j$.

TABLA 17.19 Niveles de optimismo

	Estados de la naturaleza		o_i
Alternativas	**Terminal en A**	**Terminal en B**	
Comprar el terreno en el sitio A	22	-26	22
Comprar el terreno en el sitio B	-17	19	19
Comprar el terreno en ambos sitios	7	-7	7
No comprar ningún terreno	0	0	0

Según el criterio Maximax la mejor alternativa a_i es comprar el terreno en el sitio A, ya que es la opción con el mayor nivel de optimismo.

Criterio de Hurwicz

Este criterio, el cual se debe al economista y matemático de origen ruso y nacionalidad polaca Leonid Hurwicz (1917 – 2008), es un punto intermedio entre el criterio de Wald y el criterio Maximax.

Este criterio consiste en seleccionar la alternativa a_i de forma tal que $S(a_i) = \max_{i=1}^{i=n} \{\alpha s_i + (1-\alpha) o_i\}$, donde $0 \le \alpha \le 1$

α es un valor elegido por quien toma la decisión, s_i es el i-ésimo nivel de seguridad y o_i el i-ésimo nivel de optimismo.

El valor $S(a_i)$ se conoce con el nombre de *media ponderada de los niveles de seguridad y optimismo* y α como índice de optimismo.

Los valores de α cercanos a 0 evidencian optimismo por parte de quien toma la decisión, sin embargo, los valores cercanos a 1 evidencian pesimismo.

Observe que cuando:

$\alpha = 0$ $S(a_i) = \max_{i=1}^{i=n} \{(0)s_i + (1-0)o_i\} = \max_{1=1}^{i=n} o_i$ Criterio Maximax

$\alpha = 1$ $S(a_i) = \max_{i=1}^{i=n} \{(1)s_i + (1-1)o_i\} = \max_{1=1}^{i=n} s_i$ Criterio de Wald

Un valor de α =0.5 parecería ser una elección adecuada.

Veamos la aplicación del criterio Hurwicz en el ejemplo de la compañía de taxis.

En la tabla 17.20 se ha calculado para cada alternativa a_i el nivel de seguridad $s_i = \min_{j=1}^{j=m} x_j$, el nivel de optimismo $o_i = \max_{j=1}^{j=m} x_j$ y $S(a_i)$.

Los cálculos se han realizado tomando un valor de α = 0.5

TABLA 17.20 Cálculo de la media ponderada

Alternativas	Estados de la naturaleza		s_i	o_i	$S(a_i)$
	Terminal en A	Terminal en B			
Comprar el terreno en el sitio A	22	-26	-26	22	-2
Comprar el terreno en el sitio B	-17	19	-17	19	1
Comprar el terreno en ambos sitios	7	-7	-7	7	0
No comprar ningún terreno	0	0	0	0	0

Para α = 0.5 el valor mayor de las medias ponderadas corresponde a la segunda alternativa, es decir, según el criterio de Hurwicz la mejor opción es comprar el terreno en el sitio B.

Criterio de Savage

Este criterio, el cual es debido al matemático estadounidense Leonard Jimmie Savage (1917 – 1971), se basa en el principio de que el resultado de una alternativa debe ser solamente comparado con los resultados del resto de las alternativas pero solo bajo el mismo estado de la naturaleza.

En este criterio, Savage utiliza un concepto al cual ya hicimos referencia.

Nos referimos al concepto de *pérdida relativa* o *pérdida de oportunidad* $\mathbf{r_{ij}}$ asociada a un resultado $\mathbf{x_{ij}}$, el cual es definido por Savage como la diferencia entre el máximo valor de los $\mathbf{x_{ij}}$ bajo el estado e_j y el valor $\mathbf{x_{ij}}$ de la alternativa $\mathbf{a_i}$ bajo el estado $\mathbf{e_j}$, es decir:

$$r_{ij} = \max_{i=1}^{i=n}\left(x_j\right) - x_j$$

Entonces el criterio de Savage consiste en seleccionar la alternativa con la que se obtenga la menor de las mayores pérdidas relativas, es decir, elegir la alternativa

$$a_k \; tal \; que \; \rho_k = \min_{i=1}^{i=n} \max_{j=1}^{j=m} r_j$$

Utilicemos el criterio en el ejemplo de la compañía de taxis. Como primer paso debemos construir la correspondiente tabla de pérdidas relativas (tabla 17.21) a partir de la tabla de decisión 17.2, la cual transcribimos a continuación para facilidad del lector.

TABLA 17.2 Tabla de decisión

Alternativas	Estados de la naturaleza	
	Terminal en A	Terminal en B
Comprar el terreno en el sitio A	22	-26
Comprar el terreno en el sitio B	-17	19
Comprar el terreno en ambos sitios	7	-7
No comprar ningún terreno	0	0

TABLA 17.21 Pérdidas relativas y el mínimo de éstas para cada alternativa

	Estados de la naturaleza		
Alternativas	**Terminal en A**	**Terminal en B**	ρ_i
Comprar el terreno en el sitio A	0	45	45
Comprar el terreno en el sitio B	39	0	39
Comprar el terreno en ambos sitios	15	26	26
No comprar ningún terreno	22	19	22

Analicemos a continuación cómo se calculó esta última tabla.

El mayor resultado de la columna Terminal A de la tabla de decisión 17.2 es 22, por tanto, las pérdidas relativas correspondientes a esa columna son:

22 - 22 = 0

22 - (-17) = 22 + 17 = 39

22 - 7 = 15

22 - 0 = 22

En la columna Termina B el mayor valor es 19, por tanto, las pérdidas relativas correspondientes a esa columna son:

19 - (-26) = 19 + 26 = 45

19 - 19 = 0

19 - (-7) = 19 + 7 = 26

19 - 0 = 19

El menor valor de ρ_i en la tabla 17.21 se obtiene para la cuarta alternativa, es decir, según el criterio de Savage no se debe comprar ninguno de los terrenos.

Criterio de Laplace

Este criterio, propuesto por el astrónomo, físico y matemático francés Pierre Simón Laplace (1749 – 1827), se basa en el principio de que todos los estados de la naturaleza tienen la misma probabilidad de ocurrencia, y por tanto, un problema de decisión con **m** estados de la naturaleza tendrá una probabilidad de ocurrencia igual a $\frac{1}{m}$.

De lo anterior se desprende que cada alternativa tendrá un valor esperado igual a $\sum_{j=1}^{m} \frac{1}{m} x_{ij}$. El criterio de Laplace selecciona como alternativa óptima aquella que tenga el mayor valor esperado, es decir, se elige a_k de forma tal que:

$\sum_{j=1}^{j=n} \frac{1}{m} x_{ij}$ tenga el valor mayor.

El valor esperado o media aritmética de cada una de las alternativas del ejemplo de la compañía de taxis se muestra en la última columna de la tabla 17.22.

TABLA 17.22 Cálculo del valor esperado para cada alternativa

	Estados de la naturaleza		**Valor esperado**
Alternativas	**Terminal en A**	**Terminal en B**	
Comprar el terreno en el sitio A	22	-26	-2
Comprar el terreno en el sitio B	-17	19	1
Comprar el terreno en ambos sitios	7	-7	0
No comprar ningún terreno	0	0	0

En la tabla anterior se puede apreciar que la alternativa con el mayor valor esperado es la segunda, por tanto, según el criterio de Laplace el terreno debe ser comprado en el sitio B.

A modo de resumen, en la tabla 17.23 se pueden apreciar las diferentes alternativas seleccionadas por los criterios estudiados.

TABLA 17.23 Alternativa seleccionada por cada criterio estudiado

Criterio	**Alternativa elegida**
Wald	No comprar ningún terreno
Maximax	Comprar el terreno en el sitio A
Hurwicz	Comprar el terreno en el sitio B
Savage	No comprar ningún terreno
Laplace	Comprar el terreno en el sitio B

Como se puede observar en la tabla 17.23, no existe mucha coherencia entre los criterios estudiados en cuanto a la alternativa elegida por ellos. Los criterios de Wald y Savage llegan a un mismo resultado, lo mismo que Hurwicz y Laplace, mientras que el criterio Maximax no es coincidente con ninguno de ellos.

El matemático estadounidense John Willard Milnor (1931) en relación con este tema, propuso diez axiomas para caracterizar los criterios de decisión en condiciones de incertidumbre y con un número finito de estados de la naturaleza, concluyendo que no existe ningún criterio que satisfaga los diez axiomas, y por tanto, la solución racional es contradictoria. Sin el ánimo de referirnos al contenido de los axiomas propuestos por Milnor y solo con la finalidad de mostrar la caracterización que hizo de cuatro de los criterios estudiados en párrafos anteriores, mostramos sus resultados en la tabla 17.24.

TABLA 17.24 Caracterización de criterios de decisión según Milnor

	Axiomas	**Criterios**			
		Laplace	**Wald**	**Hurwicz**	**Savage**
1	Orden	S	S	S	S
2	Simetría	S	S	S	S
3	Dominación fuerte	S	S	S	S
4	Continuidad	S	S	S	S
5	Linealidad	S	S	S	S
6	Adición de filas	S	S	S	N
7	Linealidad de columnas	S	N	N	S
8	Duplicidad de columnas	N	S	S	S
9	Convexidad	S	S	S	S
10	Adición de filas especiales	S	S	S	S

S indica que sí cumple con el axioma

N indica que no cumple con el axioma

Analicemos los axiomas que no cumple cada uno de los criterios.

- El criterio de Savage no cumple el axioma relacionado con la Adición de filas. Este axioma significa que en una tabla de decisión el orden entre dos alternativas no cambia a causa de la adición de una nueva alternativa.

Consideremos la tabla de decisión siguiente:

TABLA 17.25 Tabla de decisión

	Estados de la naturaleza	
Alternativas	e_1	e_2
a_1	21	7
a_2	11	15

Las pérdidas relativas se observan en la tabla 17.26

TABLA 17.26 Pérdidas relativas

	Estados de la naturaleza		ρ_i
Alternativas	e_1	e_2	
a_1	0	8	8
a_2	10	0	10

La alternativa óptima es a_1.

Agreguemos una tercera alternativa a la tabla 17.25.

TABLA 17.27 Tabla de decisión con una tercera alternativa

	Estados de la naturaleza	
Alternativas	e_1	e_2
a_1	21	7
a_2	11	15
a_3	9	21

Las nuevas pérdidas relativas se muestran en la tabla 17.28

TABLA 17.28 Pérdidas relativas al adicionar una tercera alternativa

	Estados de la naturaleza		ρ_i
Alternativas	e_1	e_2	
a_1	0	14	14
a_2	10	6	10
a_3	12	0	12

La alternativa óptima ahora es a_2 mientras en la situación anterior fue a_1, es decir, *el orden entre las dos alternativas cambió a causa de la adición de una nueva alternativa*, por tanto, efectivamente no cumple el axioma de *adición de filas*.

- El criterio de Wald no cumple el axioma relacionado con la *Linealidad de columnas*. Este axioma significa que en una tabla de decisión *la relación de orden establecida por el criterio no cambia si se le suma una constante cualquiera a todos los valores de un estado de la naturaleza.*

Consideremos la misma tabla de decisión 17.25 la cual transcribimos para facilidad del lector y apliquemos a los valores de la misma el criterio de Wald.

TABLA 17.25 Tabla de decisión

	Estados de la naturaleza	
Alternativas	e_1	e_2
a_1	21	7
a_2	11	15

En la tabla 17.29 se muestran los niveles de seguridad S_i calculados para cada una de las alternativas. Según estos valores el criterio de Wald escoge la alternativa $\mathbf{a_2}$ como la óptima.

TABLA 17.29 Valor de los niveles de seguridad para cada alternativa

	Estados de la naturaleza		
Alternativas	e_1	e_2	s_i
a_1	21	7	7
a_2	11	15	15

Sumemos a los valores del estado de la naturaleza $\mathbf{e_2}$ de la tabla 17.29 la constante 5 y apliquemos el criterio de Wald a la tabla de decisión resultante, la cual se muestra en la tabla 17.30.

TABLA 17.30 Valor de los niveles de seguridad para cada alternativa

	Estados de la naturaleza		
Alternativas	e_1	e_2	s_i
a_1	21	12	12
a_2	11	20	11

La alternativa óptima según el criterio de Wald es ahora $\mathbf{a_1}$, lo cual indica que dicho criterio no cumple el axioma relacionado con la *Linealidad de columnas*, ya que al sumar a $\mathbf{e_2}$ la constante 5 cambió la relación de orden establecida por el criterio.

- El criterio de Hurwicz no cumple el axioma relacionado con la *Linealidad de columnas.*

Consideremos la tabla de decisión que se muestra en la tabla 17.31, en la que se han calculado para cada alternativa los niveles de seguridad s_i, los niveles de optimismo o_i y las medias ponderadas $S(a_i)$ para un valor de $\alpha = 0.5$. Según estos valores el criterio de Hurwicz escoge la alternativa $\mathbf{a_1}$ como la óptima.

TABLA 17.31 Cálculo de la media ponderada

	Estados de la naturaleza					
Alternativas	e_1	e_2	e_3	s_i	o_i	$S(a_i)$
a_1	21	8	14	8	21	15
a_2	11	16	16	11	16	14

Sumemos a los valores del estado de la naturaleza $\mathbf{e_3}$ la constante 5 y apliquemos el criterio de Hurwicz a la tabla de decisión resultante, la cual se muestra en la tabla 17.32.

TABLA 17.32 Cálculo de la media ponderada

	Estados de la naturaleza					
Alternativas	e_1	e_2	e_3	s_i	o_i	$S(a_i)$
a_1	21	8	19	8	21	15
a_2	11	16	21	11	21	16

La alternativa óptima según el criterio de Hurwicz es ahora $\mathbf{a_2}$ lo cual indica que dicho criterio no cumple el axioma relacionado con la *Linealidad de columnas*, ya que al sumar a $\mathbf{e_3}$ la constante 5 cambió la relación de orden establecida por el criterio.

- El criterio de Laplace no cumple el axioma relacionado con la Duplicidad de columnas. Este axioma significa que en una tabla de decisión *el orden establecido por el criterio no cambia si se adiciona un nuevo estado de la naturaleza idéntico a alguno ya existente.*

TABLA 17.33 Criterio de Laplace aplicado a una tabla de decisión

	Estados de la naturaleza		**Valor esperado**
Alternativas	e_1	e_2	
a_1	21	9	15
a_2	12	16	14

El mayor valor esperado en la tabla 17.33 es 15, por tanto, según el criterio de Laplace la alternativa óptima es $\mathbf{a_1}$.

Adicionemos a esta tabla un nuevo estado de la naturaleza idéntico a e_2. El resultado de la adición y la aplicación del criterio de Laplace se muestran en la tabla 17.34.

TABLA 17.34 Método de Laplace con la adición de un nuevo estado

	Estados de la naturaleza			**Valor esperado**
Alternativas	e_1	e_2	e_3	
a_1	21	9	9	13,00
a_2	12	16	16	14,67

La alternativa óptima es ahora $\mathbf{a_2}$, por tanto, el criterio de Laplace no cumple el axioma relacionado con la *Duplicidad de columnas*, ya que *el orden establecido por el criterio cambió al adicionar un nuevo estado de la naturaleza idéntico a uno ya existente*.

17.5 Clasificación de los procesos de decisión.

Los procesos de decisión se clasifican en dependencia del nivel de conocimiento que tenga la persona encargada de decidir, con relación a los estados de la naturaleza que pueden de una manera u otra incidir en el resultado final. A esto se le llama *ambiente* o *contexto*.

De esta forma, en dependencia del contexto, los procesos de decisión se clasi-

fican:

- *En ambiente de certidumbre* cuando el encargado de la decisión conoce con exactitud los factores y variables que intervienen en el proceso.
- *En ambiente de riesgo* cuando a las consecuencias de una decisión se le puede hacer corresponder una distribución de probabilidad.
- *En ambiente de incertidumbre* cuando a las consecuencias de una toma de decisión no se le puede hacer corresponder una distribución de probabilidad.

Es decir, que en dependencia del entorno o contexto el proceso de decisión se hace bajo certidumbre, bajo riesgo o bajo incertidumbre.

Por ser el capítulo una introducción a la teoría de decisiones solamente le hemos prestado atención a los procesos de decisión bajo incertidumbre, con lo que esperamos haber preparado al lector para estudios más avanzados en este interesantísimo tema.

Ejercicios del capítulo

17.1 Un ingeniero civil, el cual es propietario de una constructora, necesita tener a su disposición un determinado número de concreteras que le permita desarrollar su trabajo de forma ininterrumpida cuando un cliente le solicita sus servicios. Por el momento el ingeniero no tiene los recursos económicos que le permitan efectuar la compra de los equipos, y en su lugar, alquila los mismos a un costo de $8 la hora y le cobra al cliente $12 por hora de utilización del equipo, es decir, con una ganancia de $4. Si un cliente requiere el uso de la concretera y el ingeniero no la tiene en existencia, deja de ganar $4 por la utilización, mientras que si tiene equipos que no son utilizados pierde $8 por cada uno de ellos por concepto de alquiler.

Adicionalmente el ingeniero dispone de un registro en el cual durante 140 días anotó la siguiente información:

Concreteras requeridas	Número de días
1	23
2	31
3	42
4	25
5	19

¿Cuántas concreteras debe tener el ingeniero en existencia para alcanzar la mayor ganancia esperada?

17.2 Considere en el ejercicio 17.1 que las concreteras se rentan a un costo de $14 la hora y se le cobra al cliente $20 por hora de utilización. Considere además que el registro del ingeniero fue el siguiente:

Concreteras requeridas	Número de días
1	32
2	45
3	63
4	33
5	27

¿Cuántas concreteras debe tener el ingeniero en existencia para alcanzar la mayor ganancia esperada?

17.3 Desarrolle el ejercicio 17.1 utilizando el método de pérdida de oportunidad o pérdida relativa.

17.4 Desarrolle el ejercicio 17.2 utilizando el método de pérdida de oportunidad

o pérdida relativa.

17.5 La siguiente tabla muestra los beneficios de cinco posibles decisiones.

	Estados de la naturaleza			
	Estado 1	**Estado 2**	**Estado 3**	**Estado 4**
Alternativa 1	100	90	-20	-45
Alternativa 2	85	80	10	-20
Alternativa 3	0	70	90	60
Alternativa 4	-30	0	40	65
Alternativa 5	-35	-10	85	120

Determine la alternativa que expresa la mejor decisión usando:

a. El criterio de Wald.
b. El criterio Maximax.
c. El criterio de Hurwicz ($\alpha = 0.5$).
d. El criterio de Savage.
e. El criterio de Laplace

17.6 La tabla que aparece a continuación muestra los beneficios para cada una de las cuatro decisiones planteadas.

	Estados de la naturaleza		
	Estado 1	**Estado 2**	**Estado 3**
Alternativa 1	27	-31	-48
Alternativa 2	-21	23	38
Alternativa 3	-35	0	45
Alternativa 4	-45	-15	53

Determine la alternativa que expresa la mejor decisión usando:

a. El criterio de Wald.
b. El criterio Maximax.
c. El criterio de Hurwicz ($\alpha = 0.5$).
d. El criterio de Savage.
e. El criterio de Laplace.

Scheffé, H. (1959*). The Analysis of Variance*. John Wiley and Sons, Inc. New York. First ed.

Stevens, S.S. (1951). *Handbook of Experimental Psychology*. New York: Wiley p.p. 1436

Andrés Venereo Bravo (Cuba, 1945)

Licenciado en Matemática por la Universidad de la Habana. Especialista en Biometría por el Instituto de Ciencia Animal de la República de Cuba. Especialista en Biometría Agrícola. Doctor en Ciencias Agrícolas en la especialidad de Estadística Matemática. Ha sido profesor invitado en diferentes universidades extranjeras. Coautor de los libros *Elementos de MX-Basic* y *Los pastos en Cuba*. Autor de *Diseño y análisis de experimentos agropecuarios*. Desde el año 1998 vive en Ecuador y en la actualidad se desempeña como Profesor Principal en la Universidad Laica Eloy Alfaro de Manabí.

2016