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

APUNTE ELECTRÓNICO

# Matemáticas IV (Estadística descriptiva e inferencial)

## Licenciatura en Informática

SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

# COLABORADORES

**DIRECTOR DE LA FCA**
Dr. Juan Alberto Adam Siade

**SECRETARIO GENERAL**
Mtro. Tomás Humberto Rubio Pérez

– – – –

**COORDINACIÓN GENERAL**
Mtra. Gabriela Montero Montiel
Jefe de la División SUAyED-FCA-UNAM

**COORDINACIÓN ACADÉMICA**
Mtro. Francisco Hernández Mendoza
FCA-UNAM

– – – –

**COAUTORES**
Mtro. Antonio Camargo Martínez
Mtro. Jorge García Castro
Mtra. Adriana Rodríguez Domínguez
Lic. Manuel García Minjares
Mtra. Rosaura Gloria Serrano Jiménez

**REVISIÓN PEDAGÓGICA**
Lic. Laura Antonia Fernández Lapray
L.P. Cecilia Hernández Reyes

**CORRECCIÓN DE ESTILO**
L.F. Francisco Vladimir Aceves Gaytán
Mtro. José Alfredo Escobar Mellado

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

**DISEÑO EDITORIAL**
Mtra. Marlene Olga Ramírez Chavero

# Matemáticas IV (Estadística Descriptiva e Inferencial)
# Apunte electrónico

Edición: agosto 2017.



Estadística descriptiva. Plan 2005/ Actualización plan 2012: 978-970-32-5318-0
Estadística inferencial: Plan 2012: En trámite
Plan de estudios 2012, actualizado 2016.



Hecho en México

# OBJETIVO GENERAL

El alumno aplicará las herramientas estadísticas que le permitan sintetizar grandes volúmenes de información para presentar informes ejecutivos que describan el comportamiento de datos, derivados del análisis e interpretación y la aplicación de modelos estadísticos.

# TEMARIO OFICIAL
## (64 horas)

| | Horas |
|---|---|
| 1. Estadística descriptiva | 8 |
| 2. Teoría de la probabilidad | 12 |
| 3. Distribuciones de probabilidad | 12 |
| 4. Distribuciones muestrales | 8 |
| 5. Pruebas de hipótesis con la distribución ji cuadrada | 8 |
| 6. Análisis de regresión lineal simple | 8 |
| 7. Análisis de series de tiempo | 8 |
| **Total** | **64** |

# INTRODUCCIÓN

En esta asignatura el estudiante estudiará lo relativo a la estadística descriptiva e inferencial.

En la **unidad 1** se estudiarán las diversas características de un conjunto de datos, desde los diferentes tipos de variables y sus escalas de medición. Se estudiará la metodología para la organización y procesamiento de datos, sus distribuciones de frecuencias absolutas y relativas, así como su presentación gráfica en histogramas, polígonos de frecuencias y ojivas. Por otra parte, se conocerán las más importantes medidas de tendencia central y de dispersión. Por último, se analizarán los teoremas de Tchebysheff y de la regla empírica.

En la **unidad 2** se estudiarán las diversas clases de probabilidad, así como los conceptos de espacio muestral y eventos. También se analizarán las reglas fundamentales de la adición y de la multiplicación. Se elaborarán e interpretarán las tablas de probabilidad conjunta y probabilidad condicional y además se conocerá y aplicará el teorema de Bayes.

La **unidad 3** comprenderá el conocimiento de las características y diferencias de las variables discretas y continuas, así como de la distribución general de una variable discreta. Además, se analizarán las principales particularidades y fórmulas de la distribución binomial, la distribución de Poisson, la distribución hipergeométrica, la distribución multinomial, la distribución normal y la distribución exponencial. Por último, se enunciará la ley de los grandes números y su interpretación.

En la **unidad 4** se estudiarán las distribuciones muestrales y el teorema central del límite, los cuales pueden ayudar a la posterior elaboración de los intervalos de confianza.

En la **unidad 5** analizaremos las pruebas de hipótesis con la distribución ji cuadrada y su aplicación.

En la **unidad 6** se investigará el análisis de regresión lineal simple para averiguar el comportamiento de las variables y sus diferentes relaciones.

En la **unidad 7** analizaremos las series de tiempo para observar su aplicación a diferentes problemas de la vida diaria de las empresas.

La estadística descriptiva e inferencial es un elemento imprescindible en la toma de decisiones, tanto en el nivel de las organizaciones privadas y gubernamentales como en el individual. En particular, los estudiantes de informática encontrarán campo fértil para aplicar métodos estadísticos en las áreas de programación y desarrollo de sistemas, entre muchas otras.

La estadística es una rama de las matemáticas, por lo que su tratamiento es formal. Esto no significa, sin embargo, que en el curso se requiera realizar demostraciones rigurosas. El enfoque que se ha adoptado es más bien pragmático, por cuanto está orientado a la aplicación de conceptos, de modo que el requisito fundamental es contar con conocimientos básicos de álgebra y manejo de hoja de cálculo.

# ESTRUCTURA CONCEPTUAL

# UNIDAD 1

# Estadística descriptiva


SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

# OBJETIVO PARTICULAR

El alumno aprenderá y aplicará el proceso estadístico para transformar datos en información útil para la toma de decisiones.

# TEMARIO DETALLADO
## (8 horas)

**1. Estadística descriptiva**

| |
|---|
| 1.1. Tabulación de datos |
| 1.2. Distribuciones de frecuencia |
| 1.3. Presentación gráfica de datos |
| 1.4. Medidas de tendencia central |
| 1.5. Medidas de dispersión |
| 1.6. Teorema de Tchebysheff y regla empírica |

# INTRODUCCIÓN

Para que la información estadística sea relevante, útil y confiable es necesario prestar atención a todas las etapas del proceso de manejo de los datos. Desde el punto de vista de la Estadística Descriptiva es importante, entonces, atender a los diferentes tipos de escalas con que pueden medirse los atributos o variables que nos interesan de un conjunto de observaciones y la forma de agrupar los datos correctamente para, a partir de aquí, aplicar los métodos estadísticos de representación gráfica, así como determinar las medidas de localización y de dispersión que nos permiten dar pasos firmes al interior de la estructura de los datos. La descripción de la información, desde el punto de vista de la estadística, constituye la parte fundamental del proceso de análisis de un conjunto de datos.

# 1.1. Tabulación de datos

Los métodos estadísticos que se utilizan dependen, fundamentalmente, del tipo de trabajo que se desee hacer. Si lo que se desea es trabajar con los datos de las poblaciones, estaremos hablando de métodos de la estadística descriptiva. Si lo que se desea es aproximar las características de una población con base en una muestra, se utilizarán las técnicas de la estadística inferencial.

| | |
|---|---|
| **Técnicas de resumen** | Nos indican la mejor manera para ordenar y agrupar la información, de forma tal que ésta tenga mayor sentido para el usuario, de una manera que los datos en bruto no lo harían. Las técnicas de agrupación de datos y preparación de tablas se incluyen dentro de las técnicas de resumen. |
| **Técnicas de presentación de datos** | Nos permiten obtener una serie de gráficas que, adecuadamente utilizadas, nos dan una idea visual e intuitiva de la información que manejamos. El alumno recuerda, sin duda, haber visto en algún periódico gráficas de barras o circulares (llamadas de *pie* o “pay”, por su pronunciación en inglés). |
| **Técnicas de obtención de parámetros** | Nos llevan a calcular indicadores numéricos que nos dan una idea de las principales características de la población. El conjunto de las 45 calificaciones que un alumno ha obtenido durante sus estudios profesionales nos pueden dar no mucha idea de su desempeño, pero si obtenemos su promedio (técnicamente llamada media aritmética) y éste es de 9.4, nos inclinaremos a pensar que es un buen estudiante. Los parámetros son números que nos sirven para representar (bosquejar una idea) de las principales características de las poblaciones. |

En cualquier estudio estadístico, los datos pueden modificarse de sujeto en sujeto. Si, por ejemplo, estamos haciendo un estudio sobre las estaturas de los estudiantes de sexto de primaria en una escuela, la estatura de cada uno de los niños y niñas será distinta, esto es, variará. Por ello decimos que la estatura es una **variable o atributo.**

Los especialistas en estadística realizan experimentos o encuestas para manejar una amplia variedad de fenómenos o características llamadas variables aleatorias.

Los **datos variables** pueden registrarse de diversas maneras, de acuerdo con los objetivos de cada estudio en particular. Podemos trabajar con cualidades de las observaciones, como por ejemplo el estado civil de una persona, o con características cuantificables, como por ejemplo la edad.

No todos los atributos se miden igual, lo que da lugar a tener diferentes escalas de medición.

**Escala para datos de tipo nominal**

Son aquellas que **no** tienen un **orden** o **dimensión preferente** o **particular** y contienen observaciones que solamente pueden clasificarse o contarse. En un estudio de preferencias sobre los colores de automóviles que escoge un determinado grupo de consumidores, se podrá decir que algunos prefieren el color rojo, otros el azul, algunos más el verde; pero no se puede decir que el magenta vaya “después” que el morado o que el azul sea “más grande” o más chico que el verde.
Para trabajar adecuadamente con escalas de **tipo nominal,** cada uno de los individuos, objetos o mediciones debe **pertenecer** a una y solamente a **una** de las **categorías** o clasificaciones que se tienen y el conjunto de esas categorías debe ser exhaustivo; es decir, tiene que contener a todos los casos posibles. Además, las categorías a que pertenecen los datos no cuentan con un orden lógico.

## Escala para datos de tipo ordinal

En esta escala, las variables sí tienen un **orden natural (**de allí su nombre) y cada uno de los datos puede localizarse dentro de alguna de las categorías disponibles. El estudiante habrá tenido oportunidad de evaluar a algún maestro, en donde las preguntas incluyen categorías como “siempre, frecuentemente, algunas veces, nunca”. Es fácil percatarse que “siempre” es más frecuente que “algunas veces” y “algunas veces” es más frecuente que “nunca”. Es decir, en las escalas de tipo ordinal se puede **establecer una gradación** u orden natural para las categorías. No se puede, sin embargo, establecer comparaciones cuantitativas entre categorías. No podemos decir, por ejemplo, que “frecuentemente” es el doble que “algunas veces” o que “nunca” es tres puntos más bajo que “frecuentemente”.
Para trabajar adecuadamente con escalas de tipo ordinal debemos recordar que las categorías son mutuamente excluyentes (cada dato puede pertenecer o una y sólo a una de las categorías) y deben ser exhaustivas (es decir, cubrir todos las posibles respuestas).

## Escalas numéricas

Estas escalas, dependiendo del manejo que se le dé a las variables, pueden ser **discretas o continuas.**
**Escalas discretas.** Son aquellas que solo pueden **aceptar determinados valores** dentro de un rango.
El número de hijos que tiene una pareja es, por ejemplo, un **dato discreto.** Una pareja puede tener 1, 2, 3 hijos, etc.; pero no tiene sentido decir que tienen 2.3657 hijos. Una persona puede tomar 1, 2, 3, 4, etc., baños por semana, pero tampoco tiene sentido decir que toma 4.31 baños por semana.
**Escalas continuas.** Son aquellas que pueden aceptar **cualquier valor** dentro de un rango y, frecuentemente, el número de decimales que se toman dependen más de la precisión del instrumento de medición que del valor del dato en sí.
Podemos decir, por ejemplo, que el peso de una persona es de 67 kg; pero si medimos con más precisión, tal vez informemos que el peso es en realidad de 67.453 kg, y si nuestra báscula es muy precisa podemos anotar un mayor número de decimales.

El objetivo del investigador condiciona fuertemente el tipo de escala que se utilizará para registrar los datos. Tomando el dato de la estatura, éste puede tener un valor puramente categórico. En algunos deportes, por ejemplo, el básquetbol, puede ser que en el equipo los candidatos a jugador se admitan a partir de determinada estatura para arriba, en tanto que de esa estatura para abajo no serían admitidos. En este caso, la variable estatura tendría solo dos valores, a saber, “aceptado” y “no aceptado” y sería una **variable nominal**. Esta misma variable, para otro estudio, puede trabajarse con una escala de tipo ordinal: “bajos de estatura”, “de mediana estatura” y “altos”. Si tomamos la misma variable y la registramos por su valor en centímetros, la estaremos trabajando como una **variable numérica.**

Dependiendo de las intenciones del investigador, se le puede registrar como variable discreta o continua (variable discreta si a una persona se le registra, por ejemplo, una estatura de 173 cm., de modo que si mide unos milímetros más o menos se redondeará al centímetro más cercano; el registro llevaría a una variable continua si el investigador anota la estatura reportada por el instrumento de medición hasta el límite de precisión de éste, por ejemplo, 173.345 cm.)

Las escalas de tipo numérico pueden tener una de dos características: las **escalas de intervalo** y las **escalas de razón.**

| | | |
|---|---|---|
| **Escalas de tipo numérico** | ***Escalas de intervalo*** | Son aquellas en las que el **cero es convencional o arbitrario.**<br>Un ejemplo de este tipo de escalas es la de los grados Celsius o centígrados que se usan para medir la temperatura. En ella el cero es el punto de congelación del agua y, sin embargo, existen temperaturas más frías que se miden mediante números negativos. En esta escala se pueden hacer comparaciones por medio de diferencias o de sumas. Podemos decir, por ejemplo, que hoy la temperatura del agua de una alberca está cuatro grados más fría que ayer; pero no se pueden hacer comparaciones por medio de porcentajes ya que no hay lugar a dividir en las escalas de intervalo. Si la temperatura ambiente el día de hoy es de diez grados, y el día de ayer fue de veinte grados, no podemos decir que hoy hace el doble de frío que ayer. Sólo podríamos decir que hoy hace más frío y que la temperatura es 10 grados menor que ayer. |
| | ***Escalas de razón*** | Son aquellas en las que el **cero absoluto sí existe.**<br>Tal es el caso de los grados Kelvin, para medir temperaturas, o algunas otras medidas que utilizamos en nuestra vida cotidiana. Encontramos un ejemplo de esta escala cuando medimos la estatura de las personas, expresada en centímetros, por ejemplo, ya que sí existe el cero absoluto, además de que sí se pueden formar cocientes que nos permiten afirmar que alguien mide el doble. |

La mayor parte de las herramientas que se aprenden en este curso son válidas para escalas numéricas, otras lo son para escalas ordinales y unas pocas (muchas de las que se ven en el tema de estadística no paramétrica) sirven para todo tipo de escalas.

**Uso de computadoras en estadística**

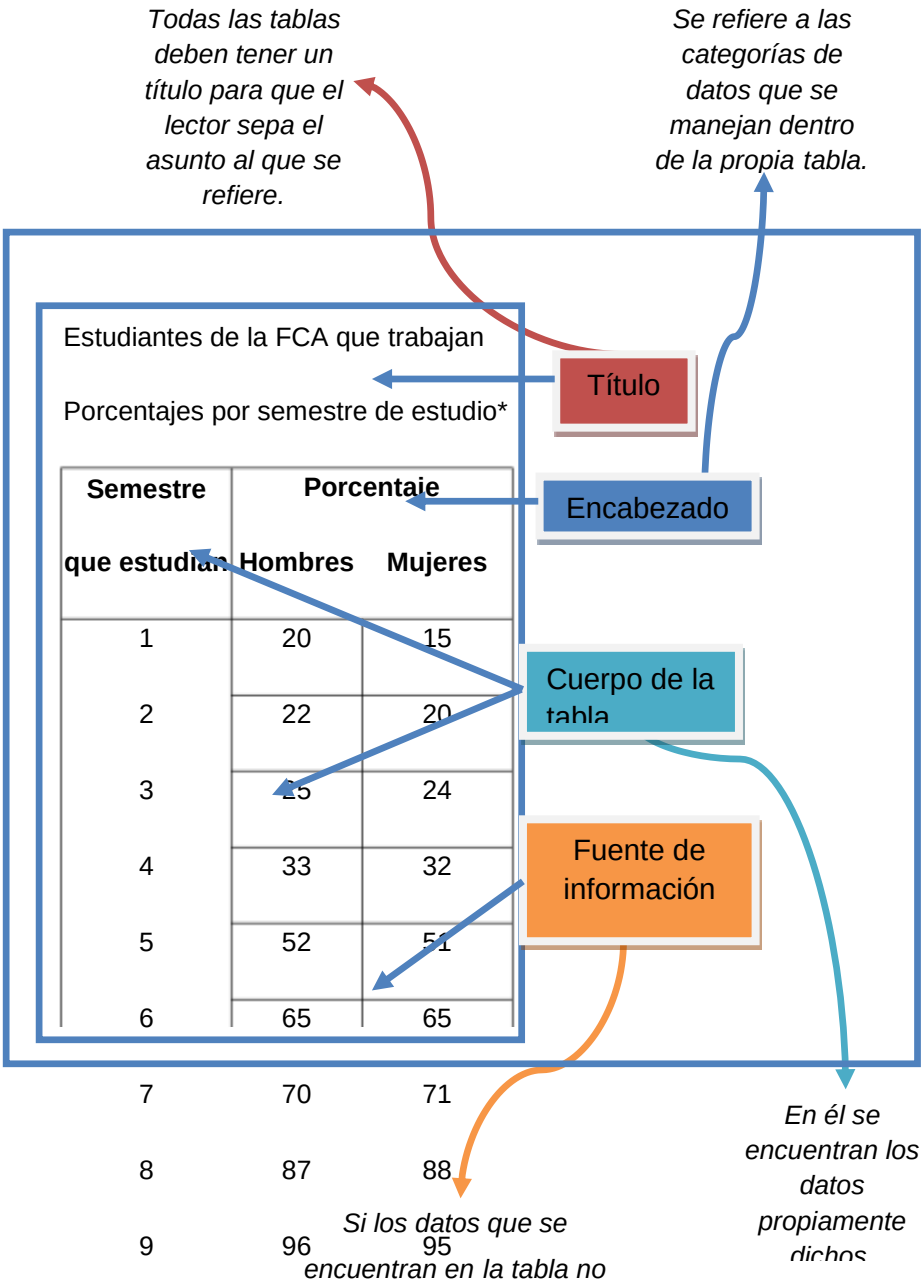
Algunas de las técnicas que se ven en este curso, y muchas que se ven en cursos más avanzados de estadística, requieren un conjunto de operaciones matemáticas que si bien no son difíciles desde el punto de vista conceptual, sí son considerablemente laboriosas por el volumen de cálculos que conllevan. Por ello, **las computadoras**, con su gran capacidad para el manejo de grandes volúmenes de información, son un **gran auxiliar**.

Existen herramientas de uso general como el **Excel** o **Lotus** que incluyen algunas funciones estadísticas y son útiles para muchas aplicaciones. Sin embargo, si se desea estudiar con mayor profundidad el uso de técnicas más avanzadas es importante contar con herramientas específicamente diseñadas para el trabajo estadístico.

Existen diversos paquetes de software en el mercado que están diseñados específicamente para ello. Entre otros se encuentran el SPSS y el SAS. Recomendamos al estudiante que ensaye el manejo de estas herramientas.

**Principales elementos de las tablas**

A continuación se presenta una tabla sencilla, tomada de un ejemplo hipotético. En ella se examinan sus principales elementos y se expresan algunos conceptos generales sobre ellos.




Estudiantes de la FCA que trabajan

Porcentajes por semestre de estudio*

| Semestre que estudian | Porcentaje Hombres | Mujeres |
|---|---|---|
| 1 | 20 | 15 |
| 2 | 22 | 20 |
| 3 | 25 | 24 |
| 4 | 33 | 32 |
| 5 | 52 | 51 |
| 6 | 65 | 65 |
| 7 | 70 | 71 |
| 8 | 87 | 88 |
| 9 | 96 | 95 |

*Fuente: Pérez José, "El trabajo en la escuela", Editorial Académica, México, 19XX

**Editorial Académica, México, 19XX**

### Tabla sencilla de datos

Independientemente de los elementos que pudieren tener las tablas, existen diversas maneras de presentar la información en ellas. No existe una clasificación absoluta de presentación de las diferentes tablas, dado que se pueden inventar varias maneras de presentar la información estadística. Empero, se puede intentar una clasificación que nos permita entender las principales presentaciones.

## Tablas simples

Relaciona una columna de categorías con una o más columnas de datos, sin más elaboración.

| FCA. Maestros de las distintas coordinaciones que han proporcionado su correo electrónico | |
|---|---|
| Coordinaciones | Número de maestros |
| Administración Básica | 23 |
| Administración Avanzada | 18 |
| Matemáticas | 34 |
| Informática | 24 |
| Derecho | 28 |
| Economía | 14 |

## Tablas de frecuencias

Es un arreglo rectangular de información en el que las columnas representan diversos conceptos, dependiendo de las intenciones de quien la elabora, pero que tiene siempre, en una de las columnas, información sobre el número de veces (frecuencia) que se presenta cierto fenómeno.

La siguiente tabla es un ejemplo de esta naturaleza. En ella, la primera columna representa las **categorías** o clases; la segunda, las **frecuencias absolutas** y, la tercera, las **frecuencias relativas**. Esta última columna recibe esa denominación porque los datos están expresados en relación con el total de la segunda columna. Las frecuencias relativas pueden expresarse en porcentaje, tal como en nuestro ejemplo, o en absoluto (es decir, sin multiplicar los valores por 100), por lo que algunos autores llaman a la frecuencia relativa "frecuencia porcentual".

| **Deportes Batista, S.A. de C.V.<br>Número de bicicletas vendidas por tienda** | | |
|---|---|---|
| **Primer trimestre de 20XX** | | |
| Tienda | Unidades | Porcentaje (%) |
| Centro | 55 | 29.1 |
| Polanco | 45 | 23.8 |
| Coapa | 42 | 22.2 |
| Tlalnepantla | 47 | 24.9 |
| **Totales** | **189** | **100.0** |

## Tablas de doble entrada

En algunos casos, se quiere presentar la información con un mayor detalle. Para ello se usan las tablas de doble entrada. Se llaman así porque la información se clasifica simultáneamente por medio de dos criterios en lugar de utilizar solamente uno. Las columnas están relacionadas con un criterio y los renglones con el otro criterio.

| Deportes Batista, S.A. de C.V. | | | | | |
|---|---|---|---|---|---|
| Bicicletas vendidas por modelo y tienda<br>Primer trimestre de 20XX | | | | | |
| | Infantil | Carrera | Montaña | Turismo | Total |
| **Centro** | 13 | 14 | 21 | 7 | **55** |
| **Polanco** | 10 | 14 | 11 | 10 | **45** |
| **Coapa** | 12 | 11 | 17 | 2 | **42** |
| **Tlalnepantla** | 9 | 8 | 13 | 17 | **47** |
| **Totales** | **44** | **47** | **62** | **36** | **189** |

Podemos observar que esta tabla, en la columna de total presenta una información idéntica a la segunda columna de la tabla de frecuencias. Sin embargo, en el cuerpo de la tabla se desglosa una información más detallada, pues nos ofrece datos sobre los modelos de bicicletas, que en la tabla de frecuencias no teníamos.

## Tablas de contingencia

Un problema frecuente es el de definir la independencia de dos métodos para clasificar eventos.

Supongamos que una empresa que envasa leche desea clasificar los defectos encontrados en la producción tanto por tipo de defecto como por el turno (matutino, vespertino o nocturno) en el que se produjo el defecto. Lo que se desea estudiar es si la evidencia de los datos (la contingencia y de allí el nombre) apoya la hipótesis de que exista una relación entre ambas clasificaciones. ¿Cómo se comporta la proporción de cada tipo de defecto de un turno a otro?

En el ejemplo de la empresa que quiere hacer este tipo de trabajo se encontró un total de 312 defectos en cuatro categorías distintas: volumen, empaque, impresión y sellado. La información encontrada se resume en la siguiente tabla.

| Lechería La Laguna, S.A. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Tabla de contingencia en la que se clasifican los defectos del empaque de leche por tipo de defecto y por turno. | | | | | | | | | | |
| **Turno** | **Volumen** | | **Empaque** | | **Impresión** | | **Sellado** | | **Totales** | |
| **Matutino** | 16 | 5.13 | 22 | 7.05 | 46 | 14.74 | 13 | 4.17 | **97** | **31.09** |
| **Vespertino** | 26 | 8.33 | 17 | 5.45 | 34 | 10.90 | 5 | 1.60 | **82** | **26.28** |
| **Nocturno** | 33 | 10.58 | 31 | 9.94 | 49 | 15.71 | 20 | 6.41 | **133** | **42.63** |
| **Totales** | **75** | **24.04** | **70** | **22.44** | **129** | **41.35** | **38** | **12.18** | **312** | **100.00** |
| | *Los números en rojo representan los porcentajes* | | | | | | | | | |

De la información de la tabla antecedente, podemos apreciar que el mayor porcentaje de errores se comete en el turno nocturno y que el área en la que la mayor proporción de defectos se da es la de impresión. Como vemos, la clasificación cruzada de una tabla de contingencia puede llevarnos a obtener conclusiones interesantes que pueden servir para la toma de decisiones.

# 1.2. Distribuciones de frecuencia

Una distribución de frecuencias o tabla de frecuencias no es más que la presentación tabular de las frecuencias o número de veces que ocurre cada característica (subclase) en las que ha sido dividida una variable. Esta característica puede estar determinada por una cualidad o un intervalo; por lo tanto, la construcción de un cuadro de frecuencia o tabla de frecuencias puede desarrollarse tanto para una variable cuantitativa como para una variable cualitativa.

**Distribución de frecuencias para variables cuantitativas**

Las variables cuantitativas o métricas pueden ser de dos tipos.

**Continua**

Cuando la variable es **continua**, la construcción de una **tabla de frecuencia presenta** como su punto de mayor importancia la determinación del **número de intervalos o clases** que la formarán.

Una clase o intervalo de clase es el elemento en la tabla que permite condensar en mayor grado un conjunto de datos con el propósito de hacer un resumen de ellos. El número de casos o mediciones que quedan dentro de un intervalo reciben el nombre de frecuencia del intervalo, que se denota generalmente como fi. La diferencia entre el extremo mayor y el menor del intervalo se llama longitud o ancho del intervalo.

La elaboración de una tabla de distribución de frecuencias se complementa, generalmente, con el cálculo de los siguientes elementos:

| Elemento | Descripción |
|---|---|
| **Marca de clase** | Está constituida por el punto medio del intervalo de clase. Para calcularla es necesario sumar los dos límites del intervalo y dividirlos entre dos |
| **Frecuencia acumulada de la clase** | Se llama así al número resultante de sumar la frecuencia de la clase i con la frecuencia de las clases que la anteceden. Se denota generalmente como fi. La última clase o intervalo en la tabla contiene como frecuencia acumulada el total de los datos. |
| **Frecuencia relativa de la clase** | Es el cociente entre la frecuencia absoluta (fi) de la clase i y el número total de datos. Esta frecuencia muestra la proporción del número de casos que se han presentado en el intervalo "i" respecto al total de casos en la investigación. |
| **Frecuencia acumulada relativa de la clase** | Es el cociente entre la frecuencia acumulada de la clase i y el número total de datos. Esta frecuencia muestra la proporción del número de casos que se han acumulado hasta el intervalo i respecto al total de casos en la investigación |

**Discretas**

En el caso de variables discretas, la construcción de una tabla de distribución de frecuencias sigue los lineamientos establecidos para una variable continua con la salvedad de que en este tipo de tablas no existen intervalos ni marcas de clase, lo cual simplifica la construcción de la tabla.

La construcción de tablas de frecuencia para variables cualitativas o no métricas requiere sólo del conteo del número de elementos o individuos que se encuentran dentro de cierta cualidad o, bien, dentro de determinada característica.

**Cuadros estadísticos**

El resultado del proceso de tabulación o condensación de datos se presenta en lo que en estadística se llaman cuadros estadísticos, también conocidos con el nombre incorrecto de tablas estadísticas, producto de la traducción inglesa.

Con base en el uso que el investigador le dé a un cuadro estadístico, éstos pueden ser clasificados en dos tipos: cuadros de trabajo y cuadros de referencia**.**

**Cuadros de trabajo**

Los **cuadros de trabajo** son aquellos estadísticos que contienen datos producto de una tabulación. En otras palabras, son cuadros depositarios de datos que son utilizados por el investigador para obtener, a partir de ellos, las medidas estadísticas requeridas.

**Cuadros de referencia**

Los **cuadros de referencia** tienen como finalidad ayudar al investigador en el análisis formal de las interrelaciones que tienen las variables que están en estudio, es decir, contienen información ya procesada de cuadros de trabajo (proporciones, porcentajes, tasas, coeficientes, etc.).

La construcción de cuadros estadísticos de trabajo o de cuadros de referencia requiere prácticamente de los mismos elementos en su elaboración, pues ambos presentan las mismas características estructurales, por lo que los elementos que a continuación se describen deberán ser utilizados en la conformación de éstos indistintamente.

1. **Número del cuadro.** Es el primer elemento de todo cuadro estadístico. Tiene como objeto permitir una fácil y rápida referencia al mismo.

**Cuadro 1.1**

2. **Título.** Es el segundo elemento del cuadro estadístico. En él se deberá indicar el contenido del cuadro, su circunscripción espacial, el periodo o espacio temporal y las unidades en las que están expresados los datos.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

3. **Nota en el título (encabezado).** Elemento complementario del título. Se emplea sólo en aquellos cuadros en los que se requiere proporcionar información relativa al cuadro como un todo o a la parte principal del mismo.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

4. **Casillas cabeceras.** Contienen la denominación de cada característica o variable que se clasifica.

   **Cuadro 1.1** Distribución de alumnos por días de ausencia

   Mes base enero

| En algunos casos se especifica el nombre del atributo |
| --- |
| |

5. **Columnas.** Son las subdivisiones verticales de las casillas cabeceras. Se incluyen tantas columnas en una casilla cabecera como categorías le correspondan.

   **Cuadro 1.1** Distribución de alumnos por días de ausencia

   Mes base enero

| | |
| --- | --- |
| | |

6. **Renglones.** Son las divisiones horizontales que corresponden a cada criterio en que es clasificada una variable.

**Cuadro 1.1** Distribución de alumnos por días de ausencia
Mes base enero

| | |
|---|---|
| | |
| | |
| | |
| | |
| | |

7. **Espacio entre renglones.** Tienen por objeto hacer más clara la presentación de los datos, facilitando así su lectura.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

| | |
|---|---|
| | |
| | |
| | |
| | |
| | |
| | |
| | |

8. **Líneas de cabecera.** Son las líneas que se trazan para dividir las casillas de cabecera de los renglones.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

| | |
|---|---|

| | |
|---|---|
| | |
| | |
| | |
| | |

9. **Cabeza del cuadro.** Está formada por el conjunto de casillas, cabecera y encabezados de columnas.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

| **Ausencia (valores de variable)** | **Número de alumnos (Frecuencia)** |
|---|---|
| | |
| | |
| | |
| | |
| | |

**10.** Casillas. Es la intersección que forman cada columna con cada renglón en el cuadro. Las casillas contienen datos o bien los resultados de cálculos efectuados con ellos.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

| **Ausencia (valores de variable)** | **Número de alumnos (Frecuencia)** |
|---|---|
| | *CASILLA* |
| | |
| | |

11. **Cuerpo del cuadro.** Está formado por todos los datos sin considerar la cabeza del cuadro y los renglones de totales.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

| **Ausencia (valores de variable)** | **Número de alumnos (Frecuencia)** |
|---|---|
| 0 | 11<br>4 |
| 1 | 4 |
| 2 | 2 |
| 3 | 2 |
| 4 | 1 |
| 5 | 1 |
| | |

12. **Renglón de totales.** Es un elemento opcional en los cuadros estadísticos.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

| **Ausencia (valores de variable)** | **Número de alumnos (Frecuencia)** |
|---|---|
| 0 | 11<br>4 |
| 1 | 4 |
| 2 | 2 |
| 3 | 2 |
| 4 | 1 |
| 5 | 1 |
| **Total** | 21 |

13. **Línea final de cuadro.** Es la línea que se traza al final del cuerpo del cuadro y en su caso al final del renglón de totales.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

| **Ausencia (valores de variable)** | **Número de alumnos (Frecuencia)** |
|---|---|
| 0 | 11<br>4 |
| 1 | 4 |
| 2 | 2 |
| 3 | 2 |
| 4 | 1 |
| 5 | 1 |
| **Total** | 21 |

14. **Notas al pie del cuadro.** Se usan para calificar o explicar un elemento particular en el cuadro que presente una característica distinta de clasificación.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

| **Ausencia (valores de variable)** | **Número de alumnos (Frecuencia)** |
|---|---|
| 0 | 11<br>4 |
| 1 | 4 |
| 2 | 2 |
| 3 | 2 |
| 4 | 1 |
| 5 | 1 |
| **Total** | 21 |

Nota: No se tiene registrado ningún caso con más de 5 ausencias.

**15.** Fuente. Es el último elemento de un cuadro estadístico. Tiene por objeto indicar el origen de los datos.

**Cuadro 1.1** Distribución de alumnos por días de ausencia

Mes base enero

| **Ausencia (valores de variable)** | **Número de alumnos (Frecuencia)** |
|---|---|
| 0 | 11<br>4 |
| 1 | 4 |
| 2 | 2 |
| 3 | 2 |
| 4 | 1 |
| 5 | 1 |
| **Total** | 21 |

Nota: No se tiene registrado ningún caso con más de 5 ausencias.

**Fuente: Informe mensual de actividades. Mes enero 2007**

La presentación de datos cualitativos suele hacerse de forma análoga a la de las variables, indicando las distintas clases o atributos observados y sus frecuencias de aparición, tal como se recoge en la tabla siguiente sobre color de pelo en un grupo de 100 turistas italianos:

| Color de pelo | Número de personas |
|---|---|
| Negro | 60 |
| Rubio | 25 |
| Castaño | 15 |

**Frecuencias absolutas y relativas**

La **frecuencia absoluta** es el número que indica cuántas veces el valor correspondiente de una variable de medición (dato) se presenta en la muestra y también se le conoce simplemente como frecuencia de ese valor de “X” (dato) en la muestra.

Si ahora dividimos la frecuencia absoluta entre el tamaño de la muestra “n” obtenemos la **frecuencia relativa** correspondiente.

A manera de teorema podemos decir que la frecuencia relativa es por lo menos igual a 0 y, cuando más, igual a 1. Además, la suma de todas las frecuencias relativas en una muestra siempre es igual a 1.

# 1.3. Presentación gráfica de datos

Es importante construir gráficas de diversos tipos que permitan explicar más fácilmente el comportamiento de los datos en estudio. Una **gráfica** permite **mostrar, explicar, interpretar y analizar** de manera sencilla, clara y efectiva los datos estadísticos mediante formas geométricas tales como líneas, áreas, volúmenes, superficies, etc. Las gráficas permiten además la comparación de magnitudes, tendencias y relaciones entre los valores que adquiere una variable.

“Un dibujo vale más que diez mil palabras”, dice el viejo proverbio chino, este principio es tan cierto con respecto a números como a dibujos. Frecuentemente, es posible resumir toda la información importante que se tiene de una gran cantidad de datos en un dibujo sencillo. Así, uno de los métodos más ampliamente utilizados para representar datos es mediante gráficas.

## Histogramas y polígonos de frecuencias

Un **histograma de frecuencias** es un gráfico de rectángulos que tiene su base en el eje de las abscisas (eje horizontal o eje de las equis), con anchura igual cuando se trata de representar el comportamiento de una variable discreta y anchura proporcional a la longitud del intervalo cuando se desea representar una variable continúa. En este último caso, el punto central de la base de los rectángulos equivale al punto medio de cada clase.

Las alturas de los rectángulos ubicadas en el eje de las ordenadas (de las Y o eje vertical) corresponde a las frecuencias de las clases. El área de los rectángulos así formados es proporcional a las frecuencias de las clases.

Los histogramas de frecuencias pueden **construirse** no sólo con las **frecuencias absolutas**, sino también con **las frecuencias acumuladas** y las **frecuencias relativas**. En este último caso, el histograma recibe el nombre de Histograma de frecuencias relativas, Histograma de porcentajes o Histograma de proporciones, según el caso.

El histograma es similar al diagrama de barras o rectángulos, aunque con una diferencia importante: mientras que en los diagramas sólo estamos interesados en las alturas de las barras o rectángulos, en el histograma son fundamentales tanto la altura como la base de los rectángulos, haciendo el área del rectángulo proporcional a su frecuencia.

Como ya se indicó, las variables cualitativas no tienen intervalos de clase por carecer éstos de sentido. Tampoco en ellas se calcula la frecuencia acumulada; por lo tanto, para las variables cualitativas sólo existe la construcción de los histogramas de frecuencia absoluta y los histogramas porcentuales o de frecuencia relativa. Para variables cualitativas no existe polígono de frecuencias.

Pasos a seguir para la elaboración de un diagrama de frecuencias (o polígono de frecuencias) y un histograma.

Considera el siguiente conjunto de datos:

| | | |
|---|---|---|
| | 8.98.39.28.49.1<br>8.68.99.18.88.8<br>8.89.18.98.78.8<br>8.99.08.68.78.4<br>8.69.08.88.99.1<br>9.49.09.29.18.8<br>9.19.39.09.28.8<br>9.78.99.78.39.3<br>8.9 8.8 9.3 8.5 8.9<br>8.39.28.28.98.7<br>8.98.88.58.48.0<br>8.58.78.78.88.8<br>8.38.68.79.08.7<br>8.48.88.48.69.0<br>9.38.88.58.79.6<br>8.59.19.08.89.1<br>8.68.68.49.18.5<br>9.19.28.88.58.3<br>9.38.68.78.79.1<br>8.88.79.09.08.5<br>8.58.88.98.29.0<br>9.08.78.78.99.4<br>8.38.69.28.78.7<br>8.79.78.99.28.8<br>8.38.68.58.69.7 | |
| Máximo | **9.79.79.79.29.2** | **máximo = 9.7** |
| Mínimo | **8.38.38.28.28.0** | **mínimo = 8.0** |

**Paso 1.** Cuenta el número de datos en la población o muestra; en este caso son 125 lecturas, por lo tanto, n=125.

**Paso 2.** Calcula el rango de los datos (R).

Para determinar el rango de los datos lo único que se debe hacer es encontrar el número mayor y el número menor de las 125 lecturas que se tienen en la tabla. Para hacer esto, el doctor Kaouru Ishikawa recomendó lo siguiente:

Se toman filas o columnas, en este caso columnas, y se identifica tanto el valor más grande como el más pequeño por columna. Se anotan los resultados en dos renglones, uno para los valores máximos y otro para los mínimos y de entre estos números se determina nuevamente el mayor y el menor, mismos que serán identificados como el *máximo y mínimo* de las lecturas en la tabla. En este caso: MÁX = 9.7 y MÍN = 8.0. El rango (R) es la diferencia entre éstos valores, por lo que R = MÁX – MÍN = 9.7 – 8.0 = 1.7.

**Paso 3.** Determina el número de clases, celdas o intervalos.

En la construcción de un diagrama de frecuencias o de un histograma es necesario encasillar las lecturas. Si bien existe una expresión matemática para el cálculo del número de clases que debe tener la distribución de frecuencias, hay un camino más práctico, el cual señala que el número de clases no debe ser menor que 6 ni mayor que 15. En este sentido, si “Q” es la cantidad de clases que tendrá el histograma; se recomienda lo siguiente:

| Número de lecturas | Número de clases |
|---|---|
| **< 50** | **6 - 8** |
| **50 - 100** | **9 - 11** |
| **100 - 250** | **8 - 13** |
| **> 250** | **10 - 15** |

**Paso 4.** Determina el ancho “c” del intervalo.

Para este caso utilizamos la siguiente fórmula:

$$C = \frac{R}{Q} = \frac{1.7}{10} = 0.17$$

Generalmente es necesario redondear “c” para trabajar con números más cómodos. En esta ocasión daremos un valor de c=0.20 unidades el cual debe mantenerse constante a lo largo del rango, que en este caso es de R=1.7

**Paso 5.** Establece los límites de clase.

En muchos casos esto sucede automáticamente y depende de la costumbre. Por ejemplo, si se le pregunta su edad a una persona, ésta contestará con el número de años que tiene. En este caso, el ancho de clase es automáticamente de un año, aunque la persona haya cumplido años ayer o hace 11 meses. En otras instancias, la resolución en los instrumentos de medición es la que determina el ancho de clase, aun cuando se siga una regla general para la normalización del histograma. En el ejemplo, la lectura menor fue de 8.0 por lo que se podría tomar éste como el límite inferior de la primera clase, y al sumar al valor de 8.0 el ancho de clase “c” se tendría el límite inferior del segundo intervalo y así sucesivamente hasta que todos los valores de la tabla queden contenidos.

**Paso 6.** Construye la distribución de frecuencias:

| Clase | Límite de clase | Marca de clase | *Frecuencia* | Total |
|---|---|---|---|---|
| **1** | 8.00-819 | 8.1 | | 1 |
| **2** | 8.20-839 | 8.3 | IIIII IIII | 9 |
| **3** | 8.40-859 | 8.5 | IIIII IIIII IIIII I | 16 |
| **4** | 8.60-879 | 8.7 | IIIII IIIII IIIII IIIII IIIII II | 27 |
| **5** | 8.80-8.99 | 8.9 | IIIII IIIII IIIII IIIII IIIII IIIII I | 31 |
| **6** | 9.00-9.19 | 9.1 | IIIII IIIII IIIII IIIII III | 23 |
| **7** | 9.20-9.39 | 9.3 | IIIII IIIII II | 12 |
| **8** | 9.40-9.59 | 9.5 | II | 2 |
| **9** | 9.60-9.79 | 9.7 | IIII | 4 |
| **10** | 9.80-9.99 | 9.9 | | 0 |
| | | | Suma de “f” = N = | = 125 |

**Tabla de distribución de frecuencias**

Al graficar los datos anteriores obtenemos la siguiente figura:

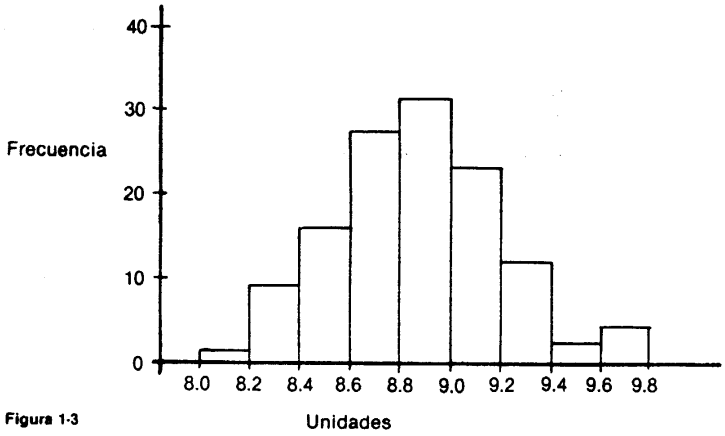



Figura 1-3

**Histograma de frecuencias**

La forma más habitual de representar la información contenida en una tabla es a partir de un sistema de ejes cartesianos. Hay, no obstante, otras formas de representar datos, como posteriormente veremos, que están básicamente orientadas a

características no cuantitativas o atributos. Para hacer más clara la exposición de las diferentes representaciones gráficas, distinguiremos las referentes a dos tipos de distribuciones:

- **Distribuciones sin agrupar**
- **Distribuciones agrupadas en intervalos**

**Gráficos para distribuciones de frecuencias no agrupadas**

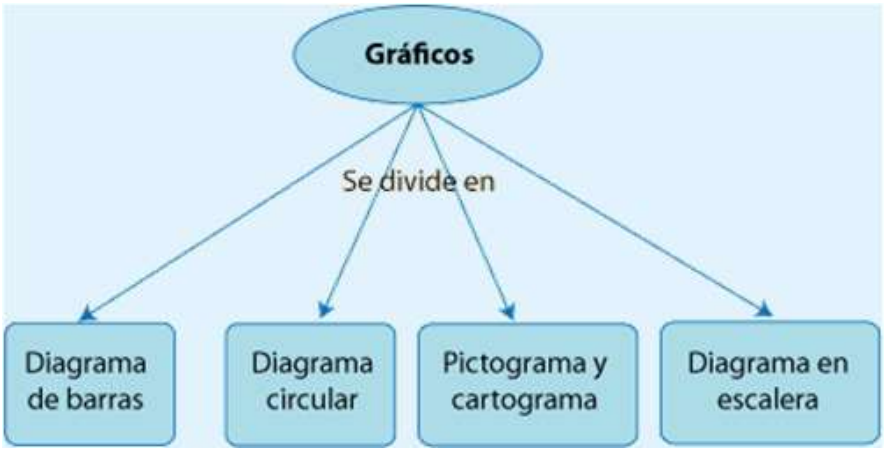
Para representar este tipo de distribuciones, los gráficos más utilizados son:

**a)** El diagrama de barras, que se emplea para distribuciones tanto de variables estadísticas como de atributos.

**b)** El diagrama circular, que es el más comúnmente utilizado para distribuciones de atributos.

**c)** El pictograma y el cartograma.

**d)** Diagrama en escalera, empleado para frecuencias acumuladas.



**a)** Diagrama de barras

Es la más sencilla de las gráficas; representa los datos mediante una barra o columna, que puede colocarse horizontal o verticalmente. Permite comparar las proporciones que guardan cada una de las partes con respecto al todo, por lo que pueden construirse usando valores absolutos, proporciones o, bien, porcentajes. Suelen utilizarse cuando se comparan gráficamente las distribuciones de iguales conceptos en dos o más periodos. Asimismo, constituye la representación gráfica más utilizada, por su capacidad para adaptarse a numerosos conjuntos de datos.

La forma de elaborar estos diagramas es la siguiente:

1. Sobre unos **ejes de coordenadas** se representan en las abscisas los diferentes valores de la variable y en las ordenadas las frecuencias.
2. Sobre cada **valor de la variable** se levanta una barra cuya altura sea la frecuencia correspondiente.
3. Esta representación será un conjunto de barras; por ello se denomina diagrama de barras.

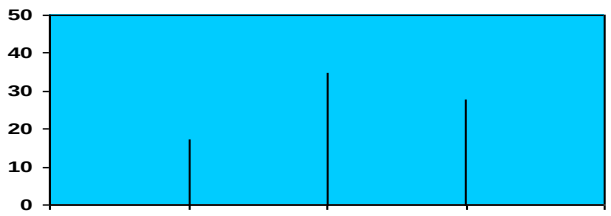



**Diagrama de barras**

A partir de este diagrama, es fácil darse cuenta en qué valores de la variable se concentra la mayor parte de las observaciones.

Una variante de este diagrama, más utilizada quizá por ser más ilustrativa, es el *diagrama de rectángulos*, que representa en el eje de las abscisas los valores de la variable y en el de las ordenadas las frecuencias. Pero ahora, sobre cada valor de la

variable, se levanta un rectángulo con base constante y altura proporcional a la frecuencia absoluta.

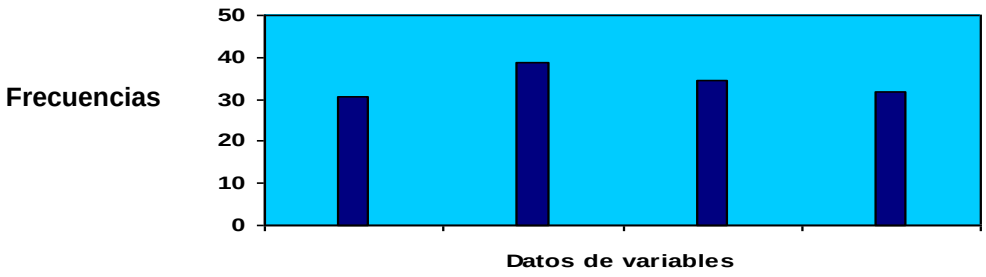



**Diagrama de rectángulos**

Aunque los datos gráficos son equivalentes, generalmente se opta por el de rectángulo por ser, a simple vista, más ilustrativo.

Además, el diagrama de rectángulos es especialmente útil cuando se desea comparar, en un mismo gráfico, el comportamiento del fenómeno en dos o más situaciones o ámbitos distintos, para lo cual podemos usar colores, uno por ámbito, y con ello obtener una visión simplificada y conjunta de lo que ocurre en ambos casos por tratar. Ejemplo de análisis comparativo que puede ser representado con rectángulos en dos tonos.

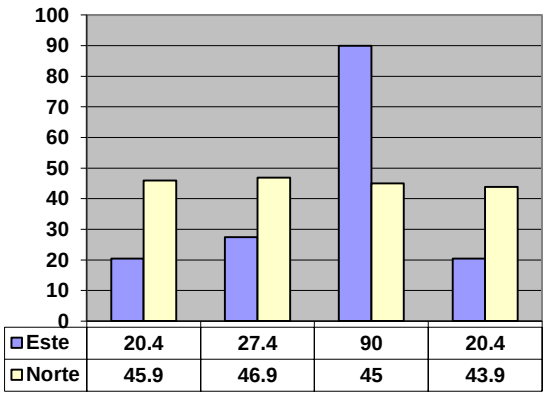


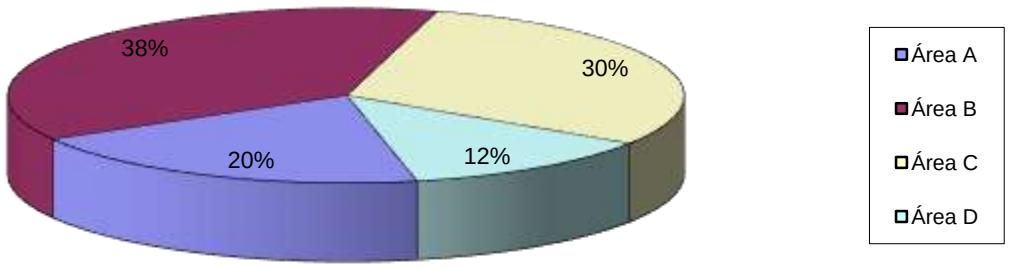

**Diagrama de rectángulos**

**b)** Diagrama circular

Esta representación gráfica es especialmente adecuada en aquellos casos en que se desea que los datos estadísticos lleguen a todo tipo de persona, incluso a las que no tienen por qué tener una formación científica.

Este tipo de diagrama muestra la importancia relativa de las diferentes partes que componen un total. La forma de elaborarlo es la siguiente:

- Se traza un círculo.
- A continuación, se divide éste en tantas partes como componentes haya; el tamaño de cada una de ellas será proporcional a la importancia relativa de cada componente. En otras palabras, como el círculo tiene 360º, éstos se reparten proporcionalmente a las frecuencias absolutas de cada componente.




**Gráfica circular o pastel**

La ventaja intrínseca de este tipo de representaciones no debe hacer olvidar que plantea ciertas desventajas que enumeramos a continuación:

1. Requiere cálculos adicionales.
2. Es más difícil comparar segmentos de un círculo que comparar alturas de un diagrama de barras.
3. No da información sobre las magnitudes absolutas, a menos que las incorporemos en cada segmento.

**c1)** Pictograma

Es otra forma de representar distribuciones de frecuencias. Consiste en tomar como unidad una silueta o símbolo que sea representativo del fenómeno que se va a estudiar.

Por ejemplo:

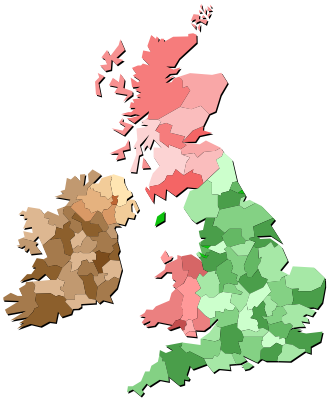
 = 100 viviendas

Y para representar 300 viviendas

 = 300 viviendas

**c2) Cartograma**

Son especialmente útiles en estudios de carácter geográfico. La forma de construirlos es la siguiente: se colorea o se raya con colores e intensidades diferentes los distintos espacios o zonas (que pueden ser comunidades autónomas, provincias, ríos, etc.) en función de la mayor o menor importancia que tenga la variable o atributo en estudio.




Fuente: Revista *Expansión*, núm. 852 (octubre 30 del 2002), p. 69.

**d) Diagrama en escalera**

Su nombre responde a que la representación tiene forma de escalera. Se utiliza para representar frecuencias acumuladas. Su construcción es similar a la del diagrama de barras; y se elabora de la forma siguiente:

- En el eje de las abscisas se miden los valores de la variable o las modalidades del atributo; en el de las ordenadas, las frecuencias absolutas acumuladas.
- Se levanta, sobre cada valor o modalidad, una barra, cuya altura es su frecuencia acumulada.
- Por último, se unen mediante líneas horizontales cada frecuencia acumulada a la barra de la siguiente.

- Los pasos anteriores conducen a la escalera; la última ordenada corresponderá al número total de observaciones.

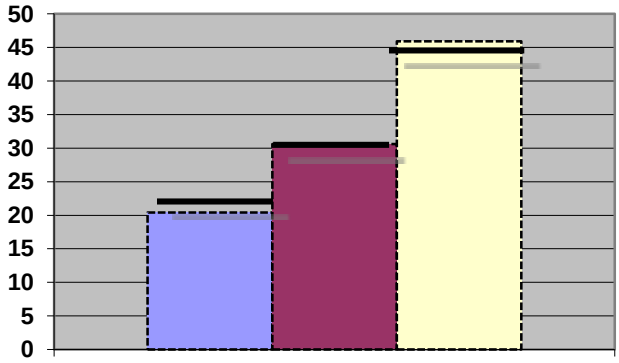


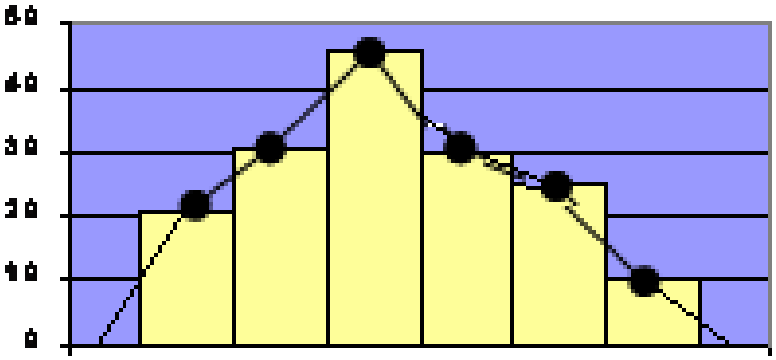

**Diagrama en escalera**

**Gráficas para distribuciones de frecuencias agrupadas en clases**

Para distribuciones agrupadas en intervalos existen básicamente tres tipos de representaciones gráficas: el histograma, el polígono de frecuencias y las ojivas.

**Polígono de frecuencias**

Es un gráfico de línea que se construye, sobre el sistema de coordenadas cartesianas, al colocar sobre cada marca de clase un punto a la altura de la frecuencia asociada a esa clase; posteriormente, estos puntos se unen por segmentos de recta. Para que el polígono quede cerrado se debe considerar un intervalo más al inicio y otro al final con frecuencias cero.



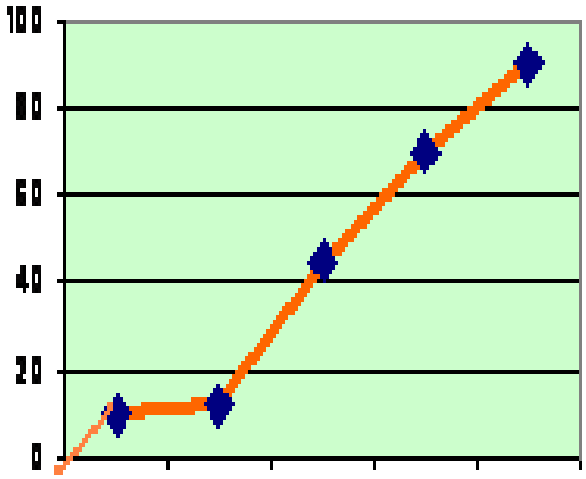
Polígono de frecuencias

**Ojivas**

Si en lugar de frecuencias absolutas utilizamos las acumuladas, obtendremos, en vez del histograma, una representación gráfica en forma de línea creciente que se conoce con el nombre de ojiva. Estos gráficos son especialmente adecuados cuando se tiene interés en saber cuántas observaciones se acumulan hasta diferentes valores de la variable, esto es, cuántas hay en la zona izquierda o inferior del límite superior de cualquier intervalo.

La ojiva es el polígono que se obtiene al unir por segmentos de recta los puntos situados a una altura igual a la frecuencia acumulada a partir de la marca de clase, en la misma forma en que se realizó para construir el polígono de frecuencias.

La ojiva también es un polígono que se puede construir con la frecuencia acumulada relativa.



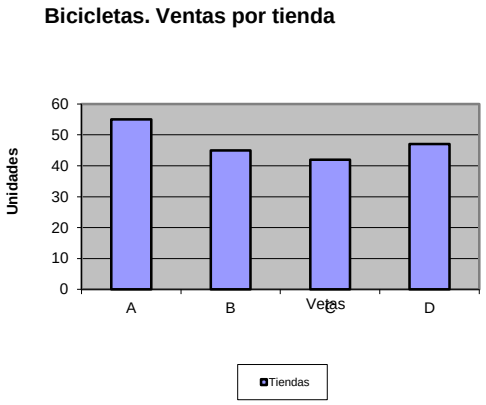

**Ojivas**

Fuente: Revista *Expansión*, núm. 852, (octubre 30 del 2002), p. 14.

En los siguientes ejemplos se observan los tipos de gráficas estudiadas:

## Columnas

Este tipo de gráficas nos permite visualizar información de categorías con mucha facilidad.



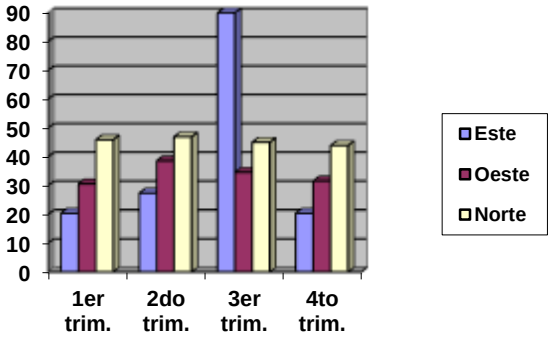




**Diagrama de columnas**

## Barras

Tiene la misma utilidad que el de columnas, pero en este caso con un formato horizontal.

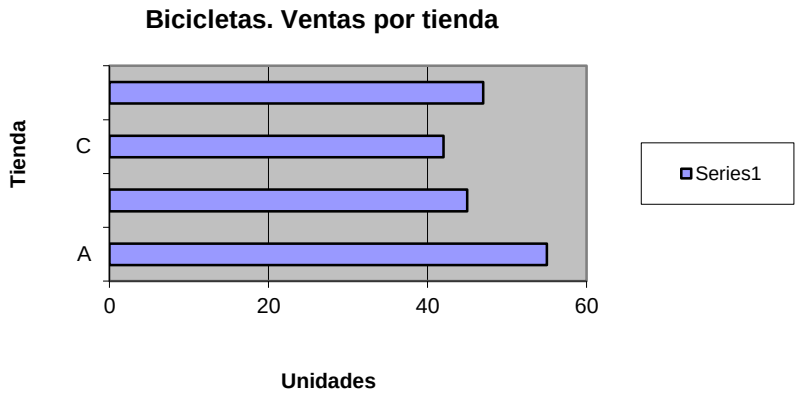



**Diagrama de barras**

## Circular

Presenta de una manera muy objetiva las proporciones que tiene cada una de las categorías en el total, como si fueran las tajadas de un pastel.


**Diagrama circular**

# 1.4 Medidas de tendencia central

Hemos visto que tanto las tablas como las gráficas pueden sernos útiles para representar y comprender información numérica. Existen, sin embargo, circunstancias en las que ni las tablas ni las gráficas nos dan información suficiente para tomar decisiones. En esos casos debemos procesar nuestros datos de diversas maneras para obtener información. A estas medidas se les llama “parámetros” de acuerdo con lo visto en la unidad 1. Se dividen en **medidas de posición y medidas de dispersión**.

**Medidas de posición**

Son aquellas que nos definen (o nos informan) del valor de datos que ocupan lugares importantes en nuestra distribución; las podemos dividir de la siguiente forma: a unas, en medidas de tendencia central y, a otras, en medidas de posición.

Las **medidas de tendencia central** son las que nos indican datos representativos de una distribución y que tienden a ubicarse en el centro de la misma.

A su vez, las medidas de posición tienen el objetivo de localizar diversos puntos de interés ubicados en diversas partes de la distribución; por ejemplo, el punto que divide la distribución en dos partes: a la izquierda (datos más pequeños), 25% de la información y a la derecha (datos más grandes), el 75% de la información. A este punto se le denomina primer cuartil o Q1.

A continuación daremos las definiciones y algunos ejemplos de las medidas de tendencia central y concluiremos el apartado con las medidas de posición.

Las medidas de tendencia central que se contemplan en este material son: la media aritmética, la mediana y la moda.

**Media aritmética**

La media aritmética es el promedio que todos conocemos desde nuestros años de infancia. Se obtiene sumando todos los datos y dividiendo el total entre el número de datos. Podemos decir entonces que la media aritmética determina cómo repartir un total entre N observaciones si el reparto es a partes iguales.

La manera formal de expresar este concepto es la siguiente:

$$\mu = \sum_{i=1}^{N} x_i \, / \, N$$

Esta expresión nos dice que la media aritmética, que está representada por la letra griega $\mu$, se obtiene sumando todos los datos a los que llamamos X subíndice i para, posteriormente, dividir el resultado entre “N”, que es el número total de datos con los que se cuenta.

Considera el siguiente ejemplo: Las calificaciones en los dos primeros semestres de un alumno que estudia la licenciatura en administración se listan a continuación:9, 10, 8, 8, 9, 7, 6, 10, 8, 8,7.

La media aritmética está dada por la siguiente expresión:

$$\mu = (9+10+8+8+9+7+6+10+8+8+7)/11$$

Haciendo las operaciones encontramos que la media aritmética es aproximadamente de 8.18.

**Mediana**

Es el valor que divide la distribución en dos partes iguales y se le conoce como Md. Para obtenerla se deben ordenar los datos (puede ser de menor a mayor o viceversa, no importa) y se encuentra el dato medio. En el caso de las calificaciones del estudiante indicadas arriba, los datos ordenados tendrían el siguiente aspecto:

6, 7, 7, 8, 8, **8**, 8, 9, 9, 10 ,10

El dato que divide la distribución a la mitad se señala con una flecha. Este dato corresponde a la mediana. Como se puede ver a la izquierda del 8 encontramos cinco datos y, a su derecha encontramos otros cinco datos. Este dato es, entonces, el correspondiente a la mediana; así, Md=8.

Si en lugar de un número impar de datos (como en nuestro ejemplo anterior), nos encontramos con un número par de observaciones, lo que se hace es promediar los dos datos medios. El procedimiento se muestra en el siguiente ejemplo:

Las ventas diarias de una pequeña tienda durante una corta temporada vacacional se consignan a continuación. Ya se ordenaron de menor a mayor para facilitar el trabajo posterior:

3,200; 3,500; 3,650; **3,720; 3,750**; 3,810; 3,850; 3,915

Puede verse fácilmente que no hay un dato central que divida la distribución en dos, por ello se toman los dos datos centrales y se promedian. En este caso la mediana es de 3,735, que es la media aritmética de los dos datos centrales.

**Moda**

Es el dato más frecuente de nuestro conjunto. En el caso de las calificaciones del estudiante el dato más frecuente es “8”, como se puede ver si repetimos nuestro conjunto de datos.

6, 7, 7, **8, 8, 8, 8**, 9, 9, 10 ,10.

En el caso de las ventas de la tienda, se puede ver que no hay dos datos iguales; por lo mismo, este conjunto de datos no tiene moda.

Puede darse el caso, en conjuntos más grandes de datos, que el “honor” de ser el valor más frecuente sea compartido por dos datos. En ese caso se afirma que la distribución es **bimodal**, pues tiene dos modas. Algunos autores llegan a hablar de distribuciones **trimodales** e incluso más.

**Cuartiles**

Así como la mediana divide la distribución de nuestros datos en dos partes iguales, existen medidas de posición llamadas **cuartiles**. Hay tres **cuartiles** en cada distribución de datos; el **primer cuartil** o Q1 divide la distribución en dos partes: a la izquierda está la cuarta parte (de allí su nombre) o el 25% de los datos. El **segundo cuartil** o Q2 se asimila a la mediana y divide la distribución de nuestros datos en dos partes iguales. El **tercer cuartil** o Q3 hace la misma función, pues divide nuestra distribución de datos en dos partes, la parte izquierda agrupa al 75% de los datos más pequeños y la parte derecha el 25% de los datos más grandes. El siguiente esquema puede aclarar la situación de los **cuartiles**:


**Posición de cuartiles**

Cada una de las barras color naranja representa un 25% de los datos.

Hay otras dos medidas de posición que se asemejan al concepto de cuartiles. Se trata de los “**deciles**” y los “**percentiles**”, sólo que éstas son medidas que en lugar de separar los datos en grupos de 25% lo hacen en grupos de 10% y de 1%, respectivamente.

Desde luego, para que los cuartiles, deciles y percentiles tengan algún sentido se requiere tener conjuntos grandes de datos.

Por ejemplo, no tiene ningún objeto hablar de percentiles si se tienen 14 datos. La manera de encontrar los cuartiles, deciles o percentiles sería, en teoría, la misma; es decir, alinear los datos de menor a mayor y contar cuál de ellos es el que cumple el requisito de dividir la distribución de la manera que queremos, pero este método es completamente impráctico, por lo que nos ocuparemos de su obtención cuando trabajemos datos agrupados.

## 1.5. Medidas de dispersión

Saber cuál es el dato central de una distribución es importante, pero también lo es saber qué tan concentrada o extendida está nuestra información. Por ejemplo, saber que una tienda tiene ingresos diarios medios de $10,000 es interesante, pero además es importante saber si todos los días esas ventas están muy cerca de los diez mil pesos o, en realidad, se alejan mucho. Enseguida damos los datos de dos tiendas que tienen la misma media de ventas diarias.

Tienda A: $10,000; $10,500; $11,000; $9,000; $9,500.

Tienda B: $10,000; $5,000; $15,000; $19,000; $1,000.

Es fácil observar que ambas tiendas tienen las mismas ventas medias ($10,000). Sin embargo, en la tienda A la planeación de flujo de efectivo es más sencilla que en la tienda B. En la primera podemos contar con un flujo más o menos constante de efectivo que nos permite afrontar los compromisos diarios; en la segunda podemos tener un flujo muy abundante o casi nada. Eso nos lleva a tener que prever cómo invertir excedentes temporales y cómo cubrir faltantes en el corto plazo.

Las medidas que nos permiten cuantificar la dispersión de los datos son cuatro: **el rango** o **recorrido**, **la varianza**, **la desviación estándar** y **el coeficiente de variación**. A continuación definimos cada una de ellas.

**Rango o recorrido**

Es la diferencia entre el dato mayor y el dato menor. En el ejemplo de las tiendas sus rangos son:

Tienda A: 11,000 – 9,000 = 2,000.

Tienda B: 19,000 – 1,000 = 18,000.

El rango se expresa frecuentemente con la siguiente fórmula:

$$R = X_M - X_m$$

En esta fórmula R representa al rango; $X_M$ al dato mayor y $X_m$ al dato menor.

El rango es una medida de dispersión que es muy fácil de obtener, pero es un tanto burda, pues solamente toma en cuenta los datos extremos y **no considera los datos que están en medio**. Para tomar en cuenta todos los datos se inventaron las medidas de dispersión que son la varianza y la desviación estándar.

**Varianza y desviación estándar**

Supongamos las ventas de las siguientes dos tiendas:

Tienda C: $5,000; $10,000; $10,000; $10,000; $15,000.

Tienda D: $5,000; $6,000; $10,000; $14,000; $15,000.

Ambas tiendas tienen una media de $10,000 y un rango de $10,000, como fácilmente el alumno puede comprobar; sin embargo, podemos darnos cuenta de que en la tienda D la información está un poco más dispersa que en la tienda C, pues en esta última, si exceptuamos los valores extremos, todos los demás son diez mil; en cambio, en la tienda D existe una mayor diversidad de valores.

Un enfoque que nos puede permitir tomar en cuenta todos los datos es el siguiente:

Supongamos que deseamos saber qué tan alejado está cada uno de los datos de la media. Para ello podemos sacar la diferencia entre cada uno de los datos y esa media para, posteriormente, promediar todas esas diferencias y ver, en promedio, qué tan alejado está cada dato de la media ya citada. En la siguiente tabla se realiza ese trabajo.

| Tienda C | | Tienda D | |
|---|---|---|---|
| **Datos** | **Cada dato menos la media** | **Datos** | **Cada dato menos la media** |
| 5,000 | 5,000 – 10,000 = –5,000 | 5,000 | 5,000 – 10,000 = –5,000 |
| 10,000 | 10,000 – 10,000 = 0 | 6,000 | 6,000 – 10,000 = –4,000 |
| 10,000 | 10,000 – 10,000 = 0 | 10,000 | 10,000 – 10,000 = 0 |
| 10,000 | 10,000 – 10,000 = 0 | 14,000 | 14,000 – 10,000 = 4,000 |
| 15,000 | 15,000 – 10,000 = 5,000 | 15,000 | 15,000 – 10,000 = 5,000 |
| | **Suma = 0** | | **Suma = 0** |

**Tabla de desviaciones de datos**

Como se puede apreciar la suma de las diferencias entre la media y cada dato tiene como resultado el valor cero, por lo que, entonces, se elevan las diferencias al cuadrado para que los resultados siempre sean positivos.

A continuación se muestra este trabajo y la suma correspondiente.

| Tienda C | | | Tienda D | | |
|---|---|---|---|---|---|
| **Datos** | **Cada dato menos la media** | **Cuadrado de lo anterior** | **Datos** | **Cada dato menos la media** | **Cuadrado de lo anterior** |
| 5,000 | 5,000 | 25,000,000 | 5,000 | –5,000 | 25,000,000 |
| 10,000 | 0 | 0 | 6,000 | –4,000 | 16,000,000 |
| 10,000 | 0 | 0 | 10,000 | 0 | 0 |
| 10,000 | 0 | 0 | 14,000 | 4,000 | 16,000,000 |
| 15,000 | 5,000 | 25,000,000 | 15,000 | 5,000 | 25,000,000 |
| **SUMA** | **0** | **50,000,000** | **SUMA** | **0** | **82,000,000** |

**Tabla de desviaciones cuadráticas**

En este caso, ya la suma de las diferencias entre cada dato y la media elevadas al cuadrado nos da un valor diferente de cero con el que podemos trabajar. A este último dato (el de la suma), dividido entre el número total de datos lo conocemos como varianza (o variancia, según el libro que se consulte).

De acuerdo con lo anterior, tenemos que la varianza de los datos de la tienda C es igual a 50, 000,000/5, es decir 10, 000,000. Siguiendo el mismo procedimiento podemos obtener la varianza de la tienda D, que es igual a 82, 000,000/5, es decir, 16, 500,000.

Es en este punto cuando nos podemos percatar que la varianza de la tienda D es mayor que la de la tienda C, por lo que la información de la primera de ellas (D) está más dispersa que la información de la segunda (C). En resumen:

> La varianza es la medida de dispersión que corresponde al promedio aritmético de las desviaciones cuadráticas de cada valor de la variable, con respecto a la media de los datos.

La expresión algebraica que corresponde a este concepto es la siguiente:

$$\sigma^2 = \sum_{1}^{N} (x_i - \mu)^2 / N$$

En donde:

$\sigma^2$ es la varianza de datos.

$\sum$ indica una sumatoria.

$x_i$ variable o dato.

$\mu$ media de datos.

$N$ número de datos en una población.

La **varianza** es una medida muy importante y tiene interesantes aplicaciones teóricas. Sin embargo, es difícil de comprender de manera intuitiva, entre otras cosas porque al elevar las diferencias entre el dato y la media al cuadrado, las unidades de medida también se elevan al cuadrado y no es nada fácil captar lo que significan, por ejemplo, pesos al cuadrado (o en algún otro problema focos al cuadrado). Por ello se determinó obtener la raíz cuadrada de la varianza. De esta manera las unidades vuelven a expresarse de la manera original y su sentido es menos difícil de captar.

La raíz cuadrada de la varianza recibe el nombre de **desviación estándar o desviación típica**.

En el caso de nuestras tiendas, las desviaciones estándar son para la tienda C $3,162.28 y para la tienda D $4,062.02.

La fórmula para la desviación estándar es:

$$\sigma = \sqrt{\sum_{1}^{N} (x_i - \mu)^2 / N}$$

El alumno podrá observar que la sigma ya no está elevada al cuadrado, lo que es lógico, pues si la varianza es sigma al cuadrado, la raíz cuadrada de la misma es, simplemente sigma. Es importante precisar que ésta es la fórmula de la desviación estándar para una población.

En estadística inferencial es importante distinguir los símbolos para una muestra y para una población. La desviación estándar para una muestra tiene una fórmula cuyo denominador es (n–1) siendo “n” el tamaño de la muestra.

El estudiante deberá notar que al total de la población se le denota con “N” mayúscula y al total de datos de la muestra se le denota con “n” minúscula.

**El coeficiente de variación**

Dos poblaciones pueden tener la misma desviación estándar y, sin embargo, podemos percatarnos intuitivamente que la dispersión no es la misma para efectos de una toma de decisiones.

El siguiente ejemplo aclara estos conceptos.

Un comercializador de maíz vende su producto de dos maneras distintas:

a) En costales de 50 kg.

b) A granel, en sus propios camiones repartidores que cargan 5 toneladas (5,000 kg).

Para manejar el ejemplo de manera sencilla, supongamos que en un día determinado solamente vendió tres costales y que además salieron tres camiones cargados; para verificar el trabajo de los operarios, se pesaron tanto unos como otros en presencia de un supervisor. Sus pesos, la media de los mismos y sus desviaciones estándar aparecen en la siguiente tabla (como ejercicio, el alumno puede comprobar las medias y las desviaciones estándar calculándolas él mismo):

| Peso de los costales | Peso de los camiones |
|---|---|
| **40 Kg** | **4,990 Kg** |
| **50 Kg** | **5,000 Kg** |
| **60 Kg** | **5,010 Kg** |

**Tabla de datos**

- Media de los costales 50 kg.
- Media de los camiones 5,000 kg.
- Desviación estándar de los costales 8.165 kg.
- Desviación estándar de los camiones 8.165 kg.

Podemos percatarnos de que las variaciones en el peso de los camiones son muy razonables, dado el peso que transportan. En cambio, las variaciones en el peso de

los costales son muy grandes, en relación con lo que debería de ser. Los operarios que cargan los camiones pueden ser felicitados por el cuidado que ponen en su trabajo, en cambio podemos ver fácilmente que los trabajadores que llenan los costales tienen algún problema serio, a pesar de que la variación (la desviación estándar) es la misma en ambos casos.

Para formalizar esta relación entre la variación y lo que debe de ser, se trabaja el coeficiente de variación o dispersión relativa, que no es otra cosa que la desviación estándar entre la media y todo ello por cien. En fórmula lo expresamos de la siguiente manera:

$$C.V. = (\sigma / \mu)100$$

donde:

$C.V.$ coeficiente de variación.

$\sigma$ desviación estándar.

$\mu$ media de la población.

En el caso de los costales tendíamos que C.V.= (8.165/50)100=16.33, lo que nos indica que la desviación estándar del peso de los costales es de 16.33% del peso medio (una desviación significativamente grande).

Por otra parte, en el caso de los camiones, el coeficiente de variación nos arroja:

C.V.= (8.165/5000)100= 0.1633, lo que nos indica que la desviación estándar del peso de los camiones es de menos del uno por ciento del peso medio (una desviación realmente razonable).

**Datos agrupados en clases o eventos**

Cuando se tiene un fuerte volumen de información y se debe trabajar sin ayuda de un paquete de computación, no es práctico trabajar con los datos uno por uno, sino que

conviene agruparlos en subconjuntos llamados "**clases**", ya que así es más cómodo manipularlos aunque se pierde alguna precisión.

Imagine que se tienen 400 datos y el trabajo que representaría ordenarlos uno por uno para obtener la mediana. Por ello se han desarrollado técnicas que permiten el trabajo rápido mediante agrupamiento de datos. A continuación se dan algunas definiciones para, posteriormente, pasar a revisar las técnicas antes citadas.

**Clase**: Cada uno de los subconjuntos en los que dividimos nuestros datos.

**Número de clases**: Debemos definirlo con base en el número total de datos.

Hay varios criterios para establecer el número de clases. Entre ellos, que el número de clases es aproximadamente…

- la raíz cuadrada del número de datos.
- el logaritmo del número de datos entre el logaritmo de 2.

Normalmente se afirma que las clases no deben ser menores que cinco ni mayores que veinte. De cualquier manera, el responsable de trabajar con los datos puede utilizar su criterio.

A continuación se dan algunos ejemplos del número de clases que se obtienen según los dos criterios antes señalados.

| Número de datos | Número de clases | |
|---|---|---|
| | (Criterio de la raíz cuadrada) | (Criterio del logaritmo) |
| 50 | Aproximadamente 7 | 6 |
| 100 | Aproximadamente 10 | 7 |
| 150 | Aproximadamente 12 | 7 |
| 200 | Aproximadamente 14 | 8 |

**Tabla de Número de clases según número de datos**

Supongamos que tenemos 44 datos —como en el caso de la tabla que se presenta a continuación—, que corresponden a las ventas diarias de una pequeña miscelánea. Si seguimos el criterio de los logaritmos, el número de clases será: logaritmo de 44 entre

logaritmo de 2, esto es, log 44 / log 2= 1.6434/0.3010 = 5.46, es decir, aproximadamente 5 clases.

| Miscelánea "La Esperanza" Ventas de 44 días consecutivos | | | | | | | |
|---|---|---|---|---|---|---|---|
| Día | Venta | Día | Venta | Día | Venta | Día | Venta |
| 1 | 508 | 12 | 532 | 23 | 763 | 34 | 603 |
| 2 | 918 | 13 | 628 | 24 | 829 | 35 | 890 |
| 3 | 911 | 14 | 935 | 25 | 671 | 36 | 772 |
| 4 | 639 | 15 | 606 | 26 | 965 | 37 | 951 |
| 5 | 615 | 16 | 680 | 27 | 816 | 38 | 667 |
| 6 | 906 | 17 | 993 | 28 | 525 | 39 | 897 |
| 7 | 638 | 18 | 693 | 29 | 846 | 40 | 742 |
| 8 | 955 | 19 | 586 | 30 | 773 | 41 | 1000 |
| 9 | 549 | 20 | 508 | 31 | 547 | 42 | 800 |
| 10 | 603 | 21 | 885 | 32 | 624 | 43 | 747 |
| 11 | 767 | 22 | 590 | 33 | 524 | 44 | 500 |

**Tabla de ventas**

### Ancho de clase

Es el tamaño del intervalo que va a ocupar cada clase. Se considera que el ancho de clase se obtiene dividiendo el rango entre el número de clases. Así, en el ejemplo de la miscelánea nuestro dato mayor es 999.70, nuestro dato menor es 500 y anteriormente habíamos definido que necesitábamos cinco clases, por lo que el ancho de clase es el rango (499.70 o prácticamente 500) entre el número de clases (5). Por tanto, el ancho de clase es de 100.

### Límites de clase

Es el punto en el que termina una clase y comienza la siguiente. En el ejemplo del párrafo anterior podemos resumir la información de la siguiente manera:

Primera clase: comienza en 500 y termina en 600

Segunda clase: comienza en 600 y termina en 700

Tercera clase: comienza en 700 y termina en 800

Cuarta clase: comienza en 800 y termina en 900

Quinta clase: comienza en 900 y termina en 1,000

Estas clases nos permitirán clasificar nuestra información. Si un dato, por ejemplo, tiene el valor de 627.50, lo colocaremos en la segunda clase. El problema que tiene esta manera de clasificar la información es que en los casos de datos que caen exactamente en los límites de clase, no sabríamos en cuál de ellas clasificarlos. Si un dato es exactamente 700, no sabríamos si debemos asignarlo a la segunda o a la tercera clase. Para remediar esta situación existen varios caminos, pero el más práctico de ellos (y el que usaremos para los efectos de este trabajo) es el de hacer intervalos abiertos por un lado y cerrados en el otro.

Esto se logra de la siguiente manera:

| Clase | Incluye datos Iguales o mayores a: | Incluye datos menores a: |
|---|---|---|
| Primera | 500 | 600 |
| Segunda | 600 | 700 |
| Tercera | 700 | 800 |
| Cuarta | 800 | 900 |
| Quinta | 900 | 1000 |

**Tabla de clases**

Como vemos, los intervalos de cada clase están cerrados por la izquierda y abiertos por la derecha. Se puede tomar la decisión inversa y dejar abierto el intervalo del lado izquierdo y cerrado del lado derecho. Este enfoque se ejemplifica en la siguiente tabla.

| Clase | Incluye datos mayores a: | Incluye datos menores o iguales a: |
|---|---|---|
| Primera | 500 | 600 |
| Segunda | 600 | 700 |
| Tercera | 700 | 800 |
| Cuarta | 800 | 900 |
| Quinta | 900 | 1000 |

**Tabla de clases**

En lo único que se debe tener cuidado es en no excluir alguno de nuestros datos al hacer la clasificación. En el caso de la última tabla, por ejemplo, excluimos a los datos cuyo valor es exactamente de 500. Podemos dejarlo así partiendo de la base de que esto no tendrá impacto en nuestro trabajo, o bien podemos ajustar los límites para dar cabida a todos los datos. A continuación, se presenta un ejemplo de esta segunda alternativa.

| Clase | Incluye datos iguales o mayores a: | Incluye datos menores a: |
|---|---|---|
| Primera | 499.99 | 599.99 |
| Segunda | 599.99 | 699.99 |
| Tercera | 699.99 | 799.99 |
| Cuarta | 799.99 | 899.99 |
| Quinta | 899.99 | 999.99 |

**Tabla de clases**

De esta manera, tenemos contemplados todos nuestros datos. El investigador deberá definir cuál criterio prefiere con base en el rigor que desea y de las consecuencias prácticas de su decisión.

Posteriormente, conforme desarrollemos el ejemplo, se verá el impacto por elegir una u otra de las alternativas.

**Marca de clase**

La marca de clase es, por así decirlo, la representante de cada clase. Se obtiene sumando el límite inferior y el superior de cada clase y promediándolos. A la marca de clase se le conoce como Xi. En nuestro ejemplo se tendría:

| Clase | Incluye datos iguales o mayores a: | Incluye datos menores a: | Marca de clase (Xi) |
|---|---|---|---|
| Primera | 500 | 600 | **(500+600)/2=550** |
| Segunda | 600 | 700 | **(600+700)/2=650** |
| Tercera | 700 | 800 | **(700+800)/2=750** |
| Cuarta | 800 | 900 | **(800+900)/2=850** |
| Quinta | 900 | 1000 | **(900+1000)/2=950** |

**Marcas de clase**

Éstas serían las marcas si las clases se construyen como en la primera tabla de clases.

Si se aplica el criterio de la tercera tabla, las marcas quedarían como sigue:

| Clase | Incluye datos iguales o mayores a: | Incluye datos menores a: | Marca e clase (Xi) |
|---|---|---|---|
| Primera | 499.99 | 599.99 | (499.99+599.99)/2=549.99 |
| Segunda | 599.99 | 699.99 | (599.99+699.99)/2=649.99 |
| Tercera | 699.99 | 799.99 | (699.99+799.99)/2=749.99 |
| Cuarta | 799.99 | 899.99 | (799.99+899.99)/2=849.99 |
| Quinta | 899.99 | 999.99 | (899.99+999.99)/2=949.99 |

**Marcas de clase**

Podemos ver que la diferencia entre la marca de clase de las dos primeras tablas y la tercera es de solamente un centavo. Veremos en el resto del ejemplo las consecuencias que tiene esa diferencia en el desarrollo del trabajo.

Una vez que se tiene la “armadura” o estructura en la que se van a clasificar los datos, se procede a clasificarlos. Para esto usaremos una de las clasificaciones ya especificadas:

| Clase | Incluye datos mayores a: | Incluye datos menores o iguales a: | Conteo de casos | Frecuencia en clase (Fi) |
|---|---|---|---|---|
| Primera | 500 | 600 | IIIII IIIII I | 11 |
| Segunda | 600 | 700 | IIIII IIIII I | 11 |
| Tercera | 700 | 800 | IIIII II | 7 |
| Cuarta | 800 | 900 | IIIII I | 6 |
| Quinta | 900 | 1000 | IIIII IIII | 9 |
| **Total:** | | | | **44** |

**Tabla de frecuencias**

Para calcular las medidas de tendencia central y de dispersión en **datos agrupados en clases** se utilizan fórmulas similares a las ya estudiadas y la única diferencia es que se incluyen las frecuencias de clase.

A continuación se maneja un listado y un ejemplo de aplicación:

**Medidas de tendencia central**

**a) Media:**

$$\overline{x} = \frac{\sum_{x=1}^{N} x_i f_i}{n}$$

En donde:

$x_i$ es la marca de clase.

$f_i$ es la frecuencia de clase.

$N$ es el número de clases.

$n$ es el número de datos.

**b) Mediana:**

$$Md = L_M + \frac{n/2 - F_M}{f_M} \cdot i$$

En donde:

$L_M$ es el límite inferior del intervalo que contiene a la mediana.

$F_M$ es la frecuencia acumulada hasta el intervalo que contiene a la mediana.

$f_M$ es la frecuencia absoluta del intervalo que contiene a la mediana.

$i$ es el ancho del intervalo que contiene a la mediana.

**c) Moda o modo:**

$$Mo = L_{Mo} + \frac{d_1}{d_1 + d_2} \cdot i \quad \begin{matrix} d_1 = f_{Mo} - f_1 \\ d_2 = f_{Mo} - f_2 \end{matrix}$$

En donde:

$L_{Mo}$ es límite inferior del intervalo que contiene el modo.

$d_1$ es la diferencia entre la frecuencia de clase $\left( f_{Mo} \right)$ del intervalo que contiene a la moda y la frecuencia de la clase inmediata anterior $\left( f_1 \right)$.

$d_2$ es la diferencia entre la frecuencia de clase $\left( f_{Mo} \right)$ del intervalo que contiene a la moda y la frecuencia de la clase inmediata posterior $\left( f_2 \right)$.

**Medidas de dispersión**

**a) Rango:** Es la diferencia entre el límite superior del último intervalo de clase y el límite inferior del primer intervalo de clase.

**b) Varianza:**

$$\sigma^2 = \frac{\sum \left( x_i - \bar{x} \right)^2 f_i}{n}$$

En donde:

$x_i$ es la marca de clase.

$f_i$ es la frecuencia de clase.

$\bar{x}$ es la media.

$n$ es el número de datos.

**c) Desviación estándar**:

$$\sigma = \sqrt{\frac{\sum \left( x_i - \bar{x} \right)^2 f_i}{n}}$$

**d) Coeficiente de variación:**

$$C.V. = \frac{\sigma}{x}$$

Se puede utilizar indistintamente la simbología de estadísticos o parámetros, si no es necesario distinguir que los datos provienen de una muestra o de una población. En la estadística inferencial sí es importante manejar esta distinción ya que se trabaja con muestras para inferir los parámetros poblacionales.

En el ejemplo siguiente se muestra la utilización de las fórmulas descritas:

En un laboratorio se estudiaron 110 muestras para determinar el número de bacterias por $cm^3$ de agua contaminada en diversas localidades de un estado del país. En la siguiente tabla de trabajo, se muestran las frecuencias encontradas $f_i$ y los diversos cálculos para determinar las medidas de tendencia central y de dispersión de estas muestras:

| **Límites reales** | $x_i$ | $f_i$ | $f_i acum$ | $x_i f_i$ | $(x_i - \bar{x})^2 f_i$ |
|---|---|---|---|---|---|
| 50 – 55 | 52.5 | 4 | 4 | 210.0 | 2,260.57 |
| 55 – 60 | 57.5 | 7 | 11 | 402.5 | 2,466.91 |
| 60 – 65 | 62.5 | 9 | 20 | 562.5 | 1,707.19 |
| 65 – 70 | 67.5 | 12 | 32 | 810.0 | 923.53 |
| 70 – 75 | 72.5 | 15 | 47 | 1,087.5 | 213.50 |
| Md 75 – 80 | 77.5 | 18 | 65 | 1,395.0 | 27.11 |
| Mo 80 – 85 | 82.5 | 20 | 85 | 1,650.0 | 775.58 |
| 85 – 90 | 87.5 | 13 | 98 | 1,137.5 | 1,638.67 |
| 90 – 95 | 92.5 | 7 | 105 | 647.5 | 1,843.27 |
| 95 – 100 | 97.5 | 5 | 110 | 487.5 | 2,252.99 |
| SUMA | | 110 | | 8,390.0 | 14,109.32 |

**Medidas de tendencia central**

**a) Media:**

$$\bar{x} = \frac{\sum_{x=1}^{N} x_i f_i}{n} = \frac{8,390.0}{110} = 76.27$$

El promedio de agua contaminada de todas las muestras es de 76.27 bacterias por $cm^3$.

**b) Mediana:**

$$Md = L_M + \frac{n/2 - F_M}{f_M} \cdot i = 75 + \frac{55 - 47}{18} \cdot 5 = 77.22$$

Se identifica el intervalo que contiene a la mediana (75 – 80) y las frecuencias del límite superior del intervalo anterior del que contiene a la mediana (47) y la frecuencia del propio intervalo (18).

El punto medio de estas muestras es de 77.22 bacterias por $cm^3$.

**c) Moda o modo:**

$$Mo = L_{Mo} + \frac{d_1}{d_1 + d_2} \cdot i \qquad \begin{matrix} d_1 = f_{Mo} - f_1 \\ d_2 = f_{Mo} - f_2 \end{matrix}$$

$Mo = 80 + \frac{2}{2+7} \cdot 5 = 80.11$, en donde: $\begin{matrix} d_1 = 20 - 18 = 2 \\ d_2 = 20 - 13 = 7 \end{matrix}$ y $\begin{matrix} f_{Mo} = 20 \\ f_1 = 18 \\ f_2 = 13 \\ i = 5 \end{matrix}$

El valor modal se encuentra en el intervalo 80 – 85 y exactamente corresponde a 80.11 bacterias por $cm^3$.

**Medidas de dispersión**

**a) Rango:** 100 – 50 = 50. La diferencia es de 50 bacterias por $cm^3$ entre la muestra menos contaminada y la más contaminada.

**b) Varianza:**

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2 f_i}{n} = \frac{14{,}109.32}{110} = 128.27$$

La desviación cuadrática de las muestras con respecto a su media es de 128.7 bacterias por $cm^3$.

**c) Desviación estándar:**

$$\sigma = \sqrt{128.27} = 11.32$$

La desviación lineal de las muestras con respecto a su media es de 11.32 bacterias por $cm^3$.

**d) Coeficiente de variación:**

$$V.I. = \frac{\sigma}{\overline{x}} = \frac{11.32}{76.27} = 0.148 = 14.8\%$$

Este resultado indica que el promedio de la desviación de los datos con respecto a su media se encuentran en un porcentaje aceptable (<25%) para utilizar esta distribución para fines estadísticos.

# 1.6. Teorema de Tchebysheff y regla empírica

El teorema de Tchebysheff y la regla empírica nos permiten inferir el porcentaje de elementos que deben quedar dentro de una cantidad específica de desviaciones estándar respecto a la media. Ambas herramientas se utilizan principalmente para estimar el número aproximado de datos que se encuentran en determinadas áreas de la distribución de datos.

**Teorema de Tchebysheff o (Chebyshev).**

Cuando menos $1-\frac{1}{k^2}$ de los elementos en cualquier conjunto de datos debe estar a menos de “k” desviaciones estándar de separación respecto a la media, “k” puede ser cualquier valor mayor que 1.

Por ejemplo, veamos algunas implicaciones de este teorema con k=2, 3, y 4 desviaciones estándar:

- cuando menos el 0.75 o 75% de los elementos deben estar a menos de z=2 desviaciones estándar del promedio.
- cuando menos el 0.89 u 89% de los elementos deben estar a menos de z=3 desviaciones estándar del promedio.
- cuando menos el 0.94 o 94% de los elementos deben estar a menos de z=4 desviaciones estándar del promedio.

**Ejemplo 1**. Supongamos que las calificaciones de 100 alumnos en un examen parcial de estadística tuvieron un promedio de 70 y una desviación estándar de 5. ¿Cuántos alumnos tuvieron calificaciones entre 60 y 80? ¿Cuántos entre 58 y 82?

Solución:

Para las calificaciones entre 60 y 80 vemos que el valor de 60 está a 2 desviaciones estándar abajo del promedio y que el valor de 80 está a 2 desviaciones estándar arriba.

Al aplicar el teorema de Tchebysheff, cuando menos el 0.75 o 75% de los elementos deben tener valores a menos de dos desviaciones estándar del promedio. Así, cuando menos 75 de los 100 alumnos deben haber obtenido calificaciones entre 60 y 80.

Para las calificaciones entre 58 y 82, el cociente (58-70)/5=2.4 indica que 50 está a 2.4 desviaciones estándar abajo del promedio, en tanto que (82-70)/5=2.4 indica que 82 está a 2.4 desviaciones estándar arriba del promedio. Al aplicar el teorema de Tchebysheff con z=2.4 tenemos que:

$$1-\frac{1}{k^2}=1-\frac{1}{2.4^2}=0.826$$

Cuando menos 82.6% de los alumnos deben tener calificaciones entre 58 y 82.

Como podemos ver, en el teorema de Tchebysheff se requiere que **z sea mayor que uno**, pero no necesariamente debe ser un entero.

Una de las ventajas del teorema de Tchebysheff es que se aplica a cualquier conjunto de datos, independientemente de la forma de la distribución de los mismos.

Sin embargo, en las aplicaciones prácticas se ha encontrado que muchos conjuntos de datos tienen una distribución en forma de colina o de campana, en cuyo caso se dice que tienen una distribución normal.

Cuando se cree que los datos tienen aproximadamente esa distribución se puede aplicar la regla empírica para determinar el porcentaje de elementos que debe estar dentro de determinada cantidad de desviaciones estándar respecto del promedio.

**La regla empírica**

La regla empírica dice que para conjuntos de datos que se distribuyen de una manera normal (en forma de campana):

- aproximadamente 68% de los elementos están a menos de una desviación estándar de la media.
- aproximadamente 95% de los elementos están a menos de dos desviaciones estándar de la media.
- casi todos los elementos están a menos de tres desviaciones estándar de la media.

**Ejemplo 2**: En una línea de producción se llenan, automáticamente, envases de plástico con detergente líquido. Con frecuencia, los pesos de llenado tienen una distribución en forma de campana. Si el peso promedio de llenado es de 16 onzas y la desviación estándar es de 0.25 onzas, se puede aplicar la regla empírica para sacar las siguientes conclusiones:

- aproximadamente 68% de los envases llenos tienen entre 15.75 y 16.25 onzas (esto es, a menos de una desviación estándar del promedio).
- aproximadamente 95% de los envases llenos tienen entre 15.50 y 16.50 onzas (esto es, a menos de dos desviaciones estándar del promedio).
- casi todos los envases llenos tienen entre 15.25 y 16.75 onzas (esto es, a menos de tres desviaciones estándar del promedio).

El estudio y conocimiento de una adecuada recolección, análisis y procesamiento de datos, constituyen una plataforma básica para profundizar en otros requerimientos estadísticos de orden superior.

La presentación gráfica de datos es muy útil para visualizar su comportamiento y distribución y también para determinar la posición de las medidas de tendencia central y la magnitud de su dispersión.

Por lo tanto el dominio que se alcance para calcular estas medidas de datos no agrupados y datos agrupados en clases, así como su correcta interpretación, ayudarán a tomar mejores decisiones en cualquier ámbito personal, social o profesional.

# RESUMEN

La estadística descriptiva es una herramienta matemática que conjuga una serie de indicadores numéricos y gráficos, así como los procedimientos con que éstos se construyen, para descubrir y describir, en forma abreviada y a través de símbolos precisos, la estructura inmersa en el conjunto de datos. Se dice que se conoce la estructura cuando se sabe:

**a)** Lo que ocurre en ciertos puntos específicos de la distribución de los datos.

**b)** En qué medida los valores de las observaciones difieren.

**c)** La forma general de la distribución de los datos.

La confiabilidad y relevancia de los indicadores depende en buena medida de una adecuada definición del objeto bajo estudio y de la medición correcta de sus atributos. De hecho, se puede decir que según la manera en que se midan los atributos dependerá el tipo de indicador que se puede construir.

# BIBLIOGRAFÍA

## SUGERIDA

| Autor | Capítulo | Páginas |
|---|---|---|
| Berenson, Levine, Krehbiel (2001) | 1. Introducción y recopilación de datos.<br>Secciones:<br>1.7 Tipos de datos. | 9-11 |
| | 2.1 Organización de datos numéricos. | 40-45 |
| | 2. Presentación de datos en tablas y gráficas.<br>Secciones:<br>2.2 Tablas y gráficas para datos numéricos. | 45-57 |
| | 2.3 Tablas y gráficas para datos categóricos. | 57-65 |
| | 2.4 Tablas y gráficas para datos bivariados. | 65-70 |
| | 2.5 Excelencia gráfica. | 70-78 |
| | 3. Resumen y descripción de datos numéricos.<br>Secciones:<br>3.1 Exploración de datos numéricos y sus propiedades. | 102-103 |
| | 3.2 Medidas de tendencia central, variación y forma. | 103-127 |
| | 3.4 Obtención de medidas descriptivas de resumen a partir de una población. | 133-139 |

| | | |
|---|---|---|
| Levin y Rubin (2004) | 2. Agrupación y presentación de datos para expresar significados: tablas y gráficas.<br>Secciones:<br>2.1 ¿Cómo podemos ordenar los datos? | 8-11 |
| | 2.3 Ordenamiento de datos en arreglos de datos y distribuciones de frecuencias. | 12-20 |
| | 2.4 Construcción de una distribución de frecuencias. | 20-9 |
| | 2.5 Representación gráfica de distribuciones de frecuencias. | 29-41 |
| | 3. Medidas de tendencia central y dispersión en distribuciones de frecuencias.<br>Secciones:<br>3.2 Representación gráfica de distribuciones de frecuencias. | 29-41 |
| | 3.5 Una cuarta medida de tendencia central: la mediana. | 77-83 |
| | 3.6 Una medida final de tendencia central: la moda. | 84-89 |
| | 3.7 Dispersión: ¿por qué es importante? | 89-91 |
| | 3.8 Rangos: medidas de dispersión útiles. | 91-95 |
| | 3.9 Dispersión: medidas de dispersión promedio. | 96-107 |
| | 3.10 Dispersión relativa: el coeficiente de variación. | 107-112 |
| Lind, Marchal, Wathen (2008) | 1. ¿Qué es la estadística?<br>Sección:<br>Tipos de variables. | 8-9 |
| | Niveles de medición. | 9-13 |

| | | |
|---|---|---|
| | 2. Descripción de datos: tablas de frecuencias, distribuciones de frecuencias y su representación.<br>Secciones:<br>Construcción de una tabla de frecuencias. | 22-27 |
| | Construcción de distribuciones de frecuencias: datos cuantitativos. | 28-32 |
| | Representación gráfica de una distribución de frecuencias. | 36-39 |
| | 3. Descripción de datos: medidas numéricas<br>Secciones:<br>La media poblacional. | 57-58 |
| | Media de una muestra. | 58-59 |
| | Propiedades de la media aritmética. | 59-61 |
| | Mediana. | 62-64 |
| | Moda. | 64-65 |
| | Posiciones relativas de la media, la mediana y la moda. | 67-68 |
| | ¿Por qué estudiar la dispersión? | 71-73 |
| | Medidas de dispersión. | 73-80 |
| | Interpretación y usos de la desviación estándar. | 81-83 |
| | La media y la desviación estándar de datos agrupados. | 84-87 |

Berenson, Mark L., David M. Levine, y Timothy C Krehbiel. (2001), *Estadística para administración.* 2ª edición, México: Prentice Hall, 734 pp.

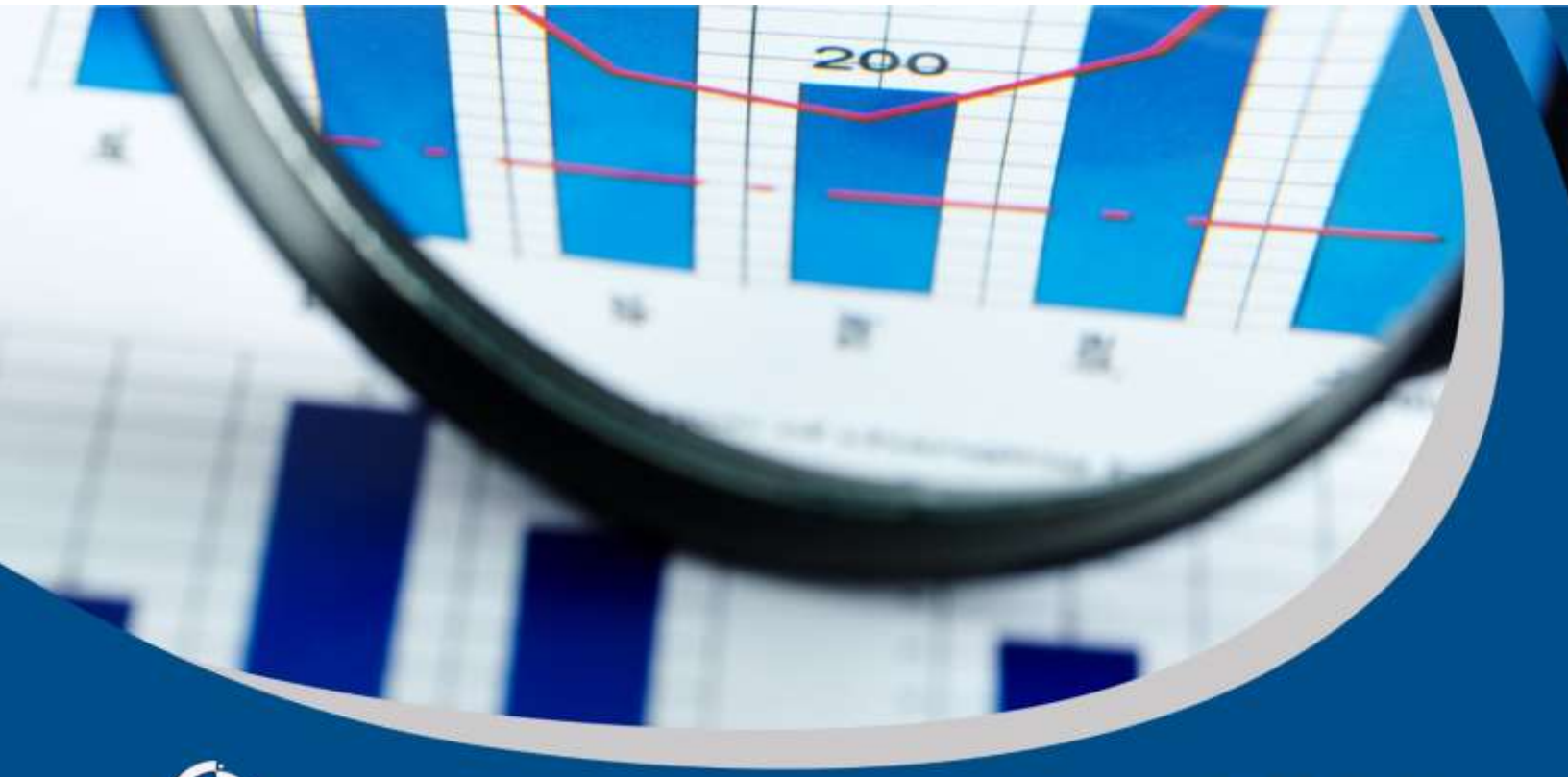
Levin, Richard I. y David S. Rubin. (2004), *Estadística para administración y economía*. 7ª edición, México: Pearson Educación Prentice Hall, 826 pp.

Lind, Douglas A., Marchal, William G., Wathen, Samuel, A. (2008), *Estadística aplicada a los negocios y la economía*. 13ª edición, México: McGraw-Hill Interamericana, 859 pp.

Licenciatura: Informática

# UNIDAD 2

# Teoría de la probabilidad




SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

# OBJETIVO PARTICULAR

El alumno identificará los diferentes enfoques de probabilidad y su interpretación para la toma de decisiones.

# TEMARIO DETALLADO

## (12 horas)

**2. Teoría de la probabilidad**

- 2.1. Interpretaciones de la probabilidad
  - 2.1.1. Teórica o clásica
  - 2.1.2. La probabilidad como frecuencia relativa
  - 2.1.3. Interpretación subjetiva de la probabilidad
- 2.2. Espacio muestral y eventos
- 2.3. Los axiomas de la probabilidad
- 2.4. La regla de la suma de probabilidades
- 2.5. Tablas de contingencias y probabilidad condicional
- 2.6. Independencia estadística
- 2.7. La regla de multiplicación de probabilidades
- 2.8. Teorema de Bayes

# INTRODUCCIÓN

Dicen que solamente existen dos cosas en la vida que con toda seguridad habremos de enfrentar: los impuestos y la muerte. Todos los demás eventos pueden o no sucedernos; es decir, tenemos un cierto nivel de duda sobre su ocurrencia. Para tratar de cuantificar el nivel de duda (o de certeza) que tenemos de que ocurra un determinado fenómeno se creó la teoría de la probabilidad. En esta unidad nos concentraremos en lo que se conoce como probabilidad básica.

En ella no existen muchas fórmulas a las cuales recurrir, aunque sí existen desde luego algunas expresiones algebraicas. La mayor parte de los problemas se resuelven mediante la aplicación de un reducido conjunto de principios básicos y de algo de ingenio. Para ello es indispensable entender claramente el problema en sí, por lo que la lectura cuidadosa y crítica es indispensable.

A reserva de ahondar más en el tema, podemos adelantar que **la probabilidad siempre es un número entre cero y uno**. Mientras más probable sea la ocurrencia de un evento más se acercará a uno; mientras más improbable sea, se acercará más a cero. Las razones de ello se explican en la siguiente sección de este tema.

Es necesario, por último, advertir sobre la presentación de datos. Al ser la probabilidad un número entre cero y uno *es frecuente expresarla en porcentaje.* A la mayoría se nos facilita más la comprensión cuando la cantidad está expresada de esta última manera. Si decimos, por ejemplo, que la probabilidad de que llueva hoy es de 10%, damos la misma información que si decimos que la probabilidad de que llueva hoy es de 0.10. Ambas maneras de presentar la información son equivalentes.

# 2.1. Interpretaciones de la probabilidad

Para determinar la probabilidad de un suceso podemos tomar dos enfoques. El primero de ellos se denomina objetivo y tiene, a su vez, dos enfoques, que a continuación se detallan.

## 2.1.1. Teórica o clásica

En el enfoque **teórico**, **clásico** o ***a priori*** (es decir, antes de que ocurran las cosas) se parte de la base de que se conocen todos los resultados posibles y a cada uno de ellos se les asigna una probabilidad de manera directa sin hacer experimento o medición alguna.

Frecuentemente decimos que al arrojar una moneda existen 50% de probabilidades de que salga águila y 50% de probabilidades de que salga sol, basándonos en el hecho de que la moneda tiene dos caras y que ambas tienen las mismas probabilidades de salir. Igual camino seguimos al asignar a cada cara de un dado la probabilidad de un sexto de salir. Razonamos que el dado tiene seis caras y por tanto, si el dado es legal, cada una de ellas tiene las mismas probabilidades.

## 2.1.2. La probabilidad como frecuencia relativa

También se le conoce como enfoque ***a posteriori*** (es decir, a la luz de lo ocurrido) y al igual que el enfoque anterior es un paradigma objetivo.

Para asignarle probabilidad a un suceso se experimenta antes y a partir de los resultados se determinan las frecuencias con que ocurren los diversos resultados. En el caso de la moneda, este enfoque nos recomendaría hacer un número muy grande de “volados”, por ejemplo diez mil, y con base en ellos definir la probabilidad de una y otra cara. Si decimos, por ejemplo, que la probabilidad de que salga águila es de 4888/10000, damos a entender que lanzamos la moneda diez mil veces y que en 4888 ocasiones el resultado fue águila. Estamos entonces aplicando la probabilidad *a posteriori*.

En ejemplos menos triviales, las compañías de seguros desarrollan tablas de mortalidad de las personas para diferentes edades y circunstancias con base en sus experiencias. Ese es un caso de aplicación del enfoque *a posteriori*.

## 2.1.3. Interpretación subjetiva de la probabilidad

La **probabilidad subjetiva** es una cuestión de opinión. Dos personas, por ejemplo, pueden asignar diferentes probabilidades a un mismo evento, aun cuando tengan la misma información. Tal **diversidad de opiniones** se puede ver en las proyecciones económicas que hacen los asesores en inversiones y los economistas para los años venideros.

Aunque muchos de estos individuos trabajan con los mismos datos, ellos se forman distintas opiniones acerca de las condiciones económicas más probables. Tales proyecciones son inherentemente subjetivas.

También se presenta cuando no existen antecedentes para determinarla (como en el caso de las tablas actuariales de las compañías de seguros) ni una base lógica para fijarla *a priori*.

Si pensamos, por ejemplo, en la final del campeonato mundial de fútbol del 2002, en la que se enfrentaron Brasil y Alemania, vemos que no había historia previa de enfrentamientos entre los dos equipos y había tantos factores en juego que difícilmente se podía dar una probabilidad sobre las bases que anteriormente llamamos objetivas; por lo mismo, se debe recurrir al juicio de las personas para definir las probabilidades. A esta manera de fijar probabilidades se le llama, por este hecho, probabilidad subjetiva.

## 2.2. Espacio muestral y eventos

Para trabajar con comodidad la probabilidad, vale la pena expresar algunos conceptos básicos que necesitaremos para el desarrollo del tema.

Conceptos estadísticos

**Experimento:** es aquel proceso que da lugar a una medición o a una observación.

**Experimento aleatorio:** es aquel experimento cuyo resultado es producto de la suerte o del azar. Por ejemplo, el experimento de arrojar un dado.

**Evento:** el resultado de un experimento.

De estos tres conceptos podemos desprender un cuarto, el concepto de **evento aleatorio** que no es sino el resultado de un experimento aleatorio. Por ejemplo, si el experimento es arrojar un dado, por el sólo hecho de que no podemos anticipar qué cara mostrará éste al detenerse, podemos decir que el experimento es aleatorio. Uno de los resultados posibles es que salga un número par. Tal resultado es un evento aleatorio.

Para referirnos a los eventos aleatorios usaremos letras mayúsculas. De este modo podemos decir que:

A es el evento de que al arrojar un dado salga un número non.

B es el evento de que al arrojar un dado salga un número par.

Como es claro, podemos definir varios eventos aleatorios respecto del mismo experimento. Algunos de ellos tendrían la característica de que encierran a su vez varias posibilidades (en el evento A quedan incluidas las posibilidades “que salga 1”, “que salga 3” o “que salga 5”).

En este contexto, conviene distinguir eventos simples de eventos compuestos:

Los **eventos simples** son aquéllos que ya no pueden descomponerse en otros más sencillos. Otra manera de denominar a los eventos simples es la de “puntos muestrales”. Esta denominación es útil cuando se trata de representar gráficamente los problemas de probabilidad pues cada evento simple (punto muestral) se representa efectivamente como un punto.

Los **eventos compuestos** incluyen varias posibilidades por lo que pueden descomponerse en eventos sencillos.

Por ejemplo, el evento A mencionado anteriormente se puede descomponer en los siguientes eventos:

E1: el evento de que al arrojar un dado salga un uno.

E2: el evento de que al arrojar un dado salga un tres.

E3: el evento de que al arrojar un dado salga un cinco.

A su vez, E1, E2 y E3 son eventos sencillos.

Ante la interrogante de qué eventos consideraremos en un experimento aleatorio dado debemos contestar que esto depende de la perspectiva que tengamos respecto del experimento aleatorio. Si estamos jugando a los dados y las apuestas sólo consideran el obtener un número par o un número impar o non, entonces los únicos resultados que nos interesarán serán precisamente esos dos: obtener número par o número impar.

Con esto damos lugar a un concepto adicional básico.

| **Espacio muestral** |
| --- |
| • Es el conjunto de todos los resultados posibles, en función de nuestra perspectiva del experimento aleatorio. También se le conoce como evento universo. |

En suma, ante un experimento aleatorio cualquiera tenemos varias alternativas para definir eventos cuya probabilidad pueda sernos de interés.

Por ejemplo, si tenemos una colectividad de 47 estudiantes egresados, entre Contadores, Administradores e Informáticos de ambos sexos, y de esa colectividad seleccionamos al azar a una persona, puede ser que nos interesen las probabilidades de los siguientes eventos:

**a)** Que la persona seleccionada haya estudiado contaduría.
**b)** Que la persona seleccionada haya estudiado administración o contaduría.
**c)** Que la persona seleccionada no haya estudiado administración.
**d)** Que la persona seleccionada sea mujer y haya estudiado informática.
**e)** Que la persona seleccionada sea hombre pero que no haya estudiado administración.

Como puede verse, en los incisos anteriores no sólo estamos manejando diversos eventos sino que además estamos incorporando relaciones entre ellos. Tales relaciones se pueden establecer de manera más eficiente recurriendo a la estructura formal de la teoría de conjuntos, esto es, incorporando los diagramas de Venn-Euler, la terminología de conjuntos, así como las operaciones que has aprendido a realizar

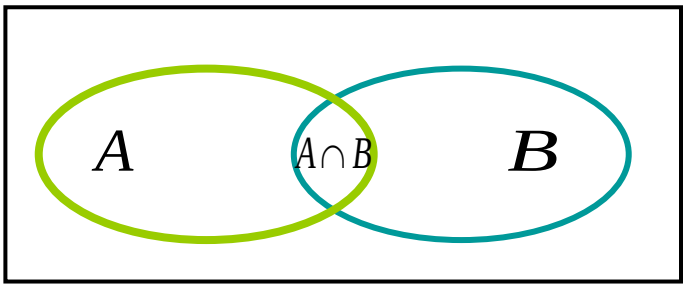
con ellos en cursos anteriores —como la unión, la intersección, el complemento, la diferencia, entre otras— son por entero aplicables al caso de los eventos, en el contexto de la teoría de la probabilidad.

Estos elementos junto con algunas definiciones que se detallan a continuación nos permitirán trabajar adecuadamente los problemas de probabilidad que enfrentaremos.

Si definimos a los eventos A y B como resultados de un experimento aleatorio y recordamos que todos los **eventos posibles** (el conjunto universal) constituyen el **espacio muestral** y representamos éste como S, tenemos que la unión de A con B es un evento que contiene todos los puntos muestrales que pertenecen al evento A y/o que pertenecen al evento B. Podemos hacer uso de la notación de conjuntos para escribir: $A \cup B$.

La probabilidad de $A \cup B$ es la probabilidad de que suceda el evento A o de que suceda el evento B o de que ambos sucedan conjuntamente. Por otra parte, tenemos que la intersección de A y B es la situación en que ambos, A y B, suceden conjuntamente, esto es en forma simultánea. La intersección se denota en la simbología de conjuntos como $A \cap B$.

A manera de resumen en la siguiente tabla te mostramos cuatro operaciones que




Eventos simultáneos.

serán muy útiles para manejar eventos aleatorios y su equivalencia con operaciones lógicas.

| Operación Lógica | Operación en conjuntos |
|---|---|
| o | Unión (U) |
| y | Intersección (∩) |
| no | Complemento (‘ )<br>Diferencia (–) |

Si en nuestro ejemplo de los egresados incorporamos estas operaciones y llamamos C al evento “egresado de contaduría”, A al evento “egresado de administración”, I al evento “egresado de informática”, M al evento “mujer” y H al evento “hombre”, tendríamos que nuestro interés es conocer las siguientes probabilidades:

**a)** Probabilidad de C
**b)** Probabilidad de A U C
**c)** Probabilidad de $A^c$
**d)** Probabilidad de M ∩ I
**e)** Probabilidad de H – $A^c$

Si además, adoptamos la convención de usar la letra P para no escribir todo el texto “probabilidad de“, y encerramos entre paréntesis el evento de interés, nuestras preguntas quedarían de la siguiente manera:

**a)** P(C)
**b)** P(A U C)
**c)** P($A^c$)
**d)** P(M ∩ I)
**e)** P(H – $A^c$)

Esta es la forma en que manejaremos relaciones entre eventos y denotaremos probabilidades.

# 2.3. Los axiomas de la probabilidad

Los elementos hasta ahora expuestos nos permiten dar ya una definición más formal de probabilidad en el contexto frecuentista:

Sea A un evento cualquiera; N el número de veces que repetimos un experimento en el que puede ocurrir el evento A; $n_A$ el número de veces que efectivamente se presenta el evento A; y P(A) la probabilidad de que se presente el evento A.

Entonces tenemos que **P(A)=** $\lim_{N \to \infty} \left( \frac{n_A}{N} \right)$

Es decir, que la probabilidad de que ocurra el evento A, resulta de dividir el número de veces que A efectivamente apareció entre el número de veces que se intentó el experimento. (La expresión N → ∞ se lee «N tiende a infinito» y quiere decir que el experimento se intentó muchas veces).

Podemos ver que el menor valor que puede tener P(A) es de cero, en el caso de que en todos los experimentos intentados A no apareciera ni una sola vez. El mayor valor que puede tener P(A) es de uno, en el caso de que en todos los experimentos intentados el evento en cuestión apareciera todas las veces, pues en ese caso $n_A$ sería igual a N y todo número dividido entre sí mismo es igual a 1.

En todos los demás casos, la probabilidad de ocurrencia estará entre estos dos números extremos y por eso podemos decir que la **probabilidad de ocurrencia** de

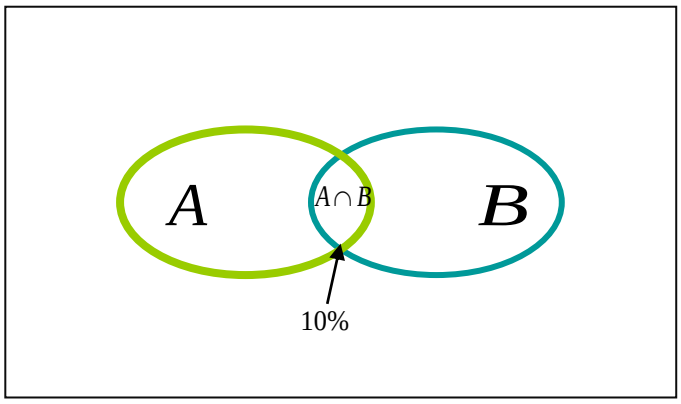
cualquier evento estará entre cero y uno. Ésta es la justificación de la afirmación análoga que se realizó al principio de la unidad y también la justificación de la afirmación que se hace frecuentemente de que la probabilidad se expresa como la frecuencia relativa de un evento; es decir, relativa al total de experimentos que se intentaron.

Consideremos el siguiente ejemplo.

**Ejemplo 1**. En una investigación de mercado se encontró que entre los integrantes de un club, 30% de los hombres usan loción para después de afeitarse, en tanto que 40% de ellos utiliza desodorante y 10% utiliza ambos productos. Si elegimos al azar a un varón de ese club, ¿qué probabilidades existen de que utilice desodorante o de que use loción para después de afeitarse?

Solución:

> Es evidente que la probabilidad que buscamos es un número positivo ya que de entre los integrantes del club sí hay varones que usan desodorante además de que también hay varones que usan loción. Es evidente además que la probabilidad que buscamos será menor a uno porque no todos usan loción y no todos usan desodorante.
>
> Por otro lado, si hacemos que A represente el evento «El sujeto usa loción para después de afeitarse», y que B represente el evento «El sujeto usa desodorante», podemos intentar una representación gráfica empleando diagramas de Venn-Euler como sigue.
>
> Cuando nos preguntan por la probabilidad de que la persona seleccionada al azar utilice desodorante o de que use loción para después de afeitar, sabemos que tal pregunta en lenguaje probabilístico se transforma en:




**P(AUB)**

Intrínsecamente la pregunta se refiere a aquellos elementos que se encuentran en A o se encuentran en B, esto es, en el interior del óvalo verde o en el interior del óvalo azul. De acuerdo con los datos, 30% de los casos se encuentran en A y 40% en B, por lo que al sumar tendríamos que aparentemente hay 70% de integrantes del club que se encuentran en la unión de ambos eventos, sólo que de ese 70% hay un 10% que es común, precisamente el porcentaje de casos que se encuentra en la intersección. Este 10% ya ha sido contado una vez al considerar el porcentaje de casos en A y fue incluido otra vez al considerar el porcentaje de casos en B, de manera que se le ha contado dos veces. Por lo tanto, para determinar el número de casos que están en la unión de A con B, debemos efectivamente considerar el 30% que está en A, el 40% que está en B, pero además debemos descontar el 10% que está en la intersección para que los elementos que están en dicha intersección sean contados sólo una vez.

De esta manera, P (AUB) = 30% + 40% – 10%.

P (AUB) = 60%

Esto quiere decir que existe 60% de probabilidades de que un socio de este club elegido al azar use alguno de los dos productos.

Las situaciones que hemos discutido dentro de este tema ilustran tres postulados básicos de la probabilidad, a los que se conoce como **Axiomas de probabilidad**, lo que en lenguaje matemático significa que son proposiciones que por su carácter evidente no requieren demostración. Constituyen, por decirlo de alguna manera, "las reglas del juego", sin importar si estamos trabajando una probabilidad subjetiva o empírica, o si seguimos los postulados de la probabilidad clásica.

Estos axiomas, que constituyen el cimiento de la teoría moderna de probabilidades y fueron propuestos por el matemático ruso Kolmogorov, se expresan de manera formal en los siguientes términos:

**1)** Para todo evento A, $P(A) \geq 0$
**2)** Si Ω representa el evento universo, entonces $P(\Omega) = 1$
**3)** Dados dos eventos A y B, ocurre que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Claramente, el primer axioma nos indica que no hay probabilidades negativas y, el segundo, que ningún evento tiene una probabilidad mayor a uno.

A partir de ellos se tienen otros resultados de suyo importantes, tales como:

**a)** $P(\varphi) = 0$, donde φ representa el conjunto vacío.
b) $P(A^c) = 1 - P(A)$

En el segundo de estos resultados estamos haciendo referencia a **eventos complementarios.** Si Ω es el evento universo, entonces para todo evento A existe un evento complemento constituido por todos aquellos resultados del espacio muestral que no están en A, con la propiedad de que $A \cup A^c = \Omega$, por lo que $P(A \cup A^c) = P(\Omega)$, de modo que $P(A \cup A^c) = 1$.

En consecuencia, de acuerdo con el axioma (3),

$$P(A \cup A^c) = P(A) + P(A^c) - P(A \cap A^c),$$

$$\rightarrow 1 = P(A) + P(A^c) - P(A \cap A^c),$$

Sin embargo, $P(A \cap A^c) = P(\phi)$ y de acuerdo con el resultado (a), esta probabilidad es cero. Por lo tanto,

$$1 = P(A) + P(A^c),$$

de donde al despejar queda:

$$P(A^c) = 1 - P(A)$$

**Ejemplo 2.** Sea el experimento aleatorio que consiste en arrojar dos dados y sea Ω el espacio muestral que contiene todos los resultados posibles de sumar los puntos obtenidos. Se definen además los eventos A como el hecho de que el tiro sume menos de cuatro y B como el hecho de que la suma sea número par. Se desea determinar las probabilidades siguientes:

- **a)** $P(A^c)$
- **b)** $P(B)$
- **c)** $P(A \cup B)$

Solución:

Claramente,

$\Omega = \{2,3,4,5,6,7,8,9,10,11,12\}$,

$A = \{2,3\}$;

$B = \{2,4,6,8,10,12\}$.

Entonces,

**a)** De acuerdo con lo anterior, $A^c$= {4,5,6,7,8,9,10,11,12}, de donde se sigue que $P(A^c) = 9/11$. Alternativamente, $P(A^c) = 1 – P(A)$, donde $P(A) = 2/11$, por lo que $P(A^c) = (11–2)/11 = 9/11$, lo que confirma el resultado.

**b)**Es inmediato que P(B) = 6/11

**c)**Aplicando el axioma 3, se tiene que:

$$P (AUB) = P (A) + P (B) – P (A \cap B),$$

donde $A \cap B = \{2\}$ por lo que $P(A \cap B\} = 1/11$.

Finalmente,

$$P(A \cup B) = 2/11 + 6/11 – 1/11$$

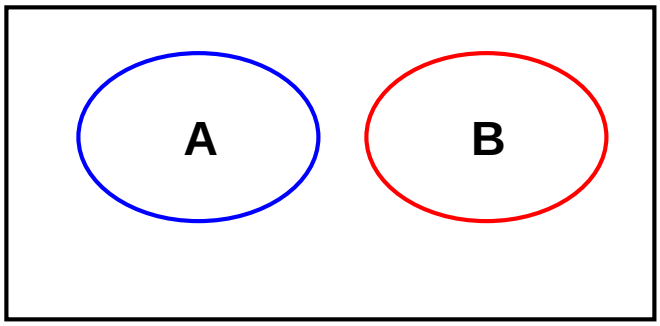
$$P(A \cup B) = 7/11$$

# 2.4. La regla de la suma de probabilidades

En el tema anterior se introdujo el axioma tres de probabilidad aplicable a cualquier pareja de eventos probabilísticos. Ahora, consideraremos un caso particular. Para ello, incorporamos primero un concepto adicional.

**Eventos mutuamente excluyentes**. Son aquellos eventos que si se produce uno de ellos, no puede producirse el otro. Dicho en el lenguaje

de los conjuntos, podemos afirmar que si dos eventos son mutuamente excluyentes, la intersección de ellos está vacía. En terminología de conjuntos también se dice que estos eventos son disjuntos.




**Eventos mutuamente excluyentes.**

**Ejemplo 1**: Sea Ω el espacio de resultados que resulta de considerar la suma de los puntos que se obtienen al arrojar dos dados.

Sea A: La suma de puntos de los dos dados es de 12.

Sea B: Aparece por lo menos un “uno” en los dados arrojados.

Se desea determinar las siguientes probabilidades:

**a)** $P(A \cap B)$

**b)** P(A U B)

Solución:

Vemos que es imposible que ocurran A y B simultáneamente, pues para que la suma de los puntos sea doce debe ocurrir que en ambos dados salga ”seis”, pero si uno de los dos dados tiene “uno” como resultado, la suma máxima que se puede lograr es de “siete”. Los eventos son mutuamente excluyentes y, por lo tanto, $P(A \cap B) = 0$.

Al aplicar el axioma 3 tenemos,

$P(AUB) = P(A) + P(B) – P(A \cap B)$,

$P(AUB) = 1/36 + 11/36 – 0$

$P(AUB) = 12/36$

Como puede verse, el impacto de que A y B sean mutuamente excluyentes es tal que para determinar la probabilidad de la unión de dos eventos sólo debemos sumar las probabilidades de cada evento individualmente considerado.

En el caso en que A y B sean mutuamente excluyentes, esto es, cuando su intersección es vacía, la probabilidad de la unión de dos eventos es la suma de las probabilidades de los eventos tomados individualmente.

$$P(A \cup B) = P(A) + P(B) \qquad \text{si } A \cap B = \varphi$$

Si tenemos varios eventos mutuamente excluyentes en el espacio de eventos Ω y queremos saber cuál es la probabilidad de que ocurra cualquiera de ellos, la pregunta que estaríamos planteando se refiere a la probabilidad de la unión de los mismos. Al ser eventos mutuamente excluyentes, la intersección está vacía y la probabilidad de

ocurrencia es simplemente la suma o adición de las probabilidades individuales; es por ello que a esta regla se la conoce como **regla de la adición**.

El siguiente ejemplo nos ayudará a dejar en claro estos conceptos.

**Ejemplo 2:** En un club deportivo, 20% de los socios pertenece al equipo de natación y 10% al equipo de waterpolo. Ningún socio pertenece a ambos equipos simultáneamente. Diga cuál es la probabilidad, si elegimos al azar un socio del club, de que sea integrante de alguno de los dos equipos.

Solución:

> El cálculo de probabilidades aparece a continuación. El estudiante debe tener en mente que, dado que ningún socio pertenece a los dos equipos simultáneamente, la intersección está vacía y por lo mismo su probabilidad es cero.

$$P(A \cup B) = 0.20 + 0.10 – 0.0 = 0.30$$

# 2.5. Tablas de contingencias y probabilidad condicional

En muchas circunstancias encontramos que la probabilidad de ocurrencia de un evento se ve modificada por la ocurrencia de otro evento. Por ejemplo, la probabilidad de pasar un examen depende del hecho de que el estudiante haya estudiado para el mismo.

En este tema nos abocaremos a analizar este tipo de situaciones. Para ello es conveniente introducir dos conceptos preliminares.

**Probabilidad simple (marginal)**

En un experimento cualquiera, la probabilidad simple de un evento es la que tiene éste, sin considerar las conexiones que pueda tener con otros eventos. También se le llama **probabilidad marginal**.

Repasemos a continuación el procedimiento para calcular la probabilidad simple o marginal de un evento.

1. Definimos el experimento.
2. Hacemos la lista de todos los eventos simples asociados con el experimento que definió (es decir, haga la lista de todos los puntos muestrales).
3. Asignamos probabilidades a cada uno de los puntos muestrales. La suma total de las probabilidades de todos los puntos muestrales debe ser igual a la unidad.
4. Definimos el evento que le interesa como un conjunto de puntos muestrales.
5. Encontramos la probabilidad del evento que le interesa sumando la probabilidad de los puntos muestrales que lo componen.

A continuación, se dan varios ejemplos que nos permitirán comprender mejor este procedimiento.

**Ejemplo 1**.

1. El experimento consiste en arrojar un dado normal y bien balanceado de seis caras.

2. Todos los resultados posibles (los eventos simples o puntos muestrales) se listan a continuación:

   E1: que salga un uno
   E2: que salga un dos
   E3: que salga un tres
   E4: que salga un cuatro
   E5: que salga un cinco
   E6: que salga un seis

3. Para asignar probabilidades a cada evento, es razonable darle la misma probabilidad a cada evento simple; si hay seis resultados posibles, también resulta razonable darle 1/6 a cada uno.

4. A continuación, definimos tres eventos de interés y los definimos como conjuntos de puntos muestrales.

   **a.** Evento A: que salga un número menor a cuatro. Se compone de los eventos E1, E2 y E3.
   **b.** Evento B: que salga un número par. Se compone de los eventos E2, E4, E6.
   **c.** Evento C: que salga un número mayor que seis. Ningún evento lo compone.

5. Calculamos las probabilidades solicitadas:

- La probabilidad de A es la suma de las probabilidades de E1, E2 y E3: 1/6+1/6+1/6 = 3/6 = 1/2.
- La probabilidad de B es la suma de las probabilidades de E2, E4, E6: 1/6+1/6+1/6 = 3/6 = 1/2.
- La probabilidad de C es de cero, pues no existe ningún evento que lo componga.

**Ejemplo 2**. El comité directivo de la sociedad de padres de familia de una escuela primaria está compuesto por cinco personas: tres mujeres y dos hombres. Se van a elegir al azar dos miembros del comité para solicitar al delegado que ponga una patrulla a vigilar durante la salida de los niños. ¿Cuál es la probabilidad de que el comité esté compuesto por un hombre y una mujer?

Solución:

El experimento es elegir al azar dos personas de las cuales tres son mujeres y dos son hombres.

Para listar todos los eventos simples simbolizaremos a las mujeres con una M y los hombres con una H. Así, el comité directivo está compuesto por: M1, M2, M3, H1 y H2, donde M1 es la primera mujer, M2 la segunda, H1 el primer hombre y así sucesivamente.

Los eventos simples factibles se listan a continuación:

M1M2; M1M3; M1H1; M1H2

M2M3; M2H1; M2H2;

M3H1; M3H2;

H1H2.

Vemos que pueden darse 10 pares distintos. Si cada par es elegido al azar, es razonable suponer que todos ellos tienen la misma probabilidad de ser

seleccionados, por ello podemos afirmar que cada par tiene una probabilidad de 1/10 de ser seleccionado.

Por otro lado, las parejas que están constituidas por un hombre y una mujer son: M1H1 M1H2; M2H1; M2H2; M3H1 y M3H2; es decir, seis de los diez pares posibles.

La probabilidad de nuestro evento de interés es entonces, de seis veces un décimo o 6/10. Expresada en porcentaje, esta probabilidad será de 60%.

**Ejemplo 3**. Una tienda de electrodomésticos va a recibir un embarque de seis refrigeradores, de los cuales dos están descompuestos. El dueño de la tienda someterá a prueba dos refrigeradores al recibir el embarque y solamente lo aceptará si ninguno de ellos presenta fallas. Nos interesa saber cuál es la probabilidad de que acepte el embarque.

Solución:

El experimento es elegir dos refrigeradores al azar para ver si funcionan o no.

Si llamamos B al refrigerador que trabaja bien y D al descompuesto, podemos listar a todos los refrigeradores del embarque de la siguiente manera:

B1, B2, B3, B4, D1, D2.

A continuación listamos todos los eventos posibles (es decir, todos los pares diferentes que se pueden elegir). Los eventos simples de interés (aquellos en que los dos refrigeradores están en buen estado) están resaltados.

B1B2; B1B3; B1B4**; B1D1; B1D2;**

B2B3; B2B4; **B2D1; B2D2;**

B3B4**; B3D1; B3D2;**

**B4D1; B4D2**

**D1D2**

Vemos que existen quince eventos posibles, de los cuales en seis se presenta el caso de que ambos refrigeradores estén en buen estado. Si, como en lo ejemplos anteriores, asignamos una probabilidad igual a todos los eventos simples (en este caso 1/15), tendremos que la probabilidad de aceptar el embarque es 6/15.

**Probabilidad conjunta**

En muchas ocasiones estaremos enfrentando problemas en los que nuestros eventos de interés estarán definidos por la ocurrencia de dos o más eventos simples.

Tomemos el caso del siguiente ejemplo.

**Ejemplo 4**. Consideremos el caso de una pareja que tiene dos hijos, situación respecto de la cual definimos los siguientes eventos de interés:

Evento A: La pareja tiene por lo menos un varón.

Evento B: La pareja tiene por lo menos una niña.

Nuestros eventos de interés se pueden expresar de la siguiente manera:

Evento A: Ocurre si se tiene varón y varón, varón y mujer en ese orden, o mujer y varón en ese orden.

Evento B: Ocurre si se tiene mujer y mujer, varón y mujer en ese orden o mujer y varón en ese orden.

Como puede verse, para que ocurra el evento A deben ocurrir dos cosas simultáneamente. Bien sea:

Varón **y** varón, o

Varón **y** mujer, o

Mujer **y** varón.

Si definimos los eventos simples V: varón y M: mujer, tendríamos que cada una de las posibilidades que se tienen para que ocurra el evento A implica la ocurrencia de dos o más eventos simples.

Algo similar puede decirse en relación al evento B.

> Cuando los eventos de interés implican la ocurrencia de dos o más eventos simples de manera simultánea, decimos que estamos en presencia de una **probabilidad conjunta**.

El lector puede confirmar que en el ejemplo 3 también estábamos en presencia de probabilidades conjuntas, aunque por la perspectiva que se adoptó aparecían como simples.

$$P\left(B / A^{'}\right) = \frac{0.10}{0.40} = 0.25 = 25\%$$

**Probabilidad condicional**

Dados dos eventos podemos preguntarnos por la probabilidad de uno de ellos bajo el supuesto de que el otro ya ocurrió. Al inicio de este tema, por ejemplo, se planteaba la situación respecto de la probabilidad de pasar un examen si el estudiante realmente estudió para dicho examen. Este tipo de situaciones dan lugar a la **probabilidad condicional**.

La probabilidad condicional de que ocurra el evento B dado que otro evento A ya ocurrió es:

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Es decir, la probabilidad de B dado que A ya ocurrió es igual a la probabilidad de que ocurran ambos eventos simultáneamente (la probabilidad conjunta) dividido por la probabilidad de que ocurra A (la probabilidad marginal), que en este caso es el evento antecedente.

El siguiente ejemplo nos ayudará a clarificar estas ideas.

**Ejemplo 5**. Sea el evento A: Amanece nublado en la región X

De acuerdo con información meteorológica recopilada a lo largo de muchos días, se sabe que:

Amanece nublado y llueve el 40% de los días.

Amanece nublado y no llueve el 20% de los días.

Amanece despejado y llueve el 10% de los días.

Amanece despejado y no llueve el 30% de los días.

Dado lo anterior, la probabilidad de que llueva en la tarde, es la suma de las probabilidades de que llueva tanto si amaneció despejado como si amaneció nublado. Es decir, 40% más 10%, o sea, 50%. La probabilidad de que no llueva es su complemento, en este caso, también, 50%.

Deseamos averiguar lo siguiente.

a) La probabilidad de que llueva en la tarde dado que amaneció nublado.

b) La probabilidad de que llueva en la tarde dado que amaneció despejado.

Solución:

En el inciso “a” deseamos saber la probabilidad de B dado que A. Con la información que tenemos podemos sustituir directamente en la expresión para la probabilidad condicional.

La probabilidad condicional de que ocurra B dado que A ya ocurrió es:

$$P(B/A) = \frac{0.40}{0.60} = 0.667 = 66.7\%$$

Es decir, que la probabilidad de que llueva, dado que amaneció nublado, es de 67%. Podemos percatarnos a simple vista de que el hecho de que amanezca nublado efectivamente afecta la probabilidad de que llueva en la tarde. Recordemos que la probabilidad marginal de que llueva (sin tener antecedentes) es de 50%.

En el inciso (b) deseamos conoce la probabilidad de que llueva en la tarde dado que amaneció despejado, esto es, buscamos B dado que $A^c$ ya ocurrió. Como amanece nublado 60% de los días y despejado 40% de ellos, podemos sustituir en la fórmula.

$$P(B/A') = \frac{0.10}{0.40} = 0.25 = 25\%$$

Vemos que, si la probabilidad de que llueva cuando amaneció nublado es de 50% y la probabilidad de que llueva estando despejado es de sólo el 25%, el hecho de que amanezca despejado efectivamente afecta las probabilidades de que llueva.

**Tablas de contingencia**

Una **tabla de probabilidad conjunta** es aquella donde **se enumeran todos los eventos posibles** para una variable (u observación) **en columnas** y una segunda variable **en filas**. **El valor en cada celda es la probabilidad de ocurrencia conjunta**.

Su elaboración incluye formar una tabla de contingencia cuyos valores de cada celda se dividen entre el total de datos para obtener los valores de probabilidad correspondientes.

**Ejemplo 6**: *Se* obtiene una estadística de 300 personas, de acuerdo con su edad y sexo, que entraron en un almacén.

Tabla de contingencia de clientes

| Edad / sexo | Hombre | Mujer | Total |
|---|---|---|---|
| **Menor de 30 años** | 35 | 46 | 81 |
| **Entre 30 y 40 años** | 42 | 59 | 101 |
| **Mayor de 40 años** | 51 | 67 | 118 |
| **Total** | 128 | 172 | 300 |

Tabla de probabilidad conjunta

| Evento | Edad /sexo | Hombre $H$ | Mujer $M$ | Probabilidad marginal |
|---|---|---|---|---|
| $E_1$ | **Menor de 30 años** | 0.117 | 0.153 | 0.270 |
| $E_2$ | **Entre 30 y 40 años** | 0.140 | 0.197 | 0.337 |
| $E_3$ | **Mayor de 40 años** | 0.170 | 0.223 | 0.393 |
| **Probabilidad marginal** | | 0.427 | 0.573 | 1.000 |

Con esta información se desea obtener la probabilidad de que la siguiente persona que entre al almacén sea:

a) Un hombre menor de 30 años.

b) Una mujer.

c) Una persona de más de 40 años.

d) Habiendo entrado una mujer, que tenga entre 30 y 40 años.

e) Habiendo entrado un hombre, que tenga menos de 30 años.

f) Sea mujer dado que tiene entre 30 y 40 años.

Solución:

**a)** $P(E_1 \cap H) = 0.117 = 11.7\%$

**b)** $P(M) = 0.573 = 57.3\%$

**c)** $P(E_3) = .393 = 39.3\%$

**d)** $P(E_2 / M) = \dfrac{P(E_2 \cap M)}{P(M)} = \dfrac{0.197}{0.573} = 0.344 = 34.4\%$

**e)** $P(E_1 / H) = \dfrac{P(E_1 \cap H)}{P(H)} = \dfrac{0.117}{0.427} = 0.274 = 27.4\%$

**f)** $P(M / E_2) = \dfrac{P(E_2 \cap M)}{P(E_2)} = \dfrac{0.197}{0.337} = 0.585 = 58.5\%$

Las ideas que hemos presentado en esta sección nos permiten reformular la probabilidad marginal como la probabilidad incondicional de un evento particular simple, que consiste en una suma de probabilidades conjuntas. Si en el ejercicio anterior se desea calcular la probabilidad de que el siguiente cliente sea un hombre, esto podría hacerse a partir de probabilidades conjuntas, como sigue:

$$P(H) = P(H \cap E_1) + P(H \cap E_2) + P(H \cap E_3)$$

o sea:

$$P(H) = 0.117 + 0.140 + 0.170 = 0.427 = 42.7\%$$

## 2.6. Independencia estadística

Sean dos eventos A y B del espacio de eventos Ω; decimos que **A y B son independientes en sentido probabilístico si la probabilidad de que ocurra A no influye en la probabilidad de que ocurra B y, simultáneamente, la probabilidad de que ocurra B no influye en la probabilidad de que ocurra A**. En caso contrario decimos que los eventos son dependientes. Esto lo expresamos simbólicamente del siguiente modo:

Para considerar que A y B son independientes se deben cumplir las dos condiciones siguientes:

$$P(B/A) = P(B) \;\; y \;\; P(A/B) = P(A)$$

Es decir, el hecho de que ocurra un evento no modifica la probabilidad de que ocurra el otro, sin importar quien sea condición de quien.

Consideremos el siguiente ejemplo.

**Ejemplo 1.** Una tienda de departamentos ha solicitado a un despacho de consultoría que aplique un cuestionario para medir si su propaganda estática tenía impactos distintos según el grupo de edad del público. Como parte del estudio el despacho

entrevistó a 150 mujeres, a las cuáles se les preguntó si recordaban haber visto dicha propaganda. Los resultados se muestran a continuación:

| | Sí la recuerdan | No la recuerdan | Total |
|---|---|---|---|
| **Menores de 40 años** | **40** | **30** | **70** |
| **40 o más años de edad** | **20** | **60** | **80** |
| **Total** | **60** | **90** | **150** |

Sean los eventos siguientes:

S es el evento «Sí recuerda la propaganda»

N es el evento «No recuerda la propaganda»

J es el evento «Menor de 40 años de edad»

E es el evento «40 o más años de edad»

Se desea saber

**a)** Si los eventos S y J son independientes en sentido probabilístico
**b)** Si los eventos N y E son independientes en sentido probabilístico

Solución:

**a)** Para saber si los eventos son independientes basta calcular P(S) y P(S | J) y comparar.
De acuerdo con los datos de la tabla,

$$P(S) = 60/150,$$

Por su parte, para determinar el valor de P(S | J) observamos que al ser J la condición, podemos modificar el universo de resultados y restringirlo sólo a

aquéllos que cumplen con dicha condición. Así, el nuevo universo es de sólo 70 casos, de los cuales 40 recuerdan la propaganda. En consecuencia,

$$P(S \mid J) = 40/70$$

Es inmediato que las probabilidades no son iguales, por lo que podemos afirmar que S y J no son independientes.

**b)** Al igual que en el inciso anterior, para saber si los eventos son independientes basta calcular P(N) y P(N | E) y comparar.
De acuerdo con los datos de la tabla,

$$P(N) = 90/150,$$

Por su parte, para determinar el valor de P(N | E) observamos que al ser E la condición, podemos modificar el universo de resultados y restringirlo sólo a aquéllos que cumplen con dicha condición. Así, el nuevo universo es de sólo 80 casos, de los cuales 60 recuerdan la propaganda. En consecuencia,

$$P(N \mid E) = 60/80$$

Es inmediato que las probabilidades no son iguales, por lo que podemos afirmar que N y E no son independientes en sentido probabilístico.

El lector puede confirmar que las otras parejas de eventos tampoco son independientes.

# 2.7. La regla de multiplicación de probabilidades

Recordemos que en general,

$$P(B/A) = \frac{P(B \cap A)}{P(A)}$$

Si A y B son independientes probabilísticamente, P(B | A) = P(B), por lo que:

$$P(B) = \frac{P(B \cap A)}{P(A)}$$

De aquí se sigue que:

$$P(A \cap B) = P(A)P(B)$$

Podemos decir en consecuencia que **si dos eventos son estocásticamente independientes, entonces su probabilidad conjunta es igual al producto de sus probabilidades marginales, y a la inversa, si la probabilidad conjunta de dos eventos es igual al producto de sus probabilidades marginales entonces esos dos eventos son independientes probabilísticamente**.

A este resultado se le conoce como la **regla de la multiplicación de probabilidades**.

Dos eventos A y B son independientes probabilísticamente si y sólo si

$$P(A \cap B) = P(A)P(B)$$

Consideremos un ejemplo sencillo.

**Ejemplo 1**. Se arroja una moneda tres veces. Se desea determinar la probabilidad de obtener cara, cruz y cara en ese orden.

Solución:

Sea C el evento «sale cara»y X el evento «sale cruz».

Se desea determinar P(C, X, C). Por otro lado, nuestra experiencia —asumiendo que la moneda es legal— nos dice que la probabilidad de obtener cruz o cara en un determinado lanzamiento de la moneda no se altera por la historia de los resultados anteriores. Esto significa que podemos asumir que los eventos son independientes probabilísticamente, por lo que:

$P(C, X, C) = P(C)P(X)P(C)$

Como cada probabilidad marginal es igual a 0.5, el resultado final es 0.125.

# 2.8. Teorema de Bayes

Cuando calculamos la probabilidad de B dado que A ya ocurrió, de alguna manera se piensa que el evento A es algo que sucede antes que B y que A puede ser (tal vez) causa de B o puede contribuir a su aparición. También de algún modo podemos decir que A normalmente ocurre antes que B.

Pensemos, por ejemplo, que deseamos saber la probabilidad de que un estudiante apruebe el examen parcial de estadística dado que estudió por lo menos veinte horas antes del mismo.

En algunas ocasiones sabemos que ocurrió el evento B y queremos saber cuál es la probabilidad de que haya ocurrido el evento A. En nuestro ejemplo anterior la pregunta sería cuál es la probabilidad de que el alumno haya estudiado por lo menos veinte horas dado que, efectivamente, aprobó el examen de estadística.

Esta probabilidad se encuentra aplicando una regla que se conoce como teorema de Bayes, mismo que se muestra enseguida.

$$P(A_i / B) = \frac{P(B / A_i) \cdot P(A_i)}{P(B / A_1) \cdot P(A_1) + P(B / A_2) \cdot P(A_2) + \ldots\ldots\ldots\ldots + P(B / A_k) \cdot P(A_k)}$$

En donde:

| | |
|---|---|
| $P(A_i) =$<br>Probabilidad previa | Es la probabilidad de un evento posible antes de cualquier otra información. |
| $P(B/A_i) =$<br>Probabilidad condicional | Es la probabilidad de que el evento "B" ocurra en cada posible suceso de $A_i$. |
| $P(B/A_i) \cdot P(A_i) =$<br>Probabilidad conjunta | Equivalente a la probabilidad de $(A_i \cap B)$ determinada por la regla general de la multiplicación. |
| $P(A_i/B) =$<br>Probabilidad *a posteriori* | Combina la información provista en la distribución previa con la que se ofrece a través de las probabilidades condicionales para obtener una probabilidad condicional final. |

**Ejemplo 1**: Un gerente de crédito trata con tres tipos de riesgos crediticios con sus clientes: las personas que pagan a tiempo, las que pagan tarde (morosos) y las que no pagan. Con base en datos estadísticos, las proporciones de cada grupo son 72.3%, 18.8% y 8.9%, respectivamente.

También por experiencia, el gerente de crédito sabe que el 82.4% de las personas del primer grupo son dueños de sus casas: el 53.6% de los que pagan tarde, son dueños de sus casas, y el 17.4% de los que no pagan, también son propietarios de sus casas.

El gerente de crédito desea calcular la probabilidad de que un nuevo solicitante de crédito en un futuro, si es dueño de su casa:

a) Pague a tiempo.

b) Pague tarde.

c) No pague.

d) Elaborar su tabla de probabilidades.

Solución:

Definición de eventos:

$P_1 =$ Clientes que pagan a tiempo.

$P_2 =$ Clientes que pagan tarde.

$P_3 =$ Clientes que no pagan.

$D =$ Clientes dueños de sus casas.

$D' =$ Clientes que no son dueños de sus casas.

Expresión general:

$$P(P_i / D) = \frac{P(D/P_i) \cdot P(P_i)}{P(D/P_1) \cdot P(P_1) + P(D/P_2) \cdot P(P_2) + P(D/P_3) \cdot P(P_3)}$$

Donde,

$$P_1 = 0.723$$
$$P_2 = 0.188$$
$$P_3 = 0.089$$
$$P(D/P_1) = 0.824$$
$$P(D/P_2) = 0.536$$
$$P(D/P_3) = 0.174$$

a) Probabilidad de que un nuevo solicitante pague a tiempo.

Sustituyendo en la fórmula general:

$$P(P_1 / D) = \frac{0.824 \times 0.723}{0.824 \times 0.723 + 0.536 \times 0.188 + 0.174 \times 0.089} = \frac{0.596}{0.712} = 0.837 = 83.7\%$$

Un nuevo solicitante que sea propietario de su casa tendrá un 83.7% de probabilidades de que pague a tiempo.

b) Probabilidad de que un nuevo solicitante pague tarde:

$$P(P_2 / D) = \frac{0.536 \times 0.188}{0.824 \times 0.723 + 0.536 \times 0.188 + 0.174 \times 0.089} = \frac{0.101}{0.712} = 0.142 = 14.2\%$$

Un nuevo solicitante que sea propietario de su casa tendrá un 14.2% de probabilidades de que pague tarde (cliente moroso).

c) Probabilidad de que un nuevo solicitante no pague.

$$P(P_3 / D) = \frac{0.174 \times 0.089}{0.824 \times 0.723 + 0.536 \times 0.188 + 0.174 \times 0.089} = \frac{0.015}{0.712} = 0.021 = 2.1\%$$

Un nuevo solicitante que sea propietario de su casa tendrá un 2.1% de probabilidades de que nunca pague.

Esta información es de gran utilidad para determinar si aprobar o no una solicitud de crédito.

El denominador de la fórmula representa la probabilidad marginal del evento “D”. Se puede indicar que un 71.2% de sus clientes son dueños de sus casas.

Se puede inferir también que una persona no “dueña de su casa” tendrá una probabilidad de pagar a tiempo de sólo un 16.3% o de pagar tarde un 85.8% y de no pagar de un 97.9%.

Este análisis se puede elaborar con mayor facilidad si se utiliza una tabla de probabilidades tal como se muestra:

| Evento $P_i$ | Probabilidad Previa $P(P_i)$ | Probabilidad Condicional $P(D \mid P_i)$ | Probabilidad Conjunta $P(D \mid P_i) \times P(P_i)$ | Probabilidad *a posteriori* $P(P_i \mid D)$ |
|---|---|---|---|---|
| $P_1$ | 0.723 | 0.824 | 0.596 | 0.837 |
| $P_2$ | 0.188 | 0.536 | 0.101 | 0.142 |
| $P_3$ | 0.089 | 0.174 | 0.015 | 0.021 |
| **Total** | **1.000** | | **0.712** | **1.000** |

Tabla de probabilidades del Teorema de Bayes.

El interés por el conocimiento de la teoría de la probabilidad nos permite obtener elementos de información verdaderamente útiles para su aplicación en las diversas situaciones de vida de tipo personal, profesional o social. La distinción de las variables aleatorias discretas o continuas así como las reglas de adición y de multiplicación dan como resultado una interpretación adecuada del concepto de probabilidad condicional, la cual tiene gran influencia en múltiples actividades de carácter comercial, industrial, o de servicios.

Las tablas de probabilidad conjunta son instrumentos muy valiosos para predecir el grado de probabilidad de ocurrencia de hechos supuestos de antemano. El concepto de probabilidad marginal nos conduce a comprender la probabilidad de un evento simple formado por la sumatoria de varios eventos conjuntos y es la base del Teorema de Bayes.

La utilización de este teorema nos permitirá descubrir la probabilidad de que un cierto evento haya sido la causa del evento que está ocurriendo o está por ocurrir. Los conceptos estudiados en este tema constituyen un importante soporte para el conocimiento de las distribuciones básicas de probabilidad de variables discretas o continuas que se verán más adelante.

# RESUMEN

La probabilidad es una rama de las matemáticas, cuyo desarrollo tiene su génesis en el siglo XVII, cuando se buscó contar con métodos racionales de enfrentar los juegos de azar. Se puede decir que hay tres grandes enfoques, escuelas o paradigmas de probabilidad, a saber, el clásico, el empírico y el subjetivo, ninguno de los cuales escapa al tratamiento axiomático, que es lo que da la estructura al tratamiento matemático moderno de la probabilidad. Como parte de esta estructura matemática se incorporan, además, el cálculo de probabilidades a la luz de información adicional bajo el concepto de probabilidad condicional y del teorema de Bayes.

# BIBLIOGRAFÍA

## SUGERIDA

| Autor | Capítulo | Páginas |
|---|---|---|
| Anderson, Sweeney, Williams (2005) | 4. Introducción a la probabilidad. Sección 4.2 Eventos y sus probabilidades. | 143-146 |
| | 4.3 Algunos resultados básicos de probabilidad. | 148-151 |
| | 4.4 Probabilidad condicional. | 153-156 |
| | 5. Teorema de Bayes. | 161-165 |
| Berenson, Levine y Krehbiel (2001) | 4. Probabilidad básica y distribuciones de probabilidad. Sección: 4.1 Conceptos básicos de probabilidad. | 155-165 |
| | 4.2 Probabilidad condicional. | 165-175 |
| | 4.3 Teorema de Bayes. | 175-179 |
| Levin y Rubin (2004) | 4. Probabilidad I: Ideas introductorias. Sección: 4.2 Terminología básica en probabilidad. | 129-131 |
| | 4.3 Tres tipos de probabilidad. | 131-137 |
| | 4.4 Reglas de probabilidad. | 137-143 |
| | 4.5 Probabilidades bajo condiciones de independencia estadística. | 143-148 |

| | | |
|---|---|---|
| | 4.6 Probabilidades bajo condiciones de dependencia estadística. | 151-155 |
| | 4.7 Revisión de las estimaciones anteriores de probabilidades: teorema de Bayes. | 158-165 |
| Lind, Marchal, Wathen (2008) | 5. Estudio de los conceptos de la probabilidad.<br>Secciones:<br>¿Qué es la probabilidad? | 140–141 |
| | Enfoques para asignar probabilidades. | 142-147 |
| | Algunas reglas para calcular probabilidades. | 147-156 |
| | Tablas de contingencias. | 156-158 |
| | Teorema de Bayes. | 161-165 |

Anderson, David R.; Sweeney, Dennis J.; Williams, Thomas A. (2005). *Estadística para administración y economía*, 8ª edición, México: International Thomson Editores, pp. 888 más apéndices.

Berenson, Mark L., David M. Levine, y Timothy C. Krehbiel. (2001). *Estadística para administración*, 2ª edición, México: Prentice Hall, 734 pp.

Levin, Richard I. y David S Rubin. (2004). *Estadística para administración y economía*, 7ª edición, México: Pearson Educación Prentice Hall, pp. 826 más anexos.

Lind, Douglas A., Marchal, William G. y Wathen, Samuel, A. (2008). *Estadística aplicada a los negocios y la economía,* 13ª edición, México: McGraw-Hill Interamericana, 859 pp.

# Unidad 3

# Distribuciones de probabilidad


SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

# OBJETIVO PARTICULAR

El alumno aplicará las diferentes distribuciones de probabilidad y su interpretación en la solución de problemas.

# TEMARIO DETALLADO
## (12 horas)

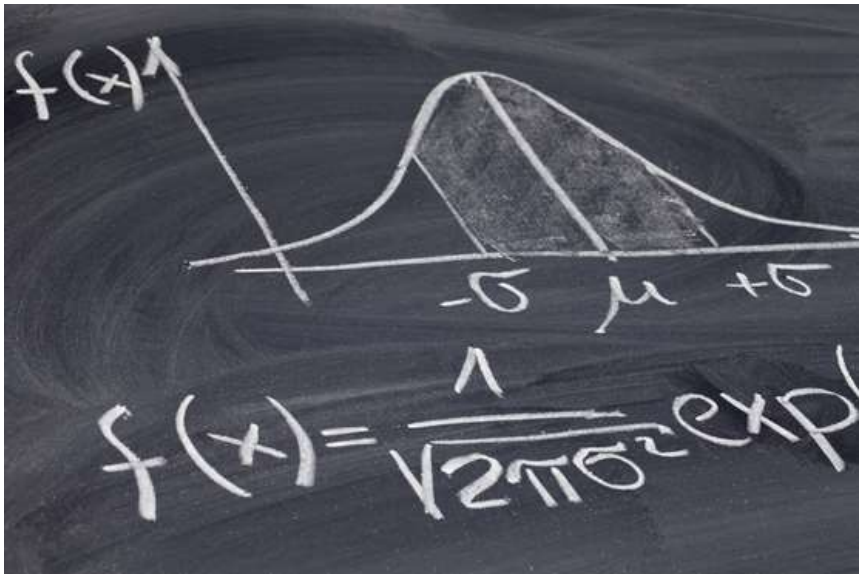
**3. Distribuciones de probabilidad**

- 3.1. Variables aleatorias, discretas y continuas
- 3.2. Media y varianza de una distribución de probabilidad
- 3.3. Distribuciones de probabilidad de variables discretas
  - 3.3.1. Distribución binomial
  - 3.3.2. Distribución de Poisson
  - 3.3.3. La distribución de Poisson como aproximación de la distribución binomial
  - 3.3.4. Distribución hipergeométrica
  - 3.3.5. Distribución multinomial
- 3.4. Distribuciones de probabilidad de variables continuas
  - 3.4.1. Distribución normal
  - 3.4.2. Distribución exponencial
- 3.5. Ley de los grandes números

# INTRODUCCIÓN

En esta unidad se describen los diferentes tipos de distribuciones de probabilidad que existen, las técnicas para el cálculo o asignación de probabilidades aplicable para cada tipo de dato y cada situación, se analizan sus características y la aplicación de una de ellas en las diferentes situaciones que se presentan en el mundo de los negocios.

Una distribución de probabilidades da toda la gama de valores que pueden ocurrir con base en un experimento, y resulta similar a una distribución de frecuencias. Sin embargo, en vez de describir el pasado, define qué tan probable es que suceda algún evento futuro.



# 3.1. Variables aleatorias, discretas y continuas

Una **variable** es **aleatoria** si los valores que toma corresponden a los distintos resultados posibles de un experimento; por ello, el hecho de que tome un valor particular es un evento aleatorio.

La variable aleatoria considera situaciones donde los resultados pueden ser de origen cuantitativo o cualitativo, asignando en cualquier caso un número a cada posible resultado.


Por ejemplo, si el experimento consiste en seleccionar a una persona de un colectivo de n de ellas, y lo que nos interesa es el sexo, la variable aleatoria podría tomar los valores 1 si resulta ser un hombre y 2 si resulta ser una mujer. Si lo que nos interesa es la edad, entonces la variable aleatoria tiene tantos posibles valores como edades haya en la población.

En esencia, lo que hace una variable aleatoria es asignar un número a cada posible resultado del experimento.

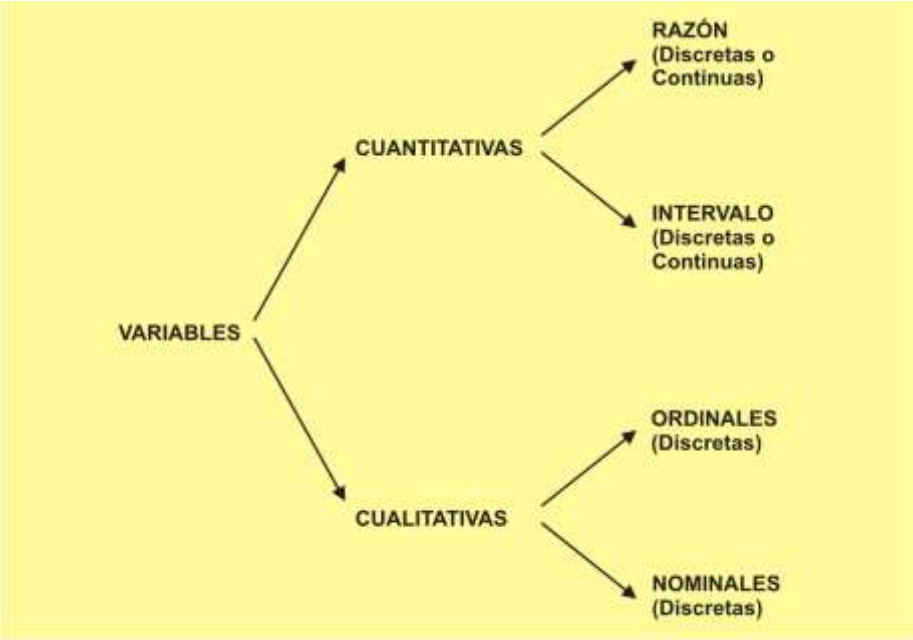
Dependiendo de esta asignación de números las variables aleatorias pueden ser discretas o continuas.

- Las **variables discretas** son aquellas que cuantifican la característica de modo tal que el número de posibles resultados se puede contar, esto es, la variable discreta toma un número finito o infinito numerable de posibles valores. Como ejemplo de este tipo de variables tenemos el número de clientes de un banco, el número de hijos de una familia, el número de alumnos en un grupo de la universidad, el número de personas en una población rural, el número de automóviles en una ciudad, etcétera.

- Las **variables continuas** son aquellas que pueden tomar cualquier valor numérico, dentro de un intervalo previamente especificado. Así, por ejemplo, la variable tiempo en una investigación podría medirse en intervalos de horas, o bien, en horas y minutos, o bien en horas, minutos y segundos según sea el requerimiento de la misma.

Desde el punto de vista de la estadística las variables aleatorias también se clasifican de acuerdo a la escala de medición inherente.

Cuando estudiaste el tema de estadística descriptiva tuviste oportunidad de aprender los conceptos de escala nominal, ordinal, de intervalo y de razón. Estas escalas generan precisamente variables aleatorias del mismo nombre. Ocurre que las variables de intervalos y de razón son cuantitativas y pueden ser discretas o continuas. Los casos nominal y ordinal se refieren a cualidades en donde la variable aleatoria al asignar un número a cada resultado asume que tales cualidades son discretas. El cuadro siguiente te proporciona un panorama general de esta situación.




VARIABLES ALEATORIAS

La clasificación de las variables anteriormente expuesta, que parte del punto de vista de la estadística, no es única, pues cada disciplina científica acostumbra hacer alguna denominación para las variables que en ella se manejan comúnmente.

Por ejemplo, en el área de las ciencias sociales es común establecer relaciones entre variables experimentales; por ello, en este campo del conocimiento, las variables se clasifican, desde el punto de vista metodológico, en dependientes e independientes.

La **variable dependiente** es aquella cuyos valores están condicionados por los valores que toma la variable independiente (o las variables independientes) con la que tiene relación.

Por lo tanto, la variable o las variables independientes son la causa iniciadora de la acción, es decir, condicionan de acuerdo con sus valores a la variable dependiente.

**Ejemplo 1.** Consideremos el comportamiento del ahorro de un individuo en una sociedad. El modelo económico que explica su ahorro podría ser:

**Ahorro = ingreso – gasto**

En este modelo, el ahorro es la variable dependiente y presentará una situación específica de acuerdo con el comportamiento que tengan las variables independientes de la relación.

Un punto importante que debes tener en mente cuando trabajes con variables aleatorias es que no sólo es importante identificarlas y clasificarlas, sino que también deben definirse adecuadamente. Para algunos autores, como Hernández, Fernández y Baptista, su definición deberá establecerse en dos niveles, especificados como nivel conceptual y nivel operacional.

**Nivel conceptual.** Consiste en definir el término o variable con otros términos. Por ejemplo, el término “poder” podría ser definido como “influir más en los demás que lo que éstos influyen en uno”. Este tipo de definición es útil, pero insuficiente para definir una variable debido a que no nos relaciona directamente con la realidad, puesto que, como puede observarse, siguen siendo conceptos.

**Nivel operacional.** Constituye el conjunto de procedimientos que describen las actividades que un observador realiza para recibir las impresiones sensoriales que indican la existencia de un concepto teórico (conceptual) en mayor o menor grado, es decir, consiste en especificar las actividades u operaciones necesarias que deben realizarse para medir una variable.

Con estas dos definiciones, estás ahora en posibilidad de acotar adecuadamente las variables para un manejo estadístico, de acuerdo con el interés que tengas en ellas, para la realización de un estudio o investigación. Mostraremos a continuación un par de ejemplos de ello.

**Ejemplo 1:**

| | |
|---|---|
| **Variable:** | "Ausentismo laboral" |
| **Nivel conceptual:** | "El grado en el cual un trabajador no se reportó a trabajar a la hora en la que estaba programado para hacerlo”. |
| **Nivel operacional:** | "Revisión de las tarjetas de asistencia al trabajo durante el último bimestre". |

**Ejemplo 2:**

| | |
|---|---|
| **Variable:** | "Sexo" |
| **Nivel conceptual:** | "Condición orgánica que distingue al macho de la hembra”. |
| **Nivel operacional:** | "Asignación de la condición orgánica: masculino o femenino". |

Finalmente, es importante mencionar que a la par que defines una variable aleatoria es importante que le asignes un nombre. Por lo general, éste es una letra mayúscula.

# 3.2. Media y varianza de una distribución de probabilidad

La distribución de probabilidad de una variable aleatoria describe cómo se distribuyen las probabilidades de los diferentes valores de la variable aleatoria. Para una **variable aleatoria discreta** "**X**", la distribución de probabilidad se describe mediante una función de probabilidad, a la que también se conoce como **función de densidad**, representada por **f(X)**, que define la probabilidad de cada valor de la variable aleatoria.


Como la probabilidad del universo (o evento universal) debe ser igual a 100%, y además cualquier evento que se defina debe estar contenido en el evento universal, cuando hablamos de cómo distribuir las probabilidades nos referimos a cómo es que se reparte este 100% de probabilidad en los diferentes eventos.

**Ejemplo 1.** Considera el experimento aleatorio que consiste en arrojar un dado dos veces y sumar los resultados de ambas caras. Se desea conocer cuál es la probabilidad de que la suma sea 7.

Solución:

La variable X puede tomar los valores del 2 al 12, inclusive, por lo que se trata de una variable aleatoria discreta. La siguiente tabla nos permitirá calcular las probabilidades de todos los eventos simples.

| Resultado | Segundo dado | | | | | |
|---|---|---|---|---|---|---|
| **Primer dado** | **1** | **2** | **3** | **4** | **5** | **6** |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **5** | **6** | **7** | **8** | **9** | **10** | **11** |
| **6** | **7** | **8** | **9** | **10** | **11** | **12** |

En ella vemos que las diagonales, a las que se ha dado diferente color, determinan el mismo valor de la suma para diferentes combinaciones de resultados de cada uno de los dos dados. Por ejemplo, si queremos saber la probabilidad de que la suma sea 7, nos fijaríamos en la diagonal amarilla y observaríamos que hay 6 formas distintas de obtener tal valor, de un total de 36, por lo que la probabilidad es 7/36.

El ejemplo nos permite darnos cuenta, además, que también podemos calcular fácilmente la probabilidad de que la suma sea menor o igual a 7 y que para ello debemos contar el número de casos que se acumulan desde la diagonal superior izquierda hasta la diagonal amarilla, que corresponde a los valores 2, 3, 4, 5 ,6 y 7. Esto es, se estaría considerando que:

$$P(X \leq 7) = P(2) + P(3) + P(4) + P(5) + P(6) + P(7)$$

Para cualquier otro resultado también estaríamos acumulando probabilidades desde la que corresponde al resultado 2 hasta el resultado tope considerado.

De este modo se construye, a partir de la función de probabilidades, otra función, a la que se denomina **función de distribución acumulativa** y que se denota como **F(*x*)**, donde la x indica el valor hasta el cual se acumulan las respectivas probabilidades. Por ejemplo, P (X≤7) corresponde a F(7).

La tabla siguiente resume la función de probabilidades y la función de distribución acumulativa para el caso del ejemplo:

| i | Función de probabilidad P( X = i ) | Función de distribución acumulativa P(X ≤ i ) |
|---|---|---|
| **2** | 1/36 | 1/36 |
| **3** | 2/36 | 1/36 + 2/36 = 3/36 |
| **4** | 3/36 | 3/36 + 3/36 = 6/36 |
| **5** | 4/36 | 6/36 + 4/36 = 10/36 |
| **6** | 5/36 | 10/36 + 5/36 = 15/36 |
| **7** | 6/36 | 15/36 + 6/36 = 21/36 |
| **8** | 5/36 | 21/36 + 5/36 = 26/36 |
| **9** | 4/36 | 26/36 + 4/36 = 30/36 |
| **10** | 3/36 | 30/36 + 3/36 = 33/36 |
| **11** | 2/36 | 33/36 + 2/36 = 35/36 |
| | | |
| **12** | 1/36 | 35/36 +1/36 = 36/36 = 1 |

Obsérvese que el valor de la función de distribución acumulativa para el último valor de la variable aleatoria acumula precisamente 100%.

**Esperanza y varianza**

Cuando se trabaja con variables aleatorias, no basta con conocer su distribución de probabilidades. También será importante obtener algunos valores típicos que resuman, de alguna forma, la información contenida en el comportamiento de la variable. De esos valores importan fundamentalmente dos: la esperanza y la varianza.

*Esperanza.*

Corresponde al valor promedio, considerando que la variable aleatoria toma los distintos valores posibles con probabilidades que no son necesariamente iguales. Por ello se calcula como la suma de los productos de cada posible valor de la variable aleatoria por la probabilidad del respectivo valor. Se le denota como μ

$$\text{Esperanza} = \mu = \sum x\left[P(X = x)\right]$$

Donde la suma corre para todos los valores x de la variable aleatoria.

*Varianza*

Es el valor esperado o esperanza de las desviaciones cuadráticas con respecto a la media μ. Se denota como $\sigma^2$ y se calcula como la suma del producto de cada desviación cuadrática por la probabilidad del respectivo valor.

$$\text{Varianza} = \sigma^2 = \sum (x - \mu)^2\left[P(X = x)\right]$$

Donde la suma corre para todos los valores x de la variable aleatoria.

La raíz cuadrada de la varianza es, desde luego, la desviación estándar.

**Ejemplo 2.** Considerando el mismo experimento del ejemplo anterior, determinar la esperanza y varianza de la variable aleatoria respectiva.

| X | Función de probabilidad $P(X = x)$ | $x\ P(X = x)$ | $(x - 7)^2$ | $(x - 7)^2\ P(X = x)$ |
|---|---|---|---|---|
| **2** | 1/36 | 2/36 | 25 | 25/36 |
| **3** | 2/36 | 6/36 | 16 | 32/36 |
| **4** | 3/36 | 12/36 | 9 | 27/36 |
| **5** | 4/36 | 20/36 | 4 | 16/36 |
| **6** | 5/36 | 30/36 | 1 | 5/36 |
| **7** | 6/36 | 42/36 | 0 | 0 |
| **8** | 5/36 | 40/36 | 1 | 5/36 |
| **9** | 4/36 | 36/36 | 4 | 16/36 |
| **10** | 3/36 | 30/36 | 9 | 27/36 |
| **11** | 2/36 | 22/36 | 16 | 32/36 |
| **12** | 1/36 | 12/36 | 25 | 25/36 |
| | **Suma** | **252/36=7** | | **260/36** |

Podemos decir entonces que al arrojar dos dados y considerar la suma de los puntos que cada uno muestra, el valor promedio será 7 con una desviación estándar de 2.69.

# 3.3. Distribuciones de probabilidad de variables discretas

Las distribuciones binomial, de Poisson, hipergeométrico y multinomial son cuatro casos de distribuciones de probabilidad de variables aleatorias discretas.

## 3.3.1. Distribución binomial

La distribución binomial se relaciona con un experimento aleatorio conocido como **experimento de Bernoulli** el cual tiene las siguientes características:

- El experimento está constituido por un número finito, n, de pruebas idénticas.
- Cada prueba tiene exactamente dos resultados posibles. A uno de ellos se le llama arbitrariamente éxito y al otro, fracaso.
- La probabilidad de éxito de cada prueba aislada es constante para todas las pruebas y recibe la denominación de “p”.
- Por medio de la distribución binomial tratamos de encontrar un número dado de éxitos en un número igual o mayor de pruebas.

Puesto que sólo hay dos resultados posibles, la probabilidad de fracaso, a la que podemos denominar q, está dada por la diferencia 1 – p, esto es, corresponde al complemento de la probabilidad de éxito, y como esta última es constante, entonces también lo es la probabilidad de fracaso.

La probabilidad de “x” éxitos en n intentos está dada por la siguiente expresión:

$$P(x)={}_nC_x\,p^x q^{n-x}$$

Esta fórmula nos dice que la probabilidad de obtener “x” número de éxitos en n pruebas (como ya se indicó arriba) está dada por la multiplicación de n combinaciones en grupos de x por la probabilidad de éxito elevada al número de éxitos deseado y por la probabilidad de fracaso elevada al número de fracasos deseados.

*Con el término **combinaciones** nos referimos al número de formas en que podemos extraer grupos de k objetos tomados de una colección de n de ellos (n ≥ k), considerando que el orden en que se toman o seleccionan no establece diferencia alguna. El símbolo nCk denota el número de tales combinaciones y se lee combinaciones de n objetos tomados en grupos de k. Operativamente,*

***nCk = n! / [x! (n-x)!]***

*El símbolo ! indica el factorial del número, de modo que*

***n! = (1)(2)(3)…(n)***

A continuación se ofrecen varios ejemplos que nos ayudarán a comprender el uso de esta distribución.

**Ejemplo 1.** Un embarque de veinte televisores incluye tres unidades defectuosas. Si se inspeccionan tres televisores al azar, indica cuál es la probabilidad de que se encuentren dos defectuosos.

Solución:

Podemos verificar si se trata de una distribución binomial mediante una lista de chequeo de cada uno de los puntos que caracterizan a esta distribución.

| Característica | Estatus | Observación |
|---|---|---|
| Hay un número finito de ensayos | SI | Cada televisor es un ensayo y hay 3 de ellos |
| Cada ensayo tiene sólo dos resultados | SI | Cada televisor puede estar defectuoso o no |
| La probabilidad de éxito es constante | SI | La probabilidad de que la unidad esté defectuosa es 3 / 20 |
| Se desea saber la probabilidad de un cierto número de éxitos | SI | Se desea saber la probabilidad de que X=2 |

Una vez que hemos confirmado que se trata de una distribución binomial aplicamos la expresión

$P(x)= nCx\ p^x q^{(n-x)}$, de modo que:

$$P(2) = {}_3C_2\ (3/20)^2(17/20)^1 = 3\ (0.0225)\ (0.85) = 0.057375$$

**Ejemplo 2.** Una pareja de recién casados planea tener tres hijos. Di cuál es la probabilidad de que los tres hijos sean varones si consideramos que la probabilidad de que el descendiente sea hombre o mujer es igual.

Solución:

Verificamos primero si se cumplen los puntos que caracterizan la distribución binomial.

Claramente, es un experimento aleatorio con tres ensayos, y en todos ellos sólo hay dos resultados posibles, cada uno con probabilidad de 0.5 en cada ensayo. Si se define como éxito que el sexo sea masculino, entonces podemos decir que se desea saber la probabilidad de que haya tres éxitos.

Entonces, el experimento lleva a una distribución binomial y,

$$P(3) = {}_3C_3\,(1/2)^3(1/2)^0 = (1/2)^3 = 0.0125$$

**Ejemplo 3**. Se sabe que el 30% de los estudiantes de secundaria en México es incapaz de localizar en un mapa el lugar donde se encuentra Afganistán. Si se entrevista a seis estudiantes de este nivel elegidos al azar:

a) ¿Cuál será la probabilidad de que exactamente dos puedan localizar este país?

b) ¿Cuál será la probabilidad de que un máximo de dos puedan localizar este país?

Solución:

Al igual que en los casos anteriores verificamos si se cumple o no que el experimento lleva a una distribución binomial.

Se trata de un experimento con seis ensayos, en cada uno de los cuales puede ocurrir que el estudiante sepa o no sepa localizar Afganistán en el mapa. Si se define como éxito que sí sepa la localización podemos decir que la probabilidad de éxito es de 0.30. Además, las probabilidades que se desea calcular se refieren al número de éxitos. Concluimos que el experimento es Bernoulli y, por lo tanto,

$$P(2) = {}_6C_2\,(0.30)^2(0.70)^4 = 15\,(0.09)\,(0.2401) = 0.324135$$

Por cuanto hace al inciso (b), la frase «un máximo de dos» significa que X toma los valores cero, uno o dos. Entonces,

$$P(X\leq 2) = P(2) + P(1) + P(0) =$$

$$= {}_6C_2\,(0.30)^2(0.70)^4 + {}_6C_1\,(0.30)^1(0.70)^5 + {}_6C_0\,(0.30)^0(0.70)^6 =$$

$= 15(0.09)(0.2401) + 6(0.30)(0.16807) + 0.1176 =$

$= 0.661941$

*Esperanza y varianza de una variable aleatoria binomial*

Consideremos de nueva cuenta el ejemplo 1.

¿Qué pasa con las probabilidades de los otros valores posibles para la variable aleatoria? Si hacemos los cálculos respectivos tendríamos:

$P(0) = {}_3C_0\ (3/20)^0(17/20)^3 = 0.614125$

$P(1) = {}_3C_1\ (3/20)^1(17/20)^2 = 3\ (0.15)(0.7225) = 0.325125$

$P(3) = {}_3C_3\ (3/20)^3(17/20)^0 = (0.003375) = 0.003375$

Si recordamos que P(2) = 0.057375, entonces podemos confirmar que

$$P(0) + P(1) + P(2) + P(3) = 1.00,$$

lo que era de esperarse puesto que los valores 0, 1, 2 y 3 constituyen el universo en el experimento en cuestión.

Con estos valores podemos determinar la esperanza y varianza de la variable aleatoria considerada. Para ello nos es útil acomodar los datos en una tabla recordando que:

$\mu = \Sigma\, x\, [P(X=x)]$, y que, $\sigma^2 = \Sigma\, (x - \mu)^2\, [P(X=x)]$

| $x$ | Función de probabilidad $P(X = x)$ | $x\, P(X = x)$ | $(x - 0.45)^2$ | $(x - 0.45)^2\, P(X = x)$ |
|---|---|---|---|---|
| **0** | 0.614125 | 0.000000 | 0.2025 | 0.124360 |
| **1** | 0.325125 | 0.325125 | 0.3025 | 0.098350 |
| **2** | 0.057375 | 0.114750 | 2.4025 | 0.137843 |
| **3** | 0.003375 | 0.010125 | 6.5025 | 0.021946 |
| | **Suma** | 0.450000 | | 0.382500 |

Entonces, la esperanza es 0.45 y la varianza 0.3825.

Si interpretamos las probabilidades anteriores en un sentido frecuentista, diríamos que si consideramos un número grande de realizaciones del experimento, por ejemplo un millón de veces, en aproximadamente 614 125 realizaciones tendremos refrigeradores sin defecto, en 325 125 veces encontraremos un refrigerador con defecto, en otras 57 375 ocasiones encontraremos dos refrigeradores con defecto y en 3 375 veces los tres refrigeradores estarían defectuosos.

Con estos datos podemos elaborar una tabla de distribución de frecuencias y calcular el promedio de refrigeradores defectuosos.

| Número de refrigeradores defectuosos (x) | Frecuencia (f) | fm |
|---|---|---|
| **0** | 614 125 | 0 |
| **1** | 325 125 | 325 125 |
| **2** | 57 375 | 114 750 |
| **3** | 3 375 | 10 125 |
| **Total** | 1 000 000 | 450 000 |

Luego,

$\mu = 450\,000 / 1\,000\,000 = 0.45$

Asimismo, podemos calcular la varianza:

$\sigma^2 = [614\,125\,(0-0.45)^2 + 325\,125\,(1-0.45)^2 + 57\,375\,(2-0.45)^2 + 3\,375\,(3–0.45)2] / 100$

$= (124\,360.313 + 98\,350.3125 + 137\,843.438 + 29\,945.9375) / 100$

$= 0.3825$

Observa que hemos seguido fielmente las lecciones de estadística descriptiva en el cálculo de μ y σ y que hemos llegado a los mismos valores que ya habíamos obtenido. Esto nos proporciona por lo menos un esquema con el cual podemos interpretar la esperanza y varianza, haciendo uso del concepto de frecuencias.

Es importante además, darse cuenta que podemos llegar a estos mismos valores de un modo más sencillo si nos percatamos que

- $\mu = 0.45$ es precisamente el resultado que se obtiene al multiplicar el número de ensayos por la probabilidad de éxito, esto es, 3(0.15)
- $\sigma^2 = 0.3825$ es precisamente el resultado que se obtiene al multiplicar el número de ensayos por la probabilidad de éxito y por la de fracaso, esto es, 3(0.15)(0.85)

En otras palabras,

Media y varianza de una variable aleatoria binomial

$\mu = np \qquad \sigma^2 = npq$

Puede ocurrir, como en el caso del ejemplo anterior, que la esperanza da un valor que no coincide con los valores posibles de la variable aleatoria. Por eso se dice que la esperanza es un valor ideal.

Por otra parte, si desglosamos cada uno de los elementos que integran la expresión del cálculo de probabilidades de la distribución binomial y consideramos las expresiones para el cálculo de la media y la varianza, tendremos que:

$$nCx = n! / [x! (n–x)!]$$

$$p^x = p^x$$

$$q^{n-x} = (1–p)^{n-x}$$

$$\text{Media} = np$$

$$\text{Varianza} = n\,p\,(1 – p)$$

Lo que nos revela que para poder calcular cualquier probabilidad con el modelo binomial o su esperanza o varianza debemos conocer los valores de *n*, el número total de ensayos, y de *p*, la probabilidad de éxito. El valor de *x*, el número de éxitos se establece de acuerdo con las necesidades del problema.

Lo anterior nos permite concluir que la distribución binomial queda completamente caracterizada cuando conocemos los valores de ***n*** y ***p***. Por esta razón a estos valores se les conoce como los **parámetros de la distribución**.

Un error que suele cometerse a propósito de la distribución binomial es considerar que sus parámetros son la esperanza y varianza de la variable aleatoria respectiva. En realidad estos dos valores se expresan en función de los parámetros.

El siguiente ejemplo nos ayudará a entender este concepto.

**Ejemplo 4*:*** De acuerdo con estudios realizados en un pequeño poblado, el 20% de la población tiene parásitos intestinales. Si se toma una muestra de 1,400 personas, ¿cuántos esperamos que tengan parásitos intestinales?

$$\text{Media} = np = 1400(0.20) = 280$$

Éste es el número promedio de elementos de la muestra que tendría ese problema.

Usando el teorema de Tchebyshev podríamos considerar que el valor real estaría a dos desviaciones estándar con un 75% de probabilidades y a tres con un 89%. De acuerdo con ello obtenemos la desviación estándar y posteriormente determinamos los intervalos.

***Teorema de Tchebyshev***

*El teorema de Tchebyshev señala que la probabilidad de que una variable aleatoria tome un valor contenido en k desviaciones estándar de la media es cuando menos 1 – 1 / $k^2$*

Desviación estándar:

$$\sigma = \sqrt{npq} = \sqrt{1,400 \times 0.20 \times 0.80} = \sqrt{224} = 14.97$$

La media más menos dos desviaciones estándar nos daría un intervalo de 250 a 310 personas que tienen problemas.

La media más menos tres desviaciones estándar nos daría un intervalo de 235 a 325 personas que podrían tener problemas.

## 3.3.2. Distribución de Poisson

Es otra distribución teórica de probabilidad de variable aleatoria discreta y tiene muchos usos en economía y comercio. Se debe al teórico francés Simeón Poisson quien la derivó en 1837 como un caso especial (límite) de la distribución binomial.

Se puede utilizar para determinar la probabilidad de un número designado de éxitos cuando los eventos ocurren en un espectro continuo de tiempo y espacio. Es semejante al proceso de Bernoulli, excepto que los eventos ocurren en un espectro continuo, de manera que al contrario del modelo binomial, se tiene un número infinito de ensayos.

Como ejemplo tenemos el número de llamadas de entrada a un conmutador en un tiempo determinado, o el número de defectos en 10 $m^2$ de tela.

En cualquier caso, sólo se requiere conocer el **número promedio de éxitos para la dimensión específica de tiempo o espacio de interés**.

Este número promedio se representa generalmente por $\lambda$ (lambda) y la fórmula de una distribución de Poisson es la siguiente:

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

En esta fórmula, x representa el número de éxitos cuya probabilidad deseamos calcular; $\lambda$ es el promedio de éxitos en un periodo de tiempo o en un cierto espacio; “e” es la base de los logaritmos naturales; y el símbolo de admiración representa el factorial del número que se trate.

**Ejemplo 1**: El manuscrito de un texto de estudio tiene un total de 40 errores en las 400 páginas de material. Los errores están distribuidos aleatoriamente a lo largo del texto. Calcular la probabilidad de que:

a) Un capítulo de 25 páginas tenga dos errores exactamente.
b) Un capítulo de 40 páginas tenga más de dos errores.
c) Una página seleccionada aleatoriamente no tenga errores.

Solución:

En cada caso debemos establecer primero el número promedio de errores. En el inciso (a) nos referiremos al número promedio por cada 25 páginas, en el inciso (b) por cada 40 páginas y en el (c) por página. Esto lo podemos hacer mediante el procedimiento de proporcionalidad directa o regla de tres.

a)Dos errores en 25 páginas

Datos

$$40 - 400$$
$$\lambda - \quad 25$$
$$\therefore \quad \lambda = 2.5$$
$$x = 2$$
$$e = 2.71828$$

$$p\left(2/2.5\right) = \frac{2.5^2 \cdot 2.71828^{-2.5}}{2!} = 0.256 = 25.6\%$$

Existe un 25.6% de probabilidad de que un capítulo de 25 páginas tenga exactamente dos errores.

b)Más de dos errores en 40 páginas

Datos

$40 - 400$

$\lambda - \quad 40$

$\therefore \quad \lambda = 4$

$x = 2$

$e = 2.71828$

$$p(0/4) = \frac{4^0 \cdot 2.71828^{-4}}{0!} = 0.018 = 1.8\%$$

$$P(1/4) = \frac{4^1 \cdot 2.71828^{-4}}{1!} = 0.073 = 7.3\%$$

$$P(2/4) = \frac{4^2 \cdot 2.71828^{-4}}{2!} = 0.146 = 14.6\%$$

$$\therefore P(>2/4) = 1 - (0.018 + 7.3 + 14.6) = 0.762 = 76.2\%$$

Existe un 76.2% de probabilidad de que un capítulo de 40 páginas tenga más de dos errores.

c) Una página no tenga errores:

Datos

$40 - 400$

$\lambda - \quad 1$

$\therefore \quad \lambda = 0.10$

$x = 2$

$e = 2.71828$

$$p(0/0.10) = \frac{0.10^0 \cdot 2.71828^{-0.10}}{0!} = 0.905 = 90.5\%$$

Existe un 90.5% de probabilidad de que una sola página seleccionada aleatoriamente no tenga errores.

Un aspecto importante de la distribución de Poisson es que su media y varianza son iguales. De hecho,

Media y Varianza de una variable aleatoria Poisson

$\mu = \lambda \qquad \sigma^2 = \lambda$

De acuerdo con lo anterior, para determinar las probabilidades en un modelo de Poisson o calcular su esperanza o varianza debemos conocer el valor de λ, esto es, del número promedio de éxitos. Éste es el parámetro de la distribución.

## 3.3.3. La distribución de Poisson como una aproximación a la distribución binomial

En un experimento de Bernoulli, tal como los que acabamos de estudiar en la distribución binomial, puede suceder que el número de ensayos sea muy grande y/o que la probabilidad de acierto sea muy pequeña y los cálculos se vuelven muy laboriosos. En estas circunstancias, podemos usar la distribución de Poisson como una aproximación a la distribución binomial.

**Ejemplo 2**: Una fábrica recibe un embarque de 1, 000,000 de rondanas. Se sabe que la probabilidad de tener una rondana defectuosa es de .001. Si obtenemos una muestra de 3000 rondanas, ¿cuál será la probabilidad de encontrar un máximo de tres defectuosas?

Solución*:*

Este ejemplo, desde el punto de vista de su estructura, corresponde a una distribución binomial. Sin embargo, dados los volúmenes y probabilidades que se manejan es conveniente trabajar con la distribución Poisson, tal como se realiza a continuación. Debemos recordar que un máximo de tres defectuosas incluye la probabilidad de encontrar una, dos y tres piezas defectuosas o ninguna.

Media: $\mu = np = 3,000 \times 0.001 = 3$

$$P(x/\mu) = \frac{\mu^x \cdot e^{-\mu}}{x!}$$

$$P(0/3) = \frac{3^0 \cdot 2.71828^{-3}}{0!} = 0.0498 = 5.0\%$$

$$P(1/3) = \frac{3^1 \cdot 2.71828^{-3}}{1!} = 0.149 = 14.9\%$$

$$P(2/3) = \frac{3^2 \cdot 2.71828^{-3}}{3!} = 0.224 = 22.4\%$$

$$P(3/3) = \frac{3^3 \cdot 2.71828^{-3}}{3!} = 0.224 = 22.4\%$$

La probabilidad de encontrar un máximo de tres piezas defectuosas está dado por la suma de las probabilidades arriba calculadas, es decir: 0.647 ó 64.7% aproximadamente.

## 3.3.4. Distribución hipergeométrica

Este es otro caso de una distribución de variable aleatoria discreta y guarda aparentemente un gran parecido con la distribución binomial, por cuanto en ambas hay un número finito de ensayos, cada uno de los cuales pertenece a uno de dos grupos (el equivalente a éxito o fracaso). Sin embargo, hay otros rasgos que distinguen claramente al modelo hipergeométrico del binomial.

Para aplicar este modelo se requiere verificar los siguientes puntos:

- Hay un población constituida por N observaciones
- La población se puede dividir en dos grupos, K y L, en el primero de los cuales hay k observaciones y en el otro N-k
- De la población se seleccionan al azar n observaciones
- Se desea determinar la probabilidad de que en la muestra haya x observaciones que pertenecen al grupo K

El lector podrá observar que en este caso no se hace ninguna mención explícita en torno a que la probabilidad de éxito sea constante, como en el caso del modelo binomial. Esto se debe a que en el modelo hipergeométrico la extracción de las observaciones no sigue un esquema con reemplazo por lo que la probabilidad de éxito ya no es constante.

Si imaginamos que hacemos la extracción de la muestra elemento por elemento, la probabilidad de que el primero en ser extraído sea del grupo K es, evidentemente, k / N. A continuación, procederíamos a extraer el segundo, pero en este caso el número de casos totales ya no sería N sino N–1 y el número de casos favorables ya no sería k sino k–1, por lo que la probabilidad de que este segundo elemento provenga del grupo K sería (k–1) / (N–1).

Claramente la probabilidad de “éxito” no es constante.

La función que permite asignar las probabilidades en el modelo hipergeométrico es:

$$P(x) = \frac{\left({}_{k}C_{x}\right)\left({}_{N-k}C_{n-x}\right)}{{}_{N}C_{n}}$$

Sus parámetros son precisamente, N, n y k. Son los parámetros porque conociendo estos valores se pueden ya calcular probabilidades con el modelo hipergeométrico.

**Ejemplo 1**. Un juez tiene ante sí 35 actas testimoniales de las cuales sabe que 18 incluyen falso testimonio. Si extrae una muestra de tamaño 10, ¿cuál es la probabilidad de que haya 5 actas con falso testimonio?

Solución:

Los datos del problema nos permiten identificar que:

N=35

n=10

k=18

x=5

$$\Rightarrow P(5) = \frac{\left({}_{18}C_{5}\right)\left({}_{17}C_{5}\right)}{{}_{35}C_{10}} = 0.2888$$

## 3.3.5. Distribución multinomial

Esta distribución de variable aleatoria discreta se aplica en situaciones en las que:

- Se extrae una muestra de N observaciones
- Las observaciones se pueden dividir en k grupos
- En cada extracción la probabilidad ($p_k$) de que el elemento seleccionado pertenezca a uno de los k diferentes grupos permanece constante
- Se desea determinar la probabilidad de que de los N elementos, $x_1$ pertenezcan al grupo 1, $n_2$ al grupo 2 y sucesivamente hasta el grupo k al que deben pertenecer $x_k$ elementos, donde $\sum x_k = N$

Como se puede apreciar, las semejanzas con la distribución binomial son claras, ya que en ambos modelos hay un número finito de ensayos y las probabilidades se mantienen constantes. La diferencia es que en el modelo multinomial la población se divide en k grupos y en el binomial sólo en dos (“éxito” o “fracaso”). En este sentido, se puede decir que la distribución binomial es un caso particular de la multinomial.

La función que permite calcular las probabilidades es:

$$P(x_1, x_2, \ldots, x_{k-1}) = \frac{N!}{x_1! x_2! x_3! \ldots x_k!} p_1^{x_1} p_2^{x_2} p_3^{x_3} \ldots p_k^{x_k}$$

donde,

$p_k$ es la probabilidad (constante) de que un elemento cualquiera de la población pertenezca al grupo k, con $\sum p_k = 1$

Sus parámetros son N y un conjunto de k–1 valores de probabilidad. Son los parámetros porque conociendo estos valores se pueden ya calcular probabilidades con el modelo multinomial.

**Ejemplo 1.** Un perito debe presentar ante una autoridad judicial 14 peritajes. Se sabe por experiencias anteriores que dicha autoridad acepta 40% de los peritajes, desecha 35% y solicita nuevos peritajes en otro 25% de los casos. El perito desea determinar la probabilidad de que le acepten 10 peritajes y le desechen sólo uno.

Solución:

Si designamos como grupo 1 el de los peritajes aceptados y como grupo 2 el de los desechados, los datos del problema nos permiten identificar que:

N=14

p1 = 0.40

p2 = 0.35

p3 = 0.25

$$\Rightarrow P(10,1,4) = \frac{14!}{10!1!4!}(0.40)^{10}(0.35)^{1}(0.25)^{4} = 0.0001$$

x1=10

x2=1

x3=4

# 3.4 Distribuciones de probabilidad de variables continuas

Para comprender la diferencia entre las variables aleatorias discretas y las continuas recordemos que las variables aleatorias continuas pueden asumir cualquier valor dentro de un intervalo de la recta numérica o de un conjunto de intervalos.

Como cualquier intervalo contiene una cantidad infinita de valores, no es posible hablar de la probabilidad de que la variable aleatoria tome un determinado valor; en lugar de ello, debemos pensar en términos de la probabilidad de que una variable aleatoria continua tome un valor dentro de un intervalo dado.

Esto significa que si X es una variable aleatoria continua, entonces por definición P(X = *x*) = 0, cualquiera que sea el valor de *x*.

Las preguntas de interés tomarán entonces alguna de las siguientes formas básicas:

- P(X ≤ a)
- P(X ≥ b)
- P(c ≤ X ≤ d)

donde a, b c y d son números reales

Aquí debe observarse que P(X ≤ a) = P(X < a), ya que como se ha hecho notar, P(X = a) = 0

Para describir las distribuciones discretas de probabilidades retomamos el concepto de una función de probabilidad f(x). Recordemos que en el caso discreto, esta función da la probabilidad de que la variable aleatoria “x” tome un valor específico. En el caso continuo, la contraparte de la función de probabilidad recibe el nombre de **función de densidad** de probabilidad que también se representa por **f(x)**. Para una variable aleatoria continua, la función de densidad de probabilidad especifica el valor de la función en cualquier valor particular de “x” sin dar como resultado directo la probabilidad de que la variable aleatoria tome un valor específico.

Para comprender esto, imagina que se tiene una variable aleatoria continua relativa a un fenómeno que puede repetirse un número muy grande de veces y que los datos se arreglan en una tabla de distribución de frecuencias con la característica especial de que los intervalos se definen de manera que sean muy finos. A continuación se graficarían los datos de la distribución formando en primera instancia un histograma, luego un polígono de frecuencias y de aquí, como paso subsecuente, una curva suavizada. Al tratar de determinar la probabilidad de que la variable tome valores en un intervalo dado se observaría que en el límite, esto es, entre más finos sean los intervalos, tal probabilidad está dada por el área bajo la curva.

La curva suavizada sería la función de densidad de probabilidad, f(x), de modo que el área entre esta curva y el eje X da la probabilidad. Esto lleva a hacer uso del cálculo integral ya que el área está dada por:

$$P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

donde el símbolo ∫ denota el proceso de integración.

Los valores que se obtienen de $P(X \leq a)$ para todos los valores posibles a, constituyen la **función de distribución acumulativa** de la variable aleatoria X, misma que se denota como $\mathbf{F_X}$.

Debe ocurrir, para que $F_X$ sea realmente una función de distribución de probabilidades, que $F_X(\infty) = 1$.

En principio parece complicado el manejo de las funciones de distribución de probabilidades en el caso continuo, particularmente si no se manejan las herramientas del cálculo integral. Sin embargo, en muchas situaciones de soluciones informáticas en las que se requieran análisis de datos, las distribuciones de probabilidad pueden ser de gran ayuda, como por ejemplo la distribución normal.

## 3.4.1. Distribución normal

Esta distribución de probabilidad también es conocida como **"Campana de Gauss"** por la forma que tiene su gráfica y en honor del matemático que la desarrolló. Tal vez dé la impresión de ser un tanto complicada, pero no debes preocuparte por ello, pues para efectos del curso, no es necesario usarla de manera analítica, sino comprender intuitivamente su significado.

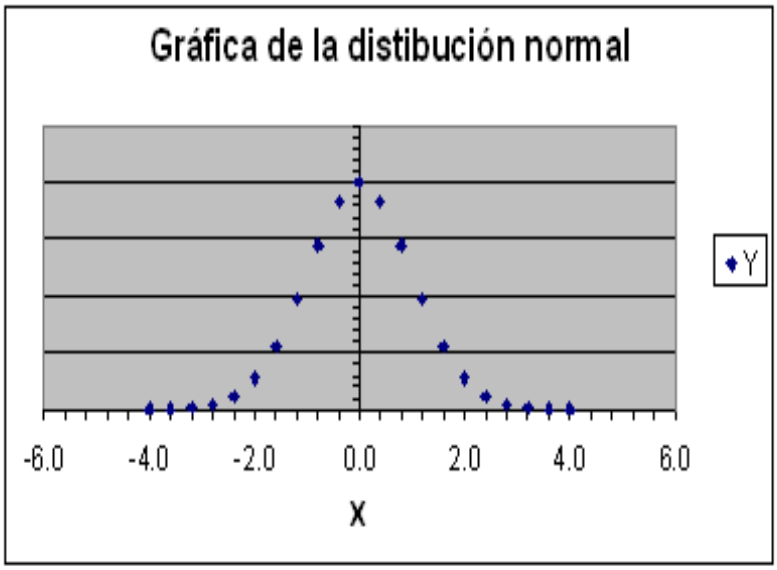
De cualquier manera se dará una breve explicación de la misma para efectos de una mejor comprensión del tema. Su función aparece a continuación.

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

En esta función la "y" es la ordenada de las coordenadas rectangulares cartesianas y representa la altura sobre el eje "x"; x es la abscisa en este sistema de coordenadas; π = 3.14159; "e" corresponde a la base de los logaritmos naturales que el estudiante ya tuvo ocasión de utilizar en la distribución de Poisson. Los símbolos μ y σ corresponden a la media y a la desviación estándar.

Podemos decir que ésta es la expresión de la ecuación normal, de la misma manera que y=mx+b es la expresión de la ecuación de la recta (en su forma cartesiana), por lo que así como podemos asignar distintos valores a m (la pendiente) y b (la ordenada al origen), para obtener una ecuación particular (p. ej. y=4x+2), de la misma manera podemos sustituir μ y σ por cualquier par de valores para obtener un caso particular de la función normal. Si lo hacemos de esa manera, por ejemplo, dándole a la media un valor de cero y a la desviación estándar un valor de 1, podemos ir asignando distintos valores a "x" (en el rango de –4 a 4, por ejemplo) para calcular los valores de "y". Una vez que se ha completado la tabla es fácil graficar en el plano cartesiano. Obtendremos una curva de forma acampanada. A continuación, se muestran tanto los puntos como la gráfica para estos valores.

| X | Y |
|---|---|
| –4.0 | 0.00013 |
| –3.6 | 0.00061 |
| –3.2 | 0.00238 |
| –2.8 | 0.00792 |
| –2.4 | 0.02239 |
| –2.0 | 0.05399 |
| –1.6 | 0.11092 |
| –1.2 | 0.19419 |
| –0.8 | 0.28969 |
| –0.4 | 0.36827 |
| 0.0 | 0.39894 |
| 0.4 | 0.36827 |
| 0.8 | 0.28969 |
| 1.2 | 0.19419 |
| 1.6 | 0.11092 |
| 2.0 | 0.05399 |
| 2.4 | 0.02239 |
| 2.8 | 0.00792 |
| 3.2 | 0.00238 |
| 3.6 | 0.00061 |
| 4.0 | 0.00013 |

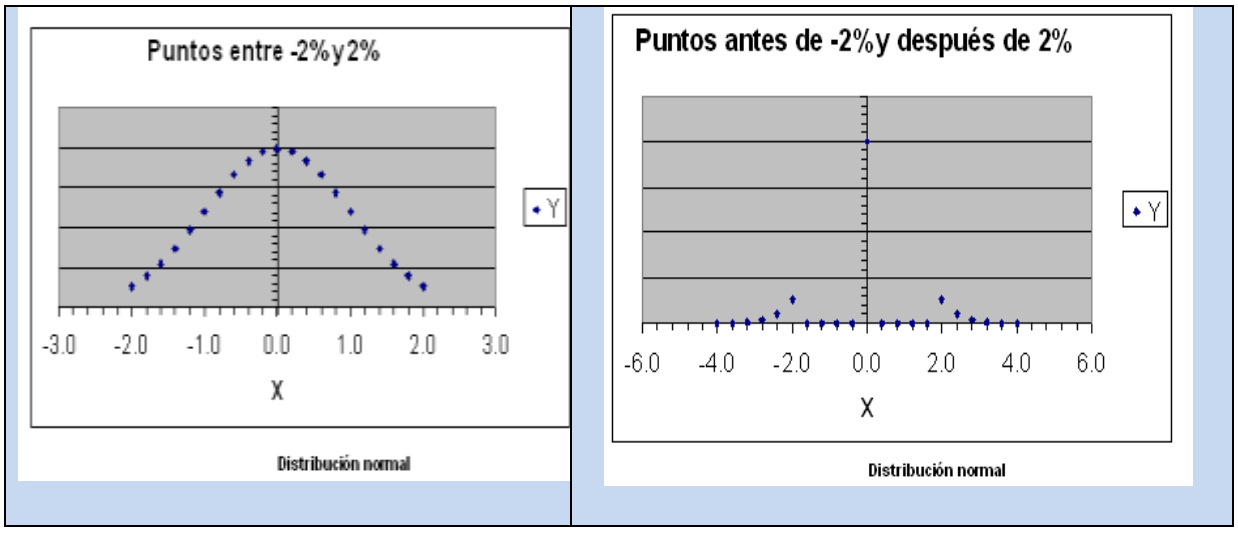



Curva·normal

Es importante mencionar que el área que se encuentra entre la curva y el eje de las abscisas es igual a la unidad o 100%. La curva normal es simétrica en relación con la media. Esto quiere decir que la parte de la curva que se encuentra a la derecha de la curva es como una imagen reflejada en un espejo de la parte que se encuentra a la izquierda de la misma. Esto es importante, pues el área que se encuentra a la izquierda de la media es igual a la que se encuentra a la derecha de la misma y ambas son iguales a 0.5 o el 50%.

Para trabajar con la distribución normal debemos unir los conceptos de área bajo la curva y de probabilidad. La probabilidad de un evento es proporcional al área bajo la curva normal que cubre ese mismo evento. Un ejemplo nos ayudará a entender estos conceptos.

Con base en la figura, vamos a suponer que el rendimiento de las acciones en la Bolsa de Valores en un mes determinado tuvo una media de 0% con una desviación estándar de 1%. (Esto se asimila a lo dicho sobre nuestra gráfica de una distribución normal con una media de cero y una desviación estándar de 1). De acuerdo con esta información, es mucho más probable encontrar acciones cuyo precio fluctúe entre –2% y 2%, que acciones con mayor fluctuación (ver las siguientes gráficas).



Para calcular probabilidades en el caso de la distribución normal se cuenta afortunadamente con valores ya tabulados para el caso en que la distribución tiene una media igual a 0 y una desviación estándar igual a 1. A esta distribución se le conoce como **distribución normal estándar**, y se le denota como **N(0,1)**. En los próximos párrafos aprenderemos a utilizar la tabla de la distribución normal estándar.

La figura “Puntos menores a –2% y 2%”, nos muestra el área que hay entre los valores –2 y 2 y además nos enseña que al ser la campana simétrica, tal área es el doble de la que hay entre los valores 0 y 2. En general, el área entre los valores –z y z es el doble de la que hay entre los valores 0 y z. ¿De qué manera nos puede ayudar la tabla a encontrar el valor de tal área?

Al examinar la tabla de la distribución normal, (el alumno puede consultar la que aparece en el apéndice de esta unidad o la de cualquier libro de estadística), podemos observar que la columna de la extrema izquierda tiene, precisamente el encabezado de “Z”. Los valores de la misma se van incrementando de un décimo en un décimo a partir de 0.0 y hasta 4.2 (en nuestra tabla, en otras puede variar). El primer renglón de la tabla también tiene valores de “Z” que se incrementan de un centésimo en un centésimo de .00 a .09. Este arreglo nos permite encontrar los valores del área bajo la curva para valores de “Z” de 0.0 a 4.29.

Así podemos ver que para Z=1 (primera columna del cuerpo de la tabla y renglón de 1.0), el área es de 0.34134. Esto quiere decir, que entre la media y una unidad de z a la derecha tenemos el 34.134% del área de toda la curva.

Por el mismo procedimiento podemos ver que para un valor de Z=1.96 (renglón de Z=1.9 y columna de Z=.06), tenemos el 0.47500 del área. Esto quiere decir que entre la media y una Z de 1.96 se encuentra el 47.5% del área bajo la curva normal. De esta manera, para cualquier valor de Z se puede encontrar el área bajo la curva.

En el caso en z=2, la tabla nos da un valor de 0.47725, por lo que el área encerrada bajo la curva entre los valores –2 y 2 es 2(0.47725) = 0.9545

La manera en que este conocimiento de la tabla de la distribución normal puede aplicarse a situaciones más relacionadas con nuestras profesiones se puede ver en el siguiente ejemplo.

**Ejemplo 1*:*** Una empresa tiene registrados en su base de datos 2000 clientes. Cada cliente debe en promedio $7000 con una desviación estándar de $1000. La distribución de los adeudos de los clientes es aproximadamente normal. Di cuantos clientes esperamos que tengan un adeudo entre $7000 y $8,500.

Solución:

Nos percatamos de que valores como 7000 o 1000 no aparecen en la tabla de la distribución normal. Es allí donde interviene la variable Z porque nos permite convertir los datos de nuestro problema en números que podemos utilizar en la tabla. Lo anterior lo podemos hacer con la siguiente fórmula:

$$Z = \frac{x_i - \mu}{\sigma}$$

En nuestro caso, nos damos cuenta de que buscamos el área bajo la curva normal entre la media, 7000, y el valor de 8500. Sustituyendo los valores en la fórmula obtenemos lo siguiente:

$$z = \frac{8,500 - 7,000}{1,000} = 1.5$$

Buscamos en la tabla de la normal el área bajo la curva para Z=1.5 y encontramos 0.43319. Esto quiere decir que aproximadamente el 43.3% de los saldos de clientes están entre los dos valores señalados.

En caso de que el cálculo de Z arroje un número negativo significa que estamos trabajando a la izquierda de la media. El siguiente ejemplo ilustra esta situación.

**Ejemplo 2:** En la misma base de datos de la empresa del ejemplo anterior deseamos saber qué proporción de la población estará entre $6,500.00 y $7,000.00.

Solución:

Como en el caso anterior, nos damos cuenta de que nos piden el valor de un área entre la media y otro número. Volvemos a calcular el valor de Z.

$$z = \frac{6,500 - 7,000}{1,000} = -0.5$$

Este valor de Z no significa un área negativa; lo único que indica es que el área buscada se encuentra a la izquierda de la media.

Aprovechando la simetría de la curva buscamos el área bajo la curva en la tabla para Z=0.5 (positivo, la tabla no maneja números negativos) y encontramos que el área es de 0.19146. Es decir que la proporción de saldos entre los dos valores considerados es de aproximadamente el 19.1%.

No siempre el área que se necesita bajo la curva normal se encuentra entre la media y cualquier otro valor. Frecuentemente son valores a lo largo de toda la curva. Por ello, es buena idea hacer un pequeño dibujo de la curva de distribución normal para localizar el o las áreas que se buscan. Esto facilita mucho la visualización del problema y, por lo mismo, su solución. A continuación, se presenta un problema en el que se ilustra esta técnica.

**Ejemplo 3:** Una pequeña población recibe, durante la época de sequía, la dotación de agua potable mediante pipas que surten del líquido a la cisterna del pueblo una vez a la semana. El consumo semanal medio es de 160 metros cúbicos con una desviación estándar de 20 metros cúbicos. Indica cuál será la probabilidad de que el suministro sea suficiente en una semana cualquiera si se surten:

a) 160 metros cúbicos.

b) 180 metros cúbicos.

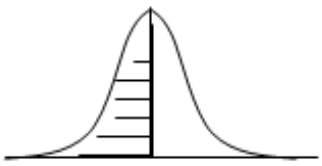
c) 200 metros cúbicos.

d) Indica asimismo cual será la probabilidad de que se acabe el agua si una semana cualquiera surten 190 metros cúbicos.

Solución*:*

a) 160 metros cúbicos.

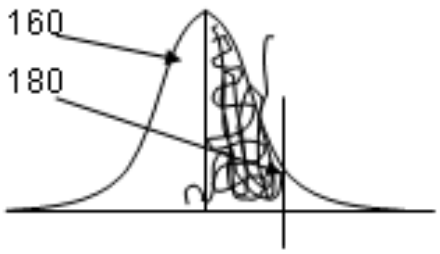
El valor de Z en este caso sería: $z = \frac{160 - 160}{20} = 0.0$ Esto nos puede desconcertar un poco; sin embargo, nos podemos dar cuenta de que si se surten 160 metros cúbicos el agua alcanzará si el consumo es menor que esa cifra. La media está en 160. Por ello el agua alcanzará en toda el área de la curva que se muestra rayada. Es decir, toda la mitad izquierda de la curva. El área de cada una de las mitades de la curva es de 0.5, por tanto, la probabilidad buscada es también de 0.5.



b) 180 metros cúbicos.

El valor de Z es $z = \frac{180 - 160}{20} = 1.0$

El área que se busca es la que está entre la media, 160, y 180. Se marca con una curva en el diagrama Si buscamos el área bajo la curva en la tabla de la normal, para z=1.0, encontraremos el valor de 0.34134. Sin embargo, debemos agregarle toda la mitad izquierda de la curva (que por el diseño de la tabla no aparece). Ese valor, como ya se comentó es de .5. Por tanto, el valor buscado es de .5 más 0.34134. Por ello la probabilidad de que el agua alcance si se surten 180 metros cúbicos es de 0.84134, es decir, aproximadamente el 84%.




| z | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |

c) 200 metros cúbicos.

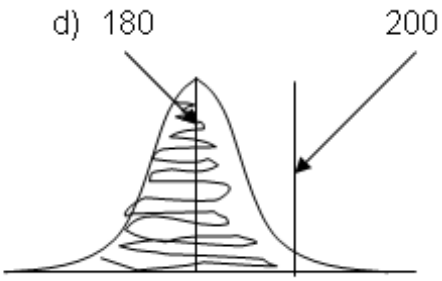
El área buscada se señala en el dibujo. Incluye la primera mitad de la curva y parte de la segunda mitad (la derecha), la que se encuentra entre la media y 200. Ya sabemos que la primera mitad de la curva tiene un área de 0.5. Para

la otra parte tenemos que encontrar el valor de Z y buscar el área correspondiente en la tabla.

Lo que nos lleva a un valor en tablas de 0.47725. Al sumar las dos partes nos

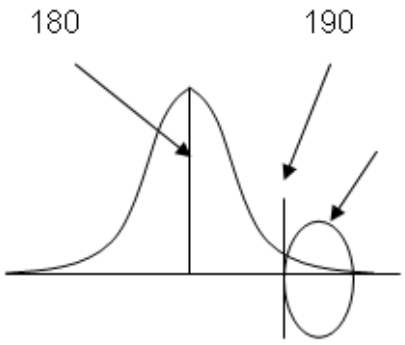
$$Z = \frac{200 - 160}{20} = 2$$

queda .97725. Es decir, si se surten 200 metros cúbicos hay una probabilidad de casi 98% de que el agua alcance.




d) Indica asimismo cual será la probabilidad de que se acabe el agua si una semana cualquiera surten 190 metros cúbicos.

La probabilidad de que se termine el agua en estas condiciones se encuentra representada en la siguiente figura.




La probabilidad de que falte el agua está representada por el área en la cola de la distribución, después del 190. La tabla no nos da directamente ese valor. Para obtenerlo debemos calcular Z para 190 y el valor del área entre la media y 190 restársela a .5 que es el área total de la parte derecha de la curva.

**Z= (190 – 160)/20= 1.5**

El área para Z= 1.5 es de 0.43319. Por tanto, la probabilidad buscada es 0.5000 – 0.43319 = 0.06681 o aproximadamente el 6.7%.

*Búsqueda de Z cuando el área bajo la curva es conocida*

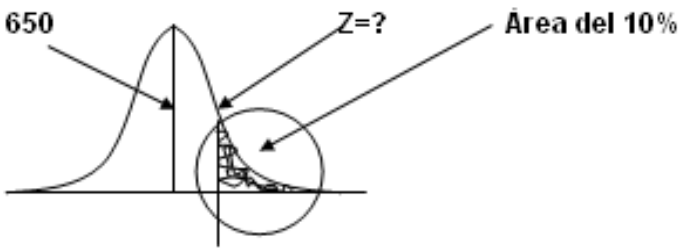
Frecuentemente el problema no es encontrar el área bajo la curva normal mediante el cálculo de Z y el acceso a la tabla para buscar el área ya mencionada. Efectivamente, a veces debemos enfrentar el problema inverso. Conocemos dicha área y deseamos conocer el valor de la variable que lo verifica. El siguiente problema ilustra esta situación.

**Ejemplo 4:** Una universidad realiza un examen de admisión a 10,000 aspirantes para asignar los lugares disponibles. La calificación media de los estudiantes es de 650 puntos sobre 1000 y la desviación estándar es de 100 puntos; las calificaciones siguen una distribución normal. Indica qué calificación mínima deberá de tener un aspirante para ser admitido si:

a) Se aceptará al 10% de los aspirantes con mejor calificación.
b) Se aceptará al 5% de aspirantes con mejor calificación.

Solución:

a) Si hacemos un pequeño esquema de la curva normal, los aspirantes aceptados representan el 10% del área que se acumula en la cola derecha de la distribución. El siguiente esquema nos dará una mejor idea.



El razonamiento que se hace es el siguiente:

Si el área que se busca es el 10% de la cola derecha, entonces el área que debemos de buscar en la tabla es lo más cercano posible al 40%, esto es 0.4000 (esto se busca en el cuerpo de la tabla, no en los encabezados que representan el valor de Z). Éste es el valor de 0.39973 y se encuentra en el renglón donde aparece un valor para Z de 1.2 y en la columna de 0.08. Eso quiere decir que el valor de Z que más se aproxima es el de 1.28. No importa si al valor de la tabla le falta un poco o se pasa un poco; la idea es que sea el más cercano posible.

| z | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.2 | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | **0.3997** | 0.4015 |

Si ya sabemos el valor de Z, calcular el valor de la calificación (es decir “x”) es un problema de álgebra elemental y se trabaja despejando la fórmula de Z, tal como se indica a continuación.

Partimos de la relación

$$Z= (X – 650) / 100$$

Observa que ya sustituimos los valores de la media y de la desviación estándar. Ahora sustituimos el valor de Z y nos queda:

$$1.28 = (X – 650) / 100$$

A continuación despejamos el valor de x

$$1.28 (100) = X – 650$$

$$128 + 650 = X$$

$$X = 778$$

En estas condiciones los aspirantes comenzarán a ser admitidos a partir de la calificación de 778 puntos en su examen de admisión.

b) El razonamiento es análogo al del inciso a. Solamente que ahora no buscamos que el área de la cola derecha sea el 10% del total sino solamente el 5% del mismo. Esto quiere decir que debemos buscar en la tabla en complemento del 5%, es decir, 45% o 0.45000. Vemos que el valor más cercano se encuentra en el renglón de Z de 1.6 y en la centésima 0.04

| z | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **1.6** | 0.4452 | 0.4463 | 0.4474 | 0.4485 | **0.4495** | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |

Esto nos indica que el valor de z que buscamos es el de 1.64. El despeje de x se lleva a efecto de manera análoga al inciso anterior, tal como a continuación se muestra.

$$1.64 = (X - 650)/100$$

$$1.64\ (100) = X - 650$$

$$164 + 650 = X$$

$$X = 814$$

En caso de que se desee mayor precisión se puede recurrir a interpolar los valores (por ejemplo, en este caso entre 1.64 y 1.65) o buscar valores más precisos en paquetes estadísticos de cómputo.

## 3.4.2. Distribución exponencial

En una distribución de Poisson los eventos ocurren en un espectro continuo de tiempo o espacio. Se considera entonces que son eventos sucesivos de modo tal que la longitud o tiempo que transcurre entre cada realización del evento es una variable aleatoria, cuya distribución de probabilidades recibe precisamente el nombre de distribución exponencial.

Ésta se aplica cuando estamos interesados en el tiempo o espacio hasta el «primer evento», el tiempo entre dos eventos sucesivos o el tiempo hasta que ocurra el primer evento después de cualquier punto aleatoriamente seleccionado. Así, se presentan dos casos:

a) La probabilidad de que el primer evento ocurra dentro del intervalo de interés. Su fórmula es: $P(T \leq t) = 1 - e^{-\lambda}$

b) La probabilidad de que el primer evento *no* ocurra dentro del intervalo de interés. Su fórmula es: $P(T > t) = e^{-\lambda}$

Donde λ es la tasa promedio de eventos por unidad de tiempo o longitud, según se trate. Como en el caso de la distribución de Poisson, su parámetro es precisamente esta tasa promedio.

**Ejemplo 1:** Un departamento de reparaciones recibe un promedio de 15 llamadas por hora. A partir de este momento, cuál es la probabilidad de que:

a) En los siguientes 5 minutos no se reciba ninguna llamada.

b) Que la primera llamada ocurra dentro de esos 5 minutos.

c) En una tabla indicar las probabilidades de ocurrencia de la primera llamada en el minuto 1, 5, 10, 15, y 30.

Solución:

a)No se reciba ninguna llamada:

Como la tasa promedio está expresada en llamadas por hora y en la pregunta se hace referencia a un periodo de 5 minutos, primero debemos hacer compatibles las unidades. Para ello, establecemos una relación de proporcionalidad directa.

$$15 - 60$$
$$\lambda - \quad 5$$
$$\therefore \quad \lambda = 1.25$$

$$P(T > t) = e^{-\lambda}$$
$$p(t > 5) = 2.71828^{-1.25} = 0.286 = 28.6\%$$

$$e = 2.71828$$

b) Primera llamada en 5 minutos:

$$P(T \leq t) = 1 - e^{-\lambda}$$
$$P(T \leq 5) = 1 - 2.71828^{-1.25} = 1 - 0.287 = 0.713 = 71.3\%$$

c) Primera llamada en 1, 5, 10, 15 y 30 minutos:

| Espacio tiempo | $\lambda$ | Probabilidad ocurra | Probabilidad No ocurra |
|---|---|---|---|
| 1 minuto | 0.25 | 0.221 | 0.779 |
| 5 minutos | 1.25 | 0.713 | 0.287 |
| 10 minutos | 2.50 | 0.918 | 0.082 |
| 15 minutos | 3.75 | 0.976 | 0.024 |
| 30 minutos | 7.50 | 0.999 | 0.001 |

# 3.5. Ley de los grandes números

La ley de los grandes números sugiere que **la probabilidad de una desviación significativa de un valor de probabilidad determinado empíricamente, a partir de uno determinado teóricamente, es menor cuanto más grande sea el número de repeticiones del experimento**.

Esta ley forma parte de lo que en la probabilidad se conoce como teoremas de límites, uno de los cuales es el teorema de De Moivre-Laplace según el cual, la distribución binomial —que se presenta en múltiples casos en los que se requiere conocer la probabilidad de ocurrencia de un número determinado de éxitos en una muestra aleatoriamente seleccionada— puede aproximarse por la distribución normal si el número de ensayos es suficientemente grande y donde el error en la aproximación disminuye en la medida en que la probabilidad de éxito se acerca a 0.5.

Desde el punto de vista de las operaciones, si lo que deseamos es calcular la probabilidad de que una variable aleatoria binomial con parámetros n y p tome valores entre a y b, entonces debemos:

- Determinar la media y desviación estándar de la variable binomial, esto es, calcular los valores de $\mu = np$, y $\sigma = (npq)^{1/2}$
- Reformular la probabilidad deseada en el contexto binomial por la probabilidad deseada en el contexto de la distribución normal, incorporando una corrección por finitud, esto es, si nuestra pregunta original es determinar el valor de $P(a \leq X \leq b)$ entonces, buscaremos aplicar la distribución normal para calcular:

$$P\left(\left[\frac{a-0.5-np}{\sqrt{npq}}\right] \leq Z \leq \left[\frac{b+0.5-np}{\sqrt{npq}}\right]\right)$$

donde los sumandos 0.5 y –0.5 constituyen la corrección por finitud.

- Emplear la tabla de la distribución normal.

Veamos un ejemplo.

**Ejemplo 1.** Se arroja una moneda legal 200 veces. Se desea saber la probabilidad de que aparezca sol más de 110 veces pero menos de 130.

Solución:

El hecho de que la moneda sea legal significa que la probabilidad de que el resultado sea sol es igual a la probabilidad de que salga águila, de modo que tanto la probabilidad de éxito como de fracaso es 0.5, y esta probabilidad no cambia de ensayo a ensayo. Podemos decir entonces que estamos en presencia de un experimento Binomial, de modo que podemos plantear el problema en los siguientes términos, donde S es la variable aleatoria que denota el número de soles:

$$P(110 < S < 130) = P(S= 111) + P(S= 112) + P(S= 113) + \ldots + P(S= 129)$$

$$= \sum_{i=111}^{129} {}_{200}C_i (0.5)^i (0.5)^{200-i}$$

El problema es que tendríamos dificultades al hacer las operaciones incluso con una calculadora. Es aquí donde resulta útil aplicar la distribución normal como aproximación a la distribución binomial.

Como n=200 y p=0.5, entonces la media es μ = 200(0.5) = 100, en tanto que la varianza es $\sigma^2$ = 200(0.5)(0.5) = 50, de modo que la desviación estándar es σ = 7.07.

En consecuencia,

$$P(111 \leq X \leq 129) = P[(110.5 - 100) / 7.07 \leq Z \leq (129.5 - 100) / 7.07]$$

$$= P(10.5 / 7.07 \leq Z \leq 29.5 / 7.07)$$

$= P\ (1.49 \leq Z \leq 4.17)$

$= 0.5 - 0.4319$

$= 0.0681$

Cuando el número de ensayos es grande pero el valor de la probabilidad de éxito se acerca a cero o a uno, esto es se aleja de 0.5, es mejor emplear la distribución de Poisson como aproximación a la binomial.

# DISTRIBUCIÓN NORMAL ESTÁNDAR (ÁREA BAJO LA CURVA)

| z | 0.00000 | 0.01000 | 0.02000 | 0.03000 | 0.04000 | 0.05000 | 0.06000 | 0.07000 | 0.08000 | 0.09000 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.00000 | 0.00399 | 0.00798 | 0.01197 | 0.01595 | 0.01994 | 0.02392 | 0.02790 | 0.03188 | 0.03586 |
| 0.1 | 0.03983 | 0.04380 | 0.04776 | 0.05172 | 0.05567 | 0.05962 | 0.06356 | 0.06749 | 0.07142 | 0.07535 |
| 0.2 | 0.07926 | 0.08317 | 0.08706 | 0.09095 | 0.09483 | 0.09871 | 0.10257 | 0.10642 | 0.11026 | 0.11409 |
| 0.3 | 0.11791 | 0.12172 | 0.12552 | 0.12930 | 0.13307 | 0.13683 | 0.14058 | 0.14431 | 0.14803 | 0.15173 |
| 0.4 | 0.15542 | 0.15910 | 0.16276 | 0.16640 | 0.17003 | 0.17364 | 0.17724 | 0.18082 | 0.18439 | 0.18793 |
| 0.5 | 0.19146 | 0.19497 | 0.19847 | 0.20194 | 0.20540 | 0.20884 | 0.21226 | 0.21566 | 0.21904 | 0.22240 |
| 0.6 | 0.22575 | 0.22907 | 0.23237 | 0.23565 | 0.23891 | 0.24215 | 0.24537 | 0.24857 | 0.25175 | 0.25490 |
| 0.7 | 0.25804 | 0.26115 | 0.26424 | 0.26730 | 0.27035 | 0.27337 | 0.27637 | 0.27935 | 0.28230 | 0.28524 |
| 0.8 | 0.28814 | 0.29103 | 0.29389 | 0.29673 | 0.29955 | 0.30234 | 0.30511 | 0.30785 | 0.31057 | 0.31327 |
| 0.9 | 0.31594 | 0.31859 | 0.32121 | 0.32381 | 0.32639 | 0.32894 | 0.33147 | 0.33398 | 0.33646 | 0.33891 |
| 1 | 0.34134 | 0.34375 | 0.34614 | 0.34849 | 0.35083 | 0.35314 | 0.35543 | 0.35769 | 0.35993 | 0.36214 |
| 1.1 | 0.36433 | 0.36650 | 0.36864 | 0.37076 | 0.37286 | 0.37493 | 0.37698 | 0.37900 | 0.38100 | 0.38298 |
| 1.2 | 0.38493 | 0.38686 | 0.38877 | 0.39065 | 0.39251 | 0.39435 | 0.39617 | 0.39796 | 0.39973 | 0.40147 |
| 1.3 | 0.40320 | 0.40490 | 0.40658 | 0.40824 | 0.40988 | 0.41149 | 0.41308 | 0.41466 | 0.41621 | 0.41774 |
| 1.4 | 0.41924 | 0.42073 | 0.42220 | 0.42364 | 0.42507 | 0.42647 | 0.42785 | 0.42922 | 0.43056 | 0.43189 |
| 1.5 | 0.43319 | 0.43448 | 0.43574 | 0.43699 | 0.43822 | 0.43943 | 0.44062 | 0.44179 | 0.44295 | 0.44408 |
| 1.6 | 0.44520 | 0.44630 | 0.44738 | 0.44845 | 0.44950 | 0.45053 | 0.45154 | 0.45254 | 0.45352 | 0.45449 |
| 1.7 | 0.45543 | 0.45637 | 0.45728 | 0.45818 | 0.45907 | 0.45994 | 0.46080 | 0.46164 | 0.46246 | 0.46327 |
| 1.8 | 0.46407 | 0.46485 | 0.46562 | 0.46638 | 0.46712 | 0.46784 | 0.46856 | 0.46926 | 0.46995 | 0.47062 |
| 1.9 | 0.47128 | 0.47193 | 0.47257 | 0.47320 | 0.47381 | 0.47441 | 0.47500 | 0.47558 | 0.47615 | 0.47670 |
| 2 | 0.47725 | 0.47778 | 0.47831 | 0.47882 | 0.47932 | 0.47982 | 0.48030 | 0.48077 | 0.48124 | 0.48169 |
| 2.1 | 0.48214 | 0.48257 | 0.48300 | 0.48341 | 0.48382 | 0.48422 | 0.48461 | 0.48500 | 0.48537 | 0.48574 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.2 | 0.48610 | 0.48645 | 0.48679 | 0.48713 | 0.48745 | 0.48778 | 0.48809 | 0.48840 | 0.48870 | 0.48899 |
| 2.3 | 0.48928 | 0.48956 | 0.48983 | 0.49010 | 0.49036 | 0.49061 | 0.49086 | 0.49111 | 0.49134 | 0.49158 |
| 2.4 | 0.49180 | 0.49202 | 0.49224 | 0.49245 | 0.49266 | 0.49286 | 0.49305 | 0.49324 | 0.49343 | 0.49361 |
| 2.5 | 0.49379 | 0.49396 | 0.49413 | 0.49430 | 0.49446 | 0.49461 | 0.49477 | 0.49492 | 0.49506 | 0.49520 |
| 2.6 | 0.49534 | 0.49547 | 0.49560 | 0.49573 | 0.49585 | 0.49598 | 0.49609 | 0.49621 | 0.49632 | 0.49643 |
| 2.7 | 0.49653 | 0.49664 | 0.49674 | 0.49683 | 0.49693 | 0.49702 | 0.49711 | 0.49720 | 0.49728 | 0.49736 |
| 2.8 | 0.49744 | 0.49752 | 0.49760 | 0.49767 | 0.49774 | 0.49781 | 0.49788 | 0.49795 | 0.49801 | 0.49807 |
| 2.9 | 0.49813 | 0.49819 | 0.49825 | 0.49831 | 0.49836 | 0.49841 | 0.49846 | 0.49851 | 0.49856 | 0.49861 |
| 3 | 0.49865 | 0.49869 | 0.49874 | 0.49878 | 0.49882 | 0.49886 | 0.49889 | 0.49893 | 0.49896 | 0.49900 |
| 3.1 | 0.49903 | 0.49906 | 0.49910 | 0.49913 | 0.49916 | 0.49918 | 0.49921 | 0.49924 | 0.49926 | 0.49929 |
| 3.2 | 0.49931 | 0.49934 | 0.49936 | 0.49938 | 0.49940 | 0.49942 | 0.49944 | 0.49946 | 0.49948 | 0.49950 |
| 3.3 | 0.49952 | 0.49953 | 0.49955 | 0.49957 | 0.49958 | 0.49960 | 0.49961 | 0.49962 | 0.49964 | 0.49965 |
| 3.4 | 0.49966 | 0.49968 | 0.49969 | 0.49970 | 0.49971 | 0.49972 | 0.49973 | 0.49974 | 0.49975 | 0.49976 |
| 3.5 | 0.49977 | 0.49978 | 0.49978 | 0.49979 | 0.49980 | 0.49981 | 0.49981 | 0.49982 | 0.49983 | 0.49983 |
| 3.6 | 0.49984 | 0.49985 | 0.49985 | 0.49986 | 0.49986 | 0.49987 | 0.49987 | 0.49988 | 0.49988 | 0.49989 |
| 3.7 | 0.49989 | 0.49990 | 0.49990 | 0.49990 | 0.49991 | 0.49991 | 0.49992 | 0.49992 | 0.49992 | 0.49992 |
| 3.8 | 0.49993 | 0.49993 | 0.49993 | 0.49994 | 0.49994 | 0.49994 | 0.49994 | 0.49995 | 0.49995 | 0.49995 |
| 3.9 | 0.49995 | 0.49995 | 0.49996 | 0.49996 | 0.49996 | 0.49996 | 0.49996 | 0.49996 | 0.49997 | 0.49997 |
| 4 | 0.49997 | 0.49997 | 0.49997 | 0.49997 | 0.49997 | 0.49997 | 0.49998 | 0.49998 | 0.49998 | 0.49998 |
| 4.1 | 0.49998 | 0.49998 | 0.49998 | 0.49998 | 0.49998 | 0.49998 | 0.49998 | 0.49998 | 0.49999 | 0.49999 |
| 4.2 | 0.49999 | 0.49999 | 0.49999 | 0.49999 | 0.49999 | 0.49999 | 0.49999 | 0.49999 | 0.49999 | 0.49999 |

# RESUMEN

Se define el concepto de variable aleatoria y se señalan sus diferentes tipos. Asimismo, se presentan los rasgos que permiten distinguir algunos modelos de distribución probabilística de variables aleatorias, tipificando los mismos a través de las expresiones analíticas de la función de probabilidad y de densidad, su esperanza matemática, su varianza y sus parámetros. Además, en el caso de la distribución normal se presenta el concepto de distribución normal estándar y se muestra el manejo de las tablas respectivas, así como el uso de esta distribución por cuanto aproximación al modelo binomial.

# BIBLIOGRAFÍA

## SUGERIDA

| Autor | Capítulo | Páginas |
|---|---|---|
| 1. Anderson, Sweeney, Williams. (2005) | 5. Distribuciones discretas de probabilidad.<br>Sección 5.3 Valor esperado y varianza. | 184-186 |
| | 5.4 Distribución de probabilidad binomial. | 189-197 |
| | 5.5 Distribución de probabilidad de Poisson. | 199-201 |
| | 5.6 Distribución de probabilidad hipergeométrica. | 203-204 |
| | 6. Distribuciones continuas de probabilidad.<br>Sección 6.2 Distribución de probabilidad normal. | 218-229 |
| | Sección 6.3 Distribución de probabilidad exponencial. | 232-234 |

| | | |
|---|---|---|
| 2. Berenson, Levine y Krehbiel. (2001) | 4. Probabilidad básica y distribuciones de probabilidad.<br>Sección 4.4 Distribución de probabilidad para una variable aleatoria. | 179-186 |
| | 4.5 Distribución binomial. | 186-194 |
| | 4.6 Distribución de Poisson. | 194-197 |
| | 4.7 Distribución normal. | 198-219 |
| 3. Hernández, Fernández, Baptista. (2006) | 6. Formación de hipótesis.<br>Sección: Definición conceptual o constitutiva. | 145-146 |
| 4. Levin y Rubin. (2004) | 5. Distribuciones de probabilidad.<br>Sección 5.1 ¿Qué es una distribución de probabilidad? | 178-181 |
| | 5.2 Variable aleatoria. | 181-187 |
| | 5.4 La distribución binomial. | 191-202 |
| | 5.5 La distribución de Poisson. | 202-208 |
| | 5.6 La distribución normal: distribución de una variable aleatoria continua. | 209-222 |
| 5. Lind, Marchal, Wathen. (2008) | 6. Distribuciones discretas de de probabilidad.<br>Secciones: ¿Qué es una distribución de probabilidad? | 181-183 |
| | Variables aleatorias. | 183-185 |

| | | |
|---|---|---|
| | Media, varianza y desviación estándar de una distribución de probabilidad. | 185-187 |
| | Distribución de probabilidad binomial. | 189-199 |
| | Distribución de probabilidad hipergeométrica. | 199-203 |
| | Distribución de probabilidad de Poisson. | 203-207 |
| | 7. Distribuciones de probabilidad continua.<br>Secciones: La familia de distribuciones de probabilidad normal. | 227-229 |
| | Distribución de probabilidad normal estándar. | 229-233 |
| | Determinación de áreas bajo la curva normal. | 233-237 |

Anderson, David R., Sweeney, Dennis J., Williams, Thomas A. (2005). *Estadística para administración y economía* (8ª. Edición). México: International Thomson Editores, 888 pp. más apéndices.

Berenson, Mark L., David M. Levine, y Timothy C. Krehbiel (2001). *Estadística para administración* (2ª Edición). México: Prentice Hall, 734 pp.

Hernández Sampieri, R., C. Fernández Collado, Lucio P Baptista (2006). *Metodología de la investigación* (4ª edición). México: McGraw-Hill Interamericana, 850 pp.

Levin, Richard I. y David S. Rubin. (2004). *Estadística para administración y economía* (7ª. Edición). México: Pearson Educación Prentice Hall, 826 pp. más anexos.

Lind, Douglas A., Marchal, William G.,Wathen, Samuel, A. (2008). *Estadística aplicada a los negocios y la economía* **(**13ª edición). México: McGraw-Hill Interamericana, 859 pp.

# UNIDAD 4

# Distribuciones muestrales


SUAyED UNAM
Sistema Universidad Abierta y Educación a Distancia

Plan 2012 2016 actualizado

# OBJETIVO PARTICULAR

El alumno identificará e interpretará los diferentes tipos de distribuciones muestrales.

# TEMARIO DETALLADO
## (8 horas)

**4. Distribuciones muestrales**

- 4.1. La distribución muestral de la media
- 4.2. El teorema central del límite
- 4.3. La distribución muestral de la proporción
- 4.4. La distribución muestral de la varianza

# INTRODUCCIÓN

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

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

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

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

# 4.1. La distribución muestral de la media

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

Variable aleatoria

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

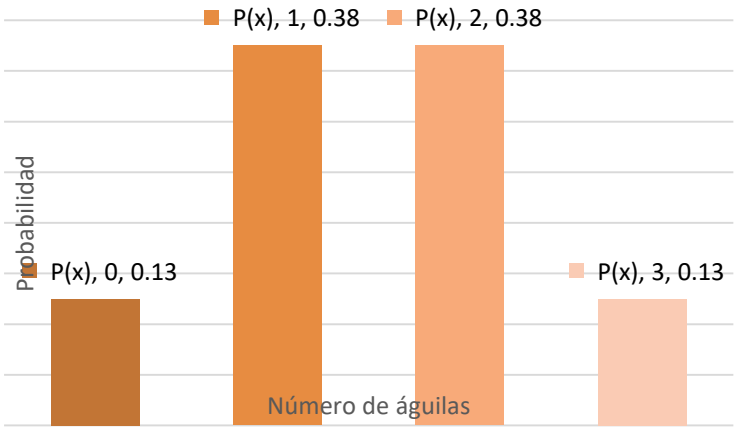
Asimismo, cada valor de la variable aleatoria tiene asociada una probabilidad de ocurrencia, que en conjunto conforman la distribución de probabilidades o simplemente la distribución de la variable aleatoria.

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

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

= 0 y X = 3; 3/8 para X = 1 y X = 2. La distribución de *X* se muestra en la siguiente figura.

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




Fuente: elaboración propia.

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

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

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

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

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

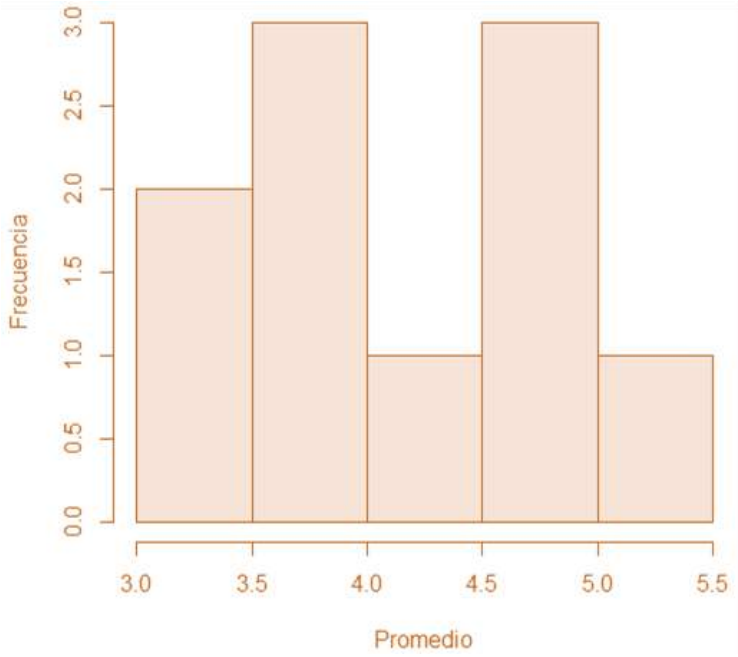
Si se extrae una muestra de tres empleados (sin reemplazo) y se calcula su promedio de antigüedad, hay $\binom{5}{3} = \mathbf{10}$ posibles resultados, los cuales se detallan en la tabla 2.

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

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

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

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




Fuente: elaboración propia.

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

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

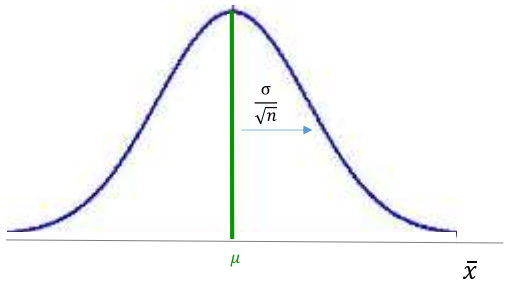
En el ejemplo anterior, la distribución muestral de la media es bimodal, lo que se debe a la poca información y dispersión de datos. ¿Si la población hubiera sido de mayor tamaño o la muestra hubiera permitido repeticiones, la distribución se habría conservado? La respuesta es no.

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

**Distribución muestral de la media**

Supóngase que se tiene una población de tamaño *N* con media μ y varianza $\sigma^2$ de la que se extrae una muestra de tamaño *n*. La distribución de la media muestral ($\bar{x}$) se aproxima a una normal con media μ y varianza $\sigma^2/n$ (figura3) en la medida que se incrementa el tamaño de la muestra (n).[1]

**Figura 3. Distribución muestral de la media**




Fuente: elaboración propia.

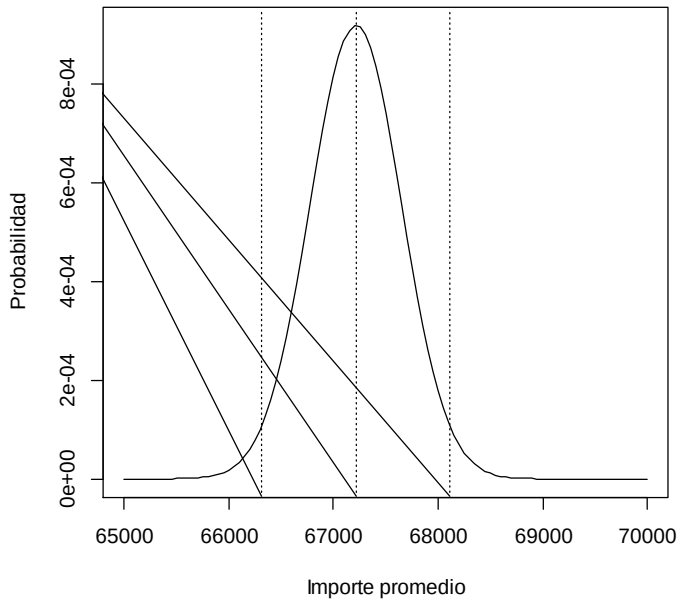
Conociendo lo anterior, puede estandarizarse esta distribución y utilizar el cálculo de una probabilidad para medir la calidad de la muestra, lo cual se ejemplifica a continuación.

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

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

Conforme a lo expuesto, la distribución muestral del promedio se aproxima a una distribución normal con media de $67,213.49 y una desviación de $\frac{\$5{,}315.22}{\sqrt{150}}$. Se busca la probabilidad de que el promedio muestral se encuentre entre $67,213.49 ± $900. En la figura 3 se muestra la región de interés.

**Figura 4. Distribución del promedio muestral de los movimientos bancarios**




Fuente: elaboración propia.

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

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

De esta manera:

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

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

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

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

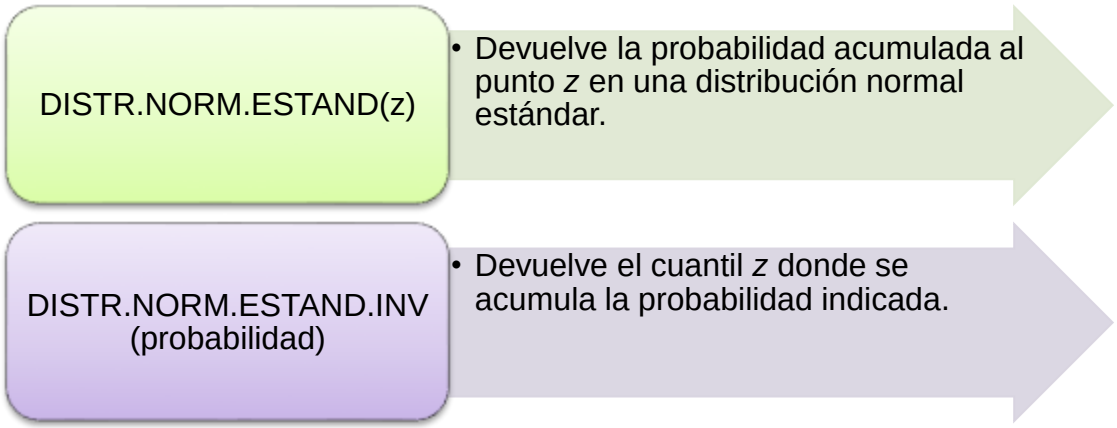
Entonces, la probabilidad buscada se calcula así:

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

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

Observación

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




**Distribución muestral de la media cuando se desconoce $\sigma^2$**

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

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

- Donde:

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

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

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

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

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

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

Función de densidad de la distribución t de Student:

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

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

- Donde:

$t_n$ = valor *t* con *n* grados de libertad
$\Gamma$ = función gamma
N = grados de libertad

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

Distr.t(x, grados de libertad, colas).

Calcula la probabilidad acumulada a partir del cuantil X considerando una o dos colas en una distribución *t* con los grados de libertad.

Distr.t(probabilidad, grados de libertad).

Calcula el cuantil a partir del cual se acumula la probabilidad de interés de una distribución *t* de dos colas, con los grados de libertad establecidos.

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

El primer paso es calcular el promedio muestral:

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

A continuación, se calcula la varianza muestral:

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

= 2,584,361.5

Por tanto, la desviación muestral es:

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

A continuación, se estandarizan los datos:

$$P(66,313.49 < X < 68,113.49)$$

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

$$P(-1.252 < t_4 < 1.252)$$

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

Entonces, la probabilidad buscada se calcula así:

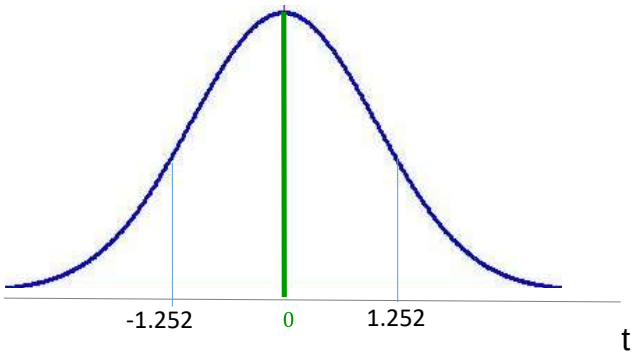
(1-Distr.t(1.252,4, 2)**)** = 0.7212

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

Observación:

La función Distr.t(1.252,4, 2)

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

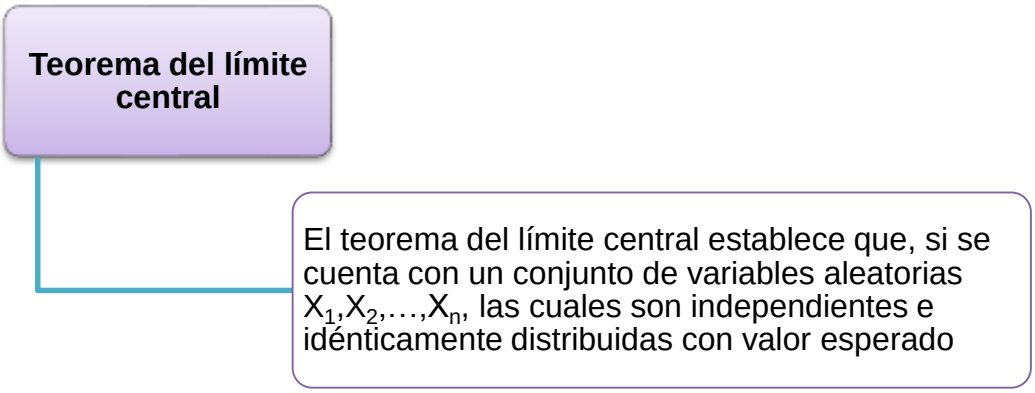



Fuente: elaboración propia.

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

# 4.2. El teorema central del límite

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




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

y varianza

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

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

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

- Donde:

$\bar{X}_n$ = Promedio de n variables
$N(\mu, \frac{\sigma^2}{n})$ = Distribución normal con media μ y varianza $\sigma^2$/ n

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

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

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

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

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

Y la varianza ($\sigma^2$):

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

Donde:

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

Así:

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

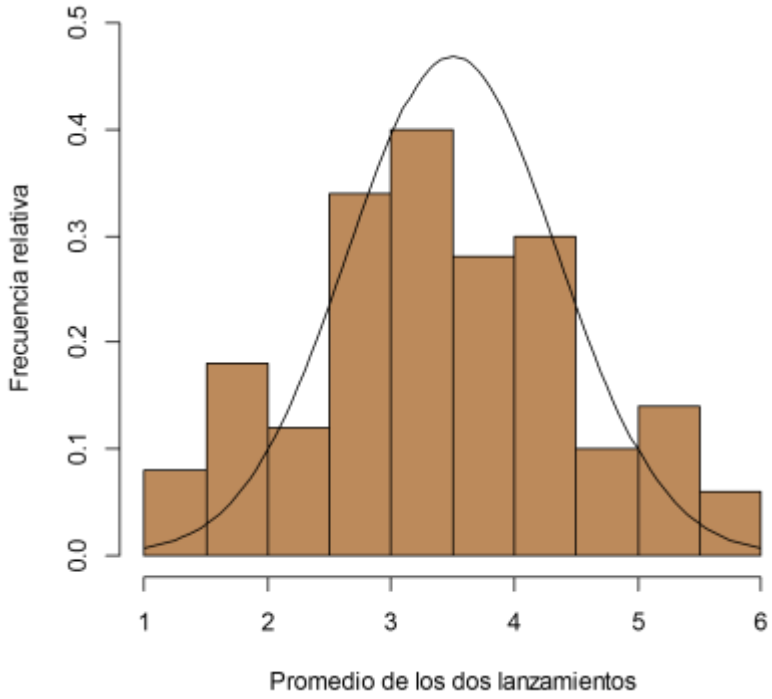
Supóngase que se lanza el dado dos veces (n = 2) y se calcula el promedio de los dos resultados y se repite este experimento 100 ocasiones (N = 100). Se obtienen los resultados que se muestran en la tabla siguiente.

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

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

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

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




Fuente: elaboración propia con empleo del paquete estadístico R.[2]

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

---

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

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

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

| | Lanzamiento | | | | | |
|---|---|---|---|---|---|---|
| Muestra | 1 | 2 | 3 | 4 | 5 | Promedio |
| 1 | 3 | 3 | 5 | 2 | 3 | 3.2 |
| 2 | 4 | 4 | 3 | 2 | 5 | 3.6 |
| 3 | 1 | 1 | 5 | 2 | 6 | 3 |
| 4 | 1 | 5 | 6 | 6 | 3 | 4.2 |
| 5 | 3 | 2 | 3 | 2 | 3 | 2.6 |
| 6 | 5 | 4 | 4 | 5 | 5 | 4.6 |
| 7 | 3 | 6 | 5 | 1 | 2 | 3.4 |
| 8 | 5 | 6 | 3 | 4 | 6 | 4.8 |
| 9 | 3 | 3 | 2 | 2 | 5 | 3 |
| 10 | 3 | 3 | 3 | 3 | 4 | 3.2 |
| 11 | 3 | 4 | 5 | 2 | 1 | 3 |
| 12 | 1 | 5 | 4 | 4 | 3 | 3.4 |
| 13 | 3 | 2 | 2 | 5 | 3 | 3 |
| 14 | 2 | 5 | 6 | 1 | 1 | 3 |
| 15 | 1 | 6 | 1 | 1 | 5 | 2.8 |
| 16 | 2 | 3 | 3 | 2 | 5 | 3 |
| 17 | 2 | 1 | 3 | 1 | 6 | 2.6 |
| 18 | 6 | 5 | 2 | 6 | 3 | 4.4 |
| 19 | 1 | 5 | 5 | 3 | 5 | 3.8 |
| 20 | 3 | 3 | 1 | 4 | 2 | 2.6 |
| 21 | 4 | 6 | 4 | 5 | 1 | 4 |
| 22 | 5 | 1 | 4 | 4 | 1 | 3 |
| 23 | 6 | 3 | 5 | 4 | 1 | 3.8 |
| 24 | 5 | 1 | 5 | 4 | 6 | 4.2 |
| 25 | 2 | 4 | 5 | 3 | 1 | 3 |
| 26 | 1 | 5 | 6 | 5 | 6 | 4.6 |
| 27 | 1 | 3 | 4 | 3 | 5 | 3.2 |
| 28 | 6 | 5 | 3 | 6 | 2 | 4.4 |
| 29 | 4 | 6 | 4 | 5 | 4 | 4.6 |
| 30 | 5 | 6 | 2 | 4 | 6 | 4.6 |
| 31 | 6 | 6 | 2 | 3 | 2 | 3.8 |
| 32 | 4 | 6 | 5 | 4 | 2 | 4.2 |
| 33 | 2 | 3 | 1 | 4 | 6 | 3.2 |
| 34 | 4 | 3 | 2 | 5 | 2 | 3.2 |
| 35 | 2 | 2 | 5 | 1 | 3 | 2.6 |
| 36 | 2 | 6 | 5 | 1 | 1 | 3 |
| 37 | 4 | 4 | 2 | 4 | 4 | 3.6 |

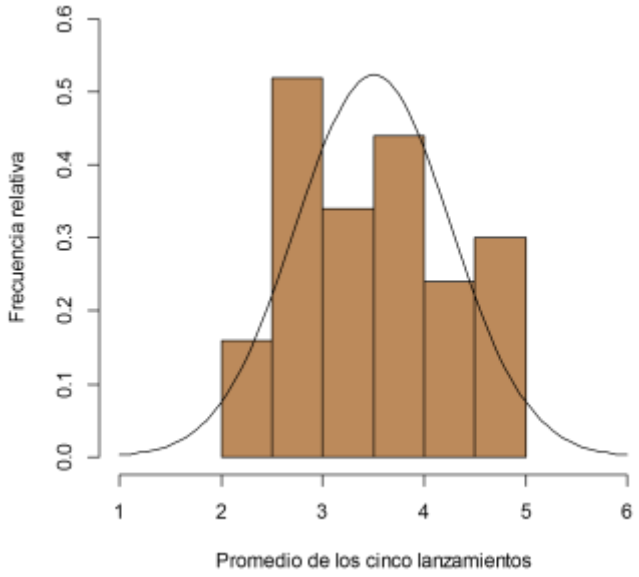
| | Lanzamiento | | | | | |
|---|---|---|---|---|---|---|
| Muestra | 1 | 2 | 3 | 4 | 5 | Promedio |
| 51 | 1 | 4 | 6 | 2 | 1 | 2.8 |
| 52 | 5 | 5 | 1 | 1 | 2 | 2.8 |
| 53 | 4 | 3 | 5 | 1 | 2 | 3 |
| 54 | 5 | 4 | 4 | 1 | 6 | 4 |
| 55 | 6 | 1 | 4 | 1 | 4 | 3.2 |
| 56 | 5 | 3 | 5 | 2 | 2 | 3.4 |
| 57 | 2 | 6 | 5 | 2 | 6 | 4.2 |
| 58 | 3 | 1 | 6 | 3 | 3 | 3.2 |
| 59 | 4 | 4 | 3 | 5 | 6 | 4.4 |
| 60 | 2 | 1 | 4 | 2 | 3 | 2.4 |
| 61 | 1 | 6 | 4 | 1 | 3 | 3 |
| 62 | 3 | 6 | 6 | 4 | 4 | 4.6 |
| 63 | 5 | 1 | 1 | 2 | 3 | 2.4 |
| 64 | 1 | 3 | 2 | 1 | 5 | 2.4 |
| 65 | 6 | 1 | 6 | 1 | 4 | 3.6 |
| 66 | 5 | 6 | 1 | 5 | 1 | 3.6 |
| 67 | 2 | 4 | 3 | 5 | 5 | 3.8 |
| 68 | 3 | 4 | 2 | 6 | 4 | 3.8 |
| 69 | 3 | 1 | 6 | 3 | 3 | 3.2 |
| 70 | 4 | 4 | 6 | 6 | 4 | 4.8 |
| 71 | 2 | 4 | 4 | 2 | 1 | 2.6 |
| 72 | 6 | 5 | 6 | 3 | 4 | 4.8 |
| 73 | 2 | 6 | 5 | 6 | 6 | 5 |
| 74 | 5 | 3 | 2 | 2 | 3 | 3 |
| 75 | 1 | 5 | 5 | 2 | 3 | 3.2 |
| 76 | 6 | 2 | 6 | 4 | 5 | 4.6 |
| 77 | 5 | 1 | 6 | 3 | 3 | 3.6 |
| 78 | 5 | 5 | 1 | 4 | 1 | 3.2 |
| 79 | 5 | 5 | 2 | 1 | 5 | 3.6 |
| 80 | 3 | 3 | 1 | 2 | 3 | 2.4 |
| 81 | 2 | 5 | 2 | 5 | 6 | 4 |
| 82 | 2 | 4 | 6 | 5 | 6 | 4.6 |
| 83 | 1 | 6 | 3 | 1 | 4 | 3 |
| 84 | 6 | 2 | 6 | 2 | 5 | 4.2 |
| 85 | 1 | 1 | 2 | 6 | 1 | 2.2 |
| 86 | 2 | 5 | 5 | 1 | 1 | 2.8 |
| 87 | 3 | 2 | 5 | 2 | 1 | 2.6 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 38 | 6 | 1 | 1 | 3 | 2 | 2.6 |
| 39 | 4 | 4 | 6 | 2 | 3 | 3.8 |
| 40 | 5 | 1 | 1 | 4 | 5 | 3.2 |
| 41 | 1 | 3 | 2 | 4 | 1 | 2.2 |
| 42 | 6 | 1 | 2 | 5 | 2 | 3.2 |
| 43 | 6 | 3 | 3 | 4 | 6 | 4.4 |
| 44 | 6 | 5 | 1 | 4 | 2 | 3.6 |
| 45 | 4 | 4 | 6 | 6 | 5 | 5 |
| 46 | 3 | 5 | 1 | 2 | 4 | 3 |
| 47 | 5 | 3 | 6 | 2 | 6 | 4.4 |
| 48 | 6 | 4 | 4 | 4 | 2 | 4 |
| 49 | 4 | 2 | 6 | 6 | 2 | 4 |
| 50 | 3 | 5 | 6 | 6 | 4 | 4.8 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 88 | 2 | 3 | 2 | 3 | 6 | 3.2 |
| 89 | 3 | 1 | 1 | 6 | 1 | 2.4 |
| 90 | 4 | 6 | 4 | 3 | 6 | 4.6 |
| 91 | 1 | 1 | 2 | 2 | 5 | 2.2 |
| 92 | 3 | 6 | 6 | 1 | 6 | 4.4 |
| 93 | 5 | 1 | 1 | 5 | 6 | 3.6 |
| 94 | 4 | 1 | 1 | 6 | 6 | 3.6 |
| 95 | 1 | 1 | 3 | 5 | 5 | 3 |
| 96 | 6 | 5 | 4 | 1 | 4 | 4 |
| 97 | 6 | 3 | 5 | 4 | 5 | 4.6 |
| 98 | 3 | 3 | 6 | 6 | 4 | 4.4 |
| 99 | 5 | 3 | 2 | 6 | 1 | 3.4 |
| 100 | 1 | 4 | 4 | 6 | 3 | 3.6 |
| | | | | | Promedio: | 3.5 |
| | | | | | Varianza: | 0.6 |

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

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

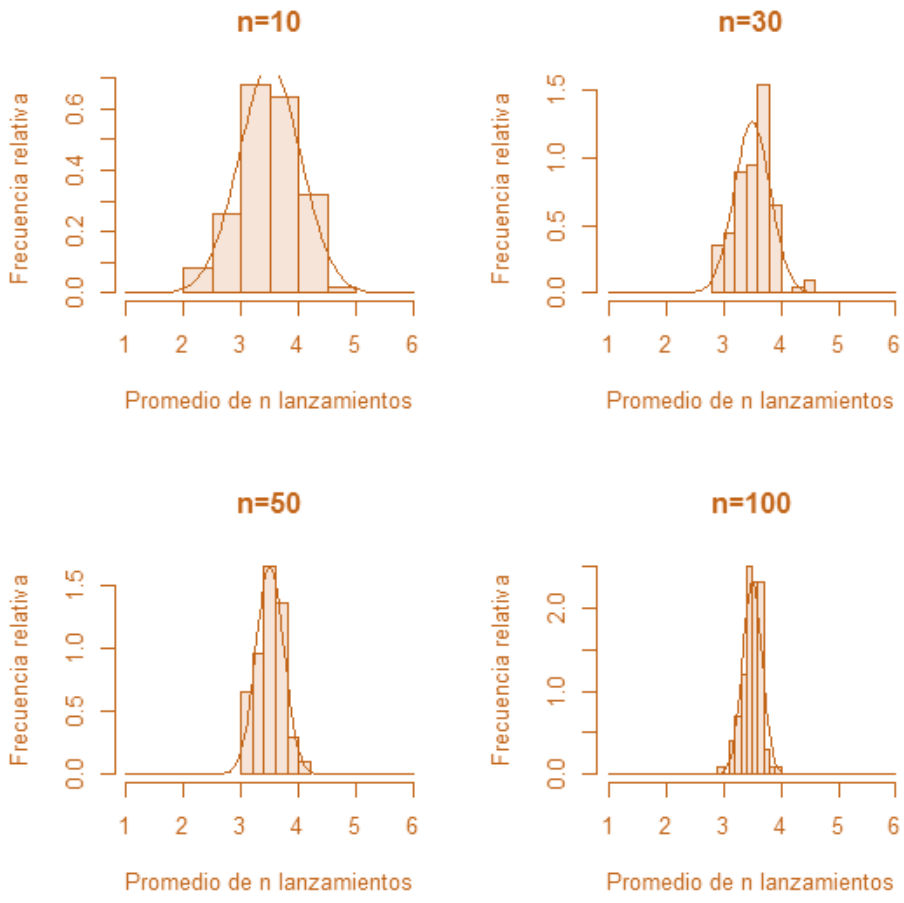



Fuente: elaboración propia con empleo del paquete estadístico R.

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

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

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




Fuente: elaboración propia con empleo del paquete estadístico R.

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

# 4.3. La distribución muestral de la proporción

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

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

- Donde:

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

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

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

y

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

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

Con lo anterior:

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

Y

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

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

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

Aplicando los resultados anteriores, para *n* considerablemente grande la distribución de $\sum_{i=1}^{n} x_i$ se aproxima a una normal con media $n\cdot P$ y varianza $n\cdot P\cdot(1 - P)$.

Si se define la siguiente variable $Y = a\cdot\sum_{i=1}^{n} x_i + b$, donde $a = \frac{1}{n}$ y b=0, entonces:

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

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

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

**Observaciones**

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

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

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

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

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

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

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

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

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

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

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

De acuerdo con una encuesta realizada a una población de 2919 egresados de licenciatura de la Facultad de Contaduría y Administración, el 80.4% considera excelentes o buenas las técnicas de enseñanza que utilizaron sus profesores durante la carrera[3]. Con la intención de conocer a mayor profundidad la metodología de enseñanza de sus docentes, la Dirección de la Facultad decide contactar a una muestra aleatoria de 100 egresados que contestaron la encuesta. ¿Cuál es la probabilidad de que el porcentaje de egresados en la muestra que juzgue excelentes o buenas las técnicas de enseñanza de sus profesores de licenciatura sea mayor a 90%?

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

De esta manera:

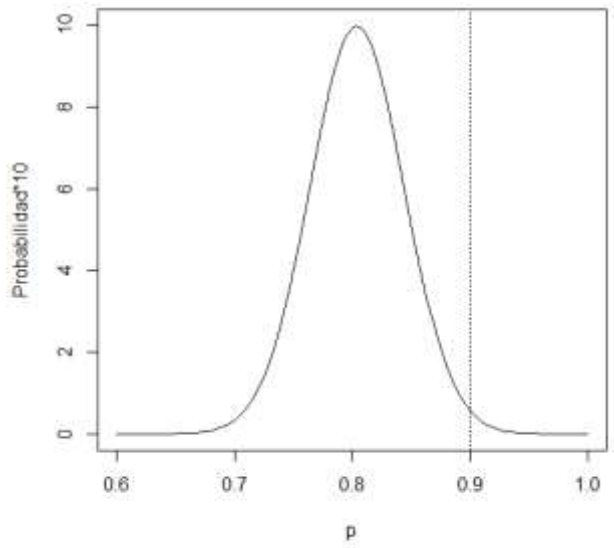
$$E(p) = P = 0.804$$

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

[3]UNAM. Dirección General de Planeación. *Perfiles de alumnos egresados del nivel licenciatura de la UNAM 2012-2013,* p. 71. www.Planeación.unam.mx/publicaciones. Consultado el 13 de julio de 2015.

Ahora, como n·P = (100)·(0.804) = 80.4 y n·(1 – P) = (100)·(1 – 0.804) = 19.6 son mayores a 5, entonces la distribución muestral de la proporción se aproxima a una normal con media 0.804 y desviación 0.04. (Véase figura 8).

**Figura 8. Distribución muestral de una proporción calculada con muestras de cien elementos**




Fuente: elaboración propia con empleo del paquete estadístico R.

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

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

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

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

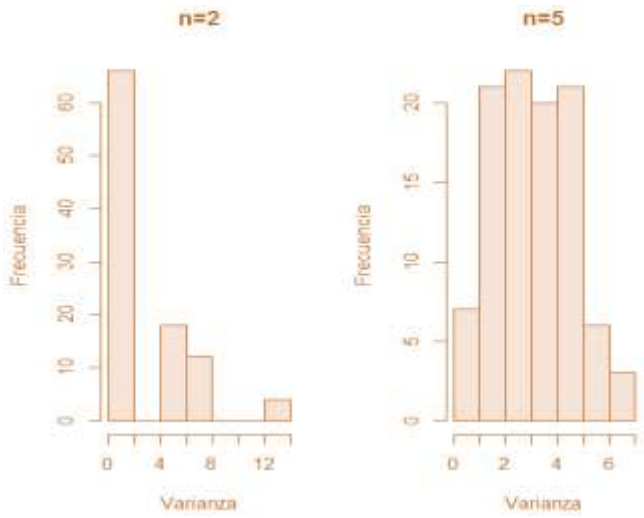
Este resultado indica que es prácticamente imposible tener en la muestra un porcentaje mayor a 90% de egresados que consideren excelentes o buenas las técnicas de enseñanza de sus profesores de licenciatura.

## 4.4. La distribución muestral de la varianza

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

En el ejemplo del subtema 2.2, se plantearon lanzamientos de un dado para mostrar el comportamiento del promedio muestral, ¿cómo sería la distribución de la varianza de 100 muestras de dos y cinco lanzamientos? (Tablas 3 y 4). En este orden, la figura 9 presenta la distribución de frecuencias de las varianzas de las 100 muestras de dos y cinco lanzamientos.

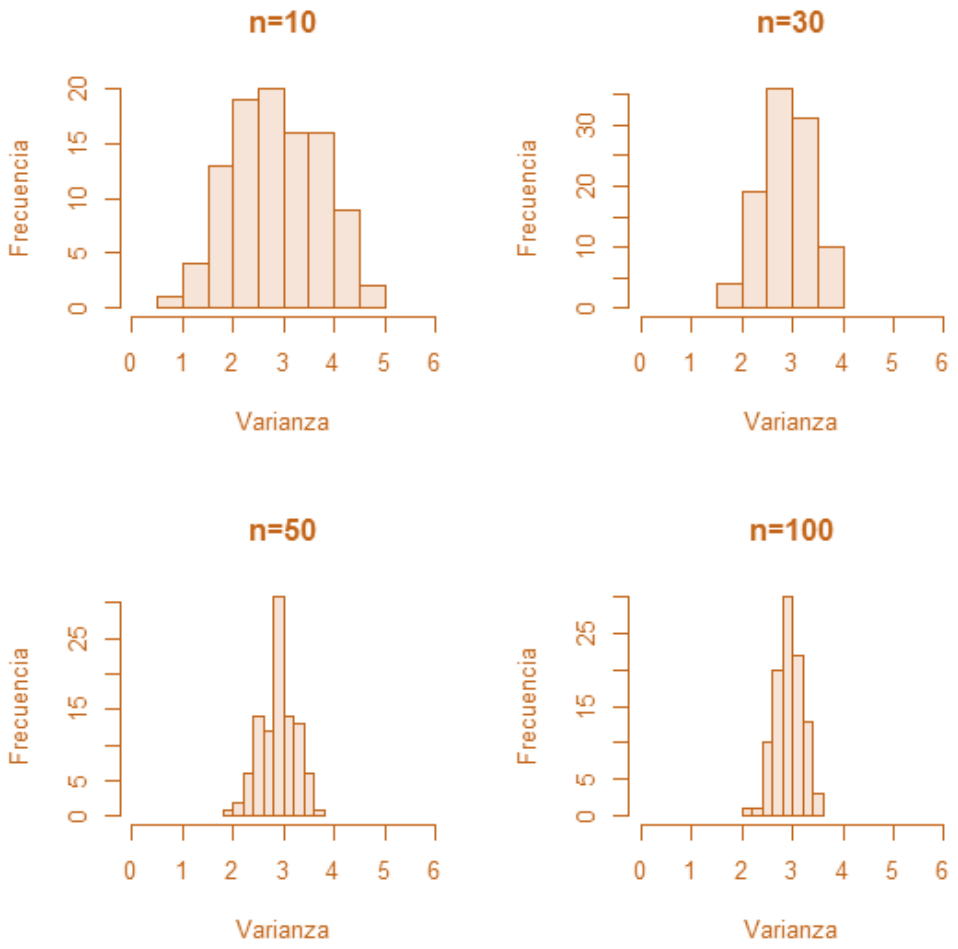
**Figura 9. Distribución de frecuencias de las varianzas de dos y cinco lanzamientos de un dado**




Fuente: elaboración propia con empleo del paquete estadístico R.

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

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




Fuente: elaboración propia con empleo del paquete estadístico R.

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

La distribución empleada para modelar la varianza muestral es $\boldsymbol{\chi^2}$(ji-cuadrada), cuya función de densidad es

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

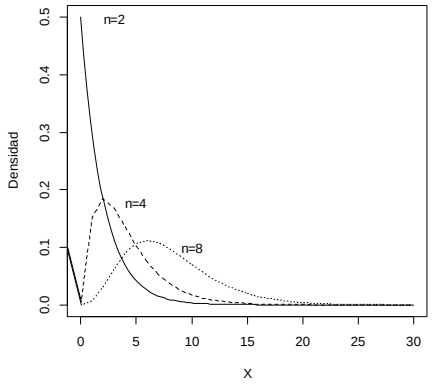
Para x >0

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

Las características de esta distribución son las siguientes:

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

**Figura 11. Ejemplo del comportamiento de una distribución $\chi^2$ con 2, 4 y 8 grados de libertad**




Fuente: elaboración propia.

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

**Resultados importantes**

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

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

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

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

**Funciones en Excel para trabajar la distribución $\chi^2$**

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

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

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

Para resolver el problema, se requiere calcular

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

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

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

**Distribución para comparar dos varianzas**

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

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

Para$x > 0$

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

Características de la distribución F:

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

**Funciones en Excel para trabajar la distribución $F$**

Excel tiene las siguientes funciones para trabajar con la distribución:

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

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

# RESUMEN

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

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

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

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

# BIBLIOGRAFÍA

## SUGERIDA

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

Anderson, S. (2012). *Estadística para negocios y economía* (11.ª ed.). México: CENGAGE Learning.

Levin, R. y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed.). México: Pearson.

Lind A. D., Marchal, G. W. y Wathen, S. (2012). *Estadística aplicada a los negocios y economía* (15.ª ed.). México: McGraw-Hill.

# UNIDAD 5

# Pruebas de hipótesis con la distribución ji cuadrada

# OBJETIVO PARTICULAR

El alumno relacionará los conceptos de prueba de hipótesis con la distribución ji cuadrada.

# TEMARIO DETALLADO
## (8 horas)

**5. Pruebas de hipótesis con la distribución ji cuadrada**

- 5.1. La distribución ji cuadrada, $\chi^2$
- 5.2. Pruebas de hipótesis para la varianza de una población
- 5.3. Prueba para la diferencia entre n proporciones
- 5.4. Pruebas de bondad de ajuste a distribuciones teóricas
  - 5.4.1. Ajuste a una distribución normal
  - 5.4.2. Ajuste a una distribución Poisson
  - 5.4.3. Ajuste a una distribución binomial
- 5.5. Pruebas sobre la independencia entre dos variables
- 5.6. Pruebas de homogeneidad

# INTRODUCCIÓN

En la unidad anterior, se dieron las bases para realizar pruebas de hipótesis para contrastar valores de parámetros de una población, como la media y una proporción. Posteriormente, se contrastaron medias, proporciones y varianzas de poblaciones independientes utilizando estadísticos de prueba con distribuciones normal, t de Student y F. Ahora, en esta unidad, se empleará otra distribución muestral, la ji cuadrada ($\chi^2$), útil no solamente para realizar pruebas relacionadas con una varianza poblacional, sino también para validar si una muestra se ajusta a una distribución teórica, si hay un cambio en una distribución, si dos variables son independientes o si dos muestras proceden de la misma población.

Primero, se expondrá la distribución $\chi^2$; después, se mostrará su uso para contrastar hipótesis relacionadas con la varianza poblacional, diferencia de proporciones, bondad de ajuste, independencia y homogeneidad.

Para el profesional egresado de la Facultad de Contaduría y Administración, el conocimiento y manejo de esta distribución le dará una herramienta adicional para una mejor toma de decisiones.

# 5.1. La distribución ji cuadrada, $\chi^2$

En la última sección de la tercera unidad, se utilizó la distribución $\chi^2$ (ji cuadrada) para estimar un intervalo para una varianza poblacional. Teóricamente, esta distribución es un caso de otra distribución conocida como *gamma*; el parámetro que determina su distribución son los grados de libertad, es decir, el número de observaciones que pueden variar libremente. Las características de esta distribución son las siguientes:

- La distribución se encuentra definida para valores positivos.
- La forma de una distribución $\chi^2$ depende de los grados de libertad (gl), por lo que hay un número infinito de distribuciones.
- El área bajo la curva es uno.
- La distribución es sesgada a la derecha.

En distribuciones muestrales, se emplea el estadístico

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

- Donde:

n = tamaño de muestra
$\sigma^2$ = varianza poblacional
$s^2$ = varianza muestral

El estadístico tiene una distribución $\chi^2$ con n – 1 grados de libertad.

Este resultado es válido si la muestra proviene de una población con distribución normal.

## 5.2. Pruebas de hipótesis para la varianza de una población

En la unidad anterior, se realizaron pruebas de hipótesis relacionadas con una media, una proporción, diferencia de medias y diferencia de proporciones, y se finalizó con pruebas entre dos varianzas. En este capítulo, se expone cómo efectuar una prueba para la varianza de una población.

Como se ha mencionado en las unidades pasadas, en ocasiones se requiere hacer inferencias sobre la varianza poblacional. Así como en la unidad anterior, en este caso se plantea una hipótesis nula y otra alternativa que involucra a la varianza, pero el estadístico de prueba es:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Y la distribución asociada es una $\chi^2$ con n – 1 grados de libertad.

A continuación, se analizan dos ejemplos.

**Ejemplo 1.**

Un *call center* tiene como criterio de calidad que la duración de sus llamadas tengan una desviación estándar de 1.5 respecto al promedio de cinco minutos El gerente del *call center* sospecha que la desviación es mayor, para confirmarlo elige una muestra de 50 llamadas y obtiene una desviación de 1.37 minutos. ¿Se puede afirmar con un nivel de confianza del 95% que la sospecha del gerente es correcta?

| Parámetro solicitado: | Datos: |
|---|---|
| $\boldsymbol{\sigma}$ | $\boldsymbol{\sigma = 1.5}$<br>$\boldsymbol{n = 50}$<br>$\boldsymbol{s = 1.37}$<br>Nivel de confianza: 95% = 0.95<br>Significancia: $\boldsymbol{\alpha = 1 - 0.95 = 0.05}$<br>Grados de libertad: n – 1 = 50 – 1 = 49 |

Hipótesis:

$$\boldsymbol{H_0 = \sigma^2 = (1.5)^2}$$
$$\boldsymbol{H_1 = \sigma^2 > (1.5)^2}$$

Cálculo del estadístico de prueba:

$$\boldsymbol{\chi^2 = \frac{(50-1)\cdot(1.37)^2}{(1.5)^2}}$$
$$\boldsymbol{\chi^2 = \frac{(49)\cdot(1.37)^2}{(1.5)^2}}$$
$$\boldsymbol{\chi^2 = \frac{(49)\cdot 1.8967}{2.25}}$$

Cálculo del punto crítico
Con el empleo de la función de Ms-Excel:

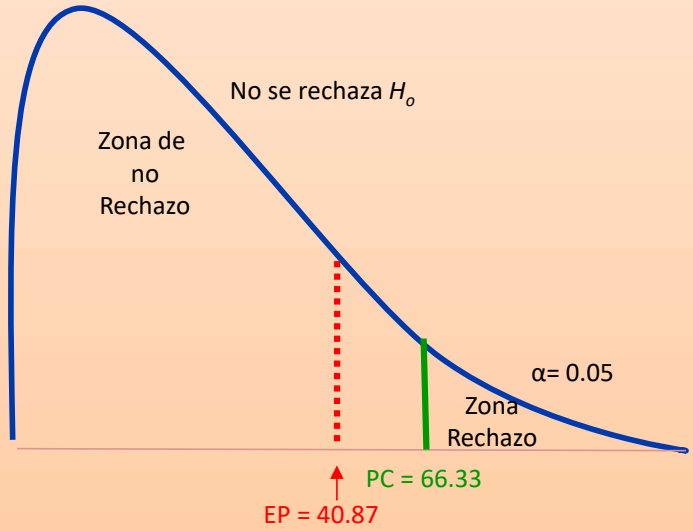
PRUEBA.CHI.INV(probabilidad,grados_de_libertad)

Se obtiene:

PRUEBA.CHI.INV(0.05,49) = 66.3386

En la figura 1, se ilustra la región donde cae el estadístico de prueba:

**Figura 1. Resultado de la prueba de hipótesis *Ho*: $\sigma^2 = 1.5$ contra *Ho*: $\sigma^2 > 1.5$**




La figura anterior muestra la distribución del estadístico de prueba asumiendo que la hipótesis nula es cierta. Como la prueba es unilateral, en este caso la región de rechazo se encuentra en el extremo derecho de la curva, a partir del punto crítico (66.33), ello significa que, si la prueba tiene un valor mayor a este punto, la hipótesis nula se rechaza. En la figura, se observa que el resultado de la prueba (40.87) es menor al punto crítico, por tanto, no se rechaza la hipótesis nula.
En conclusión, no existe evidencia estadística para rechazar la hipótesis nula, es decir, no se apoya la sospecha del gerente que la desviación estándar sea mayor a 1.5 minutos.

**Ejemplo 2.**

Una empresa realiza periódicamente una encuesta de clima laboral entre los empleados. Recientemente, varios departamentos solicitan que esta encuesta ya no se realice con la misma periodicidad, pues distrae las labores de los subordinados. En defensa de la encuesta, el director de recursos humanos sostiene que una variabilidad de 7 minutos no afecta el desempeño. Para comprobar que la variabilidad es de 7, elige una muestra de 20 empleados y obtiene un resultado de 6.7 minutos. ¿Se puede afirmar, con un nivel de confianza del 90%, que el director está en lo correcto?

| Parámetro solicitado: | Datos: |
|---|---|
| $\boldsymbol{\sigma}$ | $\boldsymbol{\sigma = 7}$<br>$\boldsymbol{n = 20}$<br>$\boldsymbol{s = 6.7}$<br>Nivel de confianza: 90% = 0.90<br>Significancia: $\boldsymbol{\alpha = 1 - 0.9 = 0.1}$<br>$\boldsymbol{\alpha = \frac{0.1}{2} = 0.05}$<br>Grados de libertad: n – 1 = 20 – 1 = 19 |

Hipótesis:

$$\boldsymbol{H_0 = \sigma^2 = (7)^2}$$
$$\boldsymbol{H_1 = \sigma^2 \neq (7)^2}$$

Cálculo del estadístico de prueba:

$$\boldsymbol{\chi^2 = \frac{(20-1)\cdot(6.7)^2}{(7)^2}}$$
$$\boldsymbol{\chi^2 = \frac{(19)\cdot 44.89}{49}}$$
$$\boldsymbol{\chi^2 = \frac{852.91}{49}}$$
$$\boldsymbol{\chi^2 = 17.4}$$

Cálculo del punto crítico
Con Excel, se obtienen los puntos críticos. Valor crítico superior:

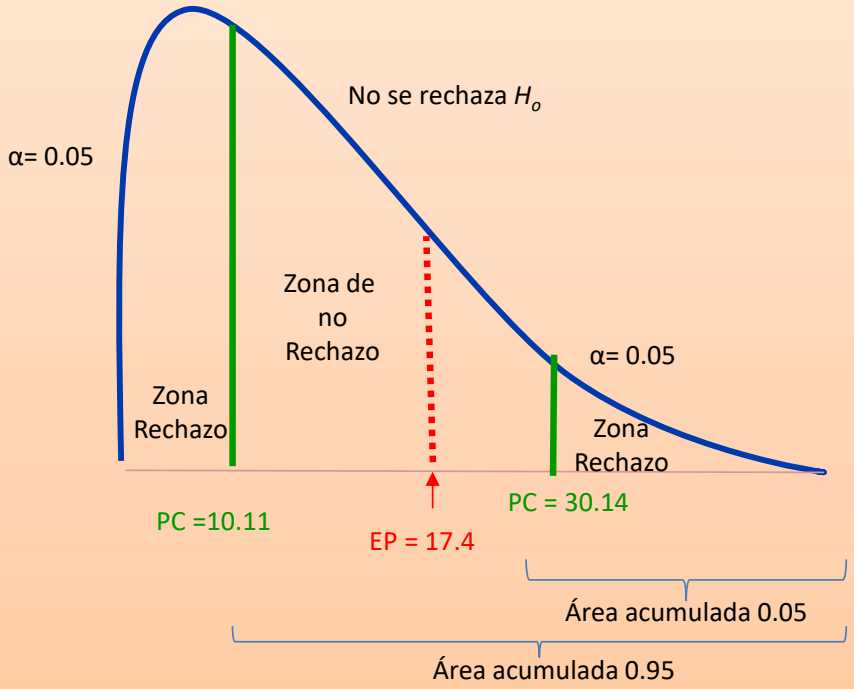
PRUEBA.CHI.INV (0.05,19) = 30.14

Valor crítico inferior:

PRUEBA.CHI.INV (0.95,19) = 10.11

En la figura 2, se ilustra la región donde cae el estadístico de prueba:

Figura 2. Resultado de la prueba de hipótesis *Ho*: = 7 contra *Ho*: ≠ 7




La figura anterior muestra la distribución del estadístico de prueba asumiendo que la hipótesis nula es cierta. Como la prueba es bilateral, la región de rechazo se encuentra en ambos extremos de la curva. La región de aceptación se halla entre los puntos críticos (10.11 y 30.14), esto significa que, si la prueba tiene un valor en esta región, la hipótesis nula se acepta. En la figura, se observa que el resultado de la prueba (17.4) se encuentra en la zona de aceptación, por tanto, no se rechaza la hipótesis nula. En conclusión, no existe evidencia estadística para rechazar la hipótesis nula: se apoya la defensa del director de recursos humanos.

# 5.3. Prueba para la diferencia entre *n* proporciones

En la sección anterior, se mostró el empleo de la distribución $\chi^2$ para hacer un contraste de hipótesis de una varianza poblacional. A partir de esta sección, se analizará su utilidad en la comparación de datos observados contra esperados, y de esta manera apoyar o no un comportamiento teórico.

Estadístico de prueba que se empleará a partir de esta sección:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Donde:

$o_i$ = valor observado
$e_i$ = valor esperado
*k* = número de categorías

Este estadístico tendrá una distribución $\chi^2$. Los grados de libertad varían según el contexto.

En esta sección, se aplicará el estadístico mencionado para apoyar o no que un conjunto de datos tiene una distribución multinomial.

En el curso de Estadística Descriptiva, se presentó la distribución binomial, la cual tiene como una de sus características que cada uno de los *n* ensayos independientes solamente ofrece dos resultados posibles manteniéndose constante la probabilidad de éxito. Cuando existen al menos tres resultados posibles, los cuales son mutuamente excluyentes y cada uno con una probabilidad de ocurrencia de manera que su suma da uno, se está frente a una distribución multinomial.

Supóngase que históricamente la proporción de estudiantes de Administración que obtiene una calificación mayor a 9 en Estadística Inferencial es 0.05; entre 8 y 9, 0.15; entre 7 y 8, 0.55; y el resto, menor a 7. Se ha propuesto un estrategia de enseñanza que se espera mejore el aprovechamiento de la materia en los estudiantes de Administración. Un grupo piloto de 140 alumnos registró los siguientes resultados:

| Nivel | Rango de calificación | Alumnos |
|---|---|---|
| **A** | 9.1-10 | 15 |
| **B** | 8.1-9.0 | 35 |
| **C** | 7.1-8.0 | 50 |
| **D** | Hasta 7.0 | 40 |
| **Total** | | 140 |

¿Se podría apoyar con un nivel de confianza de 95% que la estrategia modificó el aprovechamiento de los estudiantes de Administración en Estadística Inferencial?

Obsérvese que el tratamiento de la información se ajusta al de una distribución multinomial porque hay más de dos resultados y cada alumno nada más puede estar en una categoría. Se denotará como $p_A$, $p_B$, $p_C$ y $p_D$ a la proporción de alumnos en cada nivel, y se aplicará una prueba de hipótesis para determinar si la nueva estrategia modifica el desempeño.

La hipótesis nula y alternativa para probar si la estrategia modifica o no el desempeño es la siguiente:

$H_0: p_A = 0.05;\ p_B = 0.15;\ p_C = 0.55;\ p_D = 0.25$
$H_a: las\ proporciones\ poblacionales\ no\ son\ las\ de\ la\ hipótesis\ nula$

Asumiendo como cierta la hipótesis nula, se esperaría que los 140 alumnos se distribuyeran de la siguiente manera:

| Nivel | Rango de calificación | Proporción bajo $H_o$ | Alumnos esperados |
|---|---|---|---|
| **A** | 9.1-10 | 0.05 | $\mathbf{140 \cdot 0.05 = 7}$ |
| **B** | 8.1-9.0 | 0.15 | $\mathbf{140 \cdot 0.15 = 21}$ |
| **C** | 7.1-8.0 | 0.55 | $\mathbf{140 \cdot 0.55 = 77}$ |
| **D** | Hasta 7.0 | 0.25 | $\mathbf{140 \cdot 0.25 = 35}$ |
| **Total** | | | **140** |

Se calcula el estadístico de prueba que tendrá una distribución $\chi^2$ con k – 1 grados de libertad, en este caso, k = 4:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 = \frac{(15-7)^2}{7} + \frac{(35-21)^2}{21} + \frac{(50-77)^2}{77} + \frac{(40-35)^2}{35}$$

$$\chi^2 = 9.1 + 9.3 + 9.5 + 0.7$$

$$\chi^2 = 28.7$$

Se realiza una prueba bilateral. Con Microsoft Excel (2013), se calcula el punto crítico superior:

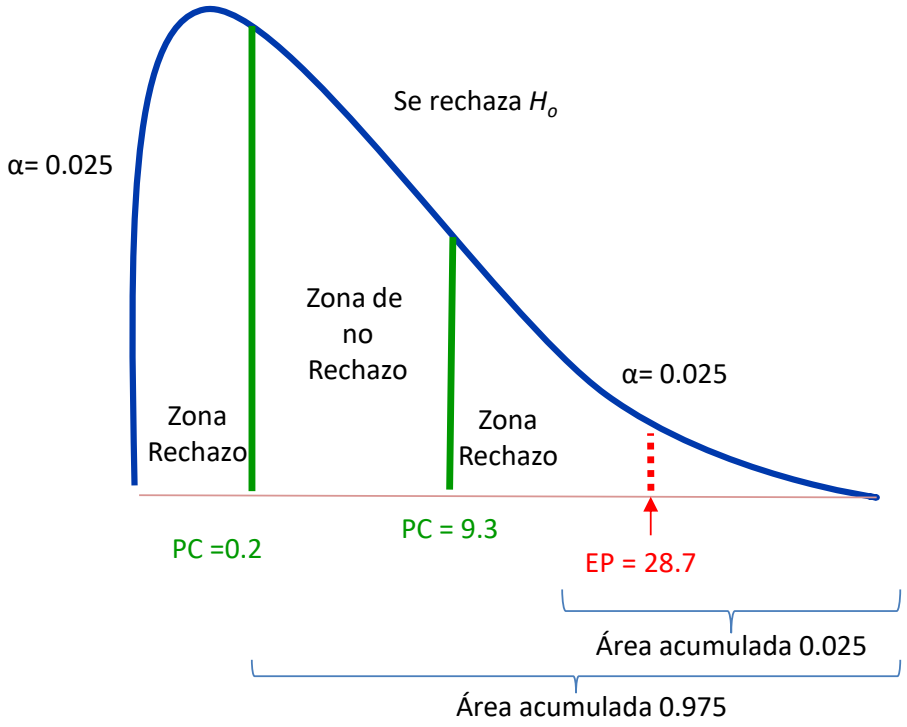
PRUEBA.CHI.INV(0.05/2,3) = 9.3

Y el inferior:

PRUEBA.CHI.INV(1-0.05/2,3) = 0.2

En la figura 3, se ilustra la región donde cae el estadístico de prueba.

**Figura 3. Resultado de la prueba de hipótesis**




Fuente: elaboración propia.

La figura anterior muestra la distribución del estadístico de prueba asumiendo que la hipótesis nula es cierta. Debido a que la prueba es bilateral, la región de rechazo se

encuentra en ambos extremos de la curva. La región de aceptación se halla entre los puntos críticos (0.2 y 9.3), lo cual significa que, si la prueba tiene un valor en esta región, la hipótesis nula se acepta. En la figura se observa que el resultado de la prueba (28.7) se sitúa en la zona de rechazo, por tanto, se rechaza la hipótesis nula.

En conclusión, hay evidencia estadística para rechazar la hipótesis nula: la estrategia modificó el aprovechamiento de los estudiantes de Administración en Estadística Inferencial.

## 5.4. Pruebas de bondad de ajuste a distribuciones teóricas

Como se ha estudiado hasta este punto, tanto las técnicas de estimación como las de contraste de hipótesis se realizan con la información de una muestra. A veces, se pretende conocer si la población de la que proviene la muestra se ajusta a una distribución teórica. En esta sección, se utilizará la distribución $\chi^2$ para probar si un conjunto de información se ajusta a una distribución Normal, Poisson o Binomial. En las tres distribuciones el proceso para realizar la prueba es similar:

Se forman categorías.

Se realizan conteos en cada categoría.

Se estima el valor esperado de elementos en cada categoría.

Se contrasta la hipótesis.

- $H_0$: los datos se ajustan a la distribución
- $H_1$: los datos no se ajustan a la distribución

Con el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Donde:

$o_i$ = valor observado
$e_i$ = valor esperado
$k$ = número de categorías

Asumiendo cierta la hipótesis nula, este estadístico tendrá una distribución $\chi^2$, con k – p – 1 grados de libertad, donde *k* es el número de categorías y *p* los parámetros de la distribución teórica.

La hipótesis nula se rechaza si el valor del estadístico de prueba resulta mayor al punto crítico de la distribución teórica.

A continuación, se muestra cómo realizar la prueba de bondad de ajuste para una distribución normal.

## 5.4.1. Ajuste a una distribución normal

Para explicar la prueba para el ajuste a una distribución normal, se utilizará el siguiente ejemplo.

Los resultados de una prueba realizada a 110 aspirantes a ocupar una plaza laboral se muestra a continuación.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 80.0 | 60.0 | 56.7 | 54.2 | 52.5 | 50.8 | 48.3 | 46.7 | 45.0 | 42.5 | 36.7 |
| 69.2 | 59.2 | 56.7 | 53.3 | 51.7 | 50.0 | 48.3 | 46.7 | 45.0 | 42.5 | 36.7 |
| 69.2 | 59.2 | 56.7 | 53.3 | 51.7 | 50.0 | 48.3 | 45.8 | 44.2 | 41.7 | 36.7 |
| 69.2 | 59.2 | 56.7 | 53.3 | 51.7 | 49.2 | 47.5 | 45.8 | 44.2 | 41.7 | 34.2 |
| 69.2 | 58.3 | 56.7 | 53.3 | 51.7 | 49.2 | 47.5 | 45.8 | 44.2 | 40.8 | 34.2 |
| 68.3 | 58.3 | 55.8 | 53.3 | 50.8 | 49.2 | 47.5 | 45.0 | 44.2 | 40.8 | 33.3 |
| 64.2 | 57.5 | 55.8 | 53.3 | 50.8 | 49.2 | 47.5 | 45.0 | 43.3 | 40.0 | 32.5 |
| 63.3 | 57.5 | 55.8 | 52.5 | 50.8 | 49.2 | 46.7 | 45.0 | 43.3 | 38.3 | 32.5 |
| 61.7 | 56.7 | 55.0 | 52.5 | 50.8 | 48.3 | 46.7 | 45.0 | 43.3 | 37.5 | 29.2 |
| 60.8 | 56.7 | 54.2 | 52.5 | 50.8 | 48.3 | 46.7 | 45.0 | 42.5 | 36.7 | 37.2 |

A fin de precisar los puntajes que deben tener los candidatos para pasar a la siguiente etapa, se quiere probar primeramente que los datos provienen de una distribución normal con un nivel de confianza de 95%.

Como la distribución normal es continua, para probar el ajuste a esta distribución, se categorizará la información en deciles.

En primer lugar, se estimarán los parámetros de la distribución (media y desviación estándar) con la información de la muestra.

El estimador de la media (μ) es el promedio muestral; y el de la desviación estándar (σ), la desviación muestral.

Así:

$$\hat{\mu} = \frac{80.0 + 69.2 + \cdots + 29.2 + 37.2}{109} = 49.7$$

- Y:

$$\hat{\sigma} = \sqrt{\frac{(80.0 - 49.7)^2 + (69.2 - 49.7)^2 + \cdots + (29.2 - 49.7)^2 + (37.2 - 49.9)^2}{109 - 1}} = 8.9$$

Se va a probar, entonces, si la información se ajusta a una distribución normal con media 49.7 y desviación estándar de 8.9.

Para realizar la prueba, se formarán 10 categorías y cada una concentrará una probabilidad de 10%. Estas categorías se determinarán con los cuantiles *z* de una distribución normal estándar; una vez conocido este valor, se procede a convertirlo en la métrica de la prueba.

En la siguiente tabla se muestran los puntos de corte.

**Tabla. Cálculo de los puntos de corte para formar las categorías que se utilizarán en la prueba de bondad de ajuste a una distribución normal**

| Corte | z | Puntaje 49.7+z·8.9 |
|---|---|---|
| 1 | -1.28 | 38.29 |
| 2 | -0.84 | 42.21 |
| 3 | -0.52 | 45.03 |
| 4 | -0.25 | 47.45 |
| 5 | 0.00 | 49.70 |
| 6 | 0.25 | 51.95 |
| 7 | 0.52 | 54.37 |
| 8 | 0.84 | 57.19 |
| 9 | 1.28 | 61.11 |

La tabla anterior consta de tres columnas: corte, z y puntaje. En la primera columna solamente se enumeran los puntos de corte que se requieren para dividir la distribución teórica en 10 partes iguales. La segunda (z) es el cuantil de una distribución normal estándar que acumula un área de 0.1 desde el último corte a la izquierda. Y la tercera es la conversión del valor del cuantil *z* a la métrica del examen. Esta conversión se fundamenta en que la distribución normal estándar se calcula así:

$$Z = \frac{X - \mu}{\sigma}$$

- Donde:

Z = variable estandarizada
X = variable original
μ = media de *X*
σ = desviación de *X*

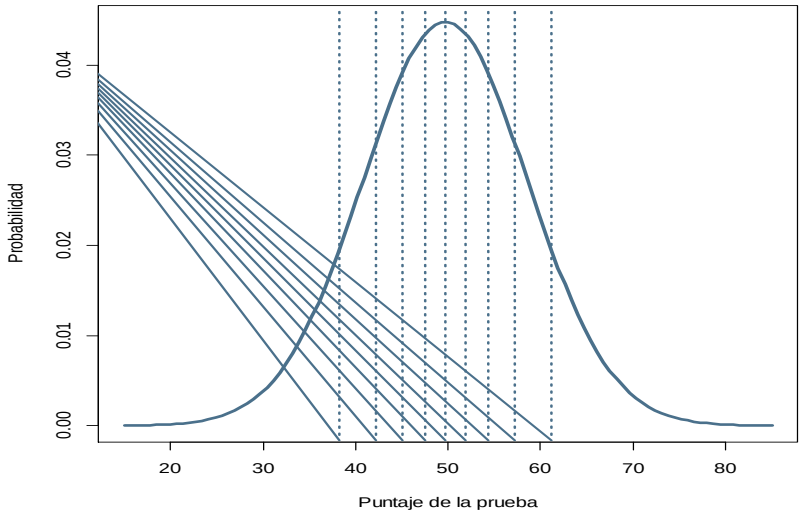
- Al despejar *X*, se obtiene:

$$X = \mu + Z \cdot \sigma$$

Así, el punto de corte en la métrica del examen se obtiene sumando al promedio (49.7) el producto del cuantil por la desviación estándar (8.9).

En la figura 4, se ilustra la segmentación de la distribución teórica con el empleo de los puntos de corte calculados.

**Figura 4. Segmentación de la distribución teórica en 10 áreas iguales**




Fuente: elaboración propia con empleo del paquete estadístico R[4]

La figura anterior muestra la segmentación en 10 áreas del mismo tamaño (0.1) de una distribución normal con media de 49.7 y desviación estándar de 8.9. El siguiente paso consiste en realizar un conteo de los aspirantes que caen en cada categoría (área) y compararlo con su número esperado: (110) (0.1) = 11. La siguiente tabla presenta las frecuencias observadas y esperadas para cada categoría.

**Tabla. Frecuencias observadas y esperadas por categoría**

| | Frecuencia | |
|---|---|---|
| **Categoría** | **Observada** | **Estimada** |
| 1 | 12 | 11 |
| 2 | 6 | 11 |
| 3 | 17 | 11 |
| 4 | 8 | 11 |
| 5 | 14 | 11 |
| 6 | 12 | 11 |
| 7 | 12 | 11 |
| 8 | 11 | 11 |
| 9 | 9 | 11 |
| 10 | 9 | 11 |

[4]R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

| | | |
|---|---|---|
| **Total** | **110** | **110** |

Una vez que se cuenta con las frecuencias observadas y estimadas para cada categoría, se procede a realizar la prueba con el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Sustituyendo los valores, se tiene:

$$\chi^2 = \frac{(12-11)^2}{11} + \frac{(6-11)^2}{11} + \cdots + \frac{(9-11)^2}{11} = 8.2$$

A partir de la hipótesis nula, el estadístico de prueba tiene una distribución $\chi^2$ con k – p – 1 grados de libertad. En este caso, k = 10 y p = 2 porque la distribución normal tiene dos parámetros (media y desviación estándar): se comparará el valor del estadístico de prueba con el punto crítico de una distribución $\chi^2$ con 10 – 2 – 1 = 7 grados de libertad que corta la curva en dos zonas: una con área de 0.05 a su derecha y la otra de 0.95.

Con Microsoft Excel (2013), se calcula el punto crítico de esta distribución así:

PRUEBA.CHI.INV(0.05, 7) = 14.07

Como el punto crítico es mayor al valor del estadístico de prueba, no se tiene evidencia estadística para rechazar la hipótesis nula. Luego, se apoya la hipótesis de que la muestra proviene de una población con distribución normal.

## 5.4.2. Ajuste a una distribución Poisson

En este apartado, se muestra un ejemplo donde se prueba la bondad de ajuste a una distribución Poisson.

En un establecimiento comercial, se han incrementado las quejas respecto a que no hay suficiente personal para atender a la clientela. Por su parte, los empleados solicitan al gerente que contrate más personal debido a que la demanda los supera. Con la intención de justificar la contratación de más personal, el gerente, durante una semana, tomó una muestra aleatoria de 60 periodos de 15 minutos y registró el número de clientes que acuden al establecimiento. Los registros son los siguientes:

| | | | | | |
|---|---|---|---|---|---|
| 10 | 6 | 9 | 8 | 12 | 9 |
| 20 | 15 | 1 | 20 | 16 | 1 |
| 14 | 16 | 18 | 0 | 19 | 9 |
| 17 | 1 | 5 | 4 | 10 | 4 |
| 10 | 20 | 13 | 10 | 16 | 19 |
| 8 | 17 | 13 | 9 | 1 | 6 |
| 5 | 10 | 15 | 10 | 14 | 9 |
| 10 | 15 | 8 | 3 | 11 | 8 |
| 18 | 17 | 14 | 17 | 12 | 9 |
| 3 | 2 | 14 | 15 | 16 | 1 |

Para realizar simulaciones, se debe estar convencido de que la distribución de las llegadas sigue una distribución Poisson. Con un nivel de confianza del 95%, se apoya la hipótesis de que las llegadas se ajustan a una distribución Poisson.

Para realizar la prueba, primero se construye una tabla de frecuencia de llegadas:

| Llegadas | Casos |
|---|---|
| 1 | 6 |
| 2 | 1 |
| 3 | 2 |
| 4 | 2 |
| 5 | 2 |
| 6 | 2 |
| 7 | 0 |
| 8 | 4 |
| 9 | 6 |
| 10 | 7 |
| 11 | 1 |
| 12 | 2 |
| 13 | 2 |
| 14 | 4 |
| 15 | 4 |
| 16 | 4 |
| 17 | 4 |
| 18 | 2 |
| 19 | 2 |
| 20 | 3 |
| **Promedio** | **10.7** |

En la primera columna de la tabla anterior, se muestra el número de llegadas registradas en periodos de 15 minutos, estas llegadas oscilan entre 1 y 20. En promedio, se registran 10.7 llegadas cada 15 minutos (este promedio se calculó utilizando el criterio de datos agrupados).

Con el propósito de no trabajar con frecuencias menores a cinco, se agruparán categorías y la tabla quedará de la siguiente forma:

| Llegadas | Casos |
|---|---|
| 1 | 6 |
| 2 a 7 | 9 |
| 8 a 9 | 10 |
| 10 y más | 35 |

La agrupación utilizada es un tanto subjetiva (normalmente, queda al criterio del investigador).

Se busca probar que la muestra proviene de una población con distribución Poisson con parámetro $\lambda = 10.7$, por lo que el siguiente paso es calcular el valor esperado de cada categoría.

| Llegadas | Casos | Probabilidad | Esperado |
|---|---|---|---|
| 1 a 7 | 15 | 0.1624 | 10 |
| 8 a 9 | 10 | 0.2096 | 13 |
| 10 y más | 35 | 0.6245 | 37 |

En la tabla anterior, se agruparon la primera y segunda categorías debido a que la frecuencia esperada de una llegada resultó cero. Para calcular la frecuencia esperada, primero se utiliza la distribución teórica Poisson con $\lambda = 10.7$. Después, la probabilidad obtenida en cada categoría se multiplica por el tamaño de la muestra (60). Una vez

que se tienen los datos observados y esperados, se realiza la prueba con el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Sustituyendo los valores, se obtiene:

$$\chi^2 = \frac{(15-10)^2}{10} + \frac{(10-13)^2}{13} + \frac{(35-37)^2}{37}$$
$$\chi^2 = 3.52$$

A partir de la hipótesis nula, el estadístico de prueba tiene una distribución $\chi^2$ con k – p – 1 grados de libertad. En este caso k = 3 y p = 1 porque la distribución Poisson tiene un parámetro, por lo que se comparará el valor del estadístico de prueba con el punto crítico de una distribución $\chi^2$ con 3 – 1 – 1 = 1 grados de libertad que corta la curva en dos zonas: una con área de 0.05 a su derecha y la otra de 0.95.

Con Microsoft Excel (2013), se calcula el punto crítico de esta distribución así:

PRUEBA.CHI.INV(0.05, 1) = 3.84

Como el punto crítico es mayor al valor del estadístico de prueba, no se tiene evidencia estadística para rechazar la hipótesis nula: se apoya la hipótesis de que la muestra proviene de una población con distribución Poisson.

## 5.4.3. Ajuste a una distribución binomial

Para finalizar el empleo de la $\chi^2$para ajustar a una distribución teórica, a continuación se presenta un ejercicio donde se desea probar que un conjunto de datos proviene de una distribución Binomial.

El expediente de un trámite se compone de cuatro documentos; si un documento está mal llenado, el expediente se clasifica como erróneo.

La auditoría realizada a la organización que elabora los expedientes mostró los siguientes resultados:

| Documentos erróneos | Expedientes |
|---|---|
| 0 | 130 |
| 1 | 150 |
| 2 | 200 |
| 3 | 120 |
| 4 | 50 |
| **Total** | **650** |

Antes de establecer alguna métrica, el auditor desea verificar que los expedientes con errores siguen una distribución binomial con un nivel de confianza del 95%.

La distribución binomial tiene dos parámetros: la probabilidad de éxito (p) y el número de ensayos (k). Si se define la variable teórica como el número de documentos con error de los cuatro que forman el trámite, k = 4 y p es la probabilidad de que un documento tenga error. Como *k* ya se conoce, el siguiente paso es estimar *p* minúscula. Para hacerlo, se calcula el promedio bajo un criterio de datos agrupados y se divide entre *k*. Realizando estas operaciones, la estimación de p = 0.42692.

Estimados los parámetros de la distribución teórica, se procede a calcular los valores esperados. Primero, se calculan las probabilidades de cada categoría y después la probabilidad calculada se multiplica por el total de expedientes.

En la siguiente tabla, se muestran las frecuencia observadas y estimadas:

| Documentos erróneos | Expedientes | Probabilidad | Esperados |
|---|---|---|---|
| 0 | 130 | 0.108 | 70 |
| 1 | 150 | 0.321 | 209 |
| 2 | 200 | 0.359 | 233 |
| 3 | 120 | 0.178 | 116 |
| 4 | 50 | 0.033 | 22 |
| **Total** | **650** | **1** | **650** |

Por último, se realiza la prueba con el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

- Sustituyendo los valores, se obtiene:

$$\chi^2 = \frac{(130-70)^2}{70} + \frac{(150-209)^2}{209} + \frac{(200-233)^2}{203} + \frac{(120-116)^2}{116} + \frac{(50-22)^2}{22}$$
$$\chi^2 = 110.1$$

A partir de la hipótesis nula, el estadístico de prueba tiene una distribución $\chi^2$ con k – p – 1 grados de libertad. En este caso, k = 5 y p = 2 porque la distribución binomial tiene dos parámetros; entonces, se comparará el valor del estadístico de prueba con el punto crítico de una distribución $\chi^2$ con 5 – 2 – 1 = 2 grados de libertad que corta la curva en dos zonas: una con área de 0.05 a su derecha, y la otra de 0.95.

Con Microsoft Excel (2013), se calcula el punto crítico de esta distribución así:

PRUEBA.CHI.INV(0.05, 2) = 5.99

Como el punto crítico es menor al valor del estadístico de prueba, no se tiene evidencia estadística para apoyar la hipótesis nula, es decir, se rechaza la hipótesis de que la muestra proviene de una población con distribución binomial.

# 5.5. Pruebas sobre la independencia entre dos variables

En las secciones 5.3 y 5.4, se mostró el uso de la distribución $\chi^2$ para realizar pruebas acerca de la distribución de una población. Otra aplicación de la distribución es para determinar independencia entre dos variables cualitativas. Por ejemplo, podría ser de interés para el gerente de marca de una bebida gaseosa determinar si existe asociación entre el apego emocional a la marca respecto al consumo del producto; o al gerente de recursos humanos de una organización le sería de utilidad identificar la asociación entre el nivel de puntualidad de los empleados respecto a su zona de residencia. A continuación, se expone el empleo de la distribución $\chi^2$ para determinar asociación entre variables.

Antes de entrar en materia, conviene repasar algunos conceptos revisados en el curso de Estadística Descriptiva referentes a probabilidad.

**Independencia de eventos**

Con frecuencia, es necesario determinar la probabilidad de dos eventos independientes. Los eventos A y B son independientes si P(A y B) = P(A)·P(B), es decir, si dos eventos son independientes, entonces, la probabilidad de que ocurran al mismo tiempo es el producto de sus probabilidades.

Como una extensión, si las variables $X_1$ y $X_2$ son independientes, su función conjunta

$$f(x_1, x_2) = f(x_1) \cdot f(x_2)$$

Para ilustrar lo anterior, se expone el siguiente ejemplo. Supóngase que la variable $X_1$ está asociada al resultado de un curso de estadística (aprobado, reprobado), donde la

probabilidad de aprobar es 0.3 y la variable $X_2$ el sexo del alumno (mujer, hombre), siendo la probabilidad que una mujer tome el curso de 0.2.

En la tabla 1, se ilustra la distribución de ambas variables.

**Tabla 1. Distribución de las variables $X_1$ y $X_2$**

| Género | Aprueba | Reprueba | $f(x_2)$ |
|---|---|---|---|
| **Mujer** | | | 0.2 |
| **Hombre** | | | 0.8 |
| $f(x_1)$ | **0.3** | **0.7** | **1.0** |

En la tabla anterior, se presentan las variables de interés: por fila se muestra los valores de la variable $X_2$ (género del alumno); y en las columnas, los valores asociados a $X_1$ (resultado del curso). En los márgenes de la tabla se encuentran las distribuciones de probabilidad de las variables $X_1$ y $X_2$, denominadas distribuciones marginales.

Si $X_1$ y $X_2$ fueran independientes, su distribución conjunta $f(x_1,x_2)$ serían los valores de las celdas de la tabla, resultado de multiplicar las distribuciones marginales.

En la tabla 2, aparece el cálculo de la distribución conjunta.

**Tabla 2. Cálculo de la distribución conjunta de $X_1$ y $X_2$**

| Género | Aprueba | Reprueba | $f(x_2)$ |
|---|---|---|---|
| **Mujer** | 0.2·0.3 = 0.06 | 0.2·0.7 = 0.14 | 0.2 |
| **Hombre** | 0.3·0.8 = 0.24 | 0.8·0.7 = 0.56 | 0.8 |

| | | | |
|---|---|---|---|
| $f(x_1)$ | **0.3** | **0.7** | **1.0** |

Los valores de cada celda de la tabla son el resultado de multiplicar el valor de la distribución marginal en la fila por el de la columna.

Con lo anterior, si el grupo se compone de 60 alumnos, ¿cuántos se esperaría observar en cada categoría? Para responder esta pregunta, se multiplica los 60 por la probabilidad conjunta correspondiente, como se muestra a continuación.

**Tabla 3. Distribución esperada de 60 alumnos conforme a la distribución conjunta de $X_1$ y $X_2$**

| Género | Aprueba | Reprueba | Total |
|---|---|---|---|
| **Mujer** | 60·0.06 = 4 | 60·0.14 = 8 | 12 |
| **Hombre** | 60·0.24 = 14 | 60·0.56 = 34 | 48 |
| **Total** | **18** | **42** | **60** |

Por último, se muestra el uso de la distribución $\chi^2$ para determinar independencia entre dos valores.

**Tablas cruzadas**

Una tabla cruzada se utiliza para clasificar observaciones de una muestra de acuerdo con dos o más características (variables cualitativas). Si las variables involucradas en la tabla son independientes, la distribución conjunta tiene una distribución $\chi^2$ con (r – 1)·(c – 1) grados de libertad, donde *r* es el número de renglones de la tabla y *c* sus columnas.

De nuevo, el estadístico de prueba es:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

A continuación, se muestra un ejemplo.

La opinión de los alumnos asignados a la licenciatura de la UNAM sobre su nivel de preparación precedente se muestra a continuación por tipo de ingreso.

**Opinión de los alumnos asignados a licenciaturas de la UNAM sobre su preparación precedente por tipo de ingreso**

| Tipo de ingreso | Excelente | Buena | Regular | Deficiente | Total |
|---|---|---|---|---|---|
| **Pase reglamentado** | 6,160 | 15,184 | 1,276 | 80 | 22,700 |
| **Concurso de selección** | 4,012 | 9,007 | 1,298 | 139 | 14,456 |
| **Total** | **10,172** | **24,191** | **2,574** | **219** | **37,156** |

Fuente: Perfiles de Aspirantes y Asignados a Bachillerato, Técnico y Licenciatura de la UNAM 2013-2014. Dirección General de Planeación. UNAM.

¿Con un nivel de confianza de 95% se apoyaría la hipótesis de que la opinión del alumno respecto a su preparación previa a la licenciatura es independiente del tipo de ingreso?

Prueba de hipótesis:

$H_0$: *La opinión sobre la preparación precedente es independiente del tipo de ingreso*

$H_a$: *Existe asociación entre la opinión sobre la preparación precedente y el tipo de ingreso*

Para responder la pregunta, primero se calculan los valores esperados: se calculan las distribuciones marginales dividiendo los totales por fila y columna entre el total general.

Por ejemplo, la proporción de alumnos que respondió excelente es $\frac{10,172}{37,156} = \mathbf{0.27}$; y la proporción de alumnos que ingresó por pase reglamentado, $\frac{22,700}{37,156} = \mathbf{0.61}$. El resto de las proporciones se muestra a continuación.

| Tipo de ingreso | Excelente | Buena | Regular | Deficiente | Total |
|---|---|---|---|---|---|
| **Pase reglamentado** | | | | | 0.61 |
| **Concurso de selección** | | | | | 0.39 |
| **Total** | **0.27** | **0.65** | **0.07** | **0.01** | **1.00** |

El siguiente paso consiste en calcular el valor esperado de cada celda de la tabla, al que se llega multiplicando el total general (37,156) por el producto de la probabilidad de la fila y de la columna. Por ejemplo, el valor esperado de alumnos de pase reglamentado que respondieron Excelente es el siguiente:

$$\mathbf{37,156 \cdot (0.61 \cdot 0.27) = 6,214}$$

De esta manera, los valores esperados se muestran a continuación.

| Tipo de ingreso | Excelente | Buena | Regular | Deficiente | Total |
|---|---|---|---|---|---|
| **Pase reglamentado** | 6,214 | 14,779 | 1,573 | 134 | 22,700 |
| **Concurso de selección** | 3,958 | 9,412 | 1,001 | 85 | 14,456 |
| **Total** | **10,172** | **24,191** | **2,574** | **219** | **37,156** |

Obtenidos los valores esperados, sigue calcular el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 = \frac{(6,160 - 6,214)^2}{6,214} + \frac{(15,184 - 14,779)^2}{14,779} + \frac{(1,276 - 1,573)^2}{1,573} + \frac{(80 - 134)^2}{134} + \frac{(4,012 - 3,958)^2}{3,958} + \frac{(9,007 - 9,412)^2}{9,412} + \frac{(1,298 - 1,001)^2}{1,001} + \frac{(139 - 85)^2}{85}$$

$$\chi^2 = 229.0608$$

Se rechazará la hipótesis nula si el estadístico de prueba es mayor al punto crítico.

Si se asume que la hipótesis nula es cierta, el estadístico de prueba tiene una distribución $\chi^2$ con (r – 1) (c – 1), donde r = 2 renglones y c = 4 columnas, por tanto, tiene (2 – 1)·(4 – 1) = 3 grados de libertad.

Con Excel, se obtiene como punto crítico:

Como el valor de la prueba es notablemente mayor al punto crítico, se rechaza la hipótesis nula: se apoya que la opinión del estudiante sobre su preparación previa se encuentra asociada a su procedencia (tipo de ingreso).

# 5.6. Pruebas de homogeneidad

En la sección precedente, se utilizó la distribución $\chi^2$ para determinar si dos variables son independientes; ahora, se empleará para comprobar que dos o más muestras son homogéneas.

Que dos o más muestras sean homogéneas significa que provienen de la misma población, por lo que es de esperarse que presenten un comportamiento similar. Supóngase que se desea realizar un estudio para determinar las causas por las que los alumnos de la carrera de Administración no tienen un buen desempeño en la materia de Estadística Inferencial. Se escogen al azar cuatro grupos (dos del turno matutino y dos del vespertino) y se obtiene la distribución de calificaciones en la materia, como se muestra a continuación.

**Tabla. Distribución de las calificaciones del curso de Estadística Inferencial en cuatro grupos de Administración**

| Grupo | Calificación del curso | | | | Total |
|---|---|---|---|---|---|
| | 5 | 6.0 a 7.5 | 7.6 a 8.5 | 8.6 a 10 | |
| **Matutino$_1$** | 7 | 50 | 9 | 6 | 72 |
| **Matutino$_2$** | 9 | 55 | 8 | 7 | 79 |
| **Vespertino$_1$** | 6 | 40 | 6 | 5 | 57 |
| **Vespertino$_2$** | 7 | 35 | 7 | 6 | 55 |
| **Total** | 29 | 180 | 30 | 24 | 263 |

La tabla anterior presenta la distribución de calificaciones de Estadística Inferencial de 263 alumnos provenientes de los cuatro grupos seleccionados: 180 (68%) tiene calificaciones entre 6 y 7.5; y 24 (9%), notas mayores a 8.5.

Antes de continuar, los académicos responsables de la investigación quieren verificar que las muestras (los grupos) sean homogéneas con un nivel de confianza de 95%

para generalizar los resultados que se obtengan, por lo que realizan una prueba de homogeneidad de muestras.

Hipótesis que contrastan:

$H_0$: las muestras son homogéneas

$H_1$: las muestras no son homogéneas

Así como se procedió para probar si dos variables son independientes, en este caso se utilizará el estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Su distribución bajo la hipótesis nula es $\chi^2$ con (r – 1)·(c – 1) grados de libertad. Para este ejemplo, la tabla cuenta con cuatro renglones (r) y cuatro columnas (c), por lo que la distribución tendrá (4 – 1)·(4 – 1) = 9 grados de libertad.

El cálculo de los valores esperados se realiza de la misma manera que la sección anterior.

**Tabla. Valores esperados conforme a la hipótesis nula**

| Grupo | Calificación del curso | | | | Total |
|---|---|---|---|---|---|
| | 5 | 6.0 a 7.5 | 7.6 a 8.5 | 8.6 a 10 | |
| **Matutino$_1$** | 7 | 50 | 8 | 7 | 72 |
| **Matutino$_2$** | 8 | 55 | 9 | 7 | 79 |
| **Vespertino$_1$** | 6 | 39 | 7 | 5 | 57 |
| **Vespertino$_2$** | 5 | 36 | 6 | 5 | 52 |

| | | | | | |
|---|---|---|---|---|---|
| **Total** | 26 | 180 | 30 | 24 | 260 |

Estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 = \frac{(7-7)^2}{7} + \frac{(50-51)^2}{51} + \frac{(9-9)^2}{9} + \frac{(6-7)^2}{7} + \cdots + \frac{(6-5)^2}{5}$$

$$\chi^2 = 1.1$$

Se rechazará la hipótesis nula si el estadístico de prueba es mayor al punto crítico.

El punto crítico de una distribución $\chi^2$ con 9 grados de libertad que separa la curva en dos regiones, una de 0.95 (izquierda del punto crítico) y otra de 0.05 (derecha del punto crítico), es el siguiente.

PRUEBA.CHI.INV(0.05, 9)=16.9

Como el valor de la prueba es menor al punto crítico, se acepta la hipótesis nula y se apoya que las muestras son homogéneas.

Como última observación, al utilizar el estadístico de prueba

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

se debe cuidar que los valores observados sean al menos de cinco. De no ser así, se sugiere juntar categorías para que se cumpla esta condición; de lo contrario, la prueba pierde precisión.

# RESUMEN

En esta unidad, se expuso la distribución $\chi^2$, su uso para contrastar hipótesis relacionadas con la varianza poblacional, diferencia de proporciones, bondad de ajuste, independencia y homogeneidad.

Se utilizaron dos estadísticos de prueba: $\boldsymbol{\chi^2 = \frac{(n-1)s^2}{\sigma^2}}$ para contrastar hipótesis relacionadas con la varianza poblacional, y $\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}}$ para el resto de las pruebas expuestas. Para que este último estadístico de prueba arroje resultados confiables, se debe observar que tanto la frecuencia observada como la esperada de las categorías sean al menos de cinco.

Como valor agregado, se utilizó Excel para el cálculo de los puntos críticos, que se ha venido practicando en unidades anteriores.

# BIBLIOGRAFÍA

## SUGERIDA

| Autor | Capítulo | Páginas |
|---|---|---|
| Anderson, S. | 11<br>12 | 449-471<br>472-505 |
| Levin, R. | 11 | 447-468 |
| Lind, D. | 17 | 648-679 |

Anderson, S. (2012). *Estadística para negocios y economía* (11.ª ed.). México: CENGAGE Learning.

Levin R. y Rubin D. (2010). *Estadística para administración y economía* (7.ª ed.). México: Pearson.

Lind A. D., Marchal G., W. y Wathen, S. (2012). *Estadística aplicada a los negocios y economía* (15.ª ed.). México: McGraw-Hill.

Licenciatura: Informática

# UNIDAD 6

# Análisis de regresión lineal simple


SUAyED
Sistema Universidad Abierta y Educación a Distancia
UNAM

Plan 2012 2016 actualizado

# OBJETIVO PARTICULAR

El alumno Interpretará los resultados obtenidos por medio del método de regresión lineal simple.

# TEMARIO DETALLADO
## (10 horas)

**6. Análisis de regresión lineal simple**

| |
|---|
| 6.1. Ecuación y recta de regresión |
| 6.2. El método de mínimos cuadrados |
| 6.3. Determinación de la ecuación de regresión |
| 6.4. El modelo de regresión y sus supuestos |
| 6.5. Inferencias estadísticas sobre la pendiente de la recta de regresión |
| 6.6. Análisis de correlación |

# INTRODUCCIÓN

Existen situaciones donde se requiere determinar si el comportamiento de cierto suceso se explica con el conocimiento de otra información. Por ejemplo, puede ser de interés conocer el impacto del número de horas de preparación para un examen de admisión a una institución de educación superior en el porcentaje de aciertos; o la afectación de los ingresos de una organización en función del presupuesto destinado a publicidad; o la duración de la batería de un dispositivo electrónico de acuerdo con el tiempo destinado a descargar tutoriales.

Para los problemas descritos en el párrafo anterior, se emplea el análisis de regresión lineal simple, técnica que trata de explicar una variable de interés o respuesta (y) en función de otra (x), mediante un modelo lineal.

En esta unidad, se mostrarán las características del modelo de regresión y el método para estimar sus parámetros. Una vez obtenidos los parámetros, se expone cómo determinar la ecuación del modelo, los supuestos que debe cumplir y la manera de realizar inferencias sobre la pendiente de la recta de regresión. La unidad concluye con el análisis de correlación lineal entre dos variables continuas.

# 6.1. Ecuación y recta de regresión

En este apartado, se tratarán los conceptos del modelo de regresión lineal simple. Para entender mejor este modelo, se repasará brevemente la ecuación de la recta.

**Ecuación de la recta**

En el plano cartesiano, la forma de describir una recta es mediante la ecuación

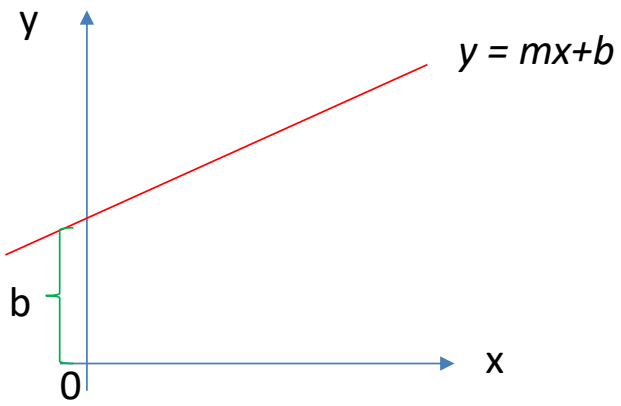
$$y = mx + b$$

- Donde:

$m$ = pendiente de la recta
$b$ = ordenada al origen o el punto donde intersecta la recta al eje Y, cuando x = 0

En la figura 1, se muestra una representación gráfica de la línea recta.

**Figura 1. Representación gráfica de la línea recta**




Fuente: elaboración propia.

La figura anterior ilustra la función de una línea recta con parámetros *m* y *b*. La pendiente *m* indica las unidades que se mueve *y* por cada unidad de cambio en *x*, y *b* es la intersección de la recta con el eje de las ordenadas.

Si m > 0, la recta tiene un ángulo de inclinación positivo; es decir, cada que aumenta *x*, aumenta *y*.

Si m < 0, la recta tiene un ángulo de inclinación negativo; es decir, cada que aumenta *x*, disminuye *y*.

Si m = 0 la recta es horizontal; es decir, cada que aumenta *x*, se mantiene constante *y* en *b*.

Para determinar la pendiente, es suficiente conocer dos puntos por donde atraviesa la recta $(x_1, y_1),(x_2, y_2)$ y aplicar la fórmula:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

Teniendo presente lo anterior, se presenta a continuación el modelo de regresión lineal simple.

**Modelo de regresión lineal**

El modelo de regresión lineal explica la relación entre una variable dependiente, a la que se denotará *y*, con otra(s) explicativa(s) a través de una ecuación de primer orden. Tanto las variables dependientes como las explicativas son observables.

Supóngase que una organización con 20 empleados realizó una evaluación del desempeño de cada empleado, y de acuerdo con el resultado se determinó un ajuste en el sueldo. Un auditor quiere explicar el incremento salarial conforme al desempeño del empleado.

En este ejemplo, el incremento salarial es la variable dependiente (y), ya que es resultado del desempeño de cada empleado (x). El incremento salarial observado del i-ésimo empleado (i = 1,2,..,20) se puede plantear de la siguiente manera:

Incremeto observado = Incremento esperado + variación (i= 1, 2, ..., 20)

Es decir, el incremento salarial observado del i-ésimo empleado tiene una parte explicable por la variable explicativa (nivel de desempeño observado) y otra no explicable, como puede ser una distracción del evaluador o su estado de salud al momento de la reunión.

Si denotamos como *y* al incremento salarial, como *x* al desempeño y como ε a la variación entre el incremento observado y estimado, entonces el incremento salarial del i-ésimo empleado (i = 1,2,..,20) se puede expresar así:

$$y_i = \mu(x_i) + \epsilon_i$$

Donde $\mu(x_i)$ representa el incremento esperado del i-ésimo empleado con su desempeño observado.

También $\mu(x_i)$ es un estimador de $y_i$ cuya estimación depende del valor de $x_i$. En el modelo de regresión lineal, la regla para estimar *y* consiste en relacionarla con *x* a través de una ecuación lineal.

Regresando al ejemplo, $\mu(x_i)$ puede expresarse así:

$$\boldsymbol{\mu(x_i) = \hat{y}_i = \beta_0 + \beta_1 x_i}$$

- Donde:

$\boldsymbol{\mu(x_i)}$ = estimador del incremento salarial del i-ésimo empleado (i=1,2,..,20) en función del desempeño observado
$\boldsymbol{\beta_0}$ = ordenada al origen de la recta de estimación
$\boldsymbol{\beta_1}$ = pendiente de la recta de estimación

Entonces, el auditor puede partir del siguiente modelo para determinar el criterio de incremento salarial de los empleados de la organización:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \epsilon_i}$$

Es el modelo de regresión lineal simple.

Ahora, cuando solamente se emplea una variable explicativa, al modelo de regresión lineal se le denomina *simple* y se modela con la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Donde:

$Y_i$ = variable dependiente o respuesta de la i-ésima observación
$\beta_0$ = intersección con el eje Y
$\beta_1$ = pendiente de la recta
$X_i$ = variable independiente o explicativa de la i-ésima observación
$\varepsilon_i$ = error no observable de la i-ésima observación
i = 1,2,.., n.

Cuando hay más de una variable explicativa, el modelo de regresión lineal es múltiple y se modela con la siguiente ecuación:

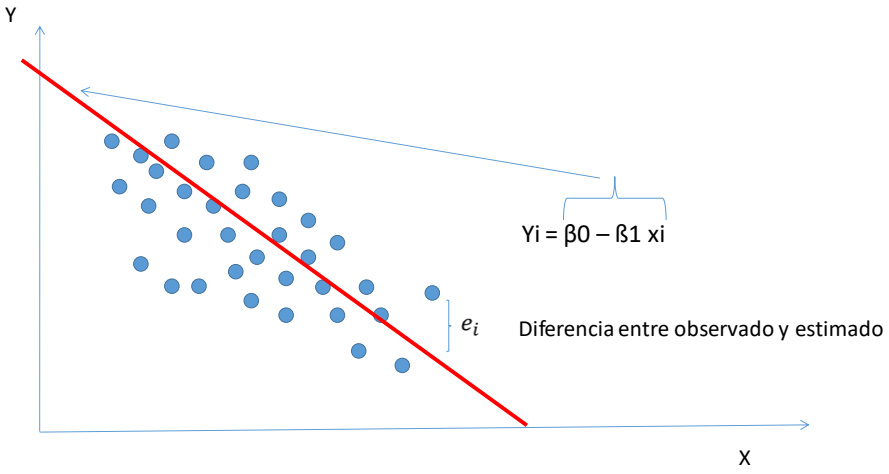
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- Donde:

$Y$ = variable dependiente o respuesta con *n* observaciones
$\beta_0$: intersección con el eje Y
$\beta_1, \beta_2, \ldots, \beta_p$ = razón de cambio de Y respecto a cada variable explicativa manteniendo el resto sin cambio.
$X_1, X_2$ y $X_p$ = variables independientes o explicativas, cada una de *n* observaciones
$\varepsilon$: error entre Y observada y estimada

Este material de estudio se enfocará al modelo de regresión lineal simple, en el cual se estima una recta que cruce a lo largo de la información con la intención de explicar el comportamiento de la variable de interés, como lo ilustra la figura 2.

**Figura 2. Ilustración del modelo de regresión lineal simple**




Fuente: elaboración propia.

La figura anterior ilustra un gráfico de dispersión donde cada punto azul representa el valor de la variable respuesta (Y) observado con el valor de la variable explicativa (X), la línea roja es la recta estimada que se ajusta al conjunto de datos, cuya ecuación es $\boldsymbol{Y_i = \beta_0 - \beta_1 X_i}$, y la diferencia entre el valor observado y el estimado con la ecuación de regresión lineal es el error.

En el ejemplo de los incrementos salariales de la organización de 20 empleados, en el eje X se representaría el desempeño del empleado; y en el eje Y, el incremento salarial. Los puntos azules serían el incremento salarial observado de cada empleado asociado a su desempeño; y la línea roja, el modelo de regresión lineal simple. En el siguiente apartado, se explica cómo calcular la recta de regresión lineal simple.

# 6.2. El método de mínimos cuadrados

En la parte final de la sección anterior, en la figura 2 se ilustró cómo la recta de regresión lineal simple atraviesa el conjunto de datos; sin embargo, el número de rectas que se pueden trazar es infinito, por lo que surge la pregunta sobre cuál es la recta conveniente. La respuesta no es difícil, dado que lo deseable es que la diferencia entre el valor estimado y observado de una observación sea la menor posible.

Partiendo del modelo para una observación cualquiera:

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Entonces, el error es la diferencia entre los valores observados y estimados:

- $y_i - \beta_0 - \beta_1 x_i = \varepsilon_i$

Error de todas las observaciones (n):

- $\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i = \sum_{i=1}^{n} \varepsilon_i$

Como se explicó en la sección anterior, la recta $\beta_0 + \beta_1 x_i$ es un valor esperado de $y_i$, por lo que la suma de las diferencias entre los valores estimados y observados se espera sea cero. Para superar este inconveniente, se procede a trabajar con los errores al cuadrado, los cuales quedan expresados así:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

La recta que se busca es de parámetros $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$ y minimiza la expresión del lado derecho. A esta metodología para obtener la recta que garantiza el menor error de estimación se le conoce como *mínimos cuadrados*.

Los valores de los parámetros $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$, por el método de mínimos cuadrados, son los siguientes:

$$\boldsymbol{\beta_0} = \bar{y} - \boldsymbol{\beta_1}\bar{x}$$

$$\boldsymbol{\beta_1} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

- Donde:

$\boldsymbol{\beta_0}$ : intersección con el eje Y
$\boldsymbol{\beta_1}$ : pendiente de la recta de regresión lineal simple
$\bar{y}$: promedio de la variable dependiente
$\bar{x}$: promedio de la variable independiente
n: número de observaciones
$x_i$: i-ésima observación de la variable independiente (i = 1,..,n)
$y_i$: i-ésima observación de la variable dependiente (i = 1,..,n)

A continuación, se muestra a manera de ejemplo cómo estimar una recta de regresión lineal simple por mínimos cuadrados.

Una PYME que imparte clases de manejo a personas de entre 30 y 65 años, para negociar las condiciones de su póliza de accidentes con la compañía de seguros que les ofrece el servicio, quiere conocer la relación entre el número de accidentes automovilísticos en la localidad donde se encuentra el negocio. La información se presenta a continuación.

**Accidentes automovilísticos por edad del conductor**

| ID | Edad | Accidentes | ID | Edad | Accidentes |
|---|---|---|---|---|---|
| 1 | 30 | 1,004 | 19 | 48 | 504 |
| 2 | 31 | 946 | 20 | 49 | 432 |
| 3 | 32 | 914 | 21 | 50 | 456 |
| 4 | 33 | 742 | 22 | 51 | 346 |
| 5 | 34 | 714 | 23 | 52 | 382 |
| 6 | 35 | 842 | 24 | 53 | 334 |
| 7 | 36 | 744 | 25 | 54 | 298 |
| 8 | 37 | 792 | 26 | 55 | 252 |
| 9 | 38 | 844 | 27 | 56 | 240 |
| 10 | 39 | 722 | 28 | 57 | 244 |
| 11 | 40 | 982 | 29 | 58 | 288 |
| 12 | 41 | 644 | 30 | 59 | 218 |
| 13 | 42 | 594 | 31 | 60 | 208 |
| 14 | 43 | 604 | 32 | 61 | 146 |
| 15 | 44 | 480 | 33 | 62 | 130 |
| 16 | 45 | 570 | 34 | 63 | 130 |
| 17 | 46 | 440 | 35 | 64 | 122 |
| 18 | 47 | 410 | 36 | 65 | 104 |

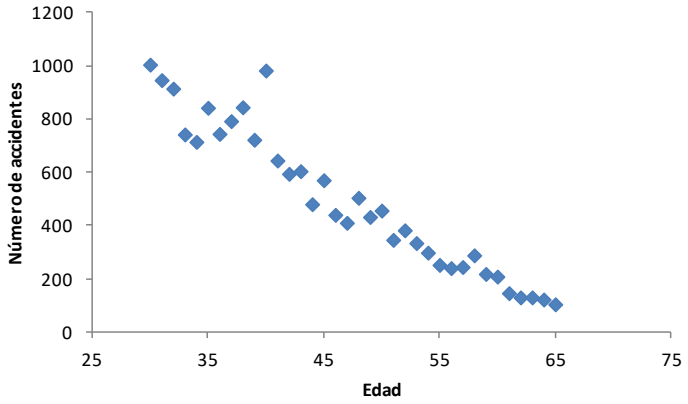
Para obtener la recta de regresión por mínimos cuadrados, se dan los siguientes pasos:

1. Determinar las variables dependientes *(Y)* e independiente(*X*).

En este problema, *Y* es el número de accidentes y *X* la edad del conductor debido a que el número de accidentes será explicado por la edad del conductor.

2. Graficar las variables *X* y *Y.*

**Gráfica 1. Número de accidentes por edad del conductor**




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

En la gráfica 1, se ilustra el número de accidentes (Y) respecto a la edad del conductor (X). Se aprecia como patrón que, conforme el conductor es mayor, el riesgo de tener un accidente disminuye.

3. Calcular los parámetros de la recta de regresión que atraviesa el conjunto de datos por mínimos cuadrados.

A continuación, se calcula la pendiente de la recta:

$$\beta_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Obsérvese que en la fórmula se requieren cinco sumas, cuyo cálculo se muestra en la siguiente tabla.

**Tabla 1. Memoria de cálculo de los elementos de la fórmula para calcular $\beta_1$ mediante mínimos cuadrados**

| | 1<br>$X_i$<br>Edad | 2<br>$Y_i$<br>Número de accidentes | 1-2<br>$X_iY_i$ | | $(1)^2$<br>$X_i^2$ | n |
|---|---|---|---|---|---|---|
| | 30 | 1004 | 30120 | | 900 | 36 |
| | 31 | 946 | 29326 | | 961 | |
| | 32 | 914 | 29248 | | 1024 | |
| | 33 | 742 | 24486 | | 1089 | |
| | 34 | 714 | 24276 | | 1156 | |
| | 35 | 842 | 29470 | | 1225 | |
| | 36 | 744 | 26784 | | 1296 | |
| | 37 | 792 | 29304 | | 1369 | |
| | 38 | 844 | 32072 | | 1444 | |
| | 39 | 722 | 28158 | | 1521 | |
| | 40 | 982 | 39280 | | 1600 | |
| | 41 | 644 | 26404 | | 1681 | |
| | 42 | 594 | 24948 | | 1764 | |
| | 43 | 604 | 25972 | | 1849 | |
| | 44 | 480 | 21120 | | 1936 | |
| | 45 | 570 | 25650 | | 2025 | |
| | 46 | 440 | 20240 | | 2116 | |
| | 47 | 410 | 19270 | | 2209 | |
| | 48 | 504 | 24192 | | 2304 | |
| | 49 | 432 | 21168 | | 2401 | |
| | 50 | 456 | 22800 | | 2500 | |
| | 51 | 346 | 17646 | | 2601 | |
| | 52 | 382 | 19864 | | 2704 | |
| | 53 | 334 | 17702 | | 2809 | |
| | 54 | 298 | 16092 | | 2916 | |
| | 55 | 252 | 13860 | | 3025 | |
| | 56 | 240 | 13440 | | 3136 | |
| | 57 | 244 | 13908 | | 3249 | |
| | 58 | 288 | 16704 | | 3364 | |
| | 59 | 218 | 12862 | | 3481 | |
| | 60 | 208 | 12480 | | 3600 | |
| | 61 | 146 | 8906 | | 3721 | |
| | 62 | 130 | 8060 | | 3844 | |
| | 63 | 130 | 8190 | | 3969 | |
| | 64 | 122 | 7808 | | 4096 | |
| | 65 | 104 | 6760 | | 4225 | |
| $\sum X_i$ | **1710** | $\sum Y_i$ | **17822** | $\sum X_i Y_i$ | **748570** | $\sum X_i^2$ | **85110** | |
| $(\sum X_i)^2$ | **2924100** | | | | | | | |
| | | $\sum X_i \sum Y_i$ | **30475620** | | | | | |

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

La tabla anterior presenta el cálculo de los elementos de la fórmula de la pendiente de la recta de regresión de mínimos cuadrados. La primera columna contiene la edad del conductor (X); la segunda, el número de accidentes reportados para cada edad (Y). La tercera columna se obtiene multiplicando las dos primeras, por ejemplo, el primer elemento de esta columna (30,120) es resultado de multiplicar el primer valor de la primera (30) por el primer valor de la segunda (1,004). La cuarta columna es resultado de multiplicar la primera por sí misma. Regresando a analizar el primer elemento (900), este se obtuvo de multiplicar por sí mismo el primer elemento de la primera columna (30). En la parte final, se encuentran las sumas y multiplicaciones que se requiere sustituir en la fórmula.

Sustituyendo, la pendiente es la siguiente:

$$\boldsymbol{\beta_{1=}} \frac{(\mathbf{36} \cdot \mathbf{748570}) - \mathbf{30475620}}{(\mathbf{36} \cdot \mathbf{85110}) - \mathbf{2924100}}$$

$$\boldsymbol{\beta_{1=}} \frac{\mathbf{26948520} - \mathbf{30475620}}{\mathbf{3063960} - \mathbf{2924100}}$$

$$\boldsymbol{\beta_{1=}} \frac{\mathbf{-3527100}}{\mathbf{139860}}$$

$$\boldsymbol{\beta_{1=}} \mathbf{-25.218}$$

Y la ordenada al origen:

$$\boldsymbol{\beta_0} = \boldsymbol{\bar{y}} - \boldsymbol{\beta_1 \bar{x}}$$

$$\boldsymbol{\bar{Y}} = \frac{\mathbf{17822}}{\mathbf{36}}$$

$$\boldsymbol{\bar{Y}} = \mathbf{495.055}$$

$$\bar{X} = \frac{17822}{36}$$

$$\bar{X} = 47.5$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_0 = 495.055 - (-25.218 \cdot 47.5)$$

$$\beta_0 = 495.055 - 1197.892$$

$$\beta_0 = 1692.948$$

De esta manera, se obtienen los parámetros de la recta de regresión lineal simple con el método de mínimos cuadrados. En la siguiente sección, se expone cómo determinar la ecuación de regresión lineal simple.

## 6.3. Determinación de la ecuación de regresión

Como se ha mencionado, el modelo de regresión lineal simple estima el valor observado de la variable dependiente (Y) a partir de la explicativa (X) con la ecuación de una recta. Una vez determinados los valores de los parámetros mediante mínimos cuadrados, la estimación de los valores de *Y* se realiza con la ecuación de regresión lineal simple:

$$\widehat{Y_i} = \beta_0 + \beta_1 X_i$$

En el ejemplo anterior, $\beta_0 = 1692.948$ (1,693) y $\beta_1 = -25.218$ (-25.2) por lo que la ecuación de regresión lineal simple es la siguiente:

$$\widehat{Y_i} = 1,693 - 25.2X_i$$

- Donde:

$\widehat{Y_i}$ = estimación del número de accidentes para conductores en la i-ésima observación. (i=1,2,…,36)
$X_i$ = edad del conductor en la i-ésima observación. (i=1,2,…,36)

En esta ecuación, $\beta_0$ indica que, cuando X = 0, se espera observar 1693 accidentes, lo que en el contexto del problema no tiene sentido, porque la edad de interés es entre 30 y 65. Por otro lado, la pendiente de la ecuación tiene una dirección negativa, esto significa que, conforme se avance en edad, se espera observar menos accidentes. El valor de la pendiente (–25.2) indica que, por cada año que aumenta la edad del conductor, el número de accidentes disminuye en 25.

# 6.4. El modelo de regresión y sus supuestos

Un aspecto fundamental cuando se trabaja con esta técnica es que el modelo de regresión lineal simple es estimado con los valores de una muestra, por lo que los valores obtenidos de $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$ son estimaciones de los parámetros de la recta con toda la población[5]. Así, el propósito del modelo no es solamente calcular los parámetros, sino realizar inferencia sobre los verdaderos valores de esos parámetros. Por lo anterior, es necesario considerar los siguientes supuestos al emplear una regresión lineal simple.

1. En el modelo de regresión lineal simple

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i} (i = 1,.., n)$$

tanto la variable dependiente (Y) como la explicativa (X) son observables.

2. El modelo es lineal en los parámetros no en las variables. Esto significa que se pueden realizar transformaciones sobre las variables originales para que haya una relación lineal, y la esencia del modelo no se pierde.

3. El error de estimación $\boldsymbol{\varepsilon_i}$ es una variable aleatoria cuyo valor esperado es cero y su varianza es $\boldsymbol{\sigma^2}$, la cual se mantiene constante en todas las observaciones y es desconocida.

[5]**Los estimadores de $\beta_0$ y $\beta_1$ son insesgados.**

4. Los errores $\varepsilon_i$ son independientes. Esto significa que, dados dos valores cualesquiera de X, $x_i$, $x_j$ ($i \neq j$), los errores $\varepsilon_i$, $\varepsilon_j$ son independientes.[6]

5. El error $\varepsilon_i$ es una variable aleatoria con distribución normal. Al ser *y* una función lineal del error, también se distribuye normalmente.

Uno de los aspectos que más se descuida al ajustar un modelo de regresión lineal simple es revisar que se cumplan los supuestos del modelo (esta revisión implica analizar el comportamiento de los residuos). Como este tema no está incluido en el plan de estudios, no se abordará; sin embargo, se sugiere profundizarlo en Anderson (2012), parte de la bibliografía citada al término de la unidad.

## 6.5. Inferencias estadísticas sobre la pendiente de la recta de regresión

Como se mencionó en la sección anterior, el propósito del modelo de regresión lineal simple no se reduce a calcular los parámetros de la recta, sino que implica realizar inferencia sobre ellos. Cuando se ajusta un modelo de regresión, la primera prueba efectuada es referente a si un modelo lineal es el adecuado para los datos, y posteriormente se hacen inferencias sobre la pendiente. En este apartado, se expondrá como llevar a cabo inferencias sobre la pendiente de la recta de regresión.

**[6]O al menos no correlacionados.**

Para establecer inferencias con la pendiente del modelo, se contrastan las siguientes hipótesis:

$H_0$: $\boldsymbol{\beta_1} = 0$
$H_1$: $\boldsymbol{\beta_1} \neq 0$

La hipótesis nula significa que el valor de la pendiente del modelo no es importante: la variable X no tiene efecto sobre Y, es decir, X no es una variable explicativa de Y.

La hipótesis alternativa plantea que el valor de la pendiente sí es importante: X tiene efecto sobre Y.

Rechazar la hipótesis nula significa que la variable X es una variable explicativa de Y. Esto implica que el modelo puede aplicarse.

El estadístico de prueba empleado para contrastar la hipótesis nula es el siguiente:

$$t = \frac{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1}{s} \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- Donde:

$\widehat{\boldsymbol{\beta}_1}$ = estimador de la pendiente de la recta de regresión
$\boldsymbol{\beta}_1$ = pendiente de la recta de regresión asumiendo cierta la hipótesis nula
s = estimador de la desviación estándar, el cual es

$$S = \sqrt{\frac{\sum (Y_i - \hat{Y})^2}{n-2}}$$

El estadístico de prueba tiene una distribución t de Student con n – 2 grados de libertad. En la figura 3, se ilustra una prueba ubicada en zona de rechazo.

**Figura 3. Ilustración de una prueba donde se rechaza la hipótesis nula**


Fuente: elaboración propia.

La figura 3 ilustra una prueba donde el estadístico de prueba se ubica en la zona de rechazo, lo que significa que la pendiente tiene un valor significativo. Al final de la unidad, se muestra un ejemplo de cómo realizar inferencias de la pendiente con Microsoft Excel (2013).

En el ejemplo de los accidentes, se mencionó que el modelo ajustado es

$$\widehat{Y_i} = 1,693 - 25.2X_i$$

La pregunta es, entonces, si los coeficientes son significativos. Para responder esto, se realiza la prueba de hipótesis, donde $H_0$ es que los coeficientes son cero (no tienen un valor significativo). El resultado de la prueba se muestra a continuación.

| | *Coeficientes* | *Error típico* | *Estadístico t* | *Probabilidad* |
|---|---|---|---|---|
| Intercepción | 1692.9 | 58.6 | 28.9 | 1.64442E-25 |
| Edad | -25.2 | 1.2 | -20.9 | 5.35232E-21 |

Fuente: Microsoft Excel (2013). Módulo de análisis de datos.

La tabla anterior muestra los valores de los coeficientes del modelo, su error, su estadístico de prueba y resaltado. Se ve la significancia de la prueba (p value), y como esta prueba es menor a 0.05, se rechaza $H_0$: los coeficientes son significativos.

## 6.6. Análisis de correlación

En el análisis de regresión lineal simple, si la variable *X* es explicativa de *Y*, entonces el modelo muestra el efecto de un cambio en *X* sobre *Y*. Un análisis complementario es el de correlación, el cual determina el grado de asociación lineal entre dos variables.

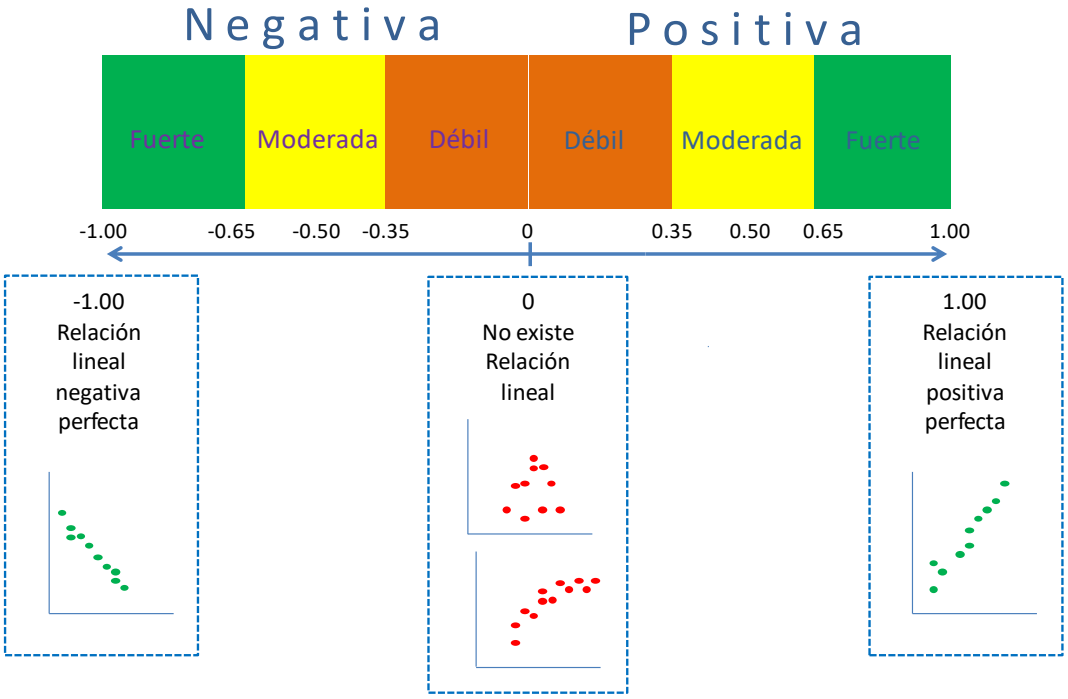
La correlación entre las variables *X* y *Y* se denota como $\rho_{xy}$, y se define como el grado en que se encuentran asociadas estas variables. El estimador de esta correlación es conocido como *coeficiente de correlación*, denotado como *r*, y su fórmula es

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

El coeficiente de correlación es un valor independiente de las unidades de las variables, lo que permite que pueda ser empleado en comparativos; toma valores entre –1 y 1 (en –1 significa que existe una asociación lineal perfecta negativa, es decir, el incremento de la variable explicativa resultará una disminución en la variable respuesta; y en 1, la asociación lineal entre las variables es perfecta y positiva, lo que implica que un aumento de la variable explicativa hará que aumente el valor de la variable respuesta). Cuando el coeficiente de correlación es cero, significa que las variables no están asociadas o que su asociación no es lineal.

La figura 4 muestra una categorización de la asociación entre dos variables en función del valor del coeficiente de correlación.

**Figura 4. Nivel de asociación de dos variables de acuerdo con el valor del coeficiente de correlación**




Fuente: elaboración propia.

En la figura anterior, se muestra cómo interpretar los niveles de asociación entre dos variables de acuerdo con el valor del coeficiente de correlación. Un valor mayor a cero indica que existe una correlación positiva; en caso contrario, la correlación es negativa. Las variables se considerarán con una asociación débil si su correlación tiene un valor absoluto entre 0 y 35; moderada, entre 35 y 65; y fuerte, mayor a 65.

Para el ejemplo del número de accidentes por edad del conductor, la correlación entre las dos variables es de –0.9633, lo que significa que la asociación entre las variables es casi negativa perfecta.

La tabla 2 muestra la memoria de cálculo de los elementos que forman parte de la fórmula de la correlación de las variables. En la parte superior de la tabla, se numera la columna (del 1 al 9) y en algunos casos, debajo de este número, se indican las columnas involucradas en la obtención de sus cifras. Por ejemplo, los valores de la columna 5 se obtienen de restarle a la edad (columna 1) el promedio de edad (columna 2). Los valores involucrados en la fórmula del coeficiente de correlación son los dos que se hallan en la parte inferior derecha, y al sustituirlos se obtiene lo siguiente:

$$r = \frac{-97075}{\sqrt{103444464748}}$$

$$r = \frac{-97075}{101707.7418}$$

$$r = -0.9633$$

Es decir, el resultado comentado.

**Tabla 2. Memoria de cálculo de los elementos de la fórmula para calcular *r* entre el número de accidentes y la edad del conductor**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| $X_i$ | $\overline{X}$ | $Y_i$ | $\overline{Y}$ | (1-2) $(X_i-\overline{X})$ | (1-2)$^2$ $(X_i-\overline{X})^2$ | (3-4) $(Y_i-\overline{Y})$ | (3-4)$^2$ $(Y_i-\overline{Y})^2$ | 5-7 |
| Edad | Promedio de X | Número de accidentes | Promedio de Y | | | | | |
| 30 | 47.5 | 1004 | 495.06 | -17.5 | 306.25 | 508.94 | 259024.45 | -8906.52778 |
| 31 | | 946 | | -16.5 | 272.25 | 450.94 | 203350.89 | -7440.58333 |
| 32 | | 914 | | -15.5 | 240.25 | 418.94 | 175514.45 | -6493.63889 |
| 33 | | 742 | | -14.5 | 210.25 | 246.94 | 60981.56 | -3580.69444 |
| 34 | | 714 | | -13.5 | 182.25 | 218.94 | 47936.67 | -2955.75 |
| 35 | | 842 | | -12.5 | 156.25 | 346.94 | 120370.45 | -4336.80556 |
| 36 | | 744 | | -11.5 | 132.25 | 248.94 | 61973.34 | -2862.86111 |
| 37 | | 792 | | -10.5 | 110.25 | 296.94 | 88176.00 | -3117.91667 |
| 38 | | 844 | | -9.5 | 90.25 | 348.94 | 121762.23 | -3314.97222 |
| 39 | | 722 | | -8.5 | 72.25 | 226.94 | 51503.78 | -1929.02778 |
| 40 | | 982 | | -7.5 | 56.25 | 486.94 | 237114.89 | -3652.08333 |
| 41 | | 644 | | -6.5 | 42.25 | 148.94 | 22184.45 | -968.138889 |
| 42 | | 594 | | -5.5 | 30.25 | 98.94 | 9790.00 | -544.194444 |
| 43 | | 604 | | -4.5 | 20.25 | 108.94 | 11868.89 | -490.25 |
| 44 | | 480 | | -3.5 | 12.25 | -15.06 | 226.67 | 52.6944444 |
| 45 | | 570 | | -2.5 | 6.25 | 74.94 | 5616.67 | -187.361111 |
| 46 | | 440 | | -1.5 | 2.25 | -55.06 | 3031.11 | 82.5833333 |
| 47 | | 410 | | -0.5 | 0.25 | -85.06 | 7234.45 | 42.5277778 |
| 48 | | 504 | | 0.5 | 0.25 | 8.94 | 80.00 | 4.47222222 |
| 49 | | 432 | | 1.5 | 2.25 | -63.06 | 3976.00 | -94.5833333 |
| 50 | | 456 | | 2.5 | 6.25 | -39.06 | 1525.34 | -976388889 |
| 51 | | 346 | | 3.5 | 12.25 | -149.06 | 22217.56 | -521.694444 |
| 52 | | 382 | | 4.5 | 20.25 | -113.06 | 12781.56 | -508.75 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 53 | 334 | 5.5 | 30.25 | -161.06 | 25938.89 | | -885.805556 | |
| 54 | 298 | 6.5 | 42.25 | -197.06 | 38830.89 | | -1280.86111 | |
| 55 | 252 | 7.5 | 56.25 | -243.06 | 59076.00 | | -1822.91667 | |
| 56 | 240 | 8.5 | 72.25 | -255.06 | 65053.34 | | -2167.97222 | |
| 57 | 244 | 9.5 | 90.25 | -251.06 | 63028.89 | | -2385.027778 | |
| 58 | 288 | 10.5 | 110.25 | -207.06 | 42872.00 | | -2174.08333 | |
| 59 | 218 | 11.5 | 132.25 | -277.06 | 76759.78 | | -3186.13889 | |
| 60 | 208 | 12.5 | 156.25 | -287.06 | 82400.89 | | -3588.19444 | |
| 61 | 146 | 13.5 | 182.25 | -349.06 | 121839.78 | | -4712.25 | |
| 62 | 130 | 14.5 | 210.25 | -365.06 | 133265.56 | | -5293.30556 | |
| 63 | 130 | 15.5 | 240.25 | -365.06 | 133265.56 | | -5658.36111 | |
| 64 | 122 | 16.5 | 272.25 | -373.06 | 139170.45 | | -6155.41667 | |
| 65 | 104 | 17.5 | 306.25 | -391.06 | 152924.45 | | -6843.47222 | |
| | | | 3885 | | 2662667 | $\sum(X_i-\bar{X})(Y_i-\bar{Y})$ | -97975 | $\sum(X_i-\bar{X})^2(Y_i-\bar{Y})^2$ |
| | | | | | | | | **10344464748** |

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

**Coeficiente de determinación $R^2$**

Para valorar el ajuste del modelo de regresión lineal simple, se considera otro coeficiente llamado *coeficiente de determinación*, denotado como $R^2$, que mide la variabilidad explicada por el modelo. Para calcular el coeficiente de determinación, se utiliza la siguiente fórmula:

$$R^2 = \frac{\sum(\widehat{Y_i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}$$

Donde:

$R^2$: coeficiente de determinación
$\widehat{Y_i}$ : i-ésima estimación de Y
$Y_i$: i-ésima observación de Y
$\overline{Y}$: promedio de Y

Para el ejemplo del número de accidentes por edad del conductor, el coeficiente de determinación del modelo ajustado entre las dos variables 0.9279, esto significa que el modelo explica en un 93% la variabilidad de la información. La tabla 3 muestra el cálculo de los elementos que intervienen en la fórmula de $R^2$.

**Tabla 3. Memoria de cálculo de los elementos de la fórmula para calcular $R^2$ entre el número de accidentes y la edad del conductor**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Xi | Yi | $\hat{Y}$i | $\bar{Y}$ | (3-4) ($\hat{Y}$i-$\bar{Y}$) | (5)2 ($\hat{Y}$i-$\bar{Y}$)2 | (2-4) (Yi-$\bar{Y}$) | (7)2 (Yi-$\bar{Y}$)2 |
| Edad | Número de accidentes | (-25.22 edad conductor) | Promedio de Y | | | | |
| 30 | 1004 | 936 | 495.06 | 441 | 194771.14 | 508.94 | 259024.448 |
| 31 | 946 | 911 | | 416 | 173147.56 | 450.94 | 203350.892 |
| 32 | 914 | 886 | | 391 | 152795.97 | 418.94 | 175514.448 |
| 33 | 742 | 861 | | 366 | 133716.35 | 246.94 | 60981.5586 |
| 34 | 714 | 836 | | 340 | 115908.70 | 218.94 | 47936.6698 |
| 35 | 842 | 810 | | 315 | 99373.03 | 346.94 | 120370.448 |
| 36 | 744 | 785 | | 290 | 84109.33 | 248.94 | 61973.3364 |
| 37 | 792 | 760 | | 265 | 70117.61 | 296.94 | 88176.0031 |
| 38 | 844 | 735 | | 240 | 57397.86 | 348.94 | 121762.225 |
| 39 | 722 | 709 | | 214 | 45950.09 | 226.94 | 51503.7809 |
| 40 | 982 | 684 | | 189 | 35774.29 | 486.94 | 237114.892 |
| 41 | 644 | 659 | | 164 | 26870.47 | 148.94 | 22184.4475 |
| 42 | 594 | 634 | | 139 | 19238.62 | 98.94 | 9790.00309 |
| 43 | 604 | 609 | | 113 | 12878.74 | 108.94 | 11868.892 |
| 44 | 480 | 583 | | 88 | 7790.85 | -15.06 | 226.669753 |
| 45 | 570 | 558 | | 63 | 3974.92 | 74.94 | 5616.66975 |
| 46 | 440 | 533 | | 38 | 1430.97 | -55.06 | 3031.1142 |
| 47 | 410 | 508 | | 13 | 159.00 | -85.06 | 7234.44753 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 48 | 504 | 482 | -13 | | 159.00 | 8.94 | 80.0030864 |
| 49 | 432 | 457 | -38 | | 1430.97 | -63.06 | 3976.00309 |
| 50 | 456 | 432 | -63 | | 3974.92 | -39.06 | 1525.33642 |
| 51 | 346 | 407 | -88 | | 7790.85 | -149.06 | 22217.5586 |
| 52 | 382 | 382 | -113 | | 12878.74 | -113.06 | 12781.5586 |
| 53 | 334 | 356 | -139 | | 19238.52 | -161.06 | 25938.892 |
| 54 | 298 | 331 | -164 | | 26870.47 | -197.06 | 38830.892 |
| 55 | 252 | 306 | -189 | | 35774.29 | -243.06 | 59076.0031 |
| 56 | 240 | 281 | -214 | | 45950.09 | -255.06 | 65053.3364 |
| 57 | 244 | 255 | -240 | | 57397.86 | -251.06 | 63028.892 |
| 58 | 288 | 230 | -265 | | 70117.61 | -207.06 | 42872.0031 |
| 59 | 218 | 205 | -290 | | 84109.33 | -277.06 | 76759.7809 |
| 60 | 208 | 180 | -315 | | 99373.03 | -287.06 | 82400.892 |
| 61 | 146 | 155 | -340 | | 115908.70 | -349.06 | 121839.781 |
| 62 | 130 | 129 | -366 | | 133716.35 | -365.06 | 133265.559 |
| 63 | 130 | 104 | -391 | | 152795.97 | -365.06 | 133265.559 |
| 64 | 122 | 79 | -416 | | 173147.56 | -373.06 | 139170.448 |
| 65 | 104 | 54 | -441 | | 194771.14 | -391.06 | 152924.448 |
| | | | | $\sum(\hat{Y}i-\bar{Y})2$ | 2470810.97 | | $\sum(Yi-\bar{Y})2$ 2662667 |

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

Así como en la tabla 2, en la parte superior de la tabla 3 se numera la columna (del 1 al 8), y en algunos casos, debajo de este número, se indican las columnas involucradas en la obtención de sus cifras. Por ejemplo, los valores de la columna 5 se obtienen de restarle a los accidentes estimados (columna 3) el promedio observado de accidentes (columna 4). Los valores involucrados en la fórmula del coeficiente de determinación son los dos que se sitúan en la parte inferior de la tabla; y al sustituirlos se obtiene:

$$R^2 = \frac{2470810.97}{2662667.89}$$

$$R^2 = 0.927945$$

Es decir, el resultado comentado.

**Análisis de regresión lineal simple con MS -Excel**

Al igual que otras técnicas de análisis, Microsoft Excel (2013) permite realizar regresión lineal simple en el módulo de *análisis de datos*. A continuación, se muestra el uso de esta herramienta con los datos del ejemplo de los accidentes registrados por edad del conductor.

Capturada la información en Excel, ir al menú de Datos, y seleccionar la opción Análisis de datos, previamente cargada. Se desplegará una ventana de diálogo de funciones para análisis, seleccionar Regresión.


Fuente: Microsoft Excel (2013).

Se despliega una nueva ventana de diálogo, donde se ingresa la información y se determina la salida que se desea obtener. En el rango Y de entrada, seleccionar los datos de la variable dependiente, es decir, el número de accidentes.


Fuente: Microsoft Excel (2013).

En el rango X de entrada, seleccionar los datos de la variable independiente, es decir, la edad.

Fuente: Microsoft Excel (2013).

En opciones de Salida, indicar el rango de salida; y en la sección de Residuales, la opción de Residuos. Finalmente, dar en Aceptar.

Es importante señalar que el nivel de significancia no se modifica. Excel toma por defecto el 95% de confianza valor, medida base para determinar si nuestro modelo y los parámetros de la ecuación lineal son o no significativos.

Fuente: Microsoft Excel (2013).

El cuadro-resumen que se proporciona es el siguiente. Algunas de las medidas señaladas con azul fueron calculadas previamente.

RESUMEN

| *Estadísticas dele regresión* | | |
|---|---|---|
| Coeficiente de correlación múltiple | 0.963299334 | **r** |
| Coeficiente de determinación R^2 | 0.927945607 | **$R^2$** |
| R^2 ajustado | 0.925826361 | |
| Error típico | 75.11890911 | **S** |
| Observaciones | 36 | **n** |

ANÁLISIS DE VARIANZA

| | *Grados de libertad* | *Suma de cuadrados* | *Promedio de los cuadrados* | *F* | *Valor crítico de F* |
|---|---|---|---|---|---|
| Regresión | 1 | 2470810.972 | 2470810.972 | 437.8657505 | 5.35232E-21 |
| Residuos | 34 | 191856.9172 | 5642.850506 | | |
| Total | 35 | 2662667.889 | | | |

| | *Coeficientes* | *Error Típico* | *Estadístico t* | *Probabilidad* | *Inferior 95%* | *Superior 95%* |
|---|---|---|---|---|---|---|
| Intercepción | 1692.948091 | 58.59935094 | 28.89021915 | 1.64442E-25 | 1573.859882 | 1812.036299 |
| Variable X1 | -25.21879022 | 1.20518512 | -20.92524195 | 5.35232E-21 | -27.66802105 | -22.76955939 |

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

Los resultados señalados con morado indican la significancia del modelo y de cada uno de los parámetros. El primero (valor crítico de F) señala que el modelo lineal es adecuado para la información que se analiza, pues es significativo por ser menor a 0.05. En el caso de los parámetros, dado que las probabilidades son menores a 0.05, se rechaza la hipótesis nula de que los parámetros no son significativos y pueden emplearse sin inconveniente en la ecuación.

Otra manera de calcular los parámetros $\beta_0$ y $\beta_1$ es con las funciones

*intersección.eje ()*
*pendiente()*

El empleo de estas funciones se ilustrará en la siguiente unidad.

# RESUMEN

Se expusieron las bases para realizar un análisis de regresión lineal simple con la información de dos variables observadas. En primer lugar, se mostró la ecuación empleada en el modelo de regresión lineal simple partiendo de un repaso de la ecuación general de la recta, y siguiendo con la metodología de mínimos cuadrados para estimar la recta que garantiza el menor error de estimación.

Calculados los parámetros del modelo, se planteó con un ejemplo la interpretación de la pendiente y se enunciaron los supuestos que debe cumplir el modelo (es habitual no comprobar esto en la práctica, por lo cual se sugiere profundizar en el análisis de los residuos).

Después se revisó la forma de realizar inferencia sobre la pendiente, y el cálculo de los coeficientes de correlación y determinación, los cuales indican, respectivamente, el grado de asociación entre las variables y la variabilidad explicada por el modelo de regresión lineal simple.

La unidad finaliza con un ejemplo de cómo ajustar un modelo de regresión lineal simple con el módulo de análisis de datos de Microsoft Excel (2013).

# BIBLIOGRAFÍA

## SUGERIDA

| Autor | Capítulo | Páginas |
|---|---|---|
| Anderson, S. | 14 | 560-641 |
| Levin, R. | 12 | 509-564 |
| Lind, D. | 13 | 461-511 |

Anderson, S. (2012). *Estadística para negocios y economía* (11.ª ed.). México: CENGAGE Learning.

Levin R. y Rubin D. (2010). *Estadística para administración y economía* (7.ª ed.). México: Pearson.

Lind A. D., Marchal G., W. y Wathen, S. (2012). *Estadística aplicada a los negocios y economía* (15.ª ed.). México: McGraw-Hill.

Licenciatura: Informática

# UNIDAD 7

# Análisis de series de tiempo


SUAyED
Sistema Universidad Abierta y Educación a Distancia
UNAM

Plan 2012 2016 actualizado

# OBJETIVO PARTICULAR

El alumno aplicará los métodos para el análisis de series de tiempo.

# TEMARIO DETALLADO
## (8 horas)

**7. Análisis de series de tiempo**

- 7.1. Los cuatro componentes de una serie de tiempo
- 7.2. Análisis gráfico de la tendencia
- 7.3. Tendencia secular
- 7.4. Variaciones estacionales
- 7.5. Variaciones cíclicas
- 7.6. Fluctuaciones irregulares
- 7.7. Modelos autorregresivos de promedios móviles

# INTRODUCCIÓN

A lo largo del curso, se ha insistido en que la estadística inferencial contribuye a la toma de decisiones que, frecuentemente, deben realizarse con información recabada en el tiempo. Por ejemplo, para un inversionista, el conocimiento de los estados de resultados de una empresa durante los últimos cinco años le ayudaría a decidir si invierte en acciones de esa compañía. O la disposición de dinero en los cajeros automáticos permitiría determinar la cantidad de efectivo que la institución bancaria debe abastecer cada semana para garantizar el servicio de sus cuentahabientes. O el historial reciente de pagos de una persona facilitaría a una micro financiera dedicada a dar créditos de autos a determinar si el individuo es sujeto de crédito.

Los ejemplos anteriores ilustran la aplicación del análisis de series de tiempo. En esta unidad, se expondrá de manera básica el empleo de esta técnica (es labor del estudiante profundizar en otras fuentes). En primer lugar, se define qué es una serie de tiempo y se exponen los componentes que suelen integrarla. Después, se muestra cómo realizar un análisis exploratorio con el apoyo de una gráfica que permita visualizar la tendencia de la serie. El siguiente punto describe algunas metodologías para trabajar la tendencia de una serie de tiempo a partir del manejo de variaciones estacionales, cíclicas y fluctuaciones irregulares. Por último, se abordan de manera breve las series estacionales y los modelos auto regresivos y de medias móviles.

# 7.1. Los cuatro componentes de una serie de tiempo

Una serie de tiempo es el registro de una variable a lo largo del tiempo realizado con una periodicidad constante, por ejemplo, de forma diaria, semanal, mensual o anual. La observación tomada en el tiempo *t* de una variable se denotará como $Y_t$.

Las series de tiempo son aplicables por lo regular en todas las áreas de conocimiento: en el índice nacional de precios al consumidor (INPC), tasa de desempleo, cotización diaria del dólar norteamericano, evolución de los niveles de colesterol de un paciente sometido a un estudio clínico en el que se estudia el efecto de un medicamento, o las calificaciones de un alumno que periódicamente es sometido a evaluaciones.

De acuerdo con la forma como se registra su información, las series se dividen en discretas o continuas. Una serie de tiempo es discreta si las observaciones son realizadas en momentos específicos, normalmente con una misma periodicidad (por ejemplo, el número anual de suscriptores a una publicación). Y es continua si las observaciones se registran de forma continua en el tiempo (como el ritmo cardiaco de un paciente durante un examen médico).

Para facilitar el estudio de las series de tiempo, se dividen en cuatro partes:

| |
|---|
| a) Componente de tendencia (T) |
| b) Componente estacional (E) |
| c) Componente cíclico (C) |
| d) Componente de fluctuaciones irregulares (I) |

Consideremos que no siempre se encuentran presentes los cuatro componentes en una serie de tiempo. En las siguientes secciones, se explicarán cada uno de estos componentes y su manejo.

Hay dos enfoques para asociar la serie de tiempo con sus componentes: aditivo y multiplicativo. En el primero, la serie de tiempo se considera que es resultado de la suma de sus componentes. De esta manera, la serie de tiempo $Y_t$ queda expresada así:

$$Y_t = T_t + E_t + C_t + I_t$$

- Donde:

$Y_t$ = valor de la serie al tiempo *t*
$T_t$ = componente de tendencia al tiempo *t*
$E_t$ = componente estacional al tiempo *t*
$C_t$ = componente de cíclico al tiempo *t*
$I_t$ = componente irregular o aleatorio al tiempo *t*

Y en el enfoque multiplicativo, la serie de tiempo se considera que es resultado de ajustar la tendencia con factores asociados a los otros componentes, por lo que la serie de tiempo $Y_t$ queda expresada así:

$$Y_t = T_t * E_t * C_t * I_t$$

- Donde:

$Y_t$ = valor de la serie al tiempo *t*
$T_t$ = componente de tendencia al tiempo *t*
$E_t$ = factor estacional al tiempo *t*
$C_t$ = factor cíclico al tiempo *t*
$I_t$ = factor irregular o aleatorio al tiempo *t*

## 7.2. Análisis gráfico de la tendencia

El primer paso para analizar una serie de tiempo es realizar, a modo de análisis exploratorio, una gráfica de líneas, donde en el eje X se ubicará el tiempo y en el eje Y el valor de la serie a lo largo del periodo. El análisis gráfico permitirá visualizar los componentes de la serie (por lo regular, la tendencia es el componente más evidente).

Una serie de tiempo muestra una tendencia si existe un crecimiento o disminución durante el periodo que se está analizando. Si la gráfica de la serie muestra un crecimiento continuo a lo largo del tiempo, se dice que la serie tiene una tendencia positiva (véase figura 1).

**Figura 1. Serie de tiempo con tendencia positiva**


Fuente: elaboración propia.

La figura anterior muestra una serie cuyo valor en general se incrementa a medida que va transcurriendo el tiempo.

Si la gráfica expresa un decrecimiento continuo a lo largo del tiempo, se dice que la serie presenta una tendencia negativa (véase figura 2).

**Figura 2. Serie de tiempo con tendencia negativa**

Fuente: elaboración propia.

La figura anterior muestra una serie cuyo valor, en general, decrece conforme transcurre el tiempo.

Una serie sin tendencia presentará variaciones alrededor de un solo valor a lo largo del tiempo, similar a lo que la presenta la figura 3.

**Figura 3. Serie de tiempo sin tendencia**

Fuente: elaboración propia.

En el análisis de series de tiempo, la realización de una gráfica es un paso casi forzado, en tanto permite conocer de forma visual su comportamiento y determinar el tratamiento que se dará a la serie. En la siguiente sección, se explicará cómo trabajar con la tendencia.

## 7.3. Tendencia secular

En el apartado anterior, se mencionó que el análisis de series de tiempo comienza con una exploración gráfica en donde se identifican los componentes más notables. Ahora, en este subtema, se explicará el componente de tendencia, que normalmente destaca más en una serie de tiempo; y para estimarla se aplicarán los métodos de regresión lineal y de promedios móviles.

La tendencia de una serie es la trayectoria o dirección que toma esa tendencia conforme avanza el tiempo. La importancia de este componente radica en que permite estimar el valor de una serie en un momento futuro. Por ejemplo, supóngase que el área de finanzas de cierta organización dedicada a realizar estudios de mercado se encuentra evaluando el presupuesto del siguiente año destinado a proporcionar un apoyo económico a los encuestadores asignados a la ciudad para traslado. Un análisis del precio del transporte público durante los últimos veinte años mostraría la manera como se ha ido incrementando, lo que permitiría establecer una estimación del precio en que se encontraría el servicio para el siguiente año.

A fin de estimar la tendencia, se acostumbra utilizar el modelo de regresión lineal simple o los promedios móviles. A continuación, se muestra en un ejemplo la aplicación de estos métodos.

**Estimación de la tendencia con el modelo de regresión lineal simple**

Con el método de regresión lineal simple, se estima una tendencia lineal al considerar que la variable dependiente es la serie y la independiente el tiempo. A continuación, se plantea un ejemplo.

Desde enero de 2013, la fábrica ABC requiere, para la producción de cierta tinta, un insumo químico, cuyo precio varía cada mes. Con la intención de diseñar un plan de adquisiciones, el área de finanzas desea estimar cuál será el precio al final del 2014, con la información de enero de 2013 a agosto de 2014.

Se muestra a continuación la información con la que cuenta el área de finanzas.

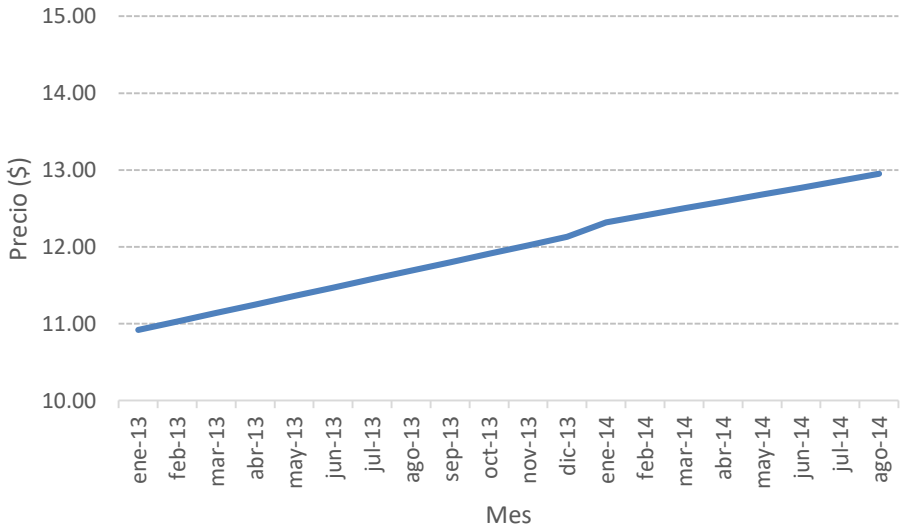
**Precio promedio del insumo durante enero de 2013 a agosto de 2014**

| Mes | Precio | Mes | Precio | Mes | Precio | Mes | Precio |
|---|---|---|---|---|---|---|---|
| ene-13 | 10.92 | jun-13 | 11.47 | nov-13 | 12.02 | abr-14 | 12.59 |
| feb-13 | 11.03 | jul-13 | 11.58 | dic-13 | 12.13 | may-14 | 12.68 |
| mar-13 | 11.14 | ago-13 | 11.69 | ene-14 | 12.32 | jun-14 | 12.77 |
| abr-13 | 11.25 | sep-13 | 11.80 | feb-14 | 12.41 | jul-14 | 12.86 |
| may-13 | 11.36 | oct-13 | 11.91 | mar-14 | 12.50 | ago-14 | 12.95 |

Precio por unidad de medida en pesos.

Se muestra en la siguiente gráfica el comportamiento del precio del insumo durante el periodo de análisis.

**Precio del insumo de enero de 2013 a agosto de 2014**



La gráfica muestra que, al comienzo del periodo de análisis, el precio de la unidad del insumo era casi de 11 pesos, y al finalizar se encuentra cerca de los 13 pesos. Se observa que, conforme han transcurrido los meses, el precio se asciende: la serie muestra una tendencia creciente. Para estimar la tendencia lineal que muestra la serie, se recurrirá al método de mínimos cuadrados (en este caso, la variable dependiente será el precio del insumo y la independiente el mes).

Antes de aplicar el método de mínimos cuadrados, se deberá realizar una adecuación a la variable independiente, que consiste en asignarle un valor numérico a cada mes. En este ejemplo, como la producción de la tinta comenzó a partir de enero de 2013, a esa observación se le asigna el valor 1, al siguiente mes el valor 2, y así sucesivamente hasta el valor 20, como se advierte en la tabla siguiente.

| X | Mes | Precio (Y) | X | Mes | Precio (Y) |
|---|---|---|---|---|---|
| **1** | ene-13 | 10.92 | 11 | nov-13 | 12.02 |
| **2** | feb-13 | 11.03 | 12 | dic-13 | 12.13 |
| **3** | mar-13 | 11.14 | 13 | ene-14 | 12.32 |
| **4** | abr-13 | 11.25 | 14 | feb-14 | 12.41 |
| **5** | may-13 | 11.36 | 15 | mar-14 | 12.50 |
| **6** | jun-13 | 11.47 | 16 | abr-14 | 12.59 |
| **7** | jul-13 | 11.58 | 17 | may-14 | 12.68 |
| **8** | ago-13 | 11.69 | 18 | jun-14 | 12.77 |
| **9** | sep-13 | 11.80 | 19 | jul-14 | 12.86 |
| **10** | oct-13 | 11.91 | 20 | ago-14 | 12.95 |

De esta manera, en el modelo se utilizarán las variables precio (Y) y X.

En la unidad anterior, se estudió cómo correr un modelo de regresión lineal simple en el módulo de análisis de datos en MS-Excel, a continuación, se utilizarán las funciones

intersección.eje()
pendiente()

para obtener los estimadores de los parámetros $\beta_0$ y $\beta_1$, respectivamente.

Para calcular $\beta_0$, en la función *intersección.eje()* se ingresan los valores de *Y*, se pone una coma y se procede a ingresar los valores de *X.*

Fuente: Microsoft Excel (2013).

Al dar *enter*, se despliega el resultado (10.82).

Para calcular $\beta_1$, en la función *pendiente ()* se ingresan también los valores de la variable *Y* y *X* (en ese orden) separados por una coma.

Fuente: Microsoft Excel (2013).

Al dar *enter*, se despliega el resultado (0.11).

Entonces, la ecuación de mínimos cuadrados es

$$\text{Precio} = 10.82 + 0.11x$$

El modelo indica que, antes de comenzar a producir la tinta (en X = 0), el precio del insumo se encontraba en $10.82, y desde ese momento, por cada mes que transcurre, el precio del insumo se eleva 11 centavos. Luego, esta ecuación es la tendencia de la serie.

Determinada la tendencia, se puede estimar el precio del insumo para los meses de septiembre a diciembre del año actual sustituyendo en la ecuación el número que corresponde al mes (21, 22, 23 o 24). De esta manera, se espera que en diciembre el precio del insumo se encuentre en $10.82 + (0.11)\cdot(24) = 13.46$.

Para calcular los pronósticos, añadimos el número de periodos que se van a pronosticar. Si se desea conocer el precio de la gasolina de los meses 21, 22 y 23 (septiembre, octubre, noviembre), aplicamos la fórmula obtenida de la regresión lineal para dichos meses.

Como una observación final a este apartado, es importante definir si, de acuerdo con el contexto de la serie a analizar, es necesario identificar un punto donde la variable independiente (X) tome el valor de 0.

**Estimación de la tendencia con el método de promedios móviles**

El método de promedios móviles (PM) consiste en construir una nueva serie con los promedios de los datos establecidos por el orden.

El orden de un promedio móvil se refiere al número de datos consecutivos a promediar. Por ejemplo, en un promedio móvil de orden dos ($PM_2$), se promedia cada conjunto de dos datos consecutivos; en uno de orden tres ($PM_3$), cada conjunto de tres datos consecutivos, y así sucesivamente.

Un promedio móvil de orden n ($PM_n$) se obtiene así:

$$PM_n = \frac{\textit{suma de los valores de los n datos más recientes}}{n}$$

Supóngase que un profesor de Estadística aplica evaluaciones mensuales a sus alumnos. Las calificaciones de los cinco exámenes realizados por un estudiante de la clase son los siguientes:

| Mes | Calificación |
|---|---|
| 1 | 7 |
| 2 | 8 |
| 3 | 7 |
| 4 | 9 |
| 5 | 6 |

El promedio móvil de orden dos ($PM_2$) se obtiene de la siguiente manera:

1. Se promedian las dos primeras calificaciones (7 y 8) y se coloca el resultado en el segundo valor de la nueva serie ($PM_2$):

| Mes | Calificación | PM2 | |
|---|---|---|---|
| 1 | 7 | | |
| 2 | 8 | 7.5 | $\frac{7+8}{2}$ |
| 3 | 7 | | |
| 4 | 9 | | |
| 5 | 6 | | |

2. El siguiente valor de la nueva serie ($PM_2$) se obtiene de promediar las calificaciones de los meses 2 y 3 (8 y 7):

| Mes | Calificación | PM2 | |
|---|---|---|---|
| 1 | 7 | | |
| 2 | 8 | 7.5 | $\frac{8+7}{2}$ |
| 3 | 7 | 7.5 | |
| 4 | 9 | | |
| 5 | 6 | | |

3. Seguir con el procedimiento hasta realizar el promedio de las últimas calificaciones (9 y 6):

| Mes | Calificación | PM2 | |
|---|---|---|---|
| 1 | 7 | | |
| 2 | 8 | 7.5 | |
| 3 | 7 | 7.5 | |
| 4 | 9 | 8 | |
| 5 | 6 | | $\frac{7+9}{2}$ |

| Mes | Calificación | PM2 | |
|---|---|---|---|
| 1 | 7 | | |
| 2 | 8 | 7.5 | |
| 3 | 7 | 7.5 | |
| 4 | 9 | 8 | |
| 5 | 6 | 7.5 | $\frac{9+6}{2}$ |

El promedio móvil de orden tres ($PM_3$) se obtiene de la siguiente manera.

1. Se promedian las primeras tres calificaciones (7, 8 y 7) y el resultado se coloca en la nueva serie $PM_3$, centrado en la segunda posición:

| Mes | Calificación | PM3 | |
|---|---|---|---|
| 1 | 7 | | |
| 2 | 8 | 7.3 | $\frac{7+8+7}{3}$ |
| 3 | 7 | | |
| 4 | 9 | | |
| 5 | 6 | | |

2. El siguiente valor de la nueva serie $PM_3$ se obtiene de promediar las calificaciones de los meses 2, 3 y 4 (8, 7 y 9):

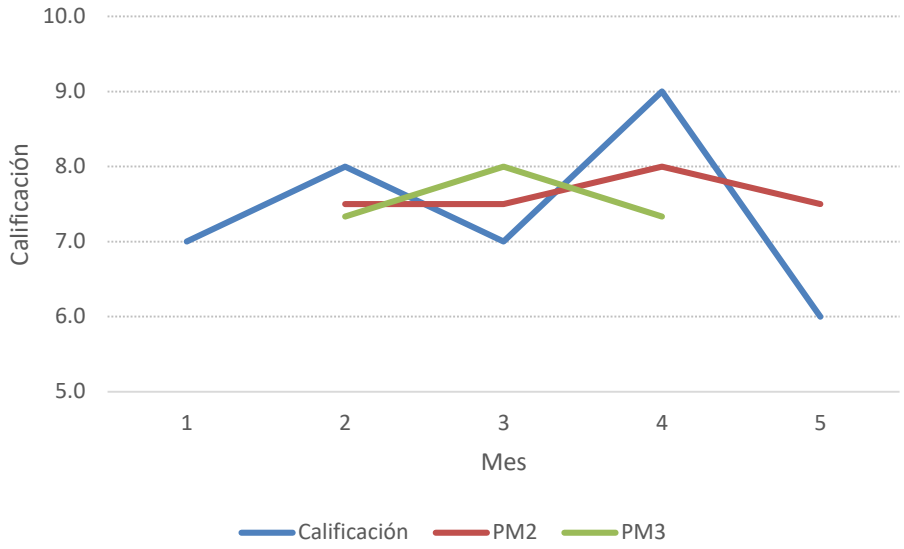
| Mes | Calificación | PM3 | |
|---|---|---|---|
| 1 | 7 | | |
| 2 | 8 | 7.3 | |
| 3 | 7 | 8.0 | $\frac{8+7+9}{3}$ |
| 4 | 9 | | |
| 5 | 6 | | |

3. Finalmente, se calcula el promedio de los últimos tres valores (7, 9 y 6) y se registra el resultado en la nueva serie:

| Mes | Calificación | PM3 | |
|---|---|---|---|
| 1 | 7 | | |
| 2 | 8 | 7.3 | |
| 3 | 7 | 8.0 | $\frac{7+9+6}{3}$ |
| 4 | 9 | 7.3 | |
| 5 | 6 | | |

Como es imposible seguir promediando tres valores, la nueva serie $PM_3$ solamente tendrá tres elementos. Es importante mencionar que, conforme aumenta el orden, la nueva serie va teniendo menos valores respecto a la serie original.

La siguiente gráfica muestra el comportamiento de las calificaciones del estudiante y los promedios móviles de orden 2 y 3.




La serie de color azul de la gráfica representa el comportamiento del estudiante en las cinco evaluaciones realizadas en el curso; la serie de color rojo es el promedio móvil de orden dos, y la serie de color gris el promedio móvil de orden tres. Los promedios móviles son un *suavizamiento* de la serie original y muestran la tendencia de la serie. Para este ejemplo, el promedio móvil de orden dos explica mejor la tendencia de las calificaciones del estudiante, y refleja que sus calificaciones se encuentran alrededor de 7.5.

Supóngase ahora que restan dos evaluaciones al curso, ¿qué calificaciones se esperan de este estudiante? Para realizar el pronóstico, se utilizará el promedio móvil de orden dos, que para este ejemplo describe mejor la tendencia, y se procederá de la siguiente manera.

1. Asumir que el último valor del promedio móvil se observará en el siguiente mes:

| Mes | Calificación | $PM_2$ |
|---|---|---|
| 1 | 7.0 | |
| 2 | 8.0 | 7.5 |
| 3 | 7.0 | 7.5 |
| 4 | 9.0 | 8.0 |
| 5 | 6.0 | 7.5 |
| 6 | 7.5 | |

2. Promediar las calificaciones de los meses 5 y 6 (6 y 7.5), y colocar el resultado en la posición 6 del promedio móvil:

| Mes | Calificación | $PM_2$ |
|---|---|---|
| 1 | 7.0 | |
| 2 | 8.0 | 7.5 |
| 3 | 7.0 | 7.5 |
| 4 | 9.0 | 8.0 |
| 5 | 6.0 | 7.5 |
| 6 | 7.5 | 6.8 |

3. Repetir el procedimiento descrito en los puntos anteriores para estimar la última calificación:

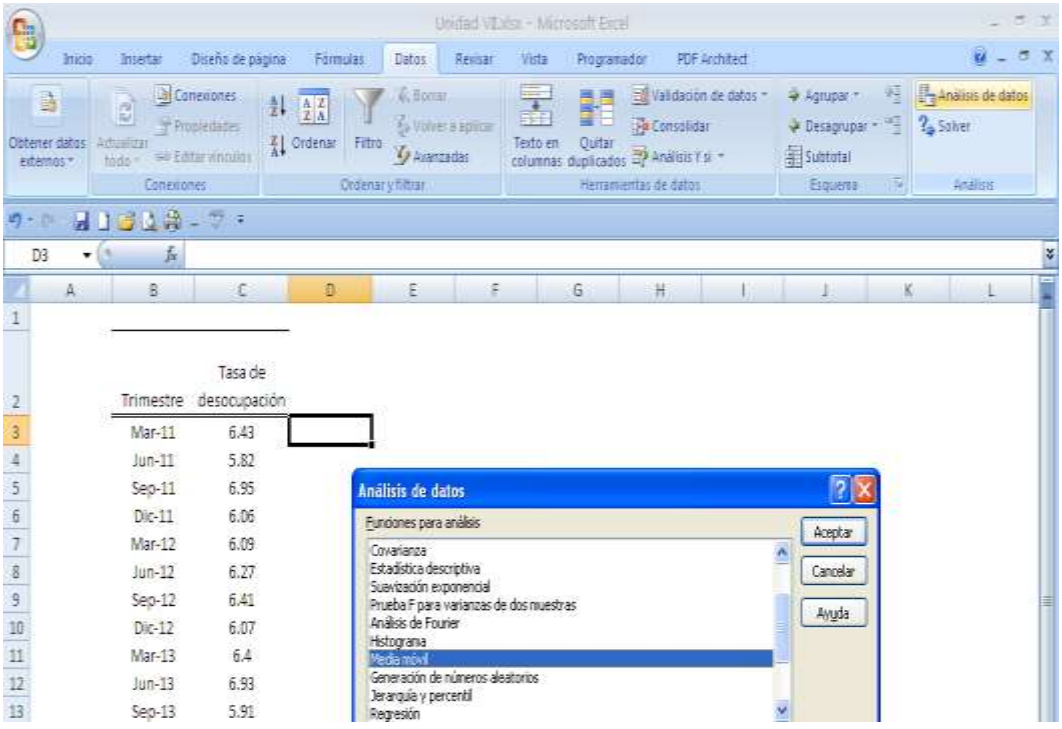
| Mes | Calificación | $PM_2$ |
|---|---|---|
| 1 | 7.0 | |
| 2 | 8.0 | 7.5 |
| 3 | 7.0 | 7.5 |
| 4 | 9.0 | 8.0 |
| 5 | 6.0 | 7.5 |
| 6 | 7.5 | 6.8 |
| 7 | 6.8 | 7.2 |

De esta manera, de acuerdo con la tendencia mostrada por el promedio móvil de orden dos, se espera que el estudiante obtenga calificaciones de 6.8 y 7.2 en las evaluaciones faltantes.

**Obtención de un promedio móvil con MS-Excel**

MS-Excel permite obtener un promedio móvil al utilizar el módulo de análisis de datos, para hacerlo se procede así.

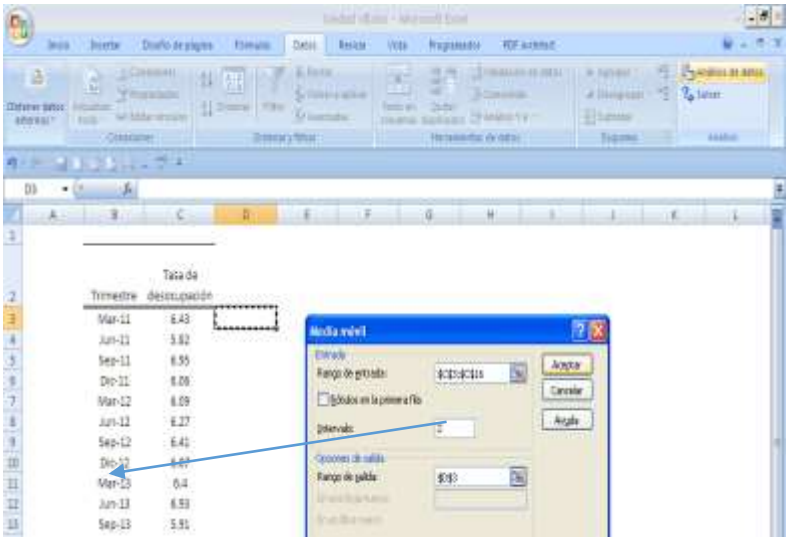
1. Acceder al menú Datos, seleccionar Análisis de datos y Media Móvil.




Fuente: Microsoft Excel (2013).

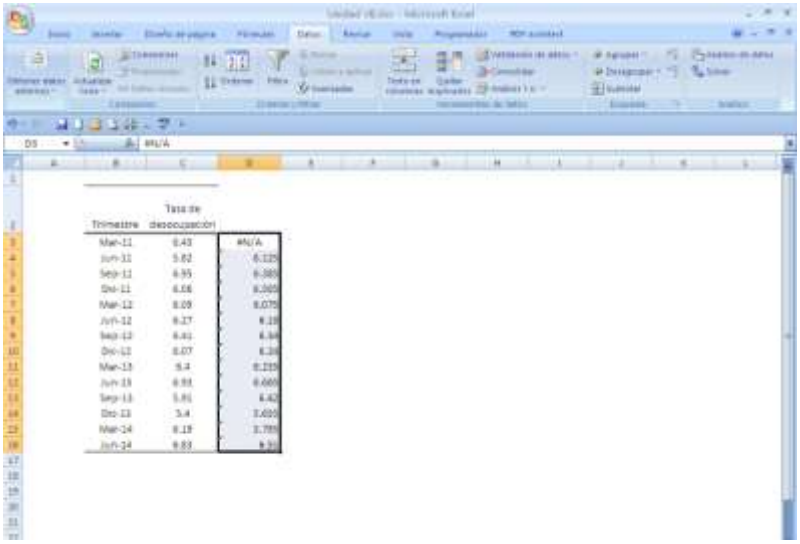
Se desplegará una ventana de diálogo que solicitará información de entrada y de salida.

En la sección Rango de entrada, seleccionar los datos de la variable; en Intervalo, indicar el rango deseado; y seleccionar el sitio en la hoja de cálculo en donde quiere que se despliegue el resultado en Rango de salida. Dar Aceptar.



Fuente: Microsoft Excel (2013).

Aparecerá una columna con los datos del promedio móvil.



Fuente: Microsoft Excel (2013).

# 7.4. Variaciones estacionales

En esta sección, se expondrá otro componente de una serie de tiempo: la estacionalidad. Una serie de tiempo tiene un comportamiento estacional si de forma periódica registra cambios a lo largo de un año. Por ejemplo, las ventas de una papelería muestran un comportamiento estacional caracterizado por un incremento durante los meses de julio y agosto, previo al comienzo del ciclo escolar del nivel básico. O la venta de pescados y mariscos crece un mes previo a las festividades de Semana Santa.

En la práctica, la manera de trabajar el componente de estacionalidad es calculando factores que se aplican a la tendencia.

A continuación, se analizará un ejemplo del tratamiento de este componente. Se muestra el indicador de comercio al por menor en México referente a artículos de papelería, libros, revistas y periódicos en el periodo, de enero de 2010 a diciembre de 2013.

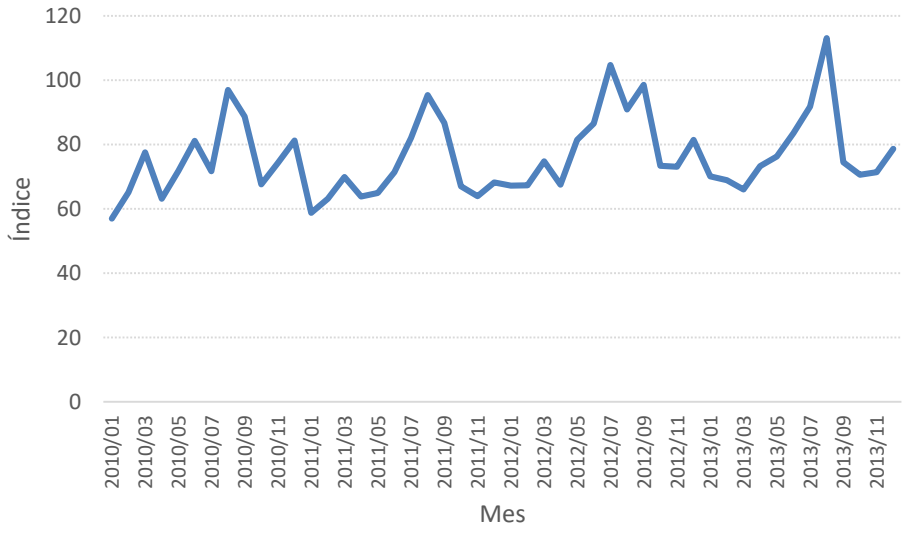
**Indicador de comercio al por menor en artículos de papelería, libros, revistas y periódicos, de enero de 2010 a diciembre de 2013**

| Mes | 2010 | 2011 | 2012 | 2013 |
|---|---|---|---|---|
| Enero | 57.0 | 58.7 | 67.2 | 70.1 |
| Febrero | 65.1 | 63.2 | 67.4 | 68.9 |
| Marzo | 77.6 | 69.9 | 74.8 | 66.0 |
| Abril | 63.1 | 63.9 | 67.5 | 73.3 |
| Mayo | 71.8 | 64.9 | 81.4 | 76.2 |
| Junio | 81.1 | 71.4 | 86.5 | 83.5 |
| Julio | 71.7 | 82.0 | 104.7 | 91.8 |
| Agosto | 96.9 | 95.4 | 90.9 | 113.1 |
| Septiembre | 88.7 | 86.7 | 98.5 | 74.4 |
| Octubre | 67.7 | 67.0 | 73.4 | 70.6 |
| Noviembre | 74.3 | 63.9 | 73.0 | 71.4 |
| Diciembre | 81.2 | 68.2 | 81.4 | 78.7 |

Base 2008.

Fuente: inegi.org.mx. fecha de consulta 7/06/2015

En la siguiente gráfica, se muestra el comportamiento de la serie.




La gráfica muestra que el índice en el periodo de análisis tiene una tendencia creciente, y alrededor de ella se aprecia que hay meses en que disminuye y meses

donde se incrementa. En consecuencia, la serie cuenta con un componente de estacionalidad.

Ahora bien, los factores de estacionalidad se calcularán de la siguiente manera.

A partir de la serie original, se construye un promedio móvil de orden 12, centrado de tal manera que se pierden los primeros y últimos seis meses.

| Mes | Índice | PM12 |
|---|---|---|
| ene-10 | 57.0 | |
| feb-10 | 65.1 | |
| mar-10 | 77.6 | |
| abr-10 | 63.1 | |
| may-10 | 71.8 | |
| jun-10 | 81.1 | |
| jul-10 | 71.7 | 74.7 |
| ago-10 | 96.9 | 74.8 |
| sep-10 | 88.7 | 74.7 |
| oct-10 | 67.7 | 74.0 |
| nov-10 | 74.3 | 74.1 |
| dic-10 | 81.2 | 73.5 |
| ene-11 | 58.7 | 72.7 |
| feb-11 | 63.2 | 73.6 |
| mar-11 | 69.9 | 73.4 |
| abr-11 | 63.9 | 73.3 |
| may-11 | 64.9 | 73.2 |
| jun-11 | 71.4 | 72.3 |
| jul-11 | 82.0 | 71.3 |
| ago-11 | 95.4 | 72.0 |
| sep-11 | 86.7 | 72.3 |
| oct-11 | 67.0 | 72.7 |
| nov-11 | 63.9 | 73.0 |
| dic-11 | 68.2 | 74.4 |
| ene-12 | 67.2 | 75.7 |
| feb-12 | 67.4 | 77.6 |
| mar-12 | 74.8 | 77.2 |
| abr-12 | 67.5 | 78.2 |
| may-12 | 81.4 | 78.7 |
| jun-12 | 86.5 | 79.5 |
| jul-12 | 104.7 | 80.6 |
| ago-12 | 90.9 | 80.8 |
| sep-12 | 98.5 | 80.9 |
| oct-12 | 73.4 | 80.2 |
| nov-12 | 73.0 | 80.7 |

| | | |
|---|---|---|
| dic-12 | 81.4 | 80.3 |
| ene-13 | 70.1 | 80.0 |
| feb-13 | 68.9 | 78.9 |
| mar-13 | 66.0 | 80.8 |
| abr-13 | 73.3 | 78.8 |
| may-13 | 76.2 | 78.5 |
| jun-13 | 83.5 | 78.4 |
| jul-13 | 91.8 | |
| ago-13 | 113.1 | |
| sep-13 | 74.4 | |
| oct-13 | 70.6 | |
| nov-13 | 71.4 | |
| dic-13 | 78.7 | |

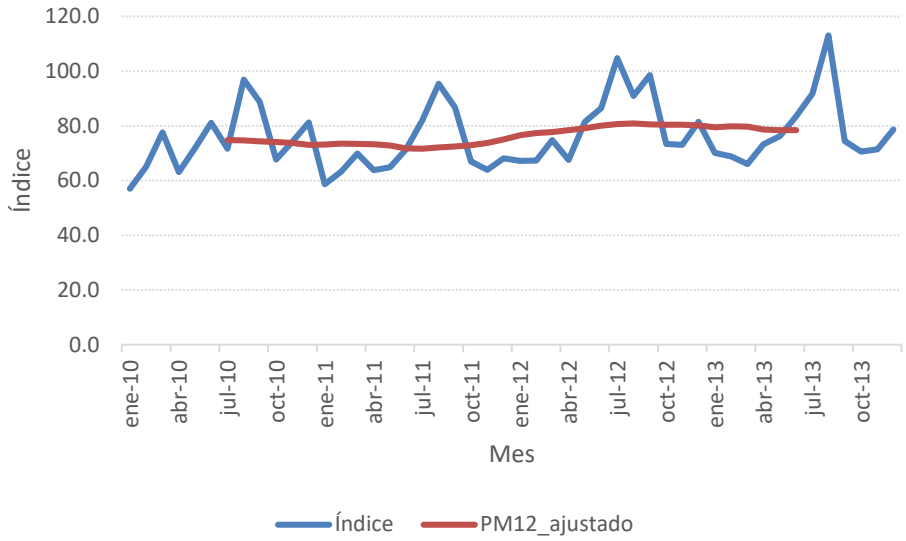
El primer punto del promedio móvil se obtiene al promediar los primeros 12 valores de la serie y se encontrará ubicado de manera que separa seis meses antes y después de él; es decir, se halla en el 15 de junio y el siguiente en el 15 de julio, para llevarlo al primero de julio se vuelve a construir un promedio móvil de orden 2.

| Mes | Índice | PM12 | PM12_ajustado |
|---|---|---|---|
| ene-10 | 57.0 | | |
| feb-10 | 65.1 | | |
| mar-10 | 77.6 | | |
| abr-10 | 63.1 | | |
| may-10 | 71.8 | | |
| jun-10 | 81.1 | | |
| jul-10 | 71.7 | 74.7 | 74.7 |
| ago-10 | 96.9 | 74.8 | 74.7 |
| sep-10 | 88.7 | 74.7 | 74.3 |
| oct-10 | 67.7 | 74.0 | 74.0 |
| nov-10 | 74.3 | 74.1 | 73.8 |
| dic-10 | 81.2 | 73.5 | 73.1 |
| ene-11 | 58.7 | 72.7 | 73.1 |
| feb-11 | 63.2 | 73.6 | 73.5 |
| mar-11 | 69.9 | 73.4 | 73.4 |
| abr-11 | 63.9 | 73.3 | 73.2 |
| may-11 | 64.9 | 73.2 | 72.8 |
| jun-11 | 71.4 | 72.3 | 71.8 |
| jul-11 | 82.0 | 71.3 | 71.6 |
| ago-11 | 95.4 | 72.0 | 72.1 |
| sep-11 | 86.7 | 72.3 | 72.5 |
| oct-11 | 67.0 | 72.7 | 72.9 |
| nov-11 | 63.9 | 73.0 | 73.7 |

| | | | |
|---|---|---|---|
| dic-11 | 68.2 | 74.4 | 75.0 |
| ene-12 | 67.2 | 75.7 | 76.6 |
| feb-12 | 67.4 | 77.6 | 77.4 |
| mar-12 | 74.8 | 77.2 | 77.7 |
| abr-12 | 67.5 | 78.2 | 78.4 |
| may-12 | 81.4 | 78.7 | 79.1 |
| jun-12 | 86.5 | 79.5 | 80.0 |
| jul-12 | 104.7 | 80.6 | 80.7 |
| ago-12 | 90.9 | 80.8 | 80.9 |
| sep-12 | 98.5 | 80.9 | 80.6 |
| oct-12 | 73.4 | 80.2 | 80.4 |
| nov-12 | 73.0 | 80.7 | 80.5 |
| dic-12 | 81.4 | 80.3 | 80.1 |
| ene-13 | 70.1 | 80.0 | 79.5 |
| feb-13 | 68.9 | 78.9 | 79.8 |
| mar-13 | 66.0 | 80.8 | 79.8 |
| abr-13 | 73.3 | 78.8 | 78.7 |
| may-13 | 76.2 | 78.5 | 78.5 |
| jun-13 | 83.5 | 78.4 | 78.4 |
| jul-13 | 91.8 | | |
| ago-13 | 113.1 | | |
| sep-13 | 74.4 | | |
| oct-13 | 70.6 | | |
| nov-13 | 71.4 | | |
| dic-13 | 78.7 | | |

El primer valor de la última serie se obtuvo al promediar los primeros dos valores del promedio móvil de orden 12. El segundo valor de la nueva serie es resultado de promediar el segundo y tercero del promedio móvil de orden 12, y así sucesivamente.

La siguiente gráfica muestra un comparativo entre el comportamiento de la serie original y el promedio móvil ajustado.




En la gráfica anterior, se plantea el comportamiento de la serie y del promedio móvil, que en este caso funciona como un eje alrededor del cual varía la serie original.

El siguiente paso es calcular la variación de cada punto respecto al promedio móvil dividiendo el valor original de la serie entre el promedio móvil.

| Mes | Índice | PM12 | PM12_ajustado | Variación |
|---|---|---|---|---|
| ene-10 | 57.0 | | | |
| feb-10 | 65.1 | | | |
| mar-10 | 77.6 | | | |
| abr-10 | 63.1 | | | |
| may-10 | 71.8 | | | |
| jun-10 | 81.1 | | | |
| jul-10 | 71.7 | 74.7 | 74.7 | 0.96 |
| ago-10 | 96.9 | 74.8 | 74.7 | 1.30 |
| sep-10 | 88.7 | 74.7 | 74.3 | 1.19 |
| oct-10 | 67.7 | 74.0 | 74.0 | 0.91 |
| nov-10 | 74.3 | 74.1 | 73.8 | 1.01 |
| dic-10 | 81.2 | 73.5 | 73.1 | 1.11 |
| ene-11 | 58.7 | 72.7 | 73.1 | 0.80 |
| feb-11 | 63.2 | 73.6 | 73.5 | 0.86 |
| mar-11 | 69.9 | 73.4 | 73.4 | 0.95 |
| abr-11 | 63.9 | 73.3 | 73.2 | 0.87 |
| may-11 | 64.9 | 73.2 | 72.8 | 0.89 |

| | | | | |
|---|---|---|---|---|
| jun-11 | 71.4 | 72.3 | 71.8 | 0.99 |
| jul-11 | 82.0 | 71.3 | 71.6 | 1.15 |
| ago-11 | 95.4 | 72.0 | 72.1 | 1.32 |
| sep-11 | 86.7 | 72.3 | 72.5 | 1.20 |
| oct-11 | 67.0 | 72.7 | 72.9 | 0.92 |
| nov-11 | 63.9 | 73.0 | 73.7 | 0.87 |
| dic-11 | 68.2 | 74.4 | 75.0 | 0.91 |
| ene-12 | 67.2 | 75.7 | 76.6 | 0.88 |
| feb-12 | 67.4 | 77.6 | 77.4 | 0.87 |
| mar-12 | 74.8 | 77.2 | 77.7 | 0.96 |
| abr-12 | 67.5 | 78.2 | 78.4 | 0.86 |
| may-12 | 81.4 | 78.7 | 79.1 | 1.03 |
| jun-12 | 86.5 | 79.5 | 80.0 | 1.08 |
| jul-12 | 104.7 | 80.6 | 80.7 | 1.30 |
| ago-12 | 90.9 | 80.8 | 80.9 | 1.12 |
| sep-12 | 98.5 | 80.9 | 80.6 | 1.22 |
| oct-12 | 73.4 | 80.2 | 80.4 | 0.91 |
| nov-12 | 73.0 | 80.7 | 80.5 | 0.91 |
| dic-12 | 81.4 | 80.3 | 80.1 | 1.02 |
| ene-13 | 70.1 | 80.0 | 79.5 | 0.88 |
| feb-13 | 68.9 | 78.9 | 79.8 | 0.86 |
| mar-13 | 66.0 | 80.8 | 79.8 | 0.83 |
| abr-13 | 73.3 | 78.8 | 78.7 | 0.93 |
| may-13 | 76.2 | 78.5 | 78.5 | 0.97 |
| jun-13 | 83.5 | 78.4 | 78.4 | 1.06 |
| jul-13 | 91.8 | | | |
| ago-13 | 113.1 | | | |
| sep-13 | 74.4 | | | |
| oct-13 | 70.6 | | | |
| nov-13 | 71.4 | | | |
| dic-13 | 78.7 | | | |

El primer valor de la serie de variaciones se obtuvo de dividir 71.7 (valor original de la serie) entre 74.7 (promedio móvil ajustado). El valor resultante de 0.96 significa que el índice observado en julio de 2010 se encontró 4% debajo del promedio. De manera similar, se procedió con el resto de los valores.

Se llega a los factores estacionales mensuales promediando todas las variaciones obtenidas en el mismo mes.

| Mes | 2010 | 2011 | 2012 | 2013 | Promedio |
|---|---|---|---|---|---|
| Enero | | 0.80 | 0.88 | 0.88 | 0.85 |
| Febrero | | 0.86 | 0.87 | 0.86 | 0.86 |
| Marzo | | 0.95 | 0.96 | 0.83 | 0.91 |
| Abril | | 0.87 | 0.86 | 0.93 | 0.89 |
| Mayo | | 0.89 | 1.03 | 0.97 | 0.96 |
| Junio | | 0.99 | 1.08 | 1.06 | 1.05 |
| Julio | 0.96 | 1.15 | 1.30 | | 1.13 |
| Agosto | 1.30 | 1.32 | 1.12 | | 1.25 |
| Septiembre | 1.19 | 1.20 | 1.22 | | 1.20 |
| Octubre | 0.91 | 0.92 | 0.91 | | 0.91 |
| Noviembre | 1.01 | 0.87 | 0.91 | | 0.93 |
| Diciembre | 1.11 | 0.91 | 1.02 | | 1.01 |

Como el promedio móvil ajustado parte de julio de 2010 y termina en junio de 2013, en cada mes se calcularon tres variaciones, que al promediarse serán los factores estacionales.

Los factores estacionales muestran una mayor actividad en los meses de julio, agosto y septiembre, donde el índice es, respectivamente, 13%, 25% y 20% mayor al promedio. La menor actividad se registra en enero y febrero, donde los factores son 0.85 y 0.86.

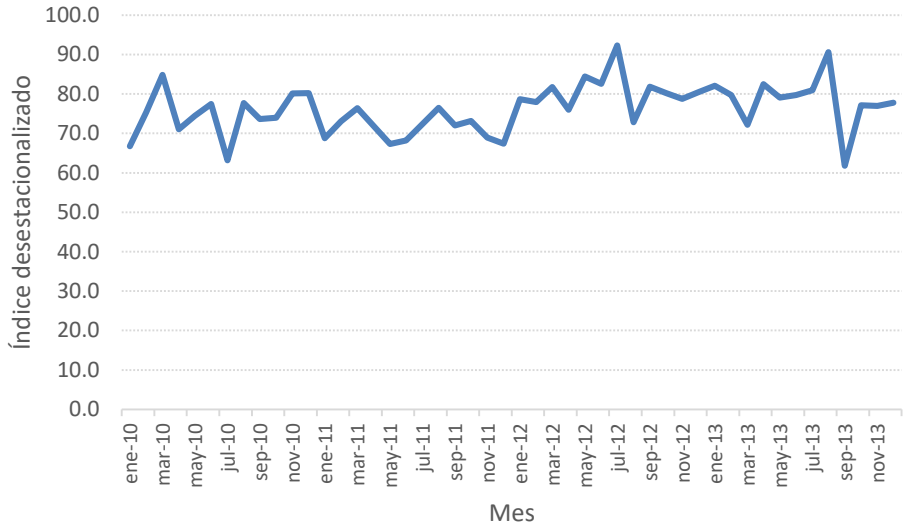
Una vez calculados los factores estacionales, sigue desestacionalizar los datos, dividiendo el valor original de la serie entre el factor que le corresponda.

| Mes | Índice | Factor | Índice desestacionalizado |
|---|---|---|---|
| ene-10 | 57.0 | 0.85 | 66.7 |
| feb-10 | 65.1 | 0.86 | 75.3 |
| mar-10 | 77.6 | 0.91 | 84.8 |
| abr-10 | 63.1 | 0.89 | 71.0 |

| | | | |
|---|---|---|---|
| may-10 | 71.8 | 0.96 | 74.4 |
| jun-10 | 81.1 | 1.05 | 77.5 |
| jul-10 | 71.7 | 1.13 | 63.2 |
| ago-10 | 96.9 | 1.25 | 77.7 |
| sep-10 | 88.7 | 1.20 | 73.7 |
| oct-10 | 67.7 | 0.91 | 73.9 |
| nov-10 | 74.3 | 0.93 | 80.1 |
| dic-10 | 81.2 | 1.01 | 80.2 |
| ene-11 | 58.7 | 0.85 | 68.8 |
| feb-11 | 63.2 | 0.86 | 73.1 |
| mar-11 | 69.9 | 0.91 | 76.4 |
| abr-11 | 63.9 | 0.89 | 71.9 |
| may-11 | 64.9 | 0.96 | 67.3 |
| jun-11 | 71.4 | 1.05 | 68.2 |
| jul-11 | 82.0 | 1.13 | 72.3 |
| ago-11 | 95.4 | 1.25 | 76.5 |
| sep-11 | 86.7 | 1.20 | 72.0 |
| oct-11 | 67.0 | 0.91 | 73.2 |
| nov-11 | 63.9 | 0.93 | 68.9 |
| dic-11 | 68.2 | 1.01 | 67.4 |
| ene-12 | 67.2 | 0.85 | 78.6 |
| feb-12 | 67.4 | 0.86 | 77.9 |
| mar-12 | 74.8 | 0.91 | 81.8 |
| abr-12 | 67.5 | 0.89 | 76.0 |
| may-12 | 81.4 | 0.96 | 84.5 |
| jun-12 | 86.5 | 1.05 | 82.6 |
| jul-12 | 104.7 | 1.13 | 92.3 |
| ago-12 | 90.9 | 1.25 | 72.9 |
| sep-12 | 98.5 | 1.20 | 81.8 |
| oct-12 | 73.4 | 0.91 | 80.2 |
| nov-12 | 73.0 | 0.93 | 78.8 |
| dic-12 | 81.4 | 1.01 | 80.5 |
| ene-13 | 70.1 | 0.85 | 82.1 |
| feb-13 | 68.9 | 0.86 | 79.7 |
| mar-13 | 66.0 | 0.91 | 72.2 |
| abr-13 | 73.3 | 0.89 | 82.5 |
| may-13 | 76.2 | 0.96 | 79.1 |
| jun-13 | 83.5 | 1.05 | 79.8 |
| jul-13 | 91.8 | 1.13 | 80.9 |
| ago-13 | 113.1 | 1.25 | 90.6 |
| sep-13 | 74.4 | 1.20 | 61.8 |
| oct-13 | 70.6 | 0.91 | 77.2 |
| nov-13 | 71.4 | 0.93 | 77.0 |
| dic-13 | 78.7 | 1.01 | 77.8 |

En la tabla anterior, los valores de la última columna son resultado de dividir el índice entre el factor.

La serie desestacionalizada queda así:




La gráfica anterior muestra los datos desestacionalizados, que no reflejan una tendencia aparente. Para confirmar lo anterior, se ajusta una regresión, la cual indica que sí existe una tendencia.

| | *Coeficientes* | *Error típico* | *Estadístico t* | *Probabilidad* |
|---|---|---|---|---|
| Intercepción | 73.0707108 | 1.82817565 | 39.969196 | 8.4506E-37 |
| Índice desestacionalizado | 0.1381169 | 0.06427523 | 2.14883546 | 0.03706385 |

Entonces, la tendencia de los datos desestacionalizados es $\boldsymbol{y_t = 73.07 + 0.14t}$

La ecuación expresa que, por cada mes transcurrido, el índice desestacionalizado se incrementa en 0.14.

Supóngase que se desea realizar un pronóstico para los siguientes cinco meses, es decir, para las observaciones 49, 50, 51, 52 y 53. Primero, se sustituyen estos valores en el modelo de la tendencia:

| **t** | 73.07 + 0.14t |
|---|---|
| 49 | 73.07 + (0.14)·(49) = 79.93 |
| 50 | 73.07 + (0.14)·(50) = 80.07 |
| 51 | 73.07 + (0.14)·(51) = 80.21 |
| 52 | 73.07 + (0.14)·(52) = 80.35 |
| 53 | 73.07 + (0.14)·(53) = 80.49 |

Los valores obtenidos se multiplican por el factor estacional:

| t | Índice desestacionalizado | Factor estacional | Pronóstico |
|---|---|---|---|
| 49 | 79.93 | 0.85 | 68.26 |
| 50 | 80.07 | 0.86 | 69.20 |
| 51 | 80.21 | 0.91 | 73.34 |
| 52 | 80.35 | 0.89 | 71.36 |
| 53 | 80.49 | 0.96 | 77.63 |

De esta manera, se alcanza el pronóstico.

# 7.5. Variaciones cíclicas

En la sección anterior, se trató cómo trabajar el componente estacional de una serie, el cual ofrece las variaciones que se presentan a lo largo de un año. Ahora, en este subtema se muestra el tratamiento de variaciones presentadas en periodos mayores a un año, los cuales son el componente de ciclicidad.

Un ciclo consiste en cambios ascendentes y descendentes en la serie respecto a su tendencia, con duración mayor a un año. Ejemplos de ello son los ciclos económicos caracterizados por un periodo de expansión y recesión, o el ciclo de vida de un producto.

Un componente cíclico tiene un comportamiento parecido al de la figura siguiente:

Fuente: elaboración propia.

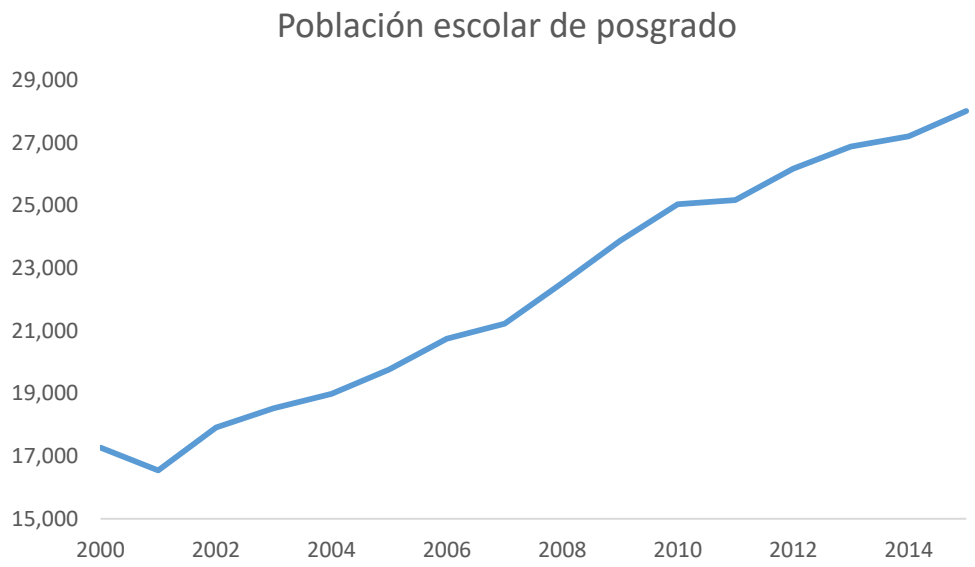
Como se muestra en la figura anterior, el ciclo se integra de dos partes: una expansiva, donde la serie aumenta de valor; y otra recesiva, donde disminuye el valor.

En cuanto al tratamiento que se dará a este componente, será bajo un enfoque aditivo. A continuación, se expone un ejemplo.

En la siguiente tabla, se muestra la población escolar de posgrado de cierta institución entre 2000 y 2015.

| Año | Población escolar de posgrado |
|---|---|
| 2000 | 17,270 |
| 2001 | 16,547 |
| 2002 | 17,910 |
| 2003 | 18,530 |
| 2004 | 18,987 |
| 2005 | 19,765 |
| 2006 | 20,747 |
| 2007 | 21,230 |
| 2008 | 22,527 |
| 2009 | 23,875 |
| 2010 | 25,036 |
| 2011 | 25,167 |
| 2012 | 26,169 |
| 2013 | 26,878 |
| 2014 | 27,210 |
| 2015 | 28,018 |

En el periodo de análisis, se advierte que la población creció de 17 270 en el año 2000 a 28 018 en 2015. Al graficar la serie, se observa el siguiente comportamiento:




La gráfica anterior manifiesta el crecimiento de la población de posgrado, caracterizado por una serie con tendencia positiva. A continuación se estimará la tendencia, con una regresión lineal simple a la serie, y se obtendrá la *y* estimada.

| Consecutivo | Año | Población escolar de posgrado | Tendencia $\hat{y}$ | | |
|---|---|---|---|---|---|
| 1 | 2000 | 17,270 | 16203 | | |
| 2 | 2001 | 16,547 | 17008 | 15397.45 | Intersección |
| 3 | 2002 | 17,910 | 17813 | 805.197059 | Pendiente |
| 4 | 2003 | 18,530 | 18618 | | |
| 5 | 2004 | 18,987 | 19423 | | |
| 6 | 2005 | 19,765 | 20229 | | |
| 7 | 2006 | 20,747 | 21034 | | |
| 8 | 2007 | 21,230 | 21839 | | |
| 9 | 2008 | 22,527 | 22644 | | |
| 10 | 2009 | 23,875 | 23449 | | |
| 11 | 2010 | 25,036 | 24255 | | |
| 12 | 2011 | 25,167 | 25060 | | |
| 13 | 2012 | 26,169 | 25865 | | |
| 14 | 2013 | 26,878 | 26670 | | |
| 15 | 2014 | 27,210 | 27475 | | |
| 16 | 2015 | 28,018 | 28281 | | |

Fuente: elaboración propia con Microsoft Excel (2013).

La tendencia de la población de posgrado se estima con la siguiente ecuación:

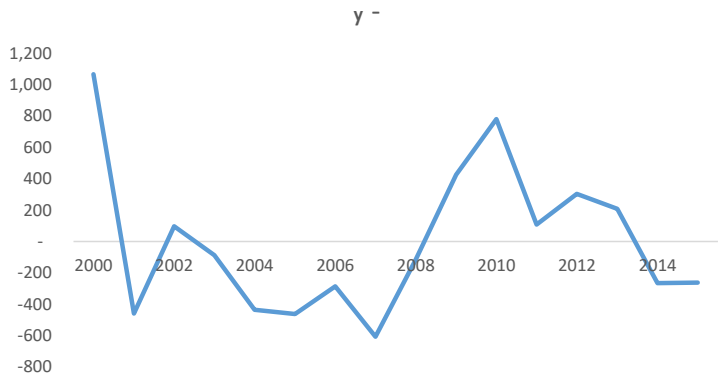
Población escolar de posgrado = 15,397 + 805 año

Enseguida, se elimina el componente de tendencia a la serie original. Para hacerlo, se resta la tendencia a los valores originales:

| Consecutivo | Año | Población escolar de posgrado | Tendencia $\hat{y}$ | Sin tendencia $y - \hat{y}$ |
|---|---|---|---|---|
| 1 | 2000 | 17,270 | 16,203 | 1,067 |
| 2 | 2001 | 16,547 | 17,008 | - 461 |
| 3 | 2002 | 17,910 | 17,813 | 97 |
| 4 | 2003 | 18,530 | 18,618 | - 88 |
| 5 | 2004 | 18,987 | 19,423 | - 436 |
| 6 | 2005 | 19,765 | 20,229 | - 464 |
| 7 | 2006 | 20,747 | 21,034 | - 287 |
| 8 | 2007 | 21,230 | 21,839 | - 609 |
| 9 | 2008 | 22,527 | 22,644 | - 117 |
| 10 | 2009 | 23,875 | 23,449 | 426 |
| 11 | 2010 | 25,036 | 24,255 | 781 |
| 12 | 2011 | 25,167 | 25,060 | 107 |
| 13 | 2012 | 26,169 | 25,865 | 304 |
| 14 | 2013 | 26,878 | 26,670 | 208 |
| 15 | 2014 | 27,210 | 27,475 | - 265 |
| 16 | 2015 | 28,018 | 28,281 | - 263 |

El cuadro anterior presenta la serie original y la tendencia calculada con el modelo de regresión. La última columna es la serie sin tendencia, resultado de restar la tendencia de la serie original.

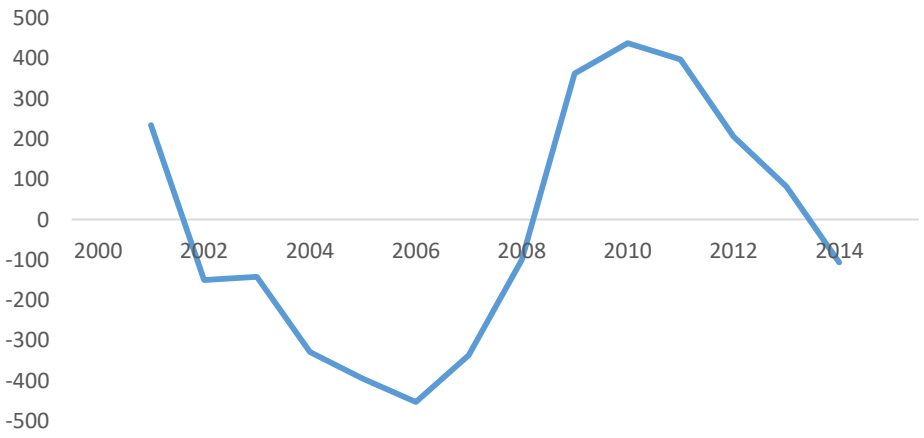
Ahora, la serie sin tendencia luce así:




La gráfica representa una serie con un comportamiento que se acerca a un ciclo. Obsérvese que aproximadamente cada tres años se cumple el ciclo, por lo que se utilizará un promedio móvil de orden 3 para obtener el componente cíclico (véase la siguiente tabla).

| Año | Sin tendencia $y - \hat{y}$ | Ciclo PM3 |
|---|---|---|
| 2000 | 1,067 | |
| 2001 | - 461 | 234 |
| 2002 | 97 | - 151 |
| 2003 | - 88 | - 143 |
| 2004 | - 436 | - 329 |
| 2005 | - 464 | - 396 |
| 2006 | - 287 | - 453 |
| 2007 | - 609 | - 338 |
| 2008 | - 117 | - 100 |
| 2009 | 426 | 363 |
| 2010 | 781 | 438 |
| 2011 | 107 | 398 |
| 2012 | 304 | 206 |
| 2013 | 208 | 82 |
| 2014 | - 265 | - 107 |
| 2015 | - 263 | |

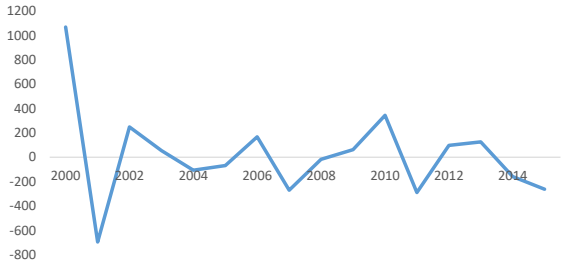
La tabla anterior expresa la nueva serie obtenida con el promedio móvil de orden 3, que al graficarse muestra el componente cíclico:




Para quitar el ciclo, se resta el componente a la serie sin tendencia:

| Año | Población escolar de posgrado | Sin tendencia $y - \hat{y}$ | Ciclo PM3 | Aleatorio $(y - \hat{y}) - PM3$ |
|---|---|---|---|---|
| 2000 | 17,270 | 1,067 | | 1067 |
| 2001 | 16,547 | - 461 | 234 | - 695 |
| 2002 | 17,910 | 97 | -151 | 248 |
| 2003 | 18,530 | - 88 | -143 | 54 |
| 2004 | 18,987 | - 436 | -329 | - 107 |
| 2005 | 19,765 | - 464 | -396 | -68 |
| 2006 | 20,747 | - 287 | -453 | 166 |
| 2007 | 21,230 | - 609 | -338 | - 271 |
| 2008 | 22,527 | - 117 | -100 | - 17 |
| 2009 | 23,875 | 426 | 363 | 62 |
| 2010 | 25,036 | 781 | 438 | 343 |
| 2011 | 25,167 | 107 | 398 | - 290 |
| 2012 | 26,169 | 304 | 206 | 98 |
| 2013 | 26,878 | 208 | 82 | 126 |
| 2014 | 27,210 | - 265 | -107 | - 159 |
| 2015 | 28,018 | - 263 | | - 263 |

El resultado es una serie irregular o aleatoria:




Supóngase que se necesita realizar un pronóstico de alumnos de posgrado del 2016 al 2019. Para hacerlo, se darán los siguientes pasos.

1. Pronosticar la tendencia en los periodos futuros con la ecuación lineal.

**Población escolar de posgrado = 15,397 + 805 año**

| | Año | Población escolar de posgrado | Tendencia | Sin tendencia $y - \hat{y}$ | Ciclo PM3 | Aleatorio $(y - \hat{y}) - PM3$ |
|---|---|---|---|---|---|---|
| 1 | 2000 | 17,270 | 16,203 | 1,067 | | 1067 |
| 2 | 2001 | 16,547 | 17,008 | - 461 | 234 | - 695 |
| 3 | 2002 | 17,910 | 17,813 | 97 | -151 | 248 |
| 4 | 2003 | 18,530 | 18,618 | - 88 | -143 | 54 |
| 5 | 2004 | 18,987 | 19,423 | - 436 | -329 | - 107 |
| 6 | 2005 | 19,765 | 20,229 | - 464 | -396 | - 68 |
| 7 | 2006 | 20,747 | 21,034 | - 287 | -453 | 166 |
| 8 | 2007 | 21,230 | 21,839 | - 609 | -338 | - 271 |
| 9 | 2008 | 22,527 | 22,644 | - 117 | -100 | - 17 |
| 10 | 2009 | 23,875 | 23,449 | 426 | 363 | 62 |
| 11 | 2010 | 25,036 | 24,255 | 781 | 438 | 343 |
| 12 | 2011 | 25,167 | 25,060 | 107 | 398 | - 290 |
| 13 | 2012 | 26,169 | 25,865 | 304 | 206 | 98 |
| 14 | 2013 | 26,878 | 26,670 | 208 | 82 | 126 |
| 15 | 2014 | 27,210 | 27,475 | - 265 | -107 | - 159 |
| 16 | 2015 | 28,018 | 28,281 | - 263 | | - 263 |
| **17** | **2016** | | 29,086 | | | |
| **18** | **2017** | | 29,891 | | | |
| **19** | **2018** | | 30,696 | | | |
| **20** | **2019** | | 31,501 | | | |
| **21** | **2020** | | 32,307 | | | |

2. Para estimar el ciclo, se recurre al procedimiento de promedio móvil: se copia el último valor (–107) de la serie Ciclo PM3 en la columna Sin tendencia, debajo del valor –263, y en ambas columnas se replican las fórmulas ya trabajadas en cada una de ellas:

| | Año | Población escolar de posgrado | Tendencia | Sin tendencia $y - \hat{y}$ | Ciclo PM3 | Aleatorio $(y - \hat{y}) - PM3$ |
|---|---|---|---|---|---|---|
| 1 | 2000 | 17,270 | 16,203 | 1,067 | | 1067 |
| 2 | 2001 | 16,547 | 17,008 | - 461 | 234 | - 695 |
| 3 | 2002 | 17,910 | 17,813 | 97 | -151 | 248 |
| 4 | 2003 | 18,530 | 18,618 | - 88 | -143 | 54 |
| 5 | 2004 | 18,987 | 19,423 | - 436 | -329 | - 107 |
| 6 | 2005 | 19,765 | 20,229 | - 464 | -396 | - 68 |
| 7 | 2006 | 20,747 | 21,034 | - 287 | -453 | 166 |
| 8 | 2007 | 21,230 | 21,839 | - 609 | -338 | - 271 |
| 9 | 2008 | 22,527 | 22,644 | - 117 | -100 | - 17 |
| 10 | 2009 | 23,875 | 23,449 | 426 | 363 | 62 |
| 11 | 2010 | 25,036 | 24,255 | 781 | 438 | 343 |
| 12 | 2011 | 25,167 | 25,060 | 107 | 398 | - 290 |
| 13 | 2012 | 26,169 | 25,865 | 304 | 206 | 98 |
| 14 | 2013 | 26,878 | 26,670 | 208 | 82 | 126 |
| 15 | 2014 | 27,210 | 27,475 | - 265 | -107 | - 159 |
| 16 | 2015 | 28,018 | 28,281 | - 263 | - 212 | - 51 |
| **17** | **2016** | | 29,086 | **- 107** | - 194 | 87 |
| **18** | **2017** | | 29,891 | - 212 | - 171 | |
| **19** | **2018** | | 30,696 | - 194 | - 192 | |
| **20** | **2019** | | 31,501 | - 171 | - 185 | |
| **21** | **2020** | | 32,307 | - 192 | | |

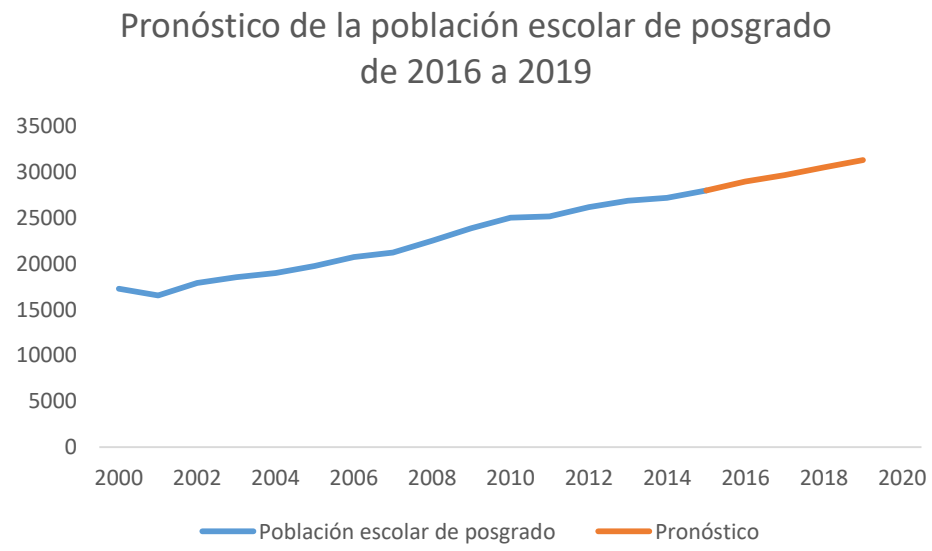
De igual manera, se replica la fórmula de la columna Aleatorio para los periodos a pronosticar:

| | Año | Población escolar de posgrado | Tendencia | Sin tendencia $y - \hat{y}$ | Ciclo PM3 | Aleatorio $(y - \hat{y}) - PM3$ |
|---|---|---|---|---|---|---|
| 1 | 2000 | 17,270 | 16,203 | 1,067 | | 1067 |
| 2 | 2001 | 16,547 | 17,008 | - 461 | 234 | - 695 |
| 3 | 2002 | 17,910 | 17,813 | 97 | -151 | 248 |
| 4 | 2003 | 18,530 | 18,618 | - 88 | -143 | 54 |
| 5 | 2004 | 18,987 | 19,423 | - 436 | -329 | - 107 |
| 6 | 2005 | 19,765 | 20,229 | - 464 | -396 | - 68 |
| 7 | 2006 | 20,747 | 21,034 | - 287 | -453 | 166 |
| 8 | 2007 | 21,230 | 21,839 | - 609 | -338 | - 271 |
| 9 | 2008 | 22,527 | 22,644 | - 117 | -100 | - 17 |
| 10 | 2009 | 23,875 | 23,449 | 426 | 363 | 62 |
| 11 | 2010 | 25,036 | 24,255 | 781 | 438 | 343 |
| 12 | 2011 | 25,167 | 25,060 | 107 | 398 | - 290 |
| 13 | 2012 | 26,169 | 25,865 | 304 | 206 | 98 |
| 14 | 2013 | 26,878 | 26,670 | 208 | 82 | 126 |
| 15 | 2014 | 27,210 | 27,475 | - 265 | -107 | - 159 |
| 16 | 2015 | 28,018 | 28,281 | - 263 | - 212 | - 51 |
| **17** | **2016** | | 29,086 | **- 107** | - 194 | 87 |
| **18** | **2017** | | 29,891 | - 212 | - 171 | - 41 |
| **19** | **2018** | | 30,696 | - 194 | - 192 | - 2 |
| **20** | **2019** | | 31,501 | - 171 | - 185 | 15 |
| **21** | **2020** | | 32,307 | - 192 | | - 192 |

Se crea una nueva columna para realizar el pronóstico, sumando los valores de las columnas Tendencia, Ciclo y Aleatorio:

| | Año | Población escolar de posgrado | Tendencia | Sin tendencia $y - \hat{y}$ | Ciclo PM3 | Aleatorio $(y - \hat{y}) - \text{PM3}$ | Pronóstico |
|---|---|---|---|---|---|---|---|
| 1 | 2000 | 17,270 | 16,203 | 1,067 | | 1067 | |
| 2 | 2001 | 16,547 | 17,008 | - 461 | 234 | - 695 | |
| 3 | 2002 | 17,910 | 17,813 | 97 | -151 | 248 | |
| 4 | 2003 | 18,530 | 18,618 | - 88 | -143 | 54 | |
| 5 | 2004 | 18,987 | 19,423 | - 436 | -329 | - 107 | |
| 6 | 2005 | 19,765 | 20,229 | - 464 | -396 | - 68 | |
| 7 | 2006 | 20,747 | 21,034 | - 287 | -453 | 166 | |
| 8 | 2007 | 21,230 | 21,839 | - 609 | -338 | - 271 | |
| 9 | 2008 | 22,527 | 22,644 | - 117 | -100 | - 17 | |
| 10 | 2009 | 23,875 | 23,449 | 426 | 363 | 62 | |
| 11 | 2010 | 25,036 | 24,255 | 781 | 438 | 343 | |
| 12 | 2011 | 25,167 | 25,060 | 107 | 398 | - 290 | |
| 13 | 2012 | 26,169 | 25,865 | 304 | 206 | 98 | |
| 14 | 2013 | 26,878 | 26,670 | 208 | 82 | 126 | |
| 15 | 2014 | 27,210 | 27,475 | - 265 | -107 | - 159 | |
| 16 | 2015 | 28,018 | 28,281 | - 263 | - 212 | - 51 | 28,018 |
| **17** | **2016** | | 29,086 | **- 107** | - 194 | 87 | 28,979 |
| **18** | **2017** | | 29,891 | - 212 | - 171 | - 41 | 29,679 |
| **19** | **2018** | | 30,696 | - 194 | - 192 | - 2 | 30,503 |
| **20** | **2019** | | 31,501 | - 171 | - 185 | 15 | 31,331 |
| **21** | **2020** | | 32,307 | - 192 | | - 192 | |

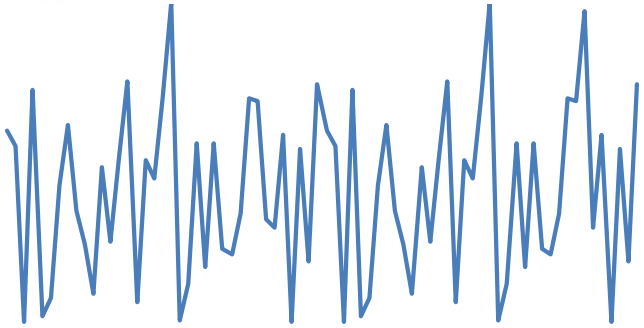
Por tanto, la población escolar aumentará de 28 018 a 31 331 alumnos entre 2016 y 2019.




# 7.6. Fluctuaciones irregulares

El último componente de una serie de tiempo es de fluctuaciones irregulares. Este componente se caracteriza por tener un comportamiento difícil de modelar, debido a que sus variaciones se deben a causas particulares que normalmente no son predecibles (por ejemplo, las variaciones en el tráfico a causa de un accidente o una manifestación).

Una serie irregular se ejemplifica en la siguiente figura, donde no se aprecia un patrón:



Fuente: elaboración propia.

En el ejemplo de la sección anterior, después de quitar los componentes de tendencia y ciclicidad, se obtuvo como resultado una serie aleatoria con la cual ya no se hizo tratamiento adicional.

En el análisis de series de tiempo, luego de quitar los componentes, se busca trabajar con series estacionarias, las cuales tienen un comportamiento constante, donde su media y varianza se mantienen a lo largo del tiempo, como lo muestra la siguiente imagen:


Fuente: elaboración propia.

Para profundizar en este tema, se sugiere consultar Hanke, J. (2010).

## 7.7. Modelos autorregresivos de promedios móviles

El empleo de estos modelos se realiza con series estacionarias. Debido a que se requieren mayores bases de probabilidad y manejo de *software* estadístico como STATA, EVIEWS, SAS, entre otros, solamente se mencionarán las principales características de estos modelos.

Los procesos autorregresivos son aquellos que se modelan en función de sus observaciones pasadas:

$$Y_t = \rho_0 + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \cdots + \rho_p Y_{t-p} + Z_t$$

Donde $\rho_k$ es la autocorrelación de rezago *k*.

Supóngase que se tiene la siguiente serie:

| t | $Y_t$ |
|---|---|
| 1 | 5 |
| 2 | 2 |
| 3 | 2 |
| 4 | 5 |
| 5 | 4 |

La autocorrelación de rezago 1, $\boldsymbol{\rho_1}$ , se calcula con los datos de la observación siguiente:

| t | $Y_t$ | $Y_{t+1}$ |
|---|---|---|
| 1 | 5 | 2 |
| 2 | 2 | 2 |
| 3 | 2 | 5 |
| 4 | 5 | 4 |
| 5 | 4 | |

En el cálculo no se considera el dato en gris.

Utilizando la fórmula COEF.DE.CORREL, de Excel, la autocorrelación de rezago 1 que se obtiene es –0.1924. El resultado indica que la observación actual tiene una correlación baja negativa con una observación anterior.

Otro proceso estacionario es el de medias móviles. En este proceso, la estimación de la observación actual se encuentra en función de los errores de las observaciones pasadas:

$$\boldsymbol{Y_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}}$$

Un proceso integrado es aquel que puede convertirse en estacionario aplicando diferencias. En cuanto al orden de integración de un proceso, es el número de diferencias que debemos aplicarle para convertirlo en estacionario.

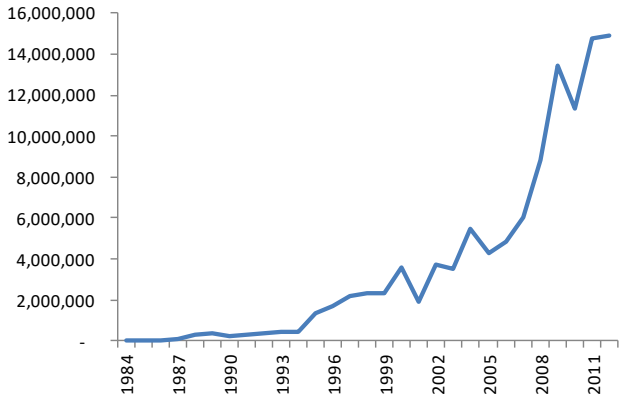
Estos modelos combinan procesos autorregresivos y de medias móviles a un proceso integrado.

Se denota ARIMA (p,d,q), donde *p* es el orden de la parte autorregresiva; *d*, el número de diferencias realizadas al modelo original para convertirla en estacionaria; y *q,* el orden de la parte de medias móviles.

| | |
|---|---|
| Se denota ARIMA (p,d,q), donde: | *p* es el orden de la parte autorregresiva; |
| | *d*, el número de diferencias realizadas al modelo original para convertirla en estacionaria; |
| | *q,* el orden de la parte de medias móviles. |

Se ejemplifica este modelo con las ventas registradas en el periodo 1984-2012 de una empresa (véase la gráfica correspondiente).

**Ventas anuales 1984-2012**

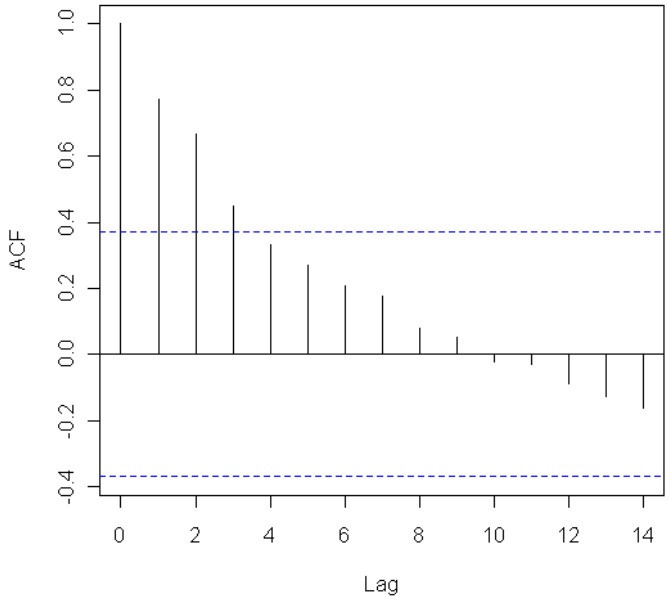



Fuente: elaboración propia.

A partir de 1994, se observa una tendencia creciente en las ventas, la cual se acentúa desde 2005.

Luego, se calculan las autocorrelaciones de la serie con diferentes rezagos y se grafica. A este gráfico se le conoce como *autocorrelograma.*

**Autocorrelograma de la serie original**

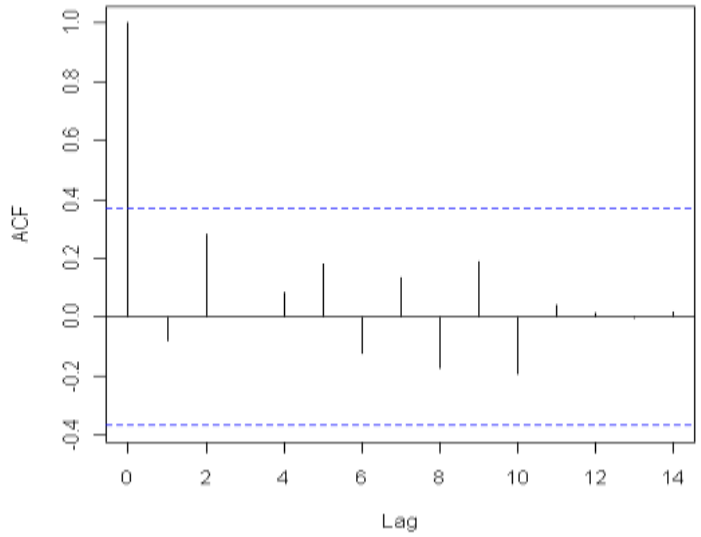



Fuente: elaboración propia. Datos procesados en el paquete estadístico R.

La gráfica anterior muestra que la observación actual está influenciada por una o dos observaciones anteriores. Después de ajustar varios modelos, se eligió un ARIMA (2, 2, 2), el que mejor se ajusta a la serie.

Para validar la calidad del modelo, se acostumbra realizar el autocorrelograma de los residuos.

**Autocorrelograma de los residuos del modelo ARIMA (2, 2, 2)**

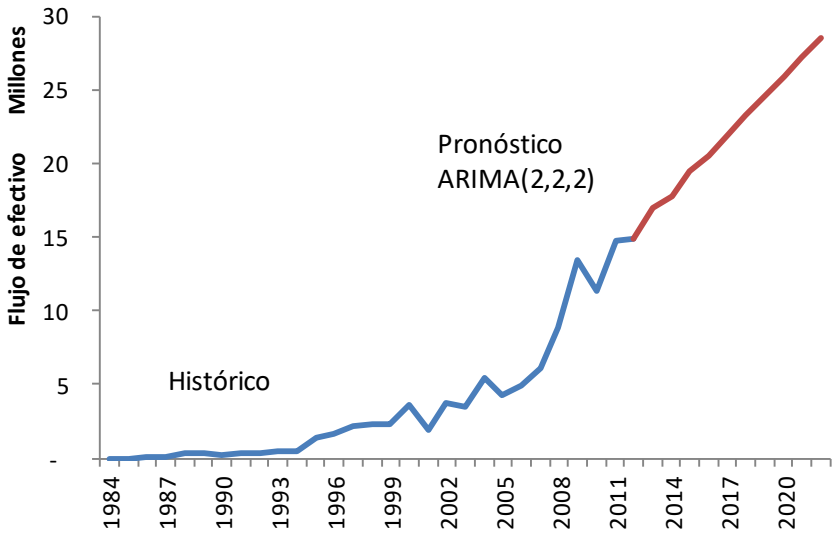



Fuente: elaboración propia. Datos procesados en el paquete estadístico R.

La gráfica anterior muestra que los residuos tienen un comportamiento de ruido blanco (aleatorio) porque se hallan dentro de la banda donde se espera caiga el 95% de las observaciones.

En la siguiente gráfica, se muestra la proyección de 10 observaciones de la serie con el empleo del modelo ARIMA (2, 2, 2).

**Pronóstico de la serie con el modelo ARIMA (2,2,2)**




Fuente: elaboración propia.

# RESUMEN

Una serie de tiempo es una observación de los valores de una variable durante un periodo, y consta de cuatro componentes: tendencia, estacionalidad, ciclicidad y un elemento irregular o aleatorio.

Una serie puede tratarse bajo dos enfoques: el aditivo y multiplicativo. En el primero, la serie se considera que es resultado de la suma de sus componentes; mientras que en el segundo, los componentes se expresan como factores que alteran la tendencia.

Para estimar la tendencia, se utilizaron los métodos de regresión lineal y promedios móviles. Para trabajar la estacionalidad, se estimaron factores aplicados a la tendencia. Para manejar la ciclicidad, se construyó una serie cíclica que, al restarse de la serie original, da como resultado una serie irregular, la cual es deseable que sea estacionaria para poder aplicar modelos autorregresivos o de medias móviles.

Por último, se expusieron los términos *irregular* y *aleatorio*, y se mencionaron las características del modelo ARIMA, cuya aplicación se ejemplificó en una serie.

# BIBLIOGRAFÍA

## SUGERIDA

| Autor | Capítulo | Páginas |
|---|---|---|
| Anderson, S. | 18 | 785-852 |
| Levin, R. | 15 | 673-718 |
| Lind, D. | 16 | 604-647 |

Anderson, S. (2012). *Estadística para negocios y economía* (11.ª ed.). México: CENGAGE Learning.

Levin R. y Rubin D. (2010). *Estadística para administración y economía* (7.ª ed.). México: Pearson.

Lind A. D., Marchal G., W. y Wathen, S. (2012). *Estadística aplicada a los negocios y economía* (15.ª ed.). México: McGraw-Hill.

Plan 2012

# 2016

actualizado