UPC
UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS
Laureate International Universities®

Estudios Profesionales para Ejecutivos - EPE

AUTORES : **WILMER PONCE**
GILBER PIÑA
CARLOS LOPEZ DE CASTILLA

COORDINADOR: **WILMER PONCE**

DOCENTES : **WILMER PONCE**
JIM SILVESTRE
ELMER ALIAGA
RAÚL ALIAGA
LUIS ÁNGEL CAJACHAHUA
CARMEN OGNIO
ENVER TARAZONA
MERCEDES TORIBIO
EBSON ALLENDE

TÍTULO: Guía de enseñanza

FECHA: Mayo 2013

Curso: Estadística para Ingeniería I

Código: CE54

Área: Ciencias

Ciclo: **2013 - I**

Capítulo 1
Estadística 1

1.1 Introducción

El término *estadística* se deriva de la palabra latina *status*, que significa estado. Los primeros usos de la estadística tienen que ver con la recopilación de datos y la elaboración de gráficas para describir diversos aspectos de un estado o un país. En 1662, Jhon Graunt publicó información estadística acerca de los nacimientos y los decesos. Al trabajo de Graunt siguieron estudios de tasas de mortalidad y de enfermedad, tamaño de poblaciones, ingresos y tasas de desempleo. En la actualidad los hogares, gobiernos y empresas se apoyan en la estadística para dirigir sus acciones.

Para muchas personas, *estadística* significa descripciones numéricas. Lo anterior puede verificarse fácilmente al escuchar, un domingo cualquiera, a un comentarista de televisión narrar un juego de fútbol. Sin embargo, en términos más precisos, la estadística es el estudio de los fenómenos aleatorios.

Uno de los aspectos más importantes en estadística es la posibilidad de obtener conclusiones basadas en los datos observados. Este proceso se conoce como *inferencia estadística*. Si una conclusión obtenida para un indicador económico importante, una posible concentración de cierto contaminante o si se pretende establecer una relación entre la incidencia de cáncer pulmonar y el fumar, es muy probable que dicha conclusión esté basada en la inferencia estadística.

1.2 Clasificación de la estadística

Estadística descriptiva

La estadística descriptiva comprende un conjunto de métodos y técnicas de recolección, organización, resumen y presentación que permiten describir apropiadamente las características de un conjunto de datos obtenidos desde una muestra o desde una población. Comprende el uso de gráficos, tablas, diagramas y criterios para el análisis.

Estadística inferencial

La estadística inferencial comprende un conjunto de métodos y técnicas que hacen posible *estimar* una o más características de una población basados en el resultado de muestras. En un procedimiento de esta naturaleza siempre existe la posibilidad de error, nunca podrá tenerse el 100% de seguridad sobre una afirmación que se base en la inferencia estadística. Sin embargo, lo que hace que la estadística sea una ciencia, separándola del arte de adivinar la fortuna, es que existe una medida de confiabilidad medida en términos de probabilidad.

1.3 Población y muestra

Para comprender la naturaleza de la inferencia estadística, es necesario entender las nociones de *población* y *muestra*. La población es la colección de toda la información posible que caracteriza a un fenómeno. Una población es cualquier colección, ya sea de un número finito de mediciones o una colección grande de datos acerca de un fenómeno de interés. Por otro lado, la muestra es un subconjunto representativo seleccionado de una población. Una muestra es *representativa* si refleja las características importantes de la población de la cual se obtuvo.

Ejemplo
El gerente de un complejo de oficinas contrata una nueva empresa de limpieza y quiere formarse una idea de cuán satisfactorio es el trabajo desarrollado por la mencionada empresa de limpieza. Para tal fin selecciona al azar 30 oficinas registrando la siguiente información:

I. Tiempo requerido para limpiar completamente la oficina.
II. Tamaño de la oficina (pequeña, mediana y grande)
III. Número de ventanas sucias por oficina.
IV. Si la oficina se encuentra alfombrada.

De acuerdo al enunciado anterior identifique la población y la muestra.

Población:

Muestra:

1.4 Parámetro y estimador

El objetivo de las técnicas de muestreo es asegurar que cada observación en la población tenga oportunidad de ser incluida en la muestra. Tales procesos de muestreo conducen a una *muestra aleatoria*. Los datos de la muestra aleatoria se usan para calcular ciertas características de la muestra denominadas *estadísticas*. Las estadísticas se usan como base para hacer inferencias acerca de alguna característica de la población, que recibe el nombre de *parámetro*.

Ejemplo.

- Según los Censos Nacionales XI de Población y VI de Vivienda 2007 ejecutados por el INEI, el 50,30% de los peruanos son mujeres, ¿este dato es un parámetro o una estadística?

  _____

- El 5 de marzo del 2011 Datum Internacional publicó una encuesta sobre intención de voto presidencial, la cual registró un 29% para Alejandro Toledo, ¿este dato es un parámetro o una estadística?

  _____

1.5 Variables

En cualquier análisis estadístico se manejan datos que provienen de la medición de una variable o variables seleccionadas en el estudio. Las *variables* son las características que interesan en los sujetos u objetos que se estudian, por ejemplo podría ser la edad de los empleados de una empresa, el monto de las ventas de determinado artículo, la ocupación de los clientes de cierto negocio, etc. Para obtener los datos correspondientes a las variables de interés, se requiere hacer una medición, como podría ser preguntar las edades de los empleados o la ocupación de los clientes, sin embargo los datos obtenidos serían de diferente tipo, pues para la primera variable serían números y para la segunda categorías. La medición se lleva a cabo en el momento en que se asigna un número al dato. Son ejemplos de variables:

1. El grado académico del jefe del hogar
2. El número de hijos de una familia
3. Ingreso familiar
4. Número de fallas de una computadora
5. Costos de un automóvil
6. Número de empleados de una oficina
7. Nivel Socioeconómico
8. Duración (min) de la conversación de una llamada telefónica
9. Cantidad de cemento vertida en una bolsa
10. Nivel de satisfacción de un nuevo producto de belleza

1.6 Tipos de variables

Las variables pueden clasificarse en:

a) **Variables cuantitativas,** son aquellas que consisten de números que representan conteos o mediciones. Las variables cuantitativas pueden ser:

   Discreta, corresponde al conteo o numeración de sucesos. Ejemplos: número de computadores en un aula, número de artículos defectuosos en un embarque, número de hijos por familia, etc.

   Continua, corresponde a "mediciones" y por tanto sus valores están comprendidos en un intervalo ya que entre dos valores existen infinitos valores intermedios. Ejemplos: contenido neto en gramos en una lata de café, diámetro de un tornillo, ingreso familiar, etc.

b) **Variables cualitativas** se dividen en diferentes categorías que se distinguen por alguna característica no numérica. Ejemplos: genero, nivel socio económico, grado de instrucción, etc.

Observación: Los datos pueden estar "codificados" numéricamente, sin que ello signifique que sea una variable cuantitativa.

Ejemplo. Clasifica las siguientes variables:

Variable	Tipo
Marca de detergente que una ama de casa usa.	
El grado de instrucción.	
Número de artículos defectuosos producidos por una máquina en un periodo de dos horas.	
La longitud de 1000 tornillos con arandela y tuerca.	

1.7 Escalas de medición

Otra forma común de clasificar las variables es mediante el uso de cuatro niveles de medición; *nominal*, *ordinal*, de *intervalo* y de *razón*. Cuando la estadística se aplica a problemas reales el nivel de medición de las variables es un factor importante para determinar el procedimiento a usar.

Escala nominal

Los diferentes valores que toma una variable medida en escala *nominal* **carecen de un orden o de un significado numérico**, no pueden utilizarse para realizar cálculos. A menudo se asignan números a las diferentes categorías pero no tienen significado computacional y cualquier promedio que se calcule carece de sentido.

Ejemplos: genero (femenino, masculino), nacionalidad (peruano, colombiano, etc.), marcas de gaseosa preferida (fanta, coca cola, pepsi etc).

Escala ordinal

La escala de medición *ordinal* permite **establecer un orden entre los diferentes valores que toma la variable**. En esta escala de medición no es posible determinar las diferencias entre los valores de la variable o tales diferencias carecen de sentido.

Ejemplos: grado de satisfacción (alto, medio, bajo), estado de salud (bueno, regular, malo), grado de instrucción (primaria, secundaria, superior).

Escala de intervalo

La escala de medición de *intervalo* tiene las propiedades de la escala ordinal pero con la característica adicional que la diferencia entre dos valores cualesquiera de la variable tiene un significado. Sin embargo, los valores en esta escala **no tienen un punto de partida natural o cero**.

Ejemplos: Temperatura (se puede medir en grado Celsius ó grado Fahrenheit), año calendario (el año puede referirse al calendario gregoriano ó calendario chino), en el caso de la escala de intervalos podemos asignar el cero a cualquier valor posible (el cero es arbitrario).

Escala de razón

La escala de medición de *razón* tiene las propiedades de la escala de intervalo con la propiedad adicional de tener **un punto de partida natural o cero**, que indica que *ausencia* de la variable (no existe nada para una variable). Para esta escala de medición, es posible establecer una relación de proporcionalidad entre sus distintos valores, es decir el cociente de los valores es significativo.

Ejemplos: peso de un recién nacido, estatura de una persona, costo de un automóvil, número de aprobados, número de artículos defectuosos.

Tipo de variable y escala de medición

TIPO DE VARIABLE: CUALITATIVA				TIPO DE VARIABLE: CUANTITATIVA			
NOMINAL		ORDINAL		INTERVALO		RAZÓN	
Ningún atributo		**Un atributo: Orden**		**Dos atributos: Orden y Distancia**		**Tres atributos: Orden, Distancia y Origen**	
Las categorías de la variable no tienen orden implícito.		Las categorías de la variable tienen orden, pero, no permite cuantificar distancia entre una y otra categoría.		Tiene intervalos iguales y que pueden ser medidos, pero no tiene punto de partida. Puede asumir valores negativos.		Tiene intervalos constantes entre un valor y otro, posee un punto de partida (cero natural) lo que indica que el cero es ausencia de la variable.	
Distrito donde vive	Género	Nivel de satisfacción	Grado de instrucción	Temperatura	Año	Talla	Defectos de un producto
Surco Barranco Miraflores	Masculino Femenino	Muy Bueno Bueno Regular Malo Muy Malo	Inicial Primaria Secundaria Superior	15 °C 10 °C 5 °C 0 °C -5 °C -10°C -15°C	500 AC 100 AC 200 DC 1290 DC	1.56 cm 1.58 cm 1.66 cm 1.51 cm	0 1 2 3 4 5
Variable cualitativa dicotómica: Tiene dos categorías Género: Masculino, Femenino **Variable cualitativa politómicas:** Tiene más dos categorías Nivel de satisfacción de un producto: Bueno, Regular, Malo				**Variable cuantitativa continua:** proviene de medir. Entre dos valores siempre existe un número intermedio. Talla, Peso, Duración de un producto, etc. **Variable cuantitativa discreta:** proviene contar. Solo se representa con números enteros. Número de hijos, Número de profesores, Cantidad de glóbulos rojos en una muestra, etc.			

Fuente: www.bioestaditico.com
Elaboración: Profesores del curso

Ejemplos

Alpha Soft es una compañía que brinda servicios informáticos a empresas que deseen tener una presencia firme en la red. Se dedica al tendido de redes LAN (Local area network), instalación de equipos, servidores, etc. Como parte de un estudio, se analiza la información correspondiente a las siguientes variables:

I. Tipos de lenguajes de programación (Cobol, Java, Informixs-4gl, etc.)
II. Cantidad de servidores por empresa.
III. Costo de las licencias de software (en dólares)
IV. Versión instalada de Office (Office XP, Office 2007, Office 2010)

Para la lista anterior identifique el tipo de variable y la escala de medición correspondiente.

Variable	Tipo	Escala
I.		
II.		
III.		
IV.		

Marque con una X según la escala de medición de las siguientes variables:

Variable	Nominal	Ordinal	Intervalo	Razón
Año de nacimiento				
Código de un alumno(a) de la UPC				
Tiempo de vida de una persona				
Número de hermanos de una persona				

1.8 Ejercicios

1. La empresa *Beta Internet* llevó a cabo un estudio entre los usuarios de Internet en Lima para poder analizar el mercado de posibles compradores a través de este servicio. Como parte del estudio realizado se analizó la información correspondiente a una muestra de 1500 usuarios con relación a las siguientes variables:

   I. Edad del principal usuario.
   II. Número de computadoras en casa con servicio de Internet.
   III. Calidad del servicio de Internet (bueno, regular, deficiente).
   IV. Tiempo diario de uso del servicio de Internet (en horas).
   V. Marca de la computadora con mayor uso del servicio de Internet.

   De acuerdo al enunciado anterior identificar la población, muestra, el tipo de variable y la escala de medición correspondiente.

2. Una entidad estatal realizó un estudio para determinar algunos indicadores socioeconómicos de los inmigrantes peruanos en Estados Unidos. El estudio se llevó a acabo aplicando encuestas a una muestra de 400 inmigrantes peruanos. Algunas de las preguntas del cuestionario fueron las siguientes:

   I. Nivel educativo alcanzado en Perú.
   II. Monto enviado (en dólares) el último mes a sus familiares en Perú.
   III. ¿Presenta alguno de los siguientes problemas? (indocumentado, discriminado, vivienda no adecuada, salud y otros).
   IV. Número de veces que ha visitado el Perú desde que viajó por primera vez a los Estados Unidos.

   De acuerdo al enunciado anterior identificar población y muestra. Para las respuestas generadas por cada una de las preguntas mostradas identificar el tipo de variable y la escala de medición correspondiente.

3. Prevenir la propagación de las fisuras por desgaste en las estructuras aeronáuticas es un elemento importante de la seguridad en la aviación. En un estudio de ingeniería para investigar las fisuras por desgaste en aviones del aeropuerto de Frankfur, se toma una muestra de 50 de aviones a los que se verifican sus alas, sometidas a cargas por unidad de superficie cíclicas reportó los siguientes resultados:

   I. El 35% de los aviones se fisuraron en el alerón izquierdo
   II. El 74% presentó defectos leves fácilmente rebatibles
   III. El 40% tuvo problemas por falta de mantenimiento y el 30% por sobrecarga
   IV. El tiempo promedio de vuelo de estos aviones fue de 5.7 años
   V. Número promedio de problemas por avión fueron 4

   Para la lista anterior identifique, la población, muestra, unidad elemental, el tipo de variable y la escala de medición correspondiente.

Capítulo 2

Estadística Descriptiva: Métodos gráficos

2.1 Introducción

Este capítulo presenta algunos procedimientos básicos para resumir un conjunto de datos cualitativos y cuantitativos. Estos resúmenes tabulares y gráficos pueden verse en informes anuales, artículos periodísticos y estudios de investigación. De alguna manera estamos familiarizados con este tipo de presentaciones por lo que es importante comprender como se preparan y aprender a interpretarlos de forma correcta.

2.2 Resumen de datos cualitativos

Tablas de distribución de frecuencias

Cuando se trabaja con grandes conjuntos de datos cualitativos es útil organizarlos y resumirlos por medio de la construcción de una tabla que muestre las diferentes categorías de la variable en estudio junto al número de veces que cada una de ellas se repite.

Ejemplo 2.1: Los datos en la hoja **Dominios** tiene información para una muestra sobre los dominios de segundo nivel registrados bajo la categoría .pe. Para construir la tabla de distribución de frecuencias es necesario contar el número de veces que aparece cada dominio.

Dominio de Segundo Nivel

Dominio	f	h	p
com.pe	285	0,570	57,0
org.pe	106	0,212	21,2
edu.pe	64	0,128	12,8
gob.pe	26	0,052	5,2
net.pe	3	0,006	0,6
tros	16	0,032	3,2
	500		

Fuente: Elaboración propia

Donde:

f Frecuencia absoluta simple.

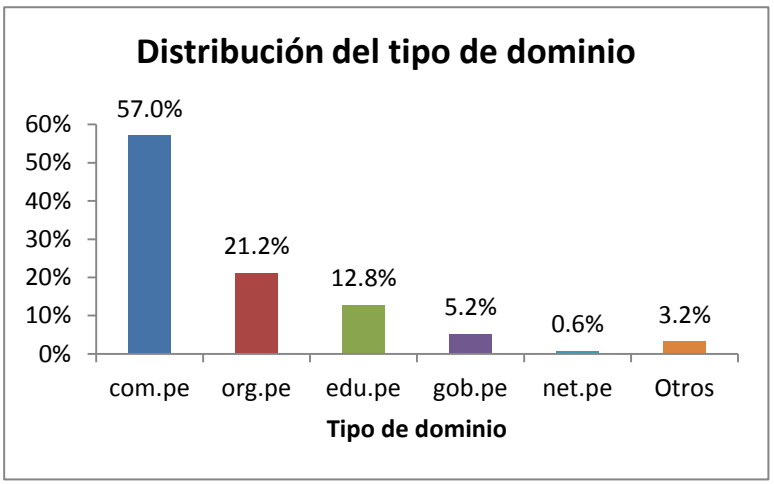
h Frecuencia relativa simple, $h = f/n$.

p Frecuencia porcentual simple, $p = 100 \times h$.

Diagrama de barras

Un diagrama de barras es una gráfica para representar un conjunto de datos cualitativos que se han resumido en una distribución de frecuencias absolutas, relativas o porcentuales. En uno de los ejes de la gráfica, por lo general el eje horizontal, se colocan las categorías de la variable y en el otro eje de la gráfica (por lo general el eje vertical) se pueden usar las frecuencias absolutas, relativas o porcentuales. Luego, se construye una barra de ancho fijo en cada categoría y cuya altura corresponda a la frecuencia utilizada en el gráfico. Las barras deben estar separadas para enfatizar el hecho que las categorías no se superponen.

Ejemplo 2.2: El diagrama de barras para los datos del Ejemplo 2.1 se muestra a continuación.

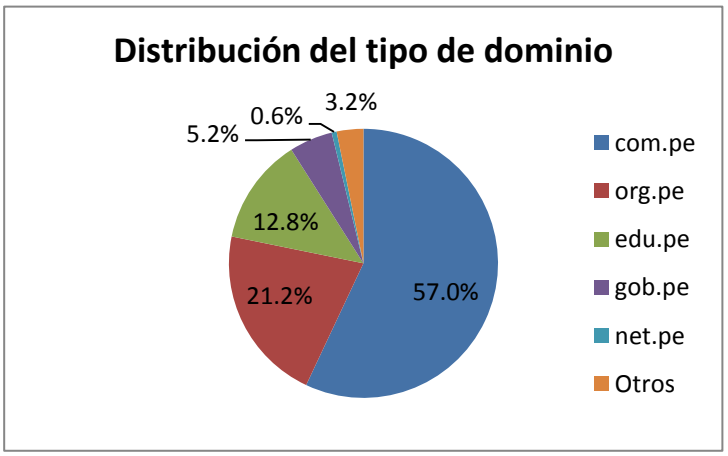


Fuente: Elaboración propia

Diagrama circular

El diagrama circular, también llamado *diagrama de pie*, es otra forma de representar la distribución de frecuencias relativas o porcentuales. Para su construcción primero se traza un círculo y luego se divide en sectores circulares de forma proporcional a la frecuencia relativa de cada categoría. Puesto que todo el círculo representa un ángulo de 360º en total cada sector es el porcentaje correspondiente de dicho total, es decir, el ángulo del sector que le corresponde a cada categoría se obtiene multiplicando 360º por la respectiva frecuencia relativa.

Ejemplo 2.3: El diagrama circular para los datos del Ejemplo 2.1 se muestra a continuación.



Fuente: Elaboración propia

Diagrama de Pareto

El nombre de Pareto fue dado por Joseph Juran (filosofía de gestión de calidad del siglo 20) en honor del economista italiano VILFREDO PARETO (1848-1923) quien realizó un estudio sobre la distribución de la riqueza, en el cual descubrió que la minoría de la población poseía la mayor parte de la riqueza y la mayoría de la población poseía la menor parte de la riqueza.

El diagrama de Pareto es una representación gráfica que permite identificar y seleccionar los aspectos prioritarios que hay que tratar en un determinado problema. También se conoce como diagrama ABC o Ley de las prioridades 20-80, que dice: "*el 80% de los problemas que ocurren en cualquier actividad son ocasionados por el 20% de los elementos que intervienen en producirlos*". Según este concepto se tiene un problema con muchas causas, podemos decir que el 20% de las causas resuelven el 80 % del problema. Esto también se conoce como "los pocos vitales y muchos triviales".

Se recomienda el uso de Pareto:

- Al buscar las causas principales de los problemas y establecer la prioridad de las soluciones
- Para identificar oportunidades para mejorar
- Para identificar un **producto o servicio** para el análisis de mejora de la calidad.
- Para analizar las diferentes agrupaciones de datos.
- Para evaluar los resultados de los cambios efectuados a un proceso comparando sucesivos diagramas obtenidos en momentos diferentes, (antes y después).

Este diagrama sirve para conseguir el mayor nivel de mejora con el menor esfuerzo posible mediante una revisión de la "minoría de características vitales" para optimizar todos los recursos disponibles. Con este análisis "descartamos a la mayoría de los triviales (sin importancia)"

Ejemplo de "minoría de características vitales"

- La minoría de rechazos que representa la mayoría de quejas de los clientes.
- La minoría de vendedores que está vinculada a la mayoría de productos rechazados.
- La minoría de problemas causantes del grueso del retraso de un proceso industrial.
- La minoría de productos que representan la mayoría de las ganancias obtenidas.
- La minoría de elementos que representan la mayor parte del costo de un inventario
- La minoría de clientes que representen la mayoría de las ventas.
- La minoría de procesos de la calidad causantes de la mayoría de pérdidas o mermas.
- etc,etc

Fuente: http: //www.elprisma.com/apuntes/ingenieria_industrial/diagramadepareto/

Los pasos para la elaboración del diagrama de Pareto son:

1. Construya una tabla de distribución de frecuencias ordenando las categorías en forma descendente respecto a la frecuencia.

2. La categoría *Otros* deberá ser colocada en la última posición, no importa cuán grande sea, porque está compuesta de un grupo de categorías cuyas frecuencias son menores en relación al valor de la variable con frecuencia más pequeña listado individualmente.
3. Agregue a la tabla de distribución de frecuencias una columna para las frecuencias acumuladas absolutas F.
4. Elabore el diagrama de Barras y agregue la línea de frecuencias acumuladas relativas $H = F/n$.

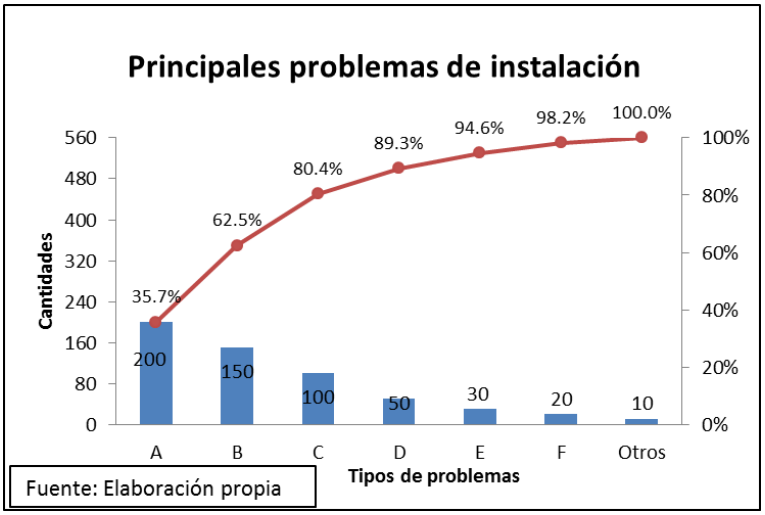
Ejemplo 2.4: Se tiene información sobre los problemas de soporte técnico encontrados durante la instalación y configuración de Exchange 2000 Server y Exchange Server 2003. Construya un gráfico apropiado para poder identificar los problemas principales. Los datos se encuentran en la hoja **Instalación**. La descripción de los problemas mencionados y la tabla de distribución de frecuencias se muestran a continuación.

Problema	Descripción
A	Error de actualización de Exchange Server 5.5
B	Permisos insuficientes
C	Problemas con el sistema de nombres de dominio (DNS)
D	Forestprep o Domainprep no finalizaron correctamente
E	Permisos de Active Directory insuficientes
F	Falta el contenedor de conexiones de Active Directory
Otros	Otros problemas

Distribución de problemas de instalación

Problema	f	h	F	H
A	200	0.357	200	0.357
B	150	0.268	350	0.625
C	100	0.179	450	0.804
D	50	0.089	500	0.893
E	30	0.054	530	0.946
F	20	0.036	550	0.982
Otros	10	0.018	560	1.000
	560			

Fuente: Elaboración propia



2.3 Resumen de datos cuantitativos

Tablas de distribución de frecuencias para datos cuantitativos discretos

Si los datos son cuantitativos discretos y no existen muchos valores diferentes es posible construir la tabla de distribución de frecuencias haciendo una lista ordenada e individual de los valores y sus respectivas frecuencias.

Ejemplo 2.5: Los datos en la hoja **Interrupciones** corresponden al número de interrupciones semanales en el sistema de red de una prestigiosa universidad registrado el último año. La tabla de distribución de frecuencias para estos datos se muestra continuación.

Número de interrupciones semanales

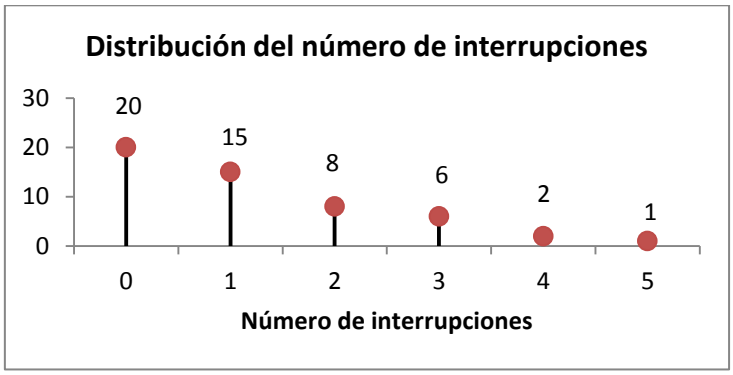
Interrupciones	f	h_i	p_i
0	20	0.384	38.5
1	15	0.288	28.8
2	8	0.153	15.4
3	6	0.115	11.5
4	2	0.038	3.8
5	1	0.019	1.9
Total	52	1	100

Fuente: Elaboración propia

Diagrama de líneas

El diagrama de líneas, también llamado *gráfico de bastones o varas*, es una representación gráfica para un conjunto de datos cuantitativos discretos resumidos en una distribución de frecuencias. En el eje horizontal se colocan los diferentes valores ordenados de la variable y en **el eje vertical se pueden utilizar las frecuencias absolutas, relativas o porcentuales**. Este gráfico se construye de manera similar al diagrama de barras con la diferencia que en lugar de columnas se utilizan líneas.

Ejemplo 2.6: El diagrama de líneas para los datos del Ejemplo 2.5 se muestra a continuación.



Fuente: Elaboración propia

Ejemplo 2.7: Para ocupar tres vacantes de digitadoras, se evaluaron a 30 candidatas para ello se tomaron varias pruebas, el objetivo de una de ellas era evaluar el número de errores cometidos en la escritura de un texto, se ha registrado número de errores de las participantes:

2	0	2	4	3	4	5	1	6	6
2	3	2	3	4	3	3	6	5	4
4	3	3	4	3	5	4	5	1	4
2	1	3	5	3	5	4	5	1	6

Construya una tabla de frecuencias y un diagrama con la información anterior e interprete el resultado.

Tablas de distribución de frecuencias para datos cuantitativos continuos

Cuando se trabajan con datos cuantitativos continuos los valores no tienden a repetir con mucha frecuencia. En este caso es preferible agrupar los datos usando intervalos de clase y considerando la cantidad de datos que se encuentran dentro de ellos. Los pasos necesarios para definir estos intervalos de clase son:

1. Calcular el *rango* (R) o *amplitud* como la diferencia entre el valor más grande y el más pequeño en el conjunto de datos que se desea agrupar. $R = x_{máx} - x_{mín}$.
   Determinar el número de intervalos según la regla de *Sturges*:
   $k = 1 + 3.322\ log_{10}n$, donde n es el número de datos a agrupar. El valor obtenido debe ser redondeado al entero más cercano.

2. Determinar el ancho de los intervalos según $w = R/k$. El valor obtenido debe redondearse por exceso a la misma cantidad de decimales que tienen los datos.

3. Determinar los límites inferior (LI) y superior (LS) de los intervalos
   $LI_1 = X_{min}$
   $LI_i = LI_{i-1} + w$
   $LS_i = LI_{i+1} = LS_{i-1} + w$

4. Para evitar que una observación pueda ser considerada en más de un intervalo, cada uno de los intervalos se considera **abierto a la izquierda y cerrado a la derecha**.

5. Finalmente, dado que los datos originales se pierden, será necesario calcular un valor que sea representativo de todas las observaciones que pertenecen a un intervalo. A este valor se le denomina marca de clase y su cálculo es:
   $$X'_i = (LI_i + LS_i) / 2$$

Ejemplo 2.8: *Alpha Systems S.A.* emprendió un estudio para determinar el comportamiento de un sistema de grabación de programas informáticos. Para que el proceso funcione adecuadamente la señal debe estar entre 9,2 y 10 voltios. Se instalaron los sistemas de grabación y se tomaron lecturas respectivas. Los datos registrados se encuentran en la hoja **Lecturas**. Los pasos a seguir son:

- $R = 10,55 - 8,05 = 2,5$
- $k = 1 + 3,322 log_{10} 70 = 7.1294 \cong$ **7** intervalos de clase
- $w = 2,5/7 = 0,3571 \cong 0,36$

Cálculo de los límites

$LI_1 = 8,05$ $LS_1 = 8,05 + 0,36 = 8,41$
$LI_2 = 8,41$ $LS_2 = 8,41 + 0,36 = 8,77$
$LI_3 = 8,77$ $LS_3 = 8,77 + 0,36 = 9,13$
.
.
.
$LI_7 = 10,21$ $LS_7 = 10,21 + 0,36 = 10,57$

Cálculo de las marcas de clase

$X'_1 = (8,05 + 8,41) / 2 = 8,23$
$X'_2 = (8,41 + 8,77) / 2 = 8,59$
$X'_3 = (8,77 + 9,13) / 2 = 8,95$
.
.
.
$X'_7 = (10,21 + 10,57) / 2 = 10,39$

Distribución de las lecturas de la señal de grabación

i	Lectura		X´	f	h	F	H
1	[8,05	8,41]	8,23	1	0,0143	1	0,0143
2	< 8,41	8,77]	8,59	4	0,0571	5	0,0714
3	< 8,77	9,13]	8,95	6	0,0857	11	0,1571
4	< 9,13	9,49]	9,31	9	0,1286	20	0,2857
5	< 9,49	9,85]	9,67	15	0,2143	35	0,5000
6	<9,85	10,21]	10,03	27	0,3857	62	0,8857
7	< 10,21	10,57]	10,39	8	0,1143	70	1,0000
				70			

Fuente: Elaboración propia

A partir de la tabla de frecuencias interpretar lo siguiente:

1. X'_2 , f_3 , F_4 , $H_5 - H_3$
   X'_2 **:** 8,59 es el valor representativo del segundo intervalo.
   f_3 **:** 6 registros de un sistema de grabación de programas informáticos tienen más 8,77 y a lo más 9,13 voltios.
   F_4 **:**20 registros de un sistema de grabación de programas informáticos tienen hasta 9,49 voltios.
   $H_5 - H_2$: **:** 0,5 – 0,071= 0,429 el 42,9% de registros de un sistema de grabación de programas informáticos tienen entre más 8,77 y a lo más 9,85 voltios.
2. ¿Cuántos registros de un sistema de grabación de programas informáticos tienen más 9,49 y como máximo 10,21 voltios?
   15 + 27 = 42 registros

3. ¿Se desea estimar que porcentaje de registros de un sistema de grabación de programas informáticos que tienen más 10,21 voltios?

   1- H_6= 1- 0,8857 = 0,1143 => 11,43%

Ejemplo 2.9: Hoy en día las empresas usan software de computadoras para proteger mejor la información electrónica crítica a través de técnicas como la encriptación, autenticación, herramientas de redacción, protección antivirus, etc. Los siguientes datos corresponden a los gastos mensuales (cientos de nuevos soles) en protección de la información para la empresa *Atlas PC*:

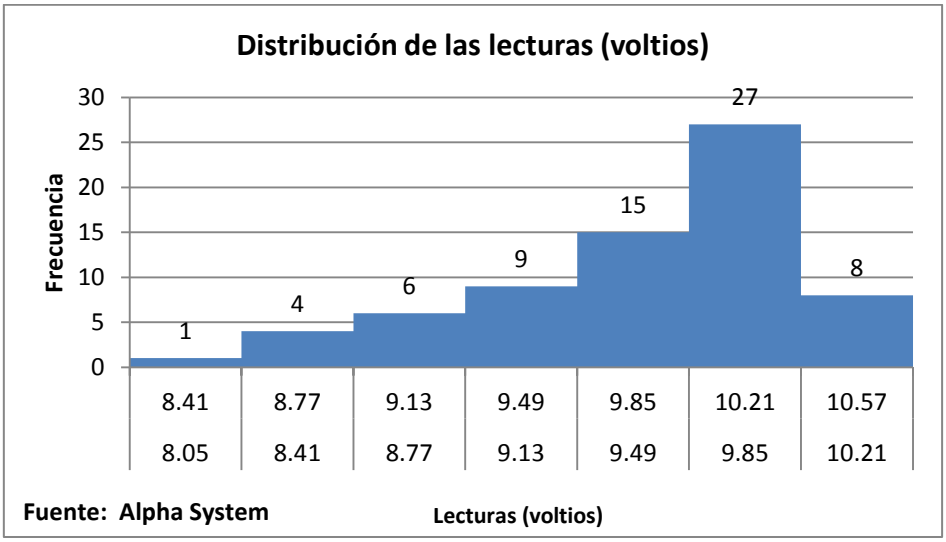
101,3	105,4	105,4	107,1	107,6	108,6	108,6	110,5	112,9	113,2	187,0	196,4
114,7	115,0	115,1	115,9	117,8	117,8	120,0	120,0	121,8	123,8	187,9	198,3
123,8	124,8	127,5	128,2	132,8	133,6	134,2	135,3	136,3	136,6	188,0	198,5
139,2	139,2	139,4	140,1	141,6	145,7	145,9	148,5	149,5	149,5	188,2	
150,2	150,3	150,4	150,8	154,0	158,0	158,4	158,6	159,1	163,7	189,6	
164,1	165,3	166,3	168,7	168,8	169,9	170,8	172,9	174,5	178,0	190,0	
179,7	179,8	179,9	180,0	181,5	183,0	183,3	184,9	186,6	186,7	194,4	

Obtenga la tabla completa de distribución de frecuencias para los gastos mensuales en protección de la información en esta empresa. Use la regla de Sturges.

Histograma

El *histograma* es la representación por medio de barras rectangulares cuya base de cada barra proporcional a la **amplitud de la clase**, su centro es la marca de clase y **la altura la su frecuencia absoluta o relativa simple**. En el eje horizontal se coloca la variable de interés y en el vertical la frecuencia correspondiente.

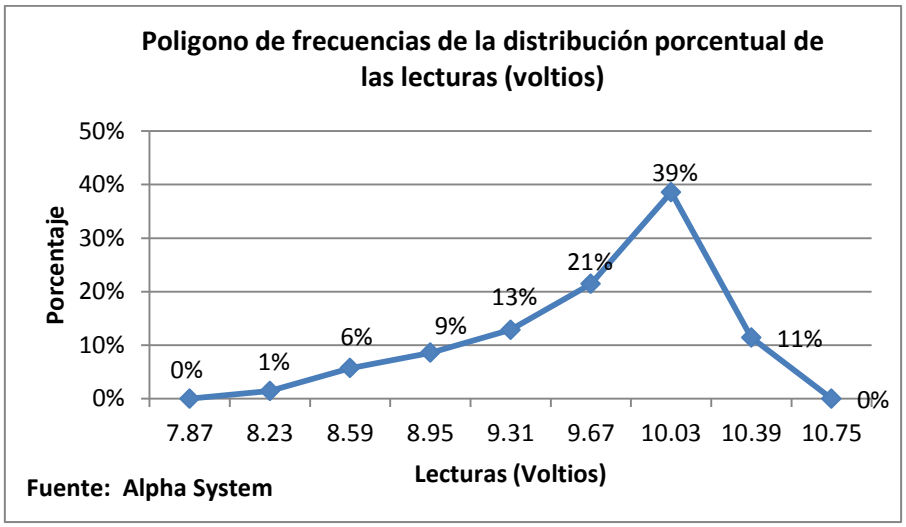
Ejemplo 2.10: El histograma para los datos agrupados del Ejemplo 2.8 se muestra a continuación.



Polígono de frecuencias

El *polígono de frecuencias* es una representación gráfica por medio de una figura poligonal cerrada. Se obtiene uniendo con segmentos de recta **el punto de intersección de cada marca de clase y su frecuencia relativa simple** correspondiente.

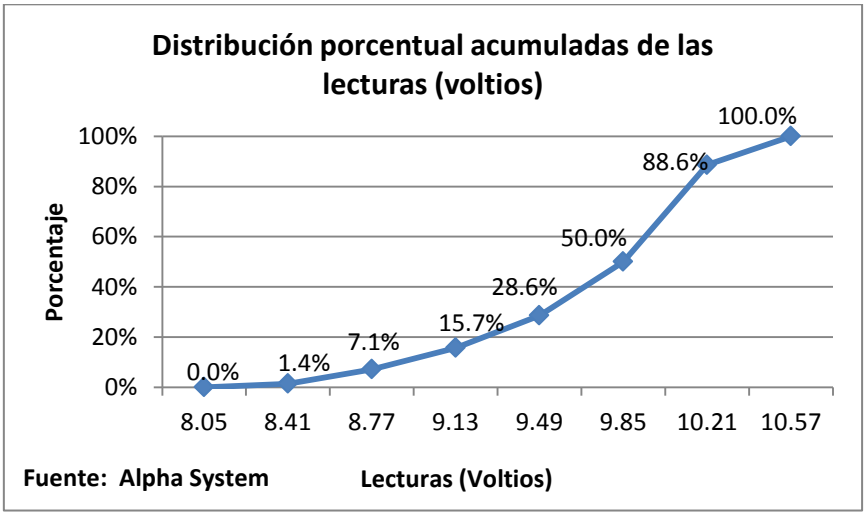
El polígono de frecuencias para los datos agrupados del Ejemplo 2.8 se muestra a continuación.



Ojiva

La *Ojiva* es la representación gráfica para las frecuencias relativas acumuladas. Se obtiene uniendo con segmentos de recta **el punto de intersección del límite superior de cada intervalo y su frecuencia relativa acumulada respectiva**. La *Ojiva* permite obtener, de manera aproximada, la proporción de los datos que pertenece a un intervalo cualquiera.

Ejemplo 2.11: La Ojiva para los datos agrupados del Ejemplo 2.8 se muestra a continuación.



2.4 Problemas

1. Los sistemas de cómputo colapsan por muchas razones, entre ellas las fallas de hardware o software, errores del operador, sobrecargas del sistema mismo y a otras causas. Los resultados obtenidos en un estudio acerca de las causas de colapso en una muestra de 200 sistemas de cómputo se encuentran en la hoja **Colapso**. Suponga que se debe priorizar entre las dos principales causas de colapso de los sistemas de cómputo. Elaborar un gráfico adecuado que permita establecer una estrategia apropiada.

2. El tiempo de inactividad es la cantidad del tiempo en el que un sistema de computadora no opera debido a una falla de hardware y software. Cuando el sistema se encuentra inactivo durante más de una hora todos los archivos de trabajo actuales se pierden. Los datos agrupados en la hoja **Inactividad** corresponden a los tiempos de inactividad (en minutos) registrados en una muestra de 30 y 28 sistemas de computadoras del turno mañana y tarde respectivamente.

i	Tiempo de inactividad		Marca de clase	Turno mañana		Turno tarde	
				f	h	f	h
1	20,5	25,5	23	1		1	
2	25,5	30,5	28	3		6	
3	30,5	35,5	33	0		3	
4	35,5	40,5	38	2		12	
5	40,5	45,5	43	20		6	
6	45,5	50,5	48	4		0	
				30		28	

   a. Construir el polígono de frecuencias para los datos agrupados del turno mañana y tarde en un solo gráfico.
   b. Elaborar un informe sustentado en el gráfico obtenido en el punto anterior.

3. Una distribuidora de circuitos integrados para computadoras vende su producto en lotes de 100 circuitos. Para la inspección de calidad se examinan 20 circuitos elegidos al azar de cada lote y se rechaza en caso de encontrar más de 4 circuitos defectuosos. Los datos en la hoja **Circuitos** muestra el número de circuitos defectuosos luego de inspeccionar 52 lotes. Construir una tabla de distribución frecuencias para la variable en estudio.

4. Se quiere comparar los resultados obtenidos por tres operadores, en cuanto a los tiempos (en minutos) utilizados en el restablecimiento de servicio de redes en el sistema de una importante entidad bancaria. Los resultados obtenidos se encuentran en la hoja **Operador**.

   a. Construir las tablas de distribución de frecuencias para los tres operadores por separado. Use la regla de Sturges.
   b. Construir los histogramas respectivos. Use la frecuencia relativa simple para cada gráfico.
   c. Elaborar un informe sustentado en los gráficos obtenidos en el punto anterior.

5. El jefe del área de sistemas de una empresa realizó un estudio para analizar el tiempo que demoran los empleados en detectar y resolver un problema informático de software o hardware. El estudio trata de estudiar las diferencias de tiempos en los turnos de trabajo A y B de una empresa. La hoja **Tiempo** muestra los datos obtenidos para una muestra de 40 problemas informáticos presentados en el turno A y 100 problemas informáticos presentados en el turno B.

   a. Construir la tabla de distribución de frecuencias para el turno A. Use la regla de Sturges.
   b. Construir los histogramas respectivos. Use la frecuencia relativa simple para cada gráfico.
   c. Elaborar un informe sustentado en los gráficos obtenidos en el punto anterior.

Capítulo 3

Estadística Descriptiva: Métodos numéricos

3.1 Introducción

En este capítulo se presentan algunos métodos numéricos de la estadística descriptiva que proporcionan algunas alternativas para resumir datos. Las medidas numéricas de tendencia central y dispersión se calculan a partir de los valores de los datos. **Si estos datos pertenecen a una muestra se denominan *estadísticos* mientras que si se calculan a partir de los datos de una población se denominan *parámetros*.**

3.2 Medidas de posición o localización

Media

Es quizás la medida numérica más importante de localización. La *media*, *media aritmética* o *promedio* de un conjunto de datos es la suma de dichos valores dividida entre el número total de datos.

Datos no agrupados

Media poblacional: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ Media muestral: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Datos agrupados

Media poblacional: $\mu = \frac{1}{N}\sum_{i=1}^{k} x_i^{/} f_i$ Media muestral: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i^{/} f_i$

Donde k es el número de intervalos, x_i es el valor del dato individual., $x_i^{/}$ la marca de clase, n es el tamaño de la muestra y N es el tamaño de la población.

Propiedades de la media

1. La media aritmética se ve afectada por valores extremos.
2. La media aritmética de una cantidad constante es la misma constante:

$$\text{Media}(k) = k$$

3. La media de una variable a la que se le suma o resta una constante equivale a la media de la variable más o menos dicha constante:

$$\text{Media}(X \pm k) = \text{Media}(X) \pm k$$

4. La media de una variable multiplicada por una constante equivale a la media de la variable multiplicada por dicha constante:

$$\text{Media}(kX) = k\text{Media}(X)$$

5. En general, para dos constantes a y b tenemos:

$$\text{Media}(aX \pm b) = a\text{Media}(X) \pm b$$

Ejemplo 3.1: Se sabe que los ingresos mensuales que perciben los trabajadores de una empresa tienen un promedio de 2800 nuevos soles, determine el nuevo ingreso promedio que se obtendría luego de realizar los siguientes aumentos:

a. Los ingresos de los trabajadores se incrementan en 250 nuevos soles.
b. Los ingresos se incrementan en un 7,5%
c. Los ingresos se incrementan en un 12,5% más una bonificación de 125 nuevos soles.

Solución

En cada caso el ingreso se denotará por X y el nuevo ingreso será representado por Y.

a) $Y = X + 250$
Media (Y) = Media(X) + 250
Media (Y) = 2800 + 250 = 3050 nuevos soles.

b) $Y = X + 0{,}075X = 1{,}075X$
Media (Y) = Media (1,075X)
Media (Y) = (1,075)(2800) = 3010 nuevos soles.

c) $Y = 1{,}125X + 125$,
Media (Y) = Media (1,125X +125)
Media (Y) = (1,125)(2800)+ 125 = 3275 nuevos soles.

Media ponderada

Se utiliza cuando los datos a promediar no tienen la misma importancia relativa dentro del conjunto total, es decir, algunos datos tiene mayor importancia, peso ó ponderación dentro del conjunto de observaciones. Se usará la siguiente expresión para su cálculo:

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

donde:

x_i : Dato individual.

w_i : Peso asignado a cada dato

Ejemplo 3.2: Una compañía vende cuatro tipos de vallas a los propietarios locales. El costo de instalación por tipo de valla se presenta en el siguiente cuadro.

Tipo de valla	Costo de instalación
Tipo A	S/. 20
Tipo B	S/. 12
Tipo C	S/. 8
Tipo D	S/. 6,5

La compañía instaló, 100 metros del tipo A, 150 metros de B, 75 metros de C y 200 de D. ¿Cuál es el costo medio del metro de valla instalado por la compañía?
Solución

$$\bar{x}_w = \frac{20(100)+12(150)+8(75)+6,5(200)}{100+150+75+200} = 10,85$$

El costo medio por concepto de instalación fue S/. 10.85

Ejemplo 3.3: Una empresa tiene tres turnos de trabajo en sus dos locales: Alpha y Beta. Los trabajadores en cada uno de estos turnos participan de acuerdo a la siguiente tabla:

Turno	Salario por hora (n.s.)	Número de horas trabajadas	
		Local Alpha	Local Beta
Mañana	12	50	40
Tarde	15	20	35
Noche	18	10	25

El salario promedio por hora en cada local es:

$$\underline{\text{Local Alpha}}: \ \bar{x}_w = \frac{12(50)+15(20)+18(10)}{50+20+10} = 13,50 \text{ nuevos soles}$$

$$\underline{\text{Local Beta}}: \ \bar{x}_w = \frac{12(40)+15(35)+18(25)}{40+35+25} = 14,55 \text{ nuevos soles}$$

Ejemplo 3.4: Los datos siguientes corresponden al número de interrupciones semanales en el sistema de red de una prestigiosa universidad registrado el último año. La tabla de distribución de frecuencias para estos datos se muestra continuación. Calcule e interprete la media del número de interrupciones semanales.

Número de interrupciones semanales

Interrupciones	f
0	20
1	15
2	8
3	6
4	2
5	1
Total	52

Fuente: Elaboración propia

Solución

$$\overline{x}_w = \frac{0(20)+1(15)+2(8)+3(6)+4(2)+5(1)}{20+15+8+6+2+1} = 1{,}19$$

El número medio de interrupciones semanales es de 1,19

Ejemplo 3.5: Los terminales de computadora tienen una pila que mantiene su configuración. Esta pila debe reponerse de vez en cuando. Se ha medido la vida útil en años de 164 estas pilas y se obtenido los siguientes resultados:

Duración (años)		X´	f
1,2	1,8		12
1,8	2,4		18
2,4	3,0		36
3,0	3,6		42
3,6	4,2		31
4,2	4,8		17
4,8	5,4		8

Solución

Primero de calcula la marca de clase, el cual es el valor representativo de cada intervalo

Duración (años)		X´	f
1,2	1,8	1,5	12
1,8	2,4	2,1	18
2,4	3,0	2,7	36
3,0	3,6	3,3	42
3,6	4,2	3,9	31
4,2	4,8	4,5	17
4,8	5,4	5,1	8

$$\overline{x}_w = \frac{1,5(12)+2,1(18)+2,7(36)+3,3(42)+3,9(31)+4,5(17)+5,1(8)}{12+18+36+42+31+17+8} = 3,2$$

El duración promedio de la vida útil de las pilas es de 3,2 años.

Mediana

La *mediana* es el valor del dato que ocupa la posición central una vez que todo el conjunto de valores se encuentra ordenado de forma ascendente. Con un número impar de datos siempre es posible encontrar un valor que ocupa la posición central. Cuando el número de datos es par existen dos valores ocupando las posiciones centrales.

Datos no agrupados

Como primer paso, los datos deben ser ordenados en orden creciente ó decreciente, luego se bebe determinar el valor que se ubica en la posición central. En caso de no coincidir el valor central con un dato, se tomará el promedio de los datos centrales. Si denotamos las observaciones ordenadas por $x_{[1]}$, $x_{[2]}$, $x_{[3]}$, ... , $x_{[n]}$, entonces:

- Para un número impar de datos, la *mediana* es el valor que ocupa la posición central.

$$\text{me} = x_{\left[\frac{n+1}{2}\right]}$$

Si *n* es impar, la mediana es el dato que ocupa la posición [(n+1)/2]

- Para un número par de datos, la *mediana* es el *promedio* de los dos valores que ocupan las posiciones centrales.

$$\text{me} = \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]} \right)$$

Si *n* es par, la mediana es el promedio de los datos que se encuentran en las posiciones [n/2] y [(n/2)+1]

Propiedades de la mediana

1. Divide a un conjunto de observaciones de dos partes iguales. El 50% con valores mayores a la mediana y el otro 50% con valores menores a la mediana.
2. La mediana no se ve afectada por valores extremos.

Ejemplo 3.6: Un centro meteorológico tiene registrado los porcentajes de humedad de una zona industrial observados en 12 días. Calcule la mediana de humedad de dicha zona basado en la información dada

92	88	95	76	80	82	89	92	93	95	83	82

Solución

Primero ordenar los datos

76	80	82	82	83	**88**	**89**	92	92	93	95	95

El total de observaciones es 12 (número par)

$$\text{me} = \frac{1}{2}\left(x_{\left\lceil \frac{12}{2} \right\rceil} + x_{\left\lceil \frac{12}{2}+1 \right\rceil} \right) = \frac{1}{2}\left(x_{[6]} + x_{[7]}\right) = \frac{1}{2}(88+89) = 88{,}5$$

El 50% de los porcentajes de humedad de una zona industrial es mayor a 88,5 y el otro 50% es menor 88,5.

Ejemplo 3.7: Si la muestra del ejemplo 3.7 hubiese sido la siguiente:

92	88	95	76	82	89	93	95	83

¿Cuál sería la mediana?

Solución

Primero se ordena los datos

76	82	83	88	**89**	92	93	95	95

El total de observaciones es 9 (número impar)

$$\text{me} = x_{\left\lceil \frac{9+1}{2} \right\rceil} = x_{[5]} = 89$$

El 50% de los porcentajes de humedad de una zona industrial es mayor a 89 y el otro 50% es menor 89.

Moda

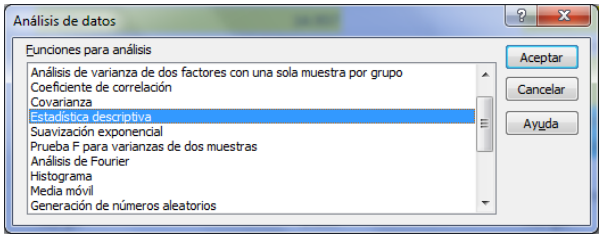
Una tercera medida de tendencia central es la *moda* y se define como el valor del dato que se repite o presenta con mayor frecuencia.

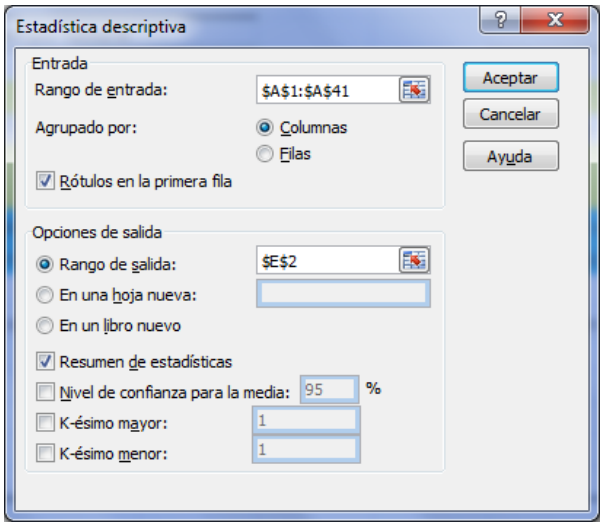
Propiedades de la moda

1. La moda puede no existir o puede existir más de una moda
2. No se ve afectada por valores extremos.
3. Se aplica tanto a información de variable cuantitativa como a cualitativa.
4. Es una medida inestable puesto que cambia de valor de una muestra a otra.

Ejemplo 3.8: Considerar el problema 5 del Capítulo 2. El jefe del área de sistemas de una empresa realizó un estudio para analizar el tiempo que demoran los empleados en detectar y resolver un problema informático de software o hardware. El estudio trata de estudiar las diferencias de tiempos en los turnos de trabajo A y B de una empresa. La hoja **Tiempo** muestra los datos obtenidos para una muestra de 40 problemas informáticos presentados en el turno A y 100 problemas informáticos presentados en el turno B.

El conjunto de valores para el turno A y B corresponde a un conjunto de datos no agrupado, en este caso la media mediana y moda se pueden obtener con Excel.
Utilizando la opción de **Análisis de datos** del menú **Datos** seleccionamos **Estadística descriptiva**.





Tiempo Turno A	
Media	**14.957**
Error típico	0.05819506
Mediana	**15.0025**
Moda	**14.872**
Desviación estándar	0.3680579
Varianza de la muestra	0.13546662
Curtosis	35.5173975
Coeficiente de asimetría	-5.74874508
Rango	2.695
Mínimo	12.75
Máximo	15.445
Suma	598.28
Cuenta	40

Tiempo Turno B	
Media	**14.69804**
Error típico	0.0065148
Mediana	**14.695**
Moda	**14.682**
Desviación estándar	0.06514799
Varianza de la muestra	0.00424426
Curtosis	-0.4749107
Coeficiente de asimetría	0.0909497
Rango	0.297
Mínimo	14.546
Máximo	14.843
Suma	1469.804
Cuenta	100

La media del tiempo de los empleados en detectar problemas informáticos de software ó hardware en el turno A es mayor que en el turno B

Ejemplo 3.9: Considerar el problema 2 del Capítulo 2. El tiempo de inactividad es la cantidad del tiempo en el que un sistema de computadora no opera debido a una falla de hardware y software. Cuando el sistema se encuentra inactivo durante más de una hora todos los archivos de trabajo actuales se pierden. Los datos agrupados en la hoja **Inactividad** corresponden a los tiempos de inactividad (en minutos) registrados en una muestra de 30 turno mañana.

i	Tiempo de inactividad		Marca de clase	Turno mañana			
				f	h	F	H
1	20,5	25,5	23	1	0,0333	1	0,0333
2	25,5	30,5	28	3	0,1000	4	0,1333
3	30,5	35,5	33	0	0,0000	4	0,1333
4	35,5	40,5	38	2	0,0667	6	0,2000
5	40,5	45,5	43	20	0,6667	26	0,8667
6	45,5	50,5	48	4	0,1333	30	1,0000
				30			

Media muestral: $\bar{x} = \frac{1}{30}\sum_{i=1}^{6} x_i^{/} f_i = \frac{1}{30}(1235) = 41{,}17$ minutos

Cuantiles

Constituyen una generalización del concepto de mediana. Así como la mediana divide a la serie de datos en dos partes con el mismo número de elementos cada una, si la división se hace en cuatro partes, o en diez partes, o en cien partes, llegamos al concepto de **cuantil**.

Cuartiles

Los cuartiles son los tres valores de la variable que dividen a un conjunto de datos ordenados en cuatro partes iguales.

Q_1: Primer cuartil, valor de la variable por debajo del cual queda 25% de los elementos de la serie de datos.
Q_2: Segundo cuartil (Mediana), valor de la variable por debajo del cual queda 50%
Q_3: Tercer cuartil, valor de la variable por debajo del cual quedan los 75% de los elementos que constituyen la serie de datos.

Deciles

Los deciles son los nueve valores que dividen la serie de datos en diez partes iguales. Los deciles dan los valores correspondientes al 10%, al 20%... y al 90% de los datos. D_5 coincide con la mediana

Percentiles

Los percentiles representan los valores de la variable que están por debajo de un porcentaje, el cual puede ser un valor de 1% a 100% (en otras palabras, el total de los datos es divido en 100 partes iguales).

La notación empleada será: P_k

Donde k es equivalente al porcentaje de datos acumulados y P_k es el valor de la variable que representa dicho porcentaje. Por ejemplo, P_8 es el valor de la variable que deja por debajo el 8% de los datos. P_{55} será entonces el valor que agrupa el 55% de los datos.

Ejemplo 3.10: Calcular el percentil 20 en el conjunto de datos no agrupados del Ejemplo 3.7. Un centro meteorológico tiene registrado los porcentajes de humedad de una zona industrial observados en 12 días.

92 88 95 76 80 82 89 92 93 95 83 82

P_{20} =81.2 haciendo uso del **EXCEL 2010 = PERCENTIL.EXC(rango,0.2)**
Aproximadamente el 20% de los porcentajes de humedad en la zona industrial fueron menores a 82 y el 80% fueron mayores a 81.2.

¿Cuál es el valor mínimo para estar considerado dentro del 30% superior de porcentajes de humedad de una zona industrial?

P_{70} = 92.1 haciendo uso del **EXCEL 2010 = PERCENTIL.EXC (rango,0.7)**
El valor mínimo para estar considerado dentro el 30% superior de porcentajes de humedad en la zona industrial es 92.1.

3.3 Medidas de dispersión

Rango

El *rango* o *amplitud* es la medida de dispersión más elemental. Se define como la diferencia entre el dato más grande y el más pequeño.

Rango poblacional: R = dato máximo – dato mínimo
Rango muestral: r = dato máximo – dato mínimo

Rango Intercuartílico (RIC)

Con frecuencia se dividen los datos en cuatro partes, cada una con aproximadamente el 25% de las observaciones. A los puntos de división se les denomina *cuartiles* y se definen como:

Primer cuartil	$Q_1 = P_{25}$
Segundo cuartil	$Q_2 = P_{50}$
Tercer cuartil	$Q_3 = P_{75}$

Una medida de dispersión que elimina la influencia de los valores extremos o atípicos de los datos es el *rango intercuartil* (**RIC**). Esta medida se define como la diferencia entre el tercer y primer cuartil. En otras palabras, el *rango intercuartil* mide la amplitud en el 50% central de los datos

$$\mathbf{RIC} = Q_3 - Q_1$$

Ejemplo 3.11: Calcule el RIC en el conjunto de datos agrupados del Ejemplo 3.7.

P_{25} =82.0	haciendo uso del	**EXCEL 2010**	**= PERCENTIL.EXC(rango,0.25)**
P_{75} =92.8	haciendo uso del	**EXCEL 2010**	**= PERCENTIL.EXC(rango,0.75)**

RIC = $Q_3 - Q_1 = P_{75} - P_{25}$ = 92,8 – 82,0 = 10,8

RIC = 10,8

Varianza

La *varianza* es una medida de dispersión que emplea todos los datos. Se basa en la diferencia entre cada dato y la media de los mismos, lo que se conoce como desviación respecto a la media. Para calcular la *varianza* las desviaciones respecto a la media se elevan al cuadrado.

Datos no agrupados

Varianza poblacional: $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

Varianza muestral: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Ejemplo 3.12: Calcule la varianza de los datos del ejemplo 3.7. Un centro meteorológico tiene registrado los porcentajes de humedad de una zona industrial observados en 12 días.

92 88 95 76 80 82 89 92 93 95 83 82

Haciendo uso del	**EXCEL 2010**	**= VAR.S(rango)**
41,2954	No tiene interpretación	

Desviación estándar

Las unidades al cuadrado asociadas con la varianza hacen difícil una interpretación de su valor numérico. La *desviación estándar* se define como la raíz cuadrada positiva de la varianza y, por consiguiente, tiene las unidades originales de los datos.

Desviación estándar poblacional: $\sigma = \sqrt{\sigma^2}$

Desviación estándar muestral: $s = \sqrt{s^2}$

Coeficiente de variación

El *coeficiente de variación* es una medida relativa de variabilidad por que evalúa la desviación estándar en relación con la media. Esta medida sirve para **comparar** dispersión en conjuntos de datos aun cuando las unidades fuesen diferentes.

Coeficiente de variación poblacional: $\text{CV} = \frac{\sigma}{\mu} \times 100\%$

Coeficiente de variación muestral: $\text{CV} = \frac{s}{\bar{x}} \times 100\%$

Coeficiente de variación	Grado de variabilidad
0% < CV < 5%	Datos muy homogéneos
5% ≤ CV < 10%	Datos homogéneos
10% ≤ CV < 15%	Datos regularmente homogéneos
15% ≤ CV < 20%	Datos regularmente heterogéneos
20% ≤ CV < 25%	Datos heterogéneos
CV ≥ 25%	Datos muy heterogéneos

Ejemplo 3.13: Considerar el problema 5 del Capítulo 2. El jefe del área de sistemas de una empresa realizó un estudio para analizar el tiempo que demoran los empleados en detectar y resolver un problema informático de software o hardware. El estudio trata de estudiar las diferencias de tiempos en los turnos de trabajo A y B de una empresa. La hoja **Tiempo** muestra los datos obtenidos para una muestra de 40 problemas informáticos presentados en el turno A y 100 problemas informáticos presentados en el turno B.

Las medidas de dispersión para el conjunto de datos no agrupados también pueden obtenerse con Excel, **a excepción del coeficiente de variación** que se obtiene a partir de los resultados previos

Tiempo Turno A	
Media	**14.957**
Error típico	0.05819506
Mediana	15.0025
Moda	14.872
Desviación estándar	**0.3680579**
Varianza de la muestra	0.13546662
Curtosis	35.5173975
Coeficiente de asimetría	-5.74874508
Rango	2.695
Mínimo	12.75
Máximo	15.445
Suma	598.28
Cuenta	40

Tiempo Turno B	
Media	**14.69804**
Error típico	0.0065148
Mediana	14.695
Moda	14.682
Desviación estándar	**0.06514799**
Varianza de la muestra	0.00424426
Curtosis	-0.4749107
Coeficiente de asimetría	0.0909497
Rango	0.297
Mínimo	14.546
Máximo	14.843
Suma	1469.804
Cuenta	100

<u>Turno A</u>

Rango muestral: r = 15,445 – 12,75 = 2,695 minutos
Varianza muestral: s^2 = 0,135466 minutos^2
Desviación estándar: s = 0,36805 minutos

El coeficiente de variación es CV = $\frac{s}{\bar{x}} = \frac{0{,}36805}{14{,}957} x100 = 2{,}460 |$

<u>Turno B</u>

Rango muestral: r = 14,843– 14,546= 0,297 minutos
Varianza muestral: s^2 = 0,004244 minutos^2
Desviación estándar: s = 0,06514 minutos

El coeficiente de variación es CV = $\frac{s}{\bar{x}} = \frac{0{,}06514}{14{,}698} x100 = 0{,}4431$

El tiempo del turno B es más homogéneo que el tiempo del turno A

3.4 Diagrama de Cajas

Un *diagrama de cajas* es un resumen gráfico de los datos con base en el resumen de cinco números: **el valor mínimo, el primer cuartil, la mediana, el tercer cuartil y el valor máximo**. Los pasos para elaborar un diagrama de cajas son los siguientes:

1. Se dibuja una caja ó rectángulo cuyos extremos se localicen en el primer y tercer cuartil. Esta caja contiene el 50% central de los datos.

2. En el punto donde se localiza la mediana se traza una línea.

3. Usando el rango intercuartil se localizan los *límites*. En un diagrama de cajas éstos se encuentran 1,5*RIC abajo del primer cuartil y 1,5*RIC arriba del tercer cuartil.

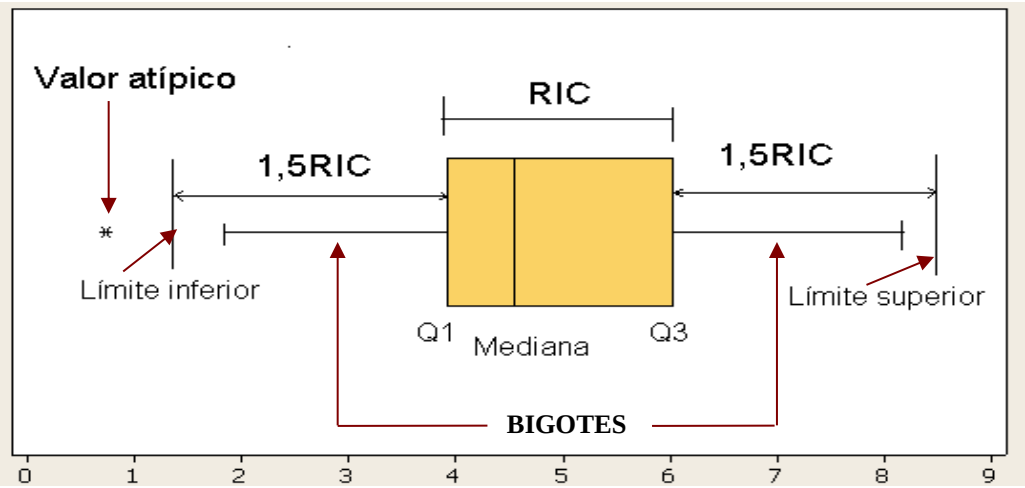
   Límite inferior: $Q_1 - 1{,}5$ RIC

   Límite superior: $Q_3 + 1{,}5$ RIC

4. Un dato se considera atípico si es mayor que el límite superior o si es menor al límite inferior.
   Se marcan con asterisco (*) las localizaciones de los datos atípicos.

5. Si no existen valores atípicos, los *bigotes* van desde los extremos de la caja hasta los valores menor y mayor dentro de los límites calculados en el paso anterior.

   Si existen valores atípicos, estos se ubican primero y luego los bigotes van desde los extremos de la caja hasta los valores menor y mayor (excluyendo los valores atípicos).



El diagrama de cajas permite determinar, en forma gráfica lo siguiente:

- Comparar las medianas de dos conjuntos de datos.
- Observar el tipo de distribución de los datos (simetría o asimetría).
- Determinar la dispersión del 50% central de los datos (tendencia central).
- Identificar la presencia de valores extremos (datos atípicos).

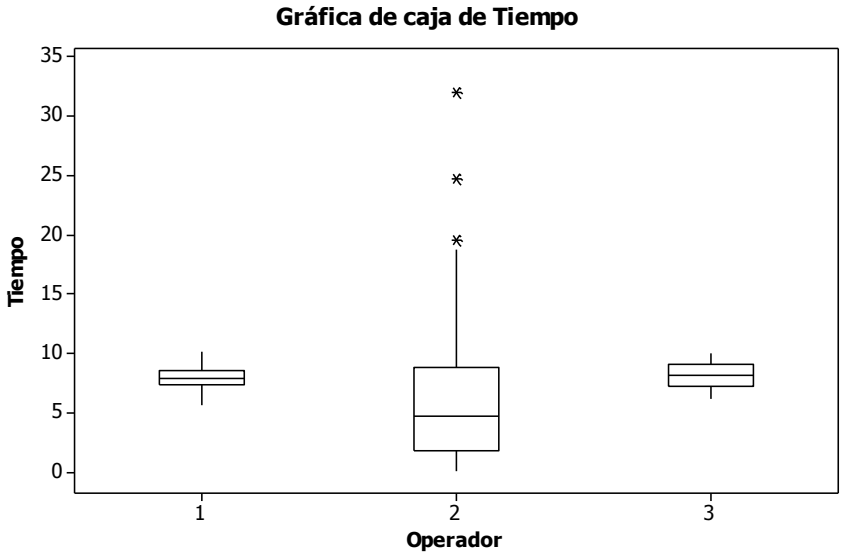
Ejemplo 3.14: El gráfico de cajas para conjunto de datos del problema 4 del capítulo 2 se muestra a continuación.

Los valores calculados para el gráfico fueron:

	Operador 1	Operador 2	Operador 3
Valor Mínimo	5,60	0,10	6,16
Percentil 25	7,34	1,81	7,16
Percentil 50	7,81	4,71	8,09
Percentil 75	8,55	8,65	9,02
Valor Máximo	10,05	31,91	9,91

	Operador 1	Operador 2	Operador 3
RIC	1,22	6,84	1,86
LI	5,515	-8,456	4,37
LS	10,375	18,913	11,81

Para el caso del operador 2, el bigote inferior se extiende hasta el valor mínimo mayor o igual al LI = -8,456 (para este caso el tiempo no puede ser negativo por lo que se puede considerar como LI= 0 minutos). El bigote superior se extiende hasta 18.73 (ordenar de mayor a menor la base de datos y observar este registro) que es el dato máximo menor o igual al LS = 18,913. Existe tres valores mayores a 18,913 estos son: 19,5 ; 24,7 ; 31,91. Estos tres datos con marcados como valores extremos con los asteriscos. Para los operadores 1 y 3 los datos se encuentran dentro del LI y LS.



Fuente: Elaboración propia

Interpretación

La caja correspondiente al operador 2, muestra tres valores extremos, correspondientes a tres registros del tiempo (minutos) utilizados en el restablecimiento de servicios de redes en el sistema de una importante agencia bancaria. Analizando los gráficos queda claro que el tiempo mediano del operador 2 es menor que los demás. También podemos observar claramente que el tiempo empleado por los operadores 1 y 3 son menos variables que el operador 2. Respecto a la asimetría en el operador 2 tiene ligera asimetría positiva (hacia la derecha si los gráficos fuesen presentados de manera horizontal).

Ejemplo 3.15: Una empresa que usa software de computadoras para proteger mejor la información electrónica. Los siguientes datos corresponden a los gastos mensuales (cientos de nuevos soles) en protección de la información. Realice un diagrama de cajas con esta información

187,9	108,6	133,6	141,6	150,8	166,3	179,8
196,4	110,5	134,2	145,7	154,0	168,7	179,9
270,8	112,9	135,3	145,9	158,0	168,8	180,0
198,5	113,2	136,3	148,5	158,4	169,9	181,5
38,4	187,0	136,6	149,5	158,6	170,8	183,0
105,4	123,8	188,0	149,5	159,1	172,9	183,3
105,4	124,8	139,2	188,2	163,7	174,5	184,9
107,1	127,5	139,2	150,2	189,6	178,0	186,6
36,8	128,2	139,4	150,3	164,1	190,0	186,7
108,6	132,8	140,1	150,4	165,3	179,7	194,4

Q1	
Q2	
Q3	
RIC	

3.5 Problemas

1. Los directivos de una empresa solicitaron a los gerentes de cada una de sus dos plantas, A y B, que informen sobre los sueldos de todos sus trabajadores correspondientes al mes de enero del año en curso. La información proporcionada (en nuevos soles) por cada una de las plantas se muestra a continuación.

Planta A	
Medida estadística	**Valor**
Media	402.50
Mediana	406.67
Moda	450.00
Varianza	10993.75

Planta B	
Medida estadística	**Valor**
Media	340.60
Mediana	330.60
Moda	290.10
Varianza	14700.84

a. ¿Se puede afirmar que el sueldo es más homogéneo en los trabajadores de la planta B?
Rpta. EL sueldo es más homogéneo en la planta A. CV_A = 26,0% y CV_B = 35,6%

b. ¿Los resultados anteriores son poblacionales o muéstrales?
Rpta. Los resultados anteriores son poblacionales.

2. La central telefónica de una empresa desea conocer el gasto mensual en que incurre debido a las llamadas que efectúan sus empleados luego de las 6 p.m., hora en que finalizan las labores. Para esto registra al azar y durante el mes de enero el tiempo de duración (en minutos) de las llamadas efectuadas entre las 6 p.m. y las 7 p.m. Los datos obtenidos, según el tipo de teléfono al que se llama, se encuentran en la hoja **Llamadas**.

a. ¿Cuál es la duración mínima de una llamada hacia un teléfono celular para considerarla dentro del grupo del 15% de las llamadas que le ocasionan mayor gasto a la empresa?
Rpta. 10 minutos

b. ¿Se puede afirmar que existe mayor homogeneidad en la duración de las llamadas realizadas hacia un teléfono celular en comparación a las realizadas hacia un teléfono fijo?
Rpta. No se puede afirmar que exista mayor homogeneidad en la duración de llamadas a teléfono celular.
CV_{fijo} = 22,5% y $CV_{celular}$ = 28,6%

c. Use el diagrama de cajas para detectar la presencia de valores atípicos en la duración de las llamadas realizadas hacia un teléfono celular.
Rpta. Existen 4 valores atípicos 1, 2, 12 y 12.

3. Un artículo es producido en tres etapas, cortado, armado y acabado. Para la etapa de cortado se requiere dos horas, cuatro para el armado y tres para el acabado; los costos por hora en cada etapa son de 2 , 4.5 y 6.5 dólares. ¿Cuál es el costo medio de producción por hora de dicho artículo?
Rpta. 4.6 dólares la hora.

4. Los datos en la hoja **Supermercados** corresponden a los montos y el tipo de pago (cheque, efectivo y tarjeta) realizados por una muestra de clientes de una cadena de supermercados. Los

directivos de esta empresa pidieron tomar la muestra para conocer el comportamiento de los pagos de sus clientes.

a. Construya en un mismo gráfico los diagramas de cajas para los montos según el tipo de pago indicando la presencia de datos atípicos en la muestra.
   Rpta. Cheque: Dos valores atípicos 5 y 95; Efectivo: No hay valores atípicos; Tarjeta: Un valor atípico 94,36

b. Según el gráfico elaborado mencione el tipo de simetría que presentan los montos según el tipo de pago.
   Rpta. Cheque: Asimétrica positiva; Efectivo: Asimétrica positiva; Tarjeta: Asimétrica Negativa

c. Obtenga las medidas de tendencia central y de dispersión para los montos pagados en cheque, efectivo y tarjeta e intérprete.

Cheque		***Efectivo***		***Tarjeta***	
Media	43.38	Media	8.84	Media	43.18
Error típico	2.63	Error típico	0.86	Error típico	4.00
Mediana	41.34	Mediana	7.41	Mediana	46.19
Moda	No existe	Moda	No existe	Moda	No existe
Desv.est.	16.62	Desv.est.	5.28	Desv.est.	18.75
Varianza	276.06	Varianza	27.88	Varianza	351.50
CV =	**38.3%**		**59.7%**		**43.4%**

Capítulo 4

Probabilidades

4.1 Introducción

A menudo las decisiones de negocios se basan en algún análisis del siguiente tipo:

- ¿Cuál es la *probabilidad* que las ventas disminuyan si se aumentan los precios?
- ¿Qué tan *factible* es que un nuevo método de ensamblado aumente la productividad?
- ¿Qué tan *probable* es que el proyecto se complete a tiempo?
- ¿Cuáles son las *posibilidades* a favor que una nueva inversión sea rentable?

La *probabilidad* es una medida de la posibilidad que tiene un evento de ocurrir. Los valores de probabilidad siempre se asignan en una escala de cero a uno. Una probabilidad cercana a cero indica que el evento tiene muy pocas posibilidades de ocurrir, mientras que una probabilidad cercana a uno indica que es casi seguro que ocurra dicho evento. La probabilidad proporciona la posibilidad de medir, expresar y analizar la incertidumbre asociada a eventos futuros.

4.2 Experimentos y espacio muestral

Dentro del punto de vista de la probabilidad se define un *experimento* como cualquier proceso que genera un conjunto de resultados bien definidos. En cualquier ejecución del experimento ocurrirá *solamente uno* de los posibles resultados experimentales.

4.3 Algunas definiciones importantes:

Experimento Aleatorio (ε)

Es una operación o acto cuyo resultado no se puede predecir con certeza y que se realiza bajo los siguientes criterios:

- Puede ser repetido bajo las mismas condiciones
- Se puede describir el número de resultados posibles.
- Se puede establecer un modelo matemático asociado a $\boldsymbol{\varepsilon}$ (esto se discutirá en el capítulo de variables aleatorias).

Ejemplo:

ε_1 : {Extraer una esfera de una urna que contiene 5 esferas de color azul y 2 rojas}

ε_2 : {Contar el número de piezas defectuosas producidas por una máquina de la producción diaria}

ε_3 : {Tiempo de vida de un foco de luz}

Espacio Muestral (Ω)

Es el conjunto de todos los resultados posibles de un experimento aleatorio E.
Ejemplo

$\Omega_1 = \{\text{azul, rojo}\}$

$\Omega_2 = \{x / x \geq 0, x \in Z\}$

$\Omega_3 = \{x / x \geq 0, x \in R\}$

Punto muestral

Es cualquier elemento de Ω. Si $\Omega_1 = \{o_1, o_2, \ldots, o_n\}$ entonces Ω tienen n puntos muestrales.
Ejemplo:

$\Omega_1 = \{azul, rojo\}$ los puntos muestrales son $o_1 = \{azul\}$, $o_2 = \{rojo\}$

Evento

Es cualquier subconjunto de un espacio muestral. Los eventos se identifican mediante letras mayúsculas.

Ejemplo:

Si se define el experimento aleatorio **ε**={Resultado de un partido de fútbol}, su espacio muestral será {ganar, empatar, perder} ; si se define el siguiente evento:
A: {Obtener al menos un punto como resultado}
A: {ganar, empatar}

Ejemplo 4.1: A continuación se presentan algunos ejemplos de experimentos y sus resultados asociados.

Experimento	Resultados experimentales
Lanzar una moneda	Cara, Sello
Seleccionar un artículo para su inspección	Defectuoso, No defectuoso
Lanzar un dado	1 , 2 , 3 , 4 , 5 , 6
Jugar un partido de fútbol	Ganar, perder, empatar

Para analizar un experimento en particular es necesario definir cuidadosamente los resultados experimentales posibles. El *espacio muestral* es el conjunto de todos los resultados posibles generados por un experimento. Cada resultado experimental se conoce también como *punto muestral* y corresponde a un elemento del espacio muestral.

Ejemplo 4.2: Si suponemos que Ω denota el espacio muestral entonces:

Experimento	Espacio muestral
Lanzar una moneda	Ω = { Cara, Sello }
Seleccionar un artículo para su inspección	Ω = { Defectuoso, No defectuoso }
Lanzar un dado	Ω = { 1 , 2 , 3 , 4 , 5 , 6 }
Jugar un partido de fútbol	Ω = { Ganar, perder, empatar }

Ejemplo 4.3: Un chip que pasa el control de calidad puede ser defectuoso (D) o no defectuoso (D´) Suponga que en cierto control se tiene tres chips, construya un espacio muestral de los posibles resultados de la condición de los tres chips.

4.4 Reglas de conteo

Un paso necesario en la asignación de probabilidades es poder identificar y contar los resultados experimentales. A continuación se analizan cuatro *reglas de conteo* que resultan útiles.

Regla de la multiplicación

Si un experimento se puede describir como una sucesión de k etapas, en las que hay n_1 resultados posibles en la primera etapa, n_2 en la segunda, etc. Entonces, la cantidad total de resultados experimentales es $n_1 \times n_2 \times \cdots \times n_k$.

Ejemplo 4.4: ¿Cuántas parejas de bailes de diferente sexo se puede formar con tres varones y cinco damas?

El experimento consiste en formar parejas de baile. La primera etapa consiste en elegir al varón y se puede hacer de $n_1 = 3$ formas. La segunda etapa consiste en escoger a la dama y se puede realizar de $n_2 = 5$ formas. Luego, se pueden formar $n_1 \times n_2 = 15$ parejas de baile.

Ejemplo 4.5: Para viajar de Tumbes a Lima se cuenta con tres posibilidades, por vía área, terrestre o marítima. De Lima a Tacna se puede viajar solo por vía aérea y marítima, por averías en la vía terrestre. ¿De cuántas maneras se puede realizar el viaje de Tumbes a Tacna pasando por Lima?

Regla de la adición

Si un primer experimento tiene n_1 resultados posibles, un segundo experimento tiene n_2 resultados posibles, etc. Entonces, la cantidad total de resultados experimentales luego de realizar el primer o segundo experimento, es decir solo uno de ellos, es $n_1 + n_2 + \ldots +$ n_k.

Ejemplo 4.6: Una persona puede viajar de Lima a Cuzco por vía aérea o terrestre y tiene a su disposición 6 líneas aéreas y 5 líneas terrestres. ¿De cuantas formas puede realizar el viaje desde Lima hasta Cuzco?

El primer experimento consiste en elegir una línea aérea y tiene $n_1 = 6$ formas posibles de hacerlo, el segundo experimento consiste en elegir una línea terrestre y tiene $n_1 = 5$ formas posibles de hacerlo. Luego, se tienen $n_1 + n_2 = 11$ formas posibles de realizar el viaje.

Ejemplo 4.7: Doña Pepa compró un pescado para cocinarlo. En su manual de cocina ella encuentra tres maneras diferentes para hacerlo al horno, dos para hacerlo frito y cuatro para prepararlo cocido. ¿De cuantas maneras diferentes puede cocinar su pescado?

Combinaciones

Las *combinaciones* permiten contar la cantidad de resultados experimentales cuando se deben seleccionar n objetos entre un total de N, por lo general más grande, donde el orden de selección *no se considera importante*. La cantidad de combinaciones de N objetos tomando n a la vez es

$$\mathrm{C}_n^N = \frac{N!}{n!(N-n)!}$$

donde $N! = N \times (N-1) \times \cdots \times 2 \times 1$, $n! = n \times (n-1) \times \cdots \times 2 \times 1$ y $0! = 1$.

Ejemplo 4.8: Se tiene un grupo de 30 estudiantes del curso Estadística EPE, ¿de cuantas maneras diferentes se podría elegir a tres delegados para que representen a la sección?

El orden de selección de los tres delegados no es importante.

Luego, existen $\mathrm{C}_3^{30} = 4060$ formas diferentes de elegirlos.

Ejemplo 4.9: De un grupo de nueve alumnos, cinco de los cuales son varones, se desea formar un grupo que consta de tres varones y dos mujeres. ¿De cuantas maneras se puede formar dicho grupo?

El orden de selección de los varones y mujeres no es importante.

Luego, existen $C_3^5 x C_2^4 = 60$ formas diferentes de elegirlos.

Permutaciones

Las *permutaciones* permiten calcular el número de resultados experimentales al seleccionar n objetos de un conjunto de N, donde es *importante el orden de selección*. Si los mismos n objetos se seleccionan en otro orden se considera que se trata de un resultado experimental distinto. El número de permutaciones de N objetos tomando n a la vez es

$$\mathrm{P}_n^N = \frac{N!}{(N-n)!}$$

Ejemplo 4.10: Se contrata un servicio de calificación para que analice y seleccione en orden las 3 mejores marcas de computadora. Si se analizan un total de 12 marcas de computadoras ¿de cuantas formas distintas puede el servicio de calificación llegar al ordenamiento final?

Para elegir las tres mejores marcas de computadora es obvio que el orden es importante. Luego, existen $\mathrm{P}_3^{12} = 1320$ formas diferentes de llegar al ordenamiento final.

Ejemplo 4.11: Un estudiante tiene ocho CDS distintos y desea acomodarlos de tal manera que tres CDS específicos siempre estén juntos.

a. ¿Cuántos arreglos diferentes son posibles?

b. ¿Cuántos arreglos diferentes serían posibles si estos tres CDS específicos nunca estén juntos?

4.5 Asignación de probabilidades

A partir de la comprensión de los conceptos de experimento y espacio muestral, veamos cómo pueden determinarse las probabilidades para los resultados experimentales. Al asignarse estas probabilidades se deben satisfacer dos requerimientos básicos de probabilidad.

1. Los valores de probabilidad asignados deben estar entre cero y uno. Si ε_i denota el i-ésimo resultado experimental y $\mathrm{P}(\varepsilon_i)$ indica su probabilidad asociada, entonces:

$$0 \leq \mathrm{P}(\varepsilon_i) \leq 1$$

2. La suma de *todas* las probabilidades asociadas a los resultados experimentales debe ser igual a uno. Si el espacio muestral tiene k resultados experimentales, entonces:

$$\sum_{i=1}^{k} \mathrm{P}(\varepsilon_i) = \mathrm{P}(\varepsilon_1) + \mathrm{P}(\varepsilon_2) + \cdots + \mathrm{P}(\varepsilon_k) = 1$$

Cualquier método para asignar valores de probabilidad a los resultados experimentales que satisfaga estos dos requerimientos y produzca medidas numéricas razonables es aceptable. En la práctica, se emplean con frecuencia el método *clásico*, el método de *frecuencia relativa* y el método *subjetivo*.

Método clásico

El método clásico de asignación de probabilidades supone que cada resultado experimental tiene la *misma posibilidad de ocurrir*. Si un experimento tiene n resultados posibles el método clásico asignaría una probabilidad de $1/n$ a cada resultado experimental.

Ejemplo 4.12: Si se considera el experimento de lanzar una moneda, en cualquiera de estos lanzamientos se observará Cara o Sello. Una suposición razonable, en caso la moneda no esté trucada, es que cualquiera de estos resultados experimentales es igualmente probable, es decir $\mathrm{P}[\text{Cara}] = \mathrm{P}[\text{Sello}] = 0,5$.

Ejemplo 4.13: Considere el experimento de lanzar un dado. Si el dado fue diseñado de modo que los seis resultados experimentales sean igualmente probables entonces se cumple que $\mathrm{P}[1] = \mathrm{P}[2] = \cdots = \mathrm{P}[6] = 1/6$.

Este método fue elaborado originalmente para analizar probabilidades en los juegos de azar, donde la suposición de resultados igualmente probables frecuentemente es razonable. Sin embargo en muchas otras situaciones esta suposición no es válida por lo que se requieren métodos alternativos para asignar las probabilidades.

La definición de probabilidad bajo el método clásico es:

$$P(A) = \frac{n(A)}{n(S)}$$

Dónde:
n(A): Es el número de elementos de un evento de interés.
n(S): Es el número total de elementos del espacio muestral.

Ejemplo 4.14: Se selecciona al azar dos chips de un grupo de 12 de los cuales cuatro son defectuosos. Y se definen los eventos:
A: Los chips seleccionados son defectuosos
B: Los chips seleccionados son no defectuosos
Determine la probabilidad de ocurrencia de A y la probabilidad de ocurrencia de B

Solución

$$P(A) = \frac{C_2^4 x C_0^8}{C_2^{12}} = 0{,}0909$$

$$P(B) = \frac{C_2^8 x C_0^4}{C_2^{12}} = 0{,}4242$$

Método de frecuencia relativa

El método de frecuencia relativa está basado en los datos observados sobre el proceso o fenómeno bajo estudio. Para su determinación se necesita obtener datos mediante pruebas, entrevistas, etc. los cuales se resumen en una tabla de resultados (la tabla de distribución de frecuencias). La frecuencia relativa de un resultado (cantidad de resultados favorables entre total de resultados) se considera como la probabilidad de ocurrencia, siempre y cuando las condiciones bajo las cuales se obtuvieron los resultados se mantengan a futuro.

Método subjetivo

Este método es apropiado cuando no se puede suponer de manera realista que todos los resultados experimentales son igualmente probables y cuando se dispone de pocos datos relevantes. Cuando se emplea este método podemos usar cualquier información disponible como nuestra experiencia o intuición. Después de considerar toda esa información se especifica un valor de probabilidad que exprese nuestro grado de creencia, en una escala de cero a uno, sobre las posibilidades de ocurrir que tiene el resultado experimental de interés. Usando el método subjetivo puede esperarse que diferentes personas asignen probabilidades diferentes al mismo resultado experimental.

4.6 Eventos y sus probabilidades

Un *evento* es una colección específica de puntos muestrales.

Ejemplo 4.15: Considere el experimento de lanzar un dado, cuyo espacio muestral es Ω = { 1 , 2 , 3 , 4 , 5 , 6 }. Si se define el evento A como obtener un número par, entonces A = { 2 , 4 , 6 }. Por tanto, si el resultado experimental fuese 2, 4 ó 6 se diría que ha ocurrido el evento A. Si se conocen las probabilidades de los puntos muestrales, la *probabilidad de un evento* es igual a la suma de las probabilidades de los puntos muestrales que la componen. Es decir que

$$P[A] = P[2] + P[4] + P[6] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0,5$$

Siempre que podamos identificar todos los puntos muestrales de un experimento y asignarles las correspondientes probabilidades podemos usar el enfoque anterior para calcular la probabilidad de un evento cualquiera. Sin embargo, en muchos experimentos la cantidad de puntos muestrales es grande y su identificación, al igual que la determinación de sus probabilidades, se vuelve demasiado compleja o, incluso, imposible. A continuación se presentan algunas relaciones básicas de probabilidad para calcular la probabilidad de un evento sin conocer todas las probabilidades individuales de los puntos muestrales que la componen. Este enfoque se basa en la teoría de conjuntos, por lo que se recomienda un breve repaso de la misma.

4.7 Algunas relaciones básicas de probabilidad

Complemento de un evento

Para un evento A cualquiera, se define su *complemento* como el evento consistente en todos los puntos muestrales que *no están* en A. El complemento del evento A se denota con A^C. La Figura 4.1 muestra una representación conocida como *Diagrama de Venn*, que ilustra el concepto del complemento. El área rectangular representa el espacio muestral para el experimento y como tal contiene todos los puntos muestrales posibles. El círculo representa el evento A y solo contiene los puntos muestrales que le pertenecen. El resto del rectángulo sombreado contienen todos los puntos muestrales que no están en el evento A, el cual por definición es el complemento de A.

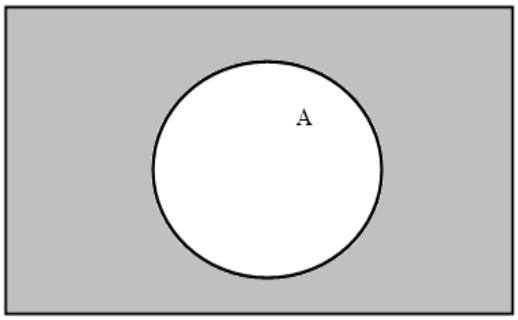


Figura 4.1

En cualquier aplicación de probabilidad, el evento A y su complemento A^C deben satisfacer la condición

$$P[A]+P[A^C]=1$$

Despejando se tiene

$$P[A]=1-P[A^C]$$

La relación anterior muestra que la probabilidad de un evento A puede calcularse si se conoce la probabilidad de su complemento.

Ejemplo 4.16: La probabilidad que un sistema de comunicación de datos tenga una selectividad elevada es 0,72, la probabilidad que tenga alta fidelidad es 0,59 y la probabilidad que tenga ambas características es 0,37. Si se definen el evento:

A = Sistema de comunicación de datos tiene una selectividad elevada
A^C = Sistema de comunicación de datos no tiene una selectividad elevada

Calcular la probabilidad que el sistema de datos no tiene selectividad elevada

$$P(A^C)=1-P(A)=1-0{,}72=0{,}28$$

Ejemplo 4.17: La probabilidad de una moneda está cargada de modo que la probabilidad de salir cara (C) sea el doble que la de sello (S). Calcule P(C) y P(S)

C = La moneda sale cara
S = La moneda sale sello

De los datos tenemos: $P(C)=2P(S)$

Por definición: $P(C)+P(S)=1$
Reemplazamos:

$$2P(S)+P(S)=1$$
$$3P(S)=1$$
$$P(S)=1/3 \Rightarrow P(C)=2/3$$

Ley de la adición

La *ley de la adición* es una relación útil cuando tenemos dos eventos y estamos interesados en conocer la probabilidad que ocurra al menos uno de ellos. Es decir, si tenemos los eventos A o B, estamos interesados en conocer la probabilidad que ocurra el evento A, el evento B o ambos simultáneamente. Antes de presentar la ley de la adición se necesita exponer dos conceptos relacionados a la combinación de eventos: la *unión* y la *intersección* de eventos.

Para dos eventos A y B, la *unión del evento A con el evento B* es el evento que contienen todos los puntos muestrales que pertenecen a *A o a B o a ambos*. La unión se

denota por $A \cup B$. El diagrama de Venn que se muestra en la Figura 4.2 describe la unión de los eventos A y B.

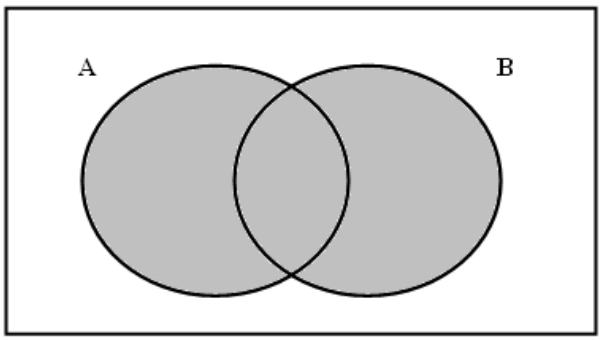


Figura 4.2

Para dos eventos A y B, la *intersección de los eventos A y B* es el evento que contienen todos los puntos muestrales que pertenecen *tanto a A como a B*, es decir, que son comunes a ambos conjuntos. La intersección se denota por $A \cap B$. El diagrama de Venn que se muestra en la Figura 4.3 describe la intersección de los eventos A y B.

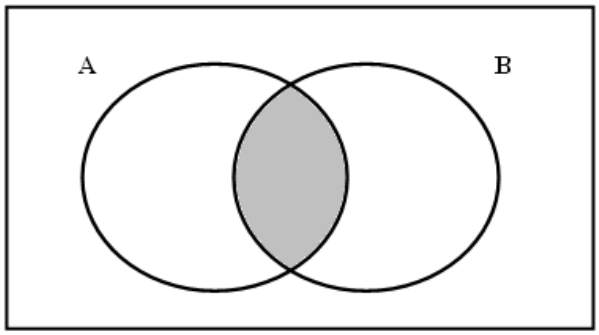


Figura 4.3

La ley de la adición permite encontrar la probabilidad de la unión de dos eventos usando la siguiente relación:

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

Ejemplo 4.18: Un instituto de investigaciones académicas desea determinar si existe relación entre el interés de un estudiante en finanzas, su habilidad en matemáticas y su género. Se selecciona una muestra de 200 estudiantes y mediante una prueba se mide su habilidad matemática y su interés en finanzas. Los resultados fueron los siguientes:

Interés en finanzas	Género						Total
	Femenino (A_1)			Masculino (A_2)			
	Habilidad en matemáticas			Habilidad en matemáticas			
	Baja (C_1)	Media (C_2)	Alta (C_3)	Baja (C_1)	Media (C_2)	Alta (C_3)	
Bajo (B_1)	25	9	6	35	6	10	91
Medio (B_2)	4	22	7	11	21	3	68
Alto (B_3)	2	5	15	3	7	9	41
Total	31	36	28	49	34	22	200

Si se selecciona un estudiante al azar:

a. ¿Cuál es la probabilidad de seleccionar un estudiante con un interés alto en finanzas?

Solución:

B_3: Interés alto en finanzas.

$$P(B_3) = \frac{41}{200} = 0{,}205$$

b. ¿Cuál es la probabilidad de que no tenga una alta habilidad en matemáticas y tenga un interés medio en finanzas?

Solución:

C_3: Habilidad alta en matemáticas.

B_2: Interés medio en finanzas.

$$P(C_3^C \cap B_2) = \frac{4+22+11+21}{200} = 0{,}29$$

c. ¿Cuál es la probabilidad de que sea hombre o tenga un bajo interés en finanzas?

Solución:

A_2: Masculino.

B_1: Interés bajo en finanzas.

$$P(A_2 \cup B_1) = P(A_2) + P(B_1) - P(A_2 \cap B_1)$$

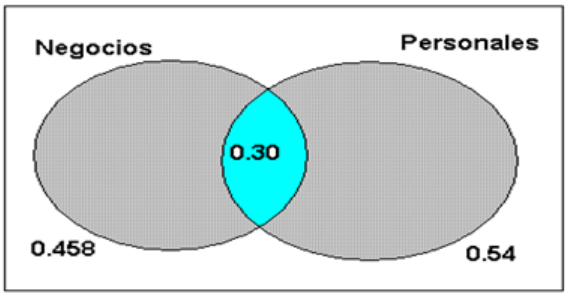
$$P(A_2 \cup B_1) = \frac{49+34+22}{200} + \frac{91}{200} - \frac{51}{200} = 0{,}725$$

Ejemplo 4.19: Una encuesta entre suscriptores de una revista local indicó que 45,8% de ellos habían rentado un automóvil por motivos de negocios, 54% por motivos personales y 30% por motivos de negocios y personales a la vez.

a. ¿Cuál es la probabilidad de que un suscriptor rente un automóvil por motivos de negocios o personales?

b. ¿Cuál es la probabilidad de que un suscriptor rente un automóvil por motivos que no sean de negocios ni personales?

Solución:



A: Rentar un automóvil por negocios

B: Rentar un automóvil por motivos personales

a. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0{,}458 + 0{,}54 - 0{,}30 = 0{,}698$

$P(A^C \cap B^C) = 1 - P(A \cup B) = 1 - 0{,}698 = 0{,}302$

Se dice que dos o más eventos son *mutuamente excluyentes* si los eventos no tienen ningún punto muestral en común, es decir, no hay puntos muestrales en la intersección

de los eventos. Para que dos eventos A y B sean mutuamente excluyentes se debe cumplir que $\Pr[A \cap B] = 0$. La Figura 4.4 proporciona un diagrama de Venn que muestra dos eventos mutuamente excluyentes.

Si los eventos A y B son mutuamente excluyentes, entonces

$$P[A \cup B] = P[A] + P[B]$$

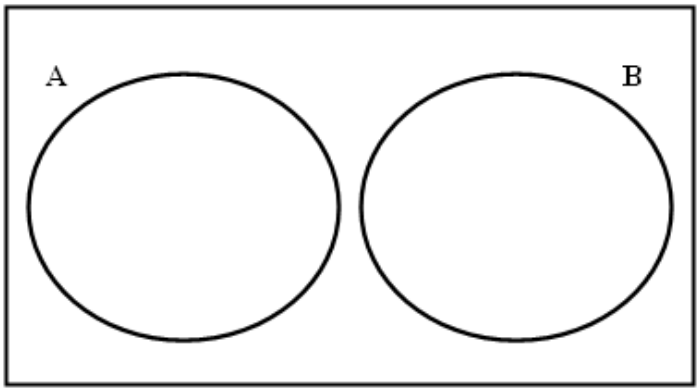


Figura 4.4

Ejemplo 4.20: La probabilidad de que una empresa alemana invierta en minería en el departamento de Junín es de 0,7; de que invierta en Cajamarca es de 0,4 y de que invierta en al menos una de ellas es de 0,8. Determine la probabilidad de que dicha empresa finalmente se localice:

a) Sólo en Cajamarca.
b) En ninguno de los lugares mencionados

Probabilidad condicional

En muchas situaciones es importante poder determinar la probabilidad de un evento cuando se sabe que ha ocurrido otro. Suponga que tenemos un evento A con probabilidad $\Pr[A]$ y que obtenemos información nueva o nos enteramos que ha ocurrido otro, denotado B. Si el evento A se relaciona con B, desearemos sacar ventaja de esta información al calcular una probabilidad nueva o revisada para el evento A.

Esta nueva probabilidad del evento A se escribe $P[A/B]$. El símbolo "/" denota el hecho que estamos considerando la probabilidad del evento A *dada la condición que ha ocurrido el evento B*. Por lo tanto, la notación $P[A/B]$ se lee "*la probabilidad de A dado B*".

Con dos eventos A y B, la probabilidad *condicional para A dado B*, y la probabilidad *condicional para B dado A* son como siguen:

$$P[A/B] = \frac{P[A \cap B]}{P[B]}$$

$$P[B/A] = \frac{P[A \cap B]}{P[A]}$$

Para que las expresiones anteriores tengan significado, en el primer caso $P[B]$ no puede ser igual a cero y en el segundo caso $P[A]$ tampoco puede ser igual a cero.

Para obtener una comprensión intuitiva del uso de las relaciones anteriores consideremos el diagrama de Venn de la figura 4.5. La región sombreada (tanto en gris claro como en gris oscuro) denota que el evento B ocurrió; la región sombreada en gris oscuro denota el evento $A \cap B$. Sabemos que una vez que ha ocurrido B, la única forma en que podemos observar también el evento A es que ocurra el evento $A \cap B$. Por lo tanto la razón entre $P[A \cap B]$ y $P[B]$ proporciona la probabilidad que observemos el evento A cuando ya ocurrió el evento B.

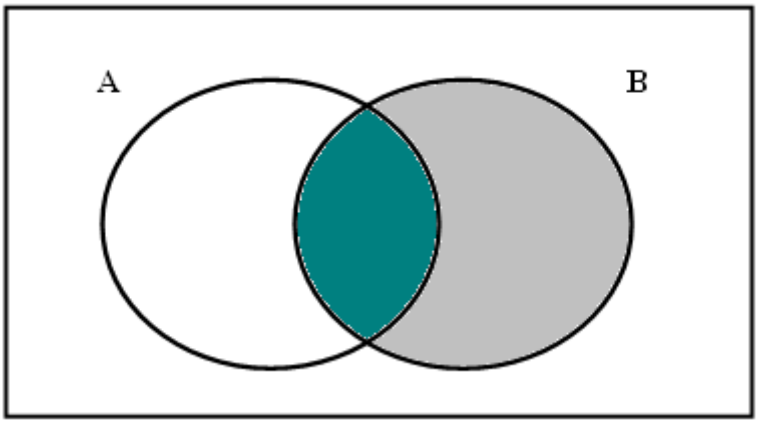


Figura 4.5

Ejemplo 4.21: Se estima que la probabilidad que la empresa A tenga éxito al comercializar su nuevo producto en el mercado es 0,75, la probabilidad que la empresa B tenga éxito al comercializar su nuevo producto en el mercado es 0,62 y la probabilidad que ambas empresas tengan éxito con sus respectivos productos es 0,45. ¿Cuál es la probabilidad que la empresa A tenga éxito al comercializar su nuevo producto, si se sabe que la empresa B también lo tuvo?

Si se definen los eventos:

A = La empresa A tiene éxito al comercializar su producto
B = La empresa B tiene éxito al comercializar su producto

Entonces $P[A/B] = \frac{P[A \cap B]}{P[B]} = \frac{0,45}{0,62} = 0,7258$.

Ejemplo 4.22: La probabilidad de que un vuelo programado normalmente salga a tiempo es 0,83; la probabilidad de que llega a tiempo es 0,82 y la probabilidad de que salga y llegue a tiempo es 0,78. Encuentre la probabilidad de que un avión:

a. Llegue a tiempo, dado que salió a tiempo

A: Vuelo programado normalmente salió a tiempo
B: Vuelo llegue a tiempo

P(A) = 0,83
P(B) = 0,82
$P(A \cap B) = 0,78$

$$P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{0,78}{0,83} = 0,9398$$

b. Salió a tiempo, dado que llegó a tiempo

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0,78}{0,82} = 0,9512$$

Eventos independientes

Dos eventos A y B son *independientes* si:

$$P[A/B] = P[A] \quad \text{ó} \quad P[B/A] = P[B]$$

De otro modo, los eventos son *dependientes*.

Ley de la multiplicación

La *ley de la multiplicación* puede usarse para encontrar la probabilidad de la intersección de dos eventos. La ley de la multiplicación se obtiene a partir de la definición de probabilidad condicional

$$P[A \cap B] = P[A/B]P[B]$$

$$P[A \cap B] = P[B/A]P[A]$$

Ejemplo 4.23: El departamento de circulación de un periódico sabe que el 84% de sus clientes se suscribe a la edición diaria (lunes a sábado). Además la probabilidad que un cliente, que ya tiene suscripción diaria, se suscriba también a la edición dominical es 0.75. ¿Cuál es la probabilidad que un cliente se suscriba tanto a la edición diaria como a la dominical del periódico?

Si se definen los eventos:

A = Un cliente se suscribe la edición diaria
B = Un cliente se suscribe la edición dominical

Entonces $P[A] = 0,84$, $P[B/A] = 0,75$. Luego

$$P[A \cap B] = P[B/A]P[A] = 0,75(0,84) = 0,63$$

Notar que para el *caso especial de eventos independientes*, la ley de la multiplicación se convierte en:

$$P[A \cap B] = P[A]P[B]$$

Por lo tanto, para calcular la probabilidad de la intersección de dos eventos independientes solo multiplicamos las probabilidades correspondientes.

Ejemplo 4.24: El éxito de un proyecto de inversión depende del trabajo de un ingeniero, un administrador y un abogado. Se sabe que la probabilidad de que el ingeniero falle en su labor es de 4%, la probabilidad de que el administrador falle es de 6% y la probabilidad de que el abogado falle es de 8%. Para que el proyecto sea exitoso, ninguno de los 3 debe fallar. Asumiendo que las labores de los tres integrantes son independientes entre sí, ¿cuál es la probabilidad de que al final el proyecto falle?

Sean los eventos

A = El ingeniero falle en su labor
B = El administrador falle en su labor
C = El abogado falle en su labor
D = El proyecto falle

$$P[D] = 1 - P[D^C] = 1 - P[A^C \cap B^C \cap C^C] = 1 - P[A^C]P[B^C]P[C^C]$$

Luego $P[D] = 1 - 0,96 \times 0,94 \times 0,92 = 0,8302$.

Ejemplo 4.25: En una caja de fusibles contiene 20 unidades, de las cuales cinco son defectuosos. Si se selecciona dos fusibles al azar uno después de otro de la caja sin reemplazar el primero, ¿Cuál es la probabilidad de que ambos fusibles estén defectuosos?

4.8 Teorema de Bayes

Una etapa importante del análisis de las probabilidades condicionales consiste en revisar dichos valores cuando se obtiene nueva información. Con esta nueva información actualizamos los valores iniciales de probabilidad encontrando probabilidades revisadas, conocidas como *probabilidades posteriores*. El teorema de Bayes proporciona un medio para hacer estos cálculos de probabilidad.

Ejemplo 4.26: Suponga una empresa que compra componentes de dos proveedores. Sean los eventos:

A_1 = Componente comprado al proveedor 1

A_2 = Componente comprado al proveedor 2

Actualmente, el 65% de los componentes comprados provienen del proveedor 1 y el resto del proveedor 2. Por lo tanto, si se elige al azar un componente, asignaríamos las probabilidades previas $P[A_1] = 0,65$ y $P[A_2] = 0,35$.

La calidad de los componentes comprados depende del proveedor. Sean los eventos:

B_1 = Componente comprado se encuentra en buenas condiciones

B_2 = Componente comprado se encuentra en malas condiciones

Las probabilidades condicionales de recibir un componente en buenas o malas condiciones (con base en datos históricos), según el proveedor, son:

$$P[B_1/A_1] = 0,98 \qquad P[B_2/A_1] = 0,02$$

$$P[B_1/A_2] = 0,95 \qquad P[B_2/A_2] = 0,05$$

El diagrama de árbol que se muestra en la Figura 4.7 describe el proceso en que esta empresa recibe un componente de unos de los proveedores mencionados y luego descubre que se encuentra en buenas o malas condiciones como un experimento de dos etapas. De los cuatro resultados experimentales posibles, dos corresponden al caso en que el componente se encuentra en buenas condiciones y dos corresponden al caso contrario.

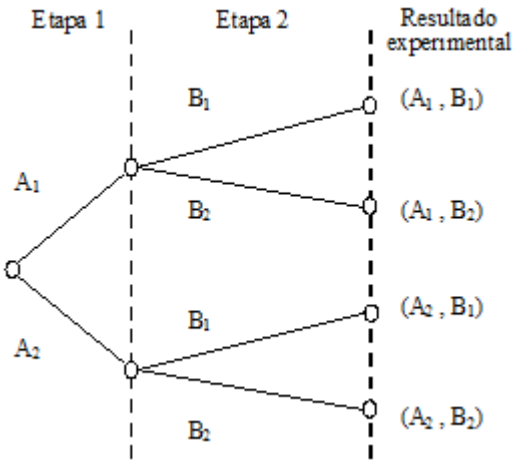


Figura 4.6

Cada uno de los resultados experimentales es la intersección de dos eventos, de modo que podemos usar la regla de la multiplicación para calcular las probabilidades. Es decir,

$$P[A_1 \cap B_1] = P[A_1]P[B_1/A_1]$$

El proceso de calcular estas probabilidades conjuntas puede describirse con lo que en ocasiones se llama *árbol de probabilidad*. También se le denomina *árbol de decisiones*, como se muestra en la Figura 4.8. De izquierda a derecha en el árbol, las probabilidades para cada una de las ramas en la primera etapa corresponden a las probabilidades iniciales, mientras que las probabilidades para cada rama en la segunda etapa son probabilidades condicionales. Para encontrar las probabilidades de cada resultado solo multiplicamos las probabilidades en las ramas que conducen al resultado. Observe que las probabilidades de los cuatro resultados experimentales suman uno.

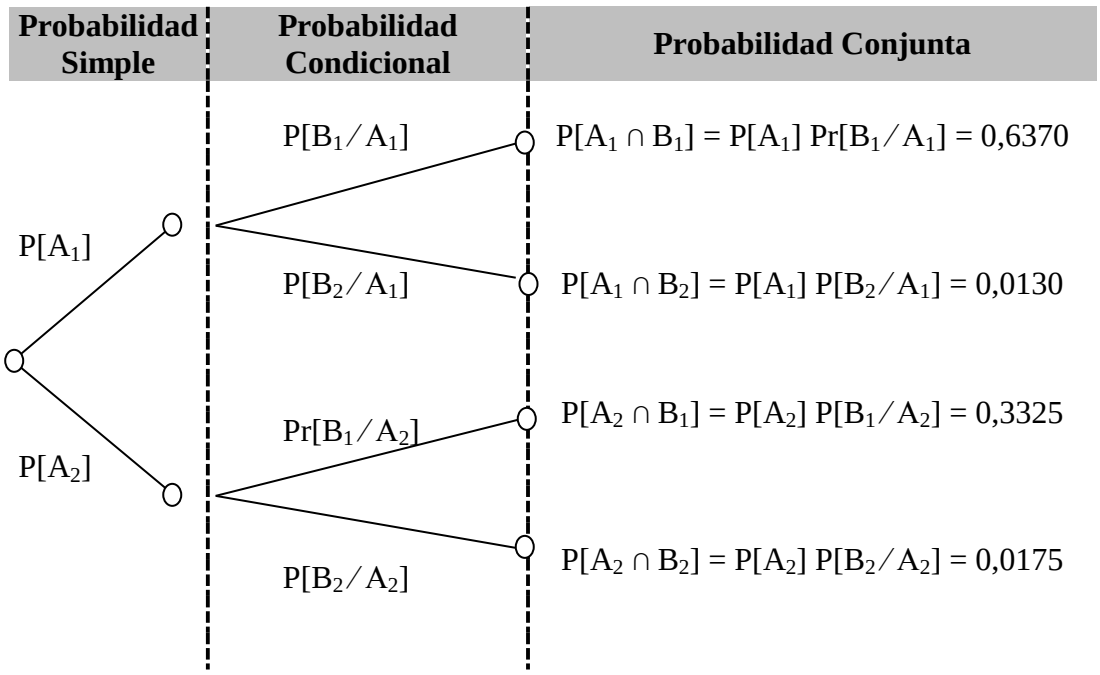


Figura 4.7

El teorema de Bayes es aplicable cuando los eventos para los que deseamos calcular probabilidades posteriores son mutuamente excluyentes y su unión es el espacio muestral completo. El teorema de Bayes puede extenderse al caso de n eventos mutuamente excluyentes A_1, A_2, $\cdots$, A_n cuya unión cubre todo el espacio muestral. En tal caso, el teorema de Bayes para el cálculo de cualquier probabilidad posterior $P[A_i/B]$ se vuelve

$$P[A_i/B] = \frac{P[A_i]P[B/A_i]}{P[B]},\ i = 1, 2, \cdots, n$$

donde $P[B] = P[A_1]P[B/A_1] + P[A_2]P[B/A_2] + \cdots + P[A_n]P[B/A_n]$ se le conoce con el nombre de *probabilidad total*.

Ejemplo 4.27: Suponga que los componentes del Ejemplo 4.16 se usan en un proceso de manufactura y que un componente en malas condiciones causa que una máquina se descomponga, ¿cuál es la probabilidad que el componente en malas condiciones provenga del proveedor 1? y ¿cuál es la probabilidad que provenga del proveedor 2? Con la información del árbol de probabilidad podemos usar el teorema de Bayes para responder a estas preguntas.

<u>Proveedor 1</u>: $P[A_1/B_2] = \dfrac{P[A_1]P[B_2/A_1]}{P[B_2]}$

<u>Proveedor 2</u>: $P[A_2/B_2] = \dfrac{P[A_2]P[B_2/A_2]}{P[B_2]}$

donde $P[B_2] = P[A_1]P[B_2/A_1] + P[A_2]P[B_2/A_2] = 0,0305$. Luego

$$P[A_1/B_2] = \frac{0,65 \times 0,02}{0,0305} = 0,4262$$

$$P[A_2/B_2] = \frac{0,35 \times 0,05}{0,0305} = 0,5738$$

Ejemplo 4.28: En un cierto laboratorio de cómputo se tienen 3 marcas de computadoras (M_1, M2 y M_3) en las proporciones 50%, 30% y 20% respectivamente. Se sabe que un cierto virus informático de Internet infecta al 5% de las computadoras de la marca 1, al 10% de la marca 2 y al 20% de la marca 3.

a. ¿Cuál es la probabilidad de una computadora seleccionada al azar esté infectada por el virus informático?
b. Si una computadora no está infectada por el virus;¿Cuál es la probabilidad de que sea de la marca 1?
c. Si cierta computadora no está infectada ¿Cuál es la probabilidad de que no sea de la marca 3?

4.9 Problemas

1. Una computadora consta de cinco celdas para ubicar un tipo de memoria electrónica. Se dispone de seis memorias en buen estado y seis en mal estado. Para que la computadora funcione es necesario que tenga como mínimo cuatro memorias en buen estado. ¿Cuál es la probabilidad de seleccionar al azar cinco memorias electrónicas del total disponible de tal forma que la computadora funcione?

2. En un intento por burlar la vigilancia en la aduana un viajero guarda junto a 10 CD's en blanco, ocho CD's con programas informáticos grabados de forma ilegal. Si el inspector de la aduana selecciona al azar tres de los CDs para verificar su contenido. ¿Cuál es la probabilidad de arrestar al viajero por posesión ilegal de programas informáticos?

3. Una planta de producción emplea 20 trabajadores en el turno de día, 15 trabajadores en el segundo turno y 10 en el turno de noche. Un consultor de control de calidad selecciona a seis trabajadores para ser entrevistados. Si la selección se hace al azar, ¿cuál es la probabilidad de que los seis trabajadores seleccionados sean del mismo turno?

4. Un artículo producido puede tener en forma independiente tres tipos de defectos; el defecto A con probabilidad 0,04, el defecto B con probabilidad 0,08 y el defecto C con probabilidad 0,05. Un artículo es seleccionado al azar para verificar su condición.
   a. ¿Cuál es la probabilidad que el artículo no presente defectos?
   b. ¿Cuál es la probabilidad que el artículo presente solo un defecto?
   c. ¿Cuál es la probabilidad que el artículo sea defectuoso?

5. La probabilidad de que el supervisor del primer turno identifique una falla en el proceso es 0,86 y la probabilidad de que el supervisor de la tarde identifique la misma falla en el proceso es 0,85, asimismo la probabilidad de que ambos identifiquen la misma falla en el proceso es 0,70, ¿Cuál es la probabilidad de que dado que el supervisor del primer turno identificó la falla, también lo haga el supervisor del turno tarde?

6. Un sistema tiene dos componentes A y B y se conecta de manera que éste funciona si cualquiera de las componentes lo hace. Se sabe que la probabilidad que el componente A funcione es 0,9, la probabilidad que el componente B funcione es 0,8 y la probabilidad que ambos funcionen es 0,72. ¿Funcionan los componentes A y B de forma independiente?

7. En la fabricación de cierto producto se presenta el defecto tipo Alpha con una probabilidad de 0,1 y el defecto tipo Beta una probabilidad de 0,5. Suponiendo que las ocurrencias de los defectos tipo Alpha y Beta son independientes:
   a. ¿Cuál es la probabilidad que un artículo no tenga ambos defectos?
   b. ¿Cuál es la probabilidad que un artículo no sea defectuoso?
   c. Calcular la probabilidad que, sabiendo que es defectuoso, tenga sólo un tipo de defecto.

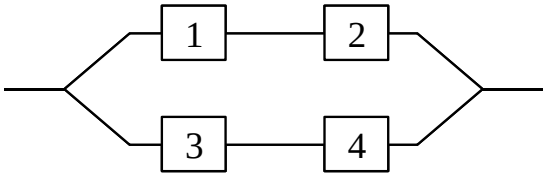
8. En una empresa que desarrolla software los programadores han sido clasificados en tres categorías: A, B y C. Se sabe que aproximadamente uno de cada 80 programas codificados por los programadores de la categoría A presentan fallas durante su ejecución. La probabilidad que un programa codificado por los programadores de la categoría B no falle es 0,975 y la probabilidad que un programa codificado por los programadores de la categoría C no falle es 0,95. Si de todos los programas desarrollados, el 50% es codificado por los programadores de la categoría A y el 30% son codificados por los programadores de la categoría B. Si se elige un programa al azar
   a. ¿Cuál es la probabilidad que presente fallas durante su ejecución?
   b. Si el programa elegido presenta fallas durante su ejecución, ¿cuál es la probabilidad que haya sido codificado por un programador de la categoría C?

9. La empresa *Omega S.A.* desea obtener un certificado de calidad de sus servicios, para lo cual debe ser visitada por dos inspectores de calidad quienes evaluarán los servicios ofrecidos por la empresa y verificarán si cumplen con los requisitos para su aprobación y certificación. Las visitas y las evaluaciones se realizan en forma independiente. La certificación de calidad se entrega sólo si ambos inspectores aprueban los servicios de la empresa *Omega S.A.*. El inspector A aprueba sólo al 2% de las empresas con algunas fallas en los servicios. El inspector B aprueba al 7% de las empresas con algunas fallas en los servicios. Si se sabe que Enigma tiene algunas fallas en sus servicios, calcule la probabilidad de que:
   a. El inspector B apruebe los servicios de *Omega S.A.*, si se sabe que el inspector A no los aprobó.
   b. *Omega S.A.* no logre la certificación.

10. Un programa de computadora utiliza cualquiera de tres subrutinas, A, B y C. Dependiendo del problema la práctica ha demostrado que la subrutina A será usada el 40% del tiempo y la subrutina B el 30% del tiempo. Si se usa la subrutina A entonces hay un 75% de probabilidad que el programa corra antes de que exceda su tiempo límite, si se usa la subrutina B hay un 50% de probabilidad de que lo haga antes de que exceda su tiempo límite, mientras que la probabilidad correspondiente para C es 65%. Si se elige un programa al azar
    a. ¿Cuál es la probabilidad que corra sin exceder el tiempo límite?
    b. ¿Cuál es la probabilidad que haya usado la subrutina B si se sabe que el programa corrió antes de exceder el tiempo límite?

11. En un curso de una facultad de ingeniería se tiene que el 20% de los alumnos son de sistemas, el 30% de electrónica, el 25% de civil y el resto de industrial. Por ciclos anteriores se sabe que los de sistemas aprueban en el 80% de los casos, los de electrónica 75%, los de civil el 72% y los de industrial en el 82% de los casos. Se elige un alumno al azar.
    a. Calcule la probabilidad de aprobar el curso y ser de industrial.
    b. Calcule la probabilidad de que sea de sistemas, si sabemos que ha aprobado el curso.

12. Un sistema consta de cuatro componentes, como se ilustra en la siguiente figura. Todo el sistema funcionará si el subsistema 1-2 funciona o el subsistema 3-4 funciona (porque los subsistemas están conectados en paralelo). Como los dos componentes de cada subsistema están conectados en serie, un subsistema funcionará sólo si ambos componentes funcionan. Si los componentes funcionan o fallan de modo independiente uno del otro y si cada uno de ellos funciona con una probabilidad de 0,9, ¿cuál es la probabilidad de que todo el sistema funcione (coeficiente de confiabilidad del sistema)?



Capítulo 5

Variables aleatorias

5.1 Introducción

En este capítulo se continúa con el estudio de la probabilidad introduciendo los conceptos de *variables aleatorias* y distribuciones de probabilidad. Se considera las distribuciones de probabilidad de variables aleatorias tanto discretas como continuas. Son particularmente interesantes las cinco distribuciones de probabilidad que se mencionan en este capítulo: Binomial, de Poisson, Uniforme, Normal y Exponencial.

5.2 Variables aleatorias

En el Capítulo 4 se define un experimento como cualquier proceso que genera resultados bien definidos. Se considera de interés el proceso de asignar *valores numéricos* a los resultados experimentales. Para hacerlo, se introduce la noción de *variable aleatoria*.

Para cualquier experimento en particular, una variable aleatoria puede definirse de modo que cada resultado experimental posible genere exactamente un valor numérico para la variable aleatoria, es decir que una variable aleatoria es la descripción numérica del resultado de un experimento.

Ejemplo 5.1:

Experimento	Variable aleatoria	Valores posibles para la variable aleatoria
Realizar 100 llamadas de ventas	Número de ventas logradas	0 , 1 , 2 , … , 100
Inspeccionar un lote de 50 artículos	Número de artículos defectuosos	0 , 1 , 2 , … , 50
Trasladarse de la UPC al hogar	Tiempo requerido (en minutos)	$15 \leq$ Tiempo ≤ 45
Administrar un restaurante	Número de clientes atendidos por día	0 , 1 , 2 , …

Una variable aleatoria puede clasificarse como discreta o continua dependiendo de los valores numéricos que pueda asumir. Una variable aleatoria que solo puede tomar una secuencia finita o infinita numerable de valores es una *variable aleatoria discreta*. El número de ventas logradas, número de artículos defectuosos y número de clientes atendidos por día son ejemplos de variables aleatorias discretas. Una variable aleatoria que toma una secuencia infinita no numerable de valores es una *variable aleatoria continua*, por ejemplo: tiempo requerido (en minutos) para trasladarse de la UPC al hogar.

5.3 Variables aleatorias discretas

Distribución de probabilidad de una variable aleatoria discreta

La *función de distribución de probabilidad*, denotada por $f(x)$, proporciona la probabilidad que la variable aleatoria discreta X tome el valor especifico x. En la construcción de una distribución de probabilidad discreta siempre deben satisfacerse dos condiciones:

- $0 \leq f(x) \leq 1$
- $\sum f(x) = 1$

Ejemplo 5.1: Una empresa recibe discos duros de su proveedor en cajas de 50 unidades. Luego de la comercialización de los discos y de acuerdo a las devoluciones realizadas por sus propios clientes ha determinado el número de discos duros defectuosos en las últimas 500 cajas compradas. La información se muestra a continuación:

Número de discos duros defectuosos	0	1	2	3	4
Número de cajas	300	100	50	25	25

Sea la variable aleatoria X = Número de discos duros defectuosos por caja de 50 unidades. La función de distribución de probabilidad se puede obtener calculando la frecuencia relativa para cada valor de X.

X	0	1	2	3	4
f(x)	0,60	0,20	0,10	0,05	0,05

Valor esperado de una variable aleatoria discreta

El *valor esperado* de una variable aleatoria discreta es un promedio ponderado de todos los valores posibles de la misma, donde los pesos son las probabilidades asociadas con los valores. El valor esperado de una variable aleatoria discreta se calcula con la siguiente expresión:

$$\mu = \mathrm{E}[X] = \sum x\, f(x)$$

El valor esperado de una variable aleatoria es el valor medio, o promedio. Para experimentos que pueden repetirse muchas veces, el valor esperado puede interpretarse como el valor promedio a largo plazo para la variable aleatoria. Sin embargo, el valor esperado no necesariamente es el número que pensamos asumirá la variable aleatoria la próxima vez que se realice el experimento.

Ejemplo 5.2: Si se desea calcular el valor esperado de la variable aleatoria definida en el ejemplo anterior.

$$\mathrm{E}(X) = \sum X\ f(x) = 0(0{,}6) + 1(0{,}2) + 2(0{,}1) + 3(0{,}05) + 4(0{,}05) = 0{,}95$$

Luego, el valor esperado del número de discos defectuosos por caja es 0,95.

Varianza de una variable aleatoria discreta

El valor esperado nos da una idea acerca del valor promedio o central para la variable aleatoria, pero es importante tener además una medida de la dispersión o variabilidad de los valores posibles de la variable aleatoria. La *varianza* es una medida usada para representar la dispersión o variabilidad en los valores de una variable aleatoria. La varianza de una variable aleatoria discreta se calcula con la siguiente expresión:

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \left[\mathrm{E}(X)\right]^2, \qquad \text{donde } \mathrm{E}(X^2) = \sum x^2 f(x)$$

Una medida de dispersión relacionada es la *desviación estándar*, σ, la cual se define como la raíz cuadrada positiva de la varianza.

$$\sigma_X = \sqrt{\mathrm{V}\left[X\right]}$$

Ejemplo 5.3: Si se desea calcular la varianza y desviación estándar de la variable aleatoria definida en el ejemplo anterior.

$$\mathrm{E}(X^2) = \sum X^2 f(x) = 0^2(0{,}6) + 1^2(0{,}2) + 2^2(0{,}1) + 3^2(0{,}05) + 4^2(0{,}05) = 1{,}85$$

Entonces: Varianza $= \mathrm{V}\left[X\right] = 1{,}85 - 0{,}95^2 = 0{,}9475$ (discos duros defectuosos)^2

Luego, Desviación Estándar de X = $\sigma_X = 0{,}9734$ discos duros defectuosos.

Ejemplo 5.4: Un jugador lanza tres monedas corrientes. Gana $5, si salen tres caras, $ 3 si salen dos caras y $1 si solamente sale una cara. De otro lado pierde $ 15 si salen tres sellos. Calcule e interprete el valor esperado del juego.

5.4 Principales variables aleatorias discretas

Distribución de probabilidad Binomial

La *distribución de probabilidad binomial* es una distribución discreta de probabilidad que se relaciona con un *experimento binomial*, el cual tiene cuatro propiedades:

1. El experimento consiste de una sucesión de **"*n*" intentos o ensayos idénticos**.
2. En cada intento o ensayo son posibles dos resultados. A uno lo llamaremos ***éxito* y al otro *fracaso***.
3. La probabilidad de obtener un éxito, **"*p*", no cambia** de un intento o ensayo a otro.
4. Los intentos o **ensayos son independientes**.

En un experimento binomial la variable aleatoria de interés se define como X = Número de *éxitos* obtenidos en los n intentos o ensayos. Esta variable aleatoria puede tomar los valores $0,1,2,\cdots,n$. La función de distribución de probabilidad asociada con esta variable aleatoria es:

$$f(x)=\mathrm{P}[X=x]=\mathrm{C}_x^n p^x (1-p)^{n-x},\ x=0,1,2,\cdots,n$$

El valor esperado: $\mathrm{E}[X]=np$; La varianza: $\mathrm{V}[X]=np(1-p)$.

Lo anterior puede escribirse como: $X \sim \mathrm{B}(n,p)$.

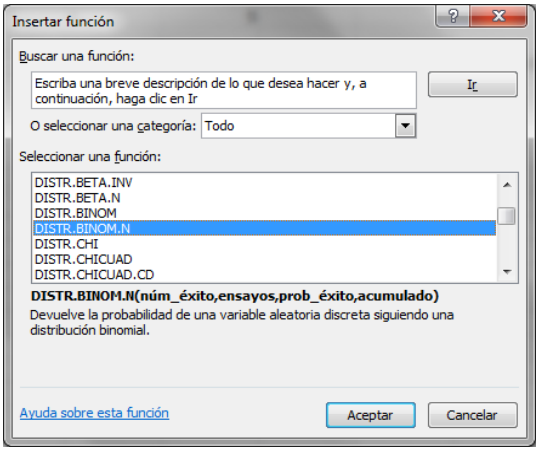
Ejemplo 5.5: La probabilidad de que cualquier símbolo particular de código se transmita erróneamente a través de un sistema de comunicaciones es 0,10. En la transmisión de los símbolos los errores ocurren de manera independiente unos de otros. Suponga que se envía un mensaje formado por 10 símbolos. ¿Cuál es la probabilidad que se transmitan erróneamente solo dos símbolos?

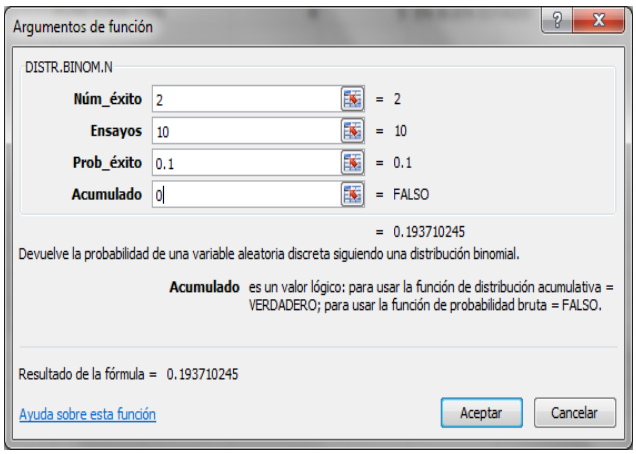
Sea X = Número de símbolos transmitidos erróneamente del total de 10, entonces $X \sim \mathrm{B}(n=10, p=0.10)$. La función de distribución de probabilidad es

$$f(x)=\mathrm{P}[X=x]=\mathrm{C}_x^{10} 0{,}10^x (0{,}90)^{10-x},\ x=0,1,2,\cdots,10$$

Luego, $f(2)=\mathrm{P}[X=2]=\mathrm{C}_2^{10} 0{,}10^2 (0{,}90)^{10-2}=0{,}1937$.

Utilizando el **MS EXCEL VERSION 2010**:
Menú **Fórmulas** dar clic en **Insertar Función,** seleccionar **DISTR.BINOM.N**

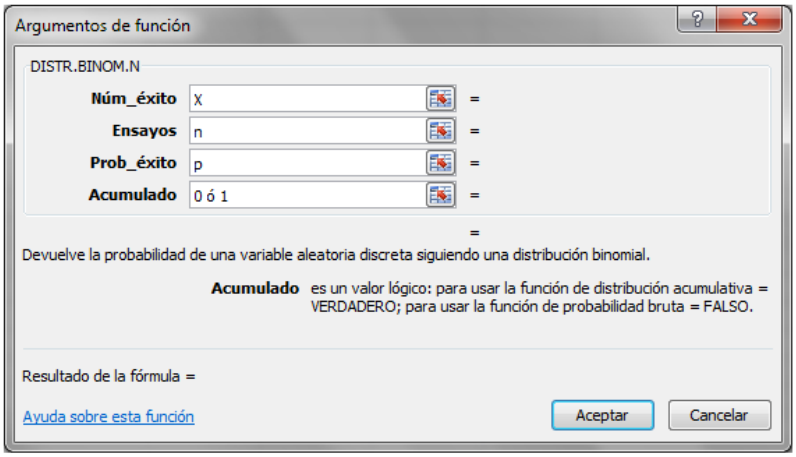




Para, este ejemplo, P(X=2) NO está acumulando por tanto en el campo Acumulado se ingresa el valor de "0" (FALSO)

P(X=2) = 0,193710

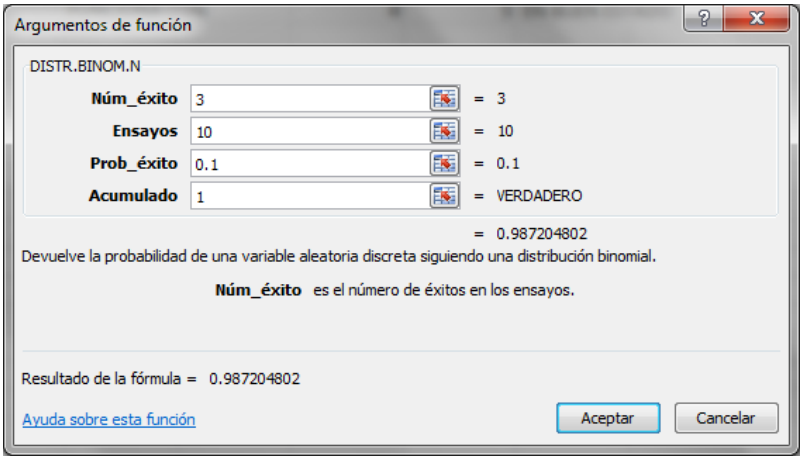
En general los campos en el **EXCEL** se ingresan de la siguiente forma:



En Acumulado: 0 ó FALSO, calcula P(X = *a*) NO ACUMULA
En Acumulado: 1 ó VERDADERO calcula P(X ≤ *a*) SI ACUMULA

Ejemplo 5.6: Con respecto al Ejemplo 5.5, ¿cuál es la probabilidad de transmitir como máximo tres símbolos erróneamente?

$$P[X \leq 3] = P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3] = 0,9872$$



Para, este ejemplo, P(X ≤ 3) SI está acumulando por tanto en el campo Acumulado se ingresa el valor de "1" (VERDADERO)

P(X ≤ 3) = 0,9872

Ejemplo 5.7: Se sabe que el 70% de los egresados de cierta facultad de una universidad se titulan al año de egresar de la universidad. El año pasado egresaron 10 de esta facultad.

a. Encuentre la función de distribución del número de egresados que se titulan al año de egresado
b. Calcule e interprete el valor esperado del número de egresados que se titulan al año
c. Calcule la probabilidad de que:
   i. Ninguno se titule
   ii. Menos de tres se titulen
   iii. Más de cuatro se titulen

Distribución de probabilidad Hipergeométrica

La *distribución de probabilidad hipergeométrica* está estrechamente relacionada con la distribución binomial, pero difieren en dos puntos: en la distribución hipergeométrica los ensayos **no son independientes** y **la probabilidad de éxito varía de ensayo a ensayo**.

En la notación usual, r denota el número de elementos considerados como éxitos que hay en una población de tamaño N, y $N - r$ denota el número de elementos considerados como fracasos que hay en dicha población. La función de probabilidad hipergeométrica se usa para calcular la probabilidad de que en una muestra aleatoria de n elementos, seleccionados sin reemplazo, se tengan x éxitos y $n - x$ fracasos. La función de distribución de probabilidad asociada con esta variable aleatoria es:

$$f(x) = \mathrm{P}[X = x] = \frac{\mathrm{C}_x^r \mathrm{C}_{n-x}^{N-r}}{\mathrm{C}_n^N}, \; x = \max(0, n + r - N), \cdots, \min(r, n)$$

El valor esperado: $\mathrm{E}[X] = n\dfrac{r}{N}$; La varianza: $\mathrm{V}[X] = n\dfrac{r}{N}\left(1 - \dfrac{r}{N}\right)\left(\dfrac{N-n}{N-1}\right)$

Lo anterior puede escribirse como: $X \sim \mathrm{H}(N, n, r)$

Ejemplo 5.8: Una empresa fabrica fusibles que empaca en cajas de 12 unidades cada una. Suponga que un inspector selecciona al azar tres de los 12 fusibles de una caja para inspeccionarlos. Si la caja contiene cinco fusibles defectuosos, ¿cuál es la probabilidad que el inspector encuentre más de un fusible defectuoso?

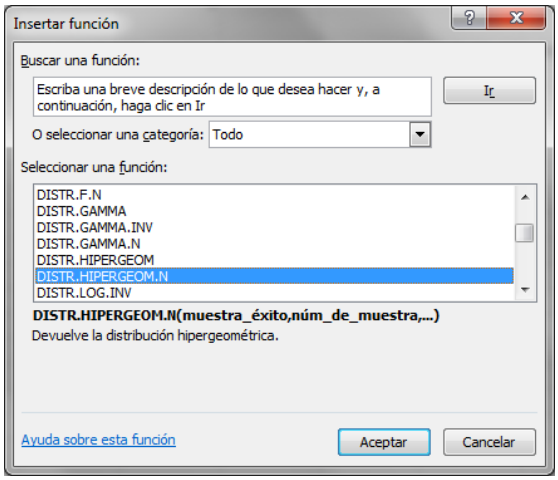
Sea X = Número de fusibles defectuosos encontrados por el inspector en la muestra de tamaño tres, entonces $X \sim \mathrm{H}(N=12, n=3, r=5)$. La función de distribución de probabilidad es:

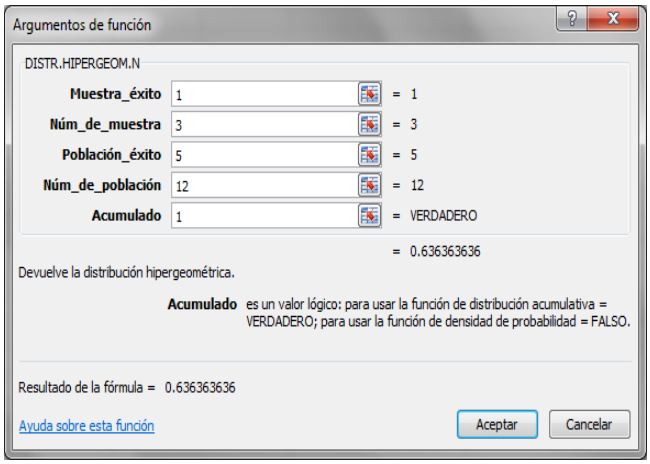
$$f(x) = \mathrm{P}[X = x] = \frac{\mathrm{C}_x^5 \mathrm{C}_{3-x}^7}{\mathrm{C}_3^{12}}, \; x = 0, \quad 1, \quad 2, \quad 3$$

Luego $\mathrm{P}[X > 1] = 1 - \mathrm{P}[X \leq 1] = 1 - \left(\frac{\mathrm{C}_0^5 \mathrm{C}_3^7}{\mathrm{C}_3^{12}} + \frac{\mathrm{C}_1^5 \mathrm{C}_2^7}{\mathrm{C}_3^{12}} \right) = 1 - 0,6364 = 0,3636$.

Utilizando el **MS EXCEL VERSION 2010**:

Menú **Fórmulas** dar clic en **Insertar Función,** seleccionar **DISTR.HIPERGEOM.N**



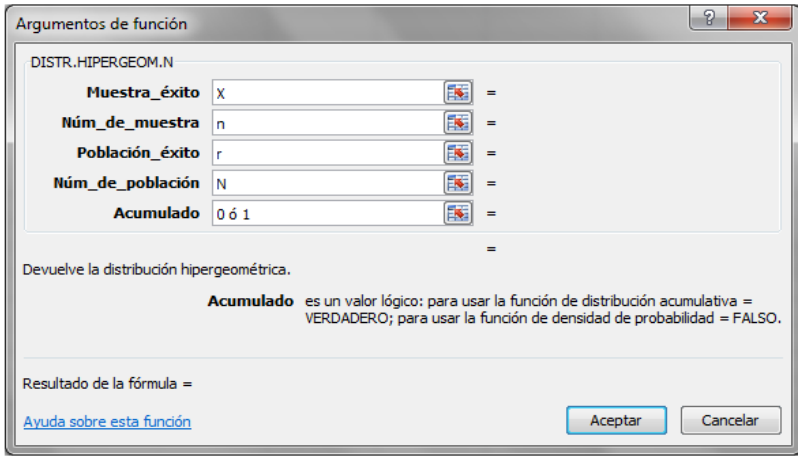


Para, este ejemplo, P(X>1) = 1- P(X ≤ 1)

P(X<=1) SI está acumulando por tanto en el campo Acumulado se ingresa el valor de "1" (VERDADERO)

P(X ≤ 1) = 0,63636
P(X > 1) = 1- 0,63636 = 0,36363

En general los campos en el **EXCEL** se ingresan de la siguiente forma:



En Acumulado: 0 ó FALSO, calcula P(X = *a*) NO ACUMULA
En Acumulado: 1 ó VERDADERO calcula P(X ≤ *a*) SI ACUMULA

Ejemplo 5.9: Lotes de 40 componentes cada uno se consideran aceptables si no contienen más de tres defectuosos. El procedimiento para muestrear el lote es la selección de cinco componentes al azar y se rechaza el lote si se encuentran un componente defectuoso.

a. Escriba la función de distribución de la variable definida como X: número de componentes defectuosos en la muestra. Interprete
b. ¿Cuál es la probabilidad de que se encuentre exactamente un componente defectuoso en la muestra si hay tres defectuosos en todo el lote?
c. ¿Cuál es la probabilidad de rechazar el lote?
d. Determine e interprete la media y la varianza de la variable aleatoria X

Distribución de probabilidad de Poisson

En esta sección se describe una variable aleatoria discreta que se usa con frecuencia para estimar la **cantidad de sucesos u ocurrencias en determinado *intervalo de tiempo* o *espacio***. Por ejemplo, la variable aleatoria de interés podría ser:

X = Número de clientes de los clientes al banco *por hora*
X = Número de reparaciones necesarias por cada *10 kilómetros de carretera*
X = Número de fugas en *100 millas de un oleoducto*.

Si se satisfacen las dos propiedades siguientes, la variable aleatoria X = Número de ocurrencias en determinado *intervalo de tiempo* o *espacio* es una variable aleatoria discreta que se describe con la *distribución de probabilidad de Poisson*:

- La probabilidad de una ocurrencia es igual en dos intervalos cualesquiera de igual longitud.
- La ocurrencia o no ocurrencia en cualquier intervalo es independiente de la ocurrencia o no ocurrencia en cualquier otro intervalo.

La función de probabilidad de Poisson es:

$$f(x) = \mathrm{P}[X = x] = \frac{e^{-\lambda}\lambda^{x}}{x!}, \; x = 0,1,2,\cdots$$

El valor esperado: E(X) = λ
La varianza: V(X) = λ

Lo anterior puede escribirse como: $X \sim \mathrm{P}(\lambda)$,

Donde: λ es la tasa o número promedio de ocurrencias en determinado *intervalo de tiempo* o *espacio* y $e \cong 2,71828$.

Ejemplo 5.10: La división de mantenimiento de telefonía está tratando de decidir cuántos reparadores necesita para proporcionar un nivel aceptable de servicio a sus clientes. El número de quejas que llegan al centro de servicio sigue una distribución de Poisson, con una tasa promedio de 30 llamadas al día. ¿Cuál es la probabilidad de recibir en el centro de servicio más de 20 quejas en medio día?

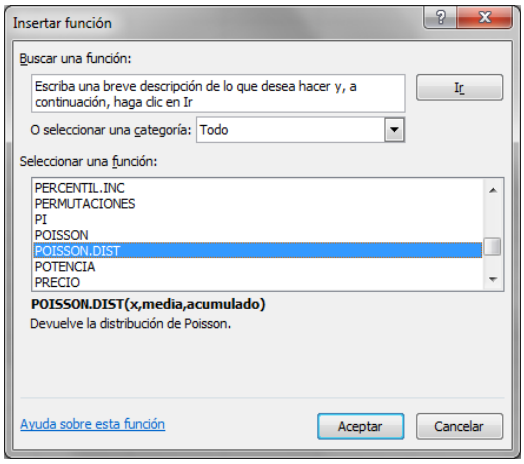
Sea X = Número de quejas que llegan al centro de servicio en *medio día*, entonces $X \sim \mathrm{P}(\lambda = 15)$. La función de distribución de probabilidad es:

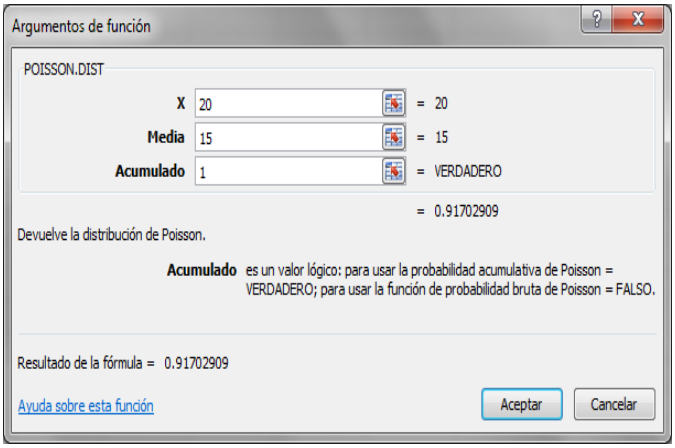
$$f(x) = \mathrm{P}[X = x] = \frac{e^{-15}15^{x}}{x!}, \; x = 0,1,2,\cdots$$

Luego, $\mathrm{P}[X > 20] = 1 - \mathrm{P}[X \le 20] = 1 - 0,9170 = 0,0830$.

Utilizando el **MS EXCEL VERSION 2010**:

Menú **Fórmulas** dar clic en **Insertar Función,** seleccionar **POISSON.DIST**



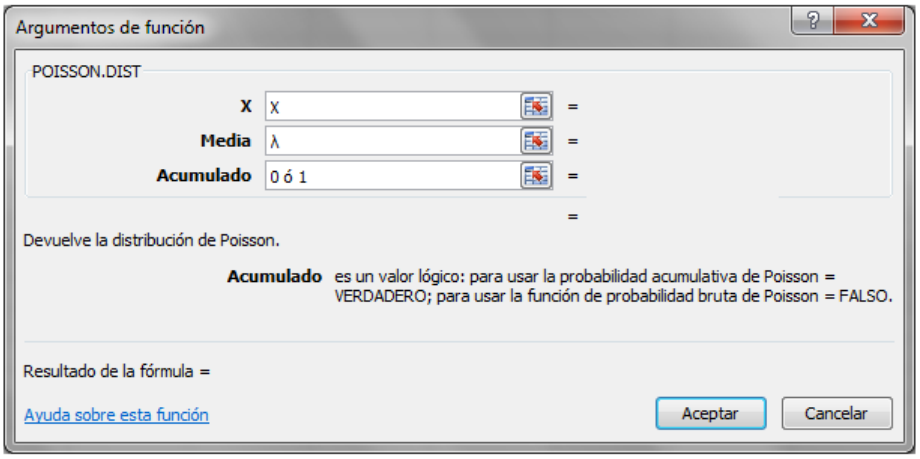


Para, este ejemplo, $P(X>20) = 1- P(X \leq 20)$

$P(X \leq 20)$ SI está acumulando por tanto en el campo Acumulado se ingresa el valor de "1" (VERDADERO)

$$\mathbf{P(X \leq 20) = 0{,}91702}$$
$$\mathbf{P(X>20) = 1- 0{,}91702= \ 0{,}08297}$$

En general los campos en el **EXCEL** se ingresan de la siguiente forma:



En Acumulado: 0 ó FALSO, calcula $P(X = a)$ NO ACUMULA
En Acumulado: 1 ó VERDADERO calcula $P(X \leq a)$ SI ACUMULA

Ejemplo 5.11: Con respecto al Ejemplo 5.6, ¿cuál es la probabilidad que lleguen al centro de servicio menos de 25 quejas al día?

Sea X = Número de quejas que llegan al centro de servicio *por día*, entonces $X \sim P(\lambda = 30)$. La función de distribución de probabilidad es:

$$f(x) = \mathrm{P}[X = x] = \frac{e^{-30}30^x}{x!}, \; x = 0,1,2,\cdots$$

Luego, $\mathrm{P}[X < 25] = \mathrm{P}[X \leq 24] = 0{,}1572$.

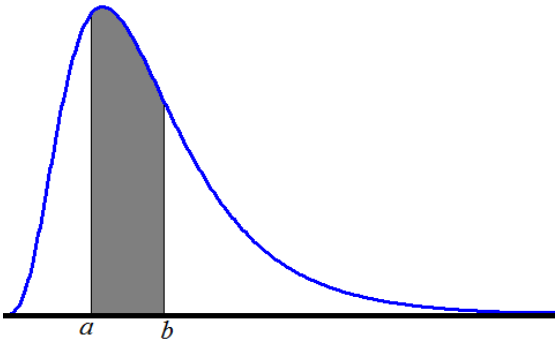
Ejemplo 5.12: Cierto tipo de tela para tapicería tiene, en promedio, dos defectos por metro. Si se supone una distribución de Poisson, ¿Cuál es la probabilidad de que un rollo de 30 metros de esta tela tenga cuatro o más defectos?

Ejemplo 5.13: Suponga que un campeonato de fútbol, en cada partido el árbitro saca un promedio de 3,5 tarjetas amarillas. Calcule la probabilidad de que en un partido cualquiera de este campeonato haya

a. Cuatro tarjetas amarillas
b. Entre cuatro y seis tarjetas amarillas
c. No más de tres tarjetas amarillas
d. Ninguna tarjeta amarilla
e. Por lo menos tres tarjetas amarillas

5.5 Variables aleatorias continuas

Una variable aleatoria continua toma un número infinito de valores: estos valores pueden asociarse a mediciones en una escala continua, de manera que no exista espacios vacíos o interrupciones. Por ejemplo la medida del punto de ebullición del agua, tomara valores cercanos a 100 grados centígrados cuando la medición se realiza a nivel del mar. En el caso de variables aleatorias sólo tiene sentido referirse a probabilidades sobre intervalos, mas no sobre valores puntuales. Por ejemplo si X es una variable aleatoria que representa el peso real de un frasco de 250 gramos de café instantáneo escogido al azar, entonces en este caso tiene sentido preguntarse sobre el valor $P(246 < X < 266)$, pero no tiene sentido escribir $P(X = 245)$, pero en la ninguna lata de café instantáneo pesará exactamente 245 gramos, luego se dirá que $P(X = 245) = 0$



La probabilidad de que la variable aleatoria *X* tome valores entre los número *a* y *b*, está dada por el área bajo la curva de *f* desde *a* hasta *b*.

5.6 Definición

Una variable aleatoria es continua si puede asumir cualquier valor en uno o más intervalos de números reales y la probabilidad que asuma un valor específico dado es cero.

Sea X una variable aleatoria continua, y sea *f* una función tal que:

1. $f(x) \geq 0 \quad , \quad \forall x \in \Re$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = \int_a^b f(x)dx$

Entonces *f* se llama función de densidad de X

Condiciones necesarias y suficientes para que una función *f* sea *función de densidad*.

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

5.7 Principales variables aleatorias continuas

Una diferencia fundamental que separa las variables aleatorias discretas de las continuas es el cálculo de las probabilidades. Para una variable aleatoria discreta, la función de probabilidad f proporciona la probabilidad que la variable aleatoria tome un valor específico x. Para las variables aleatorias continuas la función que permite obtener las probabilidades es la *función de densidad de probabilidad*, que también se denota por f. La diferencia es que esta última no permite encontrar directamente las probabilidades. No obstante, el área bajo la gráfica de f, que corresponde a un intervalo determinado, da la probabilidad que la variable aleatoria continua tome un valor en ese intervalo. Una de las implicaciones de la definición de probabilidad para variables aleatorias continuas es que la probabilidad para cualquier valor específico siempre es cero.

Distribución de probabilidad uniforme

Suponga que una variable aleatoria toma valores igualmente distribuidos sobre un intervalo. Es decir, la probabilidad que la variable aleatoria tome un valor en cada subintervalo de igual longitud es la misma. Se dice entonces que la variable aleatoria se encuentra *uniformemente distribuida* sobre dicho intervalo.

Una variable aleatoria continua X tiene distribución uniforme sobre el intervalo (a,b) si su función de densidad de probabilidad está dada por:

$$f(x) = \frac{1}{b-a}, \quad a \le x \le b$$

El valor esperado: $\mathrm{E}[X] = (a+b)/2$

La varianza $\mathrm{V}[X] = (b-a)^2/12$

Lo anterior puede escribirse como: $X \sim \mathrm{U}(a,b)$

La función de densidad de probabilidad uniforme es constante en el intervalo (a,b) tal como se muestra en la Figura 5.1. Por esta razón la distribución uniforme también se conoce como *distribución rectangular*.

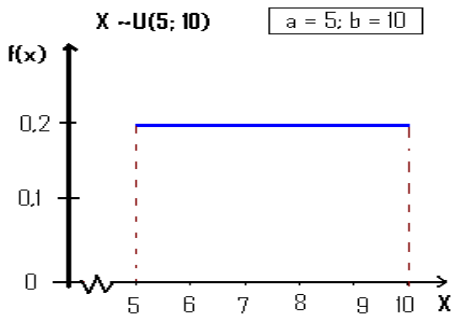


Figura 5.1

Si $a=0$ y $b=1$ se dice que la variable aleatoria X tiene distribución uniforme sobre el intervalo unitario con función de densidad de probabilidad $f(x)=1,\ 0\le x\le 1$. Esta distribución juega un papel importante en el proceso de *simulación* de valores para una variable aleatoria con una distribución específica.

Ejemplo 5.14: Una línea aérea anuncia que los vuelos de Lima a Cuzco tienen una duración promedio de una hora con 15 minutos. Si se supone que el tiempo de vuelo entre las ciudades mencionadas tiene distribución uniforme cuyo rango es igual a 18 minutos. ¿Cuál es la probabilidad de que el vuelo tenga una demora mayor a 5 minutos?

La información del problema es

$$\frac{a+b}{2}=75 \text{ y } b-a=18, \text{ entonces } a=66 \text{ y } b=84$$

Si X = Tiempo de vuelo (en minutos) de Lima a Cuzco, entonces $X\sim U(a=66, b=84)$. La función de densidad de probabilidad es:

$$f(x)=\frac{1}{18},\ 66\le x\le 84$$

Luego, $\mathrm{P}[X>80]=(84-80)\times\frac{1}{18}=0,2222.$

Ejemplo 5.15: Una tienda que vende programas de software para microcomputadoras, ha encontrado que entre el 5% y el 15% de los pedidos diarios requieren un envío especial. Suponga que, como una primera aproximación, consideramos que la distribución de pedidos (en porcentaje) en cierto día seleccionado al azar es uniforme en el intervalo que va de 5 a 15. Calcule la media, la varianza y la probabilidad de que el porcentaje de pedidos este entre 9 y 12.

X ~U(5; 15) $\mu_X=10\%$; $\sigma_X^2=25/3$; $P(9<x<12)=0,3$

Distribución de probabilidad exponencial

Una distribución continua de probabilidad que es útil para describir el tiempo necesario para realizar alguna actividad es la *distribución de probabilidad exponencial*. La función de densidad de probabilidad exponencial es:

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}; \quad x \geq 0,$$

Esperanza y varianza.

El valor esperado y la varianza de la variable X son: $E(X) = \beta; \quad V(X) = \beta^2$

Lo anterior puede escribirse como $X \sim \mathrm{E}(\beta)$,

La Figura 5.2 muestra tres distribuciones exponenciales con diferentes parámetros.

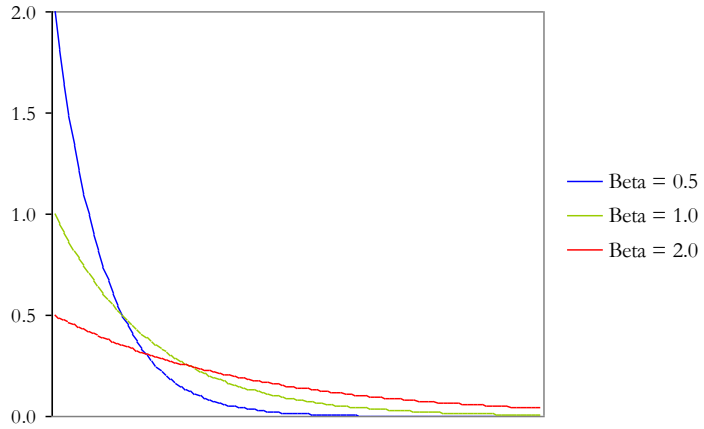


Figura 5.2

Para calcular probabilidades acumuladas en la distribución exponencial se puede utilizar la siguiente expresión:

$$\mathrm{P}[X \leq k] = 1 - e^{-\frac{k}{\beta}}$$

Ejemplo 5.16: El tiempo de duración (en horas) de un dispositivo electrónico es una variable aleatoria que tiene la siguiente función de densidad

$$f(x) = \frac{1}{50} e^{-\frac{x}{50}}, \; x \geq 0$$

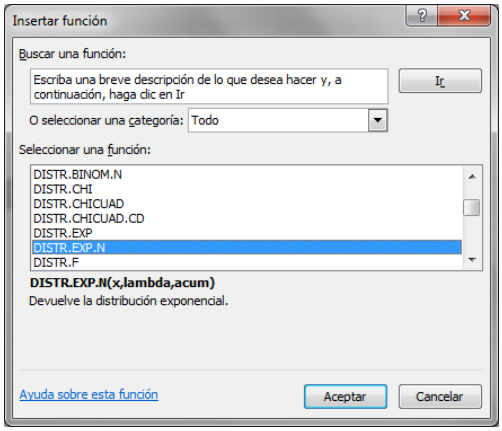
¿Cuál es la probabilidad que un dispositivo elegido al azar dure como más de 80 horas?

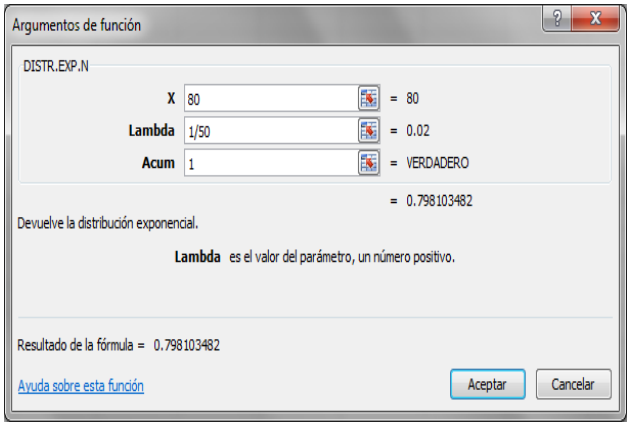
Si X = Tiempo de duración (en horas) de un dispositivo electrónico, entonces $X \sim \mathrm{Exp}(\beta = 50)$

Luego, $\mathrm{P}[X > 80] = 1 - \mathrm{P}[X \leq 80] = 1 - 0,7981 = 0,2019$.

Utilizando el **MS EXCEL VERSION 2010**:

Menú **Fórmulas** dar clic en **Insertar Función,** seleccionar **DISTR.EXP.N**



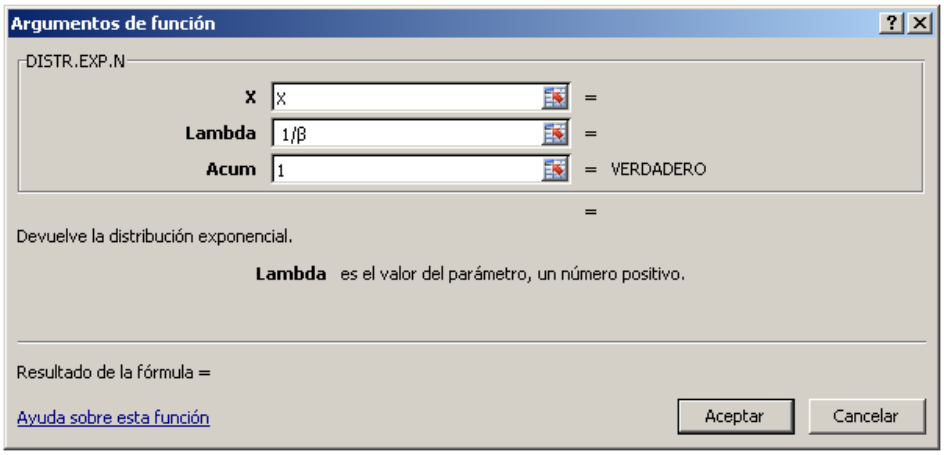


Para, este ejemplo, $P(X > 80) = 1- P(X \leq 80)$

$P(X \leq 80)$ SI está acumulando por tanto en el campo Acumulado se ingresa el valor de "1" (VERDADERO)

$$\mathbf{P(X \leq 80) = 0{,}7981}$$
$$\mathbf{P(X > 80) = 1- 0{,}7981= \ 0{,}2019}$$

En general los campos en el **EXCEL** se ingresan de la siguiente forma:



En Acum: 0 ó FALSO, devuelve el valor de la función en el punto $X = a$ (No es una probabilidad)
En Acum: 1 ó VERDADERO calcula $P(X \leq a)$ SI ACUMULA

Ejemplo 5.17: Con respecto al problema anterior, ¿cuántas horas como mínimo debe durar uno de estos componentes para ser considerado dentro del 20% de los más durables?

Ejemplo 5.18: Una de las empresas de la categoría A se llama "FOCOSA" y se dedica a la fabricación de focos eléctricos. Se tiene interés en estudiar la variable aleatoria continua X, duración dado en unidades de 100 horas. Se considera que X se distribuye según la siguiente función de densidad:

$$f(x) = \begin{cases} e^{-x} & para \quad 0 < x < \infty \\ 0 & c.c \end{cases}$$

Supóngase que el costo de fabricación de foco es de $50. El fabricante vende el artículo a $125, pero garantiza una devolución y reintegra si X < 0,9. ¿Cuál es la utilidad esperada por artículo?

Distribución de probabilidad normal

Quizás la distribución más importante de probabilidad para describir una variable aleatoria continua es la *distribución de probabilidad normal*. Esta tiene una gran variedad de aplicaciones prácticas en las que las variables aleatorias son altura y peso de personas, puntuaciones de exámenes, mediciones científicas, etc. También se utiliza en el proceso de *inferencia estadística* que es el tema principal del siguiente capítulo.

La función de densidad de probabilidad que la define es la siguiente:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

El valor esperado es: μ ; La varianza es: σ^2

Lo anterior puede escribirse como: $X \sim \mathrm{N}(\mu, \sigma^2)$

Hay algunas observaciones por mencionar acerca de las características de la distribución de probabilidad normal.

- La familia completa de distribuciones de probabilidad normales se diferencia por su media μ y su varianza σ^2.
- El punto más alto de la curva normal es la media, que también es la mediana y la moda de la distribución.
- La media de la distribución puede ser cualquier valor numérico: negativo, cero positivo. La Figura 5.2 muestra tres distribuciones normales con la misma varianza pero con tres medias distintas.
- La distribución de probabilidad normal es simétrica y su forma a la izquierda de la media es una imagen idéntica de la forma a la derecha de la media. Las colas, es decir, los extremos o los lados de la curva se prolongan al infinito en ambas direcciones y, teóricamente, nunca tocan el eje horizontal.
- La desviación estándar determina el ancho de la curva. A valores mayores de la varianza se tienen curvas más anchas y bajas que muestran una mayor dispersión en los datos. La Figura 5.3 muestra tres distribuciones normales con la misma media pero con diferentes varianzas.
- Las probabilidades para la variable aleatoria normal están dadas por áreas bajo la curva. El área total siempre es igual a uno, esto se cumple para todas las distribuciones continuas de probabilidad. Debido a que la distribución normal es simétrica, el área total bajo la curva a la izquierda de la media es 0,50 y, por consiguiente, el área total bajo la curva a la derecha de la media es también 0,50.

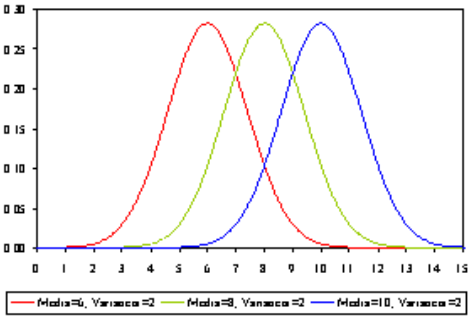


Figura 5.2

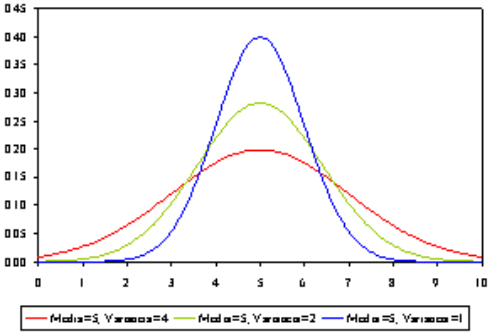


Figura 5.3

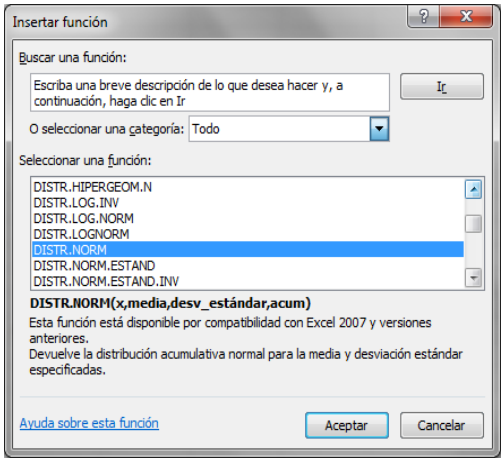
Ejemplo 5.19: Una máquina que expende bebidas ligeras está regulada de modo que descarga un promedio de 200 ml por vaso. Si la cantidad descargada por esta máquina tiene distribución normal con desviación estándar 15 ml. ¿Qué porcentaje de los vasos servidos contendrá más de 224 ml?

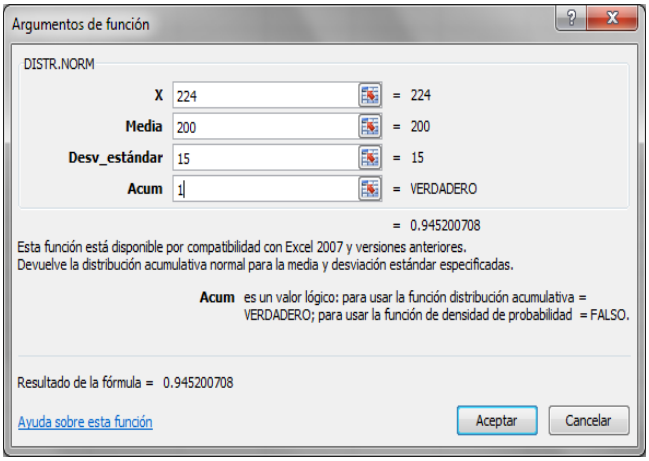
Sea $X =$ Cantidad descargada por la máquina de bebidas (en ml), entonces $X \sim \mathrm{N}(\mu = 200; \sigma = 15)$.

Luego, $\mathrm{P}[X > 224] = 1 - \mathrm{P}[X \leq 224] = 1 - 0,9452 = 0,0548$. Aproximadamente el 5,48% de los vasos contendrá más de 224 ml.

Utilizando el **MS EXCEL VERSION 2010**:

Menú **Fórmulas** dar clic en **Insertar Función,** seleccionar **DISTR.NORM**





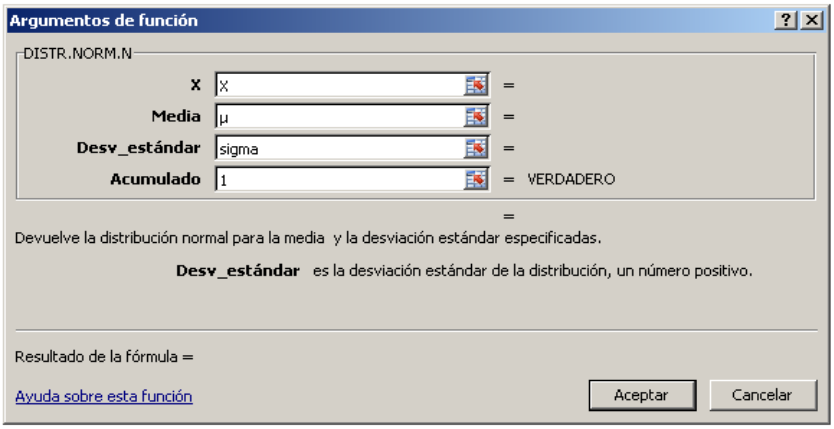
Para, este ejemplo, P(X>224) = 1- P(X $\leq$ 224)

P(X $\leq$ 224) SI está acumulando por tanto en el campo Acumulado se ingresa el valor de "1" (VERDADERO)

P(X $\leq$ 224) = 0,9452

P(X > 224) = 1- 0,9452= 0,0548

En general los campos en el **EXCEL** se ingresan de la siguiente forma:



Ejemplo 5.20: Con respecto al ejemplo anterior, ¿cuál es la probabilidad que un vaso contenga entre 191 y 209 ml?

$$\mathrm{P}[191 < X < 209] = \mathrm{P}[X \leq 209] - \mathrm{P}[X \leq 191] = 0,7257 - 0,2742 = 0,4515$$

Ejemplo 5.21: Con respecto al Ejemplo 5.14, ¿cuál es la cantidad máxima descargada por esta máquina en un vaso para poder considerarlo dentro del grupo del 30% con menores contenidos?

$$P[X \leq k] = 0,30 \text{, entonces } k = 192,13 \text{ ml}$$

Distribución de probabilidad normal estándar

Se dice que una variable aleatoria que tiene distribución normal con media cero y varianza uno tiene *distribución de probabilidad normal* estándar. Si se desea calcular una probabilidad para cualquier distribución normal, ésta puede determinarse usando la distribución normal estándar. La expresión que permite pasar de cualquier variable aleatoria normal con media μ y varianza σ^2 a la distribución normal estándar es:

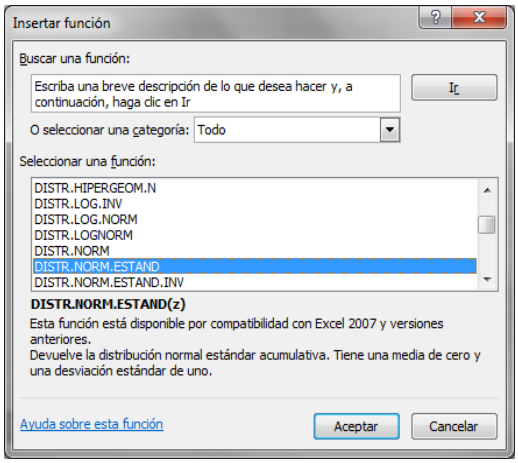
$$Z = \frac{X - \mu}{\sigma}$$

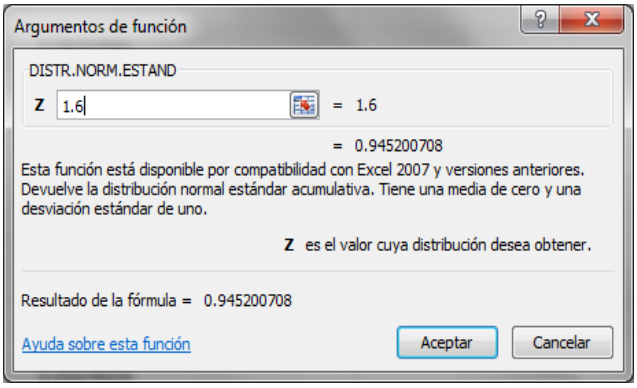
Ejemplo 5.22: Con respecto al Ejemplo 5.14, la probabilidad puede calcularse usando la distribución normal estándar.

$$P[X > 224] = P\left[Z > \frac{224 - 200}{15}\right] = P[Z > 1,6] = 1 - P[Z \leq 1,6] = 1 - 0,9452 = 0,0548$$

Utilizando el **MS EXCEL VERSION 2010**:

Menú **Fórmulas** dar clic en **Insertar Función,** seleccionar **DISTR.NORM.ESTAND**





Para, este ejemplo, P(Z>1,6) = 1- P(X $\leq$ 1,6)
P(X $\leq$ 1,6) SI está acumulando por tanto en el campo Acumulado se ingresa el valor de "1" (VERDADERO)

P(X $\leq$ 1,6) = 0,9452
P(X > 224) = 1- 0,9452= 0,0548

Ejemplo 5.23: Suponga que el sueldo mensual de los empleados de una fábrica se distribuye normalmente con media de 600 nuevos soles y una desviación estándar de 100 soles.

a. Calcule la probabilidad de que el sueldo de un empleado seleccionado en forma aleatoria sea menor que 500 nuevos soles
b. Si el 2,5% de los empleados con mayores ingresos deben pagar un impuesto. ¿a partir de que sueldo se debe pagar dicho impuesto?

Propiedad reproductiva de la distribución normal

Si $X_1, X_2, \cdots, X_k$ son k variables aleatorias independientes tales que $X_i \sim \mathrm{N}\left(\mu_i, \sigma_i^2\right)$ para $i = 1, 2, \cdots, k$ entonces la variable aleatoria

$$Y = \sum_{i=1}^{k} c_i X_i \sim \mathrm{N}\left(\mu = \sum_{i=1}^{k} c_i \mu_i, \sigma^2 = \sum_{i=1}^{k} c_i^2 \sigma_i^2\right)$$

donde $c_1, c_2, \cdots, c_k$ son constantes.

Ejemplo 5.24: La duración de las llamadas telefónicas de larga distancia realizadas desde una central telefónica tiene distribución aproximadamente normal con media y desviación estándar iguales a 130 segundos y 30 segundos respectivamente. Si en un día cualquiera se realizan 500 llamadas telefónicas desde esta central. ¿Cuál es la probabilidad que la duración total de estas llamadas supere las 18 horas?
Si se define la variable aleatoria Y = Duración total (en segundos) de las 500 llamadas telefónicas, entonces:

$$Y \sim \mathrm{N}\left(\mu = 500 \times 130, \sigma^2 = 500 \times 30^2\right)$$

$$Y \sim \mathrm{N}\left(\mu = 65000, \sigma^2 = 450000\right)$$

Luego, $\mathrm{P}\left[Y > 64800\right] = 1 - \mathrm{P}\left[Y \leq 64800\right] = 1 - 0,3828 = 0,6172$.

5.8 Problemas

1. La empresa *Enigma Soft* se dedica a la venta de productos y servicios informáticos. La siguiente tabla muestra el número de productos vendidos en las últimas 60 semanas:

Nº productos vendidos por semana	20	30	40	50	Total
N° de semanas	15	25	15	5	60

   Calcular el coeficiente de variación para la variable aleatoria X = Número de productos vendidos por semana. **Respuesta: CV= 28,342973**

2. Un negocio de computadoras que atiende pedidos por correo tiene cuatro líneas telefónicas. Sea la variable aleatoria X = Número de líneas ocupadas en un momento dado, cuya función de probabilidad es:

Nº de líneas ocupadas	0	1	2	3	4
f(x)	0,10	0,20	0,25	0,25	0,20

   Calcule el valor esperado y la varianza de X. **Respuesta: E(X)= 2,25 V(X) =1,5875**

3. Un laboratorio afirma que una droga causa efectos secundarios en una proporción de tres por cada 100 pacientes. Si se eligen cinco pacientes al azar a los que se les aplica la droga, calcular la probabilidad que
   a. Ningún paciente tenga efectos secundarios. **P(X = 0) = 0,8587**
   b. Al menos dos tengan efectos secundarios. **P(X ≥ 2) = 0,008472**
   c. ¿Cuál es el número de pacientes que el laboratorio espera sufran efectos secundarios debido a la droga en los próximos 500 pacientes? **E(x)= 15**

4. Suponga que en el Ejemplo 5,4 la probabilidad que un símbolo transmitido erróneamente se corrija al recibirse es 0,4. Sea la variable aleatoria X = Número de errores recibidos, una vez que el proceso de transmisión ha terminado, del total de 10. ¿Cuál es su distribución de probabilidad? **X~B(10; 0,4)**

5. Un ingeniero de control de calidad inspecciona una muestra tomada al azar de dos calculadoras manuales de cada lote que llega de tamaño 18 y acepta el lote si ambas están en buenas condiciones de trabajo de otra manera se inspecciona todo el lote y el costo se carga al vendedor. ¿Cuál es la probabilidad de que un lote se acepte sin inspección adicional si contiene cuatro calculadoras que no están en buenas condiciones de trabajo? **X~H(18; 2; 4) P(X=0) = 0,59477**

6. Como parte de una encuesta sobre contaminación del aire, un inspector decide examinar las emisiones de seis de los 24 camiones de una compañía. Si cuatro de los camiones de la compañía emiten cantidades excesivas de contaminantes, ¿cuál es la probabilidad de que ninguno de ellos sea parte de la muestra del inspector?
   X~H(24; 6; 4) P(X = 0) = 0,28797

7. La única cajera de una agencia bancaria sabe por experiencia que entre las cinco y las seis de la tarde, hora en que cierra el banco, llegan a su agencia en promedio dos personas por minuto siguiendo una distribución de Poisson. La cajera está obligada

a atender a todas las personas que llegan hasta las seis de la tarde. Tres minutos antes de las seis de la tarde no hay nadie en la cola y en ese momento ella recibe una llamada telefónica que la obliga a ausentarse de su puesto durante diez minutos. Calcular la probabilidad de que al volver a su puesto hayan más de tres personas en la cola.

X~P(6) ; P(X > 3) = 0,8488

8. Una compañía de teléfonos emplea cinco operadoras quienes reciben llamadas solicitando información general de los servicios de la compañía. Las llamadas son independientes y cada operadora recibe en promedio dos llamadas por minuto según la distribución de Poisson.
   a. ¿Cuál es la probabilidad de que en un periodo de un minuto, la primera operadora no reciba ninguna llamada? **Hay Independencia P(X = 0) = 0,27067**
   b. ¿Cuál es la probabilidad de que durante un periodo de dos minutos, menos de 3 de las 5 operadoras no reciban llamadas?

Y~B(5; 0,01831) P(Y < 3) = 0,999940

9. El número de imperfecciones superficiales en los tableros de plástico utilizados en el interior de automóviles tiene una distribución de Poisson con una media igual a 0,05 imperfecciones por pie cuadrado de tablero de plástico. Suponga que el interior de un automóvil contiene 10 pies^2 de tablero de plástico. Si se inspeccionan 10 autos, ¿cuál es la probabilidad que por lo menos 9 de ellos no tenga imperfecciones superficiales en sus tableros de plástico?

X~P(0,5) ; P(X=0) = 0,60653; Y~B(10; 0,60653) P(Y>=9) = 0,05044

10. La cantidad diaria de café (en litros) que sirve una máquina localizada en el vestíbulo de un aeropuerto es una variable aleatoria que tiene distribución uniforme en (a = 7, b = 10).
    a. Calcular la probabilidad que en un día cualquiera la cantidad servida de café de la máquina sea como máximo 8,6 litros. **X~U(7,10) P(X ≤ 8,6) = 0,53333**
    b. Calcular la probabilidad que en un día cualquiera la cantidad servida de café de la máquina sea por lo menos 7,8 litros. **X~U(7,10) P(X>=7,8) = 0,73333**
    c. ¿Cuál es la probabilidad que solo en dos de los siguientes cinco días la máquina haya servido más de nueve litros de café? Suponer independencia donde sea necesario.**X~U(7,10) P(X > 9) = 0,33333 Y~B(5, 0,3333) P(Y=2) = 0,329218**

11. La mayoría de programas de cómputo tienen una función para generar números aleatorios. En Excel se usa la función ALEATORIO para generar números aleatorios entre 0 y 1. Si la variable X representa el número aleatorio generado, el cual se sabe que se distribuye de manera uniforme entre 0 y 1, ¿cuál es la probabilidad de generar un número aleatorio con un valor mayor que su promedio más una desviación estándar.

P(X>E[X]+σ) = P(X > 0,78867513)= 0,21132487

12. Una máquina dispensadora de refresco llena vasos uniformemente, con una cantidad de líquido servido en cada vaso con una media de 200 mililitros y una desviación estándar de 2,88675 mililitros. Calcule la probabilidad de que la dispensadora vierta en un vaso más de 202 mililitros de refresco.

P(X > 202)= 0,16019107

13. La duración (en minutos) de una conversación telefónica de larga distancia nacional tiene distribución exponencial con promedio ocho minutos.
    a. ¿Cuál es la probabilidad que una llamada dure entre tres y diez minutos?

    P(3<X<10)= 0,4007845

    b. ¿Cuál es la probabilidad que una llamada dure más de 9 minutos?

    P(X>9)= 0,3246525

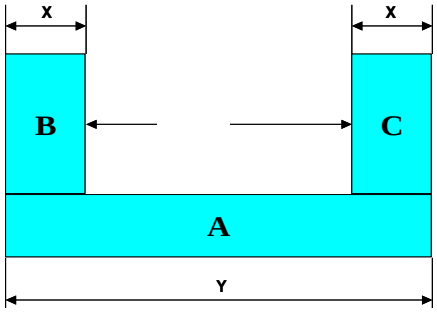
    c. ¿Cuánto debe durar como mínimo una llamada para estar considerada dentro del 10% de las de mayor duración?

    P(X ≤ k)= 0,90 => k= 18.42

    d. Si ya pasaron 5 minutos, ¿cuál es la probabilidad de que la llamada dure al menos 4 minutos más?

    P(X>=9/X>5)= (1 - P(x <= 9))/(1 - P(x <= 5)) = 0,6065

14. El tiempo de duración de cierto componente electrónico sigue una distribución normal con una media de cinco años y con una desviación estándar de un año.
    a ¿Cuál es la probabilidad que un componente elegido al azar dure más de seis años?
    b ¿Cuál es la probabilidad que un componente dure entre tres y seis años?
    c El fabricante está dispuesto a reemplazar sólo el 3% de los componentes con menores tiempos de duración. ¿Cuál es la duración que debe ofrecer el fabricante como garantía?
    d Suponga que se eligen al azar cinco componentes. ¿Cuál es la probabilidad que exactamente cuatro duren menos de siete años?

15. El consumo de los clientes de cierto restaurante es una variable aleatoria con distribución normal con una media de 25 nuevos soles y una desviación estándar de 5 nuevos soles. Se acerca fin de mes y el administrador del restaurante debe pagar una factura atrasada de 5000 nuevos soles. Suponiendo que los consumos de los clientes del restaurante son independientes. ¿Cuál es la probabilidad que el consumo de los próximos 195 clientes no sea suficiente para superar el monto de la factura atrasada?

16. Un componente en forma de "U" va a formarse con las partes A, B y C. En la figura se muestra el diagrama de las figuras ensambladas. La longitud Y de la parte A tiene una distribución normal con una media de 10 mm. y una desviación estándar de 0,1 mm. El grosor X de las partes B y C tiene una distribución normal con una media de 2 mm. y una desviación estándar de 0,04 mm. Si todas las dimensiones son independientes, ¿cuál es la probabilidad de que la abertura D sea menor que 5,9 mm.?