PROBABILIDAD Y ESTADÍSTICA

Luz Marina Pereira-González
Andrea Basantes-Andrade

2023

PROBABILIDAD Y ESTADÍSTICA

Luz Marina Pereira-González
Andrea Basantes-Andrade

2023

Autores:

PhD. Luz-Marina Pereira-González
https://orcid.org/0000-0002-7796-9819
Universidad Técnica del Norte
Docente Investigadora– Facultad de Educación, Ciencia y Tecnología
Grupo de Investigación en Educación Ciencia y Tecnología (GIECYT)

PhD. Andrea Basantes-Andrade
https://orcid.org/0000-0003-1045-2126
Universidad Técnica del Norte
Docente Investigadora – Facultad de Educación, Ciencia y Tecnología
Grupo de Investigación de Ciencias en Red (e-CIER)

Pares revisores

PhD. Pilamunga Poveda Efraín Marcelo
Universidad Técnica de Ambato
em.pilamunga@uta.edu.ec

PhD. Peñafiel Gaibor Víctor Filiberto
Universidad Técnica de Ambato
vi.penafiel @uta.edu.ec

PhD. Maryory Urdaneta Herrera
Universidad Tecnológica Israel
murdaneta@uisrael.edu.ec

Correctora de estilo
MSc. Fanny Rodas-Coloma
Docente Universidad Central del Ecuador

1a Edición 2023

ISBN: 978-9942-845-38-2

EDITORIAL UTN
IBARRA - ECUADOR

eCIER
Grupo de Investigación de Ciencias en RED
Universidad Técnica del Norte

Presentación

Este libro presenta desde una visión didáctica las nociones básicas de Probabilidad y Estadística, a fin de mejorar la comprensión de la asignatura de Estadística en la Universidad Técnica del Norte y en otras instituciones educativas.

La obra se encuentra dividida en cuatro capítulos que tienen como finalidad la incursión de los lectores en el uso de modelos de probabilidad y métodos estadísticos para analizar los datos que incidirán en la toma de decisiones. El primer capítulo inicia con algunos conceptos y terminología básica (división de la estadística, población, muestra entre otros) hasta comprender los fundamentos relacionados con la estadística descriptiva, las variables y distribución de frecuencias, la toma de datos y ordenamiento en la distribución de frecuencias, y los gráficos estadísticos.

En el segundo apartado se describe el uso de las Medidas de Tendencia Central, de Posición y de Dispersión para datos agrupados y desagrupados. El tercer capítulo, merece una atención especial, por cuanto se exponen los conceptos básicos de probabilidades, distribución binomial y distribución normal. Finalmente, el capítulo cuarto presenta el ajuste de curvas, modelos de regresión lineal simple y con datos agrupados.

Cada capítulo contiene la cantidad suficiente de ejemplos y ejercicios para que los estudiantes puedan poner en marcha la práctica teórica que esboza el contenido de este libro.

Agradecemos a los revisores pares por la examinación de la obra, sus comentarios, observaciones y sugerencias fueron de gran valía para culminar este trabajo. Asimismo, expresamos nuestro agradecimiento a las autoridades de la Universidad Técnica del Norte por el apoyo incondicional en el desarrollo investigativo y hacer posible la edición de este libro.

PhD. Luz-Marina Pereira-González
PhD. Andrea Basantes-Andrade

2023

Contenido

Capítulo 1

Introducción a la Estadística Descriptiva

CAPÍTULO I

INTRODUCCIÓN A LA ESTADÍSTICA DESCRIPTIVA

Introducción e Importancia de la Estadística

La estadística algunas veces es concebida como una ciencia y otras es considerada una disciplina, aunque todos coinciden en aceptar que es una rama de la matemática (Casella, & Berger, 2020; Giordano, & Kass, 2021; Hogg et al., 2021; Montgomery & Runger, 2022; Wasserman, 2020) que ofrece un conjunto de ideas y herramientas para tratar datos y, con base en ello, poder realizar ciertos análisis, efectuar una síntesis para presentar datos de una manera resumida y establecer una fundamentación para la toma de decisiones, Figura 1.

Figura 1
Presentación de datos para la toma de decisiones

Fuente: *Adaptado de 1455379 [Foto], de Mohamed Hassan, 30 de octubre de 2018, Pxhere,https://pxhere.com/es/photo/1455379. Creative Commons CC0.*

También podría decirse que la estadística es una rama de la matemática que proporciona métodos y procedimientos para recopilar, organizar, analizar, presentar e interpretar un conjunto de datos con el objeto de establecer conclusiones válidas y, posteriormente, realizar inferencias.

Esta última es una definición mucho más completa de lo que es la estadística, porque no solamente abarca la estadística descriptiva sino que, se extiende a la denominada estadística inferencial.

De acuerdo con lo anterior, se puede clasificar la estadística en dos grandes ramas: la estadística descriptiva, que permite recopilar, organizar, analizar, presentar e interpretar un conjunto de datos y la estadística inferencial permite establecer conclusiones válidas y realizar predicciones, se visualiza como una herramienta capaz de anticipar lo que podría suceder en el futuro, con base en un método sistemático.

Ahora ¿por qué es importante la estadística y dónde es importante? Porque donde quiera en cualquier campo del saber, en cualquier investigación que se desee realizar, uno de los métodos usuales para realizar la validación de lo investigado es, precisamente, la estadística; ya que esta rama de la matemática posee un método estructurado a través del cual la observación y el procesamiento de los datos adquiere una dimensión científica.

Lo importante de la correcta aplicación de la estadística es que permite obtener conclusiones con un determinado grado de certeza y que las condiciones de las investigaciones pueden ser replicadas, para confirmarlas o refutarlas, porque usa un método que es reconocido por investigadores a nivel internacional. En ese nivel de credibilidad que puede llegar a tener la estadística, la obtención de la muestra desempeña un rol fundamental; si se falla al momento de definir un muestreo o de tomar una muestra toda la investigación se viene abajo porque no es posible establecer conclusiones que sean válidas más allá de las unidades de estudio que se hayan empleado. Lo anterior significa que, si no se tiene una buena muestra, no se podrá obtener, de ninguna manera, un buen resultado. Hay varias técnicas que permiten seleccionar una muestra; pero, ¿qué es una muestra?

De manera intuitiva se puede afirmar que una muestra es una cantidad pequeña de un todo, que es separada a través de un método, previamente establecido y que es capaz de garantizar que esa pequeña cantidad va a exhibir características parecidas a la totalidad. Estadísticamente, puede definirse

una muestra como un subconjunto de la población; pero no un subconjunto cualquiera, sino uno que sea representativo de la población. Para poder obtener esa muestra, se deben seguir una serie de pasos que permiten garantizar que la obtención de esa muestra sea correcta desde el punto de vista estadístico. Lo importante a considerar será: el tamaño y el tipo; ya que no cualquier muestra nos garantiza que ella sea válida. Si se toman muestras muy pequeñas, es posible que estas no sean representativas de la población. En teoría, mientras mayor sea la muestra, es mejor; pero en la práctica las muestras más grandes requieren mayor inversión. Por ejemplo, en el caso de una encuesta, es más costoso aplicarla a 3000 personas que a 200. Es por ello que resulta importante que se pueda definir un tamaño adecuado y válido para poder obtener resultados que permitan sustentar algunas inferencias.

Adicionalmente, todo estudio estadístico involucra un nivel de tolerancia que se refiere a la magnitud del error estamos dispuestos a aceptar en el resultado obtenido. Esta tolerancia suele estar relacionada con la rama del conocimiento en el cual se ubique el estudio que se desea realizar. No es lo mismo desarrollar un estudio en publicidad o en comunicación, en el que se desea conocer, por ejemplo, cuáles son las preferencias de una cierta población objetivo en relación a algunos canales de televisión, a que se vaya a realizar una investigación a cerca de la prevalencia de las complicaciones graves asociadas a las diferentes vacunas contra la COVID-19; porque está claro que no debe existir un nivel de tolerancia alto cuando se trata de un riesgo de vida.

De manera general, en algunas investigaciones sociales puede admitirse hasta un 10% de error, en tanto que cuando se realizan investigaciones -fundamentalmente- en salud se pueden encontrar niveles de tolerancia máximos del 1%; mientras más pequeño es el error que el investigador está dispuesto a tolerar, más exactitud tiene el estudio.

La importancia de la estadística también reside en que permite manejar una gran cantidad de información y que ésta, a su vez, se puede presentar de una forma resumida. Lo anterior implica que cualquier persona, a través de un gráfico, puede tener una visión panorámica del comportamiento del estudio que hemos realizado.

Sin duda, la estadística constituye una poderosa herramienta cuando se realizan investigaciones pero también puede ser utilizada para distorsionar la

información o para manipular los resultados. Supongase, por ejemplo, que un artículo de opinión presentado en un periódico se titula "Los adultos mayores responden más rápido los correos electrónicos" y que el sustento presentado sean los resultados de una encuesta en la que se preguntaba a las personas ¿Considera usted que los mensajes de correo electrónico deben ser contestados de inmediato?

El resultado obtenido revelaba que las personas con 65 años o más respondían afirmativamente a la pregunta en más de un 38%, en tanto que en las personas más jóvenes los porcentajes eran mucho menores, Figura 2.

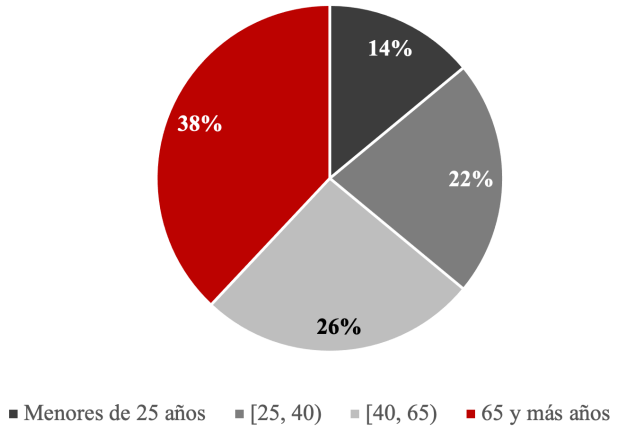
Ahora bien, el hecho de que alguien considere que los correos electrónicos deben ser contestados de inmediato no significa que esas personas respondan los emails de inmediato. En ese caso la conclusión a la que se pretende llegar, no tiene sustento en la estadística sino en una inadecuada interpretación de los resultados.

La estadística es muy antigua; sus primeros vestigios fueron encontrados en la isla de Cerdeña, en Italia, en registros de la cantidad de alimento de la que podían disponer, por ejemplo, a través de la caza. Posteriormente, se encontraron en Mesopotamia algunos monumentos que indican claramente que allí existió una forma rudimentaria de estadística que permitía contar los nacimientos, las muertes y los matrimonios. Sin embargo, el surgimiento de la estadística tuvo fines estatales; es decir, proporcionar a los Estados algún tipo de control para poder realizar censos, cobros de impuestos, así como para diagnosticar cuántas personas habitan en una determinada zona y para poseer información sin errores de nacimientos, muertes, el stock de medicinas, alimentos, y en todos los tipos de servicios que debe atender un Estado; en suma, el surgimiento de la estadística fue inducido por la necesidad que siempre se ha tenido de procesar y de recoger información.

En las mediciones siempre se presenta algún tipo de variabilidad, ello, por una parte, trae consigo errores de medición y, por la otra, establece diferencias entre un individuo y otro. El mismo proceso de realizar la medición puede ocasionar que exista algún tipo de variabilidad. Ante esta situación, la estadística descriptiva, hace posible, por ejemplo, que se pueda escoger un valor que va a ser representativo de toda la población y que corresponde a las denominadas medidas de tendencia central.

Figura 2

Distribución de porcentajes de la respuesta a ¿considera usted que los mensajes de correo electrónico deben ser contestados de inmediato?



Nota. Elaboración propia a partir de gráfico de Microsoft Office 365.

División de la Estadística

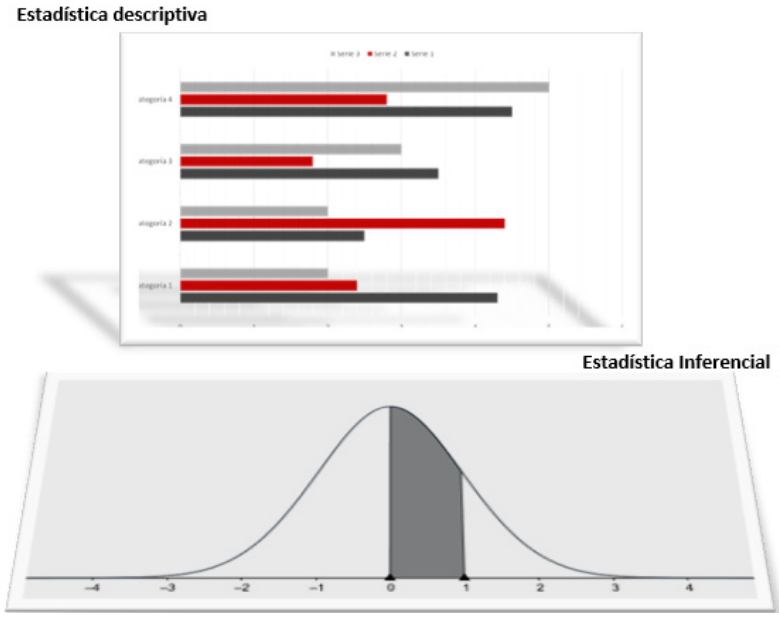
Como se indicó inicialmente en la definición, la estadística se divide en dos grandes ramas. La primera, se conoce como estadística descriptiva, que es aquella que se encarga de recoger, procesar, organizar, categorizar, analizar y presentar una determinada información. Describir, el mismo nombre lo dice, esta rama de la estadística no puede ir más allá de lo que connota el mismo proceso descriptivo: señala características, indica cómo son los datos.

Si por ejemplo tenemos un grupo de estudiantes, la estadística descriptiva, podría suministrar información, acerca de cómo es la distribución de hombres y mujeres en el grupo, cuántos son afroecuatorianos, mestizos, indígenas, blancos; así como cuántos tienen entre 21 y 30 años y cuántos entre 31 y 40, cuál es la persona más joven y la de mayor edad; quiénes están realizando una primera carrera y cuántos están cursando la segunda; cuántos ven la asignatura por primera vez y cuántos tienen segunda matrícula. Y así se podría continuar categorizando, a ese grupo de estudiantes en relación a todos los criterios que resulten relevantes para la investigación que se desea realizar.

La estadística inferencial, por su parte, forma parte de la estadística moderna, es completamente distinta a la estadística descriptiva, porque con la estadística inferencial a partir del estudio de unamuestra, bien definida,

se pueden extrapolar los resultados a la población completa; pero para ello, es absolutamente necesario que esa muestra se encuentre profundamente relacionada con la población, que exhiba sus características esenciales, es decir, que esa sea una muestra representativa de esa población, Figura 3.

Figura 3
División de la Estadística



Nota. Elaboración propia a partir de gráficos de Microsoft Office 365.

En estadística descriptiva se va a tratar el cálculo de las medidas de posición, medidas de dispersión y medidas de forma. Las medidas de posición pueden ser de tendencia central y no central. De tendencia central se tienen la media, la mediana y la moda, y de tendencia no central, los cuartiles, los percentiles y los deciles.

En medidas de dispersión se encuentran el rango, la desviación típica, la varianza y el coeficiente de variabilidad, y en las medidas de forma se estudiará, la asimetría y la curtosis que indican la forma que tiene la distribución con la que se está trabajando.

En cuanto a la estadística inferencial, existen dos elementos fundamentales: la estimación, que puede ser puntual o por intervalos, y las pruebas de hipótesis.

Población

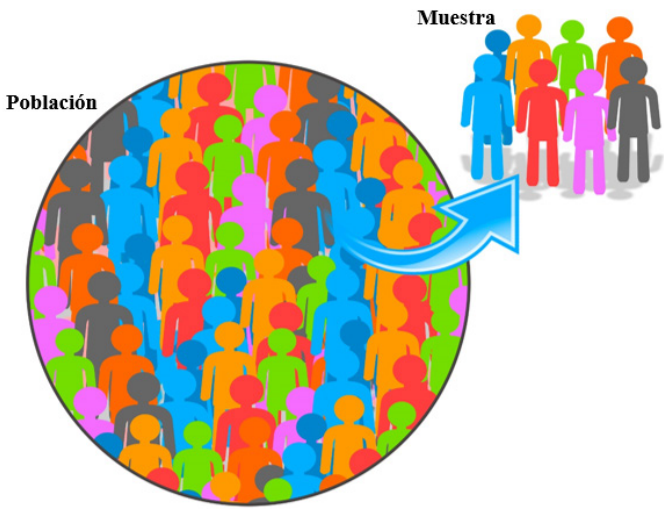
Corresponde al universo, es decir al conjunto de todos los elementos que tienen una propiedad común y que se someten a un estudio estadístico. De las poblaciones se extraen las muestras.

Las poblaciones pueden ser finitas e infinitas. Se dicen finitas cuando se puede contar el número de sus elementos y son infinitas cuando no es posible contar el número de sus elementos o cuando existen tantos elementos en ella que podemos asimilar ese tamaño al infinito. Ejemplo de una población finita podría ser un paralelo de una determinada asignatura; y de una población infinita, el número de personas menores de 30 años a nivel mundial o el número de granos de arena que existen en el mar, Figura 4.

Muestra

Se define como un subconjunto representativo de la población, figura 4, esto significa que en ese grupo de elementos, personas, animales o cosas que constituyen la muestra, estén contenidas las diferentes características que se pueden encontrar en la población.

Figura 4
Población y muestra



Nota. Elaboración propia a partir de Vector de la imagen del ícono de población, de publicdomainvectors.org, Openclipart, https://n9.cl/qeomy, Dominio público.

Las muestras pueden ser probabilísticas y no probabilísticas, las primeras son aquellas en las que cada uno de los elementos de la población tiene la misma posibilidad de ser elegido.

En la representatividad de la muestra se pueden encontrar tres elementos fundamentales: un marco y diseño muestral adecuado, una selección al azar de los elementos constitutivos (muestra probabilística) y un control riguroso del procedimiento de muestreo.

Variables y Distribución de Frecuencias

Variables y Constantes

Una variable es una característica, de interés para la investigación y que puede tomar diferentes valores, bien sea entre los elementos de una determinada población o en las diferentes condiciones en las que puede realizarse un experimento. Un ejemplo puede ser el color de los ojos: azules, grises, verdes, negros, café, etc. podría decirse entonces, que toda variable constituye una forma de resumir la información.

Las constantes, por su parte, están constituidas por las características que permanecen inmutables a lo largo del tiempo o a través de las diferentes condiciones de experimentación. Un ejemplo de constante es la cantidad de cromosomas que tiene un determinado individuo o su huella dactilar.

Tipos de Variables

Las variables pueden ser cualitativas o cuantitativas.

Las variables cualitativas son aquellas que se identifican con una determinada característica o cualidad; ejemplo, la fruta favorita de una persona.

Las variables cuantitativas, por su parte, son aquellas que pueden cuantificarse, esto es, identificarse con un número o ser representadas por una cantidad; ejemplo: el peso o la altura de un individuo.

No obstante, existen algunas variables que, dependiendo de la forma en que son medidas, pueden llegar a ser cualitativas o cuantitativas; ejemplo de ello sería la presión arterial de una persona que es una variable cuantitativa y puede ser medida en milímetros de mercurio; pero a esta variable también podemos asignarle una escala de tal forma que resulte cualitativa, este sería el caso de clasificarla como presión arterial alta, normal o baja.

Existen dos tipos de variables cuantitativas, discretas y continuas. Las variables cuantitativas son discretas cuando los valores que toman únicamente pueden ser enteros; y son continuas cuando puede tomar cualquier valor entre dos números enteros consecutivos.

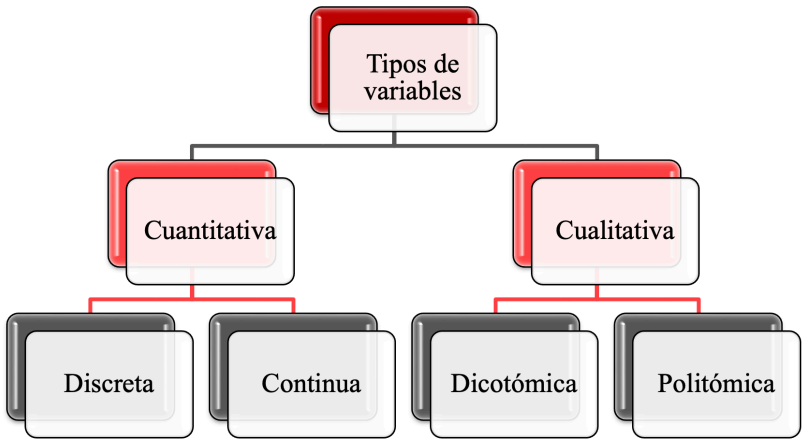
Ejemplos de variables discretas son por ejemplo el número de hermanos o la cantidad de títulos que puede tener una persona.

Variables continuas, son, por ejemplo, el peso en kilogramos de un individuo o su estatura.

Las variables cualitativas también son conocidas con el nombre de categóricas. Este tipo de variables, puede, a su vez, ser dicotómicas o politómicas.

Se dice que son dicotómicas cuando solamente pueden adquirir dos valores y son politómicas cuando la variable puede asumir más de dos valores. Ejemplo de variable dicotómica tenemos el sexo, de las personas: hombre o mujer; y ejemplo de variable politómica sería el nivel de satisfacción de un cliente que ha recibido un determinado servicio: muy satisfecho, medianamente satisfecho, poco satisfecho, Figura 5.

Figura 5
Tipos de variables



Nota. Elaboración propia a partir de SmartArt de Microsoft Office 365.

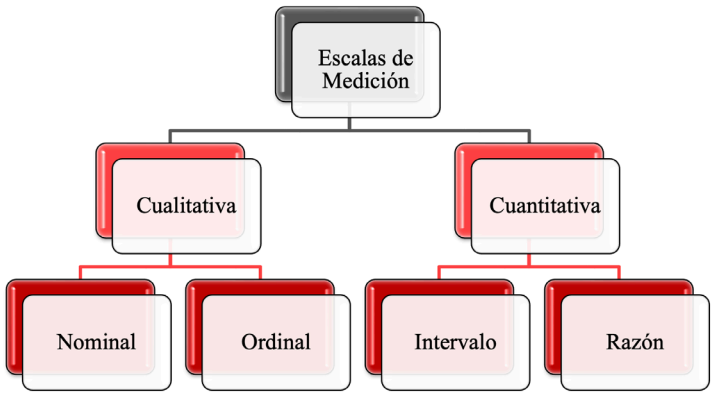
Tal como se indicó anteriormente, las variables a veces las podemos clasificar como cuantitativas o cualitativas y ello va a depender de la escala de medición que estemos utilizando. Un ejemplo de ello sería las pruebas de COVID 19. Hay pruebas cualitativas y cuantitativas.

Si realizamos una prueba de antígenos, con las cualitativas podemos saber si se tienen o no se tienen anticuerpos, en ese caso tendríamos una variable cualitativa dicotómica. Sin embargo, si lo que se desea es saber cuál es la cantidad de anticuerpos presentes en la sangre, entonces se debe utilizar una prueba cuantitativa que suministrará, en números, la cantidad exacta de anticuerpos que se tienen.

Medición de una Variable

La medición es, por definición, el grado de precisión con el que somos capaces de expresar el número o categoría asignado a una determinada variable. En el caso de variables cualitativas, un ejemplo que resulta claro es la escala utilizada para medir el dolor. Es muy difícil que una persona externa, a través de una observación simple, pueda calibrar la intensidad del dolor que puede sentir un individuo, ya que el dolor es completamente subjetivo. El tipo de variable y la escala de medición van a determinar el método estadístico que se va a emplear, de allí la importancia de poder distinguir cuando una variable es cualitativa, cuándo cuantitativa y en qué escala se va a medir. Existen cuatro escalas de medición: a) nominal, b) ordinal, c) de intervalo y d) de razón, Figura 6.

Figura 6
Escalas de medición para variables

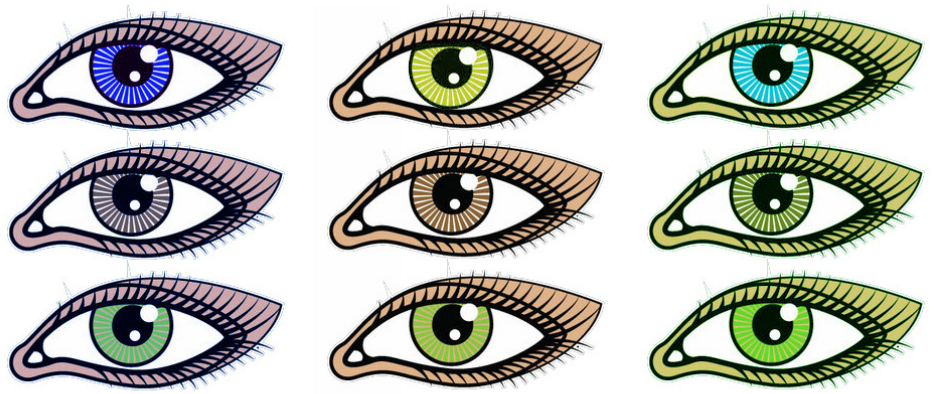


Nota. Elaboración propia a partir de SmartArt de Microsoft Office 365.

Escala Nominal

Corresponde a un tipo de medición en la cual los nombres, funcionan como etiquetas o como atributos. Por ejemplo, una etiqueta en la variable color de los ojos sería “azul” otra sería “café”, otra etiqueta sería “verde”, otra “gris”, y así se podría seguir asignando colores para tratar de cubrir toda la gama posible. Al momento de procesar los datos, es conveniente codificarlos, esto significa que a cada una de esas etiquetas se le debe asignar un número. Dicho número, contrario a lo que ocurre en el caso de las variables cuantitativas, no representa una cantidad, sino una nueva etiqueta, arbitraria, que se usa para identificar los diferentes valores que puede tomar una variable. Para el color de los ojos, un primer investigador; por ejemplo, podría asignar el número 1 a los ojos azules y el 2 a los de color violeta; y un segundo investigador podría seleccionar la etiqueta “1” para el color de ojos café y “2” para los de color celeste, Figura 7. Cuando a las variables se les puede asignar esa etiqueta o esa cualidad, que no refleja en forma alguna una cantidad, entonces se dice que la variable está medida en escala nominal.

Figura 7
Ejemplo de escala nominal



1. Azul 2. Violeta 3. Verde azulado 4. Amarillo verdoso 5. Café
6. Verde claro 7. Celeste 8. Verde mar 9. Verde lima

Nota. Elaboración propia a partir de la imagen Ojos de colores, de publicdomainvectors.org, Openclipart, https://n9.cl/3nmxm, Dominio público.

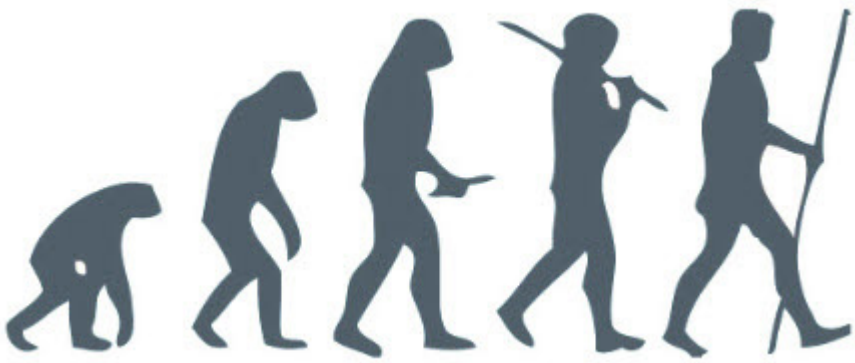
Escala Ordinal

Por su parte, la escala ordinal también corresponde a valores cualitativos, pero éstos no son arbitrarios, como en el caso de la escala nominal, sino que obedecen a un cierto orden que representa niveles o grados de expresión de la variable, pudiendo darse este orden de forma creciente o decreciente. Se puede encontrar un ejemplo de este tipo de escala en el nivel socioeconómico de una persona, pudiendo éste clasificarse en alto, medio y bajo.

Otro ejemplo de escala ordinal se encuentra en el nivel de educación alcanzado por una persona, pudiendo ser, en orden ascendente, por ejemplo: sin educación; primaria, secundaria, universitaria, maestría, doctorado y posdoctorado.

En la Figura 8 se ilustra una escala ordinal que pudiera utilizarse para representar la evolución del hombre.

Figura 8
Ejemplo de escala ordinal para la evolución del hombre



Nota. Adaptado de Human Evolution Scheme, de José Manuel Benitos, 14 de noviembre de 2009, Wikimedia Commons, https://n9.cl/vgxt0, Creative Commons CC0.

Las variables cuantitativas, por su parte, pueden tener dos escalas de medidas: de intervalo y de razón. En este caso, es importante distinguir la diferencia entre una y otra escala.

Escala de Intervalo

La característica principal de una escala de intervalo es que no se tiene un cero convencional, sino que este es asumido de manera arbitraria, Figura 9. Un ejemplo de variable medida en escala de intervalo es la temperatura ambiental. En el caso de la medición en grados Celsius, el cero fue asignado por convención al punto de fusión del agua, pero ello no significa que no haya temperatura, sino representa una convención que delimita el punto en el cual el agua pasa del estado líquido al estado sólido, Figura 10.

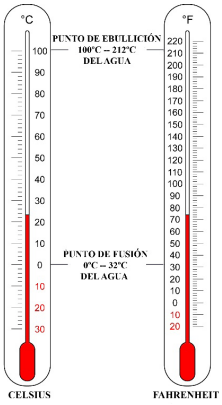
Figura 9

Características de una escala de intervalo

Nota. Elaboración propia a partir de SmartArt de Microsoft Office 365.

Figura 10

Ejemplo de escala de intervalo



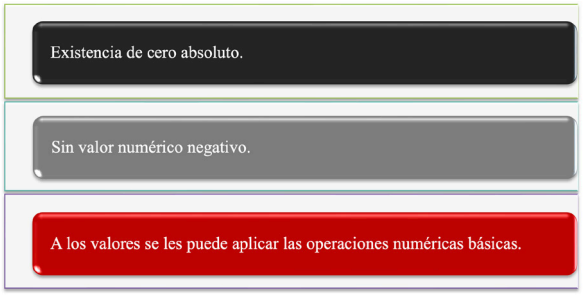
Nota. Adaptado de Thermometer, 1 de abril de 2021, Wikimedia Commons, https://n9.cl/d2h2c, Creative Commons CC0.

Escala de Razón

Cuando la medida es en escala de razón el cero significa ausencia de cantidad, es decir que ese valor tiene un significado real, lo anterior significa que en una escala de razón no pueden existir números negativos, ya que no podríamos decir que un objeto tiene un peso de -4 kilos o que un determinado volumen es de -15 litros, puesto que ninguna de esas afirmaciones tendría un sentido físico. Otra característica importante es que a los valores de la escala de razón se le puede aplicar operaciones numéricas básicas: suma, resta, multiplicación y división, Figura 11. Por ejemplo: si tenemos 5 kg, este valor lo podemos multiplicar por dos y obtendríamos 10 kg, Figura 12.

En la Tabla 1 se presentan algunos ejemplos de diferentes tipos de variables y sus escalas de medición.

Figura 11
Características de una escala de razón



Nota. Elaboración propia a partir de SmartArt de Microsoft Office 365.

Figura 12
Ejemplo de escala de razón

Nota. Elaboración propia a partir de imágenes de Openclipart, https://publicdomainvectors.org/es/vectoriales-gratuitas, Creative Commons CC0.

Tabla 1

Ejemplos de tipos de variables y sus escalas de medición

Variable	Tipo	Escala de Medición
Sexo de una persona	Cualitativa dicotómica	Nominal
Estado civil	Cualitativa politómica	Nominal
Escala de dolor	Cualitativa politómica	Ordinal
Temperatura de un cuerpo	Cuantitativa continua	Intervalo
Número de hermanos	Cuantitativa discreta	Razón
Altura de un niño	Cuantitativa continua	Razón

Toma de datos y ordenamiento: Distribución de frecuencias

Presentación de Datos: Tablas de Frecuencia

Una forma en la que se suelen presentar los datos es a través de las denominadas Tablas de frecuencia, también conocidas como distribución de frecuencias.

Las Tablas de frecuencias constituyen una forma de presentar los datos de manera resumida y se utilizan para variables cuantitativas como para variables cualitativas, con la salvedad de que estas últimas deben haber sido medidas en escala ordinal.

Arreglo Ordenado de Datos

Un concepto importante que debe manejarse en una distribución de frecuencias es el de arreglo ordenado de datos.

Considérese el siguiente conjunto de datos:

10 18 3 11 7 20 16

Un arreglo de números puede ser ordenado de dos formas: ascendente, ordenando los números de menor a mayor. En este caso, el resultado sería:

3 7 10 11 16 18 20

Sin embargo, esos datos también pueden ser ordenados de una manera descendente, es decir, comenzando por el número más grande hasta llegar al menor:

20 18 16 11 10 7 3

El resultado, en cualquiera de los dos casos, se conoce con el nombre de arreglo ordenado de datos.

Rango de un Arreglo Ordenado de Datos

El rango de un arreglo ordenado se define como la diferencia entre el número mayor, que en este caso es 20, y el número menor, que es 3. Este rango proporciona una idea acerca de la variabilidad de los datos; esto es, mientras mayor sea el valor del dato, los datos variarán en un rango mayor.

Desde el punto de vista práctico, los arreglos ordenados de datos resultan útiles cuando se tiene un número pequeño de datos; ya que, si se tienen muchos datos, puede llegar a resultar engorroso el manejo de la información. En este último caso, es preferible procesar la información agrupando los datos en intervalos.

Datos Agrupados y Datos no Agrupados

Los datos no agrupados son aquellos que se presentan como un conjunto, ordenado o no, de datos. En ellos es común que los valores que asume la variable no se repitan o, al menos, no lo hagan con mucha frecuencia. Este tipo de datos no ha recibido ningún tipo de tratamiento, sino que, por el contrario, responden al estado original en el que fueron levantados durante el proceso de investigación.

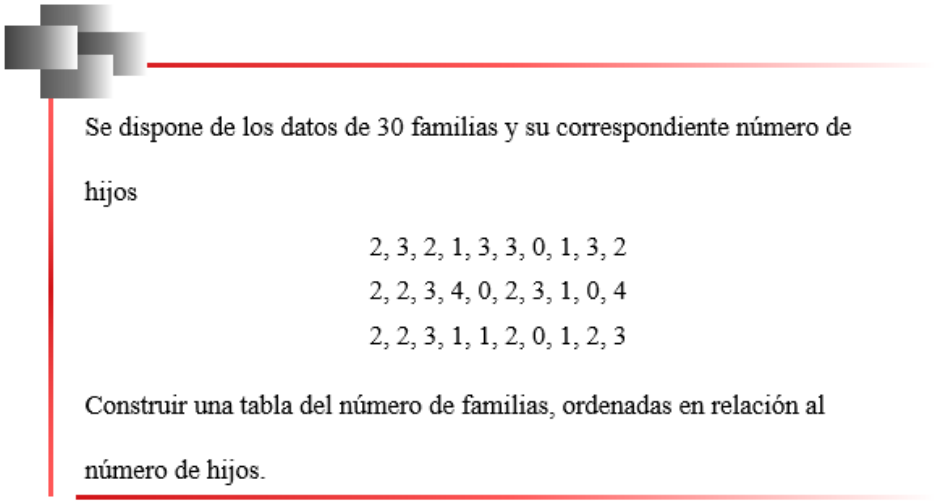
Los datos agrupados, por su parte, son aquellos que presentan clasificados según un determinado criterio. Generalmente los datos se agrupan cuando se tiene una gran cantidad de observaciones o si la variable es cuantitativa continua, ya que el estar clasificados en clases o en intervalos facilita tanto el procesamiento de los datos como su debida interpretación.

Distribución de Frecuencias para Datos no Agrupados

En la Figura 13, se presenta el enunciado correspondiente a los datos obtenidos en una encuesta realizada a un grupo de 30 familias. A los cabezas de familia se les preguntó cuál era el número de hijos que tenían. En este ejemplo se está en presencia de una variable *x* que se define como “número de hijos”.

Obsérvese que se trata de una variable cuantitativa discreta, porque la variable x sólo puede tomar valores enteros. Adicionalmente, se sabe que la variable ha sido medida en escala de razón, puesto que el cero está relacionado con una característica de ausencia: "no tienen hijos". Además, el rango, que constituye una medida de la variabilidad de los datos, es igual a cuatro, dado que el menor número de hijos, en las familias encuestadas, es cero y el mayor número de hijos es cuatro y el rango se define como la diferencia entre el valor máximo registrado en los datos y el valor mínimo.

Figura 13
Datos correspondientes al número de hijos en 30 familias encuestadas



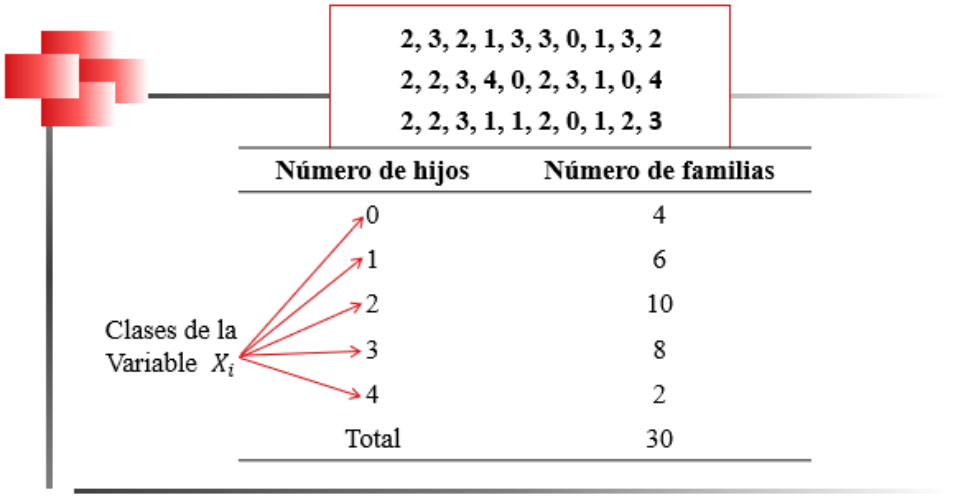
Los datos recolectados no presentan, hasta el momento, ningún tipo de tratamiento. En ese caso, se dice que se tienen datos no agrupados. Sin embargo, un análisis ligero de los datos revela que varias de estas respuestas se encuentran repetidas, esto es, que hay varias familias que tienen exactamente el mismo número de hijos.

Se puede entonces agrupar estos resultados en clases que correspondan a los distintos valores que puede tomar la variable en estudio, esto es, 0, 1, 2, 3 y 4.

En la Figura 14, se presenta el recuento de la cantidad de familias que tienen el mismo número de hijos. Los datos ubicados en la parte superior permiten verificar que hay 4 familias sin hijos y dos familias que han tenido 4 hijos.

Figura 14

Recuento del número de familias en atención a los distintos valores de las clases



Anteriormente se había señalado que la variable ***x*** correspondía a la descripción "número de hijos", y el recuento del número de familias, en una distribución de frecuencias, recibe el nombrede frecuencia absoluta.

Frecuencia Absoluta

Corresponde al número de veces que se repite un determinado evento en una muestra o en un experimento. La frecuencia absoluta se designa con una letra minúscula "***f***" a la cual se le coloca un subíndice, ***i***, para identificar que corresponde a la frecuencia absoluta de la clase ***i***.

En la Tabla 2, se presenta la distribución de frecuencia correspondiente a los datos del ejemplo. El número de familias en cada clase, corresponde a la frecuencia absoluta. Este tipo de distribución se utiliza, generalmente, cuando hay pocos datos y la variable en estudio es discreta.

Tabla 2

Distribución de frecuencias absolutas para la variable número de hijos

Xi	fi
0	4
1	6
2	10
3	8
4	2
Total	30

Frecuencia Relativa

En una distribución de frecuencias, la frecuencia relativa corresponde a la proporción de la frecuencia absoluta de una determinada clase ***i***, en relación al número total de datos que existen en la muestra. La frecuencia relativa se designa con la letra minúscula "***h***", siendo entonces h_i lafrecuencia relativa de la clase ***i***.

Frecuencia Absoluta Acumulada y Frecuencia Relativa Acumulada

Las frecuencias acumuladas se designan a través de letras mayúsculas y son dos, una corresponde a las frecuencias absolutas y la otra a las frecuencias relativas, y se identifican como. F_i y H_i, respectivamente. Para obtener las frecuencias acumuladas de una determnada clase i, basta con sumar a la frecuencia de la clase i, las frecuencias correspondientes a las clases precedentes.

Tanto la frecuencia relativa, h_i, como la frecuencia relativa acumulada, H_i, pueden ser expresadas en porcentaje; para ello, basta multiplicar por 100 el valor decimal correspondiente a la proporción.

En la Tabla 3, se presenta la distribución completa de frecuencias del ejemplo en estudio. Se observa que el valor de $h1$, es decir la frecuencia relativa correspondiente a la primera clase (número de hijos igual a cero), se obtiene al dividir la frecuencia absoluta de la primera clase, $f1$, entre el número total de familias encuestadas (30).

Resulta claro que el valor correspondiente a F_1, por estar en la primera clase, debe ser igual al valor de f_1, puesto que no existen clases anteriores. En cambio, para F_2, resulta de sumar el valor de $f_2 = 4$ con el de $f_1 = 6$.

Tabla 3

Distribución de frecuencias para la variable número de hijos

X_i	f_i	h_i	F_i	H_i
0	4	0,133	4	0,133
1	6	0,200	10	0,333
2	10	0,333	20	0,667
3	8	0,267	28	0,933
4	2	0,067	30	1,000
Total	30	1,0		

El cálculo de la frecuencia relativa acumulada H_i puede hacerse de dos formas: 1) como el cociente entre F_i y el número total de observaciones, o 2) sumando al valor de la frecuencia relativa de la clase i, las frecuencias relativas de todas las clases precedentes. Así, por ejemplo, para obtenera H_3, es necesario sumar a h_3 las frecuencias relativas de las clases 1 y 2, es decir,

$$H_3 = h_1 + h_2 + {}_h3 \qquad 1.1$$

Obsérvese que el valor en la penúltima fila de la columna de frecuencias absolutas acumuladas, F_i, de la Tabla 3, debe ser igual al número de elementos en la muestra, observaciones o mediciones realizadas. De manera similar, la penúltima fila de la columna de frecuencias relativas acumuladas, debe ser igual a la unidad, o si se encuentra expresada en porcentaje, debe ser igual al 100%.

Por otra parte, el hecho de que la Tabla de distribución de frecuencias presente los valores de la variable de manera ordenada, no es casual. Anteriormente se había señalado que la distribución de frecuencias constituía una forma de presentar, en forma resumida, la información obtenida en una investigación. Los cálculos realizados en cada una de las columnas de

la Tabla de distribuciónde frecuencias proporcionan la base para sustentar algunas interpretaciones.

Por ejemplo, si se quiere determinar el número de familias que tienen a lo sumo (como máximo) dos hijos, bastará con leer la frecuencia absoluta acumulada que corresponde a la clase cuyo valor de la variable número de hijos es igual a 2 (se debe tomar en cuenta que este valor acumula el número de hijos igual a cero, a uno y a dos). De igual forma, si en la misma fila se lee el valor de la frecuencia relativa acumulada (multiplicada por 100), se puede concluir que el 66,7% de las familias tienen, cuando mucho, dos hijos.

Otra interpretación que resulta interesante es dónde se encuentra la menor frecuencia relativa, ya que ella indicará en que clase se registra el menor porcentaje. En este caso, ese valor corresponde a la clase 5, es decir a aquellas familias que tienen cuatro hijos.

En este caso, la interpretación general que puede darse es que las familias entrevistadas no tienen un gran número de hijos, siendo tan sólo un 6,7% de ellas las que tienen cuatro hijos.

Distribución de Frecuencias para Datos Agrupados

Cuando se tiene un gran número de datos o se trata de variables cualitativas continuas, es conveniente agrupar dichos datos en intervalos o clases, ya que esto facilita la interpretación de los resultados.

Para el cálculo de los intervalos en la distribución de frecuencias, se deben seguir algunos pasos que ayudan a organizar el procedimiento:

1. Hallar valores extremos (máximo y mínimo)
2. Calcular el rango

$$R = \text{máximo} - \text{mínimo} \qquad 1.2$$

3. Calcular el número de intervalos, seleccionando el método adecuado para la obtención de las clases.

   Método de Sturges: Este método (Sturges, 1926) consiste en dividir el rango de los datos en *m* clases, donde *m* se determina mediante la fórmula

$$m = 1 + 3.322 \, .log(n) \qquad 1.3$$

donde n es el tamaño de la muestra. Una vez calculado *m*, se define la amplitud de cada clase como el rango dividido por m, y se van construyendo las clases de tal manera que la diferencia entre el límite superior e inferior de cada clase es igual a la amplitud.

Método de la raíz cuadrada: Consiste en dividir el rango de los datos en *m* clases, donde *m* se determina mediante la fórmula

$$m = \sqrt{n} \qquad 1.4$$

siendo n el tamaño de la muestra. Una vez calculado *m*, se define la amplitud de cada clase como el rango dividido por *m*, y se construyen las clases igual que en el método de Sturges, resultando de esta manera intervalos de igual amplitud. Este método no se atribuye a ningún autor en particular. Es un ejemplo de una regla empírica más general que sugiere que el número adecuado de intervalos puede estar relacionado con la raíz cuadrada del tamaño de la muestra. La fórmula se utiliza a menudo cuando se desconoce la distribución subyacente de los datos y se busca una forma rápida y fácil de determinar el número de intervalos. Es apropiada cuando el tamaño de la muestra es pequeño (Scott & Scott, 1992). En general, los resultados de los métodos de Sturges y de la raíz cuadrada son similares cuando la muestra es inferior a 50 datos.

4**. Método de Freedman-Diaconis**: Este método (Freedman & Diaconis, 1981) es similar al método de Sturges, pero en lugar de utilizar el logaritmo de la muestra, se utiliza la distancia intercuartil (Q_3- Q_1). Se define la amplitud de cada clase como:

$$a = \frac{2(Q_3 - Q_1)}{m^{1/3}} \qquad 1.5$$

donde *m* es el número de clases determinado mediante el método de Sturges.

5**. Método manual:** Este método consiste en determinar las clases de manera subjetiva, considerando el contexto específico de los datos y las características de la distribución. Por ejemplo, se pueden construir clases que representen rangos de valores significativos o que agrupen valores similares, en atención a los datos que se hayan levantado.

Es importante tener en cuenta que la elección del método para formar las clases en distribuciones intervalares puede afectar la interpretación de los resultados y la precisión de las conclusiones que se pueden

extraer. Por lo tanto, es recomendable revisar con detalle el sustento teórico de cada método para la construcción de clases y considerar la precisión requerida en el contexto específico en el que se está trabajando.

Es importante acotar que cuando el cálculo de intervalos se realiza obteniendo la raíz del número de datos, a medida que se tienen muestras más grandes, el número de intervalos se incrementa demasiado. Cuando se tiene un número de intervalos muy alto, se corre el riesgo de que algunas clases queden vacías, esto es, sin datos que pertenezcan a algún intervalo. Debido a ello, para efectos prácticos de este texto, se utilizará la fórmulade Sturges para el cálculo de intervalos.

El número de intervalos, calculado a través de la fórmula de Sturges, suele aproximarse siguiendo una regla simple que consiste en tomar en cuenta sólo el número entero que resulta del cálculo de la fórmula. Si dicho número es par, se redondea al inmediato superior; pero si el número es par, se toma directamente éste como el número de intervalos. En la Tabla 4 se incluyen algunos ejemplos del redondeo de la fórmula de Sturges.

Por otra parte, la amplitud, a, del intervalo se calcula coma la relación entre el rango y el número aproximado de intervalos, m´. Esto es,

$$a = \frac{R}{m'} \qquad 1.6$$

Donde R es igual al rango y m´ igual al número de intervalos.

Cuando el valor de la amplitud no resulta exacto, se debe calcular una amplitud real, a´, la cual siempre se redondea por exceso, es decir, al número inmediato superior que contenga tantos decimales como tengan los datos originales.

El número de decimales a tomar en el redondeo de la amplitud de un intervalo, depende del contexto en el que se esté trabajando y de la precisión requerida en los resultados.

En el caso de una variable discreta, donde los valores posibles son enteros, se suele redondear la amplitud a la unidad más cercana, o al entero más cercano si se desea una mayor precisión. Por ejemplo, si se tiene una variable que puede tomar valores enteros entre 0 y 10, y se desea construir intervalos de amplitud 2, se pueden definir los intervalos [0,2), [2,4), [4,6), [5,8), y [8,10].

En el caso de una variable continua, donde los valores posibles son infinitos, y pueden tomar cualquier valor en un rango determinado, la amplitud del intervalo se puede redondear de acuerdo a la precisión que se requiera en los resultados. En general, se recomienda redondear la amplitud a una cifra significativa de la escala de la variable. Por ejemplo, si se está trabajando con una variable de temperatura en grados Celsius, se puede redondear la amplitud a una décima o una centésima de grado Celsius, dependiendo de la precisión que interese al investigador.

La razón que justifica esta forma de realizar los redondeos de los cálculos es asegurar que, por un lado, el número de intervalos siempre resulte impar, lo que puede contribuir, como se verá más adelante, a que la distribución sea normal; y, por el otro, que la amplitud de los intervalos sea suficiente para que ningún dato vaya a quedar por fuera.

Tabla 4

Ejemplos de cálculo del número de intervalos según formula de Sturges

n	Nº de intervalos calculado por fórmula	Nº de intervalos considerados
25	5,3	7
45	6,49	7
60	6,9	7
82	7,4	7
160	8,3	9

En la Figura 15, se presentan los datos y el cálculo del rango en el conjunto valores de un ejemplo correspondiente a una variable cuantitativa continua. Siguiendo el procedimiento anteriormente descrito para la obtención de los intervalos, se tendrán los siguientes resultados:

Valor máximo = 29,85;

Valor mínimo = 10,01 Número de datos, n = 40

Rango = 29,85 – 10,01 = 19,84

Número de intervalos, m = 1 + 3,322 * log (40) = 6,32

Número de intervalos redondeado por defecto, m´ = 7

En este caso, si aplicamos la fórmula empírica, por tratarse de una muestra pequeña, se tendría:

$$m = \sqrt{40} = 6.32$$

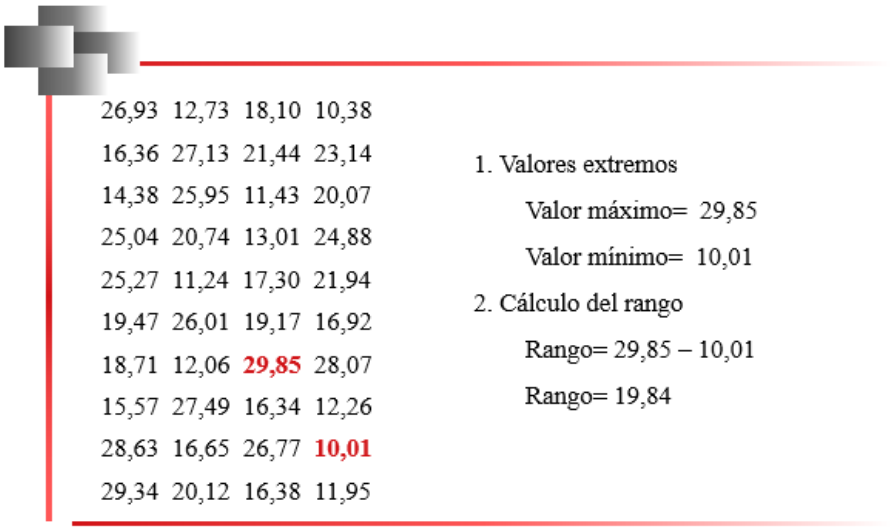
Obsérvese que tanto la Fórmula de Sturges como la empírica, proporcionan el mismo resultado.

Amplitud, a = 19,84 / 7 = 2,834;

Amplitud redondeada por exceso, a´ = 2,84. En este caso, se justifica el redondeo a dos decimales en la amplitud, porque ésta es la precisión que se tiene en los datos registrados.

Figura 15

Datos y cálculos preliminares para la distribución de frecuencias de datos agrupados



A continuación, se debe realizar el cálculo de los intervalos. Todo intervalo contiene dos números, el valor a la izquierda es conocido con el nombre de límite inferior y el valor a la derecha con el de límite superior.

El límite superior de un intervalo se obtiene sumando al límite inferior el valor de la amplitud corregida, a´. El intervalo siguiente, se forma tomando como límite inferior el límite superior de la clase anterior, y así se repite el procedimiento hasta que se haya completado el número de intervalos considerados en la distribución, m´.

Para el caso del ejemplo de estudio, considerando que el valor mínimo en los datos es de 10,01, los intervalos considerados tendrían el detalle presentado en la Tabla 5.

Tabla 5

Detalle del cálculo de intervalos en el ejemplo planteado

Clase	Intervalos	Cálculo del límite superior
1	[10,01 ; 12,85)	10,01 + 2,84
2	[12,85 ; 15,69)	12,85 + 2,84
3	[15,69 ; 18,53)	15.59 + 2,84
4	[18,53 ; 21,37)	18,53 + 2,84
5	[21,37 ; 24,21)	21,37 + 2,84
6	[24,21 ; 27,05)	24,21 + 2,84
7	[27,05 ; 29,89)	27,05 + 2,84

Obsérvese que los intervalos son cerrados en el límite inferior y abiertos en el superior (Moore, & McCabe, 2017).

Equilibrio de Colas en la Distribución de Frecuencias

En el conjunto inicial de datos, presentados en la Figura 15, se había encontrado que el mayor valor existente era de 29,85. En la Tabla 3, se puede observar que el último límite superior, correspondiente a la clase 6, excede a ese valor máximo en 0,04:

$$29{,}89 - 29{,}85 = 0{,}04 \qquad 1.7$$

Por el contrario, nótese que el límite inferior del intervalo correspondiente a la primera clase coincide de forma exacta con el valor mínimo encontrado en los datos que es de 10,01.

Lo anterior, permite concluir que hay un exceso de 0,04 unidades en el límite superior de la última clase. Este remanente puede ser distribuido, equitativamente, para que el primer intervalo inicie 0,02 unidades antes del límite inferior (esto es en 9,99) y el límite superior de la última clase cierre 0,02 unidades después del máximo valor existente en los datos (29,87).

De forma general, esta distribución del exceso debe realizarse de la siguiente manera:

$$E = \frac{\text{límite superior de la última clase} - \text{valor máximo en los datos}}{2} \qquad 1.8$$

y luego, restar el valor a repartir, E, del límite inferior de la primera clase y recalcular los límitesde todos los intervalos, Tabla 6.

Tabla 6

Detalle del recálculo de intervalos en el ejemplo planteado

Clase	Intervalos	Cálculo del límite superior
1	[9,99 ; 12,83)	9,99 + 2,84
2	[12,83 ; 15,67)	12,83 + 2,84
3	[15,67 ; 18,51)	15,67 + 2,84
4	[18,51 ; 21,35)	18,51 + 2,84
5	[21,35 ; 24,19)	21,35 + 2,84
6	[24,19 ; 27,03)	24,19 + 2,84
7	[27,03 ; 29,87)	27,03 + 2,84

Nótese que, de forma automática, el límite superior de la última clase es ahora 0,02 (E) unidades más pequeño que el valor obtenido en la distribución inicial, Tabla 5.

El procedimiento anterior de repartir en forma equitativa el remante de los datos, es conocido con el nombre de equilibrio de colas en la distribución.

Marca de Clase

La marca de clase, x_i, es un valor que representa a todo el intervalo y puede ser obtenido como la semisuma o promedio de los límites superior e inferior de dicho intervalo (Johnson & Wichern, 2007).

El resto de los valores en la Tabla de distribución de frecuencias, Tabla 5, debe ser calculado como se ha explicado en las secciones precedentes. Nuevamente, cabe acotar que si en lugar de las frecuencias relativas se desea el cálculo de las frecuencias relativas porcentuales, basta multiplicar por 100 los datos correspondientes a las columnas 4 y 6 de la Tabla 7.

Tabla 7

Distribución de frecuencias para el ejemplo de datos agrupados

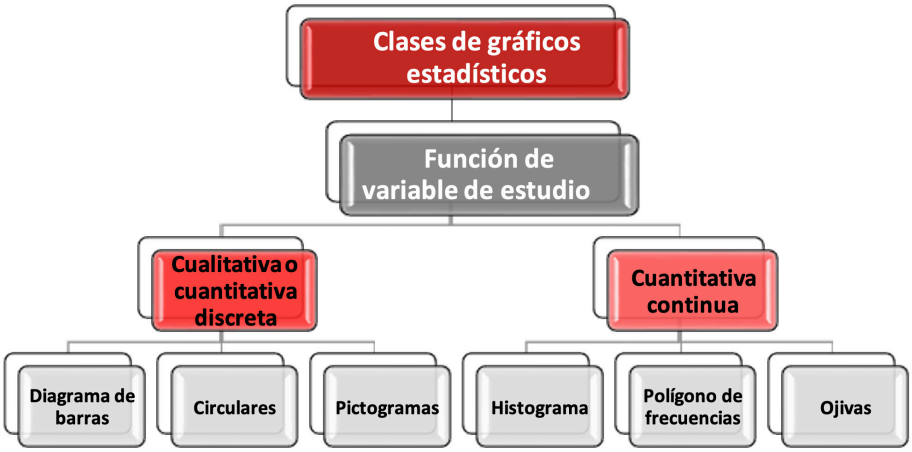
Intervalo	X_i	f_i	h_i	F_i	H_i
[9,99 ; 12,83)	11,41	8	0,2	8	0,2
[12,83 ; 15,67)	14,25	3	0,075	11	0,275
[15,67 ; 18,51)	17,09	7	0,175	18	0,45
[18,51 ; 21,35)	19,93	6	0,15	24	0,6
[21,35 ; 24,19)	22,77	3	0,075	27	0,675
[24,19 ; 27,03)	25,61	7	0,175	34	0,85
[27,03 ; 29,87)	28,45	6	0,15	40	1
Total		40	1		

Gráficos Estadísticos

Los gráficos estadísticos constituyen una poderosa herramienta de la estadística descriptiva a través de la cual se pueden presentar los datos de una manera resumida y que resulta de fácil comprensión para el lector. Entre los gráficos más utilizados, se pueden encontrar diagramas de barras, diagrama de puntos, histogramas, polígonos de frecuencia, diagrama circular, y ojivas ascendente y descendente.

En la Figura 16 se presenta el esquema correspondiente a la clasificación de los gráficos estadísticos en función del tipo de variable considerada en el estudio.

Figura 16
Clases de gráficos estadísticos en función de la variable de estudio



Nota. Elaboración propia a partir de SmartArt de Microsoft Office 365.

Diagrama de Barras

Son representaciones de barras rectangulares en el plano (Playfair, 1786), empleadas para presentar datos de variables cualitativas o cuantitativas discretas, en cuya base se coloca la categoría o el valor discreto asumido por la variable en estudio y cuya altura corresponde a la frecuencia con la que se presenta el dato en cuestión. Los diagramas de barras no muestran frecuencias acumuladas y, en ellos, la columna (o barra) con mayor altura representa la mayor frecuencia. Para obtener el número de datos de la muestra a partir de un diagrama de barras, basta con sumar todas las frecuencias absolutas, esto es, las alturas de las columnas.

Considérese el siguiente ejemplo: en una elección a concejal de un determinado Cantón se han postulado 5 candidatos cuya votación se presenta en la Tabla 8.

Tabla 8

Distribución de votos obtenidos por los candidatos a concejal

Candidato	Votos obtenidos
Candidato 1	250
Candidato 2	235
Candidato 3	318
Candidato 4	224
Candidato 5	362
Votos totales	1370

Para estos datos, colocando en el eje horizontal la variable cualitativa "candidato" y en el eje vertical los votos obtenidos por cada uno de ellos, que no serían otra cosa que las frecuencias absolutas, se obtiene el diagrama de barras que se muestra en la Figura 17. Nótese que, en este caso, la mayor frecuencia, que corresponde a la columna más alta, señala al candidato ganador de la elección.

Diagramas Circulares

Conocidos también como diagrama de sectores o gráfico de pastel (Playfair, 1801). Está formado por un círculo dividido en partes proporcionales a las frecuencias relativas de cada categoría. Son usados para representar tanto datos cualitativos como datos cuantitativos discretos.

En el caso del ejemplo anterior, se pueden calcular las frecuencias relativas dividiendo el número de votos obtenido por cada candidato entre el total de votos. En la Figura 18 se muestra el diagrama circular correspondiente al ejemplo de la elección del concejal cantonal, obsérvese que en los sectores del diagrama se muestran tanto las frecuencias absolutas como las frecuencias relativas porcentuales.

Figura 17

Diagrama de barras correspondiente a la distribución de votos de los candidatos

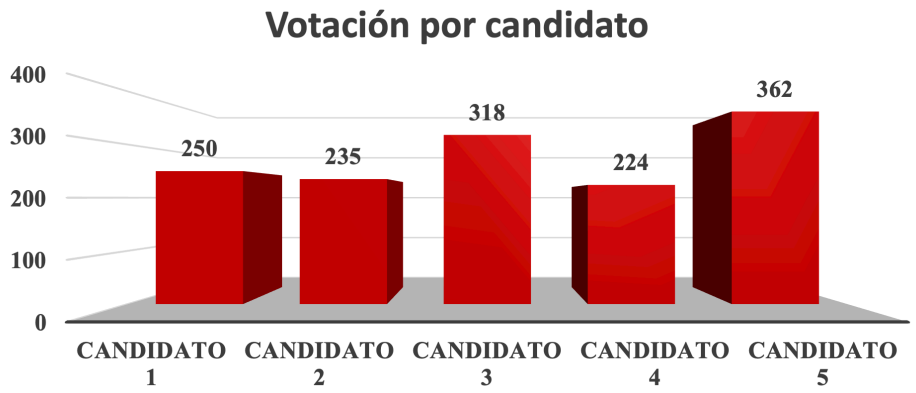
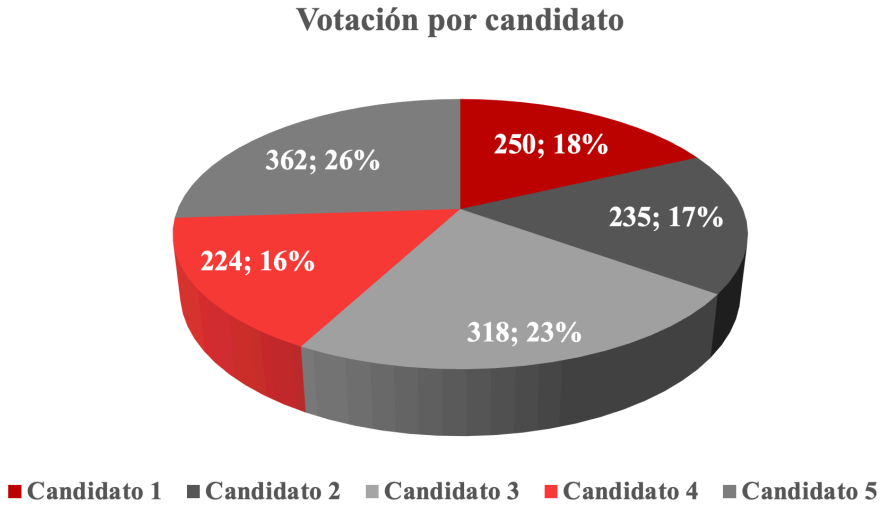


Figura 18

Diagrama circular correspondiente a la distribución de votos de los candidatos



Nota. En cada sector del diagrama se indica la cantidad de votos obtenidos por cada candidato y el porcentaje equivalente de la votación total.

Pictogramas

Corresponde a un tipo de gráfica en la que se usan imágenes relacionadas con la variable de estudio (Bertin, 1967). Las imágenes representan los datos y, en general, se usan dos tipos de representación: en la que se cambia el símbolo icónico proporcionalmente a la frecuencia que representa o en la que se utilizan diferentes imágenes para representar los cambios que presenta la variable en atención a sus distintas categorías.

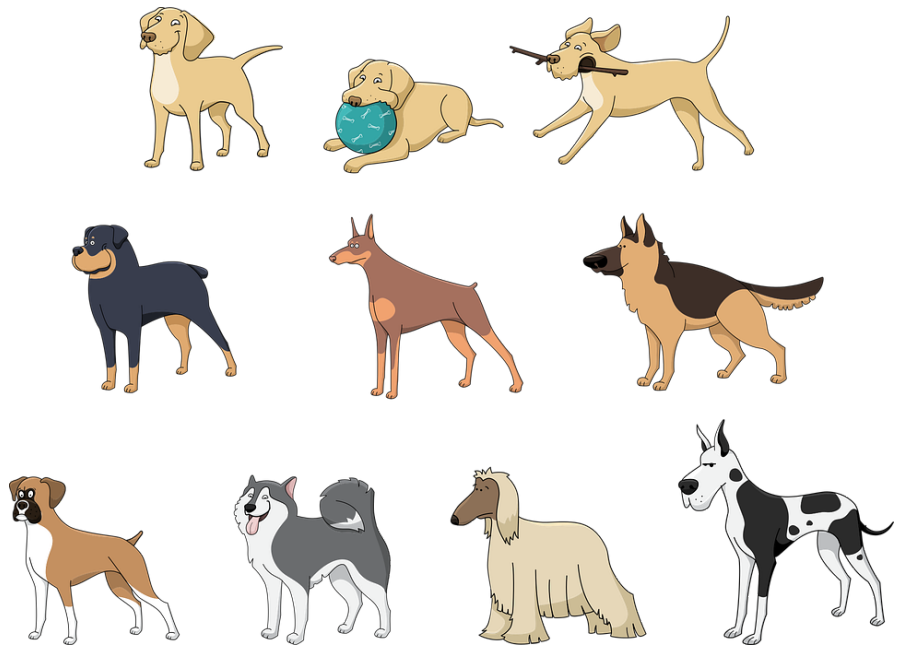
En la Figura 19 se presentan los íconos característicos para identificar la clasificación de una variable en atención al sexo del individuo, y en la Figura 20, se muestra un pictograma que podría ser usado para presentar, por ejemplo, resultados de un estudio estadístico de la esperanza de vida de los perros en atención a su raza.

Figura 19

Pictogramas utilizados para la diferenciación gráfica de la variable sexo

Nota: *Adaptado de Mujer Hombre [Pictograma] Pixabay, 19 de junio de 2014, Pixabay, https://pixabay.com/es/vectors/mujer-hombre-pictograma-separarse-310532/ Pixabay License.*

Figura 20
Ejemplificación de pictogramas que pueden ser usados para un estudio de longevidad de los perros en atención a la raza



Fuente: *Animales Mascotas [Pictograma] Pixabay, 15 de junio de 2016, Pixabay, https://pixabay.com/es/vectors/animales-mascotas-perro-1454214/ Pixabay License.*

Diagrama de Puntos

El diagrama de puntos, también conocido como dot plot en inglés, es un tipo de gráfico que utiliza puntos para representar los datos (Galton, 1886). Permite identificar con facilidad dónde se encuentran los datos y cuál es su variabilidad. De igual forma, facilita la ubicación visual de los espacios vacíos y las zonas de agrupamientos.

Considérese que se ha realizado un test de razonamiento lógico, contentivo de un total de 100 puntos, a 30 estudiantes. Los puntajes obtenidos, oscilan entre 50 y 91 puntos., Tabla 9.

Tabla 9

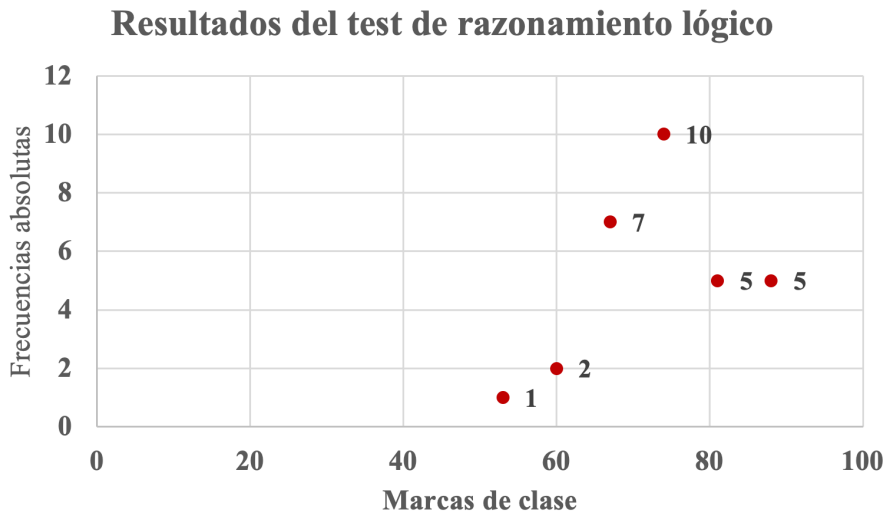
Distribución de frecuencias de puntajes obtenidos en un test de razonamiento lógico

Intervalos	X_i	f_i
[49,5 – 56,5)	53	1
[56,5 – 63,5)	60	2
[63,5 – 70,5)	67	7
[70,5– 77,5)	74	10
[77,5 – 84,5)	81	5
[84,5 – 91,5)	88	5

El diagrama de puntos puede ser obtenido a través de un gráfico de dispersión de puntos, Figura21, en cuya primera columna se colocan las marcas de clase X_i (y en la segunda columna las frecuencias absolutas f_i (especificadas en la Tabla 9.

Figura 21

Diagrama de puntos correspondiente a los datos de la Tabla 9



Histogramas

Se utiliza para representar la distribución de una variable continua (Pearson, 1895). La representación gráfica de la variable se realiza en forma de barras, pero sin espacios entre ellas. Para su construcción, sobre el eje horizontal se colocan las marcas de clases y sobre el eje vertical las frecuencias observadas. En este caso, al igual que el diagrama de barras, la sumatoria de las alturas de las columnas equivale al 100% de los datos.

Los histogramas presentan algunas diferencias con los diagramas de barras; ya que su uso se restringe a distribuciones de datos agrupados, con variables cuantitativas continuas, lo anterior implica que, a diferencia del diagrama de barras donde visualmente puede observarse una separación entre las barras, en los histogramas las barras o columnas se grafican en forma consecutiva. A partir de los datos presentados en la Tabla 10, se construye el histograma que se muestra en la Figura 22.

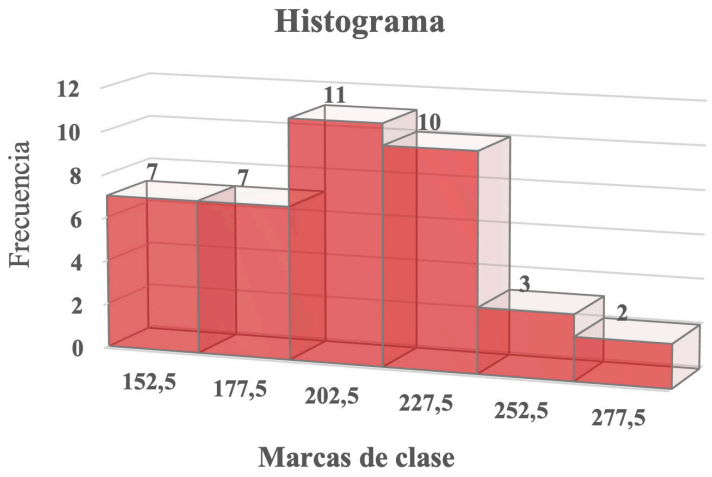
Las marcas de clase deben ser colocadas justo al centro de la base de cada columna y la altura estará definida por la altura correspondiente.

Tabla 10

Distribución de frecuencias utilizada para la construcción del histograma

Intervalos	X_i	f_i
[140 -165)	152,5	7
[165 – 190)	177,5	7
[190 – 215)	202,5	11
[215 – 240)	227,5	10
[240 – 265)	252,5	3
[265 – 290)	277,5	2

Figura 22
Histograma correspondiente a una distribución de datos cuantitativos continuos



Es importante resaltar que los histogramas sólo pueden ser usados para graficar la distribución de frecuencias de una variable continua, ya que esta puede moverse en un rango de valores infinito. Debido a esto, es necesario agrupar los datos en intervalos o clases para poder representarlos de manera gráfica.

Por otro lado, las variables discretas, que solo pueden tomar valores enteros o de un conjunto finito de valores, no necesitan ser agrupadas en intervalos o clases para ser representadas gráficamente. En su lugar, se pueden utilizar otros tipos de gráficos, como gráficos de barras o gráficos de sectores.

Polígonos de Frecuencia

Se construyen a partir de los histogramas y sólo pueden ser utilizados cuando se tienen variables continuas (Venn, 1887).

El polígono de frecuencia es una forma alternativa de representar la distribución de frecuencias que puede ser más útil en ciertas situaciones, como cuando se comparan dos o más muestras.

Los polígonos de frecuencias corresponden a una gráfica representada en el plano XY. Para su construcción en el eje horizontal se representan las marcas de clase y sobre el eje vertical se representan las frecuencias; al unir los puntos de la gráfica se forma una línea poligonal.

La característica principal que poseen es que inician y terminan sobre el eje horizontal (frecuencia cero), y se construyen uniendo con líneas rectas los puntos medios de los topes de las columnas. El punto de mayor altura del polígono representa la mayor frecuencia existente.

Para la construcción del polígono de frecuencias se crean dos intervalos ficticios de igual amplitud que el resto, pero con frecuencia absoluta igual a cero, Tabla 11.

Tabla 11

Distribución de frecuencias utilizada para la construcción del polígono de frecuencias

Intervalos	X_i	f_i
[115 -140)	127,5	0
[140 -165)	152,5	7
[165 – 190)	177,5	7
[190 – 215)	202,5	11
[215 – 240)	227,5	10
[240 – 265)	252,5	3
[265 – 290)	277,5	2
[290 – 315)	302.5	0

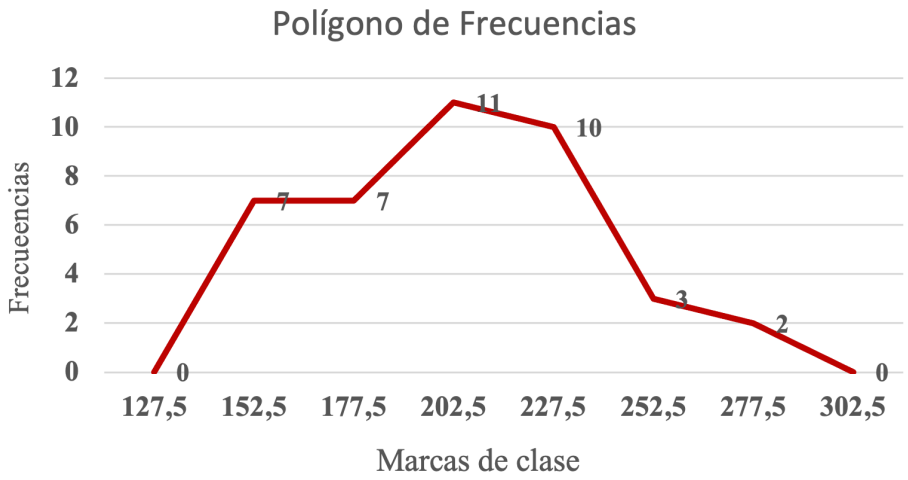
Obsérvese en la distribución de frecuencias presentada en la Tabla 11 que, para los intervalos ficticios, que han sido creados en la parte superior e inferior de la Tabla, debe realizarse el cálculo de la marca de clase y de los límites inferior y superior de la clase. Para obtener la marca de clase del primer intervalo, se resta la amplitud a la marca de clase que corresponde al

intervalo de la segunda clase. Para obtener el límite superior de la primera clase, se toma el valor del límite inferior de la clase siguiente, una vez obtenido este límite superior, se le resta la amplitud para determinar el límite inferior de esta primera clase ficticia.

Un procedimiento similar se sigue para encontrar la marca de clase y los límites de la última clase.

En la Figura 23 se presenta el polígono de frecuencia correspondiente a la Tabla 11.

Figura 23
Polígono de frecuencias



Capítulo 2

Medidas de tendencia central, de posición y de dispersión

CAPÍTULO II

MEDIDAS DE TENDENCIA CENTRAL, DE POSICIÓN Y DE DISPERSIÓN

Medidas de tendencia central

Las medidas de tendencia central, Pearson (1895), son un conjunto de estadísticos descriptivos que se utilizan para resumir y describir la distribución de una variable. Se utilizan para identificar el valor central o típico de una distribución de datos. Las tres medidas de tendencia central más comunes son la media, la mediana y la moda. Corresponde a un valor que va a ser el representante de una serie de datos que pueden estar o no agrupados.

Media Aritmética para Datos no Agrupados y Agrupados

Es la medida de tendencia central más conocida, sólo puede ser usada con variables cuantitativas. Representa el promedio de un conjunto de datos.

Es una medida sensible a los extremos, esto significa que es afectada por la existencia de datos con valores muy grandes o muy pequeños. Valores muy grandes aumentan la media y valores muy pequeños la disminuyen.

Para calcularla debe tomarse en cuenta si los datos se encuentran, o no, agrupados.

Cuando la media se calcula en una muestra, se suele designar con, $\bar{x}$ en cambio, cuando se trata de una población, se denota con la letra griega μ

Media para Datos no Agrupados

Se obtiene al sumar todos los valores de la variable y dividirlos entre el número de datos.

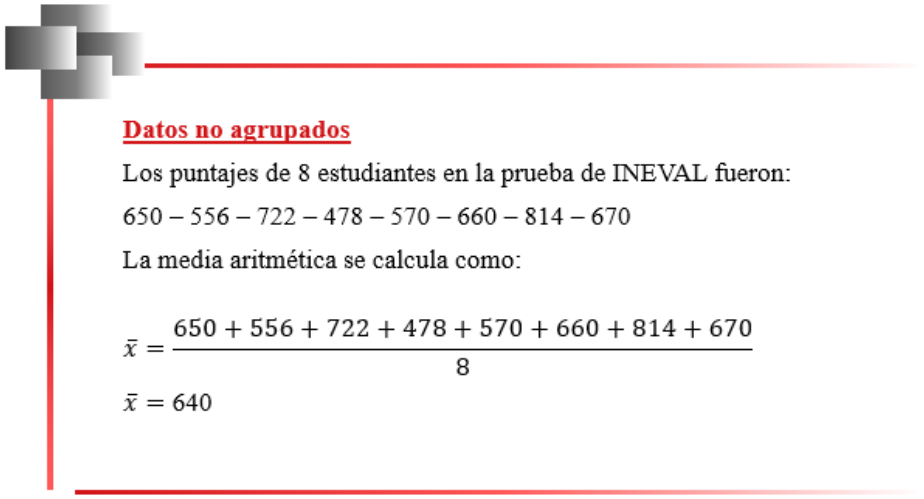
La fórmula para su cálculo es la siguiente:

$$\bar{x} = \frac{\sum x_i}{n} \qquad 2.1$$

donde x_i corresponde al valor de cada uno de los datos que forman parte de la población o la muestra, con i variando desde 1 hasta n; y n es el número

de datos u observaciones. En la Figura 24 se puede visualizar un ejemplo de cálculo de la media para datos no agrupados.

Figura 24
Ejemplo de cálculo de la media para datos no agrupados



Media para Datos Agrupados

En el cálculo de la media para datos agrupados interviene un elemento nuevo: la frecuencia absoluta, f_i. La fórmula a utilizar en este caso, es:

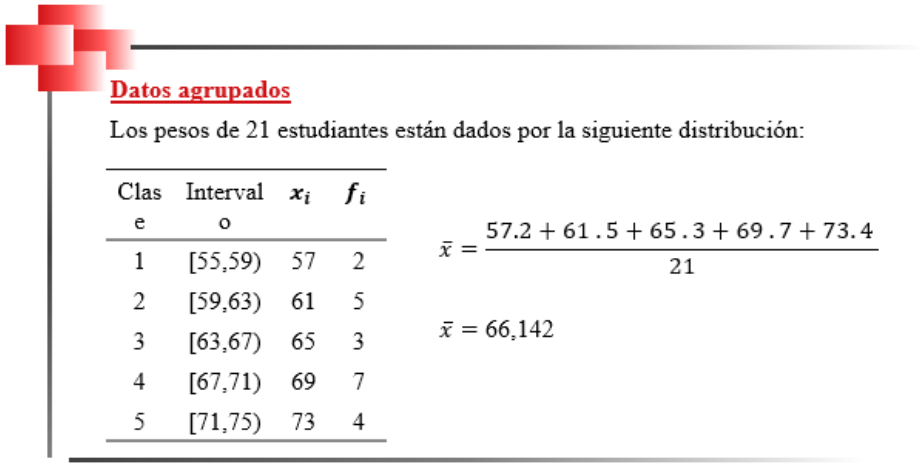
$$\bar{x} = \frac{\sum x_i \cdot f_i}{n} \qquad 2.2$$

En la Figura 25, se incluye un ejemplo de cálculo de la media para datos agrupados.

Mediana para Datos no Agrupados y Agrupados

Corresponde al valor central de un arreglo ordenado de datos La mediana, Md, no toma en cuenta ni se ve afectada por los valores extremos. Por estar ubicada al centro de los datos, deja a la izquierda y a la derecha exactamente el mismo número de observaciones. Sólo puede ser obtenida para datos cuantitativos y es comúnmente utilizada tanto en datos no agrupados, Galton (1883), como en datos agrupados (Pearson, 1895).

Figura 25
Ejemplo de cálculo de la media para datos agrupados



Mediana para Datos no Agrupados

Su forma de cálculo depende del número de datos u observaciones que existen, pudiendo este ser par o impar. Para poder determinar la mediana en datos no agrupados es absolutamente necesario que estos se encuentren ordenados de manera creciente o decreciente.

Mediana cuando el Número de Datos es Impar. Corresponde al valor del dato central. Esto implica que, en el arreglo ordenado, el valor que corresponde a la mediana deja por debajo y por arriba exactamente el mismo número de datos.

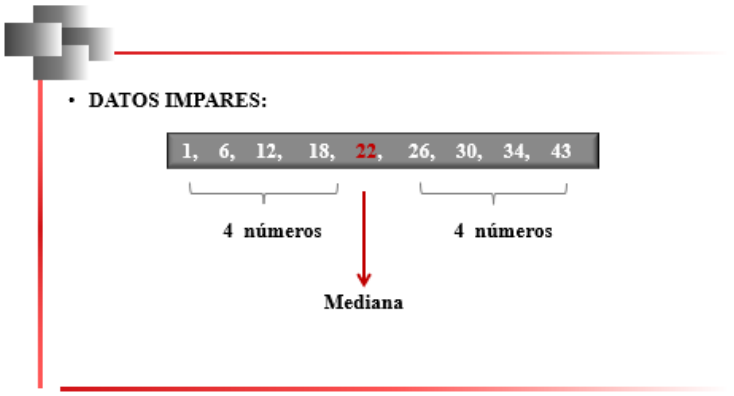
En la Figura 26, por ejemplo, se observa que existen nueve datos. Para ubicar la mediana se calcula la posición *i:*

$$i = \frac{n+1}{2} \qquad 2.3$$

que en este caso resultaría igual a 5. Como los datos están ordenados, se cuenta la posición a partir del primer dato ubicado a la izquierda, el cual ocuparía la primera posición. Exactamente en la quinta posición se encontraría ubicado el dato que corresponde a la mediana, esto es el número 22.

Obsérvese que el número 22, tiene a su izquierda 4 datos que son menores o iguales que él y, a su derecha, 4 datos que son mayores que 22.

Figura 26
Mediana en datos impares

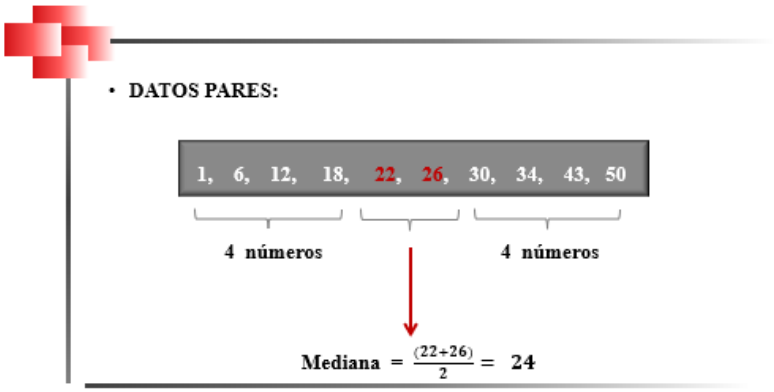


Mediana cuando el Número de Datos es Par. En este caso, la mediana es el promedio de los dos datos que se encuentran al centro del arreglo ordenado, en las posiciones i_1 e i_2:

$$i_1 = \frac{n}{2} \; ; \; i_2 = \frac{n}{2} + 1 \qquad 2.4$$

dejando a la izquierda y a la derecha el mismo número de datos, Figura 27.

Figura 27
Mediana en datos pares



Mediana para Datos Agrupados y Variable Discreta

Cuando los datos recolectados se repiten y están tabulados con su respectiva frecuencia, si el número de datos es impar, la mediana, Figura 28, estará en la posición:

$$i = \frac{n+1}{2} \qquad 2.5$$

Obsérvese que cuando el número de datos es impar, la operación anterior resulta un número par, por lo que el valor puede ser buscado de manera directa en las frecuencias absolutas acumuladas, sin hacer ninguna aproximación.

Si el número de datos es par, la mediana estará entre las posiciones.

$$\frac{n}{2} \; y \; \frac{n}{2} + 1.$$

En la Figura 29 se presenta el caso particular cuando las posiciones anteriores están acumuladas en la misma clase. Obsérvese que, como la frecuencia absoluta acumulada de la segunda clase, *F2* es 9, y la frecuencia absoluta, *f3*, de la tercera clase es 5, en la tercera clase la frecuencia absoluta acumulada, *F3*, contienen 5 posiciones: la 10, la 11, la 12, la 13, y la 14. Particularmente, allí, en esa tercera clase, se encuentran entonces las posiciones 11 y 12. Debido a ello, la tercera clase es la clase de la mediana.

Figura 28
Mediana para datos agrupados impares

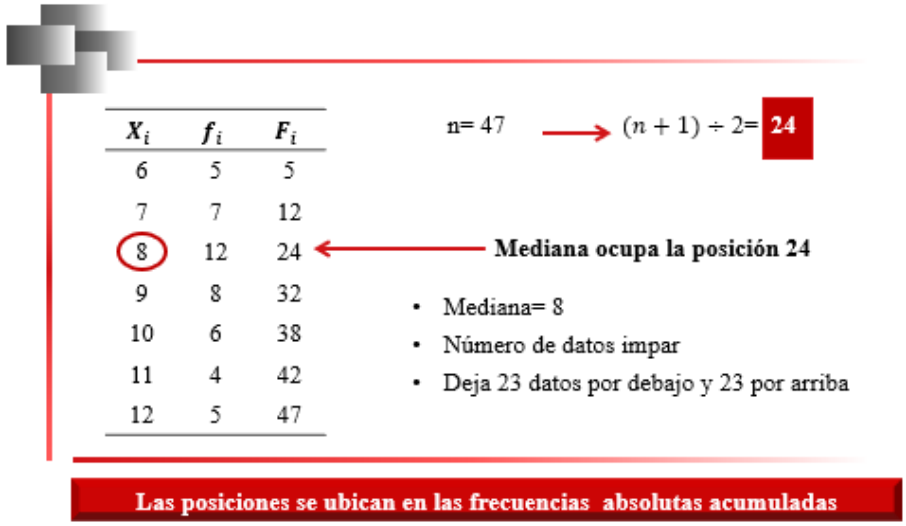
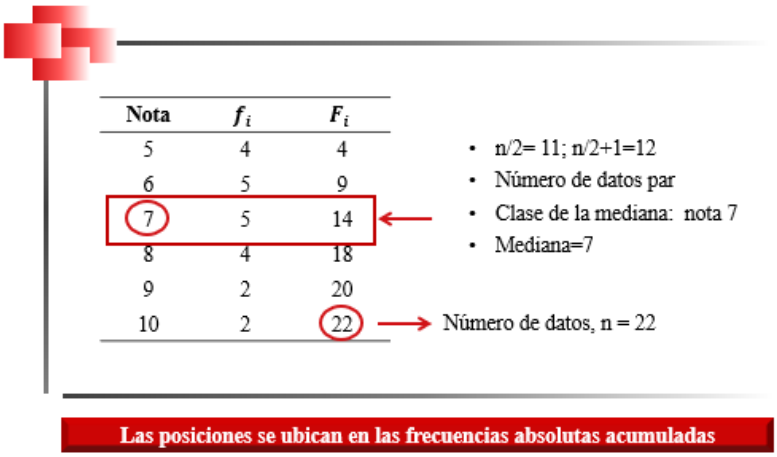


Figura 29

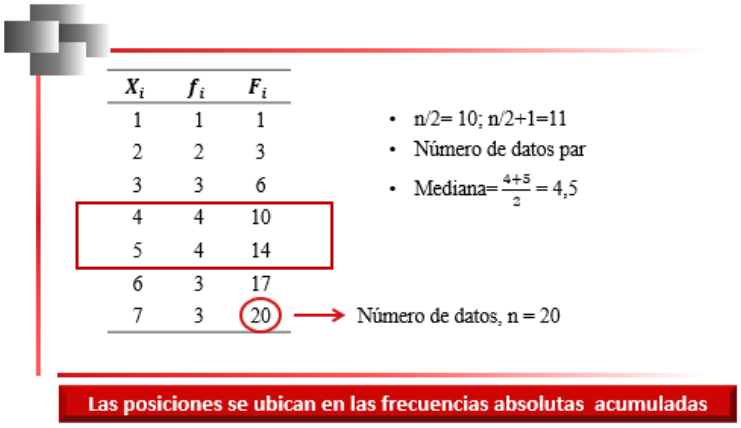
Mediana para datos agrupados pares con datos centrales en la misma clase



En la Figura 30, se presenta un ejemplo del cálculo de la mediana cuando se tiene un número par de datos y éstos se encuentran agrupados según sus respectivas frecuencias.

Figura 30

Mediana para datos agrupados pares con datos centrales en diferentes clases



En ese caso en particular, las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$ no se encuentran en la misma clase; ya que, como el número de datos es 20, $\frac{n}{2}$ corresponde a la posición 10 y $\frac{n}{2}+1$, a la posición 11. Se debe tener presente que las posiciones se deben leer en la columna de las frecuencias absolutas acumuladas, *Fi*. En la Figura 30 se puede observar que la posición 10, se encuentra en la clase 4 y la posición 11 se encuentra acumulada en la clase 5 (nótese que en

la clase 5 la frecuencia absoluta acumulada es igual a 14, lo cual indica que allí se encuentran acumuladas las posiciones 11, 12, 13, y 14).

Mediana para Datos Agrupados en Distribuciones Intervalares

Cuando se tienen distribuciones intervalares, lo primero que se debe identificar es cuál es la clase donde se encuentra la mediana. Para ello, se calcula el valor de $\frac{n}{2}$, que dará la posición de la mediana, y, luego se procede a ubicar esta posición en la columna de las frecuencias absolutas acumuladas. Una vez identificada la clase de la mediana, el cálculo de esta se realiza a través de la siguiente fórmula:

$$M_d = Lim_{\inf(i)} + \frac{\left(\frac{n}{2} - F_{i-1}\right) a}{f_i} \qquad 2.6$$

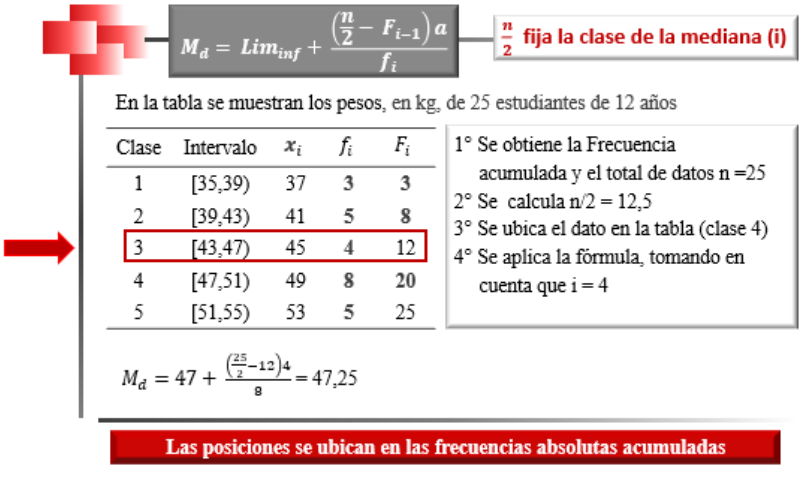
Donde es la clase de la mediana, $Lim_{inf(i)}$ corresponde al límite inferior de la clase de la mediana, n es el tamaño de la muestra, F_{i-1} es la frecuencia absoluta acumulada de la clase inmediatamente anterior a la clase que contiene a la mediana (la que se encuentra en la fila que precede a la clase de la mediana) , a es la amplitud del intervalo y f_i es la frecuencia absoluta de la clase de la mediana.

En la Figura 31, se presenta un ejemplo de cálculo de la mediana en distribuciones intervalares.

Obsérvese que en este caso la obtención de la mediana requiere de la aplicación de la fórmula correspondiente a datos agrupados en intervalos.

Figura 31

Cálculo de la mediana en distribuciones intervalares



Es importante tomar en cuenta que para fijar la clase de la mediana se debe dividir el número de datos entre dos y ubicar la posición entera en la columna de las frecuencias absolutas acumuladas (13). Sin embargo, al momento de realizar el cálculo de la mediana a través de la fórmula, el valor de n/2 debe ser introducido sin aproximación.

Moda para Datos no Agrupados y Agrupados

La moda, M_o es una medida de tendencia central que utiliza la estadística descriptiva para indicar cuál el valor de la variable que más se repite en un conjunto de datos. A diferencia de la media y de la mediana, la moda puede hallarse en variables cualitativas.

En los datos, estén agrupados o no, puede existir más de una moda. En el caso de que sólo exista una moda, la distribución se dice unimodal. Cuando se tienen dos modas es bimodal, con tres modas, bimodal y con más de tres modas, recibe el nombre de multimodal.

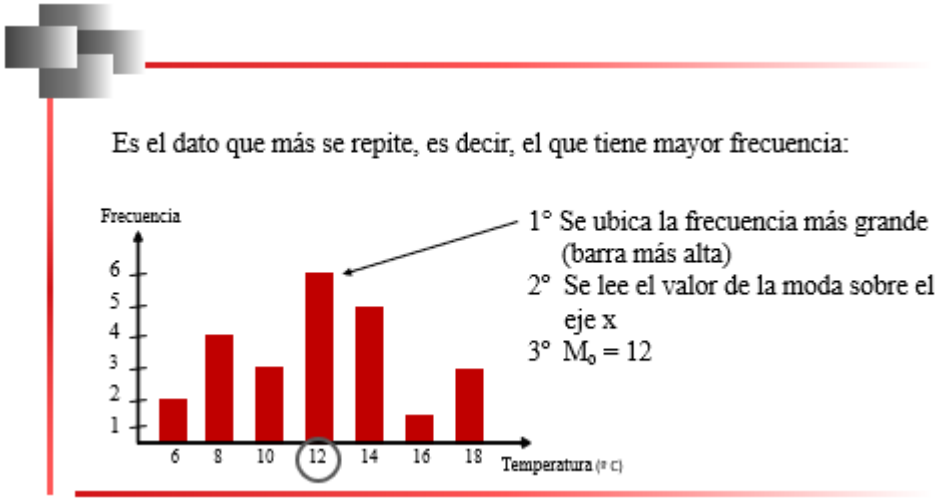
En la definición de la moda, al igual que en la mediana, no se toman en cuenta todos los datos; ya que, en este caso, lo importante es determinar cuál es la frecuencia absoluta que registra un valor mayor en todo el conjunto de observaciones.

Particularmente, esta medida de tendencia central puede ser afectada por la formación de los intervalos. La obtención de las frecuencias absolutas depende de los límites inferior y superior que se hayan definido en un intervalo, ya que ellos son los que determinan si un número pertenece a un intervalo o no. Recuérdese además que existen varios criterios para definir el número de intervalos en los que van a agruparse un conjunto de datos, por ende, al cambiar el número de intervalos los límites de cada clase serán distintos y, en consecuencia, en cada caso, un determinado dato puede llegar a pertenecer a un intervalo diferente.

En la Figura 32, se ilustra un ejemplo de la determinación de la moda a partir de un diagramade barras.

Figura 32

Moda en una distribución de variable discreta



Considérese el siguiente ejemplo:

Los datos presentados en la Tabla 12 provienen de una encuesta realizada a 25 estudiantes, de Bachillerato General Unificado, a los que se les preguntó sobre su asignatura favorita.

Tabla 12

Asignatura favorita de 25 estudiantes de Bachillerato General Unificado

Asignatura favorita				
Matemáticas	Ciencias	Literatura	Matemáticas	Ciencias
Física	Inglés	Matemáticas	Historia	Literatura
Ciencias	Literatura	Historia	Matemáticas	Inglés
Historia	Inglés	Matemáticas	Matemáticas	Ciencias
Inglés	Ciencias	Ciencias	Inglés	Historia

Para determinar la moda de los datos presentados en la Tabla 12, se cuentan las frecuencias de cada una de las respuestas:

Historia: 4	Inglés: 5	Matemáticas: 6	Literatura: 3
Ciencias: 6	Física: 1		

La moda corresponde a aquel valor que más se repite, en este caso, como se puede observar, tanto Matemáticas como Ciencias presentan una frecuencia igual a 6. Dado que éste es la mayor frecuencia encontrada en los datos y que la misma se repite para dos asignaturas, se concluye quela distribución es bimodal siendo M_{o1} = Matemáticas y M_{o2} = Ciencias.

En la Figura 33 se muestra un ejemplo de determinación de la moda en una distribución de frecuencias correspondientes a una variable discreta.

Para determinar la moda en este tipo de datos agrupados en los que no tenemos intervalos, se debe ubicar en la columna de las frecuencias absolutas el valor más alto. Obsérvese que hay dos datos que poseen la misma frecuencia, y que ésta es precisamente la frecuencia más alta. En este caso, se dice que hay dos modas, una correspondiente al número 2 y otra al número 3.

Figura 33

Determinación de la moda en distribuciones de datos agrupados, variable discreta

La siguiente tabla muestra datos discretos correspondientes a las preferencias de distintos tipos de software para realizar mapas mentales.

Número	f_i	
0	1	
1	3	
2	4	→ $M_o = 2\ y\ 3$
3	4	→ bimodal
4	3	
5	3	
6	0	

Moda para Datos en Distribuciones Intervalares

Para el cálculo de la moda en datos con distribuciones intervalares (Wackerly et al., 2008) se utiliza la siguiente fórmula:

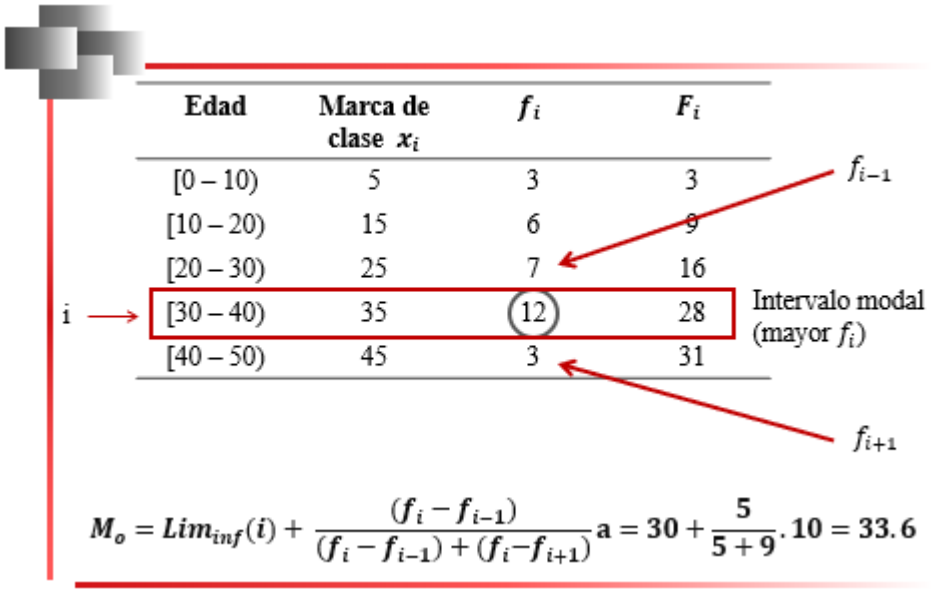
$$M_o = Lim_{inf}(i) + \frac{(f_i - f_{i-1})}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a \qquad 2.7$$

Donde $Lim_{inf} =$ límite inferior de la clase modal, $a =$ amplitud de la clase, $(f_i - f_{i-1})$ es la diferencia entre las frecuencias absolutas de la clase modal y la premodal y $(f_i - f_{i+1})$ la diferencia entre las frecuencias absolutas de la clase modal y la postmodal.

Para determinar la clase modal, se ubica, en la columna de las frecuencias absolutas, el valor más alto, Figura 34.

Figura 34

Ejemplo de determinación de clase modal y cálculo de la moda en distribuciones invervalares



En caso de que la frecuencia absoluta más alta se encuentre repetida más de una vez, la fórmula especificada para el cálculo deberá emplearse tantas veces como resultados repetidos con dicho valor se hayan encontrado en la columna de las frecuencias absolutas.

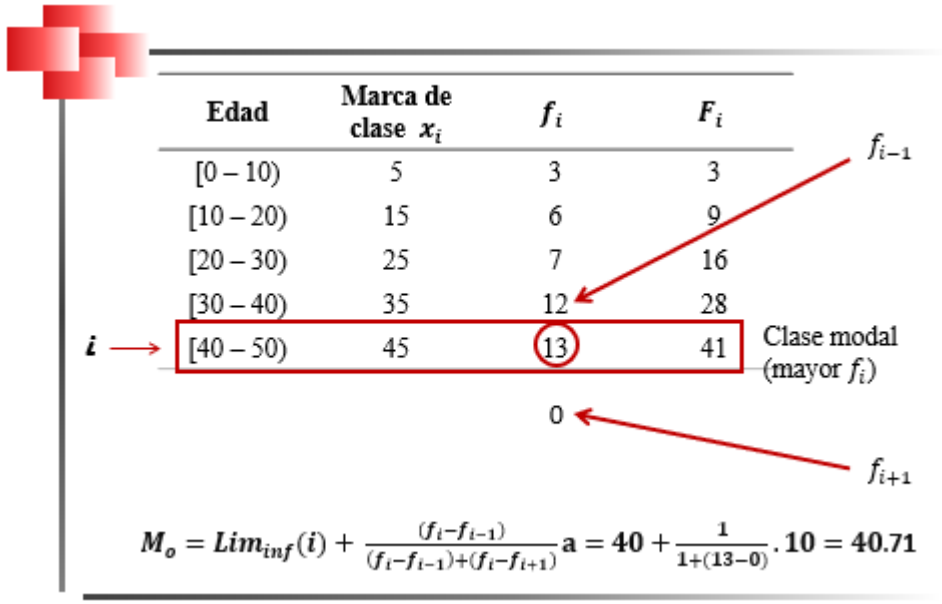
Una vez identificada la clase modal (i), la clase premodal será la clase inmediatamente anterior a la modal (i-1) y la posmodal la inmediatamente siguiente (i+1).

De presentarse el caso de que la máxima frecuencia se encuentre en la primera clase (primera fila de la Tabla de frecuencias), el valor de la frecuencia de la clase premodal, f_{i-1}, obviamente, debe ser cero, ya que no hay una clase anterior a la primera clase.

Una situación similar ocurriría si la mayor frecuencia absoluta se ubicara en la última clase de la Tabla de frecuencias. En este caso, no habría clase siguiente y, en consecuencia, la frecuencia absoluta de la clase posmodal, f_{i+1}, sería igual a cero. En la Figura 35, se ilustra un ejemplo de esta situación.

Figura 35

Ejemplo de cálculo de la moda en distribución intervalar cuando la clase modal se ubica en uno de los intervalos extremos



Medidas de posición, cuartiles, deciles y percentiles

Las medidas de posición, o cuantiles (Gnedenko & Kolmogorov, 1954), son valores capaces de dividir al conjunto de datos en un número específico de partes que poseen la misma cantidad de elementos, de tal forma que se facilite la interpretación del comportamiento de la variable en estudio.

Las fórmulas para el cálculo de percentiles, deciles y cuartiles en datos agrupados se basan en la interpolación lineal y pueden variar ligeramente dependiendo de la fuente.

A continuación, se estudiarán tres tipos de cuantiles: los percentiles, los deciles y los cuartiles, (Freund, & Simon, 2011).

Percentiles

Los percentiles son medidas de posición que se utilizan en datos cuantitativos que están ordenados de menor a mayor.

Pueden ser utilizados en variables tanto discretas como continuas.

El procedimiento para la ubicación y determinación de un determinado percentil es diferente dependiendo de si los datos se encuentran agrupados o no.

Los percentiles son un tipo de cuantil que corresponde a los valores de la muestra que permiten dividirla en 100 partes iguales. Los percentiles dan los valores correspondientes al 1%, al 2%... y al 99% de los datos e indican cómo se distribuyen estos datos comenzando desde el valor menor hasta el valor mayor.

El valor, pi ,de un percentil indica cuáles son los valores iguales o menores que ese valor pi.

La interpretación del percentil i, es que supera al i% de los datos y es superado, a su vez, por(100- i)% de los datos.

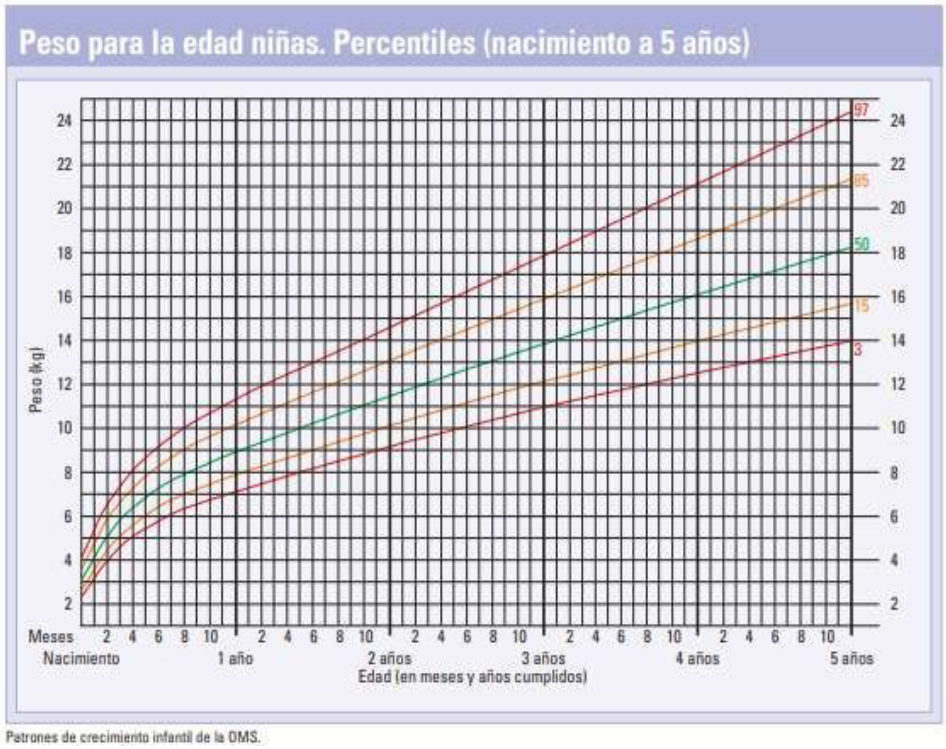
Así, el percentil 92 es superior o igual al 92% de los datos y es superado por el 8% restante; y el percentil 54, supera o es igual al 54% de los datos y es superado por el 46% restante.

Nótese que, si el valor correspondiente al P_{50} supera al 50% de los datos y es rebasado, a su vez,por el otro 50%, entonces el P_{50} coincide con la mediana.

Una de las aplicaciones más conocidas de los percentiles se tiene en Pediatría, cuando se realiza el control de las estaturas y los pesos de los niños. En la Figura 36, a manera de ejemplo, se muestra una gráfica que permite determinar diversos percentiles del peso de niñas entre 0 y 5 años.

Figura 36

Gráfica de percentiles del peso de niñas entre 0 y 5 años



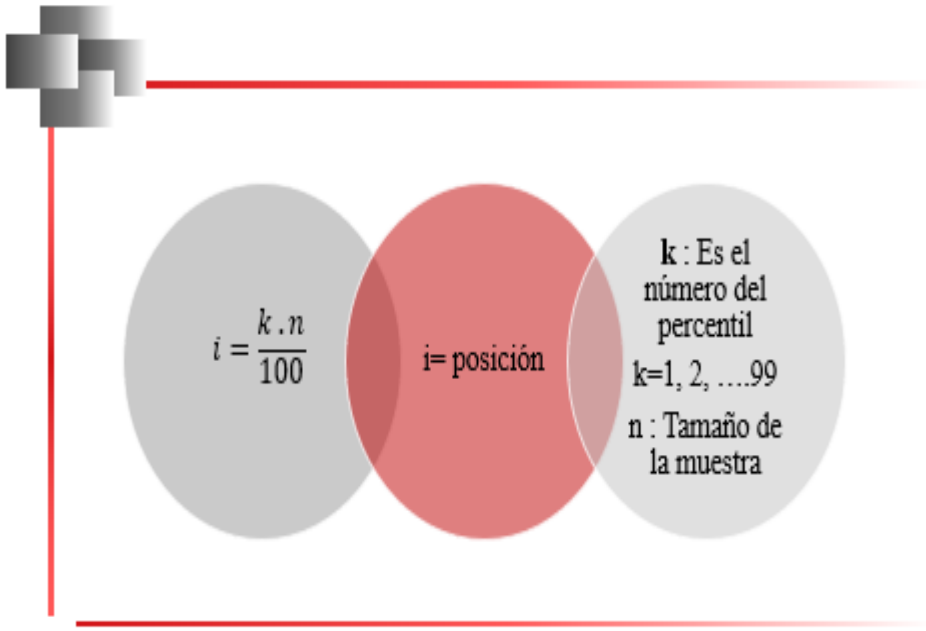
Nota. Adaptado de Patrones de crecimiento infantil de la OMS, Fundación Pediatría y Salud,2009, Asociación Española de Pediatría de Atención Primaria (https://n9.cl/tqn5i)

Nótese, por ejemplo, que una niña de 3 años que pese 14 kg, se encuentra en el percentil 50, Para encontrar el percentil de ubicación, se ubica en el eje de las x la edad de la niña y en el eje de las y el peso registrado. Se traza por la edad una recta vertical y por el peso una recta horizontal, el punto donde las rectas anteriores se intersectan cae sobre una curva que corresponde al percentil donde se encuentra ubicada la niña en cuestión.

Percentiles en Datos no Agrupados

Cuando se desea calcular cualquier cuantil, y en particular un percentil, lo primero que se debe hacer es calcular su posición. Cuando se dispone de un conjunto de datos no agrupados, para ubicar la posición se utiliza la fórmula presentada en la Figura 37.

Figura 37
Fórmula para el cálculo de la posición en percentiles

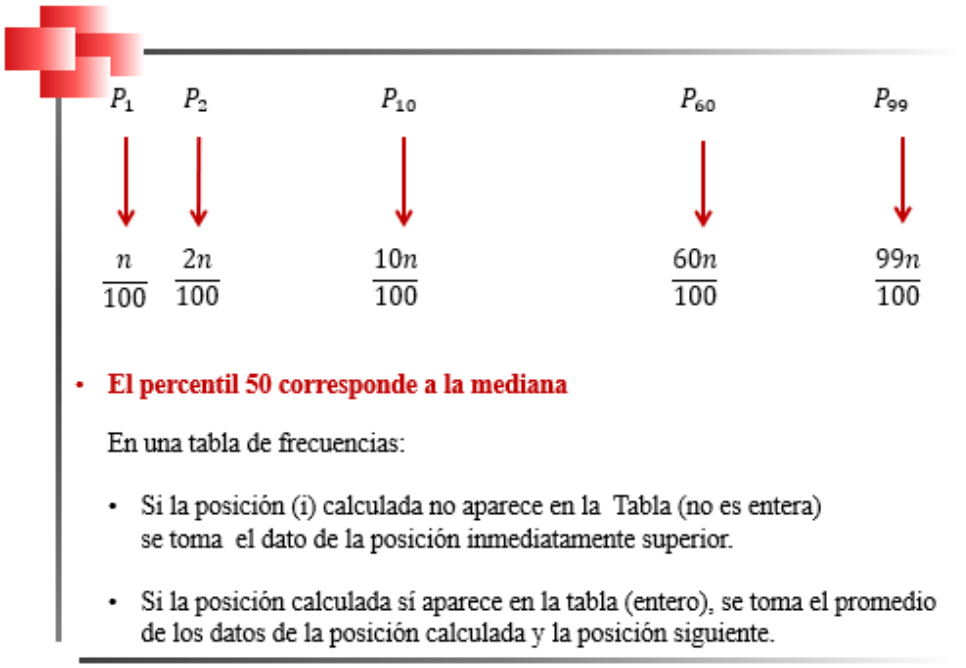


En la Figura 38, se presentan las posiciones correspondientes a algunos percentiles. En general, *el* k- ésimo *percentil* se ubica en la posición i:

$$i = \frac{k \,.\, n}{100} \qquad 2.8$$

Es de hacer notar que, dado que los percentiles dividen al conjunto de datos en 100 partes iguales, sólo existen 99 percentiles. También es importante resaltar que los datos deben estar ordenados de menor a mayor y que las posiciones corresponden a la numeración consecutiva de dichos datos.

Figura 38
Ejemplificación del cálculo de la posición de algunos percentiles



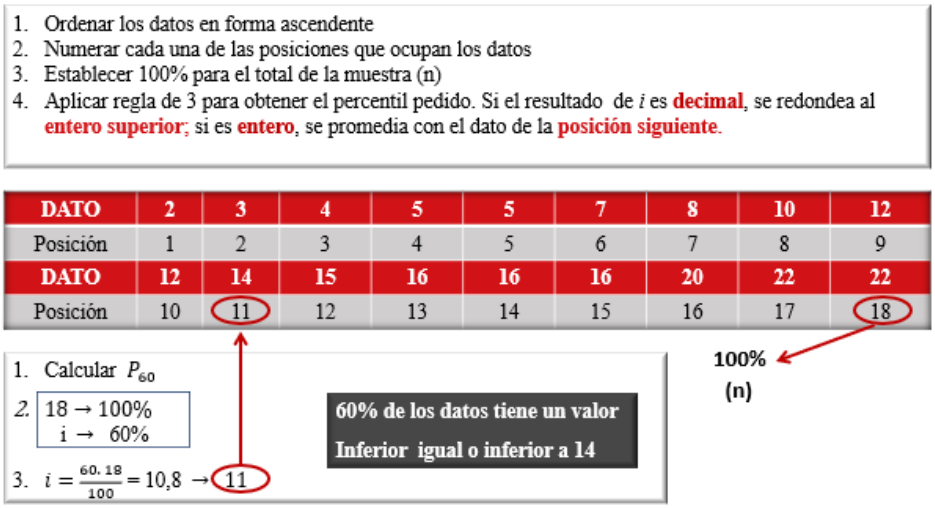
En la Figura 39, se presenta un ejemplo de cálculo de percentiles en datos no agrupados. Obsérvese que se ha identificado, de manera consecutiva, la posición de cada uno de los datos, comenzando desde el número 1.

En este ejemplo, la posición calculada resultó no ser entera, debido a ello, ya que los datos no están agrupados, se redondea dicha posición al entero inmediato superior. Resulta sencillo comprender que las posiciones corresponden a números que se utilizan para contar y que por lo tanto no pueden tener decimales. Esto no debe hacerse en el caso de que se tengan datos agrupados.

Obsérvese también que la última posición en el arreglo de datos suministra la información exacta del tamaño de la muestra.

Figura 39

Ejemplo de cálculo de percentiles en datos no agrupados cuando la posición no es entera



En la Figura 40, se incluye un ejemplo del cálculo del percentil cuando la posición calculada es entera. Obsérvese que, en este caso, se debe ubicar tanto la posición calculada, i, como la posición siguiente, i+1.

El valor del percentil estará dado por el promedio de los datos que se encuentren en ambas posiciones.

La interpretación, puesto que se trata del p_{50}, es que el promedio de los valores ubicados en las posiciones 9 y 10 corresponde a un valor que es superior al 50% de los datos y, a su vez, es inferior al 50% restante.

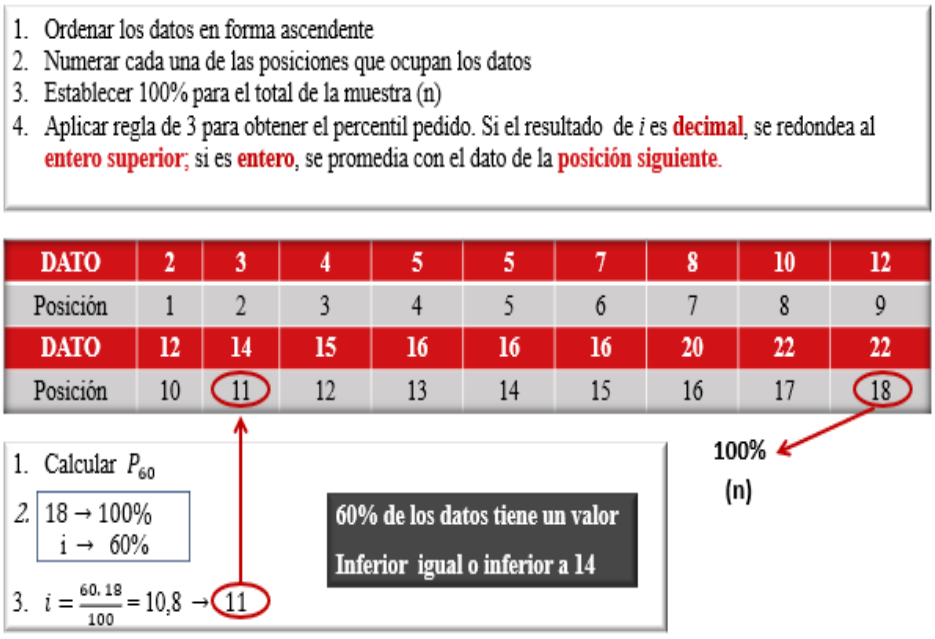
En particular, el percentil 50, $p_{50,}$ corresponde al valor de la mediana en los datos de la muestra.

Percentiles en Datos Agrupados

Cuando se desea calcular percentiles en datos agrupados en distribuciones intervalares, se utiliza la fórmula indicada en la Figura 41. Es importante señalar que en dicha fórmula el valor de la posición no se debe redondear a un valor entero, puesto que hacer esto alteraría el cálculo del percentil.

Figura 40

Ejemplo de cálculo de percentiles en datos no agrupados cuando la posición es entera



Para ubicar la clase del percentil, se procede de manera similar a cuando se calcula la mediana, sólo que ahora se debe calcular la posición a través de la fórmula señalada en la Figura 41.

Para determinar dónde se ubica la clase del percentil, se calcula la posición i, igual que cuando se tenían datos no agrupados:

$$i = \frac{k \,.\, n}{100} \qquad 2.9$$

donde k se refiere al percentil deseado, y n es el tamaño de la muestra. El valor calculado de i, debe ser ubicado en la columna correspondiente a las frecuencias absolutas acumuladas.

Figura 41

Fórmula para el cálculo de la posición en percentiles en datos agrupados en intervalos

$$P_k = Lim_{inf}(i) + \frac{\left(\frac{k\,.\,n}{100} - F_{i-1}\right) a}{f_i}$$

$Lim_{inf}(i)$: Límite inferior de la clase del percentil

F_{i-1} : Frecuencia acumulada de la clase anterior al percentil

f_i : Frecuencia absoluta de la clase del percentil

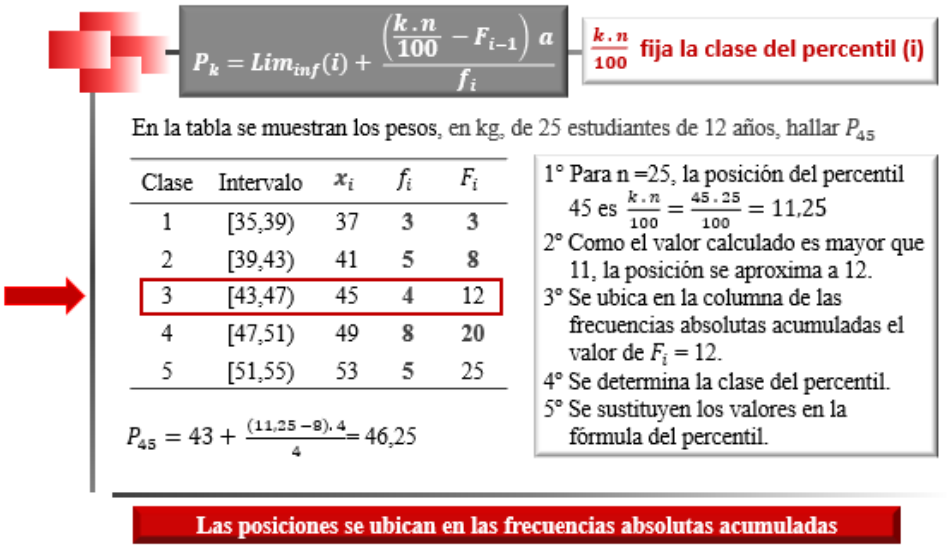
a : Amplitud del intervalo

En la Figura 42 se ilustra el cálculo del percentil 45 en una distribución intervalar. Es importante recordar que una vez que se haya determinado el valor de la posición, $\frac{k\,.\,n}{100}$, si esta no es entera, se debe aproximar al entero superior a fin de poder realizar la búsqueda de posiciones en la columna de las frecuencias absolutas acumuladas. La fila donde se encuentre la posición calculada determinará la clase que contiene al percentil buscado. Nótese en la Figura 42 que, al momento de sustituir los valores correspondientes en la fórmula del percentil para datos agrupados, el valor de $\frac{k\,.\,n}{100}$ se introduce sin aproximación, ello con el fin de obtener un cálculo más preciso. Una vez obtenido el resultado del percentil, se debe verificar que el mismo esté contenido en el intervalo de la clase del percentil.

La interpretación para el valor obtenido al calcular el percentil 45, P_{45}, es que 46,25 es mayor o igual que el 45% de los datos y que, a su vez, dicho valor es superado por el 55% de los datos restantes.

Figura 42

Ejemplo de cálculo de percentiles en distribuciones intervalares



Deciles

Son los nueve valores de la muestra que dividen a los datos en 10 partes iguales (D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9) .

Deciles para Datos no Agrupados

Para datos no agrupados, se calcula la posición del decil, i:

$$i = \frac{l \, . \, n}{10} \qquad 2.10$$

donde corresponde al número del decil l:

$$l = 1, 2, 3, 4, 5, 6 \, , 7, 8, 9 \qquad 2.11$$

y el valor de señala la posición que ocupa.

Si el resultado de la posición no es entero, se aproxima al entero **superior** y el decil D_l será el valor del dato que ocupa ese lugar.

Si el resultado de i es entero, el decil D_l será el promedio de las observaciones que ocupan los lugares i e i+1.

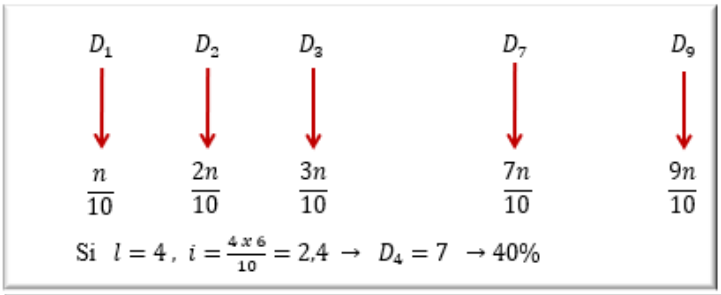
En la Figura 43 se presenta un ejemplo de cálculo del D_4 para datos no agrupados. Nótese quelos datos se encuentran ordenados y que se ha agregado una primera columna para ubicar la posición de los datos.

Una vez calculada la posición, i, puesto que se ha obtenido un valor que no es entero, éste se aproxima al número inmediato superior a fin de poder ubicarlo en la columna de posiciones. Una vez que se ha encontrado esta posición en la Tabla, en la segunda columna se lee el dato que corresponde al valor del decil buscado.

La interpretación de un decil, l, es que representa aquel valor que supera al $l * 100\%$ de los datos y que a su vez es superado por el $(1\text{-}l) * 100\%$ de los datos. Lo anterior indica que el D_5 corresponde a un valor mayor que el 50% de los datos y que, a su vez, es menor que el otro 50% de los datos. Esto significa que el Decil 5, D_5, coincide con la mediana.

Figura 43

Ejemplo de cálculo de decil en datos no agrupados



Posición	Dato
1	1
2	3
3	7
4	9
5	11
6	16

- Si la posición calculada no es entera, se toma el dato de la posición inmediatamente superior.
- Si la posición calculada es entera, se toma el promedio de los datos de la posición calculada y la posición siguiente.

Deciles para Datos Agrupados

Cuando se tienen datos agrupados, el cálculo de los deciles se realiza mediante la siguiente fórmula:

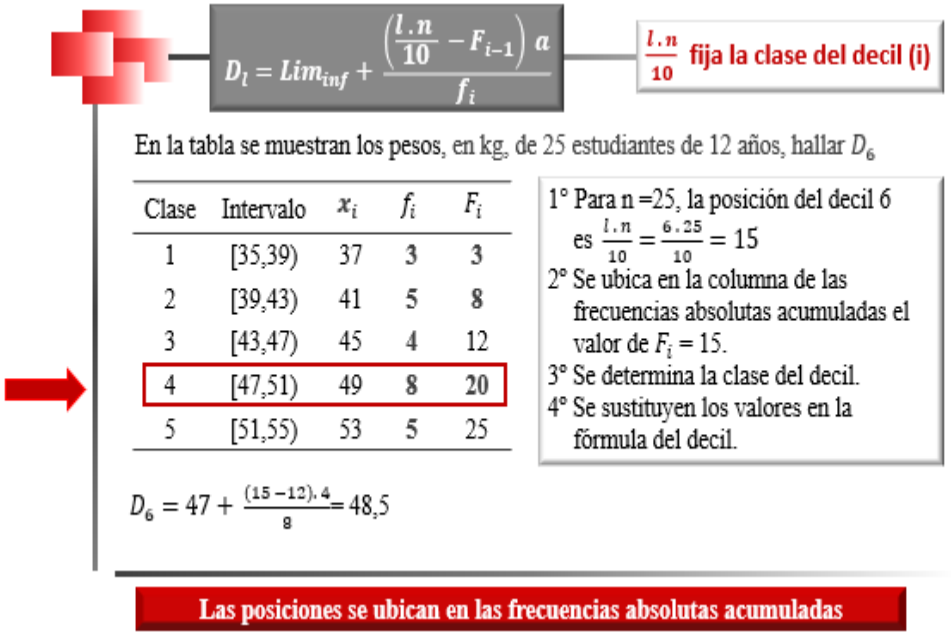
$$D_l = Lim_{inf} + \frac{\left(\frac{l \,.\, n}{10} - F_{i-1}\right) a}{f_i} \qquad 2.12$$

La clase del decil está determinada por el cálculo de la posición i.

En la Figura 44 se presenta el cálculo del decil 6, D_6, para datos agrupados en intervalos. Obsérvese que la posición calculada $\frac{l \cdot n}{10}$ es igual a 15. En este caso, se debe tener presente que en la cuarta clase, cuya frecuencia absoluta acumulada es igual a 20, se acumulan las posiciones 13, 14, 15, 16, 17, 18, 19 y 20.

Figura 44

Ejemplo de cálculo de decil en datos agrupados en distribuciones intervalares



Cuartiles

Corresponden a los tres valores que dividen a la muestra en cuatro partes iguales. Se designan a través de la letra Q acompañada del subíndice m, Q_m Donde m puede tomar los valores 1, 2, y 3.

Cada cuartil separa un 25% de los datos, así la interpretación, por ejemplo, del cuartil 1, Q_1 es que corresponde al valor que es igual a mayor que el 25% de los datos, pero que, a su vez, es inferior al 75% de ellos. El cuartil dos corresponde a la mediana $(Q_2 \equiv M_d)$.

Cuartiles para Datos no Agrupados

Recuérdese que los datos no agrupados son aquellos que no han recibido ningún tipo de tratamiento estadístico, salvo, ordenarlos de forma creciente o decreciente.

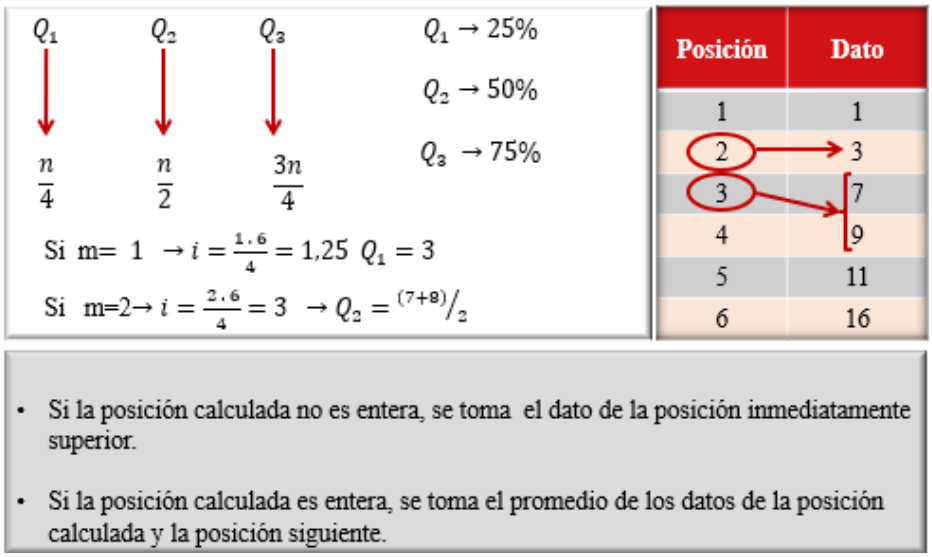
Para datos no agrupados, el cálculo de la aplicación del cuartil m, se realiza a través de la siguiente fórmula:

$$i = \frac{m \,.\, n}{4} \qquad 2.13$$

donde m es el número del cuartil, m = 1, 2, y 3, y el valor de i representa la posición del dato.

En la Figura 45 se presenta un ejemplo de cálculo de los cuartiles 1 y 2 en datos no agrupados.

Figura 45
Ejemplo de cálculo de cuartil en datos no agrupados



Obsérvese en la Figura 45 que en el cálculo del cuartil 1 la posición debe ser redondeada a 2 y el valor del cuartil se lee en la columna de los datos. $Q_1 = 3$. Para el cuartil 2, la posición resulta entera, i = 3, por lo que, para hallar el valor del cuartil 2, se deberá promediar los valores que se encuentran en las posiciones 3 y 4 de la columna de datos.

Como el cuartil por definición divide a los datos en cuatro partes iguales, resulta claro que cada parte contiene al 25% de los datos (100/4). Ahora bien, la interpretación del cuartil 2 será entonces que es aquel valor que supera al 50% de los datos (25*2) y que a su vez es superado por el 50% de los datos restantes. Es por ello que el cuartil 2 corresponde a la mediana.

Cuartiles para Datos Agrupados

Cuando los datos se encuentran agrupados en intervalos, el cálculo se realiza a través de la siguiente fórmula:

$$Q_m = Lim_{\inf(i)} + \frac{\left(\frac{m\,.\,n}{4} - F_{i-1}\right) a}{f_i} \qquad 2.14$$

donde $Lim_{\inf(i)}$ representa el límite inferior de la clase del cuartil, m el número del cuartil; f_i es la frecuencia absoluta de la clase del cuartil, F_{i-1} es la frecuencia absoluta acumulada de la clase anterior a la clase del cuartil y a es la amplitud del intervalo.

En la Tabla 13 se presenta un ejemplo de cálculo del cuartil 1 en tres muestras de diferentes tamaños.

Tabla 13

Ejemplo de cálculo del cuartil 1 en tres muestras distintas

	Muestra 1	**Muestra 2**	**Muestra 3**
n	50	80	39
(1 . n)/4	12,5	20	9,75
Ubicación	Posición 13	Entre posiciones 20 y 21	Posición 10
Valor	Dato que ocupa la posición 13	Semisuma de los datos de las posiciones 20 y 21	Dato que ocupa la posición 10

En la Figura 46 se presenta un cuadro resumen de las características principales de los cuantiles que se han estudiado. Se incluyen las fórmulas para el cálculo de la posición y para los valores de cada cuantil en el caso en que se tengan distribuciones intervalares.

Es conveniente recordar que cuando se tienen datos agrupados, la clase del cuantil está determinada por el cálculo de la posición.

Figura 46

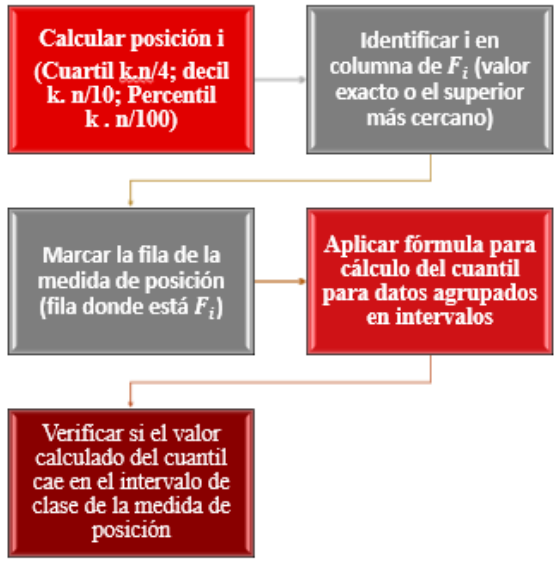
Cuadro resumen de las características de los cuantiles

PERCENTILES	DECILES	CUARTILES
$\frac{k.n}{100}$	$\frac{l.n}{10}$	$\frac{m.n}{4}$
99 percentiles	9 deciles	3 cuartiles
10 partes, 1% cada una	9 partes, 10% cada una	4 partes, 25% cada una
$P_k = Lim_{inf(i)} + \frac{\left(\frac{k.n}{100} - F_{i-1}\right) a}{f_i}$	$D_l = Lim_{inf(i)} + \frac{\left(\frac{l.n}{10} - F_{i-1}\right) a}{f_i}$	$Q_m = Lim_{inf(i)} + \frac{\left(\frac{m.n}{4} - F_{i-1}\right) a}{f_i}$

En la Figura 47, se presenta una secuencia de pasos para el cálculo de los cuantiles, específicamente de los percentiles, de los deciles y de los cuartiles; ya que hay otros. Recuérdese que, para identificar la clase del cuantil, se procede a calcular la posición i, y, posteriormente, se ubica el valor correspondiente (redondeado al entero superior si no resulta entero), en la columna de las frecuencias absolutas acumuladas.

Figura 47

Procedimiento general para el cálculo de cuantiles



En las Figuras 48 a 51, se incluye el desarrollo de un ejemplo para datos agrupados en intervalos.

Figura 48

Datos para ejemplo de cálculo de cuantiles

Determinar Q_1, D_4 y P_{30}

Salarios (I. De Clases)	f_i (No. de Empleados)	F_i
[200, 300)	85	85
[300, 400)	90	175
[400, 500)	120	295
[500, 600)	70	365
[600, 700)	62	427
[700, 800)	36	463

Figura 49

Ejemplo de cálculo de cuartiles en datos agrupados en intervalos

n= número de datos= 463 Para Q_1, $i = \frac{1 \cdot 463}{4} = 115{,}75$

$$Q_m = Lim_{inf} + \frac{\left(\frac{m \cdot n}{4} - F_{i-1}\right) a}{f_i} \quad Q_1 = 300 + \frac{(115{,}75-85).100}{90} = 334{,}2$$

Salarios (I. De Clases)	f_i (No. de Empleados)	F_i
[200, 300)	85	85
[300, 400)	90	175
[400, 500)	120	295
[500, 600)	70	365
[600, 700)	62	427
[700, 800)	36	463

Figura 50

Ejemplo de cálculo de decil en datos agrupados en intervalos

n= número de datos= 463 Para D_7, $i = \frac{7 \cdot 463}{10} = 324{,}1$

$$D_i = Lim_{inf} + \frac{\left(\frac{l \cdot n}{10} - F_{i-1}\right) a}{f_i} = 500 + \frac{(324{,}1-295) \cdot 100}{70} = 541{,}57$$

Salarios (I. De Clases)	f_i (No. de Empleados)	F_i
[200, 300)	85	85
[300, 400)	90	175
[400, 500)	120	295
[500, 600)	70	365
[600, 700)	62	427
[700, 800)	36	463

Figura 51

Ejemplo de cálculo de percentil en datos agrupados en intervalos

n= número de datos= 463 Para P_{30}, $i = \frac{30 \cdot 463}{100} = 138{,}9$

$$P_k = Lim_{\inf(i)} + \frac{\left(\frac{k \cdot n}{100} - F_{i-1}\right) a}{f_i} \quad P_{30} = 300 + \frac{(138{,}9-85) \cdot 100}{90} = 359{,}88$$

Salarios (I. De Clases)	f_i (No. de Empleados)	F_i
[200, 300)	85	85
[300, 400)	90	175
[400, 500)	120	295
[500, 600)	70	365
[600, 700)	62	427
[700, 800)	36	463

Los resultados indican que el 25% de los empleados ganan salarios por debajo de $ 334,2; que el 70%de los empleados gana bajo $ 541.57 y que el 70% de los empleados gana sobre $359,88

En la Figura 52 se muestra un ejemplo del cálculo del percentil 30 para una distribución intervalar, en donde la posición calculada, i, coincide con un valor de la columna de frecuencias absolutas acumuladas. Nótese que, debido a esto, lo más correcto en estos casos es pasar la clase del percentil a la posición i+1.

Siendo el valor de i igual a 60, la posición siguiente sería la i+1 = 61, en consecuencia, esa clase i +1 correspondería al intervalo [4,6). Recuérdese que la interpretación de la frecuencia absoluta acumulada correspondiente a ese intervalo, que tiene un valor de 100, es que allí se cuentan las observaciones comprendidas entre 61 y 100.

El cálculo del P_{30}, para este ejemplo coincide con el límite inferior de la clase del percentil.

Es importante volver a destacar que el cálculo de cualquier cuantil, en una distribución intervalar, debe caer siempre dentro del intervalo de la clase del cuantil.

Figura 52

Ejemplo de cálculo de percentil con posición existente en la columna de frecuencias absolutas acumuladas

n= número de datos= 200 Para $P_{30,}$ $i = \frac{30 \cdot 200}{100} = 60$

$$P_k = Lim_{\inf(i)} + \frac{\left(\frac{k \cdot n}{100} - F_{i-1}\right) a}{f_i} = 4 + \frac{(60-60).2}{40} = 4$$

Calificaciones	No. de estudiantes (f)	F
[0, 2)	25	25
[2, 4)	35	60
[4, 6)	40	100
[6, 8)	65	165
[8,10)	35	200

Obsérvese que si no se hubiese incrementado la posición para el P_{30}, es decir si el valor de i se hubiese mantenido en 60, la clase del percentil no habría sido la resaltada en la Figura 52, sino la precedente, es decir, la correspondiente al intervalo [2, 4). Si en estas condiciones se aplicara la fórmula para el cálculo del percentil correspondiente a datos agrupados en intervalos, se tendría:

$$P_k = Lim_{inf}(i) + \frac{\left(\frac{k\,.\,n}{100} - F_{i-1}\right)\,a}{f_i} = 2 + \frac{\left(\frac{30\,.\,200}{100} - 25\right)}{35}.2 = 4 \qquad 2.15$$

Nótese que el valor obtenido para el percentil es exactamente el mismo que se tendría si se hubiera tomado la posición i igual a 61; sin embargo, el resultado obtenido = 4, no pertenecería al intervalo de la clase del percentil, ya que en el intervalo [2,4) el 4 no está incluido.

Si, en el ejemplo de la Figura 52, adicionalmente se quisiera encontrar el percentil 85, bastaría con calcular la posición correspondiente:

$$i = \frac{85\,.\,200}{100} = 170 \qquad 2.16$$

Obsérvese que, para este valor, la clase del percentil queda ubicada en el último intervalo (que acumula los datos desde el 166 hasta el 200), para el cual:

$$P_{85} = 8 + \frac{\left(\frac{85\,.\,200}{100} - 165\right)}{35}.2 = 8{,}29 \qquad 2.17$$

Con el valor obtenido se corrobora su pertenencia al intervalo [8,10) que corresponde a la clase del percentil 85.

Medidas de Dispersión para datos agrupados y no agrupados

Las medidas de dispersión indican el alejamiento que pueden experimentar los datos con respecto a la media. Indican la variación que presenta la variable en estudio, Figura 53.

Las poblaciones de la figura poseen la misma media, pero la dispersión de los datos es distinta.

Obsérvese que hay una distribución con σ = 10 y otro con σ = 50 Mientras más pequeño sea el valor de σ las distribuciones son más empinadas y más estrechas en la base. En cambio, a medida que la desviación típica, σ se

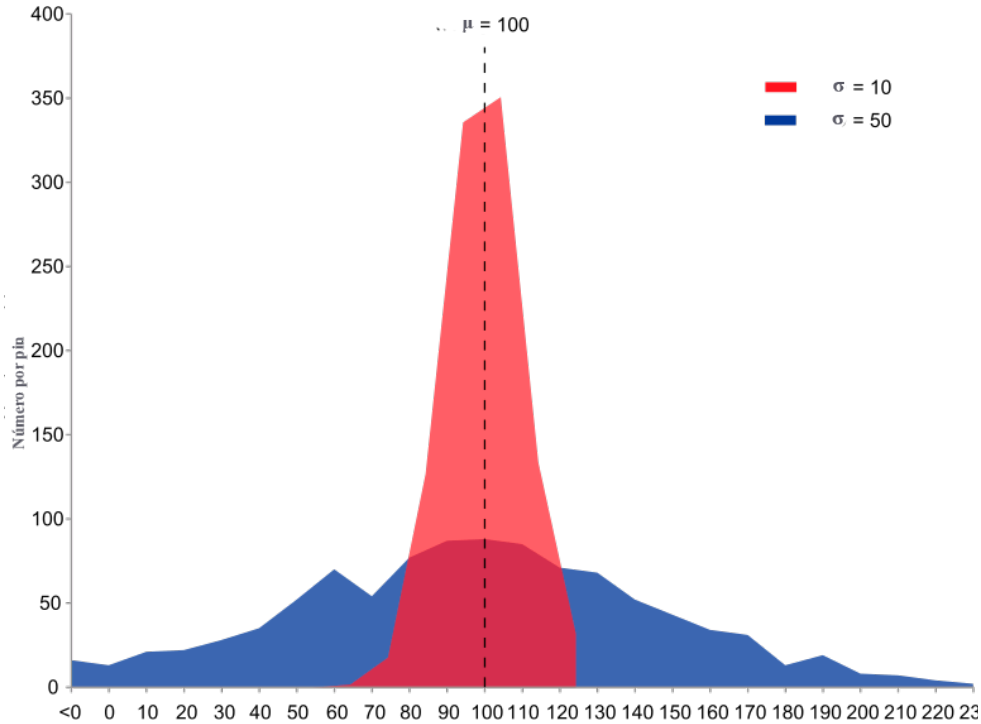
hace mayor, las distribuciones son menos altas. Pero más dispersas, esto indica que los datos son más heterogéneos.

Las medidas de dispersión (Dodge, 2003) que se estudiarán son:

1. Rango
2. Desviación típica
3. Varianza
4. Coeficiente de dispersión

Figura 53

Comparación de la dispersión presentada en dos poblaciones que tienen la misma mediaaritmética



Nota. Adaptado de Medidas de dispersión [Gráfico], JRBrown, 1 de julio de 2010, Wikipedia (https://en.wikipedia.org/wiki/Statistical_dispersion). CC BY 3.0

Rango

El rango mide la amplitud de los valores de la muestra y se calcula por la diferencia entre elvalor más elevado y el valor más bajo de la distribución estadística.

$$R = X_{máx} - X_{min} \qquad 2.18$$

En la Figura 54, se presenta un conjunto de datos. Obsérvese que los valores mínimos y máximos están resaltados. Para hallar el valor del rango, se realiza la diferencia entre el valor máximo y el valor mínimo del conjunto de datos.

Figura 54
Determinación del rango en un conjunto de datos no agrupados

108	144	103	88	91	125	109	114	108	141
130	155	105	114	123	112	110	107	121	**155**
92	112	129	137	112	100	101	112	129	106
104	116	88	123	120	136	109	119	119	118
98	**85**	104	96	115	121	99	91	100	111
102	106	123	109	151	118	87	93	115	90
134	111	129	125	116	132	117	102	109	112

R = 155 -85 = 70

Varianza para Datos no Agrupados y Agrupados

La varianza es una medida estadística de la dispersión o variabilidad de un conjunto de datos. No tiene una interpretación física, porque sus unidades vienen en una dimensión superior a la unidad de medida; esto es, si los datos se miden en centímetros, la varianza estará en centímetros cuadrados.

A pesar de ello, la varianza es la medida de dispersión más importante, porque posee unas propiedades que no son atribuibles a ninguna otra medida de dispersión.

La varianza es siempre un número no negativo debido a que procede de una relación que tiene por numerador una diferencia al cuadrado y por denominador una cantidad que refleja el tamaño de la muestra o la población estudiada.

Como la varianza es una medida de dispersión, si todos los datos de una variable en una muestra fueran iguales, la varianza sería nula.

La varianza es sensible a los valores extremos o atípicos, esto indica que si hay valores extremos muy grandes o muy pequeños en los datos, entonces la varianza será mayor que si los datos son más uniformes.

Cuando se tienen datos no agrupados, la varianza de una población puede ser obtenida a través de la relación:

$$\text{Varianza} = \frac{\sum(x_i-\mu)^2}{N} \qquad 2.19$$

Donde

$\chi\, i$ corresponde al valor del dato *i*

μ es la media de la población

N tamaño de la población

Para el caso de datos agrupados, los cuadrados individuales del numerador se verán afectados por la frecuencia:

$$\text{Varianza} = \frac{\sum(x_i-\mu)^2 f_i}{N} \qquad 2.20$$

donde *f i* corresponde a la frecuencia absoluta para la i-ésima observación.

A pesar de que la varianza no tiene una interpretación física, es una medida de dispersión que tiene propiedades matemáticas útiles y bien definidas. Por ejemplo, la varianza de la suma de dos variables aleatorias es igual a la suma de sus varianzas cuando las variables son independientes. Esto permite realizar cálculos y análisis estadísticos más complejos y sofisticados

La varianza se puede utilizar para calcular otras medidas de dispersión, como la desviación estándar y el coeficiente de variación.

Otro dato importante respecto a la varianza es que tiene una distribución probabilística definida que corresponde a la suma de los cuadrados de variables aleatorias independientes e idénticamente distribuidas con media

cero y varianza uno. Esta distribución se conoce como Chi-cuadrado y se utiliza en muchos análisis estadísticos, como en las pruebas de hipótesis y en la construcción de intervalos de confianza.

Desviación Típica o Estándar

La desviación estándar, σ es una medida muy útil porque tiene las mismas unidades de medida que la variable original, lo que la hace fácil de interpretar. Por ejemplo, si estamos midiendo el peso de los estudiantes de un determinado curso y encontramos una desviación estándar de 5 kilos, podemos interpretar que la mayoría de los estudiantes tienen un peso que varía en +/- 5 kilos de la media.

La desviación estándar corresponde al promedio de las distancias de cada dato con relación a la media. A medida que la desviación típica sea mayor, se incrementará la dispersión en los datos; por el contrario, cuando se tienen desviaciones típicas menores, existe mayor homogeneidad entre los datos de la muestra o la población en estudio.

La desviación típica puede ser interpretada como una medida de incertidumbre y su valor numérico corresponde a la raíz positiva de la varianza:

$$\sigma = \sqrt{Varianza} \qquad 2.21$$

La desviación estándar en las poblaciones suele denotarse con la letra griega sigma en minúscula, σ, mientras que para las muestras se utiliza la letra S, en mayúscula.

Para el cálculo de esta medida de dispersión, se utilizan las siguientes fórmulas (Pearson, 1895):

Para Datos No Agrupados $\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$ 2.22

Para Datos Agrupados $\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2 f_i}{N}}$ 2.23

$donde\ x$i = marca de clase si los datos están agrupados en intervalos, μ es la media y N el tamaño de la población.

En la Figura 55 se presenta un cuadro comparativo de las fórmulas que

deben ser usadas dependiendo si se trata de una medida de dispersión para una población de tamaño N o para una muestra de tamaño n.

Figura 55

Fórmulas para el cálculo de la desviación típica y de la varianza en poblaciones y muestras

		POBLACIÓN	MUESTRA
Varianza	Datos no agrupados	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
	Datos Agrupados	$\sigma^2 = \frac{\sum(x_i - \mu)^2 f_i}{N}$	$S^2 = \frac{\sum(x_i - \bar{x})^2 f_i}{n-1}$
Desviación típica	Datos no agrupados	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$
	Datos agrupados	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2 f_i}{N}}$	$S = \sqrt{\frac{\sum(x_i - \bar{x})^2 f_i}{n-1}}$

Obsérvese que, en el caso de las fórmulas para una muestra, se divide el numerador entre n-1, lo cual significa que debe disminuirse en una unidad el tamaño de la muestra. Esta desviación típica de la muestra también es conocida con el nombre de cuasidesviación.

Téngase en cuenta que, x_i, para datos no agrupados o agrupados puntualmente, representa el valor de la variable en estudio, mientras que, para datos agrupados, es la marca de clase. Por su parte, μ, es el valor de la media poblacional y $\bar{x}$, la media de la muestra; f_i, es la frecuencia absoluta correspondiente a la variable x_i; N el tamaño de la población; y n, el tamaño de la muestra.

Obsérvese que, en caso de las muestras, no se usan letras griegas ni para la desviación típica ni para la varianza y que el denominador es igual al tamaño de la muestra disminuido en una unidad.

Coeficiente de Variación de Pearson.

Esta medida de dispersión se expresa a través de la relación entre la desviación típica de una población o una muestra y su media aritmética (Pearson, 1892). Las fórmulas a utilizar para su cálculo serán:

Para una población $$CV = \frac{\sigma}{\mu}$$ 2.24

Para una muestra $$CV = \frac{S}{\bar{x}}$$ 2.25

Nótese que, debido a que la desviación típica y la media aritmética poseen las mismas unidades,cuando se realiza el cociente entre ellas el resultado es adimensional. En este hecho radica la fortaleza de esta medida de dispersión, ya que es el único que permite realizar una comparación de las variaciones existentes en muestras que se expresan en diferentes unidades.

A mayor valor del coeficiente de variación, mayor heterogeneidad existirá en los datos. Si se desea expresar el coeficiente de variación de manera porcentual, basta con multiplicar por 100 las fórmulas anteriores.

Cuando el valor del coeficiente de variación se aproxima a cero (condiciones ideales en las que no existe dispersión), la muestra es compacta (homogénea), por el contrario, si el coeficiente de variación tiende a 1- significa que el valor de la desviación típica es tan grande como la media aritmética, por lo tanto, los datos son muy dispersos (condición extrema). En general, si el coeficiente de variación de Pearson es mayor que el 30%, se considera que la media no es representativa de la población o la muestra.

Las condiciones de homogeneidad son deseables porque mientras más homogénea sea una población, o una muestra, los datos se van a distribuir en mayor medida alrededor del valor promedio y ello ofrece garantía de que la media aritmética sea un valor representativo del conjunto de datos.

En las Figura 56 y 57 se desarrolla un ejemplo de cálculo de medidas de dispersión para una muestra de datos no agrupados.

Figura 56

Datos para el cálculo de medidas de dispersión en muestra de datos no agrupados

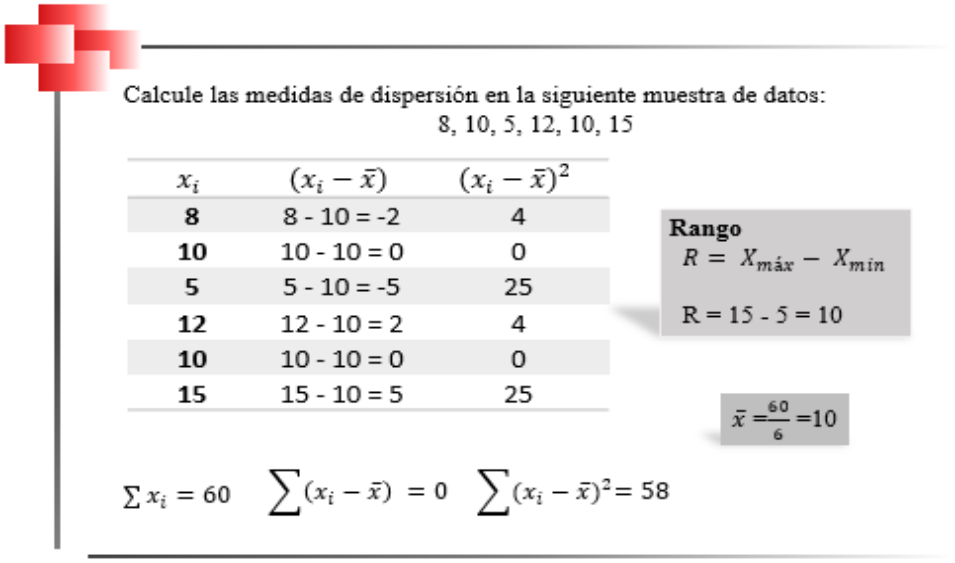
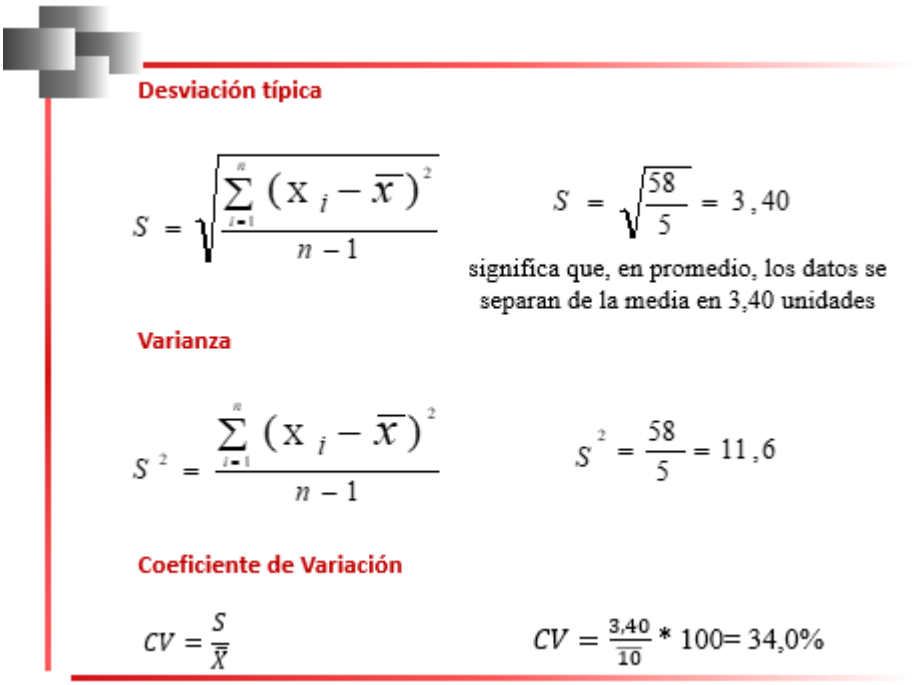


Figura 57

Cálculo de medidas de dispersión en muestra de datos no agrupados de la Figura 56



En las Figuras 58 y 59 se presenta un ejemplo de cálculo de medidas de dispersión en una población cuyos datos se encuentran agrupados en intervalos.

Para este caso, como los datos se encuentran agrupados en intervalos, el valor de corresponde a la marca de clase y no a un dato en particular. Recuérdese que esta marca de clase se obtiene promediando los límites superior e inferior de cada intervalo.

Una vez obtenida la marca de clase, se procede a calcular la media aritmética de la población. Se debe tener presente que, por tratarse de datos agrupados, la media está afectada por las frecuencias de cada clase.

Obsérvese que en la Figura 58 se ha calculado una columna que corresponde al producto entre la marca de clase y la frecuencia absoluta, $x_i \cdot f_i$, si se suman todos los productos de esta columna y se divide el total entre el número de datos, se tendrá la media de la población.

Figura 58
Datos para el cálculo de medidas de dispersión en una distribución intervalar

	x_i	f_i	$x_i \cdot f_i$	$(x_i - \mu)^2$	$(x_i - \mu)^2 f_i$
[10, 20)	15	1	15	802,59	802,59
[20, 30)	25	8	200	335,99	2687,91
[30, 40)	35	10	350	69,39	693,89
[40, 50)	45	9	405	2,79	25,10
[50, 60)	55	8	440	136,19	1089,51
[60, 70)	65	4	260	469,59	1878,36
[70, 80)	75	2	150	1002,99	2005,98
		N=42	1820		9183,33

$$\mu = \frac{\sum x_i \cdot f_i}{N} = \frac{1820}{42} = 43{,}3$$

Figura 59

Cálculo de medidas de dispersión para datos agrupados en intervalos

	$x_i . f_i$	$(x_i - \mu)^2 f_i$
[10, 20)	15	802,59
[20, 30)	200	2687,91
[30, 40)	350	693,89
[40, 50)	405	25,10
[50, 60)	440	1089,51
[60, 70)	260	1878,36
[70, 80)	150	2005,98
	1820	9183,33

$$\mu = \frac{\sum x_i . f_i}{N} \qquad \mu = \frac{1820}{42} = 43,33$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{N}} = \sqrt{\frac{9183,33}{42}} = 14,79$$

$$\sigma^2 = (14,79)^2 = 218,74$$

$$CV = \left(\frac{\sigma}{\mu}\right) . 100 = \frac{14,79}{43,33} . 100 = 34,15\%$$

En la Figura 60, se ilustra la comparación de los coeficientes de variación en tres muestras expresadas en unidades distintas. Nótese que esta comparación sólo es posible por el hecho de que el coeficiente de variación de Pearson es una medida de dispersión adimensional. De hecho, dada esta condición, es la única medida de dispersión que puede ser usada para comparar la dispersión existente entre datos de muestras o poblaciones cuyas unidades de medida son distintas.

Figura 60

Comparación de coeficientes de variación en tres muestras expresadas en unidades diferentes

	Muestra 1	Muestra 2	Muestra 3
Desviación estándar	0.28 m	3.8 kg	2.5 años
media	1.24 m	24 kg	9.7 años
CV	0.18/1.32 =0.226	3.8/24 =0.158	3.5/ 9.7=0.361

El coeficiente de variación permite comparar dispersiones entre datos expresados en escalas diferentes.

Se puede concluir que los datos de la muestra 3 tienen mayor dispersión que los de las muestras 1 y 2, siendo la muestra 2 la que posee los datos más homogéneos.

El valor del CV de la muestra 3 permite concluir que la media no es representativa.

A continuación, en la Figura 61, se ilustra un ejemplo de la aplicación práctica de las medidas de dispersión en la toma de decisiones.

Supóngase que se tiene una competencia femenina de videojuegos educativos a nivel estatal y se desea enviar a una representante al concurso. Se cuenta con tres posibles candidatas a las que se ha registrado su rendimiento en cada uno de los videojuegos. El concurso estipula que serán evaluadas las destrezas de las participantes sólo en tres de los videojuegos, los cuales serán seleccionados al azar del grupo total de 10 videojuegos. Con base en los conceptos de estadística descriptiva, seleccione a cuál de las tres participantes debe enviarse al concurso a fin de tener la mejor representación.

Obsérvese que los promedios de los puntajes obtenidos en los 10 videojuegos son exactamente los mismos en las tres candidatas (***μ=53***), por lo que la elección será aquella concursante que tenga la menor dispersión entre los puntajes obtenidos en cada videojuego, a fin de garantizar que, sin importar cuáles sean los tres videos elegidos para la prueba, la representante de la institución pueda obtener el mejor rendimiento.

Figura 61

Comparación de medidas de tendencia central y de dispersión entres grupos de datos

Videojuego	Mercedes	Carla	Verónica
1	50	100	20
2	50	0	20
3	60	80	80
4	50	70	70
5	60	30	40
6	50	60	80
7	50	80	70
8	50	70	30
9	60	10	40
10	50	30	80
	$\mu = 53$ $\sigma = 4.58$	$\mu = 53$ $\sigma = 31,64$	$\mu = 5,3$ $\sigma = 24,10$

$$\mu = \frac{\sum x_i}{N} \qquad \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

La representante a ser enviada al concurso debe ser Mercedes; ya que sus puntajes en los videojuegos presentan mucha mayor homogeneidad que las otras dos concursantes y, de esta manera, se tiene menor incertidumbre en los resultados que se obtendrán.

Capítulo 3

Distribución de probabilidades

CAPÍTULO III

DISTRIBUCIÓN DE PROBABILIDADES

Conceptos Básicos de Probabilidades

Técnicas de Conteo: Permutación y Combinación

En estadística, las técnicas de conteo son empleadas para determinar el total de posibilidades existentes al momento de realizar un determinado experimento.

Cuando hay pocos datos, es sencillo determinar de cuántas maneras distintas se les puede colocar, de tal forma que todos los posibles arreglos resulten diferentes.

Por ejemplo, si se quiere saber cuántos números distintos se pueden obtener cambiando la posición de los dígitos 1, 2 y 3, se podría construir con facilidad el siguiente resultado:

$$123 - 132 - 213 - 231 - 312 - 321 \qquad 3.1$$

Sin embargo, a medida que el número de datos se incrementa o se involucran condiciones específicas en relación a la posición que debería ocupar cada elemento, la tarea comienza a adquirir un mayor grado de complejidad.

En este sentido, la permutación y la combinación son técnicas de conteo que permiten determinar el número de posibilidades existentes en la disposición de ciertos elementos (Pascal, 1665).

Antes de comenzar a definir los conceptos de permutación y combinación, es necesario abordar un concepto base en la teoría de probabilidades.

Factorial de n (n!)

El factorial de un número entero, n, se forma como el producto de los números consecutivos iniciando en 1 y terminando en n (Kramp, 1808):

$$n! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot \ldots n \qquad 3.2$$

Por definición, el factorial de 0! Es igual a 1: 0! = 1 De acuerdo con lo anterior, se tendría que:

$$8! = 1.\ 2.\ 3.\ 4.\ 5.\ 6.\ 7.\ 8 \qquad 3.3$$

Aunque para efectos prácticos, suele escribirse al revés:

$$8! = 8.\ 7.\ 6.\ 5.\ 4.\ 3.\ 2.\ 1 \qquad 3.4$$

La razón que justifica el hecho de esta escritura inversa es que cuando se tienen cocientes entre factoriales, la forma de simplificarlos es abrir y cerrar los factoriales. Lo anterior indica que el factorial de 8, por ejemplo, puede escribirse de varias maneras:

$$\begin{gathered} 8! = 8.\ 7! \\ 8! = 8.\ 7.\ 6! \\ 8! = 8.\ 7.\ 6.\ 5! \\ 8! = 8.\ 7.\ 6.\ 5.\ 4! \\ 8! = 8\ .\ 7\ .\ 6\ .\ 5\ .\ 4\ .\ 3! \\ 8! = 8\ .\ 7\ .\ 6\ .\ 5\ .\ 4\ .\ 3\ .\ 2! \end{gathered} \qquad 3.5$$

Permutaciones

Permutar significa cambiar el orden en el que se colocan dos o más elementos (Feller, 1950). Esta referencia a cambio de posición significa que un concepto clave para identificar las permutaciones es que en ellas importa el orden en el que se disponen dichos elementos (objetos, animales o cosas).

En la Figura 62, se presenta la imagen correspondiente a un alfabeto. Nótese que este es un ejemplo típico de elementos en los que se debe aplicar la permutación, ya que el orden en el que se coloquen las letras va a cambiar por completo su significado.

Considérese, por ejemplo, las letras AES, en ese caso, como el orden en que se coloquen las letras es relevante, se aplica la permutación que consiste en hallar todas las disposiciones posibles de las letras que pueden conseguirse cambiando el orden de los elementos, siempre que no existir repetición:

$$\text{EAS-ESA-SEA-SAE-AES-ASE} \qquad 3.6$$

Tipos de Permutación. Pueden existir tres tipos de permutación: Permutación sin repetición, permutación con repetición y permutaciones circulares.

Figura 62

Alfabeto, ejemplo de elementos que pertenecen a una permutación

Fuente: Letras A B C [Imagen] Gerd Altmann 25 de mayo de 2020, Pixabay,https://pixabay.com/es/illustrations/letras-a-b-c-capacitaci%C3%B3n-alfabeto-5216916/ Pixabay License.

Las permutaciones sin repetición y con repetición, hacen referencia, como su nombre lo indica, al hecho de que puedan o no repetirse los elementos en la combinación. Cuando la repetición es posible, entonces hay que considerar todas las disposiciones que resultan involucradas, aún aquellas en las que puede repetirse el mismo elemento para todas las posiciones.

Por ejemplo, en el caso visto anteriormente donde se tienen las letras AES, tendrían que considerarse, entre otras, las disposiciones repetidas de una sola letra: AAA- EEE- SSS.

Permutaciones sin Repetición. En las permutaciones sin repetición, es posible distinguir dos casos: 1) Aquellos en los que en la permutación entran todos los elementos y 2) aquellos en los que, del total de elementos disponibles, n, se van formando grupos de k elementos.

En el primer caso, para hallar el número total de posibilidades, se utiliza la fórmula:

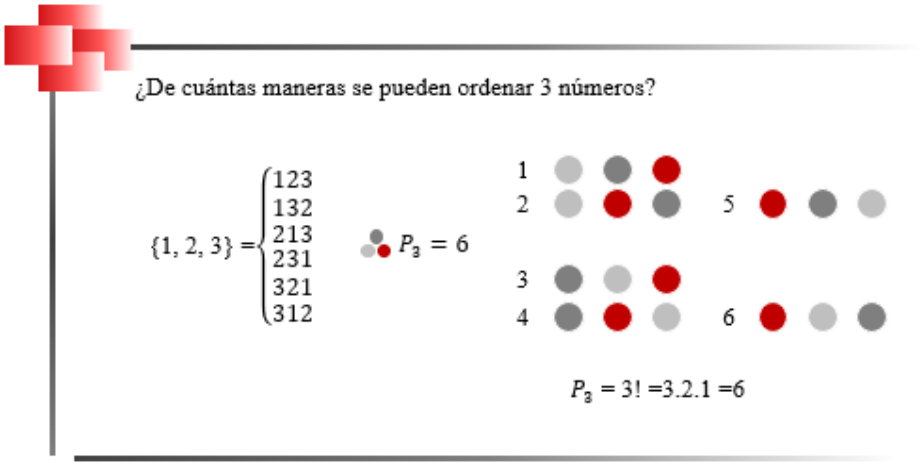
$$P_n = n! \qquad 3.7$$

Por el contrario, cuando se considera un grupo k, del total de elementos disponibles para la permutación, el cálculo de posibilidades totales se calcula a través de la fórmula:

$$P_{k,n} = \frac{n!}{(n-k)!} \qquad 3.8$$

Sean, por ejemplo, los números 1, 2 y 3, Figura 63.

Figura 63
Permutación sin repetición de 3 números



Si lo que se desea es calcular el número de disposiciones totales sin repetición, incluyendo, al mismo tiempo, todos los elementos, se tendría:

$$P_3 = 3! = 6 \qquad 3.9$$

$$123 - 132 - 213 - 231 - 312 \text{ -}321 \qquad 3.10$$

Considérese ahora el caso en el que se desea calcular el número de posibilidades de tomar, sin repetición, grupos de 2 números, teniendo a disposición los números 1, 2 y 3.

En este caso, el cálculo, resultaría de la siguiente forma:

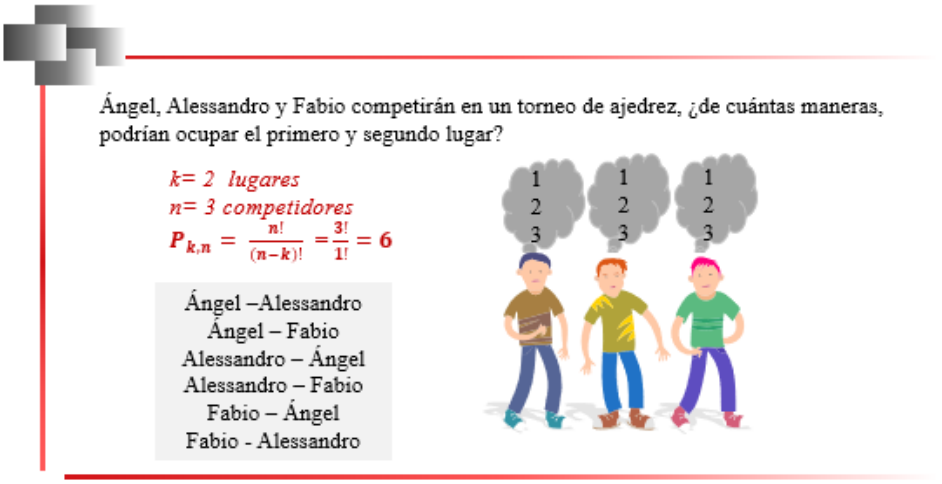
$$P_{2,3} = \frac{3!}{(3-2)!} = \frac{6}{1} = 6 \qquad 3.11$$

$$12 - 21 - 13 - 31 - 23 - 32 \qquad 3.12$$

En la Figura 64 se ilustra otro ejemplo de permutaciones de n elementos con k posiciones.

Figura 64

Ejemplo de permutaciones sin repetición, de n elementos con k posiciones



Nota. Imagen incluida: Tres adolescentes de la historieta [Imagen],Publicdomainvectors. org, 18 de diciembre de 2017, Publicdomainvectors.org(https://publicdomainvectors.org/ es/vectoriales-gratuitas/Tres-adolescentes-de-la-historie ta/68522. html). CC0 1.0.

Permutaciones con Repetición. En permutaciones con repetición, también se distinguen dos casos: 1) cuando no hay elementos repetidos en el grupo, pero es posible repetir la selección y 2) cuando hay elementos repetidos en el grupo y cada uno de ellos se repite k_1, k_2, k_3… veces.

Para ilustrar el primer caso, considérense que se desea generar una clave de 8 caracteres alfanuméricos para el ingreso a un examen en línea. Se dispone de las siguientes letras y números A, B, C, 1, 2, 3. Obsérvese que ninguno de los elementos está repetido; pero en la generación de la clave sí es posible repetir cualquiera de los caracteres.

Como la estructura de la clave tiene ocho espacios, cada uno debe serllenado por una letra o un número. Dado que todos los elementos pueden repetirse, todos ellos pueden ser utilizados en cada uno de los espacios a llenar para la generación de la clave de 8 caracteres alfanuméricos. Entonces, para el primer carácter alfanumérico se dispondrá de 6 opciones (A, B, C, 1, 2 y 3), pero, a su vez, las mismas 6 opciones estarán disponibles para cualquiera de los otros siete caracteres.

Lo anterior indica, que la fórmula a utilizar para generar la claves, será:

$$\boldsymbol{PR_k = n^k} \qquad 3.13$$

donde n sería el número de elementos disponibles y k el número de caracteres alfanuméricos necesarios para la clave que se desea generar (número de veces que se repite el experimento).

Realizando el cálculo, el número de posibilidades que se tendría para generar la clave del examen, sería:

$$\boldsymbol{PR_8 = 6^8 = 1.679.616} \qquad 3.14$$

Para el segundo caso, considérese, como ejemplo, la palabra "pepperoni". Acá se tendría quen =9, puesto que existen 9 letras en la palabra. Se trata de una permutación, porque importa el orden en el que son colocadas las letras, y, además, es una permutación con repetición, porque hayletras que se repiten en la palabra.

Se especifica el número de veces que se encuentran letras repetidas:

$$p = 3 \text{ veces};\ \ e = 2 \text{ veces} \qquad 3.15$$

A continuación, se aplica la fórmula para permutaciones con repetición:

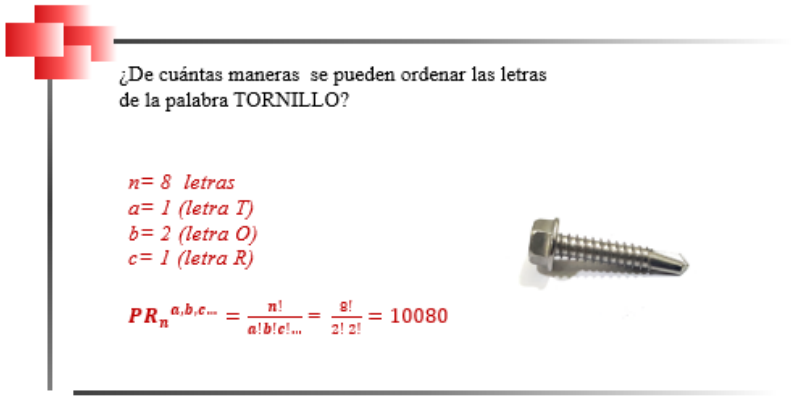
$$PR_9^{3,2} = \frac{9}{3!2!} = \frac{9.8.7.6.5.4.3!}{3!\ 2!} = 30240 \qquad 3.16$$

Lo anterior indica que existen 30240 posibilidades distintas de colocar las letras de la palabra "pepperoni".

En la Figura 65, se presenta otro ejemplo de permutaciones con repetición de elementos.

Figura 65

Ejemplo de permutaciones con repetición de elementos dentro del conjunto de datos



Nota. Imagen incluida: Tornillo autoperforante de cabeza hexagonal [Imagen], ChZaheer, 25 de agosto de 2016, Pixabay (https://pixabay.com/es/photos/tornillo-autoperforante-de-cabeza-hexago-1614969/). Pixabay license.

Permutaciones circulares. En este tipo de permutación los elementos se disponen formando un círculo sin repeticiones. Es el caso típico de personas sentadas a la mesa. Una vez que se ubica el primer elemento, queda determinado el inicio y el fin del círculo. Ahora bien, cuando se tienen n elementos y se ubica el primero de ellos, al segundo le quedan disponibles n-1 lugares para su ubicación. Una vez que se haya dispuesto la ubicación del segundo elemento, el tercero tendrá n-2 opciones para su ubicación, y así sucesivamente.

Una vez que todos los elementos se han colocado en sus respectivas posiciones, podrá observarse que, si todos los elementos se desplazan de manera ordenada una posición, por ejemplo a su derecha, cada elemento continuará teniendo a su izquierda y a su derecha exactamente el mismo elemento que tenía en la posición anterior. Esta rotación podría realizarse n veces sin alterar el resultado final siempre que se conserve el orden. Lo anterior indica que estas no son nuevas permutaciones, se trata de la misma permutación, pero girada una posición con respecto a la anterior.

El razonamiento precedente, matemáticamente, podría ser expresado de la siguiente forma:

$$PC_n = \frac{n!}{n} \qquad 3.17$$

de donde, simplificando, puesto que

$$n! = (n-1)!\ n \qquad 3.18$$

se obtendría la fórmula general para el cálculo de posibilidades de las permutaciones circulares:

$$\boldsymbol{PC_n = P_{n-1} = (n-1)!} \qquad 3.19$$

Supóngase que un padre desea sentarse a la mesa con sus dos hijos, Figura 66. Para calcular elnúmero de permutaciones posibles, bastará con aplicar la fórmula para las permutaciones circulares con n = 3 elementos (el padre y los dos hijos):

$$\boldsymbol{PC_3 = P_{3-1} = (3-1)! = 2} \qquad 3.20$$

Está claro que las permutaciones que pueden presentarse son solo dos: la primera que el padre a la mesa, en cualquier posición y a que a la derecha tenga a su hija y a la izquierda a su hijo; la segunda posibilidad existente en esta permutación circular es que el padre se ubique en cualquier posición y que a su derecha quede sentado el hijo y su hija se siente a su izquierda. Obviamente, a medida que se tienen más elementos, el número de posibilidades se irá incrementando de acuerdo al factorial correspondiente.

Figura 66
Disposición inicial de 3 elementos en permutaciones circulares

Nota: Niños japoneses alrededor de una mesa [Imagen]. Publicdomainvectors.org, 2 de julio de 2018, Publicdomainvectors.org (https://publicdomainvectors.org/es/vectoriales-gratuitas/Ni%C3%B1os-japoneses-alrededor-de-la-mesa/75318.html). CC0 1.0.

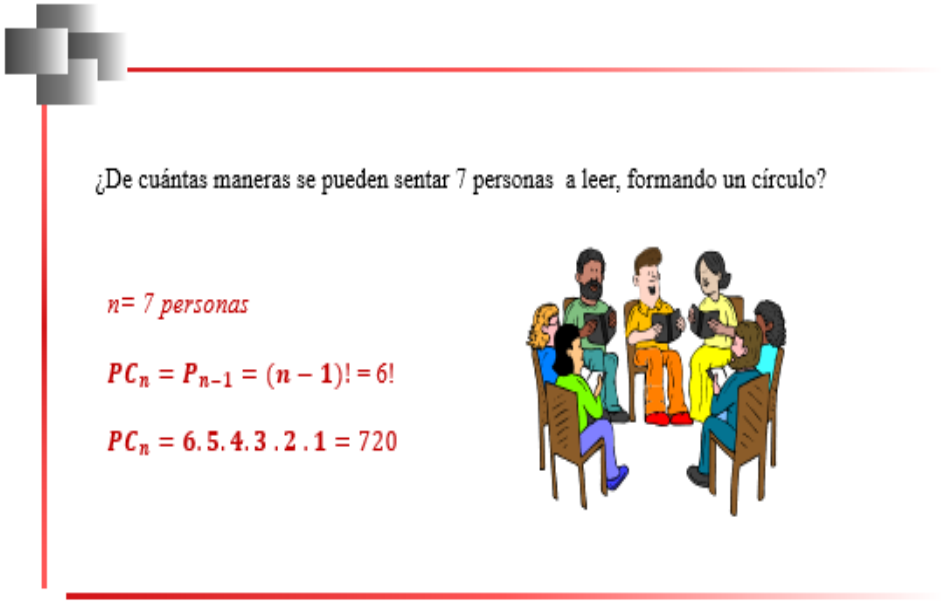
En la Figura 67 se ilustra otro ejemplo de permutaciones circulares y en la Figura 68 se presenta un resumen de los diferentes tipos de permutación, así como de las fórmulas que deben ser aplicadas en cada caso.

Combinaciones

Es una técnica de conteo que permite determinar el número de grupos que se pueden formar a partir de un conjunto dado de elementos (Feller, 1971). En las combinaciones, a diferencia de las permutaciones,el orden en el que se colocan los elementos no importa.

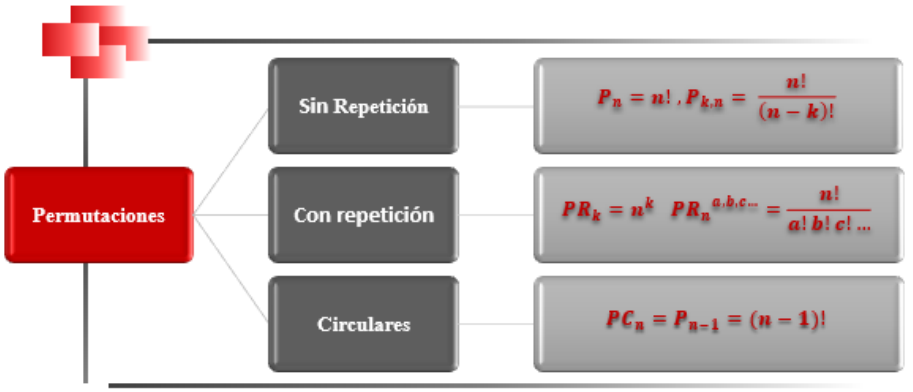
Tipos de Combinaciones. Hay dos tipos de combinaciones: 1) Combinaciones sin repetición y 2) Combinaciones con repetición.

Figura 67
Ejemplo de permutaciones circulares



Nota. Imagen incluida: Group of people sitting and reading books [Imagen], (https://publicdomainvectors.org/en/free-clipart/Vector-drawing-of-people-sitting-and-reading-a-book/18842.html, 11 de diciembre de 2013, Publicdomainvectors.org, t.ly/fTqr). CC0 1.0

Figura 68
Tipos de permutación



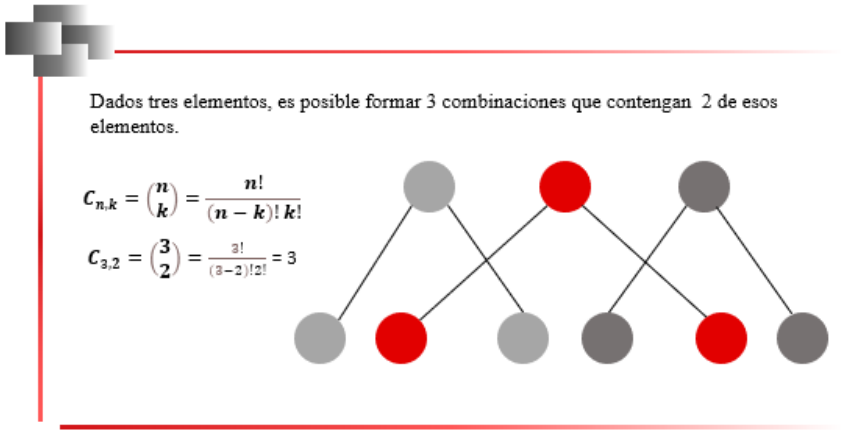
Combinaciones sin Repetición. También conocidas como combinaciones ordinarias. Presentan tres características fundamentales: 1) no importa el orden de selección de los elementos considerados en la combinación; 2) No entran todos los elementos en la selección; 3) Los elementos no se repiten.

La fórmula que permite calcular las combinaciones sin repetición es:

$$C_{n,k} = \binom{n}{k} = \frac{n!}{(n-k)!k!} \qquad 3.21$$

donde n es el número total de elementos; y k el tamaño de los grupos a ser seleccionados pararealizar las combinaciones, Figura 69.

Figura 69
Combinaciones de tres elementos, tomados de dos en dos



$C_{n,k}$ recibe el nombre de número combinatorio y se lee: combinaciones de n elementos tomados de k en k (Castillo Manrique & Gijarro Garvi, 2006, p.398), o "de n se elige k" (Devore, 2012, p.67).

Un ejemplo de una combinación ordinaria sería escoger de un grupo de 5 personas, tres representantes para presentar un proyecto. En este caso, se tendrían combinaciones de 3 elementos a ser escogidos de un total de 5:

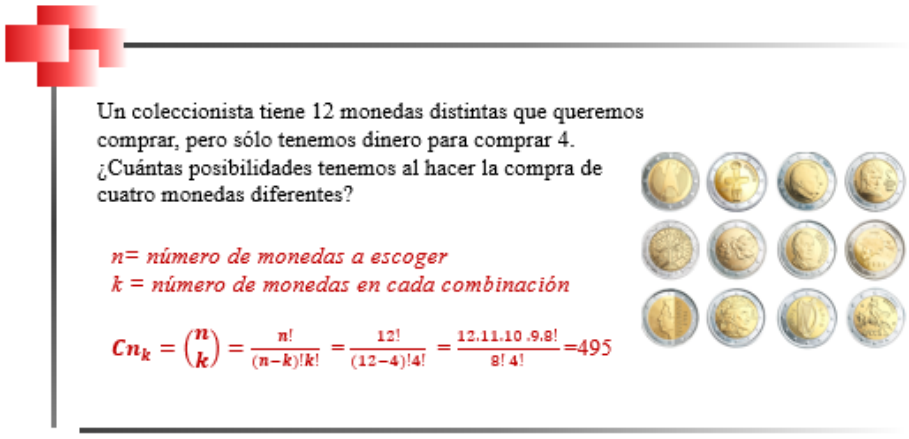
$$C_{5,3} = \binom{5}{3} = \frac{5!}{(5-3)!3!} \qquad 3.22$$

$$C_{5,3} = \frac{5.4.3!}{2!3!} = 10 \qquad 3.23$$

En la Figura 70, se ilustra otro ejemplo del cálculo de combinaciones ordinarias.

Figura 70

Ejemplo de combinaciones sin repetición



Nota. Imagen incluida: Monedas de metal [Imagen], Pixnio, s.f., Pixnio (https://pixnio.com/es/objetos/billetes/monedas-metal/monedas-de-metal-oro-negocio-mo neda-coleccion-ganancias-ingresos). CC0 1.0

Combinaciones con Repetición. En las combinaciones con repetición se utiliza la fórmula:

$$CR_n^k = \frac{(n+k-1)}{(n-1)!k!} \qquad 3.24$$

se lee “combinaciones con repetición de n elementos, tomados de k en k” (Wilhelmi,2004, p.56).

Considérese el siguiente ejemplo: se dispone de un mazo de naipes españoles que posee 40 cartas, ¿cuántos grupos de 5 barajas se pueden formar suponiendo que cada vez que se extrae una carta ésta se vuelve a reponer?

En este caso se tiene una combinación con repetición. Se trata de una combinación porque no importa el orden en el que se extraigan las cartas.

Por ejemplo, si en un grupo se han sacado un 7 de oros, un 2 de bastos, un 5 de oros, un rey de espadas y un 6 de copas, da lo mismo el orden en que hayan sido extraídos del mazo.

La combinación es con repetición, porque una vez que se saca una carta, ésta es devuelta al mazo pudiendo ser extraída nuevamente para formar parte del grupo.

Para calcular el número de grupos posibles, se hará uso de la fórmula para combinaciones con repetición, con n = 40 (número total de barajas) y k = 5 (número de cartas en cada grupo).

$$\boldsymbol{CR_{40}^{5} = \frac{(40+5-1)}{(40-1)!5!} = \frac{44!}{39!5!}} \qquad 3.25$$

$$\boldsymbol{CR_{40}^{5} = \frac{44.43.42.41.40.39!}{39!\,5.4.3.2.1}} \qquad 3.26$$

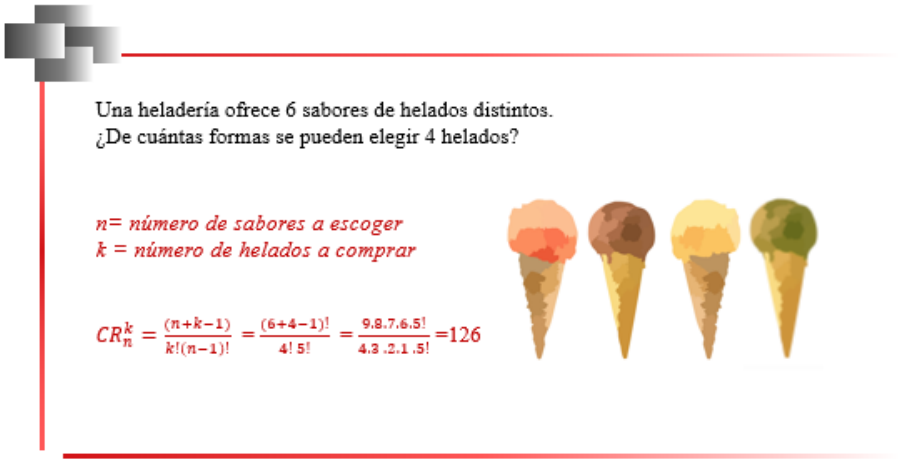
$$\boldsymbol{CR_{40}^{5} = \frac{44.43.42.41.40}{5.4.3.2.1} = 1.086.008} \qquad 3.27$$

Recuérdese que para simplificar el cálculo de los factoriales, es conveniente que el factorial del numerador se cierre, justamente, en el mayor factorial que exista en el denominador.

Nótese que, en el ejemplo anterior el factorial del numerador (44!) se abre hasta el mayor factorial que existe en el denominador (39), de esa manera hay dos factoriales, uno en el numerador y otro en el denominador, que son idénticos, pudiendo ser simplificados para que la división quede sin factoriales.

En la Figura 71 se presenta otro ejemplo de combinaciones con repetición.

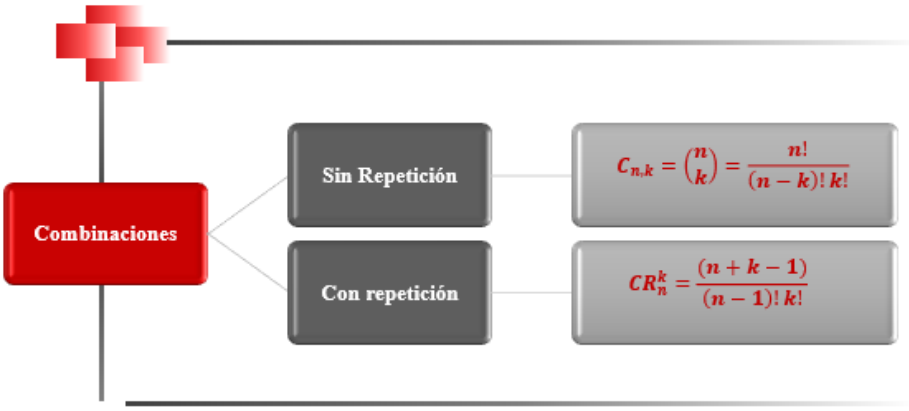
Figura 71
Ejemplo de combinaciones con repetición



Nota. Imagen incluida adaptada de Conos de helado postre vainilla chocolate fresa [Imagen], Pixabay, 3 de agosto de 2014. Pixabay t.ly/ApUv). Pixabay license.

En la Figura 72, se presenta un resumen de los tipos de combinación y las fórmulas a ser usadas.

Figura 72
Tipos de combinación



Probabilidad de Eventos: Sucesos, Eventos, Espacio Muestral

La probabilidad es una herramienta de la estadística inferencial que hace posible la obtención de conclusiones sobre una determinada población con base en el estudio de una muestra (Mendenhall et al., 2010, p128). Consiste en cuantificar la posibilidad de que un suceso tenga éxito cuando en su aparición interviene el azar.

Experimento

Constituye un suceso cuyo resultado carece de certeza (Molina & Rodrigo, 2010), un proceso a través del cual se logra obtener, por observación o medición, el resultado de un fenómeno.

En este caso, a priori, es decir, antes de comenzar el experimento, ya se conocen todos los posibles resultados. Por ejemplo, antes de lanzar un dado se sabe que solamente existen 6 posibilidades: que en la cara superior se tenga un 1, un 2, un 3, un 4, un 5 o un 6, no hay más. Sin embargo, el resultado exacto no es predecible, es decir, no se puede asegurar, con total certeza, que en la primera vez que se lance el dado va a salir, por ejemplo, un 3. Es imposible que se pueda predecir cuál va a ser el resultado exacto, siempre y cuando, obviamente, se trate de un dado normal, esto es, que no haya sido alterado para forzar algún resultado en su lanzamiento.

Evento simple

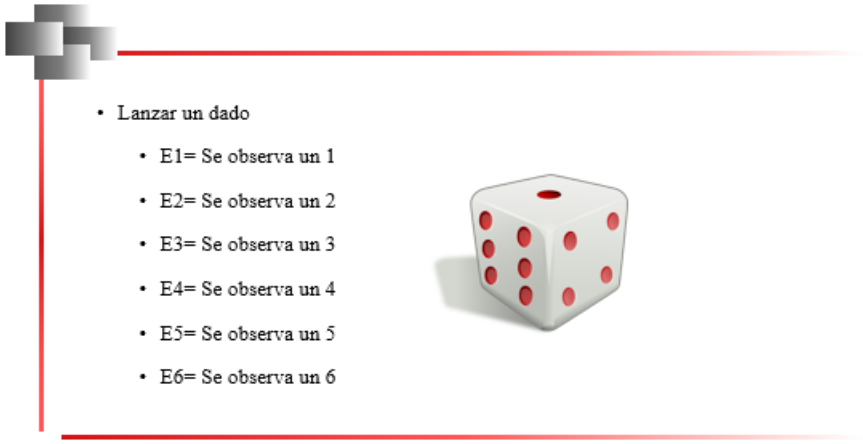
Un evento simple es el resultado que se puede observar cuando se realiza una sola vez el experimento (DeGroot & Schervish, 2011). Por ejemplo, se lanza una moneda, una sola vez, y se observa si sale cara o sale cruz.El evento simple es que salga una cara. El evento simple también es conocido como evento o suceso elemental, es un evento con un solo resultado; es decir, corresponde al resultado de un solo experimento.

En el ejemplo anterior, un evento simple en el lanzamiento de un dado sería obtener un 3 en lacara superior. En ese caso, se escribe: E = { 3 }.

En la Figura 73, se ilustran algunos ejemplos de eventos simples obtenidos al lanzar un dado y observar la cara superior.

Figura 73

Eventos simples al lanzar un dado y observar la cara superior del cubo



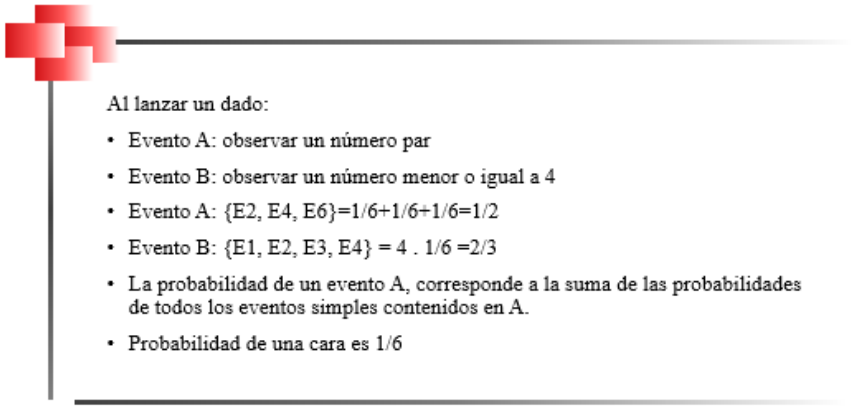
Nota. Imagen incluida adaptada de Dado [Imagen], Pixabay, 15 de octubre de 2013. Pixabay (https://pixabay.com/es/vectors/dado-cubo-morir-juego-suerte-152179/). Pixabay license.

Evento

Es un conjunto formado por uno o más resultados de un experimento definido (DeGroot & Schervish, 2011). Los eventos suelen designarse con letras mayúsculas. Los eventos están formados por eventos simples. En la Figura 74 se ilustran algunos ejemplos de eventos para el experimento de lanzar un dado y observar la cara superior.

Figura 74

Ejemplos de eventos al lanzar un dado, según el resultado de la cara superior



Espacio Muestral (Ω)

Se denomina espacio muestral al conjunto de resultados posibles que puede tener unexperimento (Spiegel & Stephens, 2009), Figura 75.

Si se tiene un experimento que consiste en lanzar tres veces una moneda al aire, el espaciomuestral sería:

$$\Omega = \{(\text{cara, cara, cara}), (\text{cara, cara, sello}), (\text{cara, sello, cara}), (\text{sello, cara, cara}), (\text{sello, sello, cara}), (\text{sello, cara, sello}), (\text{cara, sello, sello}), (\text{sello, sello, sello})\} \quad 3.28$$

Un evento, podría ser "obtener un sello en el primer lanzamiento, con lo cual se tendría:

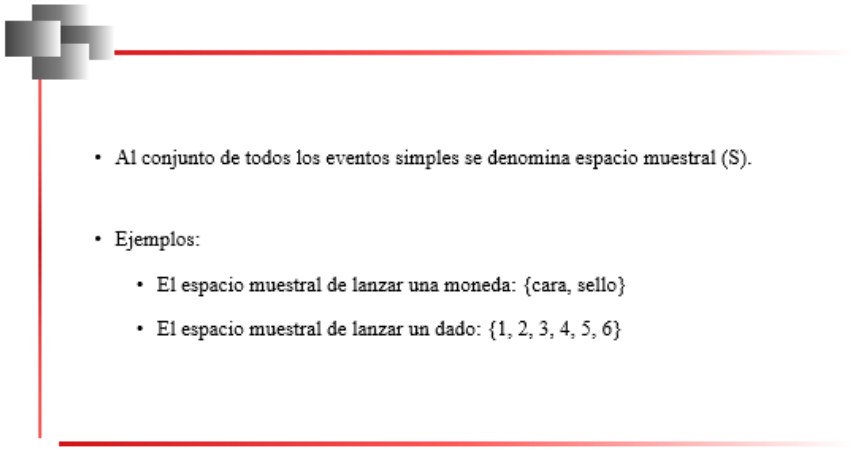
$$A = \{(\text{sello, cara, cara}), (\text{sello, sello, cara}), (\text{sello, cara, sello}), (\text{sello, sello, sello})\} \quad 3.29$$

Obsérvese que el evento A es un subconjunto del espacio muestral Ω:

$$E = \{(\text{sello, cara, cara})\} \quad 3.30$$

Figura 75

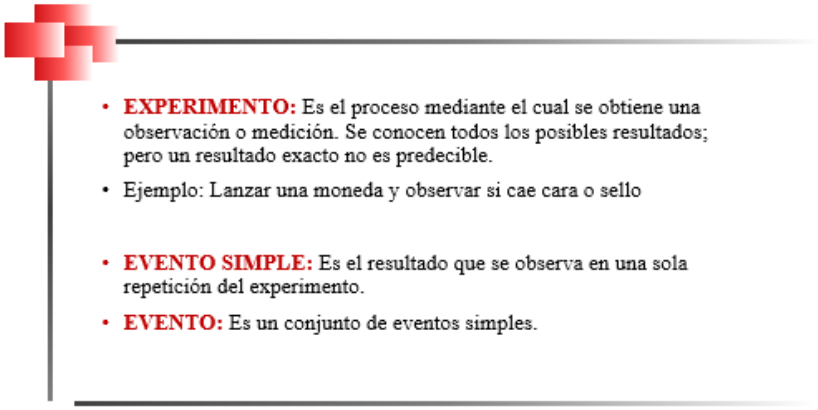
Concepto de espacio muestral y ejemplos



En la Figura 76, se presenta, a manera de resumen, los conceptos de experimento, evento simple y evento.

Figura 76

Conceptos de experimento, evento simple y evento

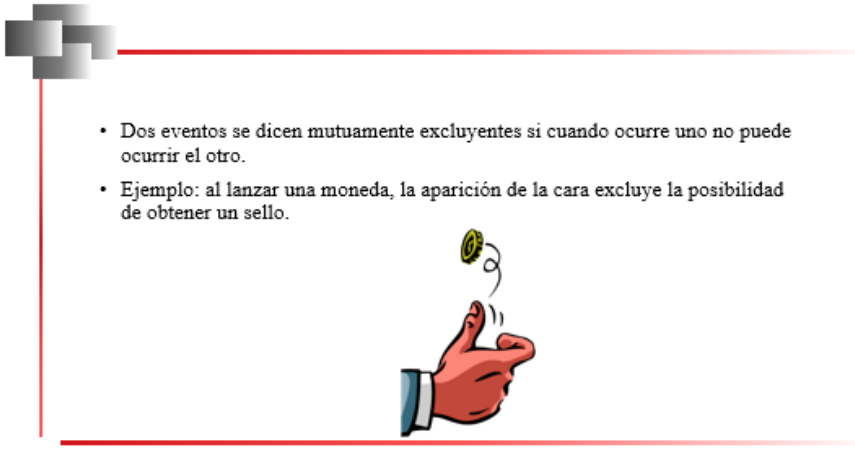


Eventos Mutuamente Excluyentes

Dos eventos se dicen mutuamente excluyentes cuando la ocurrencia del uno impide el suceso del otro (Freedman et al., 2007) . En el caso de la moneda, los eventos $E_1 = cara$ y $E_2 = sello$, son mutuamente excluyentes, porque al lanzar una moneda y obtener un resultado cara, ya no es posible la obtención de un resultado sello, Figura 77.

Figura 77

Eventos mutuamente excluyentes



Nota. Imagen incluida adaptada de Vector illustration of decision making hand flipping [Imagen], Pngkey, s.f., Pngkey (https://www.pngkey.com/detail/u2e6y3a9q8w7u2u2_vector-illustration-of-decision-making-hand-flipping-coin/). License personal use.

Los eventos mutuamente excluyentes se caracterizan porque es imposible que ocurran al mismo tiempo. Otro ejemplo sería un foco encendido o pagado, ya que la existencia de un estado, en un momento particular, descarta al otro.

En el caso de lanzamiento de un dado, cuando sale un determinado número en la cara superior, se excluye toda posibilidad de que se obtenga, se forma simultánea otro número en dicha cara.

Los eventos excluyentes también reciben el nombre de eventos disjuntos.

Cálculo de Probabilidades

Si en un espacio muestral, Ω, se tienen **N** resultados con la misma posibilidad de ocurrencia, y.además se dispone de un evento A, que está formado por **p** elementos de Ω, entonces, la probabilidad de ocurrencia de A, (Bayes, 1763), está dada por:

$$P(A) = \frac{p}{N} = \frac{número\ de\ éxitos}{número\ de\ resultados\ posibles\ en\ \Omega} \quad 3.31$$

Lo anterior indica que, si se desea calcular la probabilidad de hallar una cara en el lanzamiento de una moneda, se tendría un espacio muestral dado por:

$$\Omega = \{cara, sello\} \quad 3.32$$

Entonces, N = 2 y como sólo puede existir un éxito, que correspondería al caso favorable de que salga una cara, se tendría que la probabilidad de obtener una cara en el lanzamiento de una moneda, sería:

$$P(A) = \frac{1}{2} \quad 3.33$$

Si, por ejemplo, se desea hallar la probabilidad de obtener un número par, en la cara superior,al lanzar un dado, se tendría:

$$\Omega = \{1, 2, 3, 4, 5, 6\} \quad 3.34$$

$$p = 3 \text{ (obtener un 2, un 4 o un 6)} \quad 3.35$$

de donde

$$P(A) = \frac{3}{6} = \frac{1}{2} \quad 3.36$$

De acuerdo al valor que pueda tomar p, el número de éxitos o casos favorables al estudio, sepueden presentar tres casos que se presentan en la Tabla 14.

Tabla 14

Relación entre los casos favorables y los resultados posibles en el cálculo de la probabilidad

Casos favorables	Probabilidad	Ocurrencia de evento
p = N	P(A) = 100%	Siempre
$0 < p < N$	$0 < P(A) < 100\%$	Incierta
p = 0	P(A) = 0	Nunca

Operaciones con Eventos Aleatorios

Hay ciertas operaciones que se pueden realizar entre eventos y que, a su vez, traen como consecuencia que se formen nuevos eventos (Walpole, *et al.*, 2012). Estas operaciones son:

1. Unión
2. Intersección
3. Complemento
4. Diferencia.

Unión entre Eventos (A ∪ B)

La unión de un evento A con un evento B, da origen a la formación de un nuevo evento, A U B, cuyos elementos serán los elementos de A más los elementos de B. En el caso de existencia de eventos repetidos, éstos se incluyen una sola vez.

Sean

A ={ números pares entre 1 y 10} ={2, 4, 6, 8, 10} 3.37

y

B ={números primos entre 1 y 10}={1, 3, 5, 7} 3.38

entonces,

$$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8, 10\} \quad 3.39$$

Intersección entre Eventos ($A \cap B$)

El evento intersección entre A y B, $A \cap B$, es un evento que contiene los elementos comunes entre A y B.

Sean

$$A = \{\text{números impares entre 1 y 10}\} = \{1, 3, 5, 7, 9\} \quad 3.40$$

y

$$B = \{\text{números primos entre 1 y 10}\} = \{2, 3, 5, 7\} \quad 3.41$$

entonces, $A \cap B$ contiene a todos los elementos que, al mismo tiempo, se encuentran tanto en el evento A como en el evento B.

$$A \cap B = \{3, 5, 7\} \quad 3.42$$

Complemento ($\bar{A}$)

Sea un espacio muestral, Ω, el complemento de un evento A, $\bar{A}$, estará formado por aquellos elementos que están presentes en el espacio muestral, pero no en A. Dicho en otras palabras, $\bar{A}$, está formado por los elementos que le faltan a A, para ser igual a Ω.

Sean

$$\Omega = \{\text{“los meses del año}\} \quad 3.43$$

y

$$A = \{\text{“meses del año con más de 30 días”}\} \quad 3.44$$

entonces, el complemento del evento A, $\bar{A}$, estará formado por todos los meses del año que tienen un número de días diferente de 31, esto es:

$$\bar{A} = \{\text{febrero, abril, junio, septiembre, noviembre}\} \quad 3.45$$

Diferencia(A - B)

El evento diferencia entre A y B, está compuesto por todos los elementos que están en A, pero no están en B. También puede ser escrito como .

Sean

A = {"números del 1 al 8"}={1, 2, 3, 4, 5, 6, 7, 8} 3.46

y

B = {"números menores que 10 que son divisibles por 2"}={2, 4, 6, 8} 3.47

entonces,

$$A - B = \{1, 3, 5, 7\} \quad 3.48$$

En la Figura 78 se presenta un resumen de las operaciones con eventos: unión, intersección. complemento y diferencia; y en la Figura 79, se incluye la representación gráfica de las operaciones.

Reglas de Probabilidad: Adición y Multiplicación

Las reglas de probabilidad son principios matemáticos que se utilizan para calcular la probabilidad de que ocurra un evento. Se aplican a conjuntos de eventos o experimentos aleatorios, y se utilizan para determinar la probabilidad de que un evento en particular ocurra dentro de ese conjunto.

Son utilizadas en operaciones de adición y multiplicación. La probabilidad de la suma se denota como $\boldsymbol{P(A \text{ o } B)}$

Regla de Probabilidad de la Suma o Adición de Eventos, $\boldsymbol{P(A \text{ o } B)}$

Al momento de realizar la adición de probabilidades, es necesario tomar en cuanta si los eventos son o no mutuamente excluyentes.

Figura 78
Operaciones con eventos

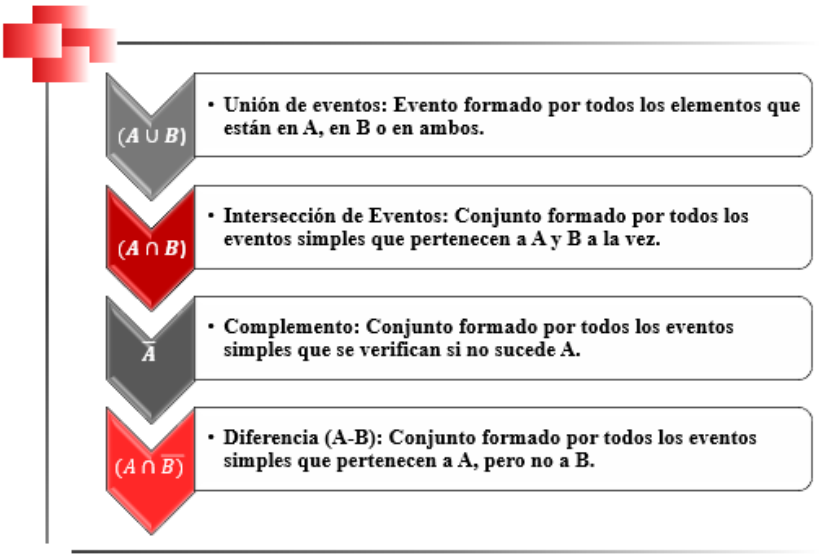
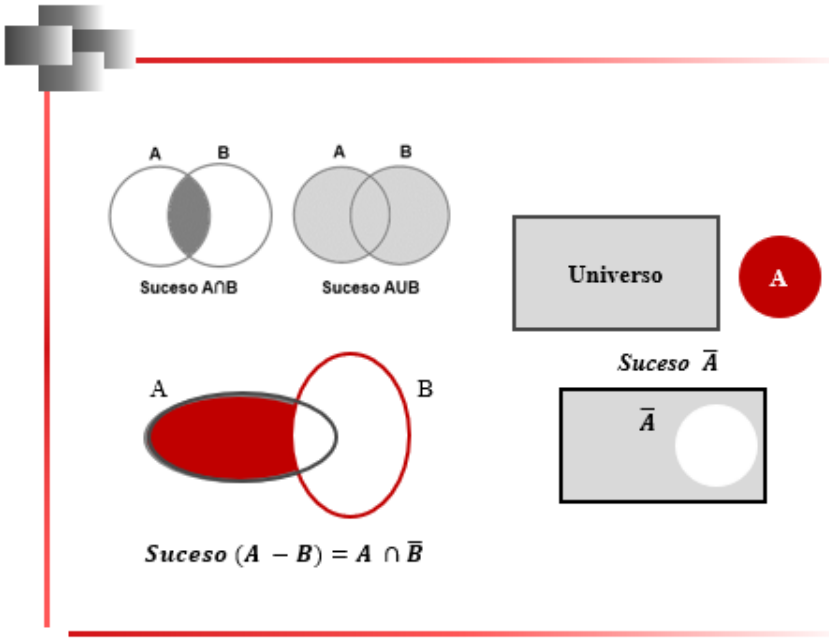


Figura 79
Representación gráfica de las operaciones entre eventos



Regla de Probabilidad de la Suma de Eventos Mutuamente Excluyentes

Cuando los eventos son mutuamente excluyentes, la probabilidad de la suma de eventos es igualala suma de las probabilidades de ambos eventos. Esto es:

$$P(A \circ B) = P(A) + P(B) \qquad 3.49$$

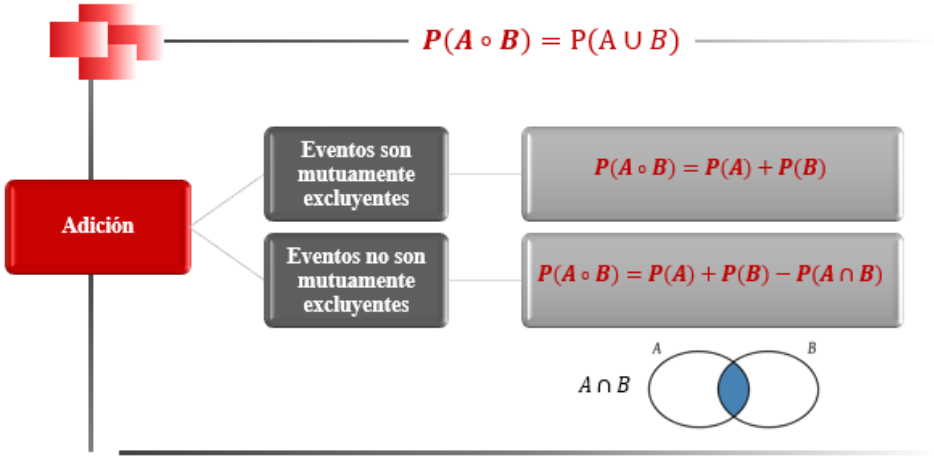
Regla de Probabilidad de la Suma de Eventos que no son Mutuamente Excluyentes

Cuando los eventos no son mutuamente excluyentes, esto es, cuando la ocurrencia de un evento,A, no anula la posibilidad de ocurrencia de un segundo evento, B, entonces, la probabilidad de lasuma de eventos es igual a la diferencia entre la suma de las probabilidades de ambos eventos y laprobabilidad de la intersección entre ellos (Jeffreys, 1939):

$$P(A\ o\ B) = P(A) + P(B) - P(A \cap B\ ) \qquad 3.49$$

En la Figura 80 se presenta un resumen de las fórmulas que deben ser empleadas para hallar la probabilidad de la suma de eventos.

Figura 80
Probabilidad de la suma de eventos



En la Figura 81 se incluye un ejemplo de cálculo de la probabilidad de la suma para eventos excluyentes. Obsérvese que, en el ejemplo dado, se trata de la probabilidad de la suma de tres eventos, por lo que vale aclarar que las fórmulas planteadas no sólo pueden ser aplicadas cuando se desea encontrar la probabilidad de la suma de dos eventos, sino que son válidas para hallar la probabilidad de la suma de cualquier cantidad de eventos.

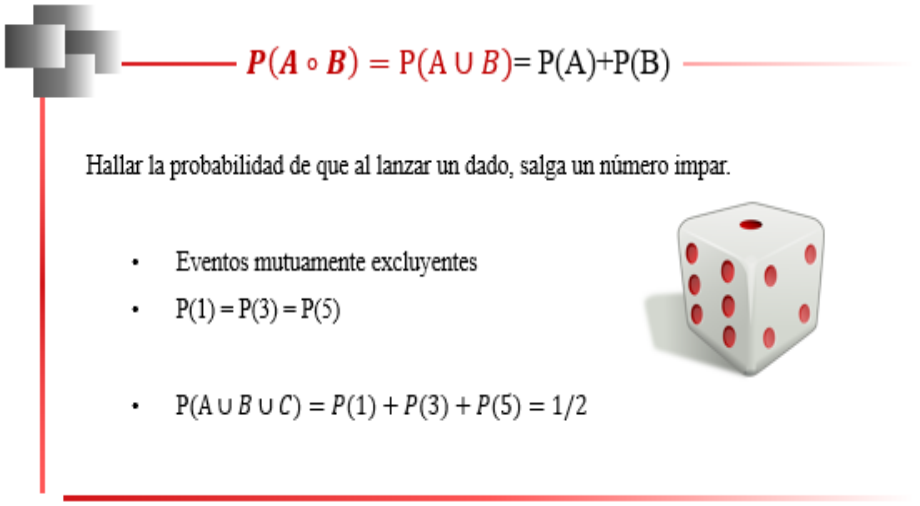
Lo anterior se puede comprender fácilmente si se piensa que la probabilidad de la suma de más de dos eventos no es más que la aplicación sucesiva de la fórmula para el cálculo de la probabilidad cuando se tienen dos eventos, esto es:

$$P(A \cup B \cup C) = P(A \cup B) + P(C) \qquad 3.50$$

lo anterior implica que la suma de probabilidades cumple la propiedad asociativa.

Figura 81

Ejemplo de cálculo de probabilidad de la suma de eventos mutuamente excluyentes



Nota. Imagen incluida adaptada de Dado [Imagen], Pixabay, 15 de octubre de 2013. Pixabay (https://pixabay.com/es/vectors/dado-cubo-morir-juego-suerte-152179/). Pixabay license.

A continuación, se plantea un ejemplo para el cálculo de la probabilidad de la suma de eventos:

los postres favoritos de Ángel David son fresas con crema y duraznos en almíbar. La probabilidad diaria de que coma fresas con crema es de 0.4 y la probabilidad de que coma duraznosen almíbar es de 0.3; sin embargo, la probabilidad de que coma los dos postres juntos, en el mismo día, es de 0.1. ¿Cuál es la probabilidad de que, en un día cualquiera, Ángel David coma uno de los dos postres?

$$P(A \circ B) = P(A) + P(B) - P(A \cap B) \qquad 3.51$$

En este caso, los eventos no son excluyentes, pues, según el enunciado, Ángel David, puede llegar a comer los dos postres en un mismo día.

Se definen los eventos:

$$F=\{\text{come fresas con crema}\} \qquad 3.52$$

$$D=\{\text{come durazno en almíbar}\} \qquad 3.53$$

Con P(F) = 0.4, P(D) = 0.3 y $P(F \cap D) = 0.1$.

Como se trata de eventos no excluyentes:

$$P(F \text{ o } D) = P(F \cup D) = P(F) + P(D) - P(F \cap D) \qquad 3.54$$

$$P(F \cup D) = 0.4 + 0.3 - 0.1 = 0.6 \qquad 3.55$$

De donde se concluye que la probabilidad de que Ángel David coma fresas con crema o duraznos en almíbar, es de 60%.

Regla de Probabilidad del Producto o Multiplicación de Eventos, $\boldsymbol{P(A \text{ o } B)}$. Para encontrar la probabilidad del producto de eventos es necesario considerar si dichos eventos son o no independientes (Boole, 1854).

Eventos Independientes y Eventos Dependientes. Dos eventos se dicen independientes cuando la ocurrencia del uno no tiene ninguna influencia sobre la probabilidad del otro evento (Bacchini *et al*., 2018, p.19) y viceversa (Mendenhall *et al*.,2010, p.149) Cuando, por el contrario, la ocurrencia de un evento altera la probabilidad de ocurrencia del otro, se dice que los eventos son dependientes.

Como ejemplo se puede plantear la extracción de una carta de un mazo de 40 naipes. Si la carta una vez seleccionada se devuelve al mazo, los eventos son independientes, ya que ambas cartas tienen la probabilidad de 1/40; en cambio, si la carta no se repone, la probabilidad de la primera carta es 1/40, pero la probabilidad de la segunda carta es de 1/39 (ya no están los 40 naipes porque se extrajo uno). En este segundo caso, los eventos son dependientes.

Regla de Probabilidad del Producto de Eventos Independientes. Cuando dos eventos son independientes, la probabilidad del producto entre ellos, P(A∩ B), puede ser hallada a través de la siguiente fórmula (Laplace, 1812):

$$\boldsymbol{P(A\ y\ B) = P(A.B) = P(A)P(B)} \qquad 3.56$$

Considérese el siguiente ejemplo: la probabilidad de que un estudiante de la Universidad Xhable inglés es de 0.2; mientras que la probabilidad de que haya aprobado Realidad Nacional es de 0.8. ¿cuál es la probabilidad de que un estudiante hable inglés y, además, haya aprobado Realidad Nacional?

$$\boldsymbol{P(A\ y\ B)} = \mathrm{P(A} \cap B) \qquad 3.57$$

En este caso, los eventos son independientes, ya que el hecho de que un estudiante tenga un segundo idioma no incide de manera alguna en que ese mismo estudiante apruebe otra asignatura que no está relacionada con el inglés. Entonces, se tiene:

$$\mathrm{P(I)} = \text{Probabilidad de que hable inglés} = 0.2 \qquad 3.58$$

$$\mathrm{P(RN)} = \text{Probabilidad de que haya aprobado RN} = 0.8 \qquad 3.59$$

de donde:

$$\mathrm{P(A} \cap B) = \mathrm{P(I).\ P(RN)} = 0.2 * 0.8 = 0.16 \qquad 3.60$$

Lo anterior significa que la probabilidad de que un estudiante hable inglés y, al mismo tiempo haya aprobado la asignatura Realidad Nacional, es del 16%.

Regla de Probabilidad del Producto de Eventos Dependientes. Cuando dos eventos son dependientes, la probabilidad del producto de ambos even-

tos se puede calcular utilizando una de las siguientes fórmulas (Lipschutz, 1965):

$$\boldsymbol{P(A\ y\ B) = P(A.B) = P(A)P(B/A)} \qquad 3.61$$

en donde $\boldsymbol{P(B/A)}$ es la probabilidad de que ocurra un evento B, considerando que ya ocurrió unevento A. En este caso, es requisito indispensable que la probabilidad del evento A sea diferente de cero, $P(A) \neq 0$.

$$\boldsymbol{P(A\ y\ B) = P(A.B) = P(B)P(A/B)} \qquad 3.62$$

en donde $\boldsymbol{P(B/A)}$ es la probabilidad de que ocurra un evento A, considerando que ya ocurrió unevento B. En este caso, es requisito indispensable que la probabilidad del evento B sea diferente de cero, $P(B) \neq 0$.

La probabilidad del producto de eventos, no es más que la probabilidad de la intersección de dichos eventos.

$$P(A.B) = P(A \cap B) \qquad 3.63$$

En la Figura 82 se presenta un resumen de las reglas de probabilidad para el producto de eventos. Es de hacer notar que, cuando se tienen más de dos eventos, la probabilidad debe ser visualizada como el producto consecutivo de eventos.

En la Figura 83 se incluye un ejemplo que permite determinar analíticamente si dos eventos son independientes o dependientes.

Figura 82
Regla de probabilidad del producto de eventos

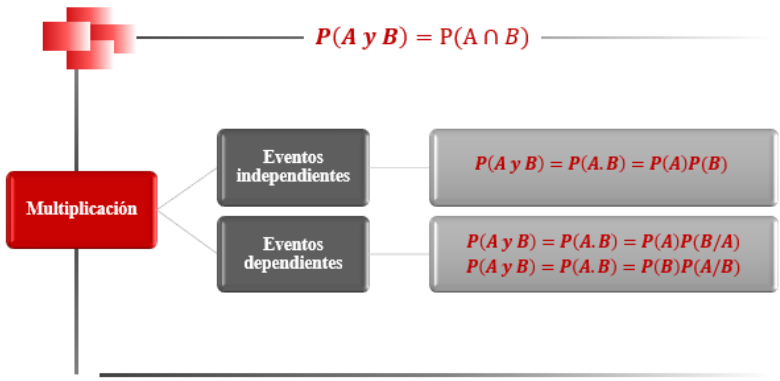
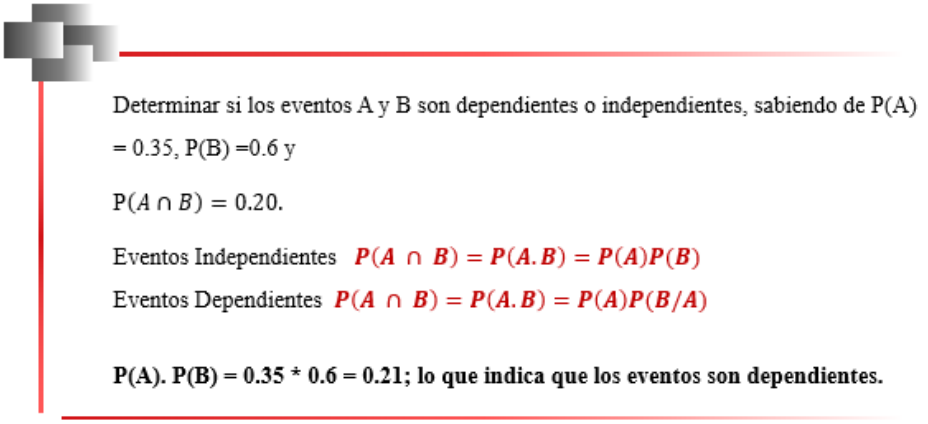


Figura 83

Ejemplo que permite determinar si dos eventos son independientes o no



Probabilidad Condicional y Teorema de Bayes

La probabilidad condicional es una medida de la probabilidad de que ocurra un evento, dado que ya ha ocurrido otro. En otras palabras, se trata de calcular la probabilidad de un evento A, suponiendo que ya ha ocurrido otro evento B.

Probabilidad Condicional

Para todo par de eventos A y B, que pertenecen a un mismo espacio muestral, la probabilidad condicional se define como:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \qquad 3.64$$

Corresponde a la probabilidad de que ocurra un evento A, considerando que ya ocurrió un evento B., siempre que P(B)> 0, (Kolmogorov, 1950).

Considérese el siguiente ejemplo: Lucy tiene en una caja 10 caramelos 3 de fresa 5 de limón y 2 de piña. Lucy les pide a Alejandra y a Fabiana que cada un saque un caramelo de la caja sin mirar dentro de ella y que no vuelvan a depositarlo adentro. Si el caramelo que extrajo Alejandra es de limón, ¿cuál es la probabilidad de que Fabiana saque otro caramelo de limón?

Ya que al inicio hay 10 caramelos dentro de la caja y 5 de ellos son de limón, la probabilidadde que Alejandra extraiga un caramelo de limón es de 5/10 = ½. Entonces, podemos definir ese primer evento A, como:

A: que Alejandra saque un caramelo de limón 3.65

Una vez que ha sucedido el evento A, se puede definir un segundo evento B:

B: que Fabiana saque un segundo caramelo y también sea de limón 3.66

La probabilidad de sacar un segundo caramelo de limón, habiendo sacado un primer caramelo del mismo sabor, no es otra cosa que una probabilidad condicional: P(B/A).

Ahora bien, de una manera intuitiva se puede razonar que, una vez que Alejandra haya sacado el primer caramelo, el espacio muestral ha variado porque ya sólo quedan 9 caramelos disponibles; y puesto que ese caramelo era de limón, el número de caramelos de este sabor ya no es 5, sino 4. De lo anterior se sigue que:

$$P(B/A) = 4/9 \qquad 3.67$$

es decir, los 4 caramelos de limón que quedan una vez que se extrajo el primero, entre los 8 caramelos que quedan en total, incluyendo todos los sabores.

Aplicando la regla de probabilidad del producto de eventos dependientes, indicada en la Figura 82, se obtiene que:

$$P(A \cap B) = P(A).P(^{B}/_{A}) \qquad 3.68$$

de donde

$$P(A \cap B) = \ . \qquad 3.69$$

Lo anterior indica que la probabilidad de que Fabiana saque un segundo caramelo de limón, una vez que Alejandra extrajo un primer caramelo de igual sabor, es igual a 22,22%.

Teorema de Bayes

El teorema de Bayes es utilizado para calcular la probabilidad de un evento del cual se posee una información a priori, es decir una probabilidad que es asignada antes de que se recolecten los datos empíricos (Lind et al., 2012, p.167). En la Figura 84 se presenta el teorema de Bayes, el cual puede ser utilizado para calcular probabilidades condicionadas de eventos que se suceden en etapas.

Figura 84
Teorema de Bayes

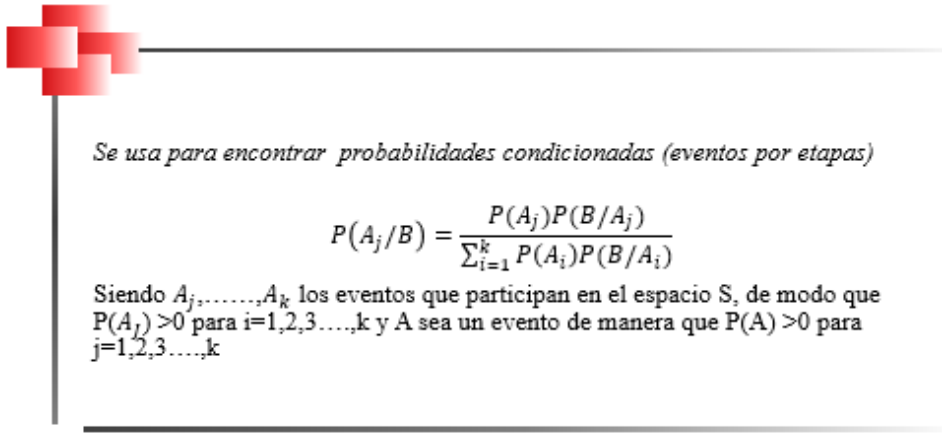


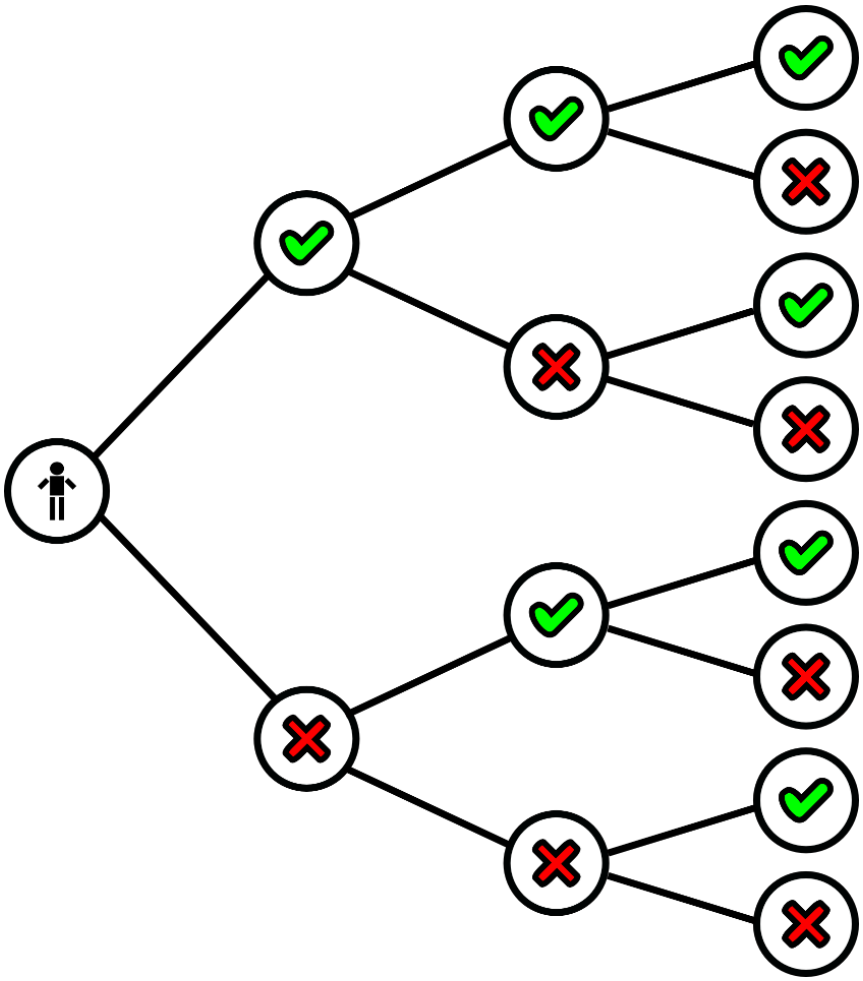
Diagrama de Árbol

Es una herramienta que facilita el cálculo de probabilidades, experimentos que se realizan en varias etapas, a través de la representación de un árbol en el que se esquematizan todas las posibilidades. Es, por tanto, también, una técnica de conteo (Feller, 1950; Devore, 2012, p.65).

La esquematización del diagrama de árbol se hace por etapas o generaciones, partiendo de un punto a la izquierda del cual van saliendo "las ramas" que son líneas rectas, trazadas de izquierdaa derecha, una por cada posibilidad de cada nodo, en cada etapa. Un nodo, es cualquier punto del que salen una o varias ramas y marca el inicio de una nueva generación.

La suma de las probabilidades de las ramas de cada nodo debe ser igual a 1. En la Figura 85 se presenta un árbol genérico de probabilidades. Obsérvese que, a partir del punto de inicio del árbol cada rama termina en un nodo donde se especifica una de las posibilidades correspondiente a la primera generación o primera etapa.

Figura 85
Árbol de probabilidades



Nota. Adaptada de Binario árbol [Imagen], Pixabay, 15 de abril de 2012. Pixabay(https://pixabay.com/es/vectors/binario-%C3%A1rbol-datos-estructura-34975/). Pixabaylicense.

Obsérvese en la Figura 85 que existen nodos que son paralelos, es decir que se encuentran sobre una misma vertical, éstos corresponden a cada una de las etapas en las que se desarrolla el experimento.

En las Figura 86 a 89 se presenta un ejemplo de aplicación del árbol de probabilidades.

Figura 86
Ejemplo de construcción del diagrama de árbol

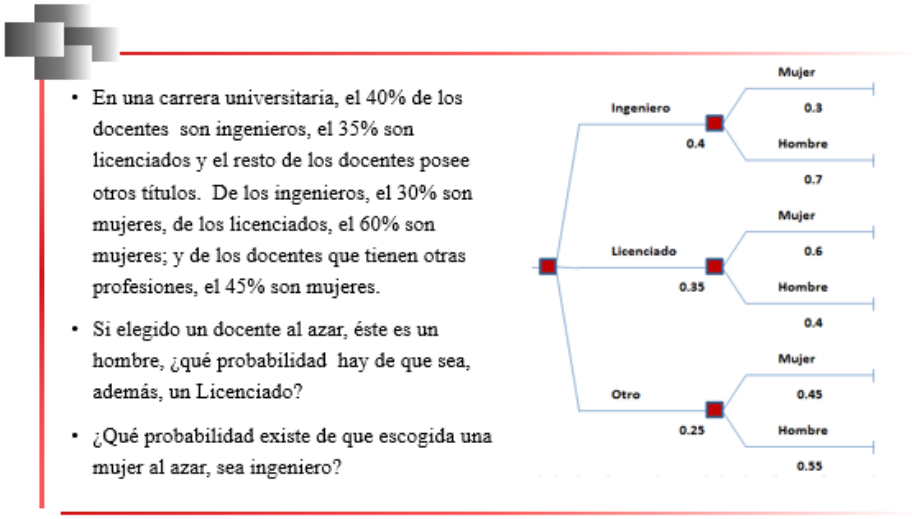


Figura 87
Aplicación del Teorema de Bayes al ejercicio planteado en la Figura 86

$$P(A_j/B) = \frac{P(A_j)P(B/A_j)}{\sum_{i=1}^{k} P(A_i)P(B/A_i)}$$

$$P(Lic./Hombre) = \frac{P(Lic.)P(\frac{Hombre}{Licenciado})}{P(Ingeniero).P\left(\frac{Hombre}{Ingeniero}\right) + P(Lic.).P\left(\frac{Hombre}{Licenciado}\right) + P(Otro).P\left(\frac{Hombre}{Otro}\right)}$$

La probabilidad de que el docente sea licenciado, se obtiene de la rama de primera generación: P(Licenciado) = 0,35; la probabilidad de que dado que es hombre, sea un licenciado, se obtiene de la rama de segunda generación, Hombre, cuya primera generación es Licenciado: P(Hombre/Licenciado)= 0.4.

Figura 88

Cálculo de la probabilidad pedida en el ejercicio planteado en la Figura 86

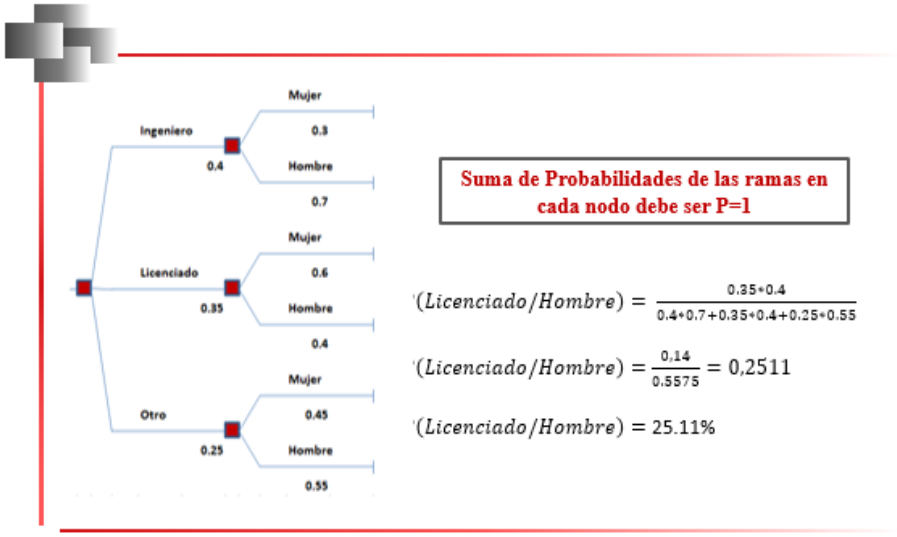
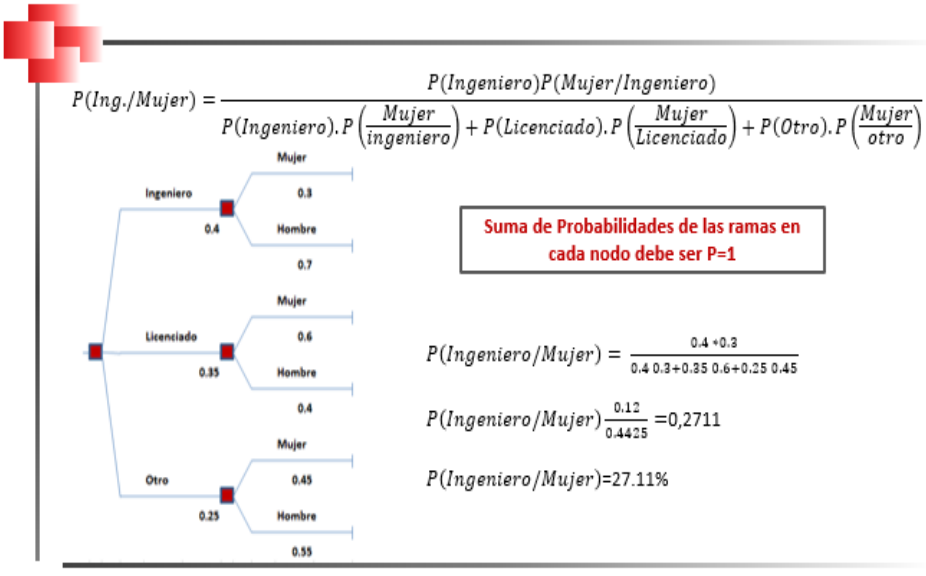


Figura 89

Cálculo de la probabilidad pedida en el ejercicio planteado en la Figura 86



Distribución binomial y Distribución Normal

Conceptos de Variable Aleatoria y Distribución de Probabilidad

Variable Aleatoria

Una variable es aleatoria si su valor proviene del resultado de un ensayo o experimento y en el que cada uno de los resultados obtenidos puede ser relacionado con un número especificando una regla de asociación. Matemáticamente, una variable aleatoria es una función cuyo dominio se corresponde con el espacio muestral y cuya imagen es el conjunto de los números reales (Devore,2012, p.93)

Como ejemplo se puede citar el tiempo que espera un cliente para pagar su compra en la fila de un supermercado (Santa María & Buccino, 2019, p.52).

Distribución de Probabilidad

Al conjunto de valores posibles que puede tener una variable aleatoria, la relación que existe entre dichos valores, y sus probabilidades correspondientes, se denomina distribución de probabilidad (Zylberberg, 2005).

La distribución de probabilidades de una variable está relacionada con su distribución de frecuencias y corresponde a una distribución de frecuencias teórica. Su forma da cuenta de la manera como se espera que se distribuyan los resultados.

Una distribución de probabilidad contiene todos los sucesos posibles en un experimento y la probabilidad asociada a cada uno de ellos.

Características de una función de probabilidad

1. La probabilidad de un evento siempre es positiva y su valor oscila entre 0 y 1.
2. Los resultados corresponden a eventos que son mutuamente excluyentes.
3. La lista de eventos considerados es exhaustiva y por ello la suma de todas lasprobabilidades debe ser igual a 1.

Distribución Binomial

Es un tipo de distribución de probabilidad de variable discreta (Laplace, 1814). Considera que la probabilidad es la misma en todos los experimentos realizados y que el resultado obtenido en un ensayo es independiente del anterior.

En una distribución binomial se tiene un experimento cuyo resultado es la ocurrencia o no de un determinado evento (Canavos, 1988, p.89). En la distribución binomial ***p,*** determina la probabilidad de que ocurra el evento en un solo experimento, y recibe el nombre de probabilidad de éxito; y ***q*** es la probabilidad de que el suceso no ocurra en un solo experimento (probabilidad de fracaso). Sólo hay dos resultados posibles en cada ensayo.

El cálculo de la probabilidad, asociada a un cierto valor de la variable aleatoria x, es:

$$P(x = k) = \frac{n!}{k!(n-k)!} p^k q^{n-k} \qquad 3.70$$

donde q = 1 – p; k = 0, 1 , 2, 3, … n (número de éxitos); y n es el número de ensayos.

Media de una Distribución Binomial

La media de una distribución binomial, (De Moivre, 1733), es denominada número esperado y se obtiene a través de la fórmula:

$$E(X) = \mu = n.p \qquad 3.71$$

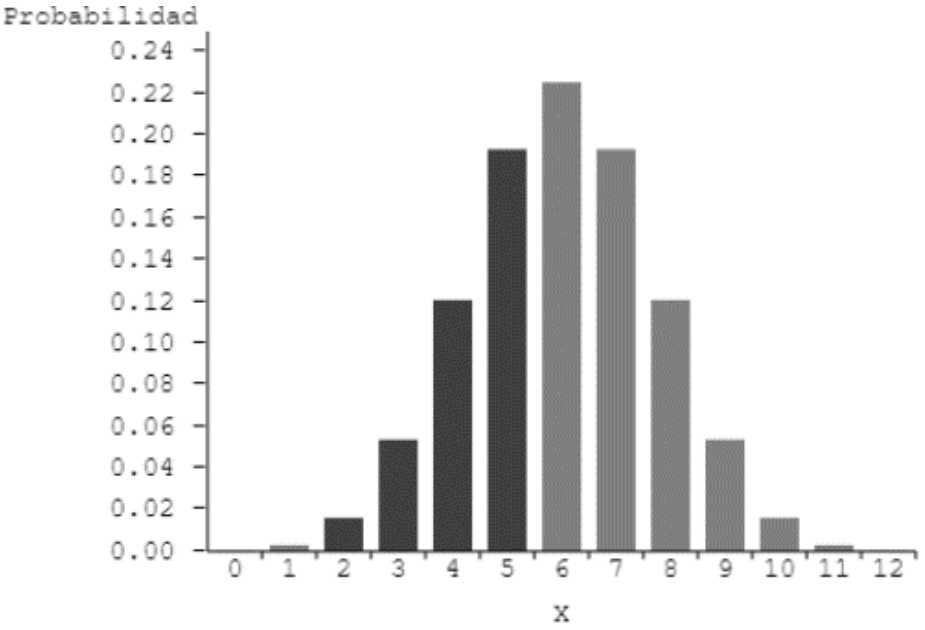
Desviación Típica de una Distribución Binomial

Conocida también como desviación estándar, es igual a:

$$\sigma = \sqrt{n.p.q} \qquad 3.72$$

En la Figura 90 se presenta un diagrama de distribución binomial correspondiente a p=0,5, n=12 y , $P\ (2 \leq x \leq 5)$. Las barras más oscuras corresponden a los valores de x=2, x=3 x=4 y x=5.

Figura 90
Distribución binomial, $n=12, p=0,5, P(2 \leq x \leq 5)$



Nota. Adaptada de Binario árbol [Imagen], Calculadorasonline, 23 de noviembre de 2022, Calculadoras de matemática (t.ly/kDWL). CC BY 3.0

Ejercicio de Probabilidad Binomial. Considérese el siguiente ejemplo para ilustrar el cálculo de probabilidades en la distribución binomial:

PowerPoint es usado, aproximadamente, en un 80% de presentaciones académicas. Hallar la probabilidad de que: a) en un grupo de 6 docentes, 2 utilicen PowerPoint para hacer sus diapositivas de clase; b) haya entre 2 y 4 docentes que usen el software; c) a lo sumo 4 docentes usen PowerPoint; d) al menos 3 docentes hagan sus diapositivas con PowerPoint.

Lo primero que se debe hacer es extraer los datos del enunciado: n = 6; p =0.8 y k =2. Como la suma de probabilidades de p y q debe ser igual a 1, entonces, ya que la probabilidad de éxito es 0.8, q = 0.2.

a) Aplicando la fórmula correspondiente a la distribución binomial, se obtiene:

$$P(x = 2) = \frac{6!}{2!\,(6-2)!}\, 0.8^2 0.2^4 \qquad 3.73$$

$$P(x = 2) = \frac{6.5.4!}{2!(4)!}\, 0.8^2 0.2^4 = 0.01536 \quad 3.74$$

$$P(x = 2) = 1.54\% \quad 3.75$$

La respuesta entonces, es que la probabilidad de que del grupo considerado de 6 docentes hayaexactamente 2 que usen PowerPoint para sus diapositivas de clase, es de 1,54%.

b).Hallar la probabilidad de que entre 2 y 4 docentes usen el software, significa encontrar la probabilidad de que $P(2 \leq x \leq 4)$ Se debe recordar que la probabilidad binomial es de variable discreta, con lo cual, se tiene que:

$$P(2 \leq x \leq 4) = P(x = 2) + P(x = 3) + P(x = 4) \quad 3.76$$

Aplicando la formula general, $P(x = k) = \frac{n!}{k!(n-k)!}\, p^k q^{n-k}$ se halla que

$$P(x = 2) = 1.54\% \text{ (parte a) del ejercicio)} \quad 3.77$$

$$P(x = 3) = \frac{\mathbf{6!}}{\mathbf{3!\,(3)!}}\, 0.8^3 0.2^3 = 0.0819 \quad 3.78$$

$$P(x = 4) = \frac{\mathbf{6!}}{\mathbf{4!\,(2)!}}\, 0.8^4 0.2^2 = 0.2458 \quad 3.79$$

de donde

$$\mathrm{P}(2 \leq x \leq 4) = 0.0154 + 0.0819 + 0.2458 = 0{,}3431 \quad 3.80$$

$$\mathrm{P}(2 \leq x \leq 4) = 34{,}31\% \quad 3.81$$

c) Para determinar la probabilidad de que, a lo sumo, 4 docentes usen PowerPoint, se debe recordar que esa expresión “a lo sumo” se refiere a un tope, cuando mucho 4 docentes, es decir,la probabilidad pedida es:

$$P(x \leq 4) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) \quad 3.82$$

Dado que en la parte b) del ejercicio ya se ha calculado la probabilidad P(2 $\leq x \leq 4$), se puede reescribir la probabilidad pedida de la siguiente manera:

$$P(x \leq 4) = P(x = 0) + P(x = 1) + P(2 \leq x \leq 4) \quad 3.83$$

se calculan ahora $P(x = 0) y\, P(x = 1)$:

$$P(x = 0) = \frac{6!}{0!\,(6)!}\, 0.8^0 0.2^6 = 0.0001 \quad 3.84$$

$$P(x = 1) = \frac{6!}{1!(5)!}\, 0.8^1 0.2^5 = 0.0015 \quad 3.85$$

luego,

$$P(x \leq 4) = P(x = 0) + P(x = 1) + P(2 \leq x \leq 4) \quad 3.86$$

$$P(x \leq 4) = 0.0001 + 0.0015 + 0{,}3431 = 0{,}3446 \quad 3.87$$

$$P(x \leq 4) = 34{,}47\% \quad 3.88$$

Cuando menos tres docentes utilicen PowerPoint para hacer sus diapositivas:

$$P(x \geq 3) = P(x = 3) + P(x = 4) + P(x = 5) + P(x = 6) \quad 3.89$$

Puesto que anteriormente se ha calculado restan por calcular

$$P(x = 5) = \frac{6!}{1!\,(5)!}\, 0.8^1 0.2^5 = 0.3932 \quad 3.90$$

$$P(x = 6) = \frac{6!}{6!(0)!}\, 0.8^6 0.2^0 = 0.2621 \quad 3.91$$

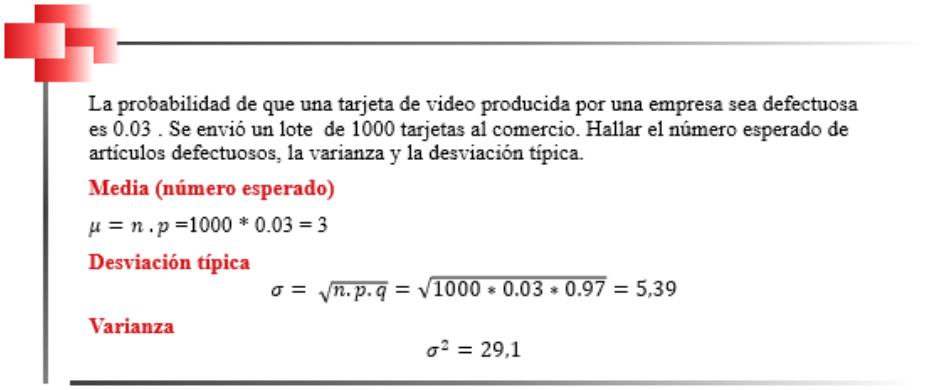
de donde

$$P(x \geq 3) = 0.0819 + 0.2458 + 0.3932 + 0.2621 = 0.983 \quad 3.92$$

En la Figura 91 se ilustra un ejemplo de cálculo de la media, la desviación típica y la mediana en una distribución binomial.

Figura 91

Ejemplo de cálculo de media, desviación típica y varianza en distribuciones binomiales



Distribución Normal

Corresponde a una distribución de variable aleatoria continua, x (Gauss, 1809). Es la más importante de las distribuciones de probabilidad, tanto en el ámbito científico como en el tecnológico, siendo uno de los requisitos fundamentales para muchas de las pruebas del contraste de hipótesis. Su gráfica tiene forma de campana y recibe el nombre de campana de Gauss o distribución gaussiana.

Una distribución normal depende de los valores de la media y de la desviación típica de la población, es decir, se define a través de dichos parámetros.

Si las observaciones de una variable aleatoria X, se pueden aproximar a una distribución normal, se escribe:

$$X \sim N(\mu, \sigma) \qquad 3.93$$

Características de la Distribución Normal

1. La distribución tiene forma de campana, es simétrica respecto a la media y se extiende desde $-\infty$ hasta $+\infty$.
2. Tiene una asíntota horizontal tanto por la izquierda como por la derecha.

3. La curva presenta dos puntos de inflexión, uno en

$$x = \mu + \sigma \qquad 3.94$$

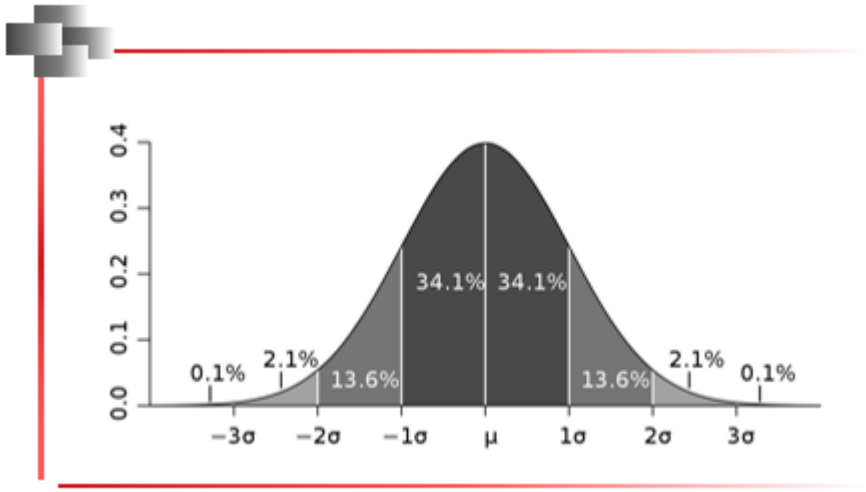
y, por simetría, otro en

$$x = \mu - \sigma \qquad 3.95$$

4. El área definida por el eje horizontal, las rectas $x = \mu + \sigma$ y $x = \mu - \sigma$ y la campana de Gauss, es igual a 0.6826, es decir 68.26% del área total. El área definida por el eje horizontal, las rectas $x = \mu + 2\sigma$ y $x = \mu$ - 2σ y la campana de Gauss, es igual a 0.9544, es decir 95.44% del área total, Figura 92.

Figura 92

Distribución normal, porcentajes de probabilidades



Nota. Adaptada de Distribución normal [Imagen], Kanijoman, 10 de febrero de 2012, Flickr (https://www.flickr.com/photos/23925401@N06/6850657337). CC BY 2.0

5. La curva presenta un máximo en

$$x = \mu \qquad 3.96$$

y *s*e cumple que la media, la mediana y la moda,tienen el mismo valor.

6. X puede tomar cualquiera de un número infinito de valores., y ya que x es una variable aleatoria continua, la probabilidad para $x = x_{\sigma}$ es igual a cero.

7. La suma de todas probabilidades asociadas a los distintos valores de x, es 1.
8. En las distribuciones de variables aleatorias continuas, el área es igual a la probabilidad.

Ejemplos de distribuciones normales, pueden ser estaturas y pesos, lapso de vida útil de un instrumento, error experimental de un laboratorio, etc.

Distribución Normal Tipificada o Estándar

Los valores de las probabilidades de una distribución normal requieren del cálculo de la funciónde probabilidad, pero para facilitar el hallazgo de los valores, éstos se encuentran tabulados en la conocida como "Tabla de distribución normal" (Gauss, 1809). En general, una distribución normal, N(μ, σ), puede tener cualquier valor para μ y para σ, por lo que se requeriría una Tabla cada vez que estos valores cambiaran. Para simplificar ese proceso, y que las probabilidades pudieran ser halladas en una sola Tabla, se procedió a tipificar o estandarizar la variable x, con el objeto de hallar una nueva variable tipificada, Z, que además es adimensional. Cuando se tipifica la variable x, la nueva distribución normal tiene $\mu = 0$ y $\sigma = 1$.

El objetivo de estandarizar la variable es que sólo se necesite una Tabla para poder encontrar cualquier valor de probabilidad. La variable tipificada responde a la relación:

$$Z = \frac{x - \mu}{\sigma} \qquad 3.97$$

Con la creación de aplicaciones en línea que permiten el cálculo de las probabilidades de la distribución normal, la Tabla de distribución normal ha caído en desuso, aunque a efectos prácticos,se continúa trabajando con la distribución normal tipificada.

Para efectos académicos, en la Figura 93 se presenta una transcripción de la Tabla de distribución normal tipificada.

Obsérvese que en la gráfica de distribución normal, situada en la esquina superior derecha, se encuentra coloreada toda la cola izquierda de la campana, ello indica que en la Tabla se encuentran tabuladas las probabilidades "menores que".

También en la Figura 93, para Z=0 la probabilidad es igual a 0.5, y a partir de Z=3 el valor dela probabilidad es muy cercano a 1. Lo anterior pone en

relevancia dos cosas: La Tabla sólo da probabilidades para $Z \geq 0$ y el área bajo la curva de una distribución normal es igual a 1.

Para hallar una probabilidad dentro de la Tabla debe encontrarse el valor de Z; para ello, en la primera (o en la última) columna, se ubica, del valor calculado de Z, para un determinado x de interés, el número entero acompañado del primer decimal y, posteriormente, en la fila superior, seubica el segundo decimal del valor de Z.

Supóngase que, la variable altura promedio de un grupo de estudiantes se puede aproximar auna distribución normal N(1.68m, 0.12m). La variable tipificada sería:

$$Z = \frac{x-1.68}{0.12}, \text{ con N(0,1)} \qquad 3.98$$

Para x $\leq$1,73 m, el valor correspondiente, calculado con la expresión anterior, sería Z $\leq$ 0.42.

Para encontrar el valor de la probabilidad correspondiente a Z=0.42, se ubica, en la columna izquierda de la Tabla, el entero y el primer decimal de Z, es decir 0.4 y, a continuación, en la fila superior, el segundo decimal, es decir, 0.02. Posteriormente, se busca la intersección entre la fila y la columna y allí se leerá la probabilidad correspondiente a la variable estandarizada, Figura 94:

$$P(Z \leq 0.42) = 0{,}6628 \qquad 3.99$$

Figura 93

Tabla de distribución normal estándar

$\mu = Media$

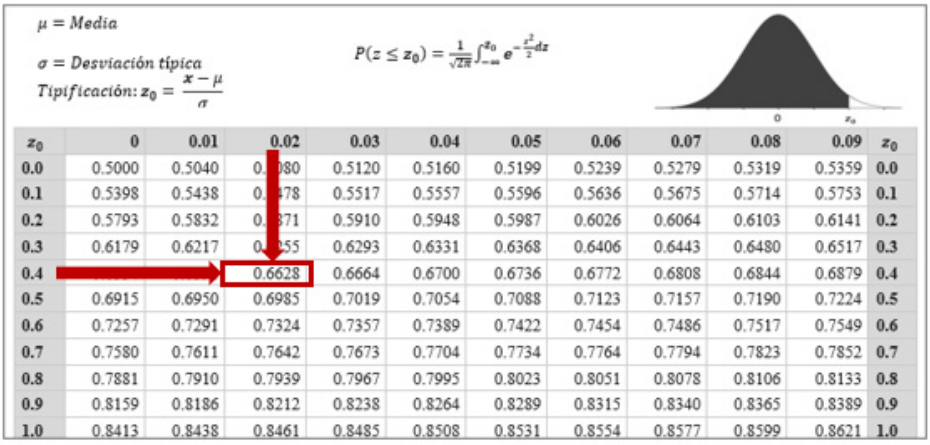
$\sigma = Desviación\ típica$

$Tipificación: z_0 = \frac{x-\mu}{\sigma}$

$P(z \leq z_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_0} e^{-\frac{z^2}{2}} dz$

z_0	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z_0
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	**0.0**
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	**0.1**
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	**0.2**
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	**0.3**
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	**0.4**
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	**0.5**
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	**0.6**
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	**0.7**
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	**0.8**
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	**0.9**
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	**1.0**
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	**1.1**
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	**1.2**
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	**1.3**
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	**1.4**
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	**1.5**
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	**1.6**
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	**1.7**
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	**1.8**
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767	**1.9**
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	**2.0**
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857	**2.1**
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890	**2.2**
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	**2.3**
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	**2.4**
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	**2.5**
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964	**2.6**
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974	**2.7**
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	**2.8**
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	**2.9**
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900	**3.0**
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929	**3.1**
3.2	0.99931	0.99943	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950	**3.2**
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965	**3.3**
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976	**3.4**
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983	**3.5**
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989	**3.6**
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992	**3.7**
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995	**3.8**
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997	**3.9**

Figura 94
Detalle de ubicación de la probabilidad para Z = 0.42



$\mu = Media$

$\sigma = Desviación\ típica$

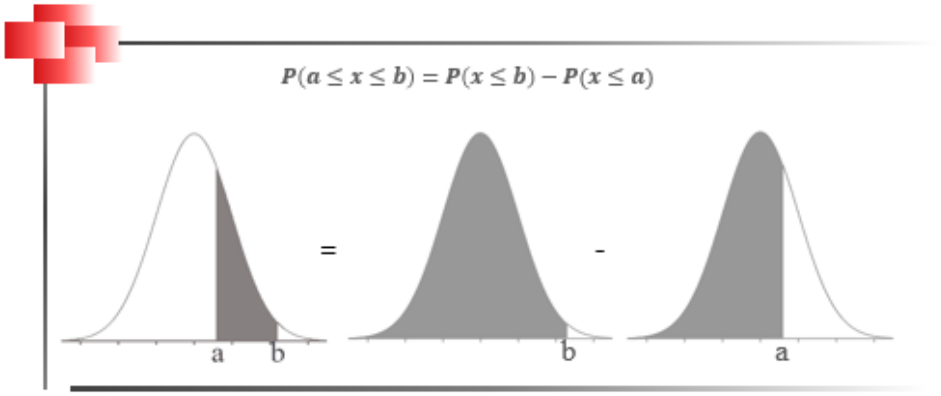
$Tipificación: z_0 = \frac{x-\mu}{\sigma}$

$P(z \le z_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_0} e^{-\frac{t^2}{2}} dz$

z_0	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z_0
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	0.0
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.1
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.2
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	0.3
0.4			0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	0.4
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	0.5
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.6
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	0.7
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.8
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	0.9
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	1.0

Por otra parte, la probabilidad de x tome cualquier valor dentro de un intervalo [a, b], es igual alárea bajo la curva entre los puntos a y b, Figura 95.

Figura 95
Probabilidad de que x tome un valor en el intervalo [a, b]



Como se había señalado anteriormente, si *x* es igual a un valor en particular, por ejemplo *x* = *a*; como no hay área arriba del punto en la distribución de probabilidad para una variable aleatoria continua, la probabilidad es igual a cero, Figura 96.

Lo anterior indica que:

$$\boldsymbol{P}(\mathrm{x} = \boldsymbol{a}) = \boldsymbol{0} \text{ para variables aleatorias continuas} \qquad 3.100$$

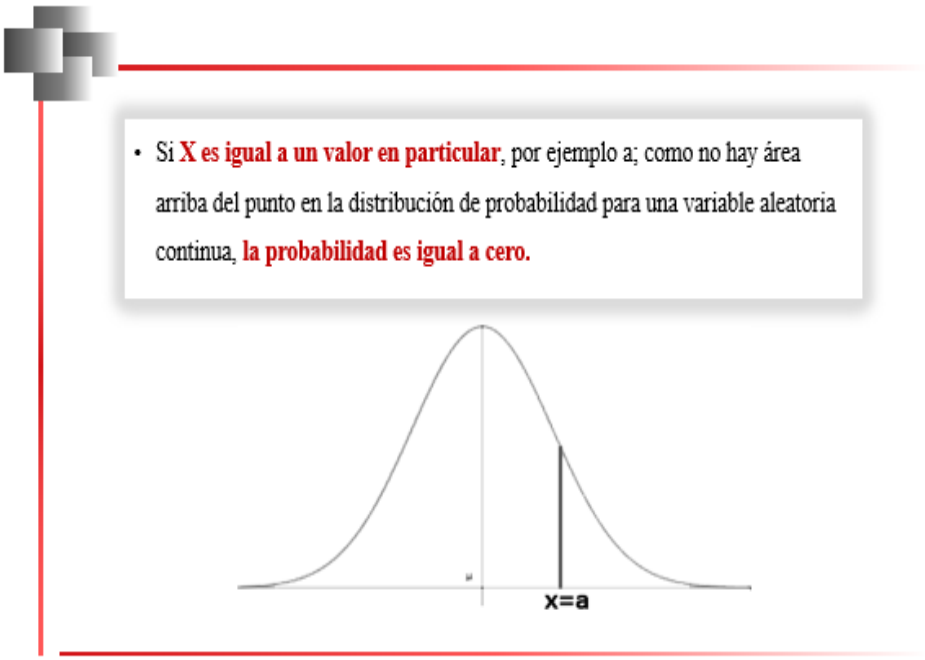
de donde se sigue que:

$$\boldsymbol{P}(\boldsymbol{x} \geq \boldsymbol{a}) = \boldsymbol{P}(\boldsymbol{x} > \boldsymbol{a}) \qquad 3.101$$

$$\boldsymbol{P}(\boldsymbol{x} \leq \boldsymbol{a}) = \boldsymbol{P}(\boldsymbol{x} < \boldsymbol{a}) \qquad 3.102$$

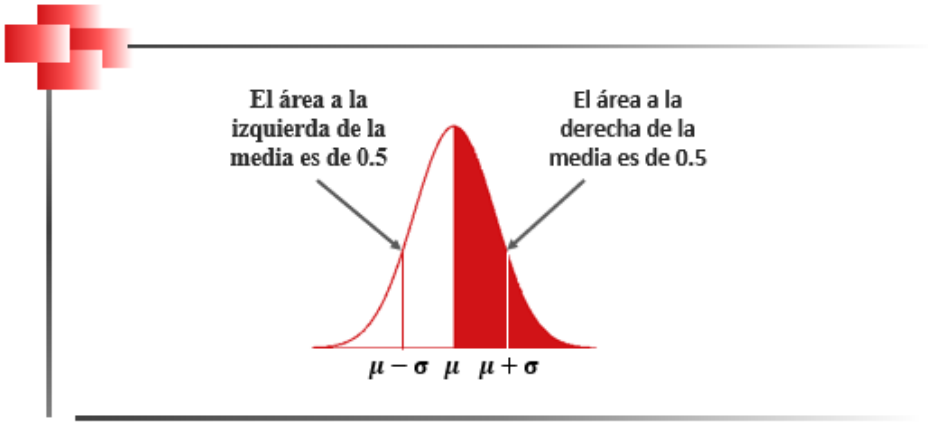
La media de la distribución normal, μ, se localiza en el centro de la distribución, esto indica que a cada lado de la media queda un área igual a 0.5; es decir, que la probabilidad es del 50%, Figura 97.

Figura 96
Probabilidad de



Por otra parte gráficamente, a medida que σ aumenta, se reduce la altura de la curva.

Figura 97
Distribución de áreas a cada lado de la media



Cuando una variable aleatoria, x, está estandarizada, su valor se expresa como el número dedesviaciones estándar, σ, que se encuentran a la izquierda o a la derecha de la media, μ.

Dado que $Z = \frac{x-\mu}{\sigma}$, si se despeja x, se obtiene que

$$x = \mu + Z\sigma \qquad 3.103$$

Esto quiere decir que cuando:

$$x < \mu, \ Z < 0 \qquad 3.104$$

$$x > \mu, \ Z > 0 \qquad 3.105$$

$$x = \mu, \ Z = 0 \qquad 3.106$$

Recuérdese que, cuando la distribución normal está estandarizada, $\mu = 0$ y $\sigma = 1$.

Como se comentó anteriormente, la Tabla de distribución normal, incluida en la Figura 93, por ser de cola izquierda, sólo da probabilidades menores que un determinado número, es decir $P(x \leq a)$, *o lo que es igual,* $P(x < a)$.

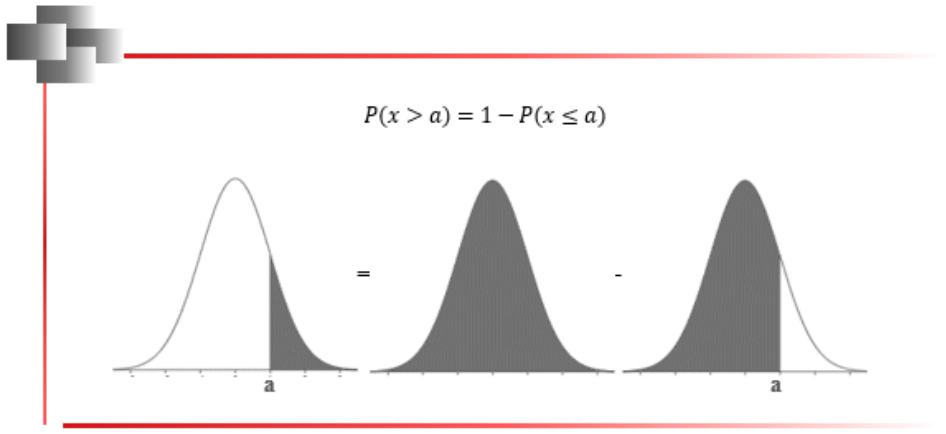
Si se desea hallar una probabilidad "mayor que", Figura 98, se puede utilizar la Tabla utilizando la diferencia de áreas, ya que se sabe que el área total es igual a 1:

$$P(x \geq a) = P(x > a) \qquad 3.107$$

$$P(x > a) = 1 - P(x \leq a) = 1 - P(x < a) \qquad 3.108$$

Otra forma para calcular las probabilidades correspondientes a una distribución normal, es utilizar algún software disponible en línea hay varios, entre ellos, se puede citar el de Domínguez y Domínguez (2006-2020), denominado CalEst, en su versión 4.4, bajo licencia de Software propietario, pero que ofrece la posibilidad de hacer uso de la Graficadora de Distribución Normal, que permite calcular de manera sencilla probabilidades mayores, probabilidades menores y probabilidades en un intervalo.

Figura 98
Planteamiento gráfico de probabilidad "mayor que"



A efectos académicos, la Graficadora de Distribución Normal sólo será usada para incluir las gráficas en el desarrollo de los ejercicios; no así para el cálculo de las probabilidades, a cuyo efecto se continuará utilizando la Tabla de distribución normal estándar, que fue incluida en la Figura 93.

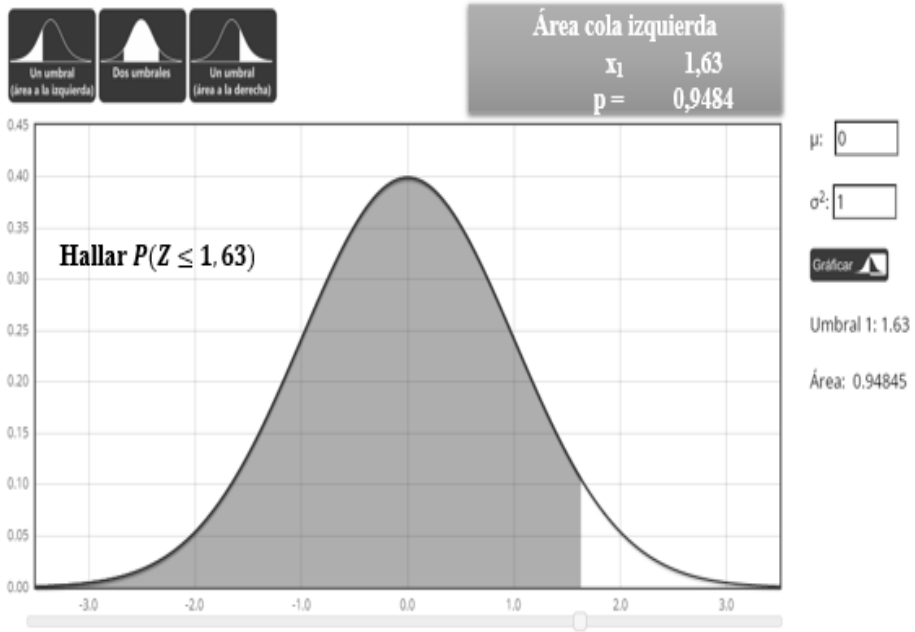
En la Figura 99 se presenta un ejemplo de cálculo de una probabilidad "menor que", la cual, dado que, además, la variable se encuentra tipificada, puede ser leída directamente en la Tabla de distribución normal, $P(Z < 1{,}63)$.

Para encontrar el valor de probabilidad en la Tabla de la Figura 93, se busca en la columna izquierda el valor de Z = 1.6 y se traza una línea horizontal, luego se ubica en la fila superior el valor de 0.03 y se traza una línea vertical, en el punto donde se encuentran las líneas anteriormente trazadas se lee el valor de la probabilidad, $P(Z < 1{,}63) = 0.9484$. Recuérdese, una vez más, que la Tablade distribución normal de la Figura 93, sólo da probabilidades positivas y "menores que". Compruébese además que la Gráfica de Distribución Normal del software CalEst, arroja un resultado idéntico para la probabilidad pedida, Figura 99.

En la Figura 100, se ilustra gráficamente cómo hallar la probabilidad cuando Z es menor o igualque un número negativo "a", $P(Z \leq -a)$.

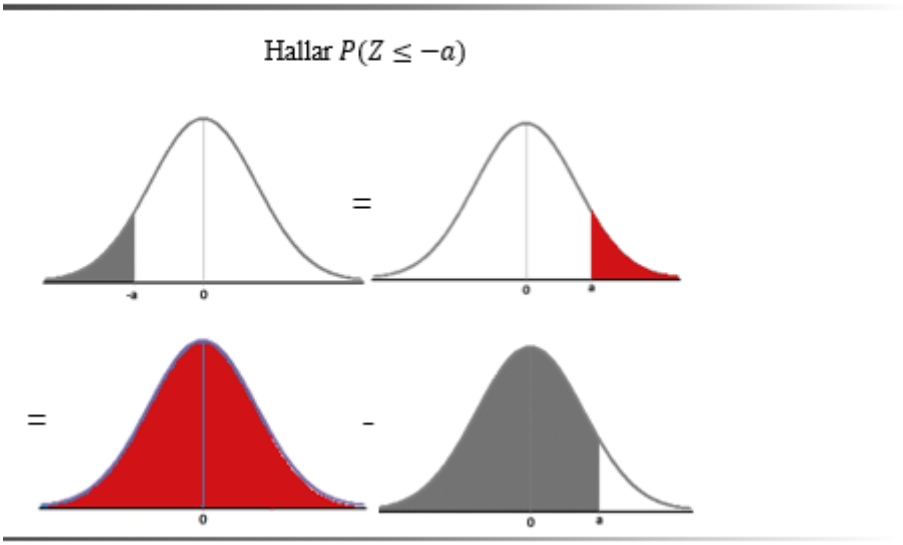
Figura 99

Ejemplo de cálculo de una probabilidad "menor que".



Nota. Gráfica construida con la Graficadora de Distribución Normal [software en línea] (Domínguez & Domínguez, 2006-2020) (http://www.calest.com/Graficadora.aspx).

Figura 100
Planteamiento gráfico para hallar $P(Z \leq -a)$



Para calcular la probabilidad de un número sea negativo, $P(Z \leq -a)$, se debe recordar que la distribución normal es simétrica, y que, por lo tanto, se cumple que:

$$P(Z \leq -a) = P(Z \geq a) = P(Z > a) \qquad 3.109$$

Obsérvese que lo anterior es como proyectar el área pedida sobre un **espejo** (Triolla, 2018, Rossman, A. J., & Chance, 2018).

De esta forma, se ha eliminado el inconveniente presentado por el hecho de que en la Tabla de la Figura 93 no se pueden leer probabilidades para valores de Z negativos; pero se ha presentado ahora otro problema: la Tabla tampoco proporciona valores de probabilidad para Z mayor que un cierto valor, "a". La solución de este segundo problema también es sencilla: calcular **el complemento** de la probabilidad pedida. Esto es:

$$P(Z > a) = 1 - P(Z \leq a) = 1 - P(Z < a) \qquad 3.110$$

En la ecuación anterior, "1" representa el área total bajo la curva de Gauss (que es equivalente a la probabilidad del ciento por ciento) y $P(Z < a)$ puede ser leída directamente en la Tabla de la Figura 93.

Para el caso de que se requiera calcular la probabilidad mayor que de un número negativo, $P(Z > -a)$, Figura 101, se presenta una doble restricción para la lectura de la probabilidad, ya que la Tabla de distribución normal estandarizada no contiene valores negativos de Z ni probabilidades "mayores que".

En este caso es necesario recordar que la distribución normal tiene la propiedad de simetría, por lo tanto, calcular la probabilidad de que Z sea mayor que un valor negativo es igual a calcular la probabilidad de que Z sea menor que :

$$P(Z >) = P(Z < a) \qquad 3.111$$

En la Figura 102 se presenta la transformación que se debe realizar en la gráfica de la Figura 101, para que la probabilidad $P(Z > -a)$ pueda ser calculada en forma directa en la Tabla Normal.

Básicamente, como se mencionó anteriormente, hay dos cosas que se pueden hacer cuando se busca la probabilidad de un valor de Z que no aparece en la Tabla:

1. Simular la proyección de la gráfica original en un espejo. Este procedimiento se aplica paravalores de Z que sean negativos.
2. Hallar el complemento de la probabilidad pedida, que se refiere a restar de la unidad la probabilidad con signo contrario a la inicial

$$P(Z >) = P(Z < a) \qquad 3.112$$

Figura 101

Gráfico correspondiente a $P(Z \leq -a)$

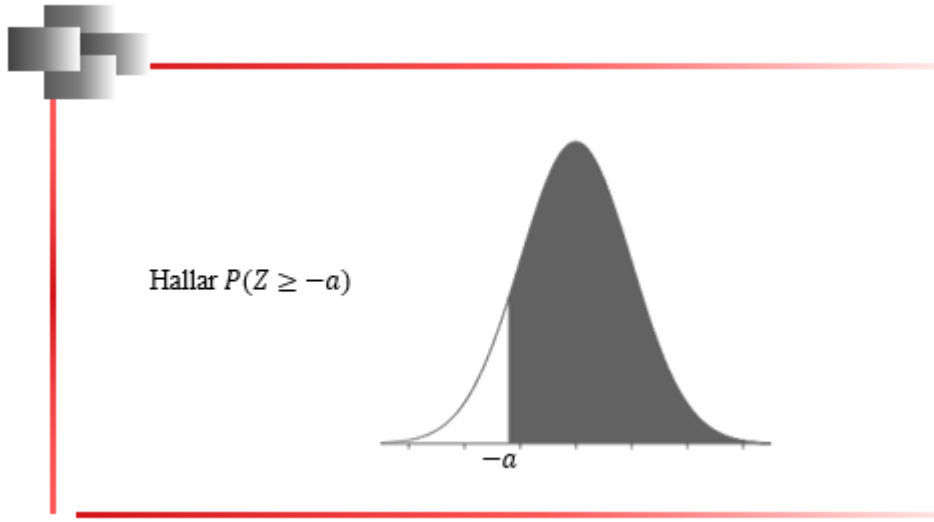
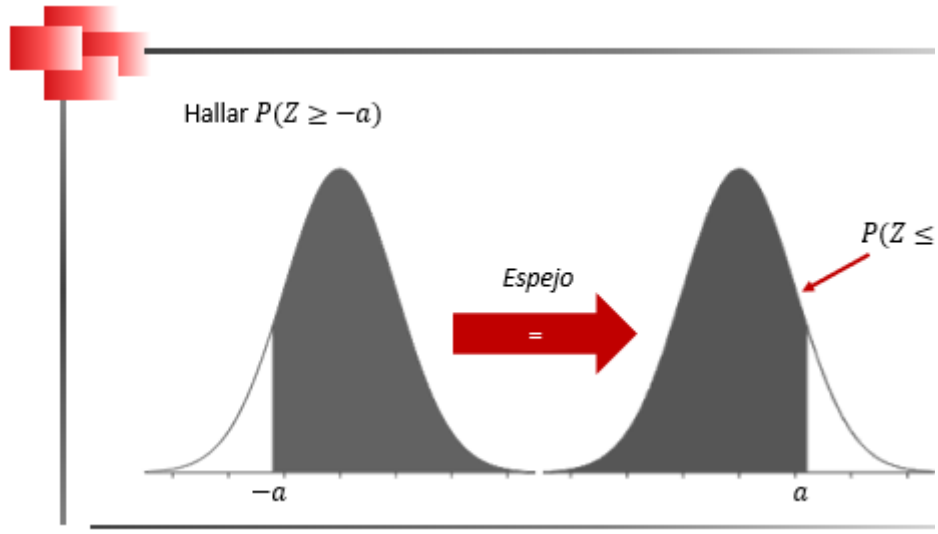


Figura 102

Transformación de $P(Z < a)$ *para el cálculo directo en la Tabla Normal*



En la Figura 103 se presenta la gráfica correspondiente a $P(-b \leq Z \leq -a)$ y en la Figura 104, se ilustra la aplicación de la técnica del espejo para el cálculo de la probabilidad en la Tabla.

Figura 103
Gráfica correspondiente a $P(-b \leq Z \leq -a)$

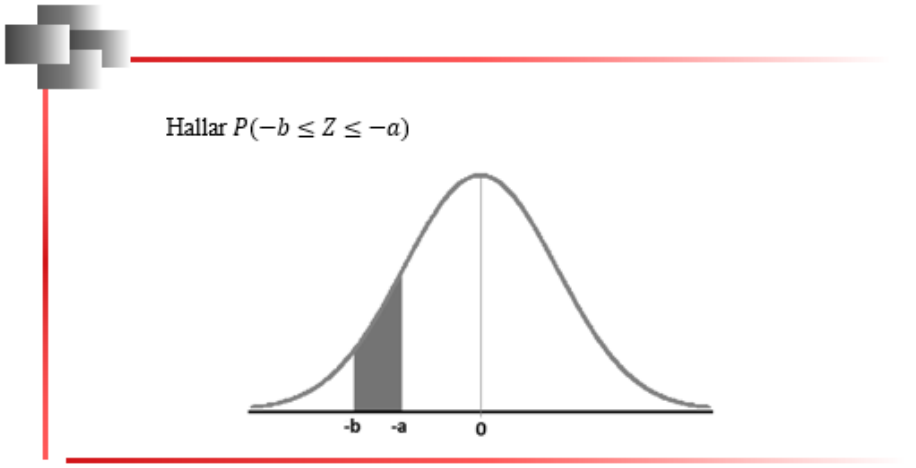
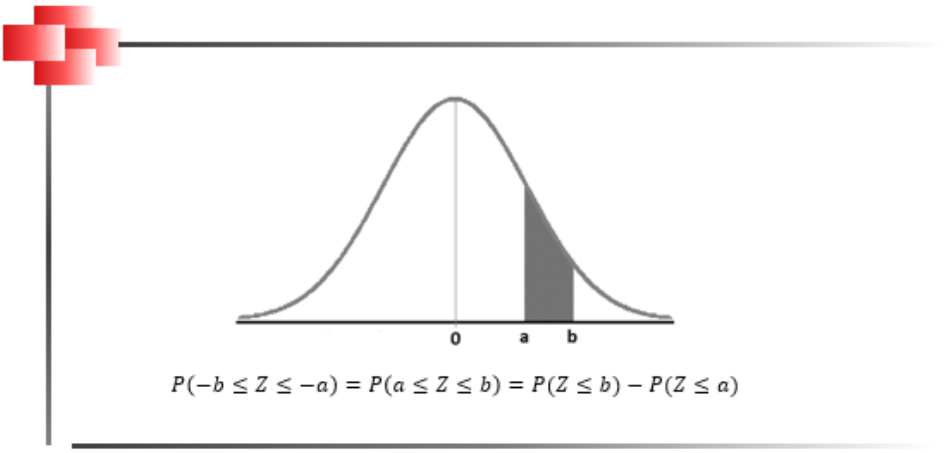
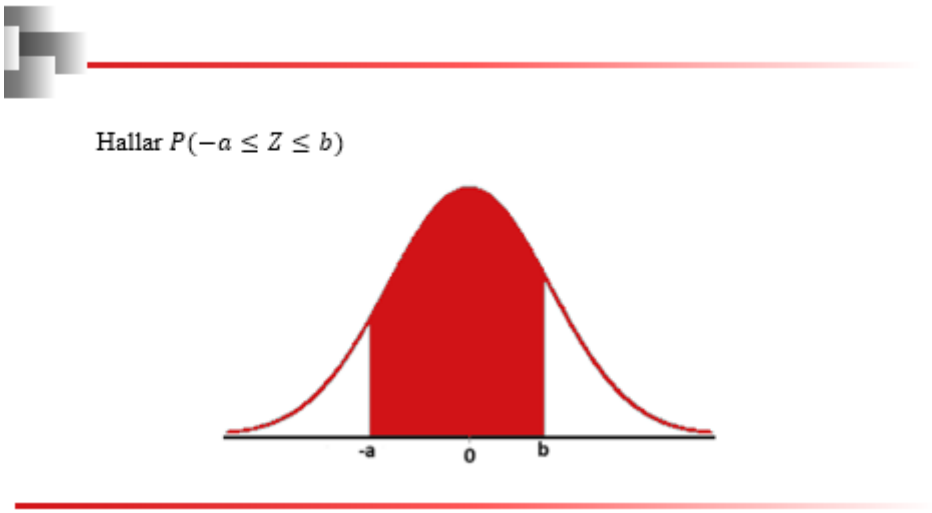


Figura 104
Transformación de $P(-b \leq Z \leq -a)$*para el cálculo directo en la Tabla Normal*



En la Figura 105 se muestra el área involucrada en el cálculo de la $P(-a \leq Z \leq b)$.Obsérvese que, en este caso el intervalo a considerar, comienza en un número negativo (antes de la media) y culmina en un número positivo de la distribución normal estandarizada.

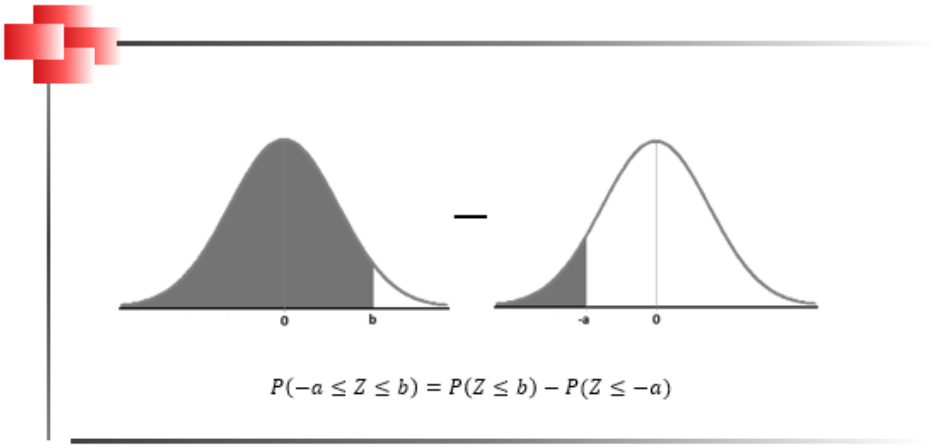
Figura 105
Gráfica correspondiente a $P(-a \leq Z \leq b)$



Para hallar la probabilidad pedida, se debe realizar el mismo planteamiento que se hace con cualquier intervalo, Figura 106, es decir:

$$P(-a \leq Z \leq b) = P(Z \leq b) - P(Z \leq -a) \qquad 3.112$$

Figura 106
Gráfica de la probabilidad $P(Z \leq b)$ - $P(Z \leq a)$

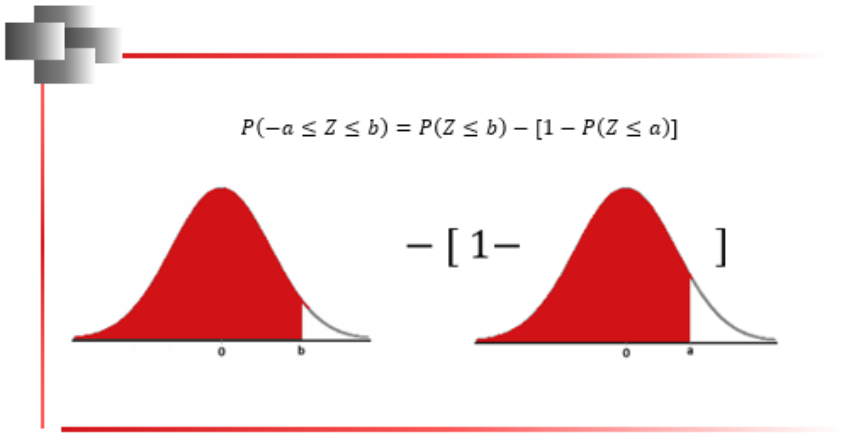


Con la probabilidad $P(Z \leq b)$ no se tiene ningún problema, puede ser leída directamente sobrela Tabla normal una vez definido el valor de b. Sin embargo, como dicha Tabla sólo tiene tabulados valores positivos, aplicamos las técnicas del espejo y del complemento, Figura 107, con lo cual se tendría:

$$P(-a \leq Z \leq b) = P(Z \leq b) - [1 - P(Z \leq a)] \qquad 3.112$$

Figura 107

Aplicación de técnicas de espejo y complemento para el cálculo de $P(-a \leq Z \leq b)$



Es recomendable que cada vez que se vaya a resolver un ejercicio de cálculo de probabilidades con el uso de tablas, se haga un gráfico para poder realizar un planteamiento correcto.

Considérese ahora el siguiente ejemplo:

Las calificaciones de los estudiantes de una determinada asignatura siguen una distribución normal de $\mu = 6, y\ \sigma = 2$. Se desea calcular la probabilidad de obtener una nota aprobatoria $P(7 \leq x \leq 10)$

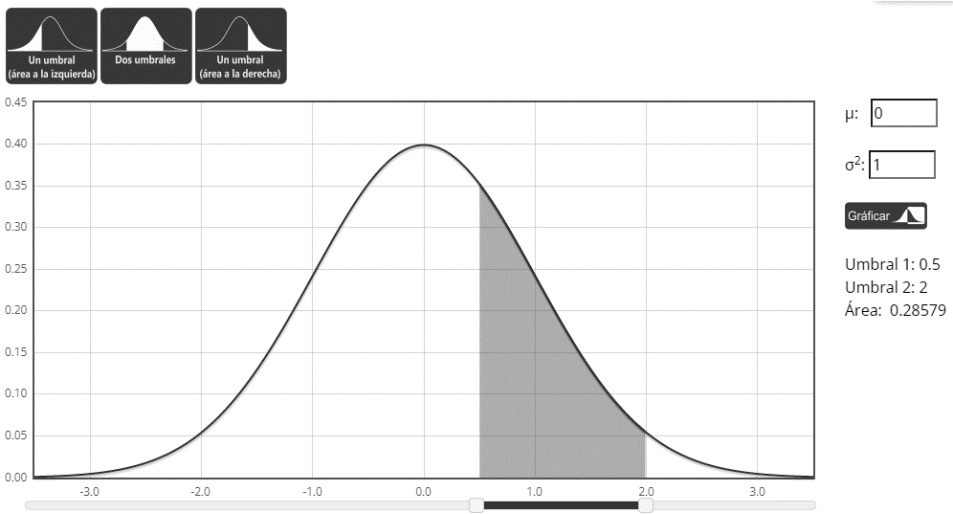
Lo primero que se debe hacer es tipificar la variable, para ello se usa la fórmula:

$$Z = \frac{x-\mu}{\sigma} = \frac{x-6}{2} = {}^{x}/_{2} - 3 \qquad 3.113$$

De donde, $P(0.5 \leq Z \leq 2)$. Como se trata de dos números positivos, se tiene la situación típica del cálculo de probabilidades de un intervalo, Figura 108:

$$P(0.5 \leq Z \leq 2) = P(Z \leq 2) - P(Z \leq 0.5) = 0.28579 \qquad 3.114$$

Figura 108
Uso del software CalEst para el cálculo de probabilidades



Nota. Gráfica construida con la Graficadora de Distribución Normal [software en línea] (Domínguez & Domínguez, 2006-2020) (http://www.calest.com/Graficadora.aspx).

En la Figura 109, Figura 110, Figura 111 y Figura 112, se incluyen otros ejemplos que incluyen diferentes tipos de probabilidades. Es importante recordar que cuando se desea calcular la probabilidad de un valor de Z comprendido entre dos números, se debe restar la probabilidad de ambos extremos.

$$P(a \leq Z \leq b) = P(Z \leq a) - P(Z \leq b) \qquad 3.115$$

Figura 109

Cálculo de la probabilidad de que $x > 1.3$, *con* $\mu = 1.18$ *y* $\sigma = 0.63$

1. Extraer los datos

   $x = 1.3$
   $\mu = 1.18$
   $\sigma = 0.63$

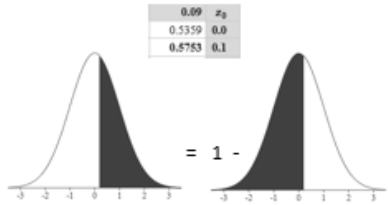
2. Estandarizar la variable

   $Z = \frac{x-\mu}{\sigma} = \frac{1.3-1.18}{0.53} = 0.19$

3. Plantear probabilidad en términos de Z

   $P\ (x > 1.3) = P(Z\ > 0.19)$

4. Construir gráfica



5. Calcular la probabilidad

   $P(Z\ > 0.19) = 1 - P(Z < 0.19) = 0.4247$

Figura 110

Cálculo de la probabilidad de que $1.38 \leq x \leq 2.57$, *con* $\mu = 1.18$ *y* $\sigma = 0.63$

1. Extraer los datos

   $x_1 = 1.38, \quad x_2 = 2.55$
   $\mu = 4.29$
   $\sigma = 1.26$

2. Estandarizar las variables

   $Z_1 = \frac{x-\mu}{\sigma} = \frac{1.38-4.29}{1.26} = -2.31$

   $Z_2 = \frac{x-\mu}{\sigma} = \frac{2.55-4.29}{1.26} = -1.38$

3. Plantear probabilidad en términos de Z

   $P\ (1.38 \leq x \leq 2.55)$

   $=$

   $P(-2.31 \leq Z \leq -1.38)$

z_0	0	0.01
2.3	0.9893	0.9896

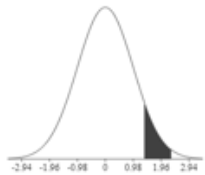
0.08	0.09	z_0
0.9162	0.9177	1.3

4. Construir gráfica (Ver Figura 104)

   $P(-2.31 \leq Z \leq -1.38)$

   $=$

   $P(2.31 \leq Z \leq 1.38)$



5. Calcular la probabilidad

   $P(2.31 \leq Z \leq 1.38)$

   $=$

   $P\ (Z \leq 2.31) - P\ (Z \leq 1.38)$

   $=$

   $0.9896 - 0.9162 = 0.0734$

Figura 111

Cálculo de la probabilidad de que $2.84 \leq x \leq 3.35$, *con* $\mu = 2.96$ *y* $\sigma = 0.98$

1. Extraer los datos

$x_1 = 2.84, \quad x_2 = 3.35$
$\mu = 2.96$
$\sigma = 0.98$

2. Estandarizar las variables

$$Z_1 = \frac{x-\mu}{\sigma} = \frac{2.84-2.96}{0.98} = -0.12$$

$$Z_2 = \frac{x-\mu}{\sigma} = \frac{3.35-2.96}{0.98} = 0.40$$

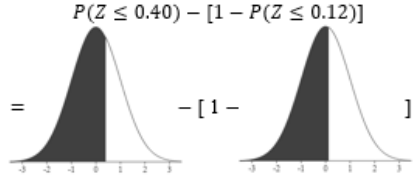
z_0	0
0.0	0.5000
0.1	0.5398
0.2	0.5793
0.3	0.6179
0.4	**0.6554**

3. Plantear probabilidad en términos de Z

$$P(2.84 \leq x \leq 3.35)$$
$$=$$
$$P(-0.12 \leq Z \leq 0.40)$$

4. Construir gráfica (Ver Figura 107)

$$P(-0.12 \leq Z \leq 0.40)$$
$$=$$
$$P(Z \leq 0.40) - [1 - P(Z \leq 0.12)]$$



z_0	0	0.01	0.02
0.0	0.5000	0.5040	0.5080
0.1	0.5398	0.5438	**0.5478**

5. Calcular la probabilidad

$$P(-0.12 \leq Z \leq 0.40)$$
$$=$$
$$0.6554 - [1 - 0.5478] = 0.2032$$

Figura 112

Cálculo de la probabilidad de que $x > 8{,}41$ *o* $x \leq 6.18$, *con* $\mu = 7.1$ *y* $\sigma = 2.73$

1. Extraer los datos

$x_1 = 8.41, \quad x_2 = 6.18$
$\mu = 7.1$
$\sigma = 2.73$

2. Estandarizar las variables

$$Z_1 = \frac{x-\mu}{\sigma} = \frac{9.29-7,1}{2.73} = 0.48$$

$$Z_2 = \frac{x-\mu}{\sigma} = \frac{6.18-7.1}{2.73} = -0.34$$

3. Plantear probabilidad en términos de Z

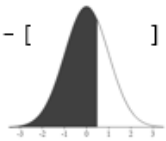
$$P(x > 8.41) = P(Z > 0.48)$$
$$P(x \leq 6.18) = P(Z \leq -0.34)$$

0.08	0.09	z_0
0.6844	0.6879	**0.4**

z_0	0.03	0.04
0.3	0.6293	**0.6331**

4. Construir gráficas

$P(Z > 0.48) = 1 - [$]



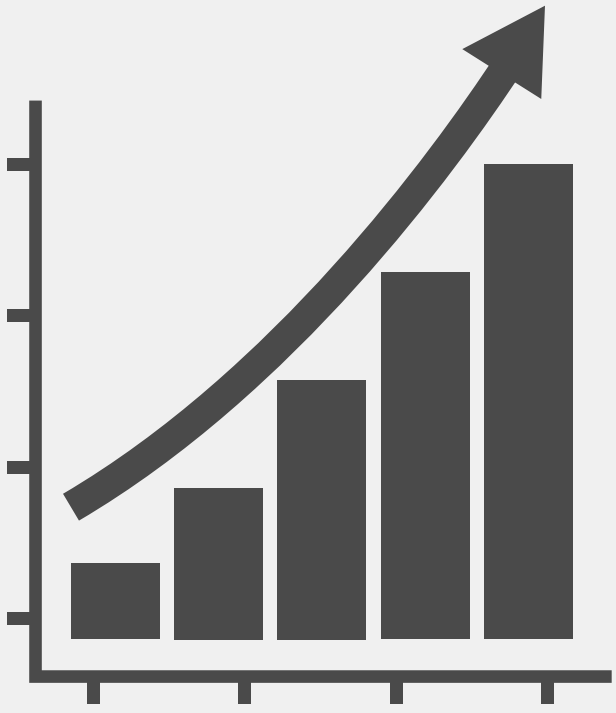
$P(Z \leq -0.34) = P(Z \geq 0.34)$

5. Calcular la probabilidad

$$P(Z > 0.48) + P(Z \leq -0.34)$$
$$[1 - P(Z \leq 0.48)] + [1 - P(Z<0.34)]$$
$$= 1 - 0.6844 + 1 - 0.6331 = 0.6825$$

Capítulo 4

Ajuste de curvas



CAPÍTULO IV

AJUSTE DE CURVAS

Ajuste por Mínimos Cuadrados

Modelo de Regresión Lineal Simple

Un modelo de regresión lineal simple es un modelo matemático que busca establecer una relación lineal entre una variable independiente (denominada x) y una variable dependiente (denominada y), mediante el uso de una ecuación de la forma $y = mx + b$.

Con este modelo se busca que, a partir de un conjunto de observaciones de ambas variables, se puedan estimar los valores de los coeficientes m y b que mejor describan la relación lineal entre ambas variables; donde m es la pendiente de la recta y b corresponde a la ordenada en el origen (valor de y cuando x =0).

El modelo de regresión lineal simple fue propuesto originalmente por el matemático francés Adrien-Marie Legendre en el año 1805, aunque fue el estadístico británico Francis Galton (1886) quien lo popularizó en el campo de la estadística en el siglo XIX, utilizando el término "regresión" para describir la tendencia de las variables a "regresar" hacia su media. Desde entonces, el modelo de regresión lineal simple ha sido ampliamente utilizado en la estadística y otras disciplinas como herramienta para describir y predecir las relaciones entre variables.

La creación de un modelo de regresión lineal consiste en hallar una ecuación de una recta (modelo de regresión) que sea capaz de explicar, como se señaló anteriormente, la relación lineal que existe entre dos variables, x e y.

La variable ***y,*** además de variable dependiente, recibe el nombre de variable estimada o de respuesta. La variable ***x***, la independiente, también es conocida como regresora, explicativa o predictora.

Recta de Regresión Ajustada (Mínimos Cuadrados)

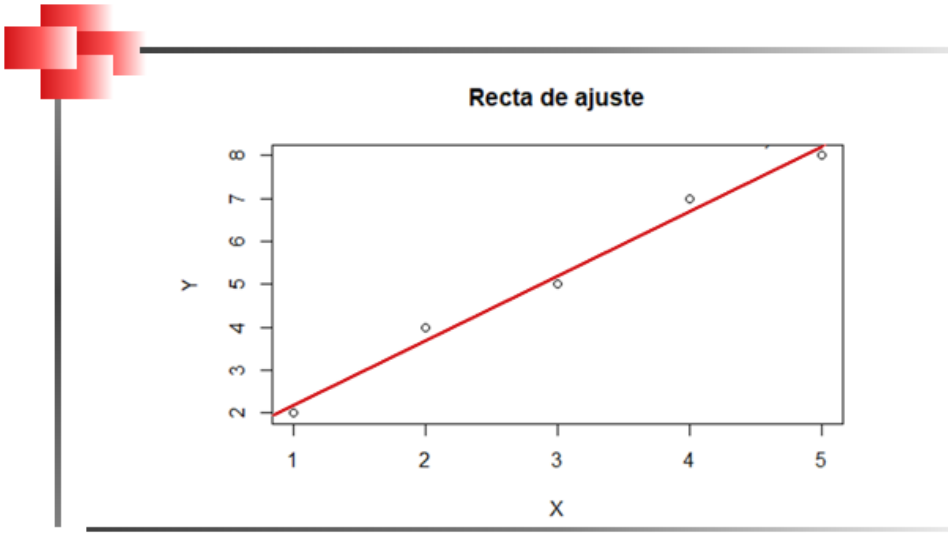
Corresponde a la recta que mejor se ajusta a los datos. Esta recta pasa por el punto $\bar{x}\ \bar{y}$, que recibe el nombre de centro de gravedad, donde

$$\bar{x} = \frac{\sum x_i}{n}; \quad \bar{y} = \frac{\sum y_i}{n} \qquad 4.1$$

La estimación de la pendiente y la ordenada en el origen que definen a la recta se hace de tal forma que "se minimice la suma de los cuadrados de las desviaciones de las observaciones respecto de la recta" (Baccini *et al.,* 2018, p.206).

En la Figura 113 se ilustra la aplicación del método de mínimos cuadrados en la obtención de una recta de regresión.

Figura 113
Método de mínimos cuadrados para ajuste de recta de regresión



Nota. Gráfico obtenido en software libre RStudio, con programación propia.

Lo que se hace con esta técnica de mínimos cuadrados es definir la distancia, medida sobre el eje y, de cada punto observado en relación a la recta de ajuste propuesta. La ecuación de dicha recta quedará definida cuando la suma de todas esas distancias adquiera el valor más pequeño posible, esto es, se haga mínima.

Cuando se plantea una regresión lineal simple entre dos variables, pueden presentarse varias situaciones:

1. Que se presente una relación directa entre las variables, esto es, que a medida que aumenta la variable explicativa (x), aumenta la variable dependiente (y).

2. Que se presente una relación indirecta entre la variable regresora (x) y la variable de respuesta (y), esto es, que a medida que aumenta el valor de *x*, disminuya el valor de *y*.

3. Que no haya relación entre las variables.

En la Figura 114 se presentan tres gráficas que ilustran la relación que puede existir entre la variable independiente y la dependiente.

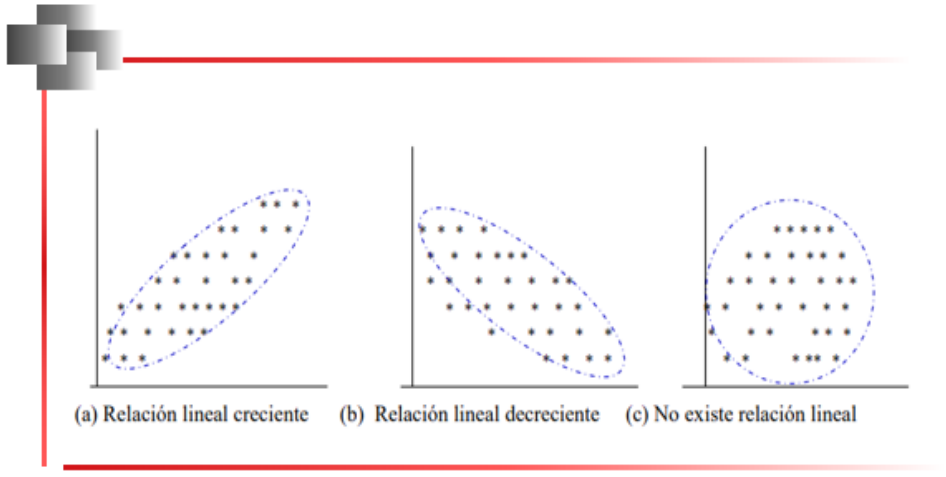
El signo de la pendiente, como en toda recta, indica si la recta es ascendente o descendente; pero no señala de ninguna forma la relación que puede existir entre las variables.

La primera gráfica de la Figura 114 corresponde a una recta de ajuste con pendiente positiva y la segunda, a una recta con pendiente negativa.

Obsérvese que en la tercera gráfica se presenta una nube de puntos en la que no es posible encontrar una recta que se ajuste a ellos, en ese caso, se dice que las variables no se encuentran relacionadas.

Figura 114

Naturaleza y fuerza de la relación entre las variables x e y



Nota. Adaptado de Gráfico 5.2 , de Nolberto Sifuentes & Ponce Aruneri, 2018, p.150.

Covarianza de una Muestra con dos Variables

Se calcula como la media aritmética de los productos de las desviaciones de cada una de lasvariables respecto a sus medias respectivas:

$$S_{xy} = \frac{1}{n}\sum x_i . y_i - \bar{x}\,\bar{y} \qquad 4.2$$

Si $S_{xy} > 0$ la relación entre las variables es directa,

Si $S_{xy} < 0$ la relación entre las variables es inversa,

Si S_{xy} =0, no existe correlación.

Varianzas de las Variables Independiente y Dependiente

La varianza para la variable independiente (***x***) se calcula como la diferencia entre la suma de cuadrados de los valores de la variable independiente y el cuadrado de la media de esa variable (Fisher, 1925):

$$S_x^2 = \frac{1}{n}\sum {x_i}^2 - (\bar{x})^2 \qquad 4.2$$

Para la variable dependiente (***y***) la varianza está definida a través de la siguiente ecuación:

$$S_y^2 = \frac{1}{n}\sum {y_i}^2 - (\bar{y})^2 \qquad 4.3$$

Correlación Lineal

Estimación de los Coeficientes de Regresión

En el caso de una regresión lineal simple, los datos, a través del método de mínimos cuadrados, son ajustados a una recta. Una de las formas en que se puede escribir la ecuación de una recta es a través de la pendiente-ordenada en el origen:

$$\hat{y} = mx + b \qquad 4.4$$

donde

$$m = \frac{S_{xy}}{S_x{}^2} \qquad 4.5$$

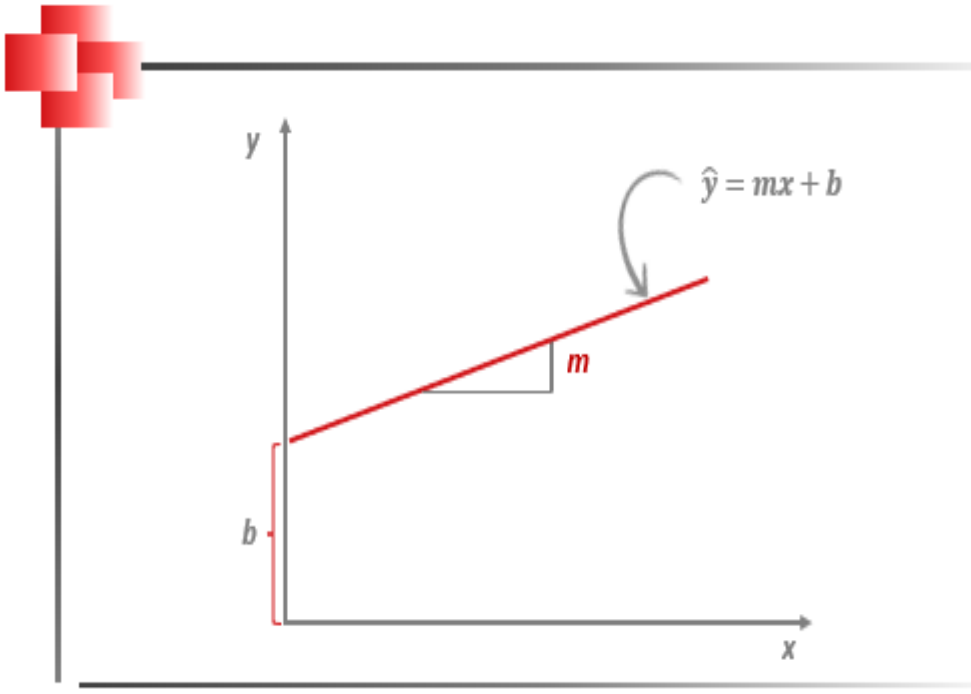
$$b = \bar{y} - m\bar{x} \qquad 4.6$$

S_{xy} es la covarianza, S_x^2 la varianza para la variable independiente; $\bar{y}$ es la variable estimada, x la variable explicativa, m la pendiente de la recta, y b la ordenada en el origen, Galton (1886).

Obsérvese que, para calcular la ordenada en el origen (b), se deben haber calculado previamente la pendiente, m, y las medias de la variable independiente, $\bar{x}$ y de la variable dependiente, $\bar{y}$.

En la Figura 115, se presenta un gráfico en el que se identifican los elementos de la recta de regresión.

Figura 115
Elementos de la recta de regresión



Análisis de Correlación

El análisis de correlación es una técnica estadística que se utiliza para evaluar la relación entre dos variables cuantitativas continuas. Permite medir la fuerza y la dirección de la relación entre las variables, lo que puede ayudar a identificar patrones y tendencias en los datos. En general, el análisis de correlación se utiliza para determinar si existe una relación significativa entre las variables, y si es así, cuál es su naturaleza y magnitud.

Es importante destacar que el análisis de correlación no implica necesariamente una relación causal entre las variables, sino simplemente una asociación o relación entre ellas. Por lo tanto, se debe tener en cuenta otros factores y variables que puedan influir en los resultados de la correlación.

Coeficiente de Correlación Lineal de Pearson

Para realizar el análisis de correlación correspondiente a una regresión lineal simple, se calcula el coeficiente de correlación (Galton, 1888):

$$r = \frac{S_{xy}}{\sqrt{S_x{}^2}\sqrt{S_y{}^2}} \qquad 4.7$$

El valor de este coeficiente, debe estar comprendido entre -1 y 1, ya que la covarianza siempre es menor o igual que el producto de las desviaciones típicas. Este coeficiente es el encargado de medir la fuerza con que se podrían relacionar dos variables si se realizara un ajuste lineal.

Interpretación del Coeficiente de Correlación

1. Si r toma valores cercanos a -1, la correlación es fuerte e inversa.
2. Si r toma valores cercanos a 1, la correlación es fuerte y directa.
3. Si r toma valores cercanos a cero, la correlación es débil.
4. Si r = 1 o r = -1, la correlación es perfecta, es decir, hay dependencia funcional.

Los valores específicos del coeficiente de correlación que se emplean como medida de la fuerza de asociación (tamaño del efecto), Cohen (1988). son los siguientes:

0: asociación nula

0.1: asociación pequeña

0.3: asociación mediana

0.5: asociación moderada

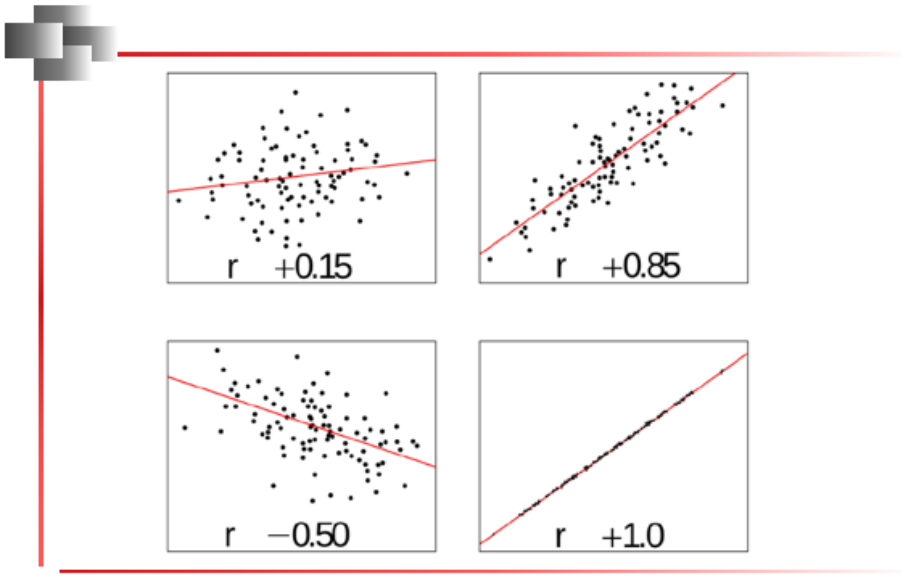
0.7: asociación alta

0.9: asociación muy alta

Es necesario acotar que los valores anteriores también pueden ser negativos, en cuyo caso se tendría el mismo tamaño del efecto, pero el signo negativo estaría indicando que la relación entre variables es inversamente proporcional, es decir que mientras los valores de una de las variables aumentan los de la otra disminuyen y viceversa. En las figuras 116 y 117 se presentan algunos ejemplos de los diagramas de dispersión correspondientes a diferentes valores del coeficiente de correlación.

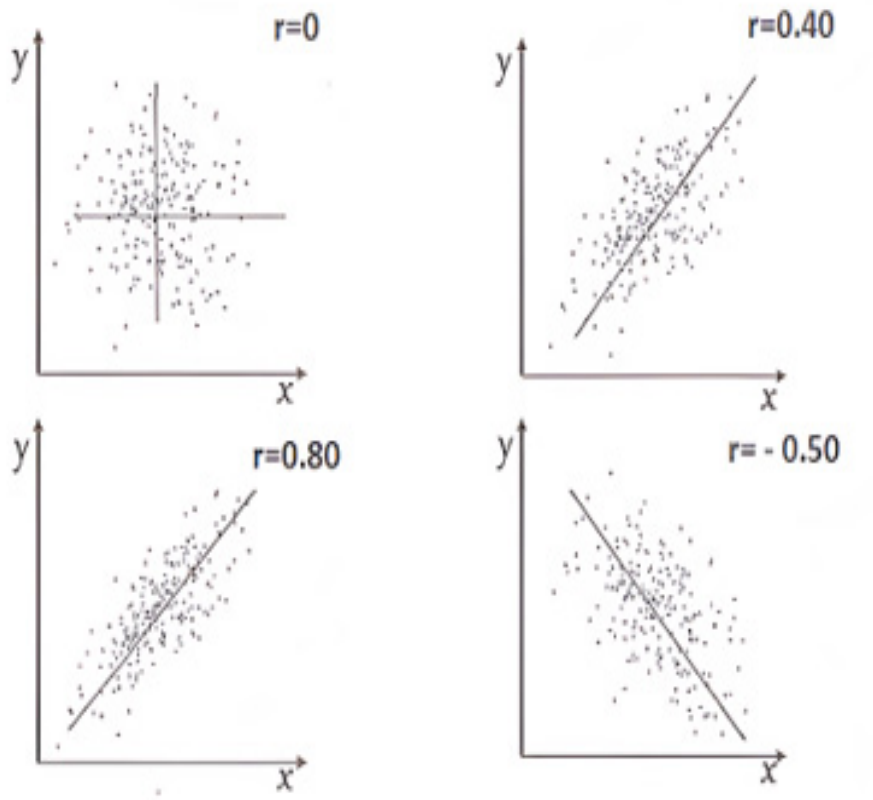
Figura 116

Ejemplos de ajustes para diferentes valores del coeficiente de correlación



Nota. Adaptado de Coeficiente de correlación en R de Vivaelsoftwarelibre, 2018, VivaelsoftwareLibre (t.ly/KDly). CC BY-SA 4.0.

Figura 117
Ejemplos de recta de regresión y sus coeficientes de correlación correspondientes



Coeficiente de Determinación

Es el encargado de cuantificar la bondad del ajuste del modelo (Fisher, 1919). En el caso de un ajuste planteado a través de un modelo lineal, se calcula a través de la ecuación:

$$r^2 = \frac{(S_{xy})^2}{S_x{}^2 S_y{}^2} \qquad 4.8$$

Este coeficiente indica qué proporción de la varianza total de un conjunto de datos puede ser explicada por el modelo, es por ello que se suele expresar en porcentaje lo cual se logra multiplicando su ecuación de cálculo por 100.

Cuanto más cerca de 100% esté el coeficiente de determinación, más por-

centaje de la variable dependiente es explicada por el ajuste realizado a través de la recta de regresión simple.

Dado que el coeficiente de determinación, para el caso de un modelo lineal es igual al cuadrado del coeficiente de correlación, está claro que el valor mínimo que puede tomar es cero y el máximo uno:

4.9

$$0 \leq r^2 \leq 1$$

Error de Predicción

Cuando se realiza un ajuste, de un conjunto de datos por medio de una regresión lineal simple, va a existir un error en cada una de las predicciones, puesto que la mayoría de las veces existirá una diferencia entre la predicción del modelo y el valor de la variable dependiente medida. Lo anterior, se puede expresar de la siguiente manera:

4.10

$$y_i = mx_i + b + \varepsilon_i$$

donde es el error cometido cuando se aproxima el valor de la variable dependiente, correspondiente a la i-ésima medición, a través del valor de obtenido a través de la ecuación de regresión lineal simple (Legendre, 1805).

A este error, se le suele denominar error de predicción o residual y está definido como la diferencia entre el verdadero valor de la variable dependiente , y el valor de su predicción según la ecuación de regresión, , es decir:

4.11

$$\varepsilon_i = y_i - \widehat{y_i}$$

Ejemplos de Ajuste de Datos a un Modelo Lineal

En la Figura 118 se desarrolla un ejercicio de regresión lineal simple cuyo enunciado ha sido tomado de Berrendero (s.f.). En la Figura 119 y 120 se incluyen otros ejemplos.

Figura 118

Primer ejemplo de regresión lineal simple

Los grillos son ectotermos, por lo que sus procesos fisiológicos y su metabolismo están influidos por la temperatura. Con el fin de estudiar estas cuestiones se ha medido el número de vibraciones por segundo de las alas de un grupo de grillos a varias temperaturas.

Hallar la predicción de y_o para $x_o = 80$

Recta de ajuste
$y = mx + b = 0.216\,x - 0.615$
Para $x = 80$
$\hat{y} = 0.216\,(80) - 0.615 = 16.68$

Vibraciones por segundo	Temperatura °F
20.0	88.6
16.0	71.6
19.8	93.3
18.4	84.3
17.1	80.6
15.5	75.2
14.7	69.7
17.1	82.0
15.4	69.4
16.2	83.3
15.0	78.6
17.2	82.6
16.0	80.6
17.0	83.5
14.1	76.3

x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
88.6	20	7849.96	1772	400
71.6	16	5126.56	1145.6	256
93.3	19.8	8704.89	1847.34	392.04
84.3	18.4	7106.49	1551.12	338.56
80.6	17.1	6496.36	1378.26	292.41
75.2	15.5	5655.04	1165.6	240.25
69.7	14.7	4858.09	1024.59	216.09
82	17.1	6724	1402.2	292.41
69.4	15.4	4816.36	1068.76	237.16
83.3	16.2	6938.89	1349.46	262.44
78.6	15	6177.96	1179	225
82.6	17.2	6822.76	1420.72	295.84
80.6	16	6496.36	1289.6	256
83.5	17	6972.25	1419.5	289
76.3	14.1	5821.69	1075.83	198.81
$\bar{x}$	$\bar{y}$	$\sum x_i^2$	$\sum x_i \cdot y_i$	$\sum y_i^2$
79.973	16.633	96567.66	20089.58	4192.01

Recta de ajuste

$$y = mx + b = 0.216\,x - 0.615$$

Covarianza

$$S_{xy} = \frac{1}{n}\sum x_i . y_i - \bar{x}\,\bar{y} = 9.082$$

Varianza para la variable independiente

$$S_x^2 = \frac{1}{n}\sum x_i^2 - (\bar{x})^2 = 42.110$$

Varianza para la variable dependiente

$$S_y^2 = \frac{1}{n}\sum y_i^2 - (\bar{y})^2 = 2.800$$

Pendiente

$$m = S_{xy}/{S_x}^2 = 0.216$$

Ordenada en el origen

$$b = \bar{y} - m\,\bar{x} = -0.615$$

Coeficiente de regresión / determinación

$$r = \frac{S_{xy}}{\sqrt{{S_x}^2}\sqrt{{S_y}^2}} = 0.836 \quad r^2 = 0.700$$

Figura 119

Segundo ejemplo de regresión lineal simple

Se muestran los datos de las calificaciones acumuladas de los estudiantes en un bimestre y las notas obtenidas en el examen parcial. Realizar un ajuste de regresión simple y encontrar el error de predicción para una calificación acumulada de 8.5.

Nota acumulada	Nota Examen
6.5	4.6
8.1	7.4
10	9
9.1	7.3
8.3	6.8
9.8	7.9
4.5	1.8
6.2	6
7	7.3
8.9	8.1
9.8	9
8.9	7.6
6.5	5.8
8.5	7.8
8.6	7.2
9.4	8.1
10	10
7.4	6.1
10	9.8
9.2	8.9

Recta de ajuste

$y = mx + b = 0.887\,x - 0.066$

Para $x = 8.5$

$\hat{y} = 0.887\,(8.5) - 0.066 = 7.5$

Error de predicción

$\varepsilon = y_i - \hat{y}_i = 7.8 - 7.5 = 0.3$

x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
6.5	4.6	42.25	29.9	21.16
8.1	7.4	65.61	59.94	54.76
10	9	100	90	81
9.1	7.3	82.81	66.43	53.29
8.3	6.8	68.89	56.44	46.24
9.8	7.9	96.04	77.42	62.41
4.5	1.8	20.25	8.1	3.24
6.2	6	38.44	37.2	36
7	7.3	49	51.1	53.29
8.9	8.1	79.21	72.09	65.61
9.8	9	96.04	88.2	81
8.9	7.6	79.21	67.64	57.76
6.5	5.8	42.25	37.7	33.64
8.5	7.8	72.25	66.3	60.84
8.6	7.2	73.96	61.92	51.84
9.4	8.1	88.36	76.14	65.61
10	10	100	100	100
7.4	6.1	54.76	45.14	37.21
10	9.8	100	98	96.04
9.2	8.9	84.64	81.88	79.21
$\bar{x}$	$\bar{y}$	$\sum x_i^2$	$\sum x_i \cdot y_i$	$\sum y_i^2$
8.335	7.325	1433.97	1271.54	1140.15

Covarianza

$$S_{xy} = \frac{1}{n}\sum x_i . y_i - \bar{x}\,\bar{y} = 63.577$$

Varianza para la variable independiente

$$S_x^2 = \frac{1}{n}\sum {x_i}^2 - (\bar{x})^2 = 71.699$$

Varianza para la variable dependiente

$$S_y^2 = \frac{1}{n}\sum {y_i}^2 - (\bar{y})^2 = 57.008$$

Pendiente

$$m = S_{xy}/{S_x}^2 = 0.887$$

Ordenada en el origen

$$b = \bar{y} - m\,\bar{x} = -0.066$$

Coeficiente de regresión / determinación

$$r = \frac{S_{xy}}{\sqrt{{S_x}^2}\sqrt{{S_y}^2}} = 0.994 \quad r^2 = 0.989$$

Figura 120

Tercer ejemplo de regresión lineal simple

Se miden las estaturas de un grupo de niños y adolescentes con edades comprendidas entre 6 y 14 años. Realizar un ajuste de regresión lineal y predecir cual será la estatura de un niño de 8 años.

Recta de ajuste

$$y = mx + b = 12.413\,x + 3.619$$

Para $x = 11$

$$\hat{y} = 12.413\,x + 3.619 = 140.2$$

Edad	Estatura
6	112
7	116
8	120
9	128
9	125
10	131
10	128
11	135
11	137
12	141
12	143
13	147
13	148
14	157
14	160

x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
6	112	36	672	12544
7	116	49	812	13456
8	120	64	960	14400
9	128	81	1152	16384
9	125	81	1125	15625
10	131	100	1310	17161
10	128	100	1280	16384
11	135	121	1485	18225
11	137	121	1507	18769
12	141	144	1692	19881
12	143	144	1716	20449
13	147	169	1911	21609
13	148	169	1924	21904
14	157	196	2198	24649
14	160	196	2240	25600
$\bar{x}$	$\bar{y}$	$\sum x_i^2$	$\sum x_i \cdot y_i$	$\sum y_i^2$
10.6	135.2	1771	21984	277040

Recta de ajuste

$$y = mx + b = 12.413\,x + 3.619$$

Covarianza

$$S_{xy} = \frac{1}{n}\sum x_i . y_i - \bar{x}\,\bar{y} = 1465.7$$

Varianza para la variable independiente

$$S_x^2 = \frac{1}{n}\sum x_i^2 - (\bar{x})^2 = 118.067$$

Varianza para la variable dependiente

$$S_y^2 = \frac{1}{n}\sum y_i^2 - (\bar{y})^2 = 18469.333$$

Pendiente

$$m = S_{xy}/S_x^2 = 12.413$$

Ordenada en el origen

$$b = \bar{y} - m\,\bar{x} = 3.619$$

Coeficiente de regresión / determinación

$$r = \frac{S_{xy}}{\sqrt{S_x^2}\sqrt{S_y^2}} = 0.993 \quad r^2 = 0.985$$

Regresión Simple con Datos Agrupados

Cuando se tienen datos agrupados se deben ajustar las ecuaciones para la inclusión de la frecuencia absoluta, ésta representa el número de veces que se repite cada par (x, y), Pearson(1900).

Covarianza

$$S_{xy} = \frac{1}{n}\sum x_i . y_i f_i - \bar{x}\,\bar{y} \qquad 4.12$$

Varianza de la Variable Independiente

$$S_x^2 = \frac{1}{n}\sum x_i{}^2 f_i - (\bar{x})^2 \qquad 4.13$$

Varianza de la Variable Independiente

$$S_y^2 = \frac{1}{n}\sum y_i{}^2 f_i - (\bar{y})^2 \qquad 4.14$$

Coeficiente de Correlación Lineal

$$r = \frac{S_{xy}}{\sqrt{S_x{}^2}\sqrt{S_y{}^2}} \qquad 4.15$$

Coeficiente de Determinación

$$r^2 = \frac{(S_{xy})^2}{S_x{}^2 S_y{}^2} \qquad 4.16$$

Coeficiente de la Regresión

$$\hat{y} = mx + b \qquad 4.17$$

Donde

$$m = \frac{S_{xy}}{S_x{}^2} \qquad 4.18$$

y

$$b = \bar{y} - m\bar{x} \qquad 4.19$$

En la Figura 121 se presenta un ejemplo de cálculo de la recta de regresión simple para datos agrupados.

Figura 121

Ejemplo de regresión lineal simple para datos agrupados

Se desea realizar un ajuste de regresión lineal simple entre las variables hábitos de estudio (x) y calificación de matemáticas, de un grupo de estudiantes de educación media. Los datos y sus respectivas frecuencias absolutas se presentan en la tabla adjunta.

Determinar la fuerza de asociación entre variables.

x	y	f_i
25	45	4
25	35	7
25	25	8
35	75	3
35	55	6
35	45	14
35	35	15
45	75	2
45	65	4
45	55	16
45	45	19
45	35	6
55	75	2
55	65	5
55	15	2

Recta de ajuste

$$y = mx + b = 0.751\,x + 16.639$$

$$r^2 = 0.237$$

La fuerza de asociación es mediana

x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
6	112	36	672	12544
7	116	49	812	13456
8	120	64	960	14400
9	128	81	1152	16384
9	125	81	1125	15625
10	131	100	1310	17161
10	128	100	1280	16384
11	135	121	1485	18225
11	137	121	1507	18769
12	141	144	1692	19881
12	143	144	1716	20449
13	147	169	1911	21609
13	148	169	1924	21904
14	157	196	2198	24649
14	160	196	2240	25600
$\bar{x}$	$\bar{y}$	$\sum x_i^2$	$\sum x_i \cdot y_i$	$\sum y_i^2$
39.07	45.97	180825	209225	258625

Recta de ajuste

$$y = mx + b = 0.751\,x + 16.639$$

Covarianza

$$S_{xy} = \frac{1}{n}\sum x_i . y_i - \bar{x}\,\bar{y} = 55.329$$

Varianza para la variable independiente

$$S_x^2 = \frac{1}{n}\sum x_i{}^2 - (\bar{x})^2 = 73.694$$

Varianza para la variable dependiente

$$S_y^2 = \frac{1}{n}\sum y_i{}^2 - (\bar{y})^2 = 175.159$$

Pendiente

$$m = S_{xy}/S_x{}^2 = 0.751$$

Ordenada en el origen

$$b = \bar{y} - m\,\bar{x} = 16.639$$

Coeficiente de regresión / determinación

$$r = \frac{S_{xy}}{\sqrt{S_x{}^2}\sqrt{S_y{}^2}} = 0.487 \quad r^2 = 0.237$$

La fuerza de asociación entre variables es mediana

Referencias bibliográficas

Abreu León, J. L., Oliveró Serrat, M., Escamilla González, O., Espinosa Longi, J. (2017). DescartesJS [Software en línea]. https://n9.cl/d3n3.

Altmann, G. (25 de mayo de 2020). Letras ABC [Imagen]. Pixabay. https://n9.cl/p3xzi.

Bacchini, R. D. Vásquez, L. V., Vianco, M. J. y García Fronti, J. I. (2018). Introducción a la probabilidad y la estadística (1a. ed.). Universidad de Buenos Aires.

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. Richard Taylor and John Adamson.

Berrendero, J. R. (s.f.). Tema 3. Modelo de regresión simple [Diapositivas de PowerPoint]. Universidad Autónoma de Madrid. https://n9.cl/46ozi.

Bertin, J. (1967). La graphique et le traitement graphique de l'information. Flammarion.

Boole, G. (1854). The Calculus of Probabilities. Macmillan and Company.

Brown, J. R. (1 de julio de 2010). Medidas de dispersión [Gráfico]. Wikipedia. https://n9.cl/m7hc4.

Calculadorasonline. (30 de enero de 2022). Binario árbol [Imagen]. Calculadoras de matemática https://n9.cl/ah31l.

Calculator Online. (2021). Interpolar calculadora. Calculator Online. t.ly/TJlW.

Casella, G., & Berger, R. L. (2020). Statistical Inference (3rd ed.). Cengage Learning.

Casella, G. (2021). Statistics and probability: two sides of the same coin. International Statistical Review, 89(1), 1-12.

Castillo Manrique, I. y Guijarro Garvi, M. (2006). Estadística descriptiva y cálculo de probabilidades. Pearson, Prentice Hall.

Canavos, G. C. (1988). Probabilidad y estadística. Aplicaciones y métodos. McGraw-Hill.

Casella, G., & Berger, R. L. (2020). Statistical Inference (3rd ed.). Cengage Learning.

Cobo, E., Kostov, B., Cortés, J., González, J. A. y Muñoz, P. (Septiembre 2014). Siete posibles medias muestrales y sus respectivos ICs [Gráfico]. Bioestadística para no estadísticos. https://n9.cl/0n7zd1.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

DeGroot, M. H., & Schervish, M. J. (2011). Probability and Statistics. 4th ed. Pearson.

De los Santos, S. (s.f.). Tablas de distribución F [Tabla en línea]. El Osio De los Santos. t.ly/C6qM

Devore, J. L. (2012). Probabilidad y estadística para ingeniería y ciencias (8va ed.). Cengage Learning.

De Moivre, A. (1733). The Doctrine of Chances: A Method of Calculating the Probability of Events in Play. W. Pearson.

Dodge, Y. (2003). The Oxford dictionary of statistical terms. Oxford University Press.

Domínguez, J. y Dominguez, A. (2006-2020). CalEst (Versión 4.4) [Software en línea]. Conteck. https://n9.cl/354tn.

Freedman, D., Pisani, R., & Purves, R. (2007). Statistics. 4th ed. W. W. Norton & Company.

Feller, W. (1950). Introduction to Probability Theory and Its Applications. Vol. 1. John Wiley & Sons.

Fisher, R. A. (1919). The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh, 52, 399-433.

Fisher, R.A. (1925). Statistical Methods for Research Workers. Oliver and Boyd.

Freund, J. E., & Simon, G. A. (2011). Estadística matemática con aplicaciones. Pearson Educación.

Fuensanta, A. (16 de noviembre de 2014). IC cociente varianzas distr. Normales: cuantiles [Widget]. Geogebra. t.ly/gC0l.

Fundación Pediatría y Salud. (2009). Patrones de crecimiento infantil de la OMS [Gráfico]. Asociación Española de Pediatría de Atención Primaria. https://n9.cl/tqn5i.

Galton, F. (1883). Inquiries into Human Faculty and its Development. Macmillan.

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

Galton, F. (1888). Co-relations and their Measurement, Chiefly from Anthropometric Data. Proceedings of the Royal Society of London, 45, 135-145.

García Cebrián, M. J. (2001). Tabla N(0,1). Red educativa digital Descartes. https://n9.cl/epoz8.

García Cebrián, M.J. (2017). Inferencia estadística. Red educativa digital Descartes. https://n9.cl/d3n3.

Gauss, C. F. (1809). Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Perthes et Besser.

Giselle. (13 de septiembre de 2015). Intervalo de confianza [Gráfico]. Geogebra. https://n9.cl/casbw.

Giordano, F., & Kass, R. E. (2021). A Primer in Probability and Statistics (2nd ed.). CRC Press.

Gerd Altmann (25 de mayo de 2020). Letras A B C [Imagen]. Pixabay. https://n9.cl/qoppe.

Hassan, M. (30 de octubre de 2018). 1455379 [Foto]. Pxhere https://n9.cl/u02gs.

Hogg, R. V., McKean, J. W., & Craig, A. T. (2021). Introduction to Mathematical Statistics (8th ed.). Pearson.

Jeffreys, H. (1939). An Introduction to the Theory of Probability. Oxford University Press.

Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis (6th ed.). Pearson Education.

Kanijoman. (10 de febrero de 2012). Distribución normal [Imagen]. Flickr. https://n9.cl/mce5y.

Kolmogorov, A. (1950). Foundations of the theory of probability. Chelsea Publishing Company.

Kramp, C. (1808). Eléments d'arithmétique universelle. Mme veuve Courcier.

Laplace, P. S. (1812). Recherches sur la probabilité des jugements en matière criminelle et en matière civile. Bachelier.

Laplace, P. S. (1814). Essai philosophique sur les probabilités. Courcier.

Legendre, A. M. (1805). Nouvelles méthodes pour la détermination des orbites des comètes. Firmin Didot.

Lind, D. A., Marchal, W. G., y Wathen, S. A. (2012). Estadística aplicada a los negocios y la economía (15ª. ed.). Mc Graw Hill.

Lipschutz, S. (1965). Teoría y problemas de probabilidad. McGraw-Hill.

Medina, L. (28 de mayo de 2014). Cálculo de probabilidad distribución normal [Widget en línea]. t.ly/gIFH.

Mendenhall, W., Beaver, R. J. y Beaver, B. M. (2010). Introducción a la probabilidad y estadística. Cengage Learning.

Molina, J. G. y Rodrigo, M. F. (2010). La estadística inferencial: Algunos conceptos previos. Universidad de Valencia.

Montgomery, D. C., & Runger, G. C. (2022). Applied Statistics and Probability for Engineers (8th ed.). Wiley.

Moore, D. S., & McCabe, G. P. (2017). Introduction to the Practice of Statistics (9th ed.). W. H. Freeman.

Nolberto Sifuentes, V. A. y Ponce Aruneri, M. E. (2008). Estadística Inferencial Aplicada. Universidad de Post Grado de la Facultad de Educación de la Universidad Nacional Mayor de San Marcos.

Pascal, B. (1665). Traité du triangle arithmétique. Chez Florentin Lambert.

Pearson, K. (1892). The Grammar of Science. Adam and Charles Black.

Pearson, K. (1895). Contributions to the Mathematical Theory of Evolution. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 186, 71-110.

Pearson, K. (1900). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it Can be Reasonably Supposed to have Arisen from Random Sampling. Philosophical Magazine, 50(302), 157-175.

Pedro José. (18 de noviembre de 2015). Intervalo de confianza [Gráfico]. Geogebra. https://n9.cl/1p7up.

Playfair, W. (1786). The Commercial and Political Atlas: Representing, by Means of Stained Copper-plate Charts, the Progress of the Commerce, Revenues, Expenditure and Debts of England, During the Whole of the Eighteenth Century. T. Burton.

Playfair, W. (1801). Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe. J. Wallis.

Pixabay. (15 de abril de 2012). Binario árbol [Imagen], Pixabay https://n9.cl/h2we0

Pixabay. (15 de octubre de 2013). Dado [Imagen]. Pixabay. https://n9.cl/1e2vu.

Pixabay. (19 de junio de 2014). Mujer Hombre [Pictograma]. Pixabay. https://n9.cl/xbmjk.

Pixabay. (3 de agosto de 2014). Conos de helado postre vainilla chocolate fresa [Imagen]. Pixabay. https://n9.cl/k9z62.

Pixabay. (15 de junio de 2016). Animales Mascotas [Pictograma]. Pixabay. https://n9.cl/1wlhe.

Pixnio (s.f.). Monedas de metal [Imagen]. Pixnio. https://n9.cl/x1fya.

Pngkey. (s.f.,). Vector Illustration of Decision Making Hand Flipping [Imagen]. Pngkey. https://n9.cl/cb7fu.

Publicdomainvectors.org. (11 de diciembre de 2013). Group of People Sitting and Reading Books [Imagen]. Publicdomainvectors.org. https://n9.cl/5mvnk.

Publicdomainvectors.org. (18 de diciembre de 2017). Tres adolescentes de la historieta [Imagen]. Publicdomainvectors.org. https://n9.cl/w4ksz.

Publicdomainvectors.org. (2 de julio de 2018). Niños japoneses alrededor de una mesa [Imagen]., Publicdomainvectors.org. https://n9.cl/tkaqk.

Rivera Berrío, J.G. (s.f). Chi-Square Cálculo Chi-cuadrado crítico [Aplicación en línea]. Geogebra. t.ly/wTto.

Rossman, A. J., & Chance, B. L. (2018). Workshop Statistics: Discovery with Data (5th ed.). Wiley.

Santa María, C. R. y Buccino, C. S. (2019). Elementos de probabilidad y estadística (2ª. ed.). Universidad Nacional de Moreno.

Scott, D. W., & Scott, L. R. (1992). Histograms and the Frequency Polygon. The American Statistician, 46(2), 123-129. https://doi.org/10.1080/00031305.1992.10475843

Spiegel, M. R. y Stephens, L. J. (2009). Estadística (4ta. Ed.). Shaum.

Sturges, H. A. (1926). The choice of a class interval. Journal of the American Statistical Association, 21(153), 65-66.

Tufte, E. R. (1997). Visual and Statistical Thinking: Displays of Evidence for Making Decision. Graphics Press.

Thompson, D. (2009). Hypothesis testing and Statistical Power. The University of Oklahoma. https://n9.cl/lksxq.

Triola, M.F. (2008). Estadística (10ma. ed.). Pearson-Addison Wesley.

Triola, M. F. (2018). Elementary Statistics (13th ed.). Pearson.

Venn, J. (1887). On the Diagrammatic and Mechanical Representation of Propositions and Reasonings. Philosophical Magazine and Journal of Science, 5(4), 161-176.

Vivaelsoftwarelibre. (2018). Coeficiente de correlación en R [Imagen] Vivaelsoftwarelibre. https://n9.cl/0fvi1.

Walpole, R. E., Myers, R. H., Myers, S. L. y Ye, K. (2012). Probabilidad y estadística para ingeniería y ciencias (9na. ed.). Pearson.

Wasserman, L. (2021). All of Statistics: A Concise Course in Statistical Inference. Springer.

Wilhelmi, M. R. (2004). Combinatoria y probabilidad. Departamento de Didáctica de la Matemática, Universidad de Granada.

Zaheer, Ch. (25 de agosto de 2016). Tornillo autoperforante de cabeza hexagonal [Imagen], Pixabay. https://n9.cl/7mz4s.

Zylberberg, A. D. (2005). Probabilidad y estadística. Nueva Librería.

IBARRA - ECUADOR

Vive
sueña
construye

ISBN: 978-9942-845-38-2