|  |  |
| --- | --- |
| Título del guion | Estadística |
| Código del guion | GUIÓN\_MA\_10\_11\_CO |
| Descripción | Los problemas, herramientas y métodos de la estadística permiten tomar decisiones ante situaciones de incertidumbre, de manera que sirve de ayuda a cada individuo para producir y comunicar una mejor comprensión de la situación en un contexto particular |

[SECCIÓN 1] **1 La estadística descripctiva**

La estadística descriptiva es una rama de la estadística que se dedica a recolectar, sistematizar, ordenar, describir y analizar algunas características de una población, basándose en un conjunto de datos. Algunos ejemplos de esta rama son cuando se diseñan, aplican y analizan encuestas de preferencia de votos o cuando una empresa hace un estudio de los gustos y las necesidades para el lanzamiento de un nuevo producto. Por otro lado la estadística inferencial busca construir conclusiones generales para toda la población a partir del estudio de una muestra.

En ambos casos la estadística permite la construcción de parámetros que proporcionen una idea global cuantitativa de la población observada y, a partir de estos, se puede establecer conclusiones y tomar decisiones ante situaciones de incertidumbre.

La agrupación de datos es una herramienta de gran utilidad cuando se tiene una gran cantidad de datos. El proceso se denomina *distribución de frecuencias* dado que no solo se distribuyen los datos, sino a que tal distribución se realiza en subgrupos o clases o intervalos. Así, el total de datos aparece distribuido según sus repeticiones o *frecuencias* en tales subgrupos.

[SECCIÓN 2] **1.1 Las variables cualitativas**

Una vez se tiene un conjunto de datos éstos pueden ser sistematizados de diferentes formas, de acuerdo a sus características. Las *variables* son las diferentes características o aspectos que podemos estudiar en los individuos de una población y son necesarias para analizarla y tener alternativas de decisión.

Existen diferentes tipos de variables, las cuales pueden ser cualitativas o cuantitativas. Como su nombre lo indica, las variables cualitativas indican características no medibles o numerables de los individuos en una población, así que toman valores no numéricos. Ejemplo de tales variables son los colores del iris de los pobladores de una región, el género de las películas preferidas, el sector de residencia o el tema principal en los libros preferidos.

[SECCIÓN 2] **1.2 Las variables cuantitativas discretas**

Por su parte las variables cuantitativas recopilan atributos medibles de los individuos en una población, como pueden ser la edad, la estatura, la cantidad de hermanos, el tiempo de estudio o participación en redes sociales semanal, etc. En general, las variables cuantitativas toman valores numéricos y, debido a eso, se subclasifican en dos: discretas y continuas.

Las variables cuantitativas discretas son aquellas que toman valores del conjunto de los números naturales, como por ejemplo la cantidad de celulares que tenemos, la cantidad de ejemplares de nuestra colección favorita o la cantidad de hermanos.

[SECCIÓN 2] **1.3 Las variables cuantitativas continuas**

Entre las variables cuantitativas, se denominan variables cuantitativas continuas a aquellas que pueden tomar valores del conjunto de los números reales, es decir que se admiten todos los valores en un intervalo. Ejemplo de ellas son el peso de una persona, la estatura, el tiempo que dedica a actividades deportivas, etc.

Si los elementos objeto del estudio estadístico toman muchos valores diferentes, se agrupan formando intervalos de clase. Cada clase está delimitada por un valor inferior y otro superior, los cuales son llamados límites de la clase. Por ejemplo, si tenemos valores de un grupo de 237 datos que están entre 1 y 100, se pueden definir grupos para sintetizar la información, por medio de los intervalos: 1-25, 26-50, 51-75, 76-100. De esta forma, resumimos los datos en cuatro grupos, entre los que se distribuirán los 237 datos. En el ejemplo, los límites inferiores de las clases fueron 1, 26, 51 y 76, mientras que los límites superiores fueron 25, 50, 75 y 100.

Por su parte, la *marca de clase* es el punto medio de cada intervalo y es el valor que se utiliza para el cálculo de parámetros en contextos de variables cuantitativas, por lo cual será el valor que representará todo el intervalo de clase. Las marcas de clase de las 4 clases del ejemplo son 13, 38, 63 y 88.

Entre las formas de condensación de la información estadística, una de las más potentes y útiles para presentar información clara y concisa es la distribución de frecuencias. La tabla de frecuencias está formada por columnas en las cuales se pueden incluir intervalos de clase, frecuencias absolutas, frecuencias relativas, frecuencias absolutas acumuladas, frecuencias relativas acumuladas, marcas de clase y porcentajes. La aparición de tales columnas depende principalmente del tipo de variable estudiada.

Específicamente, los elementos de la tabla de frecuencias son:

* Los *intervalos de clase* muestran las agrupaciones o clases en que se van a distribuir los datos. El número *i* indica la cantidad total de clases, mientras que el número *n* indica la cantidad total de datos.
* La *frecuencia absoluta* (*fi*) es el número total de veces que se repiten los datos en cada clase.
* La *frecuencia relativa* (*hi*) es el cociente entre la frecuencia absoluta (*f*) en cada intervalo y el total de datos (*n*).
* La *frecuencia absoluta acumulada* (*F*) en cada clase es la suma de las frecuencias absolutas (*f*) de los valores menores o iguales a él.
* La *frecuencia relativa* acumulada (*H*) es el cociente entre la frecuencia relativa (*hi*) en cada intervalo y el total de datos (*n*).
* La *marca de clase* (*ci*) es el punto medio de cada intervalo y .solo tiene sentido para variables estadísticas cuantitativas

Por ejemplo, para obtener información acerca de los hábitos alimenticios de los estudiantes, se preguntó a cada uno de ellos por el tipo de alimento que consumía principalmente cada día. Por supuesto, se trata de una variable de tipo cualitativo.

Tomando las opciones de grupos nutricionales definidos por las sociedades nutricionales, las opciones para elegir el principal alimento consumido eran:

* Carne, pollo, huevos y frutos secos
* Pan, cereales, arroz y pasta
* Frutas y verduras
* Aceites, grasas y dulces
* Suplementos alimentarios
* Leche, yogurt y queso

Los resultados para los 36 estudiantes consultados fueron los siguientes:

|  |
| --- |
| Carne, pollo, huevos y frutos secos |
| Aceites, grasas y dulces |
| Pan, cereales, arroz y pasta |
| Pan, cereales, arroz y pasta |
| Frutas y verduras |
| Pan, cereales, arroz y pasta |
| Frutas y verduras |
| Carne, pollo, huevos y frutos secos |
| Leche, yogurt y queso |
| Carne, pollo, huevos y frutos secos |
| Frutas y verduras |
| Carne, pollo, huevos y frutos secos |
| Carne, pollo, huevos y frutos secos |
| Leche, yogurt y queso |
| Pan, cereales, arroz y pasta |
| Pan, cereales, arroz y pasta |
| Suplementos |
| Pan, cereales, arroz y pasta |
| Pan, cereales, arroz y pasta |
| Frutas y verduras |
| Aceites, grasas y dulces |
| Leche, yogurt y queso |
| Carne, pollo, huevos y frutos secos |
| Carne, pollo, huevos y frutos secos |
| Frutas y verduras |
| Suplementos |
| Pan, cereales, arroz y pasta |
| Aceites, grasas y dulces |
| Leche, yogurt y queso |
| Pan, cereales, arroz y pasta |
| Carne, pollo, huevos y frutos secos |
| Leche, yogurt y queso |
| Suplementos |
| Leche, yogurt y queso |
| Leche, yogurt y queso |
| Frutas y verduras |

Como se observa, el registro de las respuestas no permite de manera rápida ni precisa identificar el comportamiento nutricional de la población estudiada. Sin embargo, al organizar los datos en su tabla de frecuencias se obtiene:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Principal alimento consumido**  ***xi*** | **Frecuencia absoluta *(fi)*** | **Frecuencia relativa**  ***(hi= fi/n)*** | **Frecuencia absoluta acumulada *(Fi)*** | **Frecuencia relativa acumulada *(Hi= Fi/n )*** | **Porcentaje (%)**  **(*hi* escrita como porcentaje)** |
| Carne, pollo, huevos y frutos secos | 8 | 2/9 | 8 | 2/9 | 22.22% |
| Pan, cereales, arroz y pasta | 9 | 1/4 | 17 | 17/36 | 25.00% |
| Frutas y verduras | 6 | 1/6 | 23 | 23/36 | 16.67% |
| Aceites, grasas y dulces | 3 | 1/12 | 26 | 13/18 | 8.33% |
| Suplementos alimentarios | 3 | 1/12 | 29 | 29/36 | 8.33% |
| Leche, yogurt y queso | 7 | 7/36 | 36 | 1 | 19.44% |
| **TOTAL** | **36** | **1** | **1** |  | **100.00%** |

De esta manera se identifica por ejemplo que la cuarta parte de los estudiantes consume principalmente carbohidratos, y que la misma cantidad de estudiantes consume principalmente grasas o suplementos.

Por su parte, como segundo ejemplo, en una clase de 25 alumnos se ha realizado una encuesta sobre la máxima distancia a la que cada estudiante ha viajado, medida entre su actual lugar de residencia y la capital de un país extranjero adonde han viajado. En este caso se trata de una variable de tipo cuantitativo, por lo que, en este caso, la tabla contiene marcas de clase. Los resultados fueron agrupados en cinco intervalos de clase según la cantidad de kilómetros recorrida.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cantidad de kilómetros recorrida**  ***xi*** | **Marca de clase**  ***(ci)*** | **Frecuencia absoluta *(fi)*** | **Frecuencia relativa**  ***(hi= fi/n)*** | **Frecuencia absoluta acumulada *(Fi)*** | **Frecuencia relativa acumulada *(Hi= Fi/n )*** | **Porcentaje (%)**  **(*hi* escrita como porcentaje)** |
| [0, 2.000) | 1.000 | 14 | 14 / 25 = 0,56 | 14 | 14 / 25 | 56 % |
| [2.000, 4.000) | 3.000 | 5 | 5 / 25 = 0,2 | 14 + 5 = 19 | 19 / 25 | 20 % |
| [4.000, 6.000) | 5.000 | 1 | 1 / 25 = 0,04 | 19 + 1 = 20 | 20 / 25 | 4 % |
| [6.000, 8.000) | 7.000 | 2 | 2 / 25 = 0,08 | 20 + 2 = 22 | 22/ 25 | 8 % |
| [8.000, 10.000) | 9.000 | 3 | 3 / 25 = 0,12 | 22 + 3 = 25 | 25 / 25 | 12 % |
| TOTAL |  | 25 |  | 25 | 25 / 25 = 1 | 100.00% |

Programas como Microsoft Excel permiten crear las tablas de distribución, de manera que con los datos se puedan generar a su vez diagramas que presenten de manera sucinta la información. Por ejemplo en [VER](https://youtu.be/bKK0kXzwpgs) se ejemplifica una manera de generar una tabla de distribución de frecuencias.

[SECCIÓN 2] **1.4 Consolidación**

[SECCIÓN 1] **2 Los muestreos**

Un estudio estadístico siempre está basado en el análisis de un conjunto de elementos seleccionados y no de sucesos aislados, éste conjunto es llamado *población*. Algunos ejemplos son la totalidad de votantes para unas elecciones, la totalidad de bombillos producidos por una compañía o la totalidad de animales de un zoológico.

Si la población tiene un número excesivo de elementos, es necesario seleccionar un subconjunto o *muestra*, seleccionada con cuidado para que sea representativa del total de la población. El proceso de selección se denomina *muestreo****.***

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Incluir una imagen de la población y otra de la muestra para un estudio estadístico |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** | 4º ESO/Matemáticas/La estadística/La terminología estadística  [http://profesores.aulaplaneta.com/DNNPlayerPackages/Package14643/InfoGuion/cuadernoestudio/images_xml/MT_10_10_img1_small.jpg](http://profesores.aulaplaneta.com/DNNPlayerPackages/Package14643/InfoGuion/cuadernoestudio/images_xml/MT_10_10_img1_zoom.jpg) |
| **Pie de imagen** | A la izquierda, la población del estudio para saber cuántas bolsas defectuosas de frutos secos se fabrican al día y, a la derecha, la muestra con la que se trabajará. |
| **Ubicación del pie de imagen** | Inferior |

Este proceso lo hacen en Colombia las industrias del café, para conocer cuántas bolsas defectuosas se fabrican cada día. Como no se pueden abrir todas las bolsas, el departamento de control de calidad elige unas muestras de cada lote de producción.

En este ejemplo la población sería el conjunto de bolsas café fabricadas y la muestra correspondería al número de bolsas que se toman para analizar y poder encontrar los resultados del estudio de unidades defectuosos en la muestra, e inferir a partir de tal estudio la cantidad de unidades defectuosas en el conjunto de la población. El muestreo es el método que se sigue para hacer la selección de las bolsas que se examinarán, teniendo en cuenta que el subconjunto elegido (muestra) debe representar al total de los datos (población).

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Población, muestra y muestreo** |
| **Contenido** | * La población: es el conjunto de todos los elementos que son objeto de un estudio. * La muestra: es un subconjunto representativo de la población. * El muestreo es el método que se sigue para seleccionar las muestras. |

[SECCIÓN 2] **2.1 Los tipos de muestreo**

Si bien existen varias formas de seleccionar una muestra en el conjunto de una población, no cualquier selección garantiza la condición de ser representativa de la población total. Por ejemplo, en una fábrica de bombillos, hacer la selección de 100 unidades como muestra de entre 10000 unidades producidas tomando al azar cualesquiera 100 unidades, puede representar a la población total de bombillos. Sin embargo, si se hace lo mismo preguntando por la intención de voto a la Alcaldía de una ciudad, quizá sea importante seleccionar personas mayores de edad, de diferentes localidades o sectores de la ciudad, de diferentes estratos y niveles educativos, de diferentes géneros, etc., es decir que elegir cualesquiera 100 de entre cada 10.000 puede no ser buena idea.

Un elemento fundamental para que un muestreo sea representativo es que esté formado por un número razonable de elementos. Si se comete el error de hacer conclusiones muy generales a partir de la observación de sólo una parte de una determinada población, ese error se denomina *error de muestreo*. Por otra parte, si se comete el error de generalizar conclusiones hacia una población más grande a aquella en la que originalmente se hizo el muestreo tal error se denomina *error de inferencia*.

Un elemento fundamental para que un muestreo sea aleatorio es que la muestra sea escogida al azar, garantizando con ello la neutralidad del analista respecto a los individuos analizados. Todo muestreo aleatorio descansa sobre el principio de equiprobabilidad, según el cual todos los individuos que pertenezcan a la población tienen la misma probabilidad de ser elegidos para formar parte de una muestra, con lo que se garantiza el rigor del análisis posterior.

Si bien existen ejemplos de muestreos no aleatorios, como en el caso en que un maestro elige de entre todos los estudiantes de la institución solo aquellos que son sus estudiantes, los muestreos representativos en estadística descriptiva deben ser además muestreos aleatorios.

Los muestreos aleatorios más utilizados son simples, sistemáticos, estratificados, por conglomerados o polietápicos. En general, se trata de elegir de entre una población de tamaño *N*, una muestra de tamaño *n*.

**Muestreo simple**

El muestreo aleatorio simple es el más sencillo de los métodos de muestreo. A partir de un listado completo de toda la población, se realiza una asignación de un número a cada individuo de la población. Posteriormente se acude al algún medio de selección como sorteo, asignación de números aleatorios predeterminados en una lista o generados por computador para elegir la cantidad de individuos que indique el tamaño de muestra requerido. De este modo se completa el muestreo simple.

En resumen, el proceso de muestreo simple selecciona, en una población de tamaño *N* una muestra de tamaño *n*, de manera que la probabilidad de que un elemento de la población se incluya en la muestra es de *n/N*. Aunque simple, el método requiere que se conozca de antemano a toda la población para poder listarla y la selección de muestras pequeñas puede no ser representativa de la población. Puede ser útil en muestreos para estudios de calidad en objetos producto de cadenas de producción y en general en poblaciones homogéneas y tiene la ventaja de que permite hacer cálculos de medidas de variación y dispersión muy fácilmente.

Veamos un ejemplo: En el marco del proyecto institucional de ahorro y protección financiera de un colegio, 20 estudiantes de bachillerato de diferentes grados viajarán a Cartagena para representar a su colegio en el Foro Nacional de Educación. Uno de los cursos, de 30 estudiantes, deberá elegir aleatoriamente 2 representantes. Se tiene la lista de estudiantes y la respuesta ante la pregunta ¿Ahorra?

|  |  |
| --- | --- |
| Estudiante | ¿Ahorra? |
| Camilo | No |
| Tomás | No |
| Fausto | No |
| Valeria | No |
| Martín | No |
| Jimena | Sí |
| Julián | No |
| Andrés | Sí |
| Santiago | Sí |
| Luisa | No |
| Elizabeth | No |
| Juana | Sí |
| Liliana | No |
| Ana | Sí |
| Laura | No |
| Lucas | Sí |
| Alejandro | No |
| Sara | Sí |
| Mateo | No |
| Daniel | Sí |
| Laura | No |
| Natalia | Sí |
| Juan | Sí |
| Tatiana | Sí |
| Mauricio | No |
| Joaquín | Sí |
| Claudia | Sí |
| Marcela | No |
| Sebastián | No |
| Daniela | No |

Elige, mediante un muestreo aleatorio simple a los dos representantes de este curso. Para lograrlo, puedes hacer un sorteo poniendo los números del 1 al 30 en una bolsa y sacando dos de ellos, elegir aquellos dos estudiantes cuyos códigos hayan sido los números centrales en el sorteo del baloto de un día específico, generando números aleatorios y eligiendo a los dos con el número mayor, etc.

Lo primero que debe hacerse es enumerar a la población.

|  |  |  |
| --- | --- | --- |
| No. | Estudiante | ¿Ahorra? |
| 1 | Camilo | No |
| 2 | Tomás | No |
| 3 | Fausto | No |
| 4 | Valeria | No |
| 5 | Martín | No |
| 6 | Jimena | Sí |
| 7 | Julián | No |
| 8 | Andrés | Sí |
| 9 | Santiago | Sí |
| 10 | Luisa | No |
| 11 | Elizabeth | No |
| 12 | Juana | Sí |
| 13 | Liliana | No |
| 14 | Ana | Sí |
| 15 | Laura | No |
| 16 | Lucas | Sí |
| 17 | Alejandro | No |
| 18 | Sara | Sí |
| 19 | Mateo | No |
| 20 | Daniel | Sí |
| 21 | Laura | No |
| 22 | Natalia | Sí |
| 23 | Juan | Sí |
| 24 | Tatiana | Sí |
| 25 | Mauricio | No |
| 26 | Joaquín | Sí |
| 27 | Claudia | Sí |
| 28 | Marcela | No |
| 29 | Sebastián | No |
| 30 | Daniela | No |

Una primera forma de hacer la selección es generando números aleatorios del 1 al 30, usando la función de Excel “=aleatorio.entre(1;30)”. Los dos primeros números que aparezcan serán las posiciones en el listado de los estudiantes elegidos. El resultado es 7 y 23, así que Julián y Juan serán los representantes.

Para otra forma de hacer la selección, generaremos números aleatorios del 100 al 200, usando la función de Excel “=aleatorio.entre(100;200)”. Los dos estudiantes que tengan los números mayores serán los elegidos. Para lograrlo, frente al primero escribimos ““=aleatorio.entre(100;200)” y arrastramos para que frente a cada uno haya un número aleatorio.

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Generación de números aleatorios usando Excel |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen** | Asignación de números aleatorios a una población. |
| **Ubicación del pie de imagen** | Inferior |

Concluimos el muestreo eligiendo a los seleccionados. En este caso el listado final quedó:

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Estudiante | ¿Ahorra? | Aleatorio |
| 1 | Camilo | No | 162 |
| 2 | Tomás | No | 182 |
| 3 | Fausto | No | 140 |
| 4 | Valeria | No | 117 |
| 5 | Martín | No | 108 |
| 6 | Jimena | Sí | 189 |
| 7 | Julián | No | 102 |
| 8 | Andrés | Sí | 116 |
| 9 | Santiago | Sí | 186 |
| 10 | Luisa | No | 161 |
| 11 | Elizabeth | No | 158 |
| 12 | Juana | Sí | 150 |
| 13 | Liliana | No | 175 |
| 14 | Ana | Sí | 134 |
| 15 | Laura | No | 198 |
| 16 | Lucas | Sí | 178 |
| 17 | Alejandro | No | 200 |
| 18 | Sara | Sí | 160 |
| 19 | Mateo | No | 121 |
| 20 | Daniel | Sí | 130 |
| 21 | Laura | No | 173 |
| 22 | Natalia | Sí | 110 |
| 23 | Juan | Sí | 114 |
| 24 | Tatiana | Sí | 194 |
| 25 | Mauricio | No | 157 |
| 26 | Joaquín | Sí | 199 |
| 27 | Claudia | Sí | 163 |
| 28 | Marcela | No | 172 |
| 29 | Sebastián | No | 117 |
| 30 | Daniela | No | 136 |

Por lo que Alejandro y Joaquín serán los representantes, pues el criterio de selección fue elegir a los dos estudiantes que tengan los números mayores aleatorios. Aunque se obtiene un muestreo simple, un problema de este método de generación de números aleatorios y posterior ordenación es que en ocasiones se repiten los números, si bien se subsana fácilmente generándolos hasta que no suceda que hay más elementos en la muestra que los requeridos.

Otra forma, que suele ser la más empleada y correcta al generar números aleatorios para hacer el muestreo es generando números aleatorios de una cantidad grande de cifras y seleccionando las primeras que respondan a la numeración elegida. Por ejemplo se generan cinco números aleatorios entre 0 y 1’000.000 y se eligen siempre números de dos cifras hasta que tengan sentido en el listado propuesto. Para generarlos se usa la función “=aleatorio.entre(0;1000000)”, con lo que se obtiene:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 336565 | 611783 | 391602 | 654499 | 618010 |

Los números de dos cifras son 33, 65, 65, 61, 17, 83, 39, 16, 02, etc., pero como necesitamos elegir dos con sentido en nuestro listado, 17 y 16 son los primeros. Así, en este muestreo los representantes serían Alejandro y Lucas, que tienen numeración 17 y 16.

**Muestreo sistemático**

El muestreo aleatorio sistemático inicia del mismo modo que el muestreo aleatorio simple, enumerando a todos los miembros de la población. Luego se escoge un único elemento al azar y se completa la muestra eligiendo de manera periódica el resto. La diferencia radica en que, en lugar de extraer *n* números que corresponderán a la muestra, se extrae solo uno, que se llamará *i*. Se parte de ese número aleatorio *i*, que fue elegido al azar, y para completar la muestra se eligen los números *i*, *i+k*, *i+2k*, *i+3k*,..., *i+(n-1)k*, es decir se toman los individuos de *k* en *k*, donde *k* el resultado de dividir el tamaño de la población *N* entre el tamaño de la muestra *n*, con lo cual *k = N/n*. El número *k* se llama *constante de muestreo*.

El riesgo de este tipo de muestreo es que la constante de muestreo homogeniza de alguna manera a la población, y si la homogeneidad está asociada con el fenómeno de interés, las estimaciones obtenidas a partir de la muestra pueden contener sesgo de selección. Supongamos que estamos seleccionando una muestra sobre listas de 10 individuos en los que los 5 primeros son adultos y los 5 últimos son menores de edad. Al aplicar un muestreo aleatorio sistemático con constante de muestreo *k = 10* siempre se seleccionan, o sólo niños o sólo adultos, perdiendo con ello la característica de representatividad requerida en el muestreo. Por el contrario, si la población está ordenada siguiendo una tendencia conocida, este tipo de muestreo asegura una cobertura de unidades de todos los tipos.

En nuestro problema de elegir a los dos representantes del curso, el muestreo sistemático se haría de la siguiente manera: Primero, se elige al azar un número entre 1 y 30, por ejemplo usando Excel, que arroja el número 20, que llamaremos *i*. Así, el primer representante es Daniel, que tiene el número 20.

Luego, como *N = 30* y *n* = 2, entonces *k = N/n= 15*. Como *i+k=20+15=35*, pero nuestra lista llega a 30, iniciamos de nuevo hasta llegar a Martín, que tiene el número 5. Nótese que entre Daniel y Martín hay la misma cantidad de estudiantes:

|  |  |  |
| --- | --- | --- |
| No. | Estudiante | ¿Ahorra? |
| 1 | Camilo | No |
| 2 | Tomás | No |
| 3 | Fausto | No |
| 4 | Valeria | No |
| 5 | Martín | No |
| 6 | Jimena | Sí |
| 7 | Julián | No |
| 8 | Andrés | Sí |
| 9 | Santiago | Sí |
| 10 | Luisa | No |
| 11 | Elizabeth | No |
| 12 | Juana | Sí |
| 13 | Liliana | No |
| 14 | Ana | Sí |
| 15 | Laura | No |
| 16 | Lucas | Sí |
| 17 | Alejandro | No |
| 18 | Sara | Sí |
| 19 | Mateo | No |
| 20 | Daniel | Sí |
| 21 | Laura | No |
| 22 | Natalia | Sí |
| 23 | Juan | Sí |
| 24 | Tatiana | Sí |
| 25 | Mauricio | No |
| 26 | Joaquín | Sí |
| 27 | Claudia | Sí |
| 28 | Marcela | No |
| 29 | Sebastián | No |
| 30 | Daniela | No |

En general, en el muestreo aleatorio sistemático la muestra queda homogéneamente distribuida en el conjunto de la población.

**Muestreo estratificado**

¿Notaste que hasta ahora ninguna niña ha sido elegida para representar al colegio y que no se ha usado el dato que parece relevante respecto a una práctica del ahorro? Por este tipo de circunstancia, el muestreo aleatorio estratificado inicia considerando categorías diferentes entre sí, que sean homogéneas respecto a alguna característica particular con el fin de garantizar la diversidad de la población en la muestra.

En el muestreo aleatorio estratificado cada categoría o *estrato* de representación funciona de manera independiente. En cada uno de los estratos puede aplicarse un muestreo simple o sistemático para elegir los elementos de la muestra.

El muestreo aleatorio estratificado inicia reconociendo las características homogéneas que garanticen representatividad en la muestra según el tipo de estudio poblacional a realizar. Aunque en Colombia la palabra “estrato” se usa frecuentemente para notar el nivel de calidad de la zona en que se ubica una vivienda, se puede estratificar también según la profesión, el municipio de residencia, el género, el estado civil, etc. Sin embargo, como no todos los estratos tienen la misma cantidad de individuos, debe detallarse la distribución de la muestra en los diferentes estratos; tal distribución se conoce como *afijación*.

Existen varios tipos de afijación, entre los que se encuentran la afijación simple, la afijación proporcional, la afijación de mínima varianza y la afijación óptima. En la afijación simple, a cada estrato le corresponde igual número de elementos muestrales; es óptima si los estratos están homogéneamente distribuidos en la población, pero si no es así tal afijación favorece a los estratos con menor cantidad de individuos y perjudica a los que tienen mayor cantidad de individuos. En la afijación proporcional, la distribución de elementos en la muestra es directamente proporcional al tamaño de la población en cada estrato. La afijación de mínima varianza y la afijación óptima requieren cálculos previos de desviación y varianza, por lo que no las especificaremos aquí.

En el ejemplo de seleccionar los representantes del curso que viajarán a Cartagena a representar al colegio, pueden considerarse varias estratificaciones. Una puede ser por género y, la otra, por iniciativa de ahorro.

Para realizar el muestreo estratificado categorizando por género, iniciamos diferenciando la lista por las categorías

|  |  |
| --- | --- |
| **ESTRATO 1** | **ESTRATO 2** |
| |  |  |  |  | | --- | --- | --- | --- | | **No.** | **Estudiante** | **¿Ahorra?** | **Género** | | 1 | Camilo | No | H | | 2 | Tomás | No | H | | 3 | Fausto | No | H | | 4 | Martín | No | H | | 5 | Julián | No | H | | 6 | Andrés | Sí | H | | 7 | Santiago | Sí | H | | 8 | Lucas | Sí | H | | 9 | Alejandro | No | H | | 10 | Mateo | No | H | | 11 | Daniel | Sí | H | | 12 | Juan | Sí | H | | 13 | Mauricio | No | H | | 14 | Joaquín | Sí | H | | 15 | Sebastián | No | H | | |  |  |  |  | | --- | --- | --- | --- | | **No.** | **Estudiante** | **¿Ahorra?** | **Género** | | 1 | Valeria | No | M | | 2 | Jimena | Sí | M | | 3 | Luisa | No | M | | 4 | Elizabeth | No | M | | 5 | Juana | Sí | M | | 6 | Liliana | No | M | | 7 | Ana | Sí | M | | 8 | Laura | No | M | | 9 | Sara | Sí | M | | 10 | Laura | No | M | | 11 | Natalia | Sí | M | | 12 | Tatiana | Sí | M | | 13 | Claudia | Sí | M | | 14 | Marcela | No | M | | 15 | Daniela | No | M | |

Luego de ello, aplicamos un muestreo simple o sistemático a cada uno de los estratos.

Haremos un muestreo simple en cada estrato. Para el primer estrato, usando un listado de números aleatorios para cada estrato y seleccionando números de dos cifras, se obtiene:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 713578 | 97227 | 53412 | 313227 | 86868 |

Los números son 71, 35, 78, 97, 22, 75, 34, 12, 31… Así, el muestreo arroja como ganador a Juan, que tiene el número 12 de entre los niños.

Repitiendo el proceso para el segundo estrato, se generan los números aleatorios y se procede igual. Los números aleatorios son:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 730992 | 112529 | 572350 | 774040 | 69027 |

Los números son 73, 09, 92,… Así, el muestreo arroja como ganadora a Sara, que tiene el número 9, de entre las niñas. Con este muestreo estratificado, los representantes del curso son Juan y Sara, y se garantizó que haya un niño y una niña.

En este caso no se calcula afijación, porque los estratos tienen la misma cantidad de individuos, y la probabilidad de que un niño sea escogido es 1/15, que es la misma de que una niña sea escogida.

En un muestreo que categorice la iniciativa de ahorro, también, iniciamos diferenciando la lista por las categorías

|  |  |
| --- | --- |
| **ESTRATO 1** | **ESTRATO 2** |
| |  |  |  | | --- | --- | --- | | **No.** | **Estudiante** | **¿Ahorra?** | | 1 | Jimena | Sí | | 2 | Andrés | Sí | | 3 | Santiago | Sí | | 4 | Juana | Sí | | 5 | Ana | Sí | | 6 | Lucas | Sí | | 7 | Sara | Sí | | 8 | Daniel | Sí | | 9 | Natalia | Sí | | 10 | Juan | Sí | | 11 | Tatiana | Sí | | 12 | Joaquín | Sí | | 13 | Claudia | Sí | | |  |  |  | | --- | --- | --- | | **No.** | **Estudiante** | **¿Ahorra?** | | 1 | Camilo | No | | 2 | Tomás | No | | 3 | Fausto | No | | 4 | Valeria | No | | 5 | Martín | No | | 6 | Julián | No | | 7 | Luisa | No | | 8 | Elizabeth | No | | 9 | Liliana | No | | 10 | Laura | No | | 11 | Alejandro | No | | 12 | Mateo | No | | 13 | Laura | No | | 14 | Mauricio | No | | 15 | Marcela | No | | 16 | Sebastián | No | | 17 | Daniela | No | |

Luego de ello, se aplica un muestreo simple o sistemático a cada uno de los estratos.

Haremos un muestreo simple en cada estrato. Para el primer estrato, usando un listado de números aleatorios para cada estrato y seleccionando números de dos cifras, se obtiene:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 649907 | 278501 | 929878 | 42982 | 666179 |

Los números son 64, 99, 07, 27,… Así, el muestreo arroja como ganadora a Sara, que tiene el número 7 de entre los que sí ahorran.

Repitiendo el proceso para el segundo estrato, se generan los números aleatorios y se procede igual. Los números aleatorios son:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 755076 | 831991 | 719148 | 496980 | 47466 |

Los números son 75, 50, 76, 83, 19, 91, 71, 91, 48 49, 69, 80, 47, 46 y 6,… Así, el muestreo arroja como ganador a Julián, que tiene el número 6 de entre los que no ahorran. Con este muestreo estratificado, los representantes del curso son Sara y Julián, y se garantizó que haya un estudiante que ahorra y uno que no ahorra.

En este caso se aplicó una afijación simple, porque a cada estrato le correspondió igual número de elementos muestrales, que en este caso fue uno. Sin embargo, como los estratos no tienen la misma cantidad de individuos, la probabilidad de que un estudiante que ahorra sea escogido es de 1/13, que no es igual a la probabilidad de que sea escogido un estudiante que no ahorra, que es de 1/17. En este caso, los estudiantes que no ahorran tienen mayor probabilidad de ser elegidos en la muestra.

**Muestreo por conglomerados**

A diferencia de los métodos de muestreo previos, en el muestreo por conglomerados la unidad muestral no son los elementos de la población, sino subgrupos de elementos de la población que conforman una nueva unidad. Tal unidad se conoce como *conglomerado*. Así, el muestreo por conglomerados consiste en seleccionar aleatoriamente un cierto número de conglomerados y en investigar después todos los elementos pertenecientes a los conglomerados elegidos, de manera que se alcance el tamaño muestral seleccionado.

Para efectuar el muestreo por conglomerados, lo primero que se hace es fraccionar la población en subgrupos que sean convenientes para el muestreo. Tales subgrupos serán los conglomerados y serán las unidades de estudio. Luego, se seleccionan tales unidades por algún método de muestreo aleatorio simple o sistemático para finalmente constituir la muestra con todos los elementos de los conglomerados seleccionados al azar. Este tipo de muestreo es eficaz cuando la población es muy grande y dispersa y presenta la ventaja de que no se requiere tener un listado de todas las unidades poblacionales, sino tan solo el listado de conglomerados como unidades primarias de muestreo.

En el ejemplo de la selección de los representantes del curso en el Foro de Educación, podrían elegirse como conglomerados las profesiones de los padres de los estudiantes. Luego se selecciona al azar la cantidad de profesiones que completen la muestra. Suponga que en la selección al azar se eligieron como conglomerados “Militar” y otro, “Abogado”. Si hay dos estudiantes que sean hijos de militar, ellos serán los representantes del curso. Si hay más de dos, deberán seleccionarse dos de entre el total de estudiantes que cumplan la condición. Si hay uno o ninguno, habrá que escoger entre los que son hijos de abogados el segundo representante del curso.

En general los estudios poblacionales son complejos, por lo es que es de gran utilidad usar el muestreo polietápico que opera en etapas sucesivas, sobre las que se aplica el método de muestreo aleatorio que se considere más adecuado en cada etapa. En [VER](http://recursostic.educacion.es/descartes/web/materiales_didacticos/muestreo_poblaciones_ccg/tipos_muestreo.htm) hay información adicional acerca de los tipos de muestreo

[SECCIÓN 3] **2.2 Consolidación**

SECCIÓN 1] **3 Medidas estadísticas**

Luego de tener claridad respecto al tipo de variable al que refiere un estudio estadístico, de hacer el muestro que garantice que el estudio será manejable y representativo de la población y de tener el conjunto de datos –bien sea listado en detalle o representado en una tabla de frecuencias–, se caracterizan de manera cuantitativa datos más específicos respecto a la población y a la manera en que se distribuyen los datos.

Las medidas de tendencia central como la media, la moda y la mediana son utilizadas en este proceso para establecer el comportamiento, la dispersión o la agrupación del conjunto de datos.

Tal cuantificación puede medir, bien la tendencia de la población, bien su grado de dispersión, o bien los comportamientos que relacionan conjuntamente tanto la tendencia como la dispersión de los datos. Las medidas de posición central y no central y las medidas de dispersión permiten hacer la caracterización.

[SECCIÓN 2] **3.1 Medidas de tendencia central**

Las medidas de tendencia central o valores centrales indican un valor central alrededor del cual se distribuyen el resto de valores de la distribución. Los valores de posición centrales más utilizados son la *media aritmética,* la *mediana* y la *moda.*

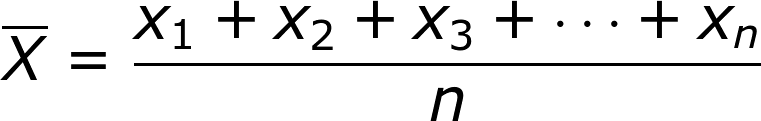
**La media aritmética**

La media aritmética representa en un único valor la cantidad total de la variable, distribuida a partes iguales entre cada observación; se obtiene al sumar todos los datos y dividir el resultado entre el número total de datos. Para representarla se utiliza la letra mayúscula *X* con una barra horizontal sobre ella: *X̄*. No debe confundirse con el uso de la letra minúscula *x* que representa cualquiera de los datos de la muestra.

La media puede ser un número que no tenga sentido en el contexto propuesto; tiene sentido en contextos continuos como el estudio de la variabilidad del peso o la altura de una muestra en una población, pero no necesariamente lo tiene en contextos de estudio de variables discretas, como por ejemplo en el caso en que nos arroja como resultado que las personas leen 1,5 libros o compran 2,5 pares de zapatos.

Los estudiantes suelen tener un acercamiento al cálculo de la media aritmética, pues es la medida de tendencia central que suele aplicarse para el cálculo de las notas en las instituciones escolares que basan las calificaciones en datos numéricos.

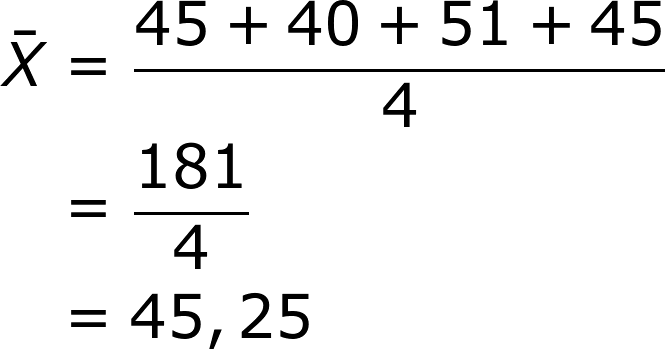
Específicamente, si *x1*, *x2*, *x3*,…, *xn*, representa nuestro conjunto de *n* datos, la media aritmética corresponde a:



donde los *xi* corresponden a cada uno de entre los *n* datos. Si se cuenta con un listado de datos en Excel sobre los cuales se desea calcular la media aritmética, se aplica la función =PROMEDIO (), variando en el conjunto de datos.

Para entender en qué consiste la media aritmética *X̄*, veamos un ejemplo:

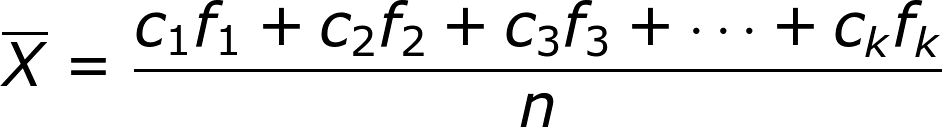
Un estudiante ha obtenido los datos del peso en kilogramos de cuatro de sus compañeros, los cuales son 45, 40, 51 y 45 Kilos. Para establecer un valor medio del peso de sus cuatro compañeros, calcula la media aritméticasumando las cantidades que representan los pesos sus compañeros. Luego, divide el entre el número total de individuos, que en este caso es 4. Entonces, el cálculo realizado en el primer momento es: *(45 + 40 + 51 + 45) = 181*, que es la suma de los pesos de los cuatro. En el segundo momento, lo divide y obtiene *181 / 4 = 45,25* kilos, lo cual corresponde a la media aritmética del conjunto de datos y se escribe: *X̄ = 45,25* kilos. En la escritura estándar ello significa:



Lo que significa que la media o promedio aritmético del peso de los cuatro estudiantes es de 45,25 kilos.

En el caso en que el conjunto de datos se presente en una tabla de frecuencias, el cálculo de media requiere el uso de las frecuencias relativas.

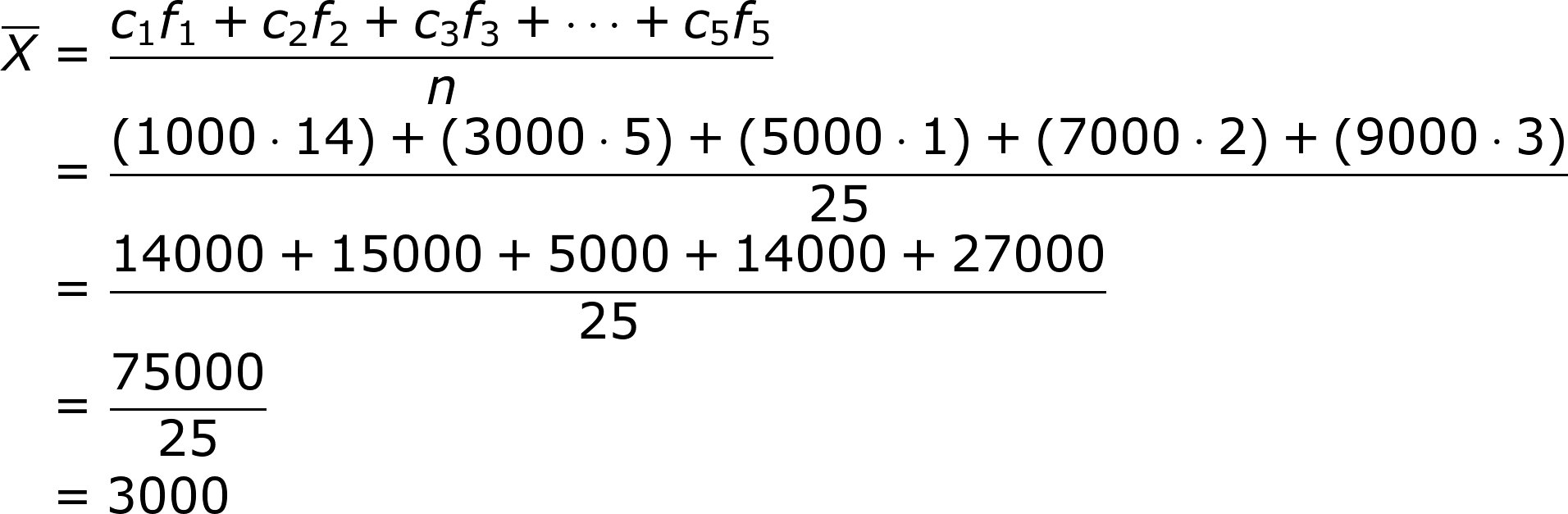
Para encontrar la media aritmética ( *X̄* ) de un conjunto de datos proveniente de una tabla de frecuencias, el cálculo consiste en sumar los productos de las marcas de clase (*ci*) por las frecuencias relativas (*fi*) de cada clase y luego dividir todo por la cantidad de datos. Así, si tenemos *k* clases con marcas de clase *c1*, *c2*, *c3*, …, *ck*  y sus frecuencias relativas respectivas *f1*, *f2*, *f3*, …, *fk*, de un conjunto de *n* datos, el resultado de la media aritmética se obtiene calculando:



Por ejemplo la tabla de frecuencias siguiente muestra los resultados de una encuesta aplicada en una clase de 25 alumnos en la que se les preguntó la máxima distancia a la que cada uno había viajado, medida entre su actual lugar de residencia y la capital del país extranjero visitado.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cantidad de kilómetros recorrida**  ***xi*** | **Marca de clase**  ***(ci)*** | **Frecuencia absoluta *(fi)*** | **Frecuencia relativa**  ***(hi= fi/n)*** | **Frecuencia absoluta acumulada *(Fi)*** | **Frecuencia relativa acumulada *(Hi= Fi/n )*** | **Porcentaje (%)**  **(*hi* escrita como porcentaje)** |
| [0, 2.000) | 1.000 | 14 | 14 / 25 = 0,56 | 14 | 14 / 25 | 56 % |
| [2.000, 4.000) | 3.000 | 5 | 5 / 25 = 0,2 | 14 + 5 = 19 | 19 / 25 | 20 % |
| [4.000, 6.000) | 5.000 | 1 | 1 / 25 = 0,04 | 19 + 1 = 20 | 20 / 25 | 4 % |
| [6.000, 8.000) | 7.000 | 2 | 2 / 25 = 0,08 | 20 + 2 = 22 | 22/ 25 | 8 % |
| [8.000, 10.000) | 9.000 | 3 | 3 / 25 = 0,12 | 22 + 3 = 25 | 25 / 25 | 12 % |
| TOTAL |  | 25 |  | 25 | 25 / 25 = 1 | 100.00% |

Para el cálculo de la distancia media a la que ha viajado el grupo de estudiantes, nótese que *n=25* y corresponde a la cantidad de datos, mientras que *k=5* indica la cantidad de clases de la tabla. Ya con ello, se establece que la distancia media es:

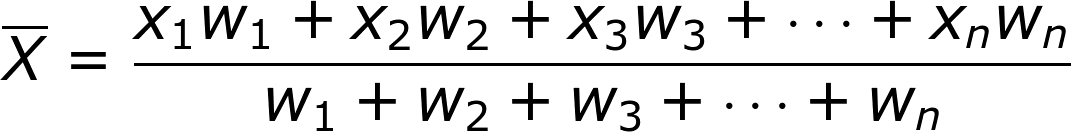


Así que el grupo de estudiantes ha viajado, en promedio, 3000 kilómetros.

**La media aritmética ponderada**

No en todas las ocasiones los datos obtenidos tienen la misma relevancia o peso para mostrar el comportamiento de los datos en la población. Cuando ello ocurre, se asignan pesos a los datos, dependiendo de su relevancia en el marco del estudio a realizar y de esa manera se obtiene la media ponderada del conjunto de datos.

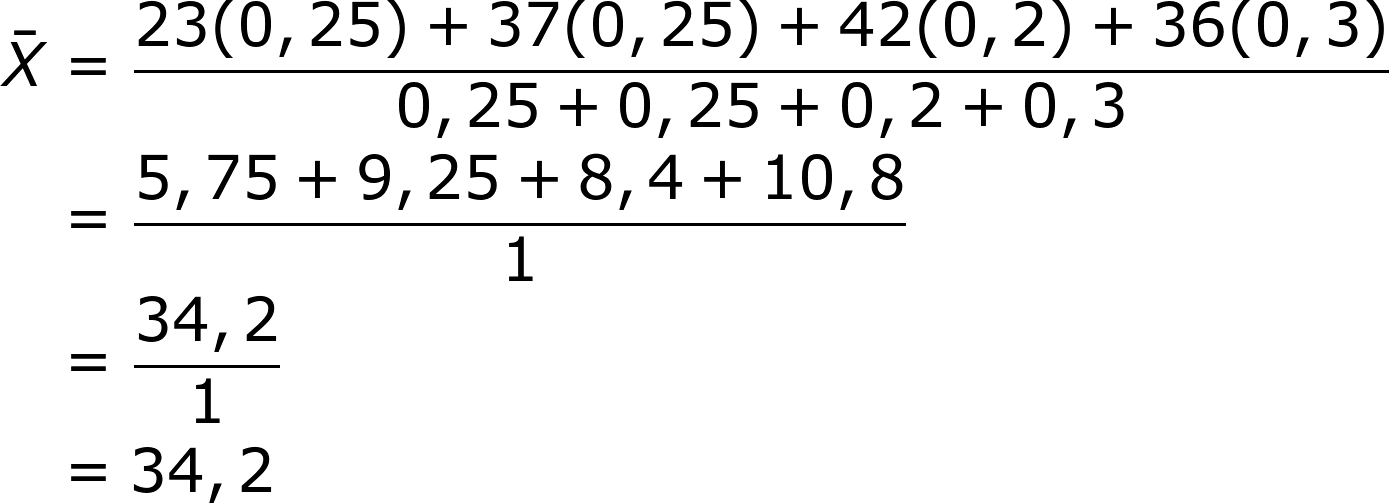
La media ponderada se calcula haciendo los productos entre los datos y su nivel de relevancia, que se expresa como un porcentaje. Más específicamente, si *x1*, *x2*, *x3*, …, *xn*, representa nuestro conjunto de *n* datos y si *w1*, *w2*, *w3*, …, *wn*, representa el conjunto de *n* pesos respectivos, la media ponderada se define de la siguiente forma:

****

Veamos el cálculo de la media ponderada en un ejemplo concreto:

Para valorar los desempeños de los estudiantes en primer semestre, una profesora hace dos pruebas escritas parciales y una prueba general final; valora además la actitud en clase. Cada una de las pruebas parciales corresponde al 25 % sobre la nota final, la actitud en clase corresponde al 20 % y -por reglamento- la prueba final corresponde al 30 % de la nota de la materia. ¿Cuál será la puntuación de un alumno que en las dos pruebas parciales sacó 23 y 37, respectivamente, y cuyas notas en el examen final y en actitud han sido 36 y 42 respectivamente?

Para conocer la respuesta se calcula la media aritmética ponderada:



Así, el desempeño del estudiante en esa materia es de 34,2. Si las notas son calculadas sobre 50 y la nota aprobatoria es 30, pasó la materia. Si las notas son calculadas sobre 100 y la nota aprobatoria es 50, perdió la materia. Si está en posgrado, las notas son calculadas sobre 50 y la nota aprobatoria es 35, así que el estudiante reprobó la materia.

En el caso en el que los datos para el cálculo de la media ponderada provengan de una tabla de distribución de frecuencias, habría que garantizar que los pesos de la ponderación varían justamente en las clases establecidas. De lo contrario, habría que expandir los datos y hacer las ponderaciones como en el caso para datos no agrupados.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **La media aritmética y ponderada** |
| **Contenido** | Llamamos media aritmética ( *X̄* ) de un conjunto *x1*, *x2*, *x3*, …, *xn* de *n* datos al cociente de la suma de todos ellos por el número *n*.    Si el cálculo de la media aritmética se hace sobre una tabla de distribución de frecuencias, con *k* clases, marcas de clase *c1*, *c2*, *c3*, …, *ck*  y frecuencias relativas correspondientes *f1*, *f2*, *f3*, …, *fk*, de un conjunto de *n* datos, la media aritmética es:    Para encontrar la media aritmética ponderada ( *X̄* ) de un conjunto *x1*, *x2*, *x3*, …, *xn*  de *n* datos, se suman los productos de estos números por sus pesos respectivos *w1*, *w2*, *w3*, …, *wn* (en %) y el resultado se divide por la suma de los pesos (que corresponde al 100 %, osea que debe sumar 1). |

**La mediana**

Otra importante medida de tendencia central es la llamada ***mediana*** de un conjunto de datos, que corresponde al valor que se encuentra **en el centro de la serie ordenada de datos**. Para representarla se utiliza la el apócope *Med*. Si la cantidad de datos es impar, la mediana es el valor central entre el conjunto de *n* datos, que deja la misma cantidad de datos a derecha e izquierda de la lista ordenada. Si la cantidad *n* de datos es impar, no habrá un único valor que está en el centro de la lista ordenada; en ese caso la mediana corresponde a la media aritmética entre los dos valores centrales en la lista ordenada. La función en Excel que permite hacer el cálculo de la mediana es = MEDIANA(), con variación en el conjunto de datos.

|  |  |
| --- | --- |
| Si *n* es par | Si *n* es impar |
| donde los *xi* están en la lista ordenada de *n* datos | donde los *xi* están en la lista ordenada de *n* datos |

Para ejemplificar el cálculo de la mediana veamos los siguientes ejemplos:

En el campeonato local de la liga colombiana de fútbol se juegan, para la misma temporada, dos fases: la fase “Todos contra todos” compuesta por 20 fechas y la fase de “finales”. La cantidad de goles que marcó un equipo en las 20 fechas de la fase “Todos contra todos” se presenta a continuación:

2, 1, 0, 1, 3, 0, 1, 1, 0, 2, 1, 0, 1, 0, 1, 1, 2, 0, 2, 1.

Para calcular la mediana de los datos, se procede de la siguiente manera:

* Se ordena la lista de datos, de menor a mayor

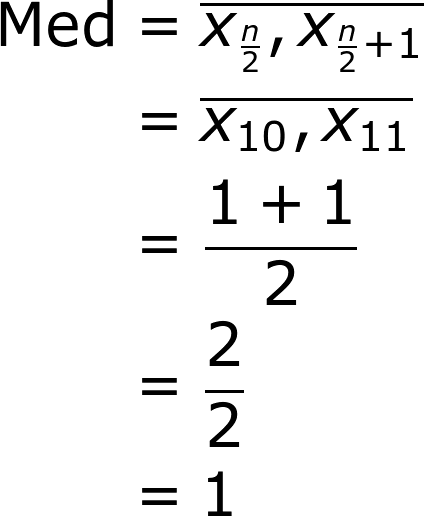
0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3

* Se identifica si la cantidad de datos es par o impar

en este caso, como *n=20*, la cantidad de datos es par

* Si la cantidad de datos es impar, se escoge el valor central de la lista ordenada. Este no es el caso, pues tenemos cantidad de datos par.
* Si la cantidad de datos es par, se escogen los dos valores centrales de la lista ordenada y se hace la media aritmética entre ellos.
* 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3

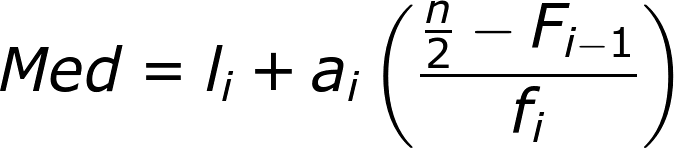
En nuestro listado, los dos valores centrales son los de las posiciones 10 y 11 de la lista ordenada, que dejan 9 datos a cada lado de la lista ordenada. La mediana es la media aritmética entre ellos dos, es decir que se obtiene:



Así, la mediana es la media aritmética o promedio de la pareja de **valores centrales 1**; esta es la **mediana** de la cantidad de goles marcados por el equipo en la fase “Todos contra todos”.

La mediana también puede ser encontrada cuando el conjunto de datos ha sido agrupado por intervalos. En este caso se debe conocer el límite inferior de la clase donde se encuentra la mediana (*li*), la frecuencia acumulada anterior a la clase donde está la mediana (*Fi-1*), la frecuencia absoluta de la clase donde está la mediana (*fi*) y la amplitud de la clase en que se encuentra la mediana (*ai*).

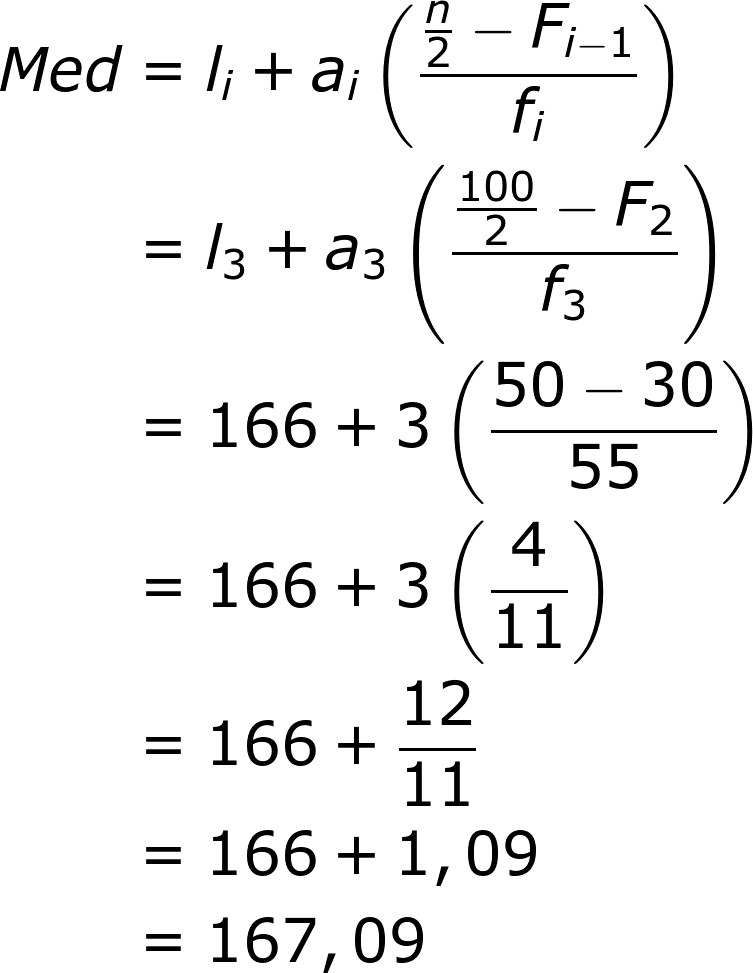
Para el cálculo se procede buscando el valor que corresponde a la posición *n / 2*, y se hace el cálculo de la mediana en la clase *i* a la que pertenezca la mediana, por medio de la siguiente ecuación:



Por ejemplo, en un estudio para identificar tallajes para las chaquetas de la promoción, se agrupan las estaturas de un grupo 100 de estudiantes en cinco intervalos:

|  |  |  |
| --- | --- | --- |
| ***Estatura (en cms)*** | ***fi*** | ***Fi*** |
| [160, 163) | 7 | 7 |
| [163, 166) | 23 | 30 |
| [166, 169) | 55 | 85 |
| [169, 172) | 10 | 95 |
| [172, 175) | 5 | 100 |

En este caso, la clase en la que se encuentra la mediana es la clase *i =* 3, pues es en esa clase en la que se sobrepasa la mitad de los datos ordenados. Así, para este caso, con *i =* 3, se tiene que el límite inferior de la tercera clase es 166, es decir *l3 = 166*; la amplitud de la tercera clase en que se encuentra la mediana es *a3 =* 3, la frecuencia acumulada anterior a la clase tres en que está la mediana es *F2 = 30*, la frecuencia absoluta de la clase 3 es *f3 = 55* y la mitad de los datos es *n/2 = 50*. Así, el cálculo de la mediana queda:



Por lo que la mediana para el conjunto de datos agrupados es 167,09 cms de estatura, lo que significa que 50 estudiantes miden menos de 167, 09 y los otros 50 miden más que 167,09.

**La moda**

La **moda** de un conjunto de datos indica la tendencia de dicho conjunto, es decir el dato más frecuente que corresponde al valor que más veces se repite. La moda se representa como *Mod* y es siempre un valor que tuvo ocurrencia en el conjunto de datos. Se trata de una medida de tendencia central que se puede calcular tanto para variables cualitativas como cuantitativas. El cálculo de la moda para datos sin agrupar se realiza haciendo el conteo de la cantidad de veces que se repite cada dato y eligiendo como moda aquel que tenga la mayor repetición. En caso de que varios datos se repitan la misma cantidad de veces, se dirá que la población tiene una distribución bimodal, trimodal y, en general, multimodal.

Por ejemplo en el siguiente par de series se presenta el resultado de una encuesta en la que se preguntó a los 29 estudiantes de un salón la edad en la que experimentaron su primer beso. Los resultados para los 11 estudiantes de género masculino fueron los siguientes:

Edades para los hombres: 13,14, 14, 13, 15, 15, 14, 15, 13, 15, 15

En este conjunto de datos se puede evidenciar que 15 es la moda o el valor más repetido.

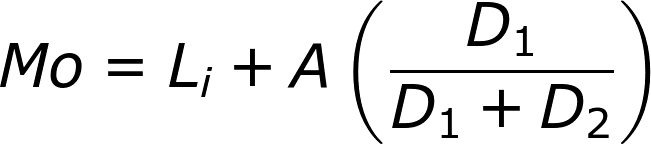
Por su parte, en el grupo de género femenino, las restantes 18 estudiantes respondieron lo siguiente:

Edades para las mujeres: 14, 13,14, 14, 13, 15, 12, 15, 14, 15, 13, 15, 15, 14, 15, 15, 14, 14.

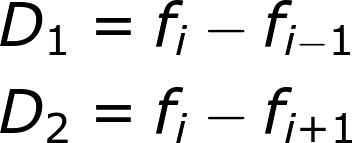
En este conjunto de datos, las edades 14 y 15 tienen la misma frecuencia, pues se repiten 7 veces cada una. Cuando existen dos modas o dos valores que tiene valores de mayor frecuencia se dice que el conjunto de datos tiene una distribución bimodal.

En el caso en que se estudie la población total del salón sin estratificarla por género, en los datos de los 29 estudiantes la distribución es unimodal, con *Mod = 15*.

Para encontrar la moda en un conjunto de datos agrupados es necesario tener en cuenta el límite inferior de la clase modal o clase en la que se presenta la máxima frecuencia (*Li*), la frecuencia absoluta de la clase modal (*fi*), la frecuencia absoluta de la clase inmediatamente inferior a la clase modal, (*fi-1*), que llamaremos *clase premodal*, la frecuencia absoluta de la clase inmediatamente posterior a la clase modal (*fi+1*), que llamaremos *clase postmodal* y la amplitud de las clases (*A*). El cálculo de la moda en este caso responde a la ecuación:



donde *D1* y *D2* representan, la diferencia entre la clase modal y la premodal, y la diferencia entre la clase modal y la postmodal respectivamente. Es decir que:

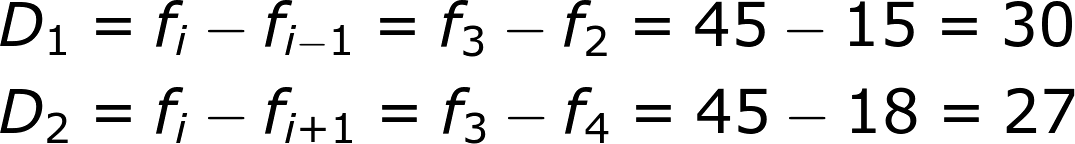


Nótese que los valores *D1* y *D2* son siempre positivos, dado que *fi* es el máximo entre las frecuencias.

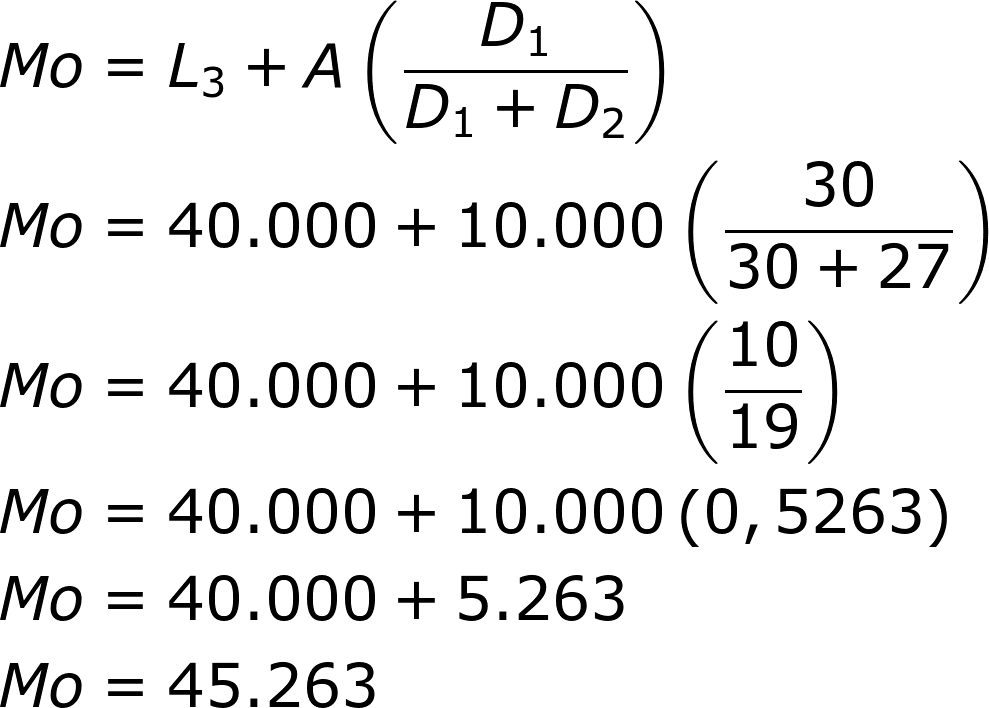
Para comprender el funcionamiento de la ecuación para el cálculo de la moda para datos agrupados, veamos el siguiente ejemplo: la tabla representa los resultados respecto a la cantidad de dinero o salario, en pesos, que gana un grupo de 100 padres de familia.

|  |  |
| --- | --- |
| **Dinero ganado en un día** | ***fi*** |
| [20.000, 30.000) | 10 |
| [30.000, 40.000) | 15 |
| [40.000, 50.000) | 45 |
| [50.000, 60.000) | 18 |
| [60.000, 70.000) | 12 |

Entonces, inicialmente observamos cuál es la clase en la que aparece la frecuencia absoluta más alta. En este caso la clase con frecuencia máxima es la tercera, es decir que *i = 3*. Ahora bien, calculando los valores de las diferencias *D1* y *D2* se obtiene:



Ya con los valores de las dos diferencias, se procede al cálculo de la moda. Así:



De esta manera, la moda o la mayor frecuencia en los datos recolectados es de $45.263, lo cual significa que la mayoría de los padres del grupo obtiene ese salario cada día.

[SECCIÓN 2] **3.2 Medidas de dispersión**

Si bien el conocimiento de las medidas de tendencia central para una población arroja información importante para lograr caracterizar dicha población, es cierto también que dos distribuciones pueden tener las mismas medidas de centralización y, aun así, ser muy diferentes.

La estadística reconoce por ejemplo que cuando la desviación de los valores observados respecto a un valor central es muy grande, las medidas de tendencia central no son representativas del total; reconoce así que es necesario caracterizar mejor la población. Para lograrlo, trata de estudiar y cuantificar la dispersión de los datos en la muestra o en la población, lo cual se conoce como ***medidas de dispersión***.

En el objetivo de completar el conocimiento de la distribución se proponen las siguientes medidas de dispersión: el rango, la desviación media, la varianza, la desviación estándar y el coeficiente de variación.

**El rango**

Una primera forma de **medir la dispersión se logra estudiando qué tan grande es el intervalo en el que varían los datos.** Para obtener esa medida se calcula el recorrido o rango, es decir, la **diferencia** entre la observación **máxima** y la **mínima en la población**, el cual se simboliza mediante la letra R.

Cuando los datos se distribuyen de manera homogénea, el rango caracteriza bien la dispersión de la población. Sin embargo, ya que el rango depende exclusivamente de las observaciones máxima y mínima, la dispersión puede estar desajustada en el caso de que dichos extremos sean valores atípicos. Como identificación de dispersión, se considera que cuanto mayor sea el rango o recorrido, más grande será la dispersión de los datos.

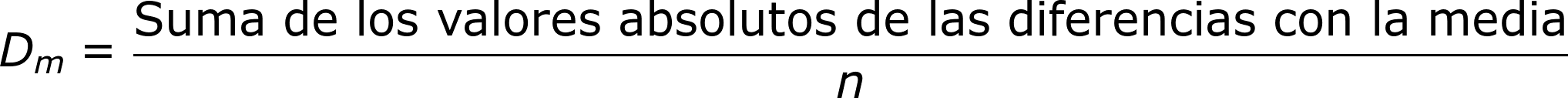
Por ejemplo, en el estudio de una población se recogieron los datos acerca de la cantidad de lotes de un artículo, producidos en un turno de la fábrica. El número mínimo de lotes fue 23 y el máximo 89. Por tanto, el rango o recorrido sería:

Rango o recorrido = 89 – 23 = 66 lotes.

Entonces, se sabe que la cantidad de lotes por turno puede variar entre 66 datos, pero no se tienen más datos acerca de la población.

**La desviación media**

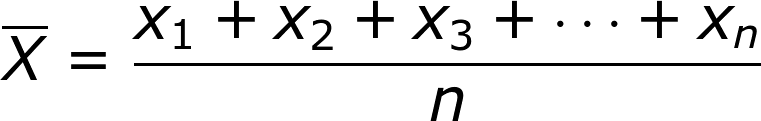
Otra forma de caracterizar la dispersión de la población se realiza midiendo qué tan lejos están las observaciones respecto a la media, las cuales se conocen como *desviaciones respecto a la media*. La media aritmética de los valores absolutos de estas desviaciones es lo que se conoce como *desviación media* de la población con *n* datos, que se representa por *Dm.* Así:



Para ejemplificar el cálculo de la desviación media veamos el siguiente ejemplo: Si la cantidad de goles que marcó un equipo en las 20 fechas de la fase “Todos contra todos” es

2, 1, 0, 1, 3, 0, 1, 1, 0, 2, 1, 0, 1, 0, 1, 1, 2, 0, 2, 1, ¿cuál es la desviación media?

Para calcularla, inicialmente se calcula la media aritmética del conjunto de datos.



Ya que la suma de los datos es 20 y que la cantidad de datos es 20, entonces la media aritmética es *X̄* = 20/20 = 1.

Las desviaciones respecto a la media son las siguientes:

|  |  |  |
| --- | --- | --- |
| **Cantidad de goles**  ***xi*** | **Desviaciones respecto a la media**  ***xi - X̄*** | **Valores absolutos de las desviaciones respecto a la media**  **| *xi - X̄|*** |
| 2 | 2 – 1 = 1 | 1 |
| 1 | 1 – 1 = 0 | 0 |
| 0 | 0 – 1 = **-1** | 1 |
| 1 | 1 – 1 = 0 | 0 |
| 3 | 3 – 1= 2 | 2 |
| 0 | 0 – 1= **-1** | 1 |
| 1 | 1 – 1= 0 | 0 |
| 1 | 1 – 1= 0 | 0 |
| 0 | 0 – 1 = **-1** | 1 |
| 2 | 2 – 1 = 1 | 1 |
| 1 | 1 – 1= 0 | 0 |
| 0 | 0 – 1 = **-1** | 1 |
| 1 | 1 – 1= 0 | 0 |
| 0 | 0 – 1= **-1** | 1 |
| 1 | 1 – 1= 0 | 0 |
| 1 | 1 – 1= 0 | 0 |
| 2 | 2 – 1 = 1 | 1 |
| 0 | 0 – 1 = **-1** | 1 |
| 2 | 2 – 1 = 1 | 1 |
| 1 | 1 – 1= 0 | 0 |
| **Desviación media = Media aritmética de los valores absolutos de las desviaciones** | | *Dm* = 12/20 = 0,6 |

Así, la desviación media del conjunto de datos es *Dm* = 0,6, que indica el grado de dispersión (o alejamiento) de los datos respecto a la media.

Nota que algunas de las diferencias resultan negativas. Por ello se requiere calcular los valores absolutos de las desviaciones, de manera que la desviación media es siempre una media aritmética de valores positivos. Cuando los datos están agrupados, se toma la marca de clase como la media aritmética de los extremos de un intervalo, lo que simplifica el cálculo de la desviación media pues de ese modo se sustituye cada intervalo por un solo número.

**La varianza**

**La varianza** (o ***σ2***) de un conjunto de valores **es la media aritmética de los cuadrados de las desviaciones** de esos valores respecto a la media aritmética. Se expresa en unidades cuadradas.

Calculemos la varianza de la población usada en el ejemplo anterior. En este caso:

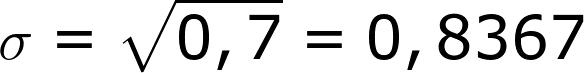
|  |  |  |
| --- | --- | --- |
| **Cantidad de goles**  ***xi*** | **Desviaciones respecto a la media**  ***xi - X̄*** | **Cuadrados de las desviaciones respecto a la media**  **( *xi - X̄ )2*** |
| 2 | 2 – 1 = 1 | 1 |
| 1 | 1 – 1 = 0 | 0 |
| 0 | 0 – 1 = **-1** | 1 |
| 1 | 1 – 1 = 0 | 0 |
| 3 | 3 – 1= 2 | 4 |
| 0 | 0 – 1= **-1** | 1 |
| 1 | 1 – 1= 0 | 0 |
| 1 | 1 – 1= 0 | 0 |
| 0 | 0 – 1 = **-1** | 1 |
| 2 | 2 – 1 = 1 | 1 |
| 1 | 1 – 1= 0 | 0 |
| 0 | 0 – 1 = **-1** | 1 |
| 1 | 1 – 1= 0 | 0 |
| 0 | 0 – 1= **-1** | 1 |
| 1 | 1 – 1= 0 | 0 |
| 1 | 1 – 1= 0 | 0 |
| 2 | 2 – 1 = 1 | 1 |
| 0 | 0 – 1 = **-1** | 1 |
| 2 | 2 – 1 = 1 | 1 |
| 1 | 1 – 1= 0 | 0 |
| **Varianza = Media aritmética de los cuadrados de las desviaciones** | | ***σ2*** = 14/20 = 0,7 |

Así que la varianza en este caso será ***σ2*** = 0,7 goles***2***

**La desviación estándar**

**Para expresar la desviación en las mismas unidades que las usadas al hacer las observaciones utilizamos como medida de dispersión la raíz cuadrada positiva de la varianza. Dicha raíz cuadrada se llama *desviación estándar*, que se representa por *σ*. La desviación estándar sirve para medir el grado de dispersión, pues cuando dos distribuciones tienen la misma media, la desviación estándar indica lo alejados que están los valores respecto a esa media.**

**La desviación estándar** para la cantidad de goles que marcó el equipo en las 20 fechas de la fase “Todos contra todos” tuvo como varianza ***σ2*** = 0,7 goles***2*, por lo que la desviación estándar para el conjunto de datos será:**



El cálculo de las medidas de dispersión para el caso en que los datos se encuentran agrupados, se sigue el procedimiento para hacer el cálculo de la media aritmética y, posteriormente, se calculan las dispersiones y sus cuadrados, a fin de calcular tanto la varianza como la desviación estándar. El procedimiento se ejemplifica a continuación:

Se ha obtenido la medida de la altura de 100 personas de una institución. Los resultados obtenidos, agrupados en intervalos han sido:

|  |  |
| --- | --- |
| **Intervalo**  **(Talla en cm)** | **Frecuencia absoluta**  ***Fi*** |
| [155, 160) | 6 |
| [160, 165) | 16 |
| [165, 170) | 24 |
| [170, 175) | 27 |
| [175, 180) | 19 |
| [180, 185) | 8 |
| ***n*** | 100 |

Inicialmente, se incluye una columna, para sustituir cada intervalo, por su respectiva marca de clase, que corresponde el punto medio del intervalo:

|  |  |  |
| --- | --- | --- |
| **Intervalo**  **(talla en cm)** | **Marca de clase**  ***xi*** | **Frecuencia absoluta**  ***fi*** |
| [155, 160) | (155, 160) / 2 = 157,5 | 6 |
| [160, 165) | (160, 165) / 2 = 162,5 | 16 |
| [165, 170) | (165, 170) / 2 = 167,5 | 24 |
| [170, 175) | (170, 175) / 2 = 172,5 | 27 |
| [175, 180) | (175, 180) / 2 = 177,5 | 19 |
| [180, 185) | (180, 185) / 2 = 182,5 | 8 |
| ***n*** | | **100** |

A continuación se utilizan las marcas de clase para encontrar la media de las estaturas medidas.

|  |  |  |  |
| --- | --- | --- | --- |
| **Intervalo**  **(talla en cm)** | **Marca de clase**  ***xi*** | **Frecuencia absoluta**  ***fi*** | ***fi* *xi*** |
| [155, 160) | (155, 160) / 2 = 157,5 | 6 | 945 |
| [160, 165) | (160, 165) / 2 = 162,5 | 16 | 2.600 |
| [165, 170) | (165, 170) / 2 = 167,5 | 24 | 4.020 |
| [170, 175) | (170, 175) / 2 = 172,5 | 27 | 4.658 |
| [175, 180) | (175, 180) / 2 = 177,5 | 19 | 3.373 |
| [180, 185) | (180, 185) / 2 = 182,5 | 8 | 1.460 |
|  |  | **100** | **17.055** |

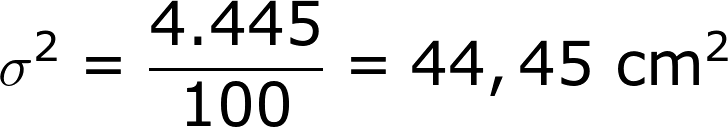
Con los resultados de la última columna y sabiendo que la cantidad *n* total de datos es *n = 100*, se obtiene la media aritmética del conjunto de datos:

http://profesores.aulaplaneta.com/DNNPlayerPackages/Package14643/InfoGuion/cuadernoestudio/images_xml/MT_10_10_formula6_resized.gif

Para realizar el cálculo de la varianza y de la desviación estándar se calculan las desviaciones y sus cuadrados. En la última columna de la derecha de la siguiente tablar se recogen los datos para poder calcular la varianza y la desviación estándar.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Intervalo**  **(talla en cm)** | **Marca de clase**  ***xi*** | **Frecuencia absoluta**  ***fi*** | ***(xi* – *X̄)2*** | ***fi* *(xi* – *X̄)2*** |
| [155, 160) | (155, 160) / 2 = 157,5 | 6 | 170 | 1.022 |
| [160, 165) | (160, 165) / 2 = 162,5 | 16 | 65 | 1.037 |
| [165, 170) | (165, 170) / 2 = 167,5 | 24 | 9 | 223 |
| [170, 175) | (170, 175) / 2 = 172,5 | 27 | 4 | 103 |
| [175, 180) | (175, 180) / 2 = 177,5 | 19 | 48 | 918 |
| [180, 185) | (180, 185) / 2 = 182,5 | 8 | 143 | 1.142 |
| ***n*** | | **100** |  | **4.445** |

En base a lo anterior, se deduce que la varianza es:



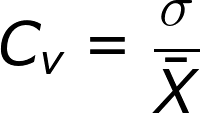
Sacando la raíz cuadrada positiva a la varianza se obtiene la desviación estándar que será entonces:



**El coeficiente de variación**

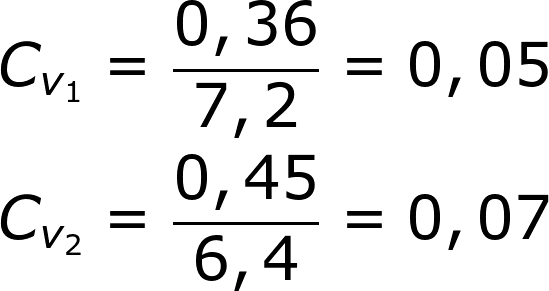
Para **comparar** dos distribuciones heterogéneas con distinta media y saber la dispersión de sus valores, podemos utilizar el **coeficiente de variación**, que normalmente se expresa en porcentaje (%). El coeficiente de variación es el cociente de la desviación estándar entre la media y se simboliza mediante Cv. Cuando se comparan dos distribuciones, la que tiene mayor coeficiente de variación es también la que tiene mayor dispersión.

El coeficiente de variación se calcula dividiendo la desviación estándar entre la media:



Por ejemplo, la nota media en matemáticas de la clase de 10º A ha sido de 7,2 y su desviación estándar de 0,36. En cambio, en la clase de 10º B ha sido de 6,4 y desviación estándar de 0,45. ¿Cuál es el coeficiente de variación de la clase de 10º A? ¿Y el de 10º B?

Aplicando la ecuación se obtiene:



Así, como el coeficiente de variación para los estudiantes del primer curso fue de 0,05, eso significa que la variación es del 5 % en 10º A. Para el segundo curso el coeficiente de variación de 0,07 indica que la variación es del 7 % en 10º B. Por lo tanto se concluye que las notas de matemáticas han estado más dispersas en la clase de 10º B.

[SECCIÓN 2] **3.3 Medidas de posición no central**

Las medidas estadísticas de posición pueden ser **valores centrales** o valores **no centrales**. En estadística tanto los primeros como los segundos se calculan para hacer la caracterización de la población. Los más comunes valores de posición central, llamados también medidas de tendencia central son la **media aritmética**, la **mediana** y la **moda.**

**Por su parte, las medidas de posición no central sirven para dividir** un conjunto de valores de una variable estadística en subconjuntos con la misma cantidad de valores, **y para establecer el comportamiento de los datos en diferentes puntos. Como estas medidas son complementarias a los valores centrales, pueden ayudar a dar información más completa del conjunto de datos,** generar diferentes representaciones de la muestra que ayuden a entender el comportamiento de la misma y a hacer inferencias acerca del comportamiento de la población de estudio. Las medidas de tendencia no central más utilizadas en estadística son los cuartiles, los quintiles, los deciles y los percentiles.

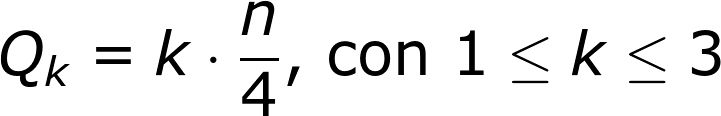
**Los cuartiles**

Los **cuartiles, como medida de posición no central,** dividen el conjunto de datos en cuatro grupos iguales. Es decir, se si tuviera la lista de 100 datos ordenados de menor a mayor, los cuartiles los dividirían en subgrupos de 25 datos; en una población de 60 datos, los cuartiles los dividirían en subconjuntos de 15 datos.

Los cuartiles son tres, Q1, Q2 y Q3. El primer cuartil es Q1, hasta el que se encuentran el primer 25% de los datos. El segundo cuartil es Q2 hasta el que se encuentra el segundo 25% de los datos, es decir el 50% de los datos y por ello coincide exactamente con la mediana. Finalmente, el tercer cuartil es Q3, hasta el que se encuentra el tercer 25% de los datos, es decir el 75 % de los datos. Así, entre el primer y el tercer cuartil, es decir entre Q1 y Q3, se encuentra ubicada la mitad de la población.

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Representación de los cuartiles sobre un segmento de recta |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** | http://dieumsnh.qfb.umich.mx/estadistica/medidasd%20de%20posicion_archivos/image002.jpg |
| **Pie de imagen** | Representación de los cuartiles |
| **Ubicación del pie de imagen** | Inferior o lateral |

La ecuación que permite calcular la posición de cada cuartil en un conjunto de *n* datos es:



La variación del parámetro *k* permite obtener la posición sobre la lista ordenada –de menor a mayor– de datos para los tres cuartiles. Así, para *k = 1*, se obtiene la posición del primer cuartil; para *k = 2*, se obtiene la posición del segundo cuartil o mediana; y para *k = 3*, se obtiene la posición del tercer cuartil.

Otra forma de decirlo es que el primer cuartil *Q1* corresponde a la mediana de la primera mitad de valores, el segundo cuartil *Q2* es la mediana general de la serie y que el tercer cuartil *Q3* corresponde a la mediana de la segunda mitad de valores.

Veamos el proceso de cálculo de los cuartiles en un ejemplo concreto: En una fábrica, la cantidad de lotes producidos por cada una de las 15 máquinas de producción ha sido anotado y se presenta en el siguiente listado:

10, 32, 19, 25, 15, 7, 17, 16, 24, 30, 12, 29, 27, 18, 27

Para hacer el cálculo de los cuartiles, se ordenan las observaciones en orden creciente, teniendo en cuenta que la cantidad de datos es *n = 15*:

7, 10, 12, **15**, 16, 17, 18, **19**, 24, 25, 27, **27**, 29, 30, 32

**A continuación se puede, bien usar la ecuación para calcular la posición de cada cuartil, o se puede iniciar** localizando la mediana *Q2* y luego calcular la posición de *Q1* y de *Q3* como medianas de la primera y segunda mitad de los datos.

Si se toma esta segunda opción, nótese que como la cantidad de datos es impar, la mediana corresponde al dato que está en el centro de la lista ordenada y que deja a cada lado la misma cantidad de datos, en este caso 7 a cada lado:

7, 10, 12, **15**, 16, 17, 18, **19**, 24, 25, 27, **27**, 29, 30, 32

entonces



**que corresponde a la mediana o segundo cuartil.**

**Para encontrar los cuartiles** *Q1* y de *Q3* se realiza el mismo procedimiento anterior, para los datos que quedan a cada lado de la mediana. Entonces, la primera mitad de los datos ordenados contiene 7 datos y la mediana de ese conjunto es el primer cuartil o *Q1*:

7, 10, 12, **15**, 16, 17, 18

Entonces



Por su parte la segunda mitad de los datos ordenados también contiene 7 datos y la mediana de ese conjunto es el tercer cuartil o *Q3*:

24, 25, 27, **27**, 29, 30, 32

Así que



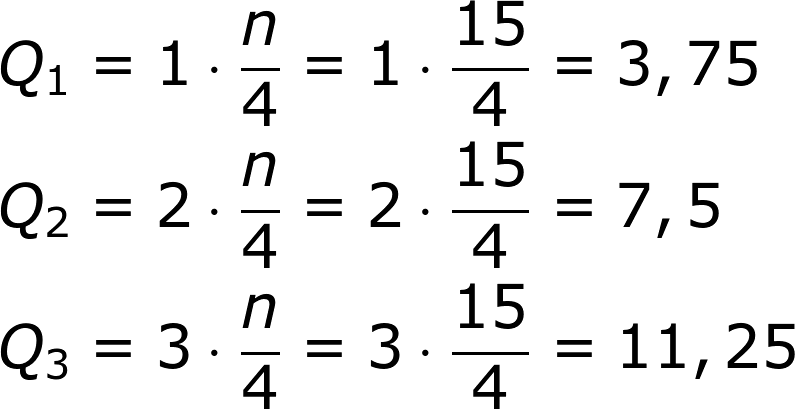
Entonces la distribución de los datos según las medidas de posición cuartílica para los 15 datos queda:

7, 10, 12, **15**, 16, 17, 18, **19**, 24, 25, 27, **27**, 29, 30, 32

*Q1 Q2 Q3*

Con lo que se concluye que la mitad de los datos está entre 15 y 27

En caso de aplicar la fórmula de los cuartiles en el mismo ejemplo, se obtiene:



Entonces el **primer cuartil** ***Q1* se ubica en la** posición 3,75, es decir, es el cuarto valor, que en este caso es 15. Ese valor corresponde a la mediana de las observaciones situadas a la **izquierda de la mediana de la totalidad**, que en este caso, es la mediana de las 7 primeras observaciones.

El segundo cuartil ***Q2*** ocupa la posición 7,5 de entre las 15 observaciones ordenadas; en este caso, es el número 19, que coincide con la que ya se había encontrado tomando el valor central del conjunto ordenado de datos.

Finalmente, el **tercer cuartil** ***Q3* se ubica en la** posición 11,25, es decir, es el doceavo valor de la lista ordenada, que en este caso es 27. Ese valor corresponde a la mediana de las observaciones situadas a la **derecha de la mediana de la totalidad**, que en este caso, es la mediana de las 7 últimas observaciones.

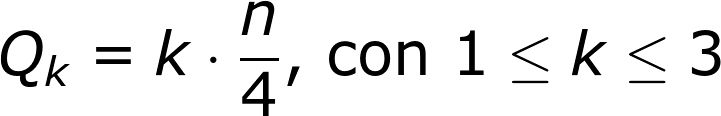
Con cualquiera de los dos métodos se obtienen los cuartiles:

7, 10, 12, **15**, 16, 17, 18, **19**, 24, 25, 27, **27**, 29, 30, 32

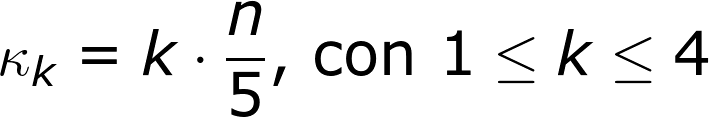
*Q1 Q2 Q3*

**Los quintiles, los deciles y los percentiles**

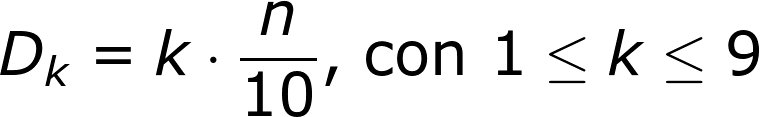
Los **cuartiles** son **3 valores** de la variable estadística que dividen los *n* datos en **4 partes iguales, que se representan con la letra Q y el subíndice correspondiente, que varía entre 1 y 3**. Los cuartiles dividen a la distribución en 4 partes, donde cada cuartil representa un 25 % de la muestra, es decir, una cuarta parte. Para calcular la posición en la lista ordenada de los datos que corresponden a cada cuartil se aplica la ecuación:



De manera análoga, los **quintiles** son los **4 valores** que dividen los *n* datos en **5 partes iguales y se representan con la letra Kappa y el subíndice correspondiente**. Los quintiles dividen a la distribución en 5 partes iguales, donde cada quintil representa un 20 % de la muestra, es decir, una quinta parte. El cálculo de la posición de los quintiles sobre la lista ordenada de datos es:



Por su parte, los **deciles** son los **9 valores** que dividen los *n* datos en **10 partes iguales. Análogamente a las medidas de posición previas se representan con una letra, en este caso la D, y el subíndice correspondiente, que varía entre 1 y 4**. Los deciles dividen a la distribución en 10 partes iguales, donde cada decil representa un 10 % de la muestra, es decir, una décima parte. El cálculo de la posición de los deciles sobre la lista ordenada de datos es:



Finalmente, se conocen como **percentiles** los **99 valores** que dividen los datos en 100 partes iguales. La representación en este caso usa la letra P y el subíndice correspondiente, que varía entre 1 y 99. Los percentiles dividen a la distribución en 100 partes iguales, donde cada percentil representa un 1 % de la muestra, es decir, una centésima parte. Para calcular la posición en la lista ordenada de los datos que corresponden a cada percentil se aplica la ecuación:



[SECCIÓN 2] **3.4 Aplicaciones de la estadística unidimensional**

En estadística la utilidad que tienen las medidas de tendencia central y no central y las medidas de posición se relacionan tanto con la necesidad de organizar y representar los datos, como con la de hacer inferencias y tomar decisiones apoyadas en los resultados obtenidos.

En la organización de los datos y en la presentación de su distribución conviene el apoyo de software, sobre todo si se trata de una gran cantidad de datos. Se usan medidas de variabilidad (rango, desviación estándar y varianza), medidas de centralidad (media aritmética, mediana, moda), medidas de no centralidad (cuartiles, quintiles, percentiles), diagramas (de cajas y bigotes), reconocimiento de valores atípicos, etc. Todo ello se hace para reconocer la distribución de los datos, bajo la comprensión de que lo importante de la representación elegida es la posibilidad de visualización de la variabilidad.

Una de las más importantes formas de presentación de la distribución de un conjunto de datos de una población es el diagrama de caja. Además, varios diagramas de caja provenientes de muestras aleatorias de una misma población permiten estudiar la variabilidad muestral y la variabilidad del azar.

Así, en el estudio estadístico no solo es importante hacer uso de lenguaje formal para expresar población, muestra, frecuencias, representatividad, variación y centralidad, sino para comprender los cambios que se dan al comparar diferentes datos provenientes de las muestras aleatorias

Entonces, las aplicaciones de la estadística unidimensional permite establecer el comportamiento de los datos de una población o de una muestra, comparar datos provenientes de diferentes muestras de una misma población o reconocer qué tanta variabilidad introduce el azar, para de ese modo poder realizar inferencias más formales generalizando el comportamiento de la muestra a la población, según las relaciones encontradas en el análisis estadístico.

**Los diagramas de cajas**

Uno de los principales usos de los cuartiles está en que sirven como marcas para hacer la representación de la distribución de los datos en diagramas de cajas que describen de manera condensada el centro y la dispersión del conjunto de datos.

Para crear un diagrama de caja de un conjunto de datos, se requiere de cinco datos esenciales:

* La **mediana o** *Q2*, para ubicar el valor central de las observaciones.
* Los **cuartiles** *Q1* y *Q3* para ubicar al 50% central del conjunto de datos. La diferencia entre la diferencia entre Q3 y *Q1* se llama rango intercuartílico y de denota como IQR.
* Las observaciones individuales **mínima (Mín) y máxima (Máx)** para indicar la dispersión.

Estos cinco números, resumen de una distribución, se expresan en un nuevo diagrama, conocido como **diagrama de caja**. Si uno de los datos se encuentra **a más de 1,5 IQR** de los extremos de la caja, se considera que es una **observación atípica** y se representa en el diagrama de caja individualmente con un punto. De ese modo se controla que los datos atípicos no “trasladen” la distribución y por ellos es de gran importancia hacer el cálculo del IQR.

Entonces, el parámetro de identificación de datos atípicos o IQR es:



y el cálculo de los valores atípicos se realiza calculando marcas atípicas inferiores y superiores. Para valores atípicos inferiores, la marca se calcula haciendo la diferencia entre *Q1* y 1,5 IQR es decir calculando:



Para valores atípicos superiores, la marca se calcula haciendo la suma entre *Q3* y 1,5 IQR es decir calculando:



Si un dato está por debajo de la marca de datos atípicos inferiores, o por encima de la marca de datos atípicos superiores, se considera un dato atípico.

El diagrama de caja se puede dibujar en posición horizontal o vertical, a condición de que el diagrama siempre incluya una escala. El diagrama de caja queda constituido al identificar si hay datos atípicos, y ubicar sobre un segmento escalado Mín, *Q1*, *Q2*, *Q3* y Máx. Luego se hace una caja alrededor de *Q1* y *Q3*, segmentos en los intervalos (Min, *Q1*) y (*Q3*, Máx) y se marcan como puntos aislados los datos atípicos. De ese modo se completa el diagrama.

Algunos ejemplos de diagramas de caja sin datos atípicos se observan en la imagen:

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Representación de algunos diagramas de caja |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** | http://www.estadisticaparatodos.es/taller/graficas/i/bigote3.gifTomada de http://www.estadisticaparatodos.es/taller/graficas/i/bigote3.gif  http://e-ducativa.catedu.es/44700165/aula/archivos/repositorio/2000/2006/html/caja_big1.png Tomada de http://e-ducativa.catedu.es/44700165/aula/archivos/repositorio//2000/2006/html/caja\_big1.png |
| **Pie de imagen** | Representación de algunos diagramas de caja |
| **Ubicación del pie de imagen** | Inferior o lateral |

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Los diagrama de caja** |
| **Contenido** | Los **diagramas de caja** son gráficos basados en **cuartiles** que son muy útiles **para la comparación de más de una distribución en un mismo gráfico**. Están compuestos por un rectángulo, la caja y dos brazos. Es un gráfico que suministra información sobre los valores **mínimo** y **máximo**, los **cuartiles** *Q1*, *Q2* o **mediana** y *Q3* y sobre la existencia de **valores atípicos** y **simetría de la distribución**. |

**Comparación de datos de la misma población**

Por ejemplo, una fábrica está realizando un estudio para identificar la productividad de su empresa en relación con los turnos de trabajo. Para ello, han tomado los datos de la producción de lotes en cada turno, durante los días hábiles que la empresa y cada turno funcionaron durante el mes anterior. Los datos han sido ordenados en orden creciente y son los siguientes:

Lotes que se han realizado en el turno de mañana de una fábrica:

23, 25, 26, 27, **28**, 3**2**, 34, 35, 37, **40**, **42**, 44, 50, 60, **75**, **85**, 87, 88, 88, 89

Lotes que se han realizado en el turno de noche:

7, 10, 12, **15**, 16, 17, 18, **19**, 24, 25, 27, **27**, 29, 30, 32

El cálculo de cuartiles para el turno de la mañana es el siguiente:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **VALOR**  **MÍNIMO** | **Q1** | **MEDIANA**  **Q2** | **Q3** | **VALOR**  **MÁXIMO** | **IQR**  **(Q3 - *Q1*)** |
| 23 | 30 | 41 | 80 | 89 | 50 |

que puede corroborarse en la lista ordenada de datos:

23, 25, 26, 27, **28**, **32**, 34, 35, 37, **40**, **42**, 44, 50, 60, **75**, **85**, 87, 88, 88, 89

*Q1 Q2 Q3*

De la misma manera, el cálculo de cuartiles para el turno de la mañana es el siguiente:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **VALOR MÍNIMO** | **Q1** | **MEDIANA**  **Q2** | **Q3** | **VALOR MÁXIMO** | **IQR**  **(Q3 - *Q1*)** |
| 7 | 15 | 19 | 27 | 32 | 12 |

que también puede corroborarse en la lista ordenada de datos:

7, 10, 12, **15**, 16, 17, 18, **19**, 24, 25, 27, **27**, 29, 30, 32

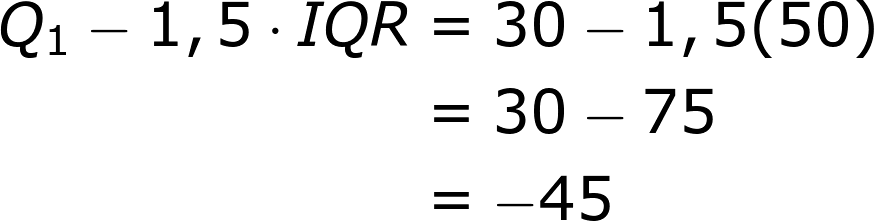
*Q1 Q2 Q3*

El IQR del turno de mañana es de 80 – 30 = 50 lotes producidos, mientras que el IQR del turno de noche: 27 – 15 = 12 lotes producidos.

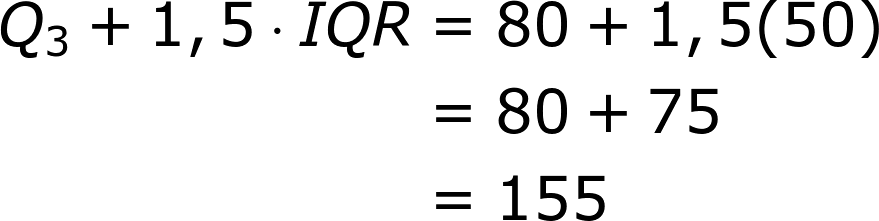
Para la construcción del diagrama de caja se dibuja un eje con la escala que se quiera utilizar. A continuación se construye un rectángulo o caja, cuyo lado izquierdo y derecho (o inferior y superior si la caja es vertical) van del primer al tercer cuartil. Por tanto, el ancho (o altura, si la caja es vertical) de la caja es la amplitud del 50 % de los datos centrales.

Debido a que si un punto se encuentra **a más de 1,5 IQR** de los extremos de la caja, se considera que es una **observación atípica**, se debe identificar si en el conjunto de datos algún dato es atípico.

Para identificar si hay valores atípicos en el turno de la mañana se calculan la marca inferior y la marca superior. La marca inferior corresponde a:

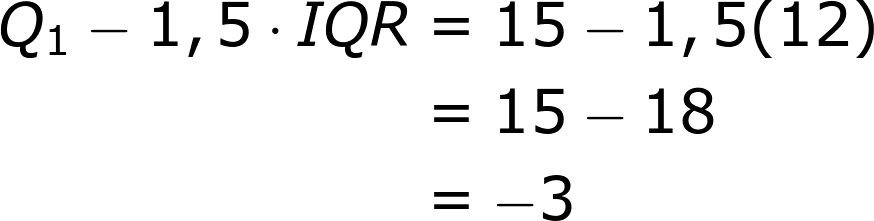


mientras que la marca inferior es:

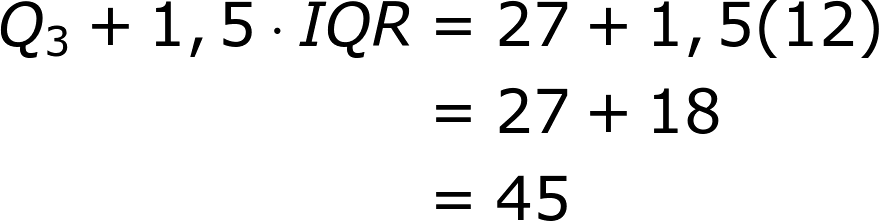


Así que como en el conjunto de datos para el turno de la mañana no hay valores inferiores a la marca atípica inferior, es decir a -45, ni datos superiores a la marca atípica superior, es decir a 155, no hay datos atípicos en el turno de la mañana.

Tampoco hay datos atípicos en el turno de la noche, donde las marcas atípicas inferior y superior son:



y



ya que en el conjunto de datos para el turno de la noche ningún dato es menor que -3 ni superior que 45.

Entonces, por último, se extienden segmentos perpendiculares a los lados derecho (superior) e izquierdo (inferior) hasta los valores máximo y mínimo respectivamente, que no sean observaciones atípicas. Los datos atípicos, de haberlos, se marcan como puntos.

Finalmente el diagrama queda:

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | 4º ESO/Matemáticas/La estadística/Las medidas de posición/Los valores no centrales |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** | 4º ESO/Matemáticas/La estadística/Las medidas de posición/Los valores no centrales    **OJO!!! Ordenaron mal los datos en el cuaderno de estudio presente en Aula Planeta y por eso algunas medidas cuartílicas son erróneas** |
| **Pie de imagen** | Diagramas de caja en un mismo gráfico para comparar el número de lotes realizados por el turno de mañana y el turno de noche. |
| **Ubicación del pie de imagen** | Inferior |

El análisis de la distribución permite concluir que en la noche hay menor producción de lotes, pero menor variación en los resultados de dicha producción.

Puedes encontrar más información teórica acerca de la estadística unidimensional en [VER](http://www.fuenterrebollo.com/Economicas2013/unidimensional-teoria.pdf) y algunos ejercicios resueltos en [VER](http://www.fuenterrebollo.com/Economicas2013/unidimensional-ejercicios.pdf).

[SECCIÓN 2] **3.5 Consolidación**

SECCIÓN 1] **4 Estadística bidimensional**

Muchas de las variables estadísticas pueden ser estudiadas aplicando la estadística unidimensional. Sin embargo, si una variable depende de o se relaciona con otra u otras, lo mejor es hacer un estudio de las dos o múltiples variables, de manera que resulte más efectivo hacer la descripción tanto del fenómeno como de sus variaciones. Cuando un fenómeno depende específicamente de dos variables, se usa en su proceso de análisis aplicando estadística bidimensional.

[SECCIÓN 2] **4.1 Variables bidimensionales**

Cuando tenemos una población en la que **a cada individuo** le corresponden los **valores de dos variables que están relacionadas entre sí**, estamos refiriéndonos a una **distribución bidimensional**. Estos valores se representan por duplas (X, Y) que varían tomando valores (*xi, yi*), es decir que la primera variable X toma los valores *x1*, *x2*, *x3*,..., *xn* y la segunda variable Y toma los valores, *y1*, *y2*, *y3*,..., *yn*.

Por ejemplo, la estadística bidimensional podría estudiar la relación entre el tiempo que una persona invierte en desplazarse a su lugar de trabajo o de estudio, en relación con la distancia entre los sitios, la influencia que tienen las deudas de una determinada familia en sus opciones de entretenimiento, o la relación que existe entre el peso y la estatura de una población.

Es importante aclarar que al considerar dos variables de una población o muestra, no necesariamente se está tratando con una variable bidimensional. Ello sucederá a condición de que las dos variables en el mismo individuo tengan algún grado de dependencia. Por ejemplo la estatura de una persona y el tiempo de antelación o retraso con que suele llegar a sus citas no son variables dependientes; para cada una podría hacerse un estudio unidimensional de manera independiente, pero no uno bidimensional.

Por el contrario si las variables son dependientes entre sí, la estadística ofrece herramientas para estudiar cómo los cambios en una variable generan cambios en la otra o para predecir cómo se comportará una variable, conocida la otra.

Existen entonces varios tipos de dependencia entre las variables:

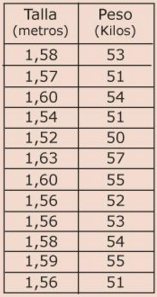
* Dependiencia funcional, que sucede cuando es posible predecir los valores de una variable a partir de los datos de la otra variable. Si hay relación funcional existe una función *f(x)* tal que *yi = f (xi)*. Por ejemplo la longitud que alcanza una barra de hierro que es calentada a una cierta temperatura o el tiempo que tarde en caer un objeto que se ha arrojado hacia el suelo desde una cierta altura son fenómenos en los que aparece dependencia funcional.
* Independencia o falta de correlación, que significa que el cambio en una de las variables no repercute en modo alguno sobre los valores que toma la otra variable, en cuyo caso se pueden analizar por separado. Por ejemplo la estatura en relación con el rendimiento académico, o el signo del zodiaco en relación con la profesión ejercida no presentan correlación.
* Dependencia estadística o correlación, que sucede cuando no es posible establecer una relación de dependencia funcional pero tampoco es posible afirmar que las variables son enteramente independientes o que les falta correlación. En ese caso se dice que las variables están en relación estadística que se conoce como correlación. Por ejemplo la tendencia a tener una expectativa de vida más larga si se realiza mayor actividad física, o el riesgo de sufrir un paro cardiaco relacionado con la cantidad de cigarrillos consumidos son fenómenos en los que hay correlación. Si bien en estos casos no se puede establecer una relación funcional en la que a tantos cigarrillos consumidos le corresponden tantas horas sin padecer un fallo cardiaco, o que a tantas horas de actividad física se aumentan tantos minutos en la expectativa de vida, la tendencia a que suceda es cierta.

[SECCIÓN 2] **4.2 Diagrama de dispersión**

En los casos en los que la dependencia entre dos variables es estadística, es decir cuando se presenta una correlación entre dos variables, la representación del conjunto de datos se puede ubicar en un plano coordenado formando un gráfico en el que los datos se tratan como coordenadas de un punto. El conjunto de todos los puntos formarían una **nube de puntos** que se conoce como **diagrama de dispersión**.

El diagrama de dispersión está conformado por duplas (X, Y), que varían tomando valores (*xi, yi*), en los que la variable X toma los valores *x1*, *x2*, *x3*,..., *xn* y la variable Y toma los valores, *y1*, *y2*, *y3*,..., *yn*.

Por ejemplo, se quiere analizar la correlación entre la variable talla (metros) y peso (kilos) de una serie de personas. Los datos se presentan en la siguiente tabla:



Tomada de 4º ESO/Matemáticas/La estadística/La estadística bidimensional

A partir de los datos se procede a ubicar las parejas ordenadas (*xi, yi*), donde las *xi* varían en el conjunto de tallas, y a cada una les corresponde un valor *yi*, que indica el peso de la persona con esa talla. De ese modo se genera la siguiente “nube de puntos” o diagrama de dispersión de los datos.

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | 4º ESO/Matemáticas/La estadística/La estadística bidimensional |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen** | Diagramas de dispersión para la correlación entre talla y peso de una población. |
| **Ubicación del pie de imagen** | Inferior |

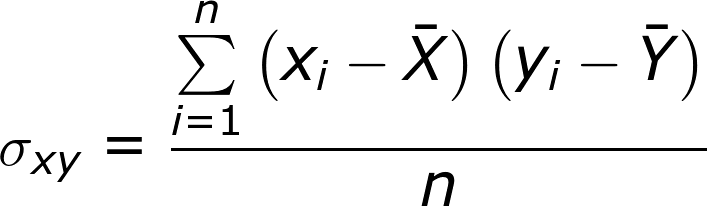
En la imagen vemos cierta correspondencia entre un peso más grande y una talla más alta, por lo que se puede decir que entre las dos variables hay correlación. Es más, en este caso es fácil observar que es posible trazar una recta que pase muy cerca de todos los puntos de la nube; el segmento de recta rojo corresponde a tal proximidad. Así se puede asegurar que, en este caso, la correlación es de tipo lineal.

[SECCIÓN 2] **4.3 Covarianza**

Aun cuando el diagrama de correlación o nube de puntos indique una cierta tendencia de comportamiento que haga presumir una correlación entre las variables, para poder asegurar que hay correlación, la estadística cuantifica el grado de correlación entre las variables. Para lograrlo, se vale de los principios aplicados en la medición de la dispersión usados en la estadística unidimensional, generalizando los cálculos de desviaciones y varianzas.

La medida que se usa para indicar el grado de correlación entre dos variables, se denomina *covarianza*. El cálculo de la covarianza corresponde a la media aritmética de los productos de las desviaciones de cada variable respecto de su media. Se representa la covarianza usando la letra griega *sigma* (σ) con subíndice XY, es decir *σxy* que se lee, “la covarianza entre las variables X y Y”.

La ecuación para el cálculo de la covarianza en el contexto bidimensional es la siguiente:



Por ejemplo, para el cálculo de la covarianza entre las variables talla y peso de una población escolar se obtienen los siguientes datos:

|  |  |
| --- | --- |
| **Talla** | **Peso** |
| ***xi*** | ***yi*** |
| 1.58 | 53 |
| 1.57 | 51 |
| 1.60 | 54 |
| 1.54 | 51 |
| 1.52 | 50 |
| 1.63 | 57 |
| 1.60 | 55 |
| 1.56 | 52 |
| 1.56 | 53 |
| 1.58 | 54 |
| 1.59 | 55 |
| 1.56 | 51 |

Para establecer la covarianza se completa la tabla con las desviaciones respecto a la media de cada variable, es decir *(xi* – *X̄ )* y *(yi* – *)* y el producto entre estas *(xi* – *X̄ )(yi* – *)*:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Talla** | **Peso** | ***(xi*** – ***X̄ )*** | ***(yi*** – ***)*** | ***(xi*** – ***X̄)(yi* –** ***)*** |
| ***xi*** | ***yi*** |
| 1.58 | 53 | 0.00583 | 0 | 0.00000 |
| 1.57 | 51 | -0.00417 | -2 | 0.00833 |
| 1.60 | 54 | 0.02583 | 1 | 0.02583 |
| 1.54 | 51 | -0.03417 | -2 | 0.06833 |
| 1.52 | 50 | -0.05417 | -3 | 0.16250 |
| 1.63 | 57 | 0.05583 | 4 | 0.22333 |
| 1.60 | 55 | 0.02583 | 2 | 0.05167 |
| 1.56 | 52 | -0.01417 | -1 | 0.01417 |
| 1.56 | 53 | -0.01417 | 0 | 0.00000 |
| 1.58 | 54 | 0.00583 | 1 | 0.00583 |
| 1.59 | 55 | 0.01583 | 2 | 0.03167 |
| 1.56 | 51 | -0.01417 | -2 | 0.02833 |

En la tabla anterior:

*X̄ = 1,5742*

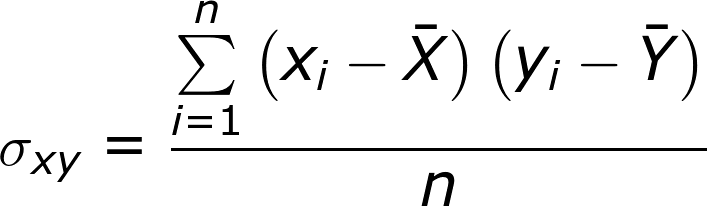
= 53

*σxy* = 0,6200 / 12 = 0,05167

Entonces la covarianza, que es la media aritmética de los valores de la última columna, entre las variables del peso y la talla es 0,05167.

[SECCIÓN 2] **4.4 Correlación**

A partir del cálculo de la covarianza se puede establecer el grado de correlación entre las variables involucradas en el conjunto de datos. La covarianza para una estadística bidimensional es:



Una primera interpretación de la covarianza se realiza por su comparación con cero. Si la covarianza es positiva, ello significa que la correlación entre las dos variables es directa. Si la correlación es negativa, significa que la correlación entre las dos variables es inversa. Finalmente, una covarianza cero significa que no existe relación de dependencia entre las variables.

Cuando la covarianza es muy cercana a cero debe tenerse cuidado, pues el signo negativo o positivo de la covarianza allí no indica correlación directa ni inversa, sino falta de relación entre las variables.

En un análisis más detallado de la covarianza, no solo por el signo sino por el número, sucede que cuanto más alta sea la covarianza, mayor es la correlación entre las variables, es decir que ambas variables crecen o decrecen simultáneamente. En el caso en que se presente una covarianza alta pero negativa, también se deduce que las variables tienen una fuerte correlación, pero inversa. Es decir que cuando una variable crece, la otra decrece y viceversa.

Otra forma de tener indicio sobre la correlación entre variables en estadística bidimensional se hace por observación del comportamiento del diagrama de dispersión. Teniendo dicho diagrama, la nube de puntos puede asemejar el comportamiento gráfico de alguna función conocida, como por ejemplo una función lineal, polinómica, exponencial etc. En esos casos hay indicio de una relación entre las variables y se dice que las variables están correlacionadas de forma lineal, cuadrática, exponencial, etc.

En el caso en que la nube de puntos no tenga la forma de funciones conocidas, la estrategia para establecer la correlación es intentar encerrar la nube de puntos entre una elipse y, en ese caso, la estrechez de la elipse es un indicador de la fuerza de la correlación. Si no se pueden encerrar los puntos de la nube en una elipse estrecha, sino cercana a una circunferencia, no existe correlación entre las variables.

Para establecer la posible correlación entre las variables “Talla” y “Peso”, contenidas en la tabla, se analiza el valor de la covarianza.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Talla** | **Peso** | ***(xi*** – ***X̄ )*** | ***(yi*** – ***)*** | ***(xi*** – ***X̄)(yi* –** ***)*** |
| ***xi*** | ***yi*** |
| 1.58 | 53 | 0.00583 | 0 | 0.00000 |
| 1.57 | 51 | -0.00417 | -2 | 0.00833 |
| 1.60 | 54 | 0.02583 | 1 | 0.02583 |
| 1.54 | 51 | -0.03417 | -2 | 0.06833 |
| 1.52 | 50 | -0.05417 | -3 | 0.16250 |
| 1.63 | 57 | 0.05583 | 4 | 0.22333 |
| 1.60 | 55 | 0.02583 | 2 | 0.05167 |
| 1.56 | 52 | -0.01417 | -1 | 0.01417 |
| 1.56 | 53 | -0.01417 | 0 | 0.00000 |
| 1.58 | 54 | 0.00583 | 1 | 0.00583 |
| 1.59 | 55 | 0.01583 | 2 | 0.03167 |
| 1.56 | 51 | -0.01417 | -2 | 0.02833 |

Entonces, como la covarianza *σxy* = 0,05167, puede decirse que hay indicio de que la correlación entre las variables peso y talla en esta población es directa. Sin embargo, como la covarianza es un valor muy cercano a cero, el indicio de correlación no es fuerte.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Interpretación de la covarianza** |
| **Contenido** | Para medir el grado de correlación entre dos variables, *σxy*, se utiliza la covarianza. Su cálculo se realiza haciendo:    La interpretación de la covarianza es la siguiente:  Si *σxy* > 0 la correlación entre las dos variables es directa, y cuanto mayor sea la covarianza, más fuerte es la correlación.  Si *σxy* < 0 la correlación entre las dos variables es inversa, y cuanto menor sea la covarianza (más alta en valor absoluto), más fuerte es la correlación.  Si *σxy* = 0 o cercana a cero, significa que no existe relación de dependencia entre las variables. |

[SECCIÓN 2] **4.5 Coeficiente de correlación lineal**

En múltiples ocasiones, debido a la cantidad y calidad de los datos, a las unidades en las que están expresadas las variables o a la cercanía de la covarianza a cero, no es posible solamente con el cálculo de la covarianza establecer una correlación estrecha entre las variables.

Además, la covarianza como medida de correlación tiene algunos problemas, entre los que se encuentran:

* Los puntos extremos de la nube determinan más fuertemente la covarianza que los datos centrales.
* Nubes o diagramas de dispersión proporcionales, como por ejemplo los obtenidos al hacer cambio de escala, tienen diferentes covarianzas, aun cuando la relación entre las dos variables es la misma.

Para subsanar estos problemas, se establece una medida adimensional que permita establecer mejor la correlación entre las variables, calculando la razón entre la covarianza bidimensional y el producto de las desviaciones estándar individuales de las variables. Dicha medida se conoce como *coeficiente de correlación lineal* que se denota con la letra *r* y cuyo cálculo es el siguiente:

http://profesores.aulaplaneta.com/DNNPlayerPackages/Package14643/InfoGuion/cuadernoestudio/images_xml/MT_10_10_formula9.gif

El cálculo de dicho coeficiente es la una de las maneras más usadas para **comparar variables bidimensionales** que estén expresadas en unidades diferentes y para lograr establecer la correlación entre ellas.

El coeficiente cumple las siguientes **propiedades**:

* Su valor está comprendido entre **−1 y 1**.
* El signo del coeficiente de correlación es el mismo que el de la covarianza:
* Si la covarianza *σxy* es positiva, la correlación es directa.
* Si la covarianza *σxy* es negativa, la correlación es inversa.
* Si la covarianza *σxy* es nula, no existe correlación.
* Si *r = 1* o *r = ‒1*, todos los puntos de la distribución están en una recta, de lo que se concluye que hay **dependencia funcional**.
* Si los valores de ***r*** **están cercanos a 1**, hay una **fuerte** y **directa** correlación lineal **positiva entre las dos variables**.
* Si los valores de ***r*** **están cercanos a −1**, hay una **fuerte** e **inversa** correlación lineal **negativa entre las dos variables**.
* Si los valores de **r** **están cercanos a 0**, la correlación es **débil**.

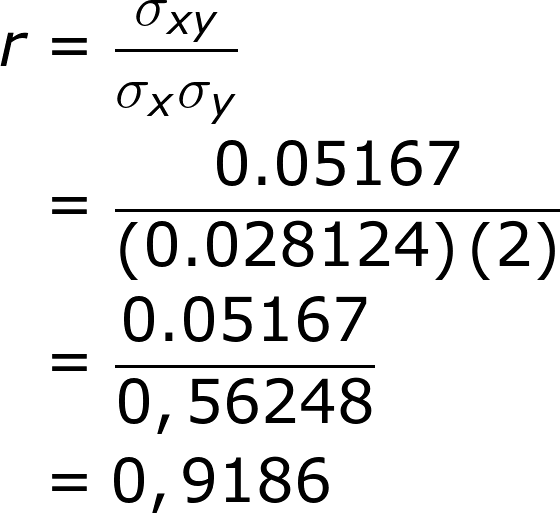
En la práctica, si el valor absoluto de *r* es mayor que 0,87 se dice que la correlación entre las dos variables es significativa.

Como ejemplo, los datos y parámetros necesarios para el cálculo del coeficiente de correlación lineal para las variables Talla y el Peso de una población son:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Talla** | **Peso** | ***(xi*** – ***X̄ )*** | ***(xi*** – ***X̄ )*²** | ***(yi*** – **y̅ *)*** | ***(yi*** – **y̅ *)*²** | ***(xi*** – ***X̄)(yi* – *y̅)*** |
| ***xi*** | ***yi*** |
| 1.58 | 53 | 0.00583 | 0.00003 | 0 | 0 | 0.00000 |
| 1.57 | 51 | -0.00417 | 0.00002 | -2 | 4 | 0.00833 |
| 1.60 | 54 | 0.02583 | 0.00067 | 1 | 1 | 0.02583 |
| 1.54 | 51 | -0.03417 | 0.00117 | -2 | 4 | 0.06833 |
| 1.52 | 50 | -0.05417 | 0.00293 | -3 | 9 | 0.16250 |
| 1.63 | 57 | 0.05583 | 0.00312 | 4 | 16 | 0.22333 |
| 1.60 | 55 | 0.02583 | 0.00067 | 2 | 4 | 0.05167 |
| 1.56 | 52 | -0.01417 | 0.00020 | -1 | 1 | 0.01417 |
| 1.56 | 53 | -0.01417 | 0.00020 | 0 | 0 | 0.00000 |
| 1.58 | 54 | 0.00583 | 0.00003 | 1 | 1 | 0.00583 |
| 1.59 | 55 | 0.01583 | 0.00025 | 2 | 4 | 0.03167 |
| 1.56 | 51 | -0.01417 | 0.00020 | -2 | 4 | 0.02833 |

De los datos de la tabla anterior, calculando las medias aritméticas correspondientes para encontrar la varianza, la desviación estándar y el coeficiente de correlación se concluye que:

Ya se conoce que en este ejemplo *σxy* = 0,05167. Las desviaciones estándar *σx* y *σy* para cada una de las variables son la raíz cuadrada de cada varianza. En este caso *σx = 0,028124*, mientras que *σy = 2*. Entonces el coeficiente de correlación lineal para las dos variables es:



Entonces, como el valor de ***r*** **está cercano a 1**, hay una **fuerte** y **directa** correlación lineal **positiva entre las variables talla y peso en la población examinada**.

[SECCIÓN 2] **4.6 Regresión lineal**

Si bien es posible que desde el diagrama de dispersión de un par de variables bidimensionales se tenga indicio de que su comportamiento se asemeja al gráfico de alguna función conocida, como por ejemplo una función lineal, polinómica, exponencial etc., es necesario establecer con claridad cuál es la curva que más se ajusta a la nube de puntos.

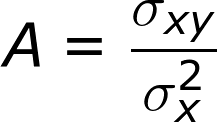
Los métodos de regresión se usan para encontrar dicha curva, tratando de encontrar una función que refleje con un alto grado de exactitud la relación entre las dos variables, con lo que posteriormente se podrán hacer inferencias acerca del comportamiento de una de las variables si se conoce el comportamiento de la otra.

La regresión lineal, como su nombre lo indica, sirve para encontrar la función lineal que mejor se ajuste a una nube de puntos. Así, se requiere determinar una recta con ecuación de la forma *y = A x + B*, que se ajuste a dicha nube.

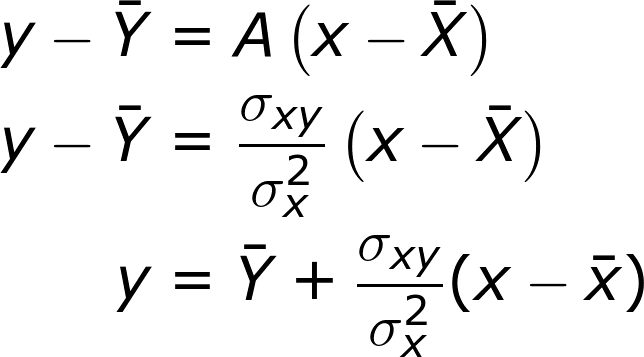
Ya que la ecuación de la recta en la forma punto - pendiente se expresa como *y - y0 = A(x – x0),* donde A es la pendiente de la recta y P = *(x0, y0 )* es un punto por el que pasa la recta, lo que se requiere es determinar esos dos elementos, es decir, un punto que pase por la recta y la pendiente de la recta.

Dos condiciones se hacen necesarias para la determinación de dicha recta:

* Que la recta pase por el punto en el que se encuentran las medias aritméticas de cada variable, es decir que la recta pase el punto (*X̄, Ῡ)*. Entonces, el punto P = *(x0, y0)* por el que pasa la recta es *P = (X̄, Ῡ )*.
* Que la distancia entre los puntos de la nube y la recta sea mínima. Ello se garantiza usando el método de mínimos cuadrados, según el cual la pendiente de la recta es:

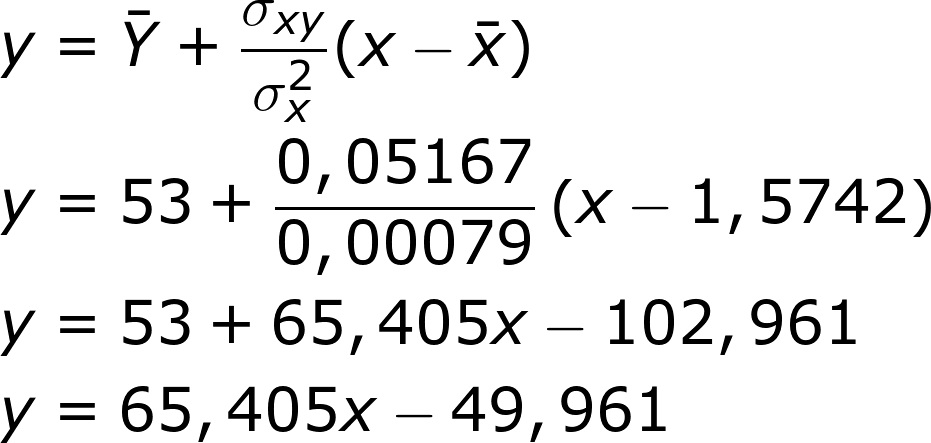


Entonces, la recta de regresión lineal de una nube de puntos para una distribución bidimensional con medias *X̄, Ῡ* está dada por:



Por ejemplo, para los datos del último ejemplo se obtuvo *X̄ = 1,5742, Ῡ = 53, σx2 = 0.00079, σxy =0.05167.*

Entonces la recta de regresión lineal que ajusta mejor los datos es:



Lo cual se verifica observando el diagrama de dispersión y la recta:

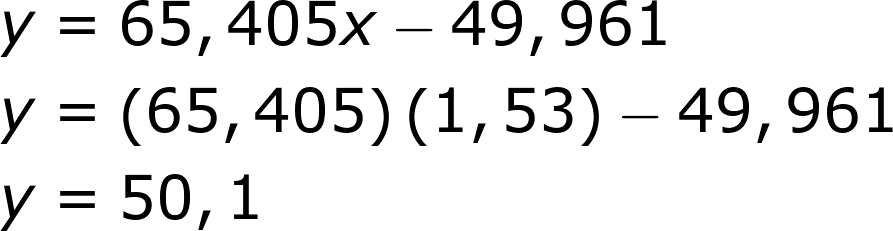
|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | 4º ESO/Matemáticas/La estadística/La estadística bidimensional |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen** | Recta de regresión lineal para la distribución de talla y peso de una población. |
| **Ubicación del pie de imagen** | Inferior |

[SECCIÓN 2] **4.7 Aplicaciones de la estadística bidimensional**

Una de las principales utilidades que tiene la estadística es la capacidad de predicción de datos desconocidos a partir de los conocidos. En ese sentido, por ejemplo la recta de regresión lineal de un par de variables correlacionadas puede utilizarse para anticipar o predecir, bien el valor de X, a partir de un valor conocido de Y, o bien el valor de Y a partir de un valor conocido de X.

El proceso por el cual se estima un valor de la variable Y a partir de un cierto valor de X, dentro de su recorrido se conoce como interpolación. La estimación de un valor de la variable Y para un cierto valor de X fuera de su recorrido se denomina extrapolación.

Por ejemplo, en el análisis de los datos de talla y peso de la población estudiada se puede establecer que una persona que mida 1,53 mt ha de pesar unos 50,1 k, pues esa es la predicción que arroja el cálculo sobre la recta de regresión lineal:



El resultado anterior se trata de una extrapolación de los datos, pues el dato de talla 1,53 no ocurrió para ninguno de los datos con los que se contaba en la población.

[SECCIÓN 2] **4.8 Consolidación**

(No es posible tener una Sección 4)

[SECCIÓN 1]**Ejercitación y competencias**

|  |  |
| --- | --- |
| **Mapa conceptual** | |
| **Código** | XX\_00\_00\_REC00 |
| **Título** | Mapa conceptual |
| **Descripción** |  |

|  |  |
| --- | --- |
| **Evaluación: recurso nuevo** | |
| **Código** | XX\_00\_00\_REC00 |
| **Título** |  |
| **Descripción** |  |

|  |  |  |
| --- | --- | --- |
| **Webs de referencia** | | |
| **Código** | XX\_00\_00\_REC00 | |
| **Web 01** | *La estadística y la propuesta de un currículo por competencias* | [*URL*](https://repensarlasmatematicas.files.wordpress.com/2014/01/s68-material-de-referencia.pdf) |
| **Web 02** | *Estadística descriptiva unidimensional* | [*URL*](http://www.fuenterrebollo.com/Economicas2013/unidimensional-teoria.pdf) |
| **Web 03** | *Estadística bidimensional* | [*URL*](http://platea.pntic.mec.es/jfgarcia/editorialsm/es4_op_b_sm_esfera/unidad15.pdf) |