|  |  |
| --- | --- |
| Título del guion | La estadística y la probabilidad |
| Código del guion | MA\_11\_06\_CO |
| Descripción | La estadística y la probabilidad son una rama de la matemática que se usan para el análisis de la información. En este tema se presentan diferentes conceptos, modelos y estrategias de análisis de variables que son usadas para tal fin. |

[SECCIÓN 1] **1 El análisis estadístico**

La estadística y la probabilidad se han convertido en herramienta fundamental para desarrollar estudios en gran cantidad de campos de trabajo y de investigación. Han permitido, a partir del diseño experimental, la recolección y análisis de información y finalmente la interpretación de los resultados obtenidos, soportar la toma de decisiones a nivel académico, empresarial y científico.

Los conceptos de **población** y **muestra** son muy importantes al involucrarse en el campo del análisis estadístico. De acuerdo a la claridad que se tenga de cada uno de ellos se genera confiabilidad en las inferencias que se contruyen.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Población, marco muestral y muestra** |
| **Contenido** | La **población** es un conjunto de elementos sobre los que se desea realizar una inferencia o construir una conclusión.  Un **marco muestral** es una lista de unidades de la población disponibles para obtener información. En muchas ocasiones el marco muestral no es igual a la población; en los casos en los cuales la población es infinita, muy grande, o imposible de contactar es necesario recurrir a listas o bases de datos existentes de elementos de la población que faciliten el contacto para obtener la información requerida.  La **muestra** corresponde a los elementos seleccionados de la población que aportan la información que permite construir inferencias acerca de la población. |

La muestra debe garantizar dos propiedades básicas que se reflejan en la credibilidad e imparcialidad de los resultados obtenidos:

**Aleatoriedad:** es el proceso de seleccionar una muestra de una población garantizando imparcialidad, razón por la cual se debe recurrir a métodos en los cuales el investigador no influye en la selección de los individuos.

El proceso al cual se recurre con mayor frecuencia para esta selección es el planteado a patir de generar números aleatorios para cada uno de los elementos del marco muestral; luego se hace una selección de aquellos individuos a los cuales se les asignó los mayores o los menores números, de acuerdo al criterio establecido por el investigador.

**Representatividad**: cuando se define claramente una población objetivo se considera si existen subgrupos que son diferentes, unos de otros, en su comportamiento y que pertenecen a la misma población. La muestra a seleccionar debe garantizar que todos aquellos subgrupos estén representados; esta característica se relaciona con la cobertura de la muestra sobre la población.

|  |  |
| --- | --- |
| **Recuerda** | |
| **Contenido** | Las **variables** son aquellas características que se quieren observar en la población y que, generalmente, se derivan de los objetivos y/o de las hipótesis planteadas en el trabajo de investigación. Al momento de definirlas es fundamental que se considere el T**ipo de variable**.   * Si se quiere observar características relacionadas con opiniones, gustos, preferencias, etc, la variable se llama **cualitativa**. * Por otra parte, si la variable a observar se puede medir en una escala numérica se llama **cuantitativa**.   Estos conceptos deben ser clarificados al momento de diseñar el proyecto de investigación ya que los métodos para analizar cada tipo de variable son diferentes. |

Al definir una variable es necesario, determinar la escala en la cual se van a medir los resultados para el caso cuantitativo, o definir los rangos de respuesta, para el caso cualitativo. Una variable no está bien definida cuando existen valores que no corresponden a una escala numérica determinada o los rangos de respuesta no incluyen resultados que se presentan en los individuos de la muestra.

[SECCIÓN 2] **1.1 Análisis de variables cualitativas**

Una variable cualitativa se puede considerar como una pregunta en la cual sus respuestas miden cualidades, características, preferencias, gustos, género de una persona, estado civil, estrato al cual pertenece, intención de voto para determinadas elecciones, preferencia por alguna comida. También se pueden considerar como preguntas cuyas respuestas se pueden clasificar en rangos pre establecidos como nivel de satisfacción (medido como: totalmente satisfecho, satisfecho, medianamente satisfecho o insatisfecho).

Por ejemplo, el nuevo gerente de una cadena de almacenes quiere determinar si existe una tendencia en la forma de pago de sus clientes. Para ello, toma una muestra de 80 compradores que llegaron a una de las tiendas y observa si cada uno de ellos hace su pago en efectivo, con tarjeta o con bonos.

[SECCIÓN 3] **1.1.1 Proporción**

El primer criterio para analizar los resultados de una variable cualitativa corresponde a la proporción de la muestra que se encuentra en cada uno de los rangos de respuesta.

Para identificar la proporción es importante tener en cuenta la construcción de una tabla de frecuencias.

|  |  |
| --- | --- |
| **Recuerda** | |
| **Contenido** | Una tabla de frecuencias corresponde al conteo de las respuestas en los rangos correspondientes. Esta tabla representa el primer informe resumido del comportamiento de la variable cualitativa a analizar.  En variables cualitativas esta tabla es única, por lo cual no se genera incertidumbre en su interpretación. |

Por ejemplo: La siguiente tabla muestra los resultados obtenidos por el gerente de la tienda, el cual quiso medir el comportamiento de la forma de pago dentro de sus clientes.

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Tabla de frecuencias de una variable cualitativa |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** | |  |  |  |  | | --- | --- | --- | --- | | Variable  Forma de Pago | Frecuencia  *f* | Frecuencia Relativa  *Fr o p* | Porcentaje  *%* | | Efectivo | 10 |  | 12.5% | | Tarjeta | 42 |  | 52.5% | | Bonos | 20 |  | 25% | | Otro | 8 |  | 10% | | TOTAL | 80 |  | 100% | |
| **Pie de imagen**  **Inferior** | La variable a analizar corresponde a Forma de Pago, por lo cual los rangos de respuesta son: Efectivo, Tarjeta, Bonos, Otro. La **frecuencia ( *f)***, la cual corresponde al número de veces que se repite cada respuesta en la muestra estudiada. Para el ejemplo, 10 de los clientes de la muestra hicieron sus pagos en Efectivo, 42 con tarjeta, 20 con Bonos y 8 con otro medio de pago. La **frecuencia relativa ( *fr)***, la cual corresponde al resultado de dividir la frecuencia *f* y el tamaño de la muestra. Se puede expresar como fracción o como decimal. |

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de proporción** |
| **Contenido** | La proporción se define como el cociente entre el conteo de uno de los rangos de respuesta y el total poblacional.  Por ejemplo, si se quiere establecer la proporción de profesores hombres, se divide el total de profesores de género masculino entre el número total de profesores. Por tanto:  Donde: X es el conteo de individuos que seleccionaron un rango de respuesta y N es el total de la población.  La proporción se puede expresar de forma fraccionaria, decimal o porcentual.  Por ejemplo, la proporción de personas propensas a adquirir una enfermedad es de 1 a 5, o de 0,20 o del 20%. |

En la tabla de frecuencias, la proporción corresponde a la frecuencia relativa.

En el ejemplo, la proporción de personas que hicieron sus pagos en efectivo es de ,o de 1 a 8, o de *0.125, o del 12.5%.* Se puede decir que 1 de cada 8 clientes hacen sus pagos en efectivo

Este valor, en variables cualitativas, corresponde a la probabilidad relativa. Es decir, la probabilidad de que un cliente pague sus compras en efectivo es de o de *0.125.*

[SECCIÓN 3] **1.1.2 Moda**

Una de las principales pretensiones que hay dentro del análisis estadístico corresponde a la identificación de tendencias que determinen el comportamiento de la población. La medida correspondiente a las variables cualitativas, y que busca este fin, es la moda

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de moda** |
| **Contenido** | La moda, , de un conjunto de datos cualitativos es el rango de respuesta de la variable con mayor frecuencia.  Corresponde al dato con mayor frecuencia en la muestra |

Para el ejemplo del gerente de la tienda, se tiene que , . En este caso se puede afirmar que la tendencia, dentro de los clientes de la tienda, es realizar el pago con tarjeta, cuya proporción es de 42 a 80, o del 0,52, o del 52,5%.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Cuando la moda no es representativa** |
| **Contenido** | Una moda significativamente cercana a los demás rangos de la variable. Es decir que varios valores obtenidos en la muestra tienen valores cercanos a la moda, sugieren que la moda no sea significativamente diferente a las demás. Por ejemplo: se midió el nivel educativo de los profesores del colegio y se encontró que 35 personas tienen formación de pregrado, 36 tienen formación de especialización y 32 tiene formación de maestría. En este caso la moda es: Especialización, sin embargo no es una tendencia de la población. |

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Cuando la moda no es única** |
| **Contenido** | Puede ser que dos valores tengan la misma proporción y corresponda a la más alta, por lo que la variable es bimodal. Es decir, es posible obtener dos valores tengan la frecuencia mayor en la muestra. En este caso no se puede hablar de la tendencia de población hacia un valor determinado.  En los casos en los cuales la mayoría, si no todos, los rangos de respuesta tengan el mismo valor máximo, la variable es amodal. Es decir, si todos los valores de variable son diferentes, no existe una tendencia de acumulación en la población. |

[SECCIÓN 3] **1.1.3 Análisis de Gráficos**

Uno de los criterios de mayor utilidad, estadísticamente hablando, para determinar tendencias, construir inferencias o pronósticos futuros del comportamiento de la población con respecto a una variable corresponde al comportamiento gráfico de la variable.

|  |  |
| --- | --- |
| **Recuerda** | |
| **Contenido** | Los gráficos que representan mejor el comportamiento de una variable cualitativa son: el diagrama circular y el diagrama de barras. |

Para el ejemplo de la variable Comportamiento de pago de los clientes de la tienda:

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Grafica circular de la variable cualitativa forma de pago |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | Hay una tendencia de los clientes de la tienda a realizar sus pagos usando tarjetas. Esta proporción corresponde a casi el 53%.  En esta gráfica los porcentajes se han aproximado al entero más cercano |

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Diagrama de barras de la variable cualitativa forma de pago |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | En estos gráficos es importante recordar que el eje horizontal no corresponde a una recta numérica. Por lo cual las barras deber estar separadas y el orden en que se ubiquen no es específico |

Cuando los valores de las proporciones son muy similares no hay una tendencia poblacional definida con respecto a dicha variable. En este caso no es posible construir conclusiones adecuadas con respecto al comportamiento de la población. Estas variables tienen alta incertidumbre en su interpretación.

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Graficas de la variable cualitativa género de una persona |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | En este caso determinar la tendencia del genero de una persona de dicha población se hace imposible, o se expone el investigador a un rango de incertidumbre bastante alto |

En el caso contrario, cuando una de las proporciones corresponde a la moda y es significativamente diferente a las demás, las conclusiones que se obtengan determinan una tendencia en la población.

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Graficas de la variable cualitativa género de una persona |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | En este caso, hay una tendencia claro hacia el género masculino. Es decir que si se esperara un nuevo miembro a la población este tendría una alta probabilidad de que sea hombre |

[SECCIÓN 2] **1.2 Análisis de variables cuantitativas**

Una de los primeros resúmenes de las variables cualitativas corresponde a la tabla de frecuencias. Este tipo de análisis se conoce regularmente como de “Datos Agrupados”, ya que el objetivo es conformar rangos, grupos o intervalos que agrupen una cierta cantidad de datos.

Es necesario aclarar que la tabla de frecuencias no es única, es posible construir muchas tablas distintas de un mismo conjunto de datos.

|  |  |
| --- | --- |
| **Recuerda** | |
| **Título** | **Cómo se elabora una tabla de frecuencias para variables cuantitativas** |
| **Contenido** | El primer paso es hallar el **rango de la distribución** restando el dato mayor del dato menor.  Luego, se determina el número de intervalos que se van a utilizar; una buena aproximación a este número se hace calculando la raíz cuadrada del número de datos en la muestra    El número de intervalos calculado debe ser un número entero y en caso de no serlo se utiliza una aproximación. Luego, se define el tamaño de cada intervalo    Finalmente, se construyen los intervalos de acuerdo a los datos recogidos, es decir, si la información obtenida contiene una cifra decimal, los límites de los intervalos también deben tenerla. |

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Qué debe tener la tabla de frecuencias** |
| **Contenido** | **La frecuencia *f***de un intervalo es la cantidad de datos que están contenidos en el rango determinado por los límites.  **La frecuencia****relativa** es el cociente entre la frecuencia y el tamaño de la muestra,  La frecuencia relativa se considera importante para la interpretación y presentación de los datos ya que proporciona un dato comparado con el total de la muestra o de la población, y no solamente el número de observaciones de un intervalo.  **La frecuencia acumulada F** de un intervalo es la sumatoria de frecuencias de los intervalos anteriores incluyendo su frecuencia.  **La frecuencia relativa acumulada** es el cociente entre la frecuencia acumulada y el tamaño de la muestra..  **Las marcas de clase** , son los puntos medios de cada intervalo. Estas marcas se utilizan para construir la gráfica del polígono de frecuencias. |

Por ejemplo, la siguiente tabla resume los resultados obtenidos por rangos de edad de una muestra de 130 aspirantes a ser gerente deportivo de un reconocido club de la ciudad:

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Tabla de frecuencias de la variable cuantitativa edad de los aspirantes al cargo |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Edad en años | *f* |  | *F* |  |  | | 25 a 30 | 7 | 0,05 | 7 | 0,05 | 27,5 | | 30 a 35 | 12 | 0,09 | 19 | 0,15 | 32,5 | | 35 a 40 | 25 | 0,19 | 44 | 0,34 | 37,5 | | 40 a 45 | 32 | 0,25 | 76 | 0,58 | 42,5 | | 45 a 50 | 26 | 0,20 | 102 | 0,78 | 47,5 | | 50 a 55 | 18 | 0,14 | 120 | 0,92 | 52,5 | | 55 a 60 | 10 | 0,08 | 130 | 1,00 | 57,5 | | Total | 130 | 1,00 |  |  |  | |
| **Pie de imagen**  **Inferior** | De la tabla se puede ver que: El 19% de los aspirantes tiene entre 35 y 40 años; El 58 % de los aspirantes al cargo tiene 45años o menos. Vale la pena aclara que si un profesor tiene 45 años, debe ser incluido en el intervalo de 40 a 45 años. |

Además de presentar un resumen de los datos obtenidos, es necesario producir un gráfico que permita describir el comportamiento de la variable. Para tal fin se utilizan dos tipos de gráficos, el histograma de frecuencias y el polígono de frecuencias

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de Histograma de Frecuencias** |
| **Contenido** | El **histograma de frecuencias** corresponde al grafico de barras. Para el caso de las variables cuantitativas, el eje horizontal representa una variable de número, por tanto existe el principio de continuidad y por ende las barras deben estar pegadas unas con otras. Es posible construir el histograma para las frecuencias y para las frecuencias acumuladas, depende del tipo de análisis que se quiera realizar. |

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Histograma de frecuencias para las edades de los aspirantes al cargo de gerente deportivo |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | El histograma permite construir una primera aproximación al comportamiento gráfico de la variable.  Si interpretación se fundamenta en la tabla de frecuencias |

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Histograma de frecuencias acumuladas para las edades de los aspirantes al cargo de gerente deportivo |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | El histograma de frecuencias acumuladas también es llamado Ojiva de la distribución.  Es una gráfica siempre creciente y determina los saltos que tiene la variable al cambiar de intervalo |

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de Polígono de Frecuencias** |
| **Contenido** | El **polígono de frecuencias** es una representación lineal del histograma en la cual se describe el comportamiento de la variable entre los intervalos. Para construir un polígono de frecuencias, se ubica en el eje horizontal la marca de clase de cada intervalo. Luego, en el eje vertical se ubican las frecuencias obtenidas en la tabla. Este gráfico es también conocido como el diagrama de líneas. Para este caso, también, es posible construir un polígono de frecuencias y un polígono de frecuencias acumuladas. |

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Polígono de frecuencias para las edades de los aspirantes al cargo de gerente deportivo |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | El polígono de frecuencias esta construir por líneas rectas, por lo cual su análisis se puede complementar con el valor de la pendiente de la recta en cada intervalo |

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Polígono de frecuencias acumuladas para las edades de los aspirantes al cargo de gerente deportivo |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | El polígono de frecuencias acumuladas permite analizar el comportamiento de la variable a través del avance en cada uno de los rangos de la variable. Es siempre creciente y permite también construir un análisis de las pendientes de las rectas que lo forman. |

[SECCIÓN 3] **1.2.1 Medidas de tendencia central: Media, mediana y moda**

Las medidas de tendencia central pretenden encontrar un punto centro o punto de acumulación de los datos. Las principales medidas de tendencia central son: el promedio aritmético o media, la mediana y la moda.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de la media aritmética** |
| **Contenido** | La **media** **aritmética o promedio** de un conjunto de datos se define como el promedio aritmético de ellos. Generalmente se denota por .  Para calcular la media: Si  *x1 , x2, . . . , xn*  es un conjunto de datos, entonces    El promedio se puede definir como el valor que representa al conjunto de datos con respecto a una variable.  Se considera la media como el valor que se espera que tenga un individuo de la población. Es por esto que se usa como elemento fundamental de pronóstico. |

Por ejemplo: Los tiempos de reacción, en minutos, de 15 personas al habérseles aplicado un estímulo externo son:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 27 | 26 | 8 | 21 | 21 |
| 14 | 6 | 18 | 17 | 19 |
| 3 | 21 | 10 | 10 | 5 |

La media de los tiempos es:

Es decir que, el tiempo medio de reacción de una persona de dicha población es de 15 minutos.

Se puede afirmar que, si una persona se somete al mismo estímulo, el tiempo que tardará en reaccionar estará alrededor de 15 minutos.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de la mediana** |
| **Contenido** | La **mediana** se define como el dato que divide un conjunto de datos en dos partes porcentualmente iguales. Es notada como  Para calcular la mediana es necesario ordenar el conjunto de datos de menor a mayor,    Luego, ubicar el punto o valor que está en el centro de ellos; para encontrar la mediana se tienen dos casos:  *Caso 1:* Si el número de datos, *n,* es impar    *Caso 2:* Si el número de datos, *n,* es par    En este caso, la mediana es el punto medio entre las dos observaciones que se ubican en el centro. |

Para el ejemplo de los tiempos de reacción al estímulo, se considera el caso 1, ya que el número de datos impar, se tiene:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | 5 | 6 | 8 | 10 | 10 | 14 | 17 | 18 | 19 | 21 | 21 | 21 | 26 | 27 |

El valor de la mediana corresponde al dato ubicado en la posición , que para el caso corresponde a 17 minutos.

Es decir que, el 50% de las personas tuvieron tiempos de reacción menores a 17 minutos, mientras que el 50% de las personas tuvieron tiempos de reacción mayores a 17 minutos.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de la moda** |
| **Contenido** | La **moda** se define como el dato que más se repite dentro del conjunto. Se denota .  Cuando se tiene una cantidad grande de datos se puede presentar con mayor frecuencia la existencia de más de una moda. En algunos casos se tiene que la moda no existe. |

La moda no se utiliza como criterio de descripción de una variable cuantitativa, por ejemplo para el caso de las medidas de los tiempos de reacción al estímulo, la moda corresponde a 21 minutos. Sin embargo la moda está determinada por solamente tres de los 15 datos. Por lo cual no es una medida que genere una tendencia clara del comportamiento de una variable.

[SECCIÓN 3] **1.2.2 Medidas de dispersión: Rango, Varianza y Desviación Estándar**

Las medidas de dispersión son valores numéricos que miden la dispersión o variabilidad entre los datos. Estas medidas se consideran como un criterio para determinar la cercanía de las observaciones. Si los datos están relativamente cerca unos de otros, con respecto a la escala en la cual se midieron, las medidas de dispersión toman valores pequeños. Si por el contrario los datos están relativamente lejanos unos de otros, las medidas de dispersión toman valores numéricos grandes.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de Rango** |
| **Contenido** | El **rango** de un conjunto de datos es la diferencia numérica entre el dato mayor del conjunto y el dato menor.    El rango se puede considerar como el intervalo cerrado en el que se encuentra la totalidad de los datos; entre más pequeño sea su valor, más cercanos estarán los datos. |

La **varianza** es una medida que pretende establecer la cercanía de cada uno de los datos con respecto a la media. Se utiliza como medida de confiabilidad del promedio.

Si su valor es pequeño, la mayoría de los datos están cerca entre sí, por lo cual el promedio es un buen individuo típico del conjunto, se constituye en un muy buen pronóstico para la variable.

Si su valor es grande, los datos no son tan cercanos, existe mucha dispersión entre ellos, por lo cual el promedio no es un buen individuo típico. Si se usara como pronóstico estaría cargado de incertidumbre.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de desviación** |
| **Contenido** | Primero, es necesario calcular las desviaciones de cada uno de los datos.  Una **desviación**  se define como la distancia que hay de cada dato al valor de la media.  Se puede concluir que si el valor de la desviación es negativo, entonces el dato correspondiente es menor que la media, y si la desviación es positiva entonces el dato correspondiente es mayor que la media. |

Ya que la media o promedio es un punto de equilibrio de los datos, la suma de las desviaciones siempre es cero.

Es por ello que para determinar una distancia promedio, se usa el valor de las desviaciones cuadradas, de tal forma que la suma no sea cero.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de Varianza muestral** |
| **Contenido** | La **varianza muestral** de un conjunto de datos *x1 , x2, . . . , xn*, notada como *S2* se define como la media aritmética o promedio de los cuadrados de las desviaciones.  Se calcula así: |

Por ejemplo, se midió el tiempo, en minutos, que tardó un medicamento en aliviar una determinada dolencia al aplicar dos medicamentos diferentes. En cada uno de los siguientes casos se registraron los tiempos para 5 personas.

|  |  |
| --- | --- |
| **MEDICAMENTO 1** | **MEDICAMENTO 2** |
| 1, 2, 1, 0.5, 0.5 | 3, 2, 0.5, 0.2, 1.3 |

El tiempo medio de reacción es:

|  |  |
| --- | --- |
| **MEDICAMENTO 1** | **MEDICAMENTO 2** |
|  |  |

Para el caso del medicamento 1 se espera que el tiempo de reacción de un paciente al que se le administre el medicamento sea de 1 minuto, mientras que para el caso del medicamento 2 es de 1.4 minutos.

Las varianzas para cada caso son:

**MEDICAMENTO 1**

**MEDICAMENTO 2**

De los resultados se puede afirmar que el medicamento 1 tiene menor variabilidad, es decir que el tiempo de reacción del individuo frente a éste estará cercano a un minuto; mientras que para el medicamento 2, la variabilidad es mayor, por tanto el tiempo de reacción no estará tan cerca de 1.4 minutos.

La varianza de los datos es una medida que relaciona la distancia de los datos con respecto a la media, sin embargo su interpretación en algunos contextos no es fácil por su forma de medición Para solucionar esta dificultad de interpretación se define una nueva medida de dispersión llamada la desviación estándar.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de Desviación Estándar** |
| **Contenido** | La desviación estándar muestral, notada como *S*, es la raíz cuadrada positiva de la varianza, |

Esta medida es la que usualmente se utiliza en la interpretación de los datos, sin embargo, la medida que permite generar análisis estadísticos formales es la varianza.

Para el caso de los medicamentos, se tiene que:

|  |  |
| --- | --- |
| **MEDICAMENTO 1** | **MEDICAMENTO 2** |
|  |  |

[SECCIÓN 3] **1.2.3 Diagramas de cajas y bigotes. Puntos inusuales**

El diagrama de cajas es un resumen gráfico en el que se describen varias de las características más destacadas de un conjunto de datos.

Algunas de las características que se muestran en un diagrama de cajas son: La identificación de datos inusuales y la dispersión de los datos con respecto a la mediana.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de cuartíles** |
| **Contenido** | Los cuartiles son valores de la variable que dividen el conjunto ordenado de datos en cuatro partes iguales. Cada una de estas partes contiene el 25% del total de los datos.  El primer cuartil, , es el número que a lo más el 25% de los datos es menor que él.  El segundo cuartil,, es la mediana.  El tercer cuartil,, es un número tal que a lo más el 75% de los datos es menor que él,  Para calcular los cuartiles se debe encontrar la mediana, es decir ; con los datos menores o iguales a  se calcula una nueva mediana que corresponde a , y con los datos mayores o iguales a  se calcula otra nueva mediana que corresponde a . |

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Cuartíles |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** |  |

El procedimiento para construir un diagrama de cajas es:

* Calcular el valor de los cuartiles. Ubicarlos en una recta numérica y construir rectángulos cuyas bases están formadas por las distancias que hay entre cuartiles

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Primer paso para la elaboración de un diagrama de cajas |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | La longitud de las cajas no necesariamente es igual. En la mayoría de los casos prácticos, estas cajas son de diferente tamaño. |

* Luego se define una medida de dispersión, llamada *rango intercuartílico*, relacionando los cuartiles uno y tres.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de Rango Inter cuartílico** |
| **Contenido** | El **rango intercuartílico** o cuarta dispersión es la diferencia entre el cuartil tres y el cuartil uno  Este rango contiene el 50% de los datos de la muestra. |

* Para la construcción de los bigotes se toman varios criterios de acuerdo al tipo de variable o a las intenciones del investigador. Sin embargo, en la mayoría de los casos la longitud de los bigotes corresponde a:
* Una vez determinada esta medida de dispersión se construye una línea desde el valor del primer cuartil hasta el valor , y de la misma forma, desde el tercer cuartil hasta

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Construcción del primer bigote |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | La longitud de los bigotes debe ser la misma en cada uno de los extremo |

Los datos que están incluidos en esta longitud son datos que se consideran dentro de un rango permitido, ya que están cubiertos dentro del rango calculado.

Finalmente, se construyen los segundos bigotes conservando la longitud calculada en el paso anterior.

Es decir, el segundo bigote izquierdo se traza desde  hasta . Mientras que el segundo bigote derecho se traza desde hasta .

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Diagrama de cajas y bigotes |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | En algunas ocasiones puede suceder que coincidan dos de los valores de los cuartíles, por lo cual, existirá tan solo una caja.  Es posible que el valor de los tres cuartiles sea el mismo. Es decir que no existirán cajas. Solamente una línea que corresponde a la mediana. En este caso la varianza es muy pequeña |

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de Datos o Puntos Inusuales** |
| **Contenido** | Un dato se llama inusual o atípico si esta fuera del diagrama de cajas y bigotes.  Son valores que influyen de manera significativa en el cálculo de las medidas de tendencia central y de dispersión que caracterizan la variable. Estos datos atípicos corresponden a valores mal tomados en el momento de la obtención de la información, o a individuos dentro de la muestra o la población que no son representativos y que por el contrario se alejan mucho de la característica que se está midiendo. |

Un ejemplo de construcción del diagrama de cajas y bigotes es el siguiente:

El reporte del número de llamadas que recibió un centro de atención al cliente de un banco, durante los últimos 20 días es:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 194 | 181 | 153 | 159 | 217 | 172 | 168 | 220 | 150 | 156 |
| 123 | 176 | 212 | 202 | 219 | 205 | 638 | 224 | 180 | 176 |

Los valores necesarios para la construcción del diagrama son:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cuartíl 1 | Mediana | Cuartíl 3 | R. Intercuartílico | Longitud del bigote |
| 165,75 | 180,5 | 213,25 | 47,5 | 71,25 |

El valor mínimo del diagrama es 38 y el valor máximo es 355.75. El diagrama correspondiente es:

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Diagrama de cajas y bigotes para el número de llamadas |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** | 50  100  150  200  250  300  350  400  450  500 |
| **Pie de imagen**  **Inferior** | Ya que existe un valor mayor a 355.75, entonces, existe un dato inusual en la muestra. Este valor corresponde a 638 |

[SECCIÓN 3] **1.2.4 Coeficiente de variación**

El coeficiente de variación se puede considerar como una medida de dispersión o de variabilidad. Su propósito es comparar la media con la variabilidad que presenta la variable.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Definición de Coeficiente de Variación** |
| **Contenido** | El **coeficiente de variación** se define como la relación porcentual que existe entre la desviación estándar y la media. |

El coeficiente de variación no tiene unidades. Se utiliza expresado como porcentaje y puede ser un valor negativo, siempre y cuando el valor de la media sea negativo.

Una forma de interpretar el coeficiente de variación corresponde a su valor. A mayor valor, significa que hay mayor heterogeneidad de los valores de la variable, a menor valor significa que hay mayor heterogeneidad de los valores de la variable.

Uno de los ediles de la zona quiere conocer el comportamiento del gasto en telefonía celular en los habitantes de uno de los barrios. Para tal fin preguntó a 12 padres cabeza de familia la cantidad de dinero, en miles de pesos, que pagó por concepto del uso de la línea celular. Los resultados son:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 21 | 18,5 | 50 | 13,5 | 23 | 19,5 | 45,5 | 37 | 25 | 12,3 | 125 | 9,5 | 16 | 14 | 32,5 |

Los valores de la media y la desviación estándar son:

,y,

Por lo cual, el coeficiente de variación es:

[SECCIÓN 2] **1.3 Análisis de tendencias**

Uno de los aspectos más importantes en el análisis estadístico está en construir inferencias acerca de los resultados obtenidos de la muestra. Sin embargo, es necesario complementarlos con el uso de técnicas combinadas que permitan un análisis en conjunto.

A continuación se relacionan algunos criterios que cumplen con este propósito.

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Relación entre el diagrama de cajas y las medidas de tendencia central** |
| **Contenido** | Si las medidas de tendencia central son iguales, media, mediana y moda, el diagrama de cajas es simétrico con respecto a la mediana. |

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Diagrama de cajas para medidas de tendencia central iguales |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | Este caso es ideal y en pocas ocasiones es posible encontrarlo en un caso en un contexto real |

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Relación entre el diagrama de cajas y las medidas de tendencia central** |
| **Contenido** | Si la mediana es mayor que la media, entonces la distribución de los datos está cargada hacia la derecha. En este caso se dice que la distribución esta sesgada hacia la derecha |

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Diagrama de cajas para mediana mayor que la media |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | La forma del diagrama de cajas está, e la mayoría de los casos, por el valor de la mediana |

|  |  |
| --- | --- |
| **Destacado** | |
| **Título** | **Relación entre el diagrama de cajas y las medidas de tendencia central** |
| **Contenido** | Si la mediana es menor que la media, entonces la distribución de los datos está cargada hacia la izquierda. En este caso se dice que la distribución esta sesgada hacia la izquierda |

|  |  |
| --- | --- |
| **Imagen (fotografía, gráfica o ilustración)** | |
| **Código** |  |
| **Descripción** | Diagrama de cajas para mediana menor que media |
| **Código Shutterstock (o URL o la ruta en AulaPlaneta)** |  |
| **Pie de imagen**  **Inferior** | La forma del diagrama de cajas está, e la mayoría de los casos, por el valor de la mediana |

Las relaciones planteadas se aplican de manera similar para la forma del histograma y el polígono de frecuencias.

[SECCIÓN 2] **1.4 Consolidación**

Actividades para afianzar lo que has aprendido en esta sección.

|  |  |
| --- | --- |
| **Practica: recurso nuevo** | |
| **Código** | RM\_01\_01\_CO |
| **Título** | Refuerza tu aprendizaje: El análisis estadístico |
| **Descripción** | Actividades sobre el análisis estadístico |