La estadística para la exploración y predicción de datos

Con frecuencia estamos interesados en obtener razonamientos válidos respecto a datos de un grupo grande de personas u objetos, pero nuestra capacidad humana para analizar al mismo tiempo grandes cantidades de datos es limitada, por lo cual se utilizan las herramientas estadísticas.

La expresión estadística puede considerarse como la colección de datos numéricos, resultado de observaciones clasificadas y ordenadas según un determinado criterio, siendo los datos valores particulares de una variable.

Los datos se pueden dividir en cualitativos y cuantitativos. Los datos cualitativos son aquellos que no representan un valor numérico, éstos pueden ser nominales u ordinales. Los nominales son los que funcionan solamente como etiquetas mientras que los ordinales son valores que se asignan de acuerdo a un orden que contiene información sobre la intensidad del atributo. Los datos cuantitativos son lo que son intrínsecamente numéricos, se pueden dividir en discreto o continuos. Los discretos representan valores enteros y con frecuencia describen conteos mientras que los continuos se pueden producir cualquier valor en un intervalo.

En el desarrollo de los métodos estadísticos la población se define como el conjunto de sucesos, personas, objetos, etc. Que le interesan a la persona que hace el estudio. La población según el tamaño puede ser finita o infinita. El concepto de infinita solo existe en teoría, ya que en la práctica no encontraremos aplicación a poblaciones de elementos infinitos como, por ejemplo, las estrellas de universo. Sin embargo, en la estadística matemática las poblaciones con un número suficientemente grande de elementos son tratados como si fueran infinitos.

Cuando la población o conjunto es muy grande, se hace difícil la observación de los caracteres a estudiar en cada uno de los elementos, debido al enorme costo que tendría la observación de toda la población y debido también al enorme trabajo y tiempo necesarios para llevar a cabo una observación exhaustiva de cada uno. Estos trabajos, inconvenientes, pueden ser superados mediante la elección de una muestra lo suficientemente representativa de la población. Una muestra se puede definir como un subconjunto de la población a la que se tiene acceso y se pueden hacer observaciones.

Una muestra aleatoria es una muestra muy representativa de la población. Se considera que cada elemento ha tenido la misma probabilidad de pertenecer a la muestra. Por lo que las conclusiones basadas en una muestra aleatoria resultan confiables.

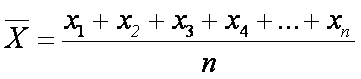
**Estadística descriptiva**

La estadística descriptiva es el conjunto de métodos y técnicas que se usan para recolectar, organizar y presentar en forma de tablas y gráficas, información numérica; incluyendo el cálculo de medidas estadísticas de tendencia centralidad y dispersión.

**Medidas de tendencia central**

Media aritmética

Es una medida de tendencia central que resulta conveniente cuando los datos no presentan valores atípicos o extremos ya que es muy sensible a estos. Se puede decir que representa el centro de gravedad de los datos. La media aritmética se calcula con la suma de los valores de los datos de la muestra divididos entre el número total de datos que conforman la muestra.



Mediana

Es una medida de tendencia central y se define como el valor que divide un conjunto de datos previamente ordenados de menor a mayor, y es el punto intermedio entre todos ellos. Esta medida es conveniente cuando la muestra presenta datos atípicos o extremos.

Se calcula ordenando los datos de menor a mayor. Si el número N de datos es impar, entonces hay un número intermedio que es el que representa la mediana. En el caso de que el número N de datos sea par, entonces hay dos datos intermedios, por lo que se calcula la media entre ellos.

Moda

En un conjunto de datos de una muestra la moda es el valor que ocurre con mayor frecuencia, por lo que es el valor más representativo.

Medidas de dispersión

Mientras que las medidas de tendencia central describen el comportamiento de los datos en una muestra o conjunto de datos, las medidas de dispersión proporcionan información acerca de cómo están distribuidos o dispersos los valores con relación a la tendencia central.

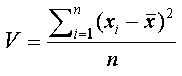
Rango

En todo conjunto de datos hay valores extremos, uno menor y otro mayor, la diferencia entre estos se le llama rango. Esta medida de dispersión es la más fácil de obtener, sin embargo se usa poco debido a que es muy influenciable por la presencia de valores extremos de poca frecuencia, lo cual conduce a apreciaciones falsas.

Varianza

Es una medida de dispersión que refleja la distancia al cuadrado de los valores en un conjunto con el valor esperado o media.

Para calcular la varianza es necesario contar con el valor de la media. Se obtiene la diferencia de cada valor del conjunto con la media y se eleva al cuadrado. Se suman todos los valores obtenidos y se dividen entre el número de datos.



Desviación estándar

Es la medida de dispersión más utilizada y muy objetiva. Cuanto mayor sea su valor, mayor es la dispersión de los datos, aunque no se puede determinar si se encuentra por encima o por debajo de la media aritmética.

Se calcula obteniendo la raíz cuadrada de la varianza.

Cuartiles

Son medidas de posición que permiten conocer los intervalos dentro de los cuales quedan representados proporcionalmente los elementos de un conjunto, se divide la distribución de frecuencias en 4 partes iguales, cada una contiene igual número de observaciones.

El primer cuartil corresponde el 25% y se designa con Q1. El segundo cuartil representa el 50% y coincide con la mediana. El tercer cuartil representa el 75% de las observaciones que están por debajo de él.

**Referencias**

Fuenlabrada, S. (2004). *Probabilidad y Estadística*. México: McGrawHill.

Ramírez, M. M. (s.f.). D*epartamento de Estadística y Centro de Estadística Aplicada.*Recuperado de ITAM http://allman.rhon.itam.mx/~mendoza/EyP.pdf

Salinas, H. (2012). Universidad de Atacama. Recuperado de http://www.mat.uda.cl/hsalinas/cursos/2010/eyp2/clase1.pdf