ESTADÍSTICA DESCRIPTIVA EN POCAS PALABRAS

(por jmd –matetam.com)

DEFINICIONES BÁSICAS

Estadística

Es la ciencia que estudia los procedimientos, técnicas y principios empleados para recolectar, organizar y analizar datos, la partir de los cuales se toman decisiones en situaciones de incertidumbre. Se divide para su estudio en descriptiva e inferencial.

Estadística descriptiva

La estadística descriptiva se encarga de la presentación, descripción, análisis e interpretación de los datos, los resume mediante medidas descriptivas.

Estadística inferencial

La estadística Inferencial estudia los métodos y principios del proceso de inferencia de propiedades de una población a partir de los datos de una muestra. Es decir, del proceso de lograr generalizaciones acerca de las propiedades de la población, a partir del conocimiento de una muestra. En síntesis, la estadística inferencial es un conjunto de técnicas para obtener conclusiones sobre la población a partir de una muestra.

Este proceso de inferencia está basado en la teoría de la probabilidad. Para que éstas inferencias sean válidas la muestra deben ser representativa de la población y la los datos de la muestra deben ser confiables. La validez de la inferencia depende entonces de la forma en que se eligen los elementos de la población en los que se miden las variables de interés, y de la medición de éstas debe ser controlada.

Pero incluso mediante esos controles, existe una variabilidad aleatoria y así las conclusiones deben ir acompañadas de una especificación de ese error aleatorio.

Variables, medición y datos

Una vez elegidas las variables de interés en una población, y habiendo sido especificada ésta, se debe establecer un procedimiento de medición y un método para elegir la muestra.

Una variable en una población es toda característica que varía de un elemento a otro de la población. Sobre esos elementos se mide la variable y el resultado son los datos muestrales.

Los datos son los valores observados de las variables medidas. Son el resultado del proceso de medición a que se someten los elementos de la muestra (también llamados unidades muestrales).

ORGANIZACIÓN Y REPRESENTACIÓN DE DATOS

Una vez obtenidos los datos deben ser sometidos a un tratamiento descriptivo, el cual consiste básicamente de organizar los datos en tablas y de representarlos gráficamente.

Ordenación de los datos. Es una forma de presentar los datos en un orden ascendente o descendente.

Agrupación de los datos en tabla de frecuencias.

Cuando los datos son discretos o continuos, y son muchos, los datos se agrupan dividiendo el rango en intervalos denominados intervalos de clase y se cuenta cuántos datos tiene cada uno de esos intervalos de clase. El resultado de esa agrupación se pone en forma de tabla de dos columnas, denominada tabla de frecuencias, anotando en la primera columna la clase y en la segunda la frecuencia correspondiente a esa clase.

Terminología de las tablas de frecuencias

Clases o intervalos de clase: Son los intervalos en que se divide el rango de los datos para formar clases (artificiales) de datos.

Límites de clase. Son los puntos extremos del intervalo de clase. Cada clase se describe con sus límites de clase, es decir, los puntos de la recta numérica en que empieza el intervalo (límite o extremo inferior) y en que acaba el intervalo (límite o extremo superior).

Límites Reales: Toman en cuenta los errores de redondeo en variables continuas y sirven para mantener una clase junto a las adyacentes (ejemplo: en datos de edad, una clase puede ser 18-20, pero realmente, dado que se da el dato de edad en número entero de años, los límites son 17 años y medio y 20 años y medio).

Anchura o tamaño de clase: es la diferencia entre los límites reales de una clase

Marca de Clase: Es el punto medio del intervalo de clase (los cuales deberían ser puntos realmente observados para facilidad de cálculo).

Frecuencia de clase: es el número de datos que pertenecen a la clase.

Frecuencia Acumulada (menor que): Es el número de datos menores que un límite de clase.

Frecuencia Relativa: Es la proporción que representa la frecuencia respecto al total de datos.

Frecuencia Acumulada Relativa: Es la proporción de datos menor que un límite de clase.

GRÁFICOS DE UNA DISTRIBUCIÓN DE FRECUENCIAS

Histograma

Es la representación gráfica de una tabla de frecuencias, y está formado por rectángulos cuya base es el tamaño de clase y su altura es tal que su área sea la frecuencia de clase.

Polígono de Frecuencias

Es el polígono que une con segmentos de recta los puntos de coordenadas (marca de clase, frecuencia de clase), cerrando en los extremos en una marca de clase inicial ficticia y una marca de clase final ficticia.

Ojiva

Es la representación gráfica de la distribución de frecuencias acumuladas. Se obtiene uniendo con segmentos de recta los puntos de coordenadas (límite real, frecuencia acumulada).

MEDIDAS DESCRIPTIVAS DE LOS DATOS

Son indicadores numéricos que resumen el conjunto de datos correspondientes a una variable. En cierta forma son representaciones numéricas del conjunto de datos, y complementan las representaciones gráficas de los datos.

Medidas de posición (tendencia central)

Son valores numéricos calculados en función de los datos y marcan un centro para éstos. Es decir, una medida de posición marca un punto alrededor del cual se agrupan los datos. Para su interpretación correcta, deben estar acompañadas de una medida de la variabilidad o la dispersión de los datos.

MEDIA ARITMÉTICA

Es la medida de posición más usada. Para datos no agrupados se calcula sumando todos los datos y dividiendo entre el número de datos. Para datos agrupados en una tabla de frecuencias se calcula sumando los productos f_iX_i (marca de clase por frecuencia de clase) y dividiendo entre el número de datos.

Propiedades de la media aritmética:

1. Puesto cada dato entra en el cálculo de la media cada uno de ellos afecta su valor. En particular es muy sensible a los datos mínimo y máximo.

2. La suma algebraica (i.e., respetando los signos) de las desviaciones de los valores individuales respecto a la media es cero.

3. La suma de los cuadrados de las desviaciones de los datos, respecto a cualquier número A, es mínimo si A es la media aritmética.

MODA

Es el valor de un conjunto de datos que ocurre con más frecuencia. Se considera como el valor más típico de una serie de datos. Para datos nominales es la única medida de posición que se puede obtener. (La moda puede no existir o no ser única.)

Cuando los datos están agrupados se denomina Clase Modal al intervalo con mayor frecuencia.

MEDIANA

Es el valor de la observación que ocupa la posición central de un conjunto de datos ordenados según su magnitud. Es el valor que está en el centro de la lista ordenada de datos, y si el número de datos es par se calcula como el promedio de los dos centrales. En otras palabras, la mediana es el valor de la variable que tiene el mismo número de datos a su izquierda que a su derecha.

Propiedades de la mediana

1. No es afectada por los datos mínimo y máximo (datos extremos).

2. Su cálculo obedece a un procedimiento y no a una fórmula.

3. Cuando los datos están agrupados en una tabla de frecuencias, primero hay que ubicar la clase mediana de acuerdo a la definición (la mediana deja la mitad de los datos a cada lado de ella), para después calcularla por interpolación.

CUANTILES

Son valores de la variable que dividen a los datos ya ordenados en n partes iguales.

Los cuartiles, dividen a los datos ordenados en cuatro partes iguales: Q1, Q2, Q3 (El segundo cuartil Q2 es la mediana.)

Los deciles, dividen a los datos ordenados en diez partes iguales: D1, D2,..., D9

Los percentiles o centiles, los dividen en cien partes iguales: P1, P2,..., P99.

Medidas de dispersión

Puesto que los datos son el resultado de medir una variable en los elementos de una muestra, necesariamente presentan una variabilidad. Y ésta puede ser grande o pequeña. Por ejemplo, decir que el promedio de edades de cuatro mujeres es 20 años puede llevar a falsas conclusiones si ese promedio no va acompañado de una medida de su dispersión.

La dispersión de los datos se mide generalmente respecto a las medidas de posición, y se define como el grado en que se distribuyen alrededor de un valor (generalmente la media aritmética). Considérese el caso en que las cuatro mujeres del ejemplo anterior sean tres niñas de 2, 3 y 5 años y su abuela de 70.

RANGO

El rango de los datos (también llamado recorrido) es la distancia entre los datos máximo y mínimo. Es la medida de dispersión más fácil de calcular pero tiene la desventaja de que puede conducir a falsas interpretaciones dado que los datos extremos podrían ser valores de la variable con probabilidades muy pequeñas de ocurrencia. Por ejemplo, si el promedio de hijos por familia en una muestra es de 3, el rango podría ser de 10 hijos debido a la ocurrencia del dato extraordinario de que una de las familias tiene 11 hijos y todas las demás tienen entre 1 y 5.

RANGO INTERDECÍLICO

Mide la dispersión del 80% de los datos centrales (después de eliminar o ignorar los datos menores que el primer decil y mayores que el noveno). Esto evita los datos extremos.

RANGO INTERCUARTÍLICO

También se le llama intervalo intercuartílico, es el rango del 50% de los datos centrales, después de ignorar el 25% inferior y el 25% superior.

VARIANZA

Una forma natural de medir la dispersión de los datos es calcular las desviaciones de cada uno de ellos con respecto a la media aritmética. Si se toman estas desviaciones en valor absoluto y se obtiene su promedio se obtiene la desviación media absoluta. Este mismo procedimiento se puede realizar con la mediana.

Sin embargo, la suma algebraica de las desviaciones de los datos respecto a la media aritmética es cero. Así que para darle la vuelta a esta suma cero de las desviaciones de las observaciones respecto a su media aritmética, se inventaron las desviaciones cuadráticas, es decir, el cuadrado de las desviaciones respecto a la media aritmética.

Al promedio de las desviaciones cuadráticas de los datos respecto a al media aritmética se le llama varianza, y es la más importante de las medidas de dispersión. Las razones son varias. Una de ellas es que, desde el punto de vista teórico, tiene propiedades más adecuadas para su manipulación matemática que la desviación media absoluta. Otra es que sin ignorar los signos de las desviaciones, da una medida de dispersión comparable a la desviación media absoluta.

DESVIACIÓN ESTÁNDAR

Como la varianza presenta el problema de que sus unidades son distintas a las de la media aritmética (si la media estuviese en metros, la varianza estaría en metros cuadrados), se ideó la desviación estándar o típica. La desviación estándar es la raíz cuadrada de la varianza –y eso resuelve el problema de las unidades.

Medidas relativas de dispersión

Cuando se necesita comparar dos o más series de datos a veces no es posible hacerlo con las medidas absolutas, ya sea porque las unidades son diferentes o porque tienen diferente media. Por esta razón es conveniente usar medidas relativas de dispersión, definidas como la razón de la dispersión absoluta entre la media aritmética.

COEFICIENTE DE VARIACIÓN

Es la medida de dispersión relativa más usada y se define como el cociente de la desviación estándar entre la media aritmética.

MEDIDAS DE ASIMETRÍA (SESGO) Y CURTOSIS

MEDIDAS DE SESGO O ASIMETRÍA

En las distribuciones cuyo histograma no es acampanado, es decir, las que se alejan de la ley de Gauss, de normalidad, la concentración de los datos no se presenta en el centro sino que tiende a concentrase a un lado de la media, es decir, la distribución no es simétrica.

En esos casos conviene calcular dos medidas adicionales de posición: la de asimetría y la de curtosis. Las medidas de asimetría muestran si en la distribución hay concentración de datos en un extremo. A este comportamiento de los datos se le llama sesgo. Es sesgo positivo si la concentración es a la izquierda, y negativo si a la derecha (la campana de Gauss es insesgada).

Las medidas de asimetría son útiles para ponderar la hipótesis de normalidad en las poblaciones estudiadas.

MEDIDAS DE CURTOSIS

La curtosis se define como el grado de “aplanamiento” de una distribución de frecuencias con respecto a la campana de Gauss. A la campana de Gauss se le llama mesocúrtica. Si la distribución es más aplanada que la campana de Gauss se dice que es platicúrtica, y si es más “picuda” se le llama leptocúrtica.

EPÍLOGO

Después de haber leído lo anterior, puede ser que el lector haya logrado visualizar que la terminología y métodos de la estadística descriptiva no son demasiados. Pero tampoco son pocos. Así que lo que sigue es entrar a los detalles de los métodos.

En resumen, lo primero que hay que hacer con los datos es organizarlos en tabla de frecuencias. Después representarlos gráficamente. El tercer paso es calcular sus medidas de centralización y de dispersión. Finalmente, el sesgo y la curtosis no son tan importantes en esta etapa descriptiva pues se pueden ver del histograma.