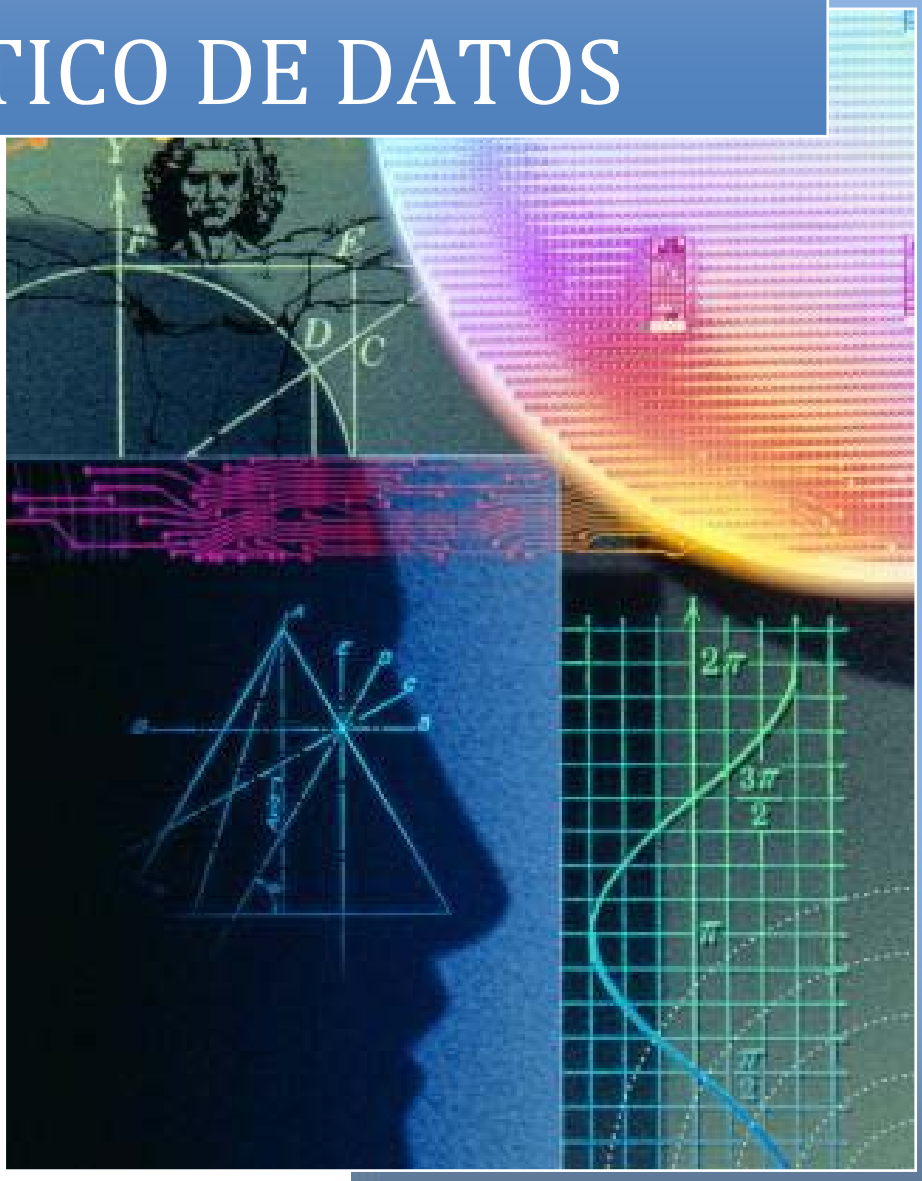
2009

CONCEPTOS FUNDAMENTALES PARA EL ANÁLISIS ESTADÍSTICO DE DATOS



Prof. Dimas Sulbarán

Conceptos Fundamentales Para el Análisis Estadístico de Datos

Dimas Sulbarán
dimas.sulbaran@gmail.com
Universidad Central de Venezuela
Escuela de Psicología

Resumen

Partiendo de las necesidades de un curso que intenta cubrir de manera razonada aquellos conceptos y métodos que se consideran básicos e imprescindibles para su posterior aplicación en cualquier campo. Este trabajo no pretende ser una guía exhaustiva que haga énfasis en teoremas y demostraciones. De manera intencionada se ha desarrollado una revisión y compilación de los aspectos fundamentales, para el dominio del discurso estadístico, que ha incluido sólo aquellos conceptos estadísticos que pueden ser de aplicación regular para cualquier investigación fundada en datos numéricos. Donde una nutrida referencia bibliográfica puede orientar al lector interesado sobre libros más específicos. También de forma intencionada se han excluido todos aquellos conceptos que, aunque muy interesantes, desde un punto de vista aplicado pueden generar una innecesaria confusión. El resultado es un documento de corta extensión, preciso y básico, para el análisis de datos cuantitativos con el auxilio de herramientas informáticas.

Palabras claves: metodología, estadística aplicada, ciencias sociales.

Índice

Introducción

La estadística es una ciencia formal y por tanto con base matemática, que enmarca a un conjunto de procedimientos diseñados para la recolección, análisis e interpretación de datos, que busca explicar condiciones regulares en fenómenos de tipo aleatorio. Es una herramienta que apoya a una amplia variedad de ciencias empíricas desde la física hasta las ciencias sociales (medicina, biología, psicología, sociología, economía, etc.) y es usada fundamentalmente para la toma de decisiones en áreas de investigación académica y comercial, tanto en instituciones públicas como privadas.

La intención de este trabajo es la de llevar información relevante y concisa al participante que se pregunta ¿Para qué sirve el análisis estadístico? En este sentido se ha hecho un esfuerzo por incorporar algunos conceptos básicos de estadística que le permitirán dominar el lenguaje fundamental dentro del discurso de la materia. Así mismo, debido a la importancia que tiene en un curso de análisis de datos "cuantitativos" o numéricos se atendió al concepto de medición y sus problemas fundamentales, la confiabilidad y la validez y los niveles de medición.

La tercera parte de este trabajo aborda de forma específica el análisis estadístico de los datos y por esta vía el análisis descriptivo e inferencial. Con relación al primero trata lo referente a las frecuencias, los cuantiles, estadísticos de tendencia central, variabilidad y forma de la distribución. Cerrando con los puntajes transformados z. Con respecto a la segunda, se han incorporado algunas de las pruebas más representativas como lo son las pruebas de contraste para dos grupos (independientes y relacionados), así como para más de dos grupos.

No siendo el interés del autor hacer una obra que equipare a manuales especializados en el área, siendo más bien nuestro propósito servir como una guía introductoria para aquellas personas que aun no ha formado altos niveles de experticia, el cual entre sus necesidades abriga una exposición sencilla de los conceptos relevantes para comprender el discurso estadístico. De forma responsable, el autor asume las limitaciones en este sentido y ha remitido al lector acucioso, interesado en el área, a una serie de prestigiosas referencias que le permitirán ahondar de forma pertinente en los temas que aquí sólo se han tratado de introducir.

La estadística

La estadística es una ciencia formal y por tanto con base en modelos matemáticos que trata los problemas con relación a la recolección, almacenamiento, organización, análisis e interpretación de datos numéricos, que funciona como una herramienta para conocer y explicar condiciones regulares en fenómenos de tipo aleatorio. Algunos autores han considerado que: "la estadística es la teoría y el método de analizar datos cuantitativos obtenidos de muestras de observaciones para estudiar y comparar fuentes de varianza de los fenómenos, para ayudar en la toma de decisiones, para aceptar o rechazar relaciones hipotetizadas entre los fenómenos, y para contribuir en la extracción de inferencias confiables a partir de observaciones empíricas" (Kerlinger y Lee, 2002 p. 232).

La estadística se suele dividir en función de una serie de criterios, donde el alcance o los objetivos del análisis estadístico define las dos categorías más gruesas, a saber: descriptiva e inferencial. Una segunda categoría considera la naturaleza del análisis en función de la atención a las relaciones entre variables y el número de variables implicado. En este sentido, se puede clasificar a la estadística en univariada, bivariada y multivariada. Finalmente, el tipo de análisis estadístico con el cual se opere deberá atender a las características de la distribución de las variables implicadas y determina las tipologías: paramétrica y no paramétrica.

Ambas ramas (descriptiva e inferencial) comprenden la estadística aplicada. Hay también una disciplina llamada estadística matemática, la cual se refiere a las bases formales de la materia. Los métodos estadísticos son un instrumento en la investigación que sirven para describir los datos, estudiar relaciones, causación y generar argumentos para la confirmación o el rechazo de hipótesis (Glass y Stanley, 1984).

Para qué sirve el análisis estadístico

Como se mencionara anteriormente, las ciencias pueden ser clasificadas en formales y empíricas. En las ciencias formales (lógica y matemática) no hay necesidad de entrar en contacto con el mundo real; basta con establecer un conjunto de postulados sobre entidades

abstractas y proceder a partir de ellos por deducción lógica. En las ciencias empíricas, sin embargo, esas leyes capaces de explicar el comportamiento de la naturaleza sólo pueden ser descubiertas y verificadas observando el mundo real.

Mientras que las leyes de la deducción lógica permiten llegar a conclusiones verdaderas a partir de premisas verdaderas, la generalización inductiva (propia de las ciencias empíricas) intenta ir desde lo que se considera que es verdad desde un conjunto reducido de observaciones hasta la afirmación de que eso mismo sucede en el conjunto total de observaciones de la misma clase. Bajo las circunstancias que definen la medición en ciencias sociales, las conclusiones a las que es posible llegar inductivamente requieren la utilización de una metodología en cierto sentido sensible a estas debilidades. Es precisamente la estadística, mediante el conjunto de procedimientos o herramientas englobadas bajo la denominación de análisis estadístico, quien proporciona a las ciencias empíricas esa metodología.

La aplicación más importante del análisis estadístico está relacionada con su atención al concepto de incertidumbre, entendida esta como la tendencia de un resultado a variar cuando se efectúan observaciones repetidas del mismo fenómeno, bajo condiciones semejantes. En situaciones deterministas, donde los factores son causa ontológica de otra (por ejemplo, la relación peso y velocidad de la caída), el álgebra o análisis matemático bastan para definir el fenómeno en cuestión. Sin embargo, cuando las condiciones están signadas por la variabilidad de las respuestas al mismo fenómeno bajo las mismas condiciones, es necesario recurrir al análisis estadístico para poder extraer conclusiones fiables y validas.

Conceptos básicos de estadística

La comprensión de toda ciencia comienza obligatoriamente por un proceso de reconocimiento de los términos característicos del discurso que define el marco interpretativo en cuestión. Es por tanto, que en lo sucesivo se hará una exposición de algunos de los términos más representativos del lenguaje estadístico.

Datos: son los hechos, medidas o números que han sido recopilados como resultados de observaciones; se deben reunir, analizar y resumir para su presentación e interpretación. Pueden ser cuantitativos (siempre numéricos) o cualitativos (que pueden ser numéricos o no, ya que son etiquetas o nombres asignados a un atributo de cada elemento. Por ejemplo: el sexo de una persona es masculino o femenino, pero podrían ser codificadas con 1 o 2 y en este caso, los números sólo servirían para indicar la categoría y no tendrían significación numérica).

Población (N): es el conjunto de todas las observaciones o de los elementos de interés en un determinado estudio que cumplen ciertas propiedades comunes. Este conjunto puede ser un número finito de datos o una colección grande (virtualmente infinita) de datos.

Muestreo: significa tomar una porción como representativa de una población o de un universo.

Muestra (n): es un subconjunto de la población, sin embargo, nos interesa que ese subconjunto seleccionado de la población sea representativo, esto significa que debe contener las características relevantes de la población en la misma proporción en que están incluidas en dicha población. Las muestras pueden ser probabilísticas (aleatoria simple, estratificadas, por conglomerados, etc.) o no probabilísticas (por juicio, por cuota, etc.).

Distribución muestral: se refiere a la función de probabilidad (de densidad de probabilidad) de un estadístico. Por tanto, una distribución muestral puede definirse como una distribución teórica que asigna una probabilidad concreta a cada uno de los valores que puede tomar un estadístico en todas las muestras del mismo tamaño que es posible extraer en una determinada población.

Parámetro: es cualquier medida descriptiva de una población, por ejemplo, la media poblacional.

Estadístico: es cualquier medida descriptiva de una muestra y se usa como base para estimar el parámetro correspondiente de la población. Por ejemplo, la media muestral.

Variable: es un carácter o propiedad de la muestra o de la población que es susceptible de ser observada y medida. A nivel estadístico las variables se clasifican como:

- Cualitativa: cuando la característica de estudio es no numérica; por ejemplo: la preferencia religiosa, el sexo, el color del cabello, el estado civil, etc.
- Cuantitativa: es aquella que asume valores numéricos acompañados de una unidad de medida; por ejemplo: calificaciones de un examen.
- Continua: es aquella que puede tomar cualquier valor dentro de un intervalo, por lo general los valores de una variable continua proceden de mediciones. Ejemplos: la estatura, la presión de aire en un caucho, etc.
- Discreta: es aquella que sólo puede tomar determinados valores en un intervalo, por lo general son números enteros, y suelen ser el resultado de un conteo. Ejemplo: el número de hijos de una familia, el número de asistencias, etc.

Medición

De acuerdo con Kerlinger y Lee (2002), la medición es una de las piedras angulares de la investigación. Todos los procedimientos sobre los cuales versa el análisis estadístico se sostienen en la medición, de manera que las implicaciones son obvias. La definición clásica de medición se atribuye a Stevens (1968), quien afirmó que "en su sentido más amplio, la medición es la asignación de valores numéricos a objetos o eventos, de acuerdo con ciertas reglas". El fin general de toda teoría de la medición, trátese de la ciencia que se trate, es estimar los errores aleatorios de las mediciones, pues toda medición, en mayor o menor grado, conlleva un cierto error. En todos los casos, por un lado, hay que estimar la cuantía de los errores cometidos al medir, y, por otro, hay que garantizar que la medición no es trivial, que tiene entidad explicativa y predictiva. En otras palabras, hay que comprobar que las mediciones son fiables y válidas.

En la actualidad, la medición de variables aparentemente poco cuantificables ha dejado de ser un obstáculo para la medición en ciencias sociales. Hoy en día, la medición no se concibe exactamente como la asignación de un numeral que exprese la magnitud de cierta propiedad. Medir en todo caso consiste en establecer con claridad las reglas de correspondencia para dos sistemas de relaciones: uno empírico (el de las propiedades que se desea medir) y otro formal (el de los números que se asignan en la medición). Por consecuente, dependiendo de la riqueza de las relaciones que se logren establecer entre los diferentes valores de una variable, existirán diferentes niveles o escalas de medida. Sobre este punto ahondaremos en lo sucesivo.

Confiabilidad

Bajo la denominación genérica de confiabilidad se agrupan todo un conjunto de métodos y técnicas utilizadas por los psicólogos para estimar el grado de precisión con el que están midiendo sus variables.

Este aspecto de la exactitud con que un instrumento mide lo que se pretende medir es lo que se denomina la confiabilidad de la medida. En este sentido, el término confiabilidad es equivalente a los de estabilidad y predictibilidad. Esta es la acepción que más comúnmente se le da a este término. Ahora bien, ¿cómo estiman los psicólogos el grado de error que hay en sus mediciones?

Existen varias maneras para estimar la confiabilidad de una medida. Sin embargo, no es el objetivo de este trabajo ahondar en el estudio de este tema, para un estudio más detallado de este punto se sugiere la lectura de Anastasi (1976). En esta sección abordaré tres de las más

conocidas: (a) confiabilidad de reaplicación de pruebas (test-retest); (b) confiabilidad de versiones equivalentes (pruebas paralelas); y (c) confiabilidad de consistencia interna (homogeneidad).

La confiabilidad, aun cuando no es la característica más importante de un instrumento de medición, requiere se le preste toda la atención que sea necesaria. Ciertamente, una alta confiabilidad, por si sola, no garantiza "buenos" resultados científicos. Pero, no puede haber "buenos" resultados científicos sin instrumentos confiables. En síntesis, la confiabilidad es una condición necesaria, pero no suficiente para obtener resultados de investigación que sean científicamente valiosos y socialmente útiles.

Validez

En esta sección, nos interesa estudiar la exactitud con que pueden hacerse mediciones significativas y adecuadas con un instrumento, en el sentido de que mida realmente el rasgo que pretende medir. Esta propiedad o característica de un instrumento de medición recibe el nombre de validez. Es decir, en sentido general, la validez de un instrumento tiene que ver con las preguntas siguientes: ¿qué miden los puntajes del test? y ¿qué predicen dichas puntuaciones? Al igual que en el caso anterior, solo se hará una presentación del tema. Para un estudio exhaustivo del mismo (ver Guilford, 1954; Nunnally, 1967; Anastasi, 1976; Magnusson, 1982).

Es este el problema de la Validez, concepto clave de la medición en las ciencias sociales. Que las mediciones sean fiables es una condición necesaria, pero no suficiente para que sean válidas. Se puede estar midiendo con gran precisión algo que no tiene ninguna capacidad explicativa o predictiva. No en vano los grandes debates acerca de la utilidad de los tests, las escalas y otras mediciones psicológicas y educativas se centran generalmente en torno al problema de su validez. Dentro de este marco general hay tres procedimientos clásicos y muy utilizados para recabar información empírica probatoria de la validez, denominados Validez de Contenido, Validez Predictiva y Validez de Constructo (ver Kerlinger y Lee, 2002).

Niveles de medición

Desde la tradición fundada por Stevens (1968) se han distinguido cuatro escalas o niveles de medida: nominal, ordinal, de intervalo y de razón. Las cuales se exponen a continuación.

La medida nominal, es considerada el nivel de medición más bajo y en las discusiones más conservadores se cuestiona que en realidad se trate de un nivel de medida. Esto obedece a

que la medición nominal consiste exclusivamente en clasificar en categorías a los sujetos u objetos que se desea medir, de modo que todos los sujetos u objetos clasificados dentro de la misma categoría sean equivalentes con relación a la propiedad, según la cual se les está clasificando.

La medida ordinal consiste en asignar a los sujetos, objetos o eventos medidos un número que permita ordenarlos en relación a una determinada propiedad con la cual los estamos asociando. Con esta clasificación nos adentramos en el campo de la medición propiamente cuantitativa y además de la clasificación de los elementos ganamos la capacidad de ordenarlos y reconocer una mayor o menor presencia de la propiedad en cuestión.

En la medida de intervalo, entramos en un tercer nivel de medición caracterizado por permitirnos establecer afirmaciones que trascienden la mera jerarquización y determinar de forma objetiva los cambios en nuestras escalas que se asocian con cambios idénticos en la propiedad medida entre un valor y el siguiente en todos los puntos de la distribución. Esta escala permite considerar un cero relativo, lo cual permite operar sobre operaciones aritméticas complejas.

La medida de razón, se considera el nivel de medida más alto y esto obedece a que es capaz de cubrir todas las bondades de los niveles de medida inferiores incorporando a la escala la presencia de un cero absoluto o natural.

La importancia de distinguir apropiadamente las diferentes escalas de medida radica fundamentalmente en que las técnicas estadísticas son sensibles a este tipo de definiciones. Por esta razón es de suma importancia conocer la problemática relacionada con las escalas de medida: el conocimiento de esta problemática puede contribuir a dilucidar si con los números disponibles, tiene o no sentido efectuar determinado tipo de operaciones.

El análisis estadístico de los datos

El análisis estadístico fue definido como el procedimiento por el cual se conseguía el almacenamiento, procesamiento e interpretación de los datos, con base a una serie de estrategias para la tabulación, resumen, análisis y contraste de los datos que fueron obtenidos de las observaciones a un conjunto de elementos. Este procedimiento, debe ser entendido siempre como un medio y no como un fín en sí mismo, por lo que el análisis que se haga siempre se entenderá y evaluará con referencia al marco del problema para el cual fue propuesto con la intención de generar respuestas o disminuir los niveles de incertidumbre. Con respecto a esto se

ha planteado una serie de criterios sobre los cuales debe versar cualquier estrategia de análisis, a saber: a) cuál es el objetivo de la investigación; b) cuál es el nivel de medición de las variables y c) cuál fue la estrategia de muestreo empleada.

En función de lo anterior, el investigador puede toparse con la necesidad de decidir según el alcance o los objetivos de su proyecto no sólo llevar a efecto un análisis descriptivo, sino disponerse a realizar un análisis inferencial. Por la atención a las relaciones entre las variables estudiadas preparar no sólo análisis univariados, sino bivariados o incluso multivariados. En tanto que, según la naturaleza de las distribuciones de sus observaciones debe estimar si ha de trabajar con estadísticos paramétricos o no paramétricos. En este punto, abordaremos los análisis que nos permiten alcanzar el conocimiento a nivel descriptivo de nuestras variables y nos acercaremos a los abordajes bivariados o de las relaciones entre dos variables.

El análisis descriptivo

La estadística descriptiva, como lo sugiere su nombre es aquel conjunto de procedimientos que nos permitirán alcanzar la descripción de los datos. Consiste en una serie de tareas que orienta los métodos de tabulación, organización, representación y resumen de datos originados a partir de los fenómenos en estudio. Los datos pueden ser resumidos numérica o gráficamente. En este segmento se ubican: a) de frecuencia y posición (cuantiles y percentiles); b) tendencia central (moda, mediana y media); c) de variabiklidad (rango, varianza y desviación típica), d) forma de la distribución (asimetría y curtosis); y e) transformaciones lineales (puntajes estandarizados). A nivel gráfico se ubican los histogramas, las barras, sectores o tortas, diagramas de caja, entre otros. De este modo, la estadística descriptiva sirve de herramienta para describir, resumir o reducir las propiedades de un conglomerado de datos para que sean más manejables y más comprensibles (Glasss y Stanley, 1984).

Frecuencias

El análisis de frecuencias en su acepción más simple, se puede definir como el estudio de la cantidad de veces que aparece cada uno de los valores de una variable en una observación dada. El análisis de frecuencias se puede aplicar a variables de cualquier nivel de medición y siempre será de gran utilidad en la exploración de los datos. Se puede clasificar según tres criterios, el primero es por las categorías de análisis, el segundo por la escala empleada para la representación de las frecuencias y el tercero por el carácter acumulativo de los sujetos. Las frecuencias se presentan como *tablas de distribución de frecuencias.*

El análisis de frecuencias se puede establecer en las siguientes categorías de análisis: a) simple; b) agrupado y c) agrupado por clases. En el análisis de frecuencias simple, los casos se presentan de forma ordenada ascendente o descendente con frecuencia de 1 para cada valor de la variable. En el análisis agrupado simple, los datos son ordenados según su valor en forma ascendente o descendente, recogiendo en un solo número para cada caso la cantidad (n) de apariciones. Finalmente, en los análisis de frecuencias agrupados por clases el investigador agrupa una determinada cantidad de valores en una categoría y recoge en un solo número la cantidad (n) de apariciones asociadas con dichos valores.

De acuerdo con la escala empleada para representar la cantidad de apariciones de los valores observados, se presentan dos opciones: a) absolutas y b) relativas. Las primeras nos muestran la cantidad de apariciones (n) en una escala bruta, por lo que puede mostrar valores para f dentro del rango de todos los enteros positivos, para cada uno de los valores observados de x. por su parte, las frecuencias relativas trabajan con una escala transformada para las puntuaciones de f, que se clasifican a su vez en: a) porciones, con valores entre 0 y 1; b) para porcentajes donde los valores pueden encontrarse entre 0 y 100. Frecuencia relativa de clase se simbolizan como "h_i" y "$\%h_i$" se obtiene al dividir la frecuencia de clase (f_i) entre el número total de observaciones (n), por lo que indica la proporción de la cantidad total de datos que pertenecen a una clase, la cual al multiplicarse por 100 arroja el porcentaje. Las formulas son:

$$h_i = \frac{f_i}{n} \quad \text{y} \quad \%\, h_i = \frac{f_i}{n} * 100$$

Finalmente, por el carácter acumulativo de los datos se consideran dos opciones posibles: a) frecuencias de clase, la cual corresponde con el esquema que se ha expuesto hasta ahora; b) frecuencias acumuladas de clase, implica el conocimiento de la cantidad de sujetos que han obtenido puntuaciones iguales o inferiores al mayor valor del intervalo de clase. Esta definición es válida para distribuciones acumuladas "menor que", por tanto sólo funciona a partir de variables con nivel de medición ordinal.

Las Distribución de frecuencias acumuladas *"menor que"* es una tabla donde se presentan las frecuencias acumuladas, para hallar esta distribución en una clase determinada lo que se hace es sumar la frecuencia de esa clase a la de las clases anteriores. Las distribuciones de frecuencias acumuladas nos permiten ver cuántas observaciones se encuentran por arriba o debajo de ciertos valores.

Determinación del intervalo de clase

Una vez que el número de observaciones se ha vuelto poco parsimonioso el interés del investigador se suele volcar hacia el empleo de datos agrupados por intervalos o clases. La selección del número de clases que se utilicen para el análisis depende primordialmente de la cantidad de datos que se tengan. Aunque en las ciencias duras se han planteado algunas formulas que intentan estandarizar el proceso de decisión, en la mayoría de los casos se reduce a una decisión arbitraria. Sin embargo, ciertas reglas deben aplicarse para que la decisión goce de validez y aceptación ante el resto de la comunidad académica. En términos generales, se recomienda que la distribución de frecuencias este conformada por al menos 5 clases y no más de 15 (si no existen suficientes clases, o si hay demasiadas, la información que se puede obtener es precaria). Entre las expresiones que se pueden utilizar para calcular el número de clases tenemos:

- $d \geq \text{Log } n / \text{Log } 2$ donde “d” es el número de clases y “n” el número total de observaciones.
- $\sqrt{n}$ donde “n” el número total de observaciones.
- $1+3{,}322 \text{ Log } n$ (regla de Sturges) y “n” el número total de observaciones.

No obstante estas reglas que en algunos casos funcionan como una referencia, no deben tomarse como un factor determinante o definitivo. Un ejemplo permitirá ilustrar lo absurda que puede resultar una decisión fundada en formulas estandarizadas, suponga que el número de observaciones que tenemos es 100, es un buen criterio agrupar las observaciones en 100 = 10 intervalos, pero si el número de observaciones fuese muy alto como por ejemplo n = 1000000, este segundo criterio nos da un número excesivo de intervalos (1000).

Un elemento que no captan las formulas a priori conocidas es el carácter situado y el significado del dato con el cual estamos trabajando. Por ejemplo, dos personas pueden encontrar la distribución de las notas de un curso que se define en un rango de 20 puntos desde 01 hasta 20. Ambos toman la decisión de trabajar con datos agrupados, uno de ellos considera conveniente trabajar con cuatro intervalos de clase, para diferenciar cuatro niveles de desempeño en el curso. Por su parte, el otro analista decide que la distribución debe mostrar dos grupos: el de los aplazados y el de los aprobados. Ambas ofertas son igualmente validas y en estos casos sólo se hizo uso del sentido común para determinar el número de intervalos.

De igual manera, siempre atendiendo a la perspicacia del investigador para captar las mejores decisiones según la situación, aunque anteriormente se recomendó que todas las clases tuviesen el mismo tamaño, existen casos donde esta regla no puede o no debe aplicarse; por ejemplo, si se tuviera a mano la lista de impuestos pagados por la población en un año, estas

cantidades (supuestas) pueden encontrarse en un intervalo de Bs. 0 a Bs. 10000000, aún a pesar de que se eligiesen 20 clases para la distribución de frecuencia, con intervalos de igual longitud, cada clase tendría una cobertura de Bs. 500000. Lo anterior daría origen a una situación en la que casi todas las observaciones caerían en las primeras clases, pero si ampliamos el número de clases equivalentes atentaríamos contra la parsimonia de la distribución de intervalos. En casos como este, es preferible trabajar con una distribución de intervalos no equivalentes a fin de seleccionar una escala más pequeña en el extremo inicial que la utilizada para el extremo superior. También sería posible reducir el número de clases que se requieren cuando unos cuantos de los valores son muchos menores o mucho mayores que el resto, mediante clases abiertas. Los anteriores, son situaciones que se deben evitar cuando sea posible ya que reducen el nivel de medición de nuestras variables y complican la aplicación de ciertos cálculos y análisis que puedan ser de interés.

Obtención de los intervalos de clase: el intervalo de clase es el recorrido de los valores que se encuentran dentro de una clase, es recomendable al elaborar la tabla que todas las clases tengan el mismo tamaño porque facilita la interpretación estadística de cualquier utilización posterior que se pueda hacer de los datos. El cálculo del intervalo de clase parte de una sencilla formula:

$$a = \frac{\text{Rgo. Total}}{n}$$

Donde **a** es la *amplitud del intervalo de clase;* **Rgo. Total** es el *rango total* de la distribución; y **n** es el *número de intervalos* que se han determinado. De este modo, obtenemos un valor que sirve de guía para establecer el tamaño de los intervalos, el valor numérico que obtengamos de la fórmula anterior lo podemos redondear dependiendo de nuestra conveniencia, pero en cualquier caso, se toma con un grado de aproximación no mayor a aquel con el que se registran los datos.

La definición de los intervalos de clase sigue unas reglas simples, las del proceso de categorización. En este sentido, el analista sólo tiene que responder a las siguientes condiciones: a) deben ser exhaustivas, lo cual implica que debe abarcar de manera eficiente el número suficiente de unidades de observación y; b) mutuamente excluyentes, se necesitan establecer límites claramente definidos para cada una de las clases, de manera que se eviten problemas como: El solapamiento entre clases (no debe existir duda en la ubicación de los datos en las clases) y que no se incluyan a todas las observaciones.

El lector encontrará al revisar literatura al respecto que los autores difieren en la forma en que toman los límites cuando construyen las tablas de distribución de frecuencias (básicamente

según el tipo de variable con la cual trabaje). Un modelo bastante popular toma los intervalos de tal manera que son cerrados en el límite inferior y abiertos en el superior, en forma de *intervalos semiabiertos*, a saber, si *x* es una observación o dato de una muestra cualquiera, y los valores a y b son el límite inferior y superior, respectivamente, de una clase o categoría se dice que *x* pertenece a dicha clase si y sólo si *x* es igual o mayor que a y menor que b. esto se expresa en símbolos matemáticos de la siguiente manera:

$$x \in [a,b) \;//\; a \leq x < b$$

En la notación anterior, el término [a,b) significa que se trata de un intervalo que comprende los valores iguales o mayores que a pero menores que b, con lo cual se denota a un intervalo semiabierto (cerrado por la izquierda y abierto por la derecha).

De modo que, para la determinación de las frecuencias de clase, los intervalos se expresan como intervalos *semiabiertos*, en donde el límite inferior de cada clase corresponde a la parte cerrada del intervalo, y el límite superior corresponde a la parte abierta del mismo. Esta conformación de los intervalos nos garantiza la eficiencia de los intervalos de clase: a) en primer lugar, garantiza que ningún dato de la muestra pertenezca a más de un intervalo, así como que ningún dato quede fuera de alguno de los intervalos obtenidos. Ejemplo: Los datos 23, 24, 18, 14, 20, 13, 38, 19, 16, 24, 11, 16, 18, 20, 23, 19, 32, 36, 15, 10, 40 son parte de un total de 80 datos que serán utilizados para construir una tabla de distribución de frecuencias, suponemos que los cálculos para determinar el número de clases y la amplitud ya fueron realizados dando como resultado que el número de clases es 6 y la amplitud es 5. La tabla resultante nos daría una presentación como la siguiente:

li - ls
[10 – 15)
[15 – 20)
[20 – 25)
[25 – 30)
[30 – 35)
[35 – 40)

Establecimiento de la marca de clase (xi): al trabajar con intervalos de clase se hace necesario un punto o valor representativo del intervalo. Si éste es acotado tomamos como marca de clase al punto medio del intervalo (se asume que los valores de la variable se distribuyen de manera uniforme dentro del intervalo). Se obtiene como un promedio aritmético entre los límites superior e inferior de cada intervalo de clase. La formula se define como:

$$X_i = \frac{Ls - Li}{2}$$

Donde, X_i es igual al intervalo de clase; L_s es el límite superior del intervalo de clase y L_i es el límite inferior del intervalo.

Cuantiles

El nombre genérico para un conjunto de estadísticos que cortan la distribución de frecuencias en un determinado número de partes iguales es el de cuantil y el mismo se define como el valor bajo el cual se encuentra una determinada proporción de los valores de una distribución. Dentro de las medidas más representativas de los cuantiles tenemos:

Deciles: Son aquellos valores que dividen en diez partes iguales a un conjunto de datos ordenados. Se representan por D1 , D2 , D3 , ...D9. De esta manera tenemos que:

- D1 (primer decil) es el valor por debajo del cual se encuentran como máximo el 10% de las observaciones, mientras que el 90% restante se sitúan por encima de él.
- D2 (segundo decil) es el valor por debajo del cual se encuentran como máximo el 20% de las observaciones, mientras que el 80% restante se sitúan por encima de él.

Y así sucesivamente.

Cuartiles: Son aquellos valores que dividen en cuatro partes iguales a un conjunto de datos ordenados. Se representan por Q1, Q2, y Q3. De esta manera tenemos que:

- Q1 (primer cuartil) es el valor por debajo del cual se sitúan a lo sumo el 25% de las observaciones y por encima de éste el 75% restante.
- Q2 (segundo cuartil) es el valor por debajo de cual se sitúan a lo sumo el 50% de las observaciones y por encima de éste el 50% restante. Está justo en el centro y corresponde a la mediana
- Q3 (tercer cuartil) es el valor por debajo del cual se sitúan a lo sumo el 75% de las observaciones y por encima de éste el 25% restante.

Observación: Hay algunas variaciones en las convenciones de cálculo de cuartiles ya que los valores reales calculados pueden variar un poco dependiendo de la convención seguida. Sin embargo, el objetivo de todos los procedimientos de cálculo de cuartiles es dividir los datos en aproximadamente cuatro partes iguales.

Percentiles:

Son aquellos valores que dividen a un conjunto de datos ordenados en cien partes iguales. Se representan por P1, P2...., P99. De esta manera tenemos que:

- P1 es el valor por debajo del cual se sitúan a lo sumo el 1% de los datos y por encima de él tenemos el 99% restante.

- P2 es el valor por debajo del cual se sitúan a lo sumo el 2% de los datos y por encima de él tenemos el 98% restante. Y así sucesivamente.

En forma genérica el p-ésimo percentil es un valor tal que por lo menos un " p" por ciento de los elementos tiene dicho valor o menos y, al menos, un (100-p) por ciento de los elementos tiene ese valor o más.

Tendencia central

Las medidas de tendencia central comprenden una serie de estadísticos cuyo propósito es el de resumir la información con respecto al comportamiento de los datos de una distribución en un solo número. Este número que, para tal fin, suele situarse hacia el centro de la distribución de datos se denomina medida o parámetro de tendencia central o de centralización. Las formas más representativas de este tipo de análisis incluyen el uso de la moda, la mediana y la media. La moda se conoce como la puntuación con mayor frecuencia en una distribución. La mediana es técnicamente el valor asociado con el percentil 50, lo cual refiere a que corta la distribución por la mitad. La media se define como el cociente que resulta de la sumatoria de todos los valores observados de x dividido entre el número total de observaciones.

Elección de una medida de tendencia central

El cálculo de cualquier estadístico es algo que en este punto debería de preocuparnos poco. No así la responsabilidad con la decisión asociada a la elección del estadístico más conveniente para nuestros análisis. Si bien las máquinas pueden realizar eficientemente los cálculos, aun no pueden llevar a efecto las reflexiones pertinentes que hasta tanto definen el trabajo del analista. A continuación, se exponen algunas consideraciones que deben tenerse presentes a la hora de elegir con qué estadístico de tendencia central debemos operar.

- En grupos muy pequeños es mejor evitar los estadísticos de resumen, en todos los casos los resultados suelen ser muy inestables.
- A diferencia de la media, la mediana no se ve afectada de manera importante por las presencia de valores atípicos en nuestra distribución.
- Algunos grupos de puntuaciones carecen de una tendencia central representativa. Esto es particularmente valido para grupos multimodales.

Variabilidad

Una vez que se han estudiado las propiedades de la distribución en términos de sus tendencias centrales, el investigador siempre necesitará responder a una serie de preguntas que atiendan a la cuestión ¿cuán diferentes son los datos observados con relación a estos valores de tendencia central? La variabilidad trata de una serie de estadísticos que intentan dar respuesta a la inquietud con respecto a cuán alejados están un conjunto de valores entre sí. Entre los estadísticos de variabilidad más utilizados se cuentan: a) el rango; b) la varianza y c) la desviación típica.

La medida más simple de dispersión es el rango o amplitud, la cual nos informa sobre la diferencia entre el valor menor y el mayor de un conjunto de valores. Esta medida presenta el problema de que puede verse afectada por la presencia de valores extremos, poco representativos. Por esta razón, se incluyeron variantes del rango que incluyen los análisis D, el rango intercuartílico y semi-intercuartílico. Para los análisis D el investigador excluye las puntuaciones extremas por denajo del percentil 10 y por encima del percentil 90. El rango intercuartílico por su parte considera los valores entre el primer y el tercer cuartil. En tanto que el rango semi-untercuartílico calcula el promedio del rango intercuartílico entre dos.

La varianza es el promedio de las diferencias de los valores individuales de x con relación a la media elevados al cuadrado y se simboliza s^2. Es uno de los estadísticos de mayor importancia por su presencia en la mayoría de las pruebas de contraste más potentes. Así como fundamento para muchos de los análisis más complejos. Sin embargo, cuando se trata de análisis descriptivos no es conveniente trabajar con este estadístico, pues se encuentra expreso en una escala diferente a la de las puntuaciones observadas lo que dificulta en gran forma su interpretación.

La desviación típica es la raíz cuadrada de la varianza y en este sentido refiere al promedio de las diferencias de las puntuaciones observadas con relación a la media del grupo, que se encuentra expresada en la misma escala que las puntuaciones originales. Lo cual facilita de manera importante su interpretación.

Forma de la distribución

Las medidas de forma de la distribución nos permiten identificar el tipo de distribución de la cual estamos extrayendo nuestras inferencias. Una vez se han respondido a preguntas con relación a la magnitud promedio del fenómeno en cuestión y cómo se dispersan los puntajes en

torno a estos promedios, algunas preguntas aflorarán como aquellas que desean conocer cuánta es la diferencia relativa entre la media y la mediana, cuando obviamente estas no son idénticas. Mientras que otras preguntas intentarán definir cuán dispersos se encuentran los datos con relación al promedio observado. Sus principales índices son la Asimetría y la Curtosis.

La asimetría es el estadístico que nos indica la medida en que los datos se distribuyen de forma uniforme alrededor del punto central (Media aritmética). La asimetría presenta tres estados diferentes (positiva, simétrica y negativa), cada uno de los cuales define de forma concisa como están distribuidos los datos respecto al eje de asimetría. Se dice que la asimetría es positiva cuando la mayoría de los sujetos se encuentran por debajo del valor de la media aritmética, la curva es Simétrica cuando se distribuyen la misma cantidad de sujetos y valores en ambos lados de la media y se conoce como asimetría negativa cuando la mayor cantidad de sujetos se aglomeran en los valores mayores a la media.

Con la curtosis se determina el grado de concentración que presentan los valores en torno a la región central o promedio de la distribución. Por medio del Coeficiente de Curtosis, podemos identificar si existe una gran concentración de valores (Leptocúrtica), una concentración normal (Mesocúrtica) ó una baja concentración (Platicúrtica).

Cuando la distribución de los datos cuenta con un coeficiente de asimetría muy cercano a 0 y un coeficiente de Curtosis con corrección de aproximadamente 0, se le denomina Curva Normal. Este criterio es de suma importancia ya que para la mayoría de los procedimientos de la estadística inferencial se requiere que los datos se distribuyan normalmente, para emplear pruebas paramétricas.

Puntuaciones z

Las puntuaciones z son las formas más populares de transformación de los valores o puntuaciones observadas, con el propósito de analizar su distancia respecto a la media, en unidades de desviación estándar. En este sentido, una puntuación estándar o z nos indica la dirección y la magnitud en que un valor individual se aleja de la media, en una escala de desviaciones estándar. Por la naturaleza de los datos, las transformaciones z sólo pueden ser llevadas a efecto en los casos de variables con un nivel de intervalo.

Estandarizar los valores permite comparar puntuaciones de dos o más distribuciones diferentes. La distribución que resulta de la transformación z no altera la forma original de la

distribución, pero sí modifica las unidades originales. La distribución transformada a puntuaciones z se establece con media 0 y desviación típica 1.

Las puntuaciones z también sirven para comparar mediciones de distintas pruebas o escalas aplicadas a los mismos sujetos. Las puntuaciones z también sirven para analizar las distancias entre puntuaciones de una misma distribución y áreas de la curva que abarcan tales distancias, o para sopesar el desempeño de un grupo de sujetos en varias pruebas.

Estadística inferencial

La estadística inferencial o inductiva, engloba una serie de estrategias que permiten la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión, desde las propiedades de ese conjunto de datos empíricos (muestra) hasta el conjunto total de datos (población) teniendo en cuenta la aleatoriedad de las observaciones. En este sentido, el análisis estadístico inferencial se encuentra altamente asociado a los conceptos de muestreo y de probabilidad. En la investigación, estas inferencias pueden tomar la forma de argumentos que permiten confirmar o rechazar las hipótesis de trabajo (prueba de hipótesis).

Referencias

Glass, G. y Stanley, J. (1984). *Métodos estadísticos aplicados a las ciencias sociales.* Bogotá: Prentice Hall Interamericana.

Hernández, R., Fernández, C. y Baptista, P. (2006). *Metodología de la Investigación* (4ª ed.). México, DF: McGraw-Hill.

Kerlinger, F. y Lee, H. (2002). *Investigación del comportamiento: Métodos de investigación en ciencias sociales* (4ª ed.). México: McGraw Hill Interamericana.

Landero, R., Gómez, M. (2006). *Estadística con SPSS y Metodología de la Investigación.* México: Trillas.

Muñiz, J. (1998). La medición de lo psicológico [versión electrónica] *Psicothema, 10* (1), 1-21.

Siegel, S. (1990). *Estadística no paramétrica* (3ra ed.). México, DF: Trillas.

Stevens, S. (1968). Measurement, statistics, and the schemapiric view. *Science, 161,* 849 – 856.