# CAPITULO 2

# 2. MARCO TEÓRICO

# ANÁLISIS MULTIVARIADO Y LOS MÉTODOS DE ESTIMACIÓN COMO HERRAMIENTAS DEL ANÁLISIS DE DATOS

# Las técnicas Estadísticas se utilizan en muchas situaciones de la vida diaria, puesto que en contacto con medios de comunicación como revistas, periódicos, radio, televisión y otros, la información se ofrece en forma de datos y son de mucha utilidad en la población en general, pero en particular, el conocimiento de la Estadística es de gran importancia para ingenieros, científicos, administradores y médicos quienes manejan y analizan datos que les permiten tomar decisiones sobre un problema en particular. Lo que demuestra la valiosa contribución que hace la Estadística en todas las áreas de la sociedad.

# Como ya se ha mencionado anteriormente, la Estadística y la Informática cada día se compenetran con mayor fuerza es por eso que en este capítulo hablaremos sobre las técnicas Estadísticas a utilizar para el análisis de los datos y el gran aporte de la Computación con las herramientas de software que permiten un mejor procesamiento y análisis de los mismos.

**2.1 Métodos De Estimación**

Para tener una mejor comprensión de los términos y las técnicas que vamos a utilizar, es conveniente hacer una explicación de los términos estadísticos.

**2.1.1 Concepto De Estadística**

El New Collegiate Dictionary de Webster, define la Estadística como *“Una rama de las matemáticas que trata de la recopilación, el análisis, la interpretación y la presentación de una gran cantidad de datos numéricos”.* Por otro lado Kendall y Stuart expresan *que “La Estadística es la rama del método científico que trata de los datos reunidos al contar o medir las propiedades de alguna población”*, según Fraser al referirse a la experimentación y las aplicaciones de la Estadística*: “La Estadística trata con métodos para obtener conclusiones a partir de los resultados de los experimentos o procesos”*. Freund relaciona a la Estadística con algo que abarca *“El conocimiento relacionado con el tomar decisiones en situaciones de incertidumbre”*. Mendenhall, Wackerly y Scheaffer reúnen todos estos conceptos y expresan *“Un examen superficial de estas definiciones indica una carencia desconcertante de uniformidad, pero todas tienen algunos elementos en común. Cada definición implica una recopilacón de datos teniendo como objetivo la inferencia. Cada una requiere la selección de un subconjunto de una gran colección de datos existente o conceptual, con el propósito de formular inferencias con respecto a las características del conjunto completo. Entonces* ***la Estadística es una teoría de la información que tiene como objetivo efectuar inferencias****.”*

**2.1.2 Población y Muestra**

El término *Población* se refiere a una colección o un conjunto de medidas de todos los elementos de un universo, acerca del que deseamos tomar decisiones o conclusiones. El subconjunto de ahí seleccionado representa una *Muestra*.

**2.1.3 La Distribución de Frecuencia y El Histograma**

Una Distribución de frecuencia es un resumen más compacto de datos que las observaciones originales.

Para construir una distribución de frecuencia, se debe dividir la gama de los datos en intervalos, los cuales se denominan intervalos de clase. Si es posible los intervalos de clase deben de ser del mismo ancho, para aumentar la información visual en la distribución de frecuencias. El número de intervalos que de clase que se utiliza depende del número de observaciones y de la dispersión de los datos. La raiz cuadrada del número de observaciones, funciona bien cuando se trata de elegir el número de intervalos de clase.

Las *Frecuencias Relativas* se determinan dividiendo la frecuecnia observada en cada intervalo de clase por el número total de observaciones.

También es de mucha utilidad presentar en forma gráfica la distribución de frecuencia, lo que se denomina *Histograma.* Para dibujar un Histograma, se usa el eje horizontal para representar la escala de medida, y se dibujan las fronteras de los intervalos de clase, el eje vertical representa la frecuencia relativa. El Histograma brinda una interpretación visual de la forma de la distribución de las mediciones, así como información acerca de la dispersión de los datos.

**2.1.4 Medidas de Tendencia Central**

La medida más común de tendencia central es la media aritmética ordinaria. Debido a que casi siempre consideramos a los datos como la muestra, nos referimos a la media aritmética como la media de la muestra. Si las observaciones de una muestra de tamaño n son x1, x2,.....xn, entonces la media de la muestra es:

La media de la muestra *X* representa el valor promedio de todas las observaciones en la muestra.

También podemos hacer el cálculo del valor promedio de todas las observaciones de una población, el cual se denomina *la media de la población,* y se denota:

En la práctica se utiliza la media de la muestra como una estimación puntual de la media de la Población.

Otra medida de tendencia central es la *mediana*, o punto en el cual la muestra se divide en dos mitades iguales.

Sean x(1), x(2),...x(n) los elemntos de una muestra arreglada en forma creciente de magnitud; esto es, x(1) es la primera observación más pequeña, x(2) denota la segunda observación más pequeña,..., y x(n) denota la observación más grande. Entonces la mediana se define como:

La Moda es la observación que ocurre con mayor frecuencia en la muestra.

**2.1.5 Medidas De Dispersión**

La medida de dispersión más importante es la *varianza de la muestra.* Si x1, x2,..xn es una muestra de n observaciones, entonces la varianza de la muestra se define como:

De igual forma, que para la muestra, hay una medida de variabilidad en la población denominada varianza de la población y se nota con la letra griega sLa raíz cuadrada sdenotará la desviación estándar de la población. Cuando la población es finita y consiste en N valores podemos definir la *varianza de la población* como:

**2.1.6 Experimentos Y Espacios Muestrales**

La teoría de la probabilidad ha sido inspirada por diferentes situaciones de la vida real en las que se realiza un experimento y el investigador observa un resultado. Este resultado no puede predecirse con certeza. A estos experimentos se los denomina ***Experimentos Aleatorios***y se nota con la letra E.

Podemos fijarnos que los experimentos aleatorios tienen algunas características comunes. En primer lugar no podemos predecir el resultado con certeza, pero sí es factible describir el conjunto de posibles resultados.

En segundo lugar, desde un punto de vista conceptual, el experimento es tal que podría repetirse en condiciones que permanezcan invariables, ocurriendo los resultados de una manera fortuita; no obstante, a medida que el número de repeticiones aumenta, surgen ciertos patrones en la frecuencia de ocurrencia de los resultados.

El conjunto de resultados posibles se llama ***Espacio Muestral*** y se nota con la letra S.

Un ***Espacio Muestral Discreto***es aquel en el que hay un número finito de resultados.

Un ***Espacio Muestral Contínuo*** es aquel que tiene resultados incontables, estos podrían ser números reales en un intervalo o pares reales contenidos en el producto de intervalos, donde las nediciones se realizan respecto a dos variables en un experimento.

**2.1.7 Eventos**

Un evento **e**, está asociado al espacio muestral del experimento. El espacio Muestral **S**, se considera el conjunto universal, de modo que **e** es simplemente un subconjunto de **E**.

**2.1.8 Variables Aleatorias**

Si **E** es un experimento que tiene el espacio muestral **S**, y **X** es una función que asigna un número real **X(e)** para todo resultado **e** que pertenece a **S**, entonces **X(e)** se denomina ***variable aleatoria.***

**2.1.8.1 Variables Aleatorias Discretas**

Si x es una variable aleatoria, asociamos un número px(xi)=Px(X=xi) con cada resultado xi, en Rx para i=1,2,..n,..., donde los números px(xi) satisfacen:

1. px(xi)0 para todo i

Observamos de inmediato que

Px(xi) = Fx(xi) – Fx(xi-1)

Y

Fx(xi) = Px(Xxi) =

La función px se llama función de probabilidad o ley de probabilidad de la variablealeatoria, y la colección de pares [(xi, px(xi)) i=1,2,...] se llama distribución de la probabilidad de X. La función px suele presentarse en forma tabular, gráfica o matemática.

**2.1.8.2 Variables Aleatorias Contínuas**

Una variable aleatoria X es contínua cuando su espacio del rango Rx consistirá en uno o más intervalos.

La función de densidad de una variable aleatoria contínua fx(x) se define como:

Y resulta que:

Notamos ahora las siguientes propiedades:

1. fx(x) para toda x que pertenece a Rx.

2.

1. Fx(x) = 0, si x no está en el rango Rx

**2.1.9 Estimación De Parámetros**

Las técnicas de inferencia estadística pueden dividirse en dos áreas principales: *Estimación de parámetros y pruebas de hipótesis.*

Si x es una variable aleatoria con distribución de probabilidad f(x), caracterizada por el parámetro desconocido , y si X1, X2, ...Xn es una muestra aleatoria de tamaño n de x, entonces la estadística =h(X1, X2, ...Xn)correspondiente a  se llama *estimador de *

Una propiedad deseable de un estimador es que debe estar en cierto sentido “cercano” al valor verdadero del parámetro desconocido. Formalmente decimos que *~~Ô~~* es un estimador insesgado del parámetro **si

E(*~~Ô~~*) = **

**2.1.10 Coeficiente de Confianza**

En muchas ocasiones el estimador puntual no proporciona suficiente infromación acerca del parámetro de interés. Una estimación del intervalo de la forma **L U** podría resultar más útil. Los puntos extremos de este intervalo representan variables aleatorias, ya que son funciones de datos de la muestra.

En general, para construir un estimador de intervalo del parámetro desconocido debemos encontrar L y U tal que:

**P { L U} = 1 - **

El intervalo resultante **L U,** se llama ***Intervalo de confianza*** del 100(1-por ciento para el parámetro desconocido a **L** y **U** se los denomina ***Límites de Confianza*** inferior y superior respectivamente, además ***1-***recibe el nombre de ***Coeficiente de Confianza.*** La interpretación del Intervalo de Confianza es que si se coleccionan muchas muestras aleatorias y se calcula un intervalo de confianza del 100(1-por ciento en el valor de cada muestra, entonces 100(1-de los intervalos contendrá el valor verdadero de 

**2.1.11 Covarianza y Correlación**

Como sabemos, E(X1) =1 y V(X1) = son la media y la varianza de X1, pueden determinarse a partir de la distribución marginal de X1. De manera similar E(X2) =2 y V(X2) = son la media y la varianza de X2,  Dos medidas que se utilizan para describir el grado de asociación entre X1 y X2 son la ***Covarianza*** de [X1, X2] y el ***Coeficiente de correlación*** ****

**Definición**

Si [X1, X2] es una variable aleatoria bidimensional, la covarianza, denotada por , es:

**Cov(X1, X2) = E[(X1 – E(x1))(X2 – E(x2))]**

Haciendo las respectivas multiplicaciones tenemos:

**Cov(X1, X2) =  E(X1 X2) – [E(x1) E(x2)]**

El coeficiente de correlación es una cantidad adimensional que mide la asociación lineal entre dos variables aleatorias.

Y se denota por:

**2.1.11.1 Teorema**

Si X1 y X2 son independientes, entonces  = 0.

El inverso de este teorema no necesariamente es cierto, pues podemos tener  = 0 sin que las variables sean independientes. En este caso se dice que las variables no están correlacionadas.

**2.1.12 Pruebas de Hipótesis**

En muchas ocasiones, se requiere decidir si se acepta o se rechaza un enunciado acerca de algún parámetro. El enunciado suele llamarse ***Hipótesis***, y el procedimiento en torno a la hipótesis recibe el nombre de ***Prueba de Hipótesis***.

Una Hipótesis Estadística es un enunciado acerca de la distribución de probabilidad de una variable aleatoria. Las hipótesis Estadísticas con mucha frecuencia involucran uno o más parámetros de esta distribución. Es importante recordar que las hipótesis son siempre enunciados relativos a la población o distribución bajo estudio, no enunciados en torno a la muestra.

Existen dos tipos de hipótesis, la hipótesis Nula que es la que nos fija o especifica un valor del parámetro de la población especificada. La hipótesis alterna especifica valores del parámetro que podrían ser más grandes o más pequeños.

* + 1. **Análisis de Correspondencia Múltiple**

El Método de Análisis de Correspondencia Múltiple es una generalización del Análisis Factorial de Correspondencias, y estudia la relación entre un número p de variables con cada una de las m modalidades de las posibles respuestas.

Es una tabla formada por ceros y unos, donde el elemento cero representa la ausencia de la característica en mención en un individuo X. Este procedimiento resulta muy útil en análisis de variables cualitativas.

El objetivo de este método es analizar la relación que existe entre cada una de las modalidades, es decir, entre cada uno de los valores o características que toma una variable.

Para efectuar este método se necesita definir la matriz **S**, de orden n x l, donde n representa el número de individuos y l el número de modalidades como variables, por lo tanto se obtiene una matriz de elementos Sij que toma valores de ceros o unos.

|  |  |
| --- | --- |
|  | *P es el n úmero de Preguntas o características a estudiar* |
|  | *Es el inverso del Número de Preguntas* |
|  | *Es el número de individuos que posee la modalidad m* |

Para construir la matriz final, se nota la matriz B de orden n x n como resultado de multiplicar St por S: B = St S. Además se define la matriz D diagonal de orden l x l, cuyos elementos son los totales por columna o modalidad (total de unos por cada columna). Y por último se nota la matriz E de orden l x l como resultado de un producto de matrices:



**2.2 Epi Info 2000**

Epi Info es una serie de programas para Microsoft Windows 95, 98, NT y Windows 2000 de distribución gratuita, usado por profesionales en Investigaciones sobre Salud Pública, Administradores de Bases de Datos para la Salud Pública y otras técnicas, y tiene su aplicación en bases de datos generales y técnicas estadísticas.

La primera versión de Epi Info, fue desarrollada en 1985, y según un estudio realizado en 1997, se registraron 145.000 copias en versión DOS de Epi Info y Epi Map en 117 países. El manual DOS y otros programas han sido traducidos en 13 idiomas diferentes.

Epi Info 2000 tiene gran compatibilidad con los estándares de esta industria, como Microsoft Access, SQL y otras bases de datos ODBC; Visual Basic Versión 6, browsers World Wide Web y HTML, además incluye un Sistema de Información Geográfica (SIG), llamado Epi Map 2000, construído en el programa MapObject de Environmental Systems Research, Inc. (ESRI), quienes son los productores de ArcView. Epi Map es compatible con datos SIG de numerosos sitios de Internet en el popular formato ESRI. En Epi Map podemos encontrar mapas de diferentes Países con los cuales se puede trabajar en forma integrada o en forma individual según las regiones en las que se encuentre dividido. Este Sistema nos ayudará a obtener una mayor explicación geográfica de los datos analizados.

Utilizando Epi Info 2000 y un computador personal, los Médicos, Epidemiólogos y todos los médicos que trabajen en Salud Pública, pueden desarrollar rápidamente un cuestionario o formulario, ingresar datos y analizarlos.

Entre las utilidades estadísticas que nos brinda este software, están el cálculo de medias, varianzas, modas, percentiles, frecuencias, correlaciones, límites de confianza, etc, utilizando además diferentes técnicas estadísticas para estimación de parámetros como regresiones, Pruebas de Hipótesis Estadísticas, Análisis de Varianza, y una aplicación en el análisis actuarial con el Análisis de Sobrevivencia de Kaplan Meier.