ð (De *estadista*); sust. f.

1. Ciencia que utiliza conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades: *la **estadística** ofrece conclusiones muy impersonales que en ocasiones no reflejan la realidad*.

2. Ciencia que tiene por objeto reunir y clasificar una serie de hechos con una determinada característica común: *este programa de **estadística** calcula automáticamente el PIB de cada país*.

3. [Por extensión] Conjunto homogéneo de datos sobre hechos o manifestaciones de cualquier tipo (social, científico, deportivo, etc.): *hay una **estadística** que dice que las mujeres españolas fuman más que los hombres*; *me gustaría tener las **estadísticas** del partido para ver los porcentajes de tiro de tombergas*.

Sinónimos

Diagrama, gráfico, esquema, censo, padrón, catastro, lista, natalidad, demografía, nupcialidad, mortalidad, criminalidad, porcentaje.

ð (**1**)[Matemáticas] **Estadística**.

Rama de las Matemáticas que se basa en la obtención de los métodos adecuados para obtener conclusiones razonables cuando hay incertidumbre. Esta ciencia tiene como principal objeto aplicar las leyes de la cantidad a hechos sociales para medir su intensidad, deducir las leyes que los rigen y hacer un predicción próxima. Existen dos ramas muy diferentes dentro de la estadística: la estadística descriptiva y la estadística matemática.

La estadística **descriptiva** se basa en la recolección de datos de una muestra representativa de una población de la que se quiere estudiar alguna característica, en su tratamiento y en la obtención de una serie de resultados y medidas matemáticas (medias, desviaciones, etc), que posteriormente se analizan y se extrae una conclusión de las posibles causas que producen la característica de la población en estudio y su relación con otros fenómenos.

La estadística **matemática** utiliza el cálculo de probabilidades para establecer previsiones y conclusiones de los fenómenos colectivos.

ð (**2**)[Técnica] **Estadística**.

Estudio del tratamiento de la información que contienen las series de datos procedentes de observaciones de fenómenos colectivos (demográficos, económicos, tecnológicos, etc), en los que el gran número de factores de variación que intervienen hace necesarios modelos probabilísticos para que a las conclusiones, leyes o decisiones basadas en los mismos, se les pueda asignar una confianza mensurable.

Evolución histórica

Aunque es muy difícil establecer los comienzos de la estadística tal y como hoy la conocemos, se puede decir que el primer precedente que se encuentra en la historia son los censos chinos que el emperador Tao ordenó elaborar hacia el año 2200 antes de Cristo. Más tarde, hacia el 555 a.C., se elaboraron los censos del Imperio Romano, y desde entonces hasta el siglo XVII muchos estados realizaron estudios sobre sus poblaciones.

En 1662 el comerciante londinense de lencería J. Graunt publica un libro, titulado *Natural and Political Observations made upon the Bills of Mortality* (Observaciones naturales y políticas extraídas de los

certificados de defunción), que es el primer intento de interpretar fenómenos biológicos y sociales de la población partiendo de datos numéricos. En su libro ponía de manifiesto la influencia que tenían las cifras de nacimientos y muertes sobre el medio social en el Londres de los años 1604 a 1661. Graunt entabló amistad con un hombre llamado Sir Willian Pety, que más tarde publicó un libro sobre lo que él llamó "la nueva ciencia de la aritmética política" que fue ampliamente difundido. A finales del siglo XVII, el famoso astrónomo E. Halley publicó un artículo titulado "Un cálculo de los grados de mortalidad de la humanidad, deducido de curiosas tablas de los nacimientos y funerales de la ciudad de Breslau". Los estudios publicados por Graunt, Pety y Halley están considerados la base fundamental de los trabajos posteriores sobre esperanza de vida que son tan utilizados hoy por las aseguradoras.

Hacia el año 1835 el astrónomo real de Bélgica A. Q. Lambert elaboró el primer censo de su país, en el que analizó la influencia sobre la mortalidad de determinados factores como la edad, el sexo y la situación económica. A este estudioso se debe el término "hombre medio" que adoptarían después otros científicos de esta materia.

Pero es a otros dos hombres relevantes, llamados Galton y Pearson, a los que se considera los "padres" de la estadística actual, ya que fueron ellos los que provocaron el paso de la estadística deductiva, estudiada en su época, a la estadística inductiva, que es la que hoy tiene más influencia en la ciencia.

Por último, cabe destacar aquí, al profesor de física, G. T. Fechner. Este científico del siglo XIX aplicó los conocimientos de estadística de su tiempo a las ciencias sociales, fundamentalmente a la psicología.

Generalidades

La estadística descriptiva incluye al conjunto de tratamientos de los datos de una muestra, de los que se extraen unos valores que sintetizan o resumen sus características más importantes, y las técnicas de representación de estos valores de forma que se facilite su análisis. Los valores que aportan gran información sobre los datos tomados son las medidas de centralización, dispersión y forma.

Se conoce con el nombre de *variable cuantitativa*, o simplemente *variable*, a aquella magnitud que toma valores mensurables. Las variables se conocen como discretas si toman valores enteros, como el número de alumnos en un aula o el número de defectos por metro en un cable eléctrico. Las variables continuas pueden variar de forma continua, como por ejemplo el peso de una persona o la longitud de una varilla.

Las *variables cualitativas* o *atributos* son aquellas cualidades que no son mensurables, por ejemplo si una determinada pieza es o no defectuosa.

La Regresión muestra la dependencia entre variables por medio de un modelo matemático que contempla tanto la parte sistemática como la aleatoria de la relación entre dichas variables. El modelo obtenido se contrasta por medio de unas pruebas estadísticas con las que se comprueban las hipótesis formuladas, y así generalizar los resultados a la población.

Medidas de centralización

Estas medidas proporcionan información sobre la tendencia central de las observaciones.

− Media: La media o media aritmética (x) es la suma de un conjunto de valores dividido entre el número total de estos datos.

$$\bar{x} = (1/n) \eth x_i$$

En el caso en el que los datos estén agrupados en intervalos, los valores de una misma clase pueden ser sustituidos por la marca de la clase correspondiente.

Siendo: xi : marca de la clase del intervalo i

Fi : frecuencia absoluta del intervalo i

$\bar{x}$

$\bar{x} = (ð\ fi\ xi) / (ð\ Fi)$

Las propiedades de la media son las siguientes:

− La media de una constante es la propia constante.

− La media de la suma o diferencia de variables es igual a la suma o diferencia de las medias de dichas variables.

− La media del producto de una constante por una variable, es igual a la constante por la media de la variable.

− La media de una combinación lineal de dos o más variables es igual a la combinación lineal de las medias de dichas variables.

− La media es el centro de gravedad de la distribución, ya que las desviaciones respecto a la media suman 0.

$\bar{x}$

$ð(xi - \bar{x}) = 0$

− Mediana: La mediana es el valor del elemento que ocupa el lugar central, si los datos están ordenados, bien de forma creciente o de forma decreciente.

− Moda: La moda es el valor más frecuente, es decir es el valor de la variable que se repite un mayor número de veces.

En el caso de una distribución totalmente simétrica, la media y la mediana coinciden. Si la media y la mediana difieren mucho significa que hay heterogeneidad entre los datos y que la distribución, por tanto será asimétrica.

Medidas de dispersión

Son las medidas que muestran la variación de los datos tomados.

− Recorrido: El Recorrido (R) es la diferencia entre el valor mayor y el valor menor de los que toma la variable.

$R = xmax - xmin$

− Varianza: La varianza (S2) es la media aritmética de los cuadrados de las desviaciones respecto a la media.

$\bar{x}$

$S2 = (1/n)\ \eth(xi - x)2$

Las propiedades de la varianza son:

− La varianza es siempre positiva o cero.

− La varianza de una constante es cero.

− La varianza de la suma o diferencia de una variable y una constante es igual a la varianza de la variable.

− La varianza de un producto de una constante por una variable es igual al cuadrado de la constante por la varianza de la variable.

− Desviación Típica: La desviación típica o estándar (S) se define por:

_____

$S = \eth(1/n)\ \eth(xi - x)2$

Las propiedades de la desviación típica son:

− La desviación típica es siempre positiva o cero.

− La desviación típica de una constante es cero.

− La desviación típica de una constante por una variable es igual a la constante por la desviación típica de la variable.

− La desviación típica de la suma o diferencia de una variable y una constante es igual a la desviación típica de la variable.

− Meda: La meda es al medida de dispersión asociada a la mediana, y es la mediana de las desviaciones absolutas.

MEDA = mediana xi − mediana

− Percentil: El percentil (p) de una variable es el menor valor superior al p% de los datos. Se utiliza para construir medidas de dispersión en los datos ordenados. Se denomina Primer Cuartil (Q1) al menor dato que supera al 25 % de los datos menores. El Segundo Cuartil (Q2) coincide con la mediana. El Tercer Cuartil (Q3) es el menor dato que supera al 75% de los datos menores. Se usa con frecuencia el rango intercuantílico, que es la diferencia entre los percentiles Q3 y Q1.

Representación gráfica de los datos

Las representaciones gráficas de una distribución de frecuencias permite obtener, de un golpe de vista, información de las características de dicha distribución.

A) Histograma

El Histograma representa la frecuencia con la que se presentan los diferentes grupos de datos de la variable objeto de estudio. Es un conjunto de rectángulos, los cuales representan a cada una de las clases. En el eje de abscisas se representan las clases definidas y en el eje de ordenadas la frecuencia de cada una de ellas.

La amplitud del intervalo de las clases se halla dividiendo el Recorrido entre el número de clases.

El Histograma proporciona mucha información respecto a la estructura de los datos. Por tanto, es importante analizar la situación del centro del Histograma y el ancho del mismo que definen la tendencia central y la variabilidad del conjunto de datos respectivamente, así como la forma del Histograma que identifica algunas de las características del proceso en estudio.

− **Distribución Simétrica Unimodal:** Se caracteriza porque cada una de las observaciones equidistantes al máximo central, tienen aproximadamente la misma frecuencia. Es típico de la mayoría de los procesos industriales.

− **Distribución Asimétrica:** Es típica de datos económicos, y de forma general en distribuciones de renta, consumo de electricidad, población, tamaño de empresas,...

− **Distribución Triangular**: Es totalmente asimétrica y se presenta al estudiar tiempos entre averías, entre llegadas, entre accidentes, o en fabricación donde existe la imposibilidad de superar un valor o bien se ha realizado una selección de 100% de alguna característica.

− **Distribución Bimodal:** Se presenta como dos distribuciones muy separadas. Suele aparecer cuando se han recopilado datos a partir de dos procesos distintos, tales como las características de una pieza suministrada por dos proveedores diferentes.

− **Distribución Rectangular:** Presenta gran variabilidad. Aparece al mezclar datos de Distribuciones Simétricas Unimodales.

− **Distribución Truncada:** Aparece al presentar datos de procesos que no cumplen las especificaciones, después de pasar un control de calidad. Puede ser, también un síntoma de una elección de un número de clases menor al adecuado.

− **Distribución sin Datos en la Zona Central:** Suele aparecer cuando los datos corresponden a un material de mala calidad, y el "material bueno" ha sido seleccionado previamente.

− **Distribución con Picos en las Colas:** Es una representación típica cuando se han sometido a un reproceso, los elementos que en un primer control cayeron fuera de tolerancias.

B) Diagrama de puntos

Este gráfico muestra un conjunto de puntos, que son la intersección de las frecuencias (representadas en el eje de ordenadas) y de los valores de la distribución (representados en el eje de abscisas).

C) Diagrama de barras

Presenta los valores posibles de los datos sin agrupar y sus frecuencias absolutas o relativas. En el eje horizontal aparecen los datos tratados y en el eje vertical las frecuencias. Sobre el eje horizontal se traza un segmento de longitud proporcional al valor de las frecuencias.

D) Polígono de frecuencias

Es un gráfico que une los puntos que representan la intersección de las marcas de clase con su frecuencia correspondiente.

Cabe mencionar también los gráficos de sectores, de rectángulos, pictogramas, etc.

Covarianza y correlación. Análisis gráfico

− **Covarianza**: Es una medida de asociación que mide la relación lineal entre las variables "x" e "y", y se define como:

$$Cov(x,y) = (1/n) * (ð(xi - \bar{x})(yi - \bar{y}))$$

En el caso de que los datos estén agrupados en clases, la fórmula de la covarianza es la siguiente:

$$Cov(x,y) = ð((xi - \bar{x})(yi - \bar{y}) f(yi, xi))$$

− **Correlación:** Es el resultado de dividir la covarianza por un término de sus mismas dimensiones, obteniendo el coeficiente de correlación (r). Sx y Sy, son las desviaciones típicas de "x" e "y" respectivamente.

$$r = (Cov(x,y)) / (Sx * Sy)$$

Las propiedades del coeficiente de correlación son las siguientes:

− Si multiplicamos "x" e "y". por constantes (aunque éstas sean distintas), el coeficiente de correlación no varía.

− Cuando no existe una relación lineal exacta entre "x" e "y", el coeficiente de correlación varía entre −1 y 1 (−1 <> r <> 1).

− Cuando no existe relación lineal r = 0 (en este caso puede ocurrir que exista otro tipo de relación no lineal).

− Si existe una relación lineal, entonces r = 1 (correlación lineal perfecta y directa) ó r = −1 (correlación lineal perfecta e inversa).

Del simple análisis gráfico se obtiene una gran información:

− Caso A: la relación es lineal, y además es positiva o directa ya que la variable "y" aumenta proporcionalmente con "x". Existe dispersión de los puntos, por lo que "r" no se acercará a la unidad.

− Caso B: Se observa que no existe ninguna relación entre las variables. No hay variaciones sistemáticas de una variable cuando varía la otra.

− Caso C: Existe una relación lineal, inversa o negativa. Los puntos están concentrados, luego "r" estará próxima a "−1".

− Caso D: Existe una relación no lineal.

Regresión simple

Cuando existe una dependencia causal entre dos variables, y se toman diversas observaciones, éstas aparecen reflejadas en una nube de puntos debido al componente aleatorio, a pesar de que pueda existir una fuerte dependencia entre ellas. El objetivo de la regresión es obtener una recta (línea de regresión, denominada así

por Galton) hacia la que tienden los puntos de un diagrama de dispersión, y va a definir la dependencia exacta entre las variables "x" e "y".

El modelo de la recta de regresión se ajusta a la expresión:

y = ðð + ðð * x + µ

Siendo "ðð + ðð * x" , la parte sistemática o explicada y "µ" la parte aleatoria o impredecible o perturbación, que engloba a todas las variables no explícitas en el modelo, las cuales tienen relevancia sobre el resultado de "y". Los números ðð, ðð se denominan parámetros de la recta y definen completamente el modelo. "ðð" recibe el nombre de coeficiente de regresión.

Para proceder a la estimación de los parámetros ðð y ðð se parte de las siguientes hipótesis:

− Linealidad.

− Homocedasticidad de las perturbaciones, es decir su variabilidad se mantiene constante.

− Independencia de las perturbaciones. Sus covarianzas deben ser nulas, para que haya ausencia de autocorrelación.

− Normalidad. Las perturbaciones deben seguir el modelo de una distribución normal (0, ð).

Si de una población se extrae una muestra, se tiene la siguiente relación:

ð

y = yi + ei

ð

yi : valores medios de la variable dependiente correspondientes a los valores xi observados (valor previsto).

ei: errores o residuos debidos a los factores aleatorios de perturbación. En Estadística, residuo = valor observado − valor previsto. Se cumple siempre que ð ei = 0

Para una función lineal:

ð ð ð

y = ð0 + ð1 x

Regresión y correlación múltiples

En la realidad es habitual que exista una dependencia causal entre más de dos variables, con una variable dependiente "y" (efecto) y varias variables independientes "x1", "x2",... (causas).

Estimación del modelo:

En una población, se considera:

ð ð ð

yi = ð0 + ð1 X1i + ... + bk xki + ði

Siendo:

ð ð ð

(ð0 + ð1 X1i + ... + ðk Xki) la parte sistemática y µi la parte aleatoria.

De la población, se obtiene una muestra:

ð ð ð

yi = ð0 + ð1 X1i + ... + ðk Xki + ei

Siendo:

yi : el valor observado

ð ð ð ð

yi = ð0 + ð1 X1i + ... + ðk Xki : el valor previsto

ei : el residuo

La estimación de los coeficientes de regresión (ð0 , ð1 , ... , ðð) se realiza por el método de los mínimos cuadrados, minimizando la suma de los residuos al cuadrado. También se deben cumplir las hipótesis de linealidad, homocedasticidad, independencia y normalidad de la parte aleatoria.

Ecuaciones temporales

Las series temporales expresan la relación entre dos variables, siendo una de ellas el "tiempo". Estas series permiten describir la evolución en el pasado de una magnitud y formular predicciones para el futuro.

Este análisis puede realizarse desde dos puntos de vista, el llamado enfoque clásico y el enfoque causal.

El enfoque causal estudia las series temporales en función de las variables que han producido dichas variaciones, ya que el tiempo es sólo el marco donde se producen los hechos, no la causa de los mismos.

El enfoque clásico analiza la serie considerando cada variable por separado y en función del tiempo; se ha convertido en un método estándar de estudio de estas series, y es aceptado de forma unánime por los estadísticos; por tanto es el que se describirá en el presente tema.

Las series temporales, también denominadas series cronológicas, crónicas o históricas son un conjunto de observaciones de una variable, la cual está relacionada con un conjunto de intervalos o instantes de tiempos. Cuando cada observación se refiere a un período, la variable se denomina "flujo" (por ejemplo, la serie mensual del consumo de electricidad a nivel nacional) y cuando cada observación se refiere a un instante, la variable se denomina "nivel" o "stock" (por ejemplo, serie diaria de las temperaturas recogidas en un punto determinado cada hora).

Las observaciones de una serie temporal tienen ciertas peculiaridades, pues con el paso del tiempo pueden perder homogeneidad a causa de factores como son la mejora de los métodos de observación estadística, las variaciones en las definiciones estadísticas, etc. Por otra parte, las observaciones temporales no son del todo

independientes y es frecuente que una observación dependa de la precedente, y se suele presentar una correlación entre una serie temporal y la misma pero retardada en el tiempo, fenómeno que se denomina autocorrelación. Por todo ello, es necesario estudiar las series temporales de forma independiente a las series de las observaciones transversales.

Al analizar una serie temporal, se puede apreciar que se producen variaciones, y éstas pueden ser:

− **Evolutivas:** el "nivel medio" de la variable está sometido a cambios muy bruscos.

En los siguientes gráficos se aprecian estas variaciones; en la "figura 1" se observa cómo la serie va pasando de unos niveles altos al principio a unos muy bajos después, para volver al final a tener unos niveles semejantes a los del principio. En la "figura 2", se aprecia cómo las observaciones van teniendo niveles cada vez más altos.

− **Estacionarias:** el "nivel medio" de la variable permanece prácticamente constante, lo que no quiere decir que no aparezcan fluctuaciones.

En el siguiente gráfico aparece una serie estacionaria, que aunque muestra fluctuaciones, el "nivel medio" permanece constante en el tiempo.

Por tanto, como es normal encontrar variables que fluctúan en el tiempo, es importante conocer las causas que producen tales variaciones, para analizar la evolución en el tiempo de un fenómeno determinado, de forma que se pueda, tanto describir el comportamiento en el pasado de la variable como formular predicciones para el futuro.

Desde el punto de vista clásico, se estudia una serie temporal considerando cada variable por separado y en función del tiempo, es decir:

$y = f(t)$

Siendo:

y : variable dependiente

t : variable independiente o explicativa

En general, las serie temporales son mensuales, trimestrales o anuales.

Partiendo de la hipótesis de que las observaciones de la variable corresponden a un intervalo de tiempo, se supone que las variaciones que se producen son el resultado del efecto de cuatro fuerzas o componentes: la tendencia, la estacionalidad, los ciclos y los accidentes.

− **Tendencia (T):** Conocida también como "tendencia secular". Determina los movimientos de la variable a largo plazo, ignorando de forma consciente las variaciones a corto y medio plazo. Se consideran, en general, más de diez años, para apreciar si la serie obedece a una ley determinada. La tendencia proporciona información sobre si la serie es estacionaria o evolutiva.

− **Estacionalidad (E):** Representa las variaciones a corto plazo, que se repiten de forma periódica en un tiempo inferior al año, y que además se van reproduciendo un año tras otro. Es independiente de la tendencia y de las fluctuaciones cíclicas.

La estacionalidad debe ser rigurosamente periódica. Suponiendo "t" el número de años y "m" el período

correspondiente:

Et = Et + m = Et + 2m =...

− Fluctuaciones cíclicas (C): Corresponden a las variaciones a medio plazo, es decir a tiempos superiores al año. Estos movimientos cíclicos no son tan regulares como los estacionales, se corresponden con movimientos de periodicidad y amplitud variables.

− Variaciones accidentales (A): Son los movimientos esporádicos que se producen de forma ocasional. Son provocados por dos tipos de factores: aleatorios (provocados por pequeños accidentes) o erráticos (como consecuencia de inundaciones, terremotos, huelgas, etc). Se corresponden con movimientos irregulares e imprevisibles de poca amplitud y no permanentes. En general, su media es nula, en un numero pequeño de meses.

La tendencia y las fluctuaciones cíclicas pueden tomarse de forma conjunta, como un movimiento coyuntural o extraestacional, que marca la evolución fundamental de la serie.

Inferencia estadística

La inferencia estadística es una parte de la Estadística que permite generar modelos probabilísticos a partir de un conjunto de observaciones.

Del conjunto se observaciones que van a ser analizadas, se eligen aleatoriamente sólo unas cuantas, que es lo que se denomina muestra, y a partir de dicha muestra se estiman los parámetros del modelo, y se contrastan las hipótesis establecidas, con el objeto de determinar si el modelo probabilístico es el adecuado al problema real que se ha planteado.

La utilidad de la inferencia estadística, consiste en que si el modelo se considera adecuado, puede usarse para la toma de decisiones o para la realización de las previsiones convenientes.

En el desarrollo del tema se utilizarán variables aleatorias, que son variables determinadas por el azar.

La inferencia estadística parte de un conjunto de observaciones de una variable, y a partir de estos datos "infiere" o genera un modelo probabilístico; por tanto es la consecuencia de la investigación empírica. La inferencia estadística es, en consecuencia, un planteamiento inductivo.

Partiendo de los datos recopilados, la inferencia estadística sigue los siguientes pasos:

− estimar los parámetros (por ejemplo la media y la desviación típica)

− hallar los intervalos de confianza, es decir el rango de valores donde es probable que se encuentren los parámetros.

− contrastar las hipótesis (por ejemplo si la media μ es igual a un valor μ0 ó no es igual a μ0).

Se entiende por población al conjunto de elementos de los que se analiza una cierta característica. La práctica dice que lo habitual, es no poder estudiar la totalidad de estos elementos, debido a diversas razones, tales como:

− Económicamente, no es rentable el análisis de toda la población, por ser excesivamente grande.

− Los elementos, no existen como tales. Como son los casos de los elementos defectuosos.

− El análisis requiere la destrucción de los elementos. Como, por ejemplo, en los ensayos destructivos.

Por todo lo mencionado anteriormente, se selecciona sólo un conjunto de los elementos, que es lo que se denomina muestra.

Se entiende por marco el patrón de la población por el cual deben regularse o contrastarse las medidas.

Muestreo aleatorio

Dentro de las técnicas de muestreo aleatorio merecen mención, el muestreo aleatorio simple, el muestreo aleatorio estratificado, el muestreo sistemático y el muestreo polietápico. Todas ellas tienen como objetivo fundamental seleccionar muestras que sean representativas de la población.

A) Muestreo Aleatorio Simple

El muestreo aleatorio simple consiste en seleccionar elementos de una población, bajo las siguientes condiciones:

−todos los elementos tienen la misma probabilidad de ser elegidos

−la población es idéntica en todas las extracciones, es decir una vez seleccionada una población, ésta se reemplaza.

La selección de las observaciones de una muestra aleatoria simple, se suele realizar mediante "números aleatorios", que son precisamente, un conjunto de números, los cuales tienen todos ellos la misma probabilidad de aparición.

Si x1, x2,...,xn es una muestra aleatoria simple de una variable discreta, la probabilidad de obtener dicha muestra se denomina "probabilidad conjunta" y es igual al producto de las probabilidades de cada observación:

P (x1, x2,...,xn) = P (x1) P (x2) ... P (xn)

Esta relación se obtiene como consecuencia de la independencia de las observaciones.

Si la variable aleatoria es continua, se establece una relación equivalente a la anterior, pero con las funciones de densidad.

f (x1, x2,...,xn) = f (x1) f (x2) ... f (xn)

Con esta técnica, cada uno de los elementos de la muestra Xi será una variable aleatoria con la misma distribución que la población de la que se ha obtenido.

Entre sí, los elementos muestrales también son variables aleatorias independientes.

Método de montecarlo

El Método de Montecarlo es una forma artificial de realizar el muestreo aleatorio simple, pues se utiliza cuando los elementos de la población no están disponibles.

Consiste en seleccionar muestras de cualquier población, siempre y cuando se conozca su distribución de probabilidad.

B) Muestreo Aleatorio Estratificado

El muestreo aleatorio estratificado se produce cuando los elementos de una población se estructuran en clases (o estratos).

Para dividir la población en clases se siguen los siguientes criterios:

− Se respetan, de forma proporcional, los tamaños relativos en la población. Es decir, si en una población existieran un 60% de mujeres y un 40% de hombres, esta proporción se respetaría en el estrato.

− Se respeta también, de forma proporcional, la variabilidad de la población en el estrato. Es decir, se toman menos elementos de estratos donde la característica tenga menor dispersión.

La muestra se elige de la siguiente manera:

−se asigna un número determinado de elementos a cada clase

−se elige, por muestreo aleatorio simple, dentro de cada clase

C) Muestreo Sistemático

Se utiliza cuando los datos de la población se presentan ordenados en listas. Dada una población de tamaño "N", si se pretende obtener una muestra de tamaño "n". Suponiendo que "k" es el número entero más cercano a N/n, la muestra sistemática se va obteniendo al azar un elemento entre los primeros "k" (mediante números aleatorios).

Si el orden del elemento elegido es nð, a continuación se toman los elementos n1+k, nð+2k, y así de forma sucesiva a intervalos fijos de "k" hasta completar la muestra.

Si en el orden en el que se presentan los elementos de la población, se encuentran más próximos los individuos que son más semejantes y se encuentran más alejados los individuos que más difieren entre sí, este método es más preciso que el muestreo aleatorio simple, porque cubre de forma más homogénea la población.

Si el orden de los elementos que figuran en las listas ha sido tomado al azar, este método es igual al muestreo aleatorio simple.

D) Muestreo Polietápico

Se utiliza cuando la población es muy heterogénea. Para seleccionar una muestra de una población se va dividiendo dicha población de forma sucesiva conforme algún criterio determinado con anterioridad. De las partes que resultaron de la primera división se eligen algunas por muestreo aleatorio simple. A su vez estas partes se subdividen en otras y de ellas se vuelve a seleccionar algunas, también por muestreo aleatorio simple.

Un ejemplo clásico de un muestreo polietápico resulta cuando se quiere seleccionar una muestra de una ciudad grande; la ciudad se divide en barrios y de ellos se eligen algunos por muestreo aleatorio simple; los barrios se dividen en calles, y dentro de ellas se seleccionan algunas también por muestreo aleatorio simple, y así sucesivamente.

Estadístico

Un Estadístico o Estimador de un parámetro poblacional ð desconocido es cualquier función que relaciona los

elementos de la muestra y que utilizaremos para estimar o contrastar el verdadero valor de ð.

ð = g (x1,...,xn)

Debido a que el estimador es una función de la variable aleatoria que representa la muestra genérica, será a su vez una variable aleatoria con su correspondiente distribución en el muestreo.

Distribuciones de variables discretas

Una variable aleatoria es discreta cuando el conjunto de sus valores posibles es finito, o bien en el caso de ser infinito es numerable. La distribución de probabilidad de variables aleatorias discretas se representa indicando los valores de la variable aleatoria y sus respectivas probabilidades.

En general se expresa mediante la función de distribución F(x) que es la probabilidad de que la variable aleatoria x tome un valor menor o igual que x0.

F(x) = P(x ð x0)

Distribución binomial

Una distribución sigue la ley binomial siempre y cuando se cumplan las siguientes hipótesis:

− Las observaciones se clasifican en dos categorías, que son además excluyentes. Por ejemplo, los elementos se pueden clasificar en aceptables o defectuosos

− Las observaciones son independientes. Esto significa que la probabilidad de que aparezca un elemento aceptable es siempre la misma y a su vez la probabilidad de aparición de un elemento defectuoso también se mantiene.

− La proporción de elementos de las dos categorías en las que se ha clasificado la población es siempre constante.

El modelo de la distribución binomial se aplica a:

− poblaciones finitas, de las que se toman elementos al azar, con reemplazamientos.

− poblaciones consideradas infinitas desde el punto de vista conceptual, como son las piezas que produce una máquina (defectuosas o aceptables), siempre que el resultado de cada momento sea independiente de lo ocurrido con anterioridad.

La variable binomial es una variable discreta, de parámetros "n" y "p" que toma los valores enteros:

x = 0,1,2,...,n

Sean los parámetros, "p" y "q" comprendidos entre 0 y 1, y siendo q=1−p, se cumple también que:

ðpi = 1 y ðqi = 1

La distribución binomial B (n,p) sigue la siguiente ley:

ð n ð

F(x) = ð x ð px qn−x

Es una distribución, en general, asimétrica. Sólo es simétrica cuando se verifica que p = 1/2

La media, varianza y desviación típica tienen las siguientes expresiones:

.Su media es igual al producto de los parámetros "n" y "p":

ð = np

.Su varianza es: 2 = npq

_____

.Su desviación típica es: = ð n p q

Un ejemplo de distribución binomial son los sondeos a una población cuyos individuos se dividen en dos categorías.

Distribución binomial negativa

La distribución binomial negativa permite hallar un número de "z" elementos de una categoría antes de que aparezca el primer elemento de la otra categoría.

Por ejemplo: "z" piezas aceptables antes de la k−ésima defectuosa.

ð k + n − 1 ð

p(z = x) = ð x ð pk qx

Distribuciones de variables continuas

Una variable aleatoria es continua cuando puede tomar cualquier valor dentro de un intervalo.

Para representar un distribución de probabilidad de variables continuas es necesario tener en cuenta los siguientes aspectos:

Si las medidas de una magnitud se representa en un Histograma y se van tomando cada vez más observaciones y se van disminuyendo el tamaño de las clases, dicho Histograma tiende a una curva que describe el comportamiento de la variable a largo plazo. Esta función límite recibe el nombre de Función de Densidad f (x).

El área del Histograma es la unidad, debido a que la suma de las frecuencias relativas es la unidad. En consecuencia:

ð

ð f(x) dx = 1

−ð

Distribución Normal

La Distribución Normal o de Gauss−Laplace es la distribución de probabilidad más importante, pues muchos procesos de medición sin errores sistemáticos se aproximan a ella.

La función de densidad de una distribución normal es la siguiente:

1 1

f(x)= ðððð exp[− ðððð (x − ð)2]

_____

ð 2 ð 2 2

La función depende de dos parámetros, de la media μ (que es también la moda y la mediana) y de la desviación típica . Cuando una variable sigue esta función de densidad es N (μ,).

Si la variable normal tiene media μ = 0 y desviación típica = 1 se denomina "normal estándar" N(0,1)

Para transformar una variable aleatoria normal "x" en una variable normal estándar "z", se realiza mediante la expresión:

x−ð

z = ðð

La distribución normal tiene como coeficiente de apuntamiento el valor 3, el cual se toma de referencia para juzgar otras distribuciones.

Distribución "t" de Student

La Distribución "t" de Student es una distribución continua que se define por la siguiente expresión:

x

tn = ðððð

_____

ðx2n /n

Correspondiendo x2n a la igualdad: x2n = x21 + x22 +...+ x2n, siendo "x" una variable aleatoria normal estándar y "n" los grados de libertad.

La distribución "t" de Student puede tomar valores negativos, pero, en general, sólo interesa su magnitud y no su signo.

Es una distribución simétrica y con mayor dispersión que la distribución normal. No obstante, cuando "n" es igual o mayor que 100, la distribución "t" es igual a la normal.

Su media y su varianza son respectivamente:

− media: μ = 0

− varianza: 2 = n / n+2, para n >> 2

Su función de densidad es:

1 ð((n + 1)/2) ð X2 ð −(n+1)/2

f(x) = ðð ðððððððð 1 + ðð

____ ð n ð

ð ð n ð(n/2)

Siendo ð la distribución gamma, que es una distribución particular de media y varianza iguales.

Distribución x2

La distribución x2 de Pearson es una distribución continua, y representa la distribución de una distancia. Se define la variable x2 con "n" grados de libertad, de la siguiente manera:

X2 = Z12 + Z22 + ... + Zn2

Siendo Zi2 variables aleatorias normales (0,1) e independientes.

La variable aleatoria x2 no puede ser nunca negativa, debido a que es una suma de cuadrados.

Su media y su varianza son respectivamente:

− media: μ = n

− varianza: 2 = 2n

Es una distribución asimétrica.

Su función de densidad es:

1

f(x)= ðððððð x(n/2)−1 e−x/2 x 0

2(n/2) ð(n/2)