Unidad V: Estadística aplicada

5.1 Inferencia estadística: Concepto, Estimación, Prueba de hipótesis

La Inferencia Estadística es la parte de la estadística matemática que se encarga del estudio de los métodos para la obtención del modelo de probabilidad (forma funcional y parámetros que determinan la función de distribución) que sigue una variable aleatoria de una determinada población, a través de una muestra (parte de la población) obtenida de la misma.

Los dos problemas fundamentales que estudia la inferencia estadística son el "Problema de la estimación" y el "Problema del contraste de hipótesis"

Cuando se conoce la forma funcional de la función de distribución que sigue la variable aleatoria objeto de estudio y sólo tenemos que estimar los parametros que la determinan, estamos en un problema de inferencia estadística paramétrica ; por el contrario cuando no se conoce la forma funcional de la distribución que sigue la variable aleatoria objeto de estudio, estamos ante un problema de inferencia estadística no paramétrica.

En lo que sigue nos vamos a limitar a problemas de inferencia estadística paramétrica, donde la variable aleatoria objeto de estudio sigue una distribución normal, y sólo tendremos que tratar de estimar los parámetros que la determinan, la media y la desviación típica.

Esta situación se presenta con frecuencia debido a que es posible a menudo conocer la forma funcional de la distribución de probabilidad, por consideraciones

teóricas, quedando únicamente indeterminados los parámetros que determinan la función de distribución.

Como las poblaciones en las que se pretende estudiar una determinada variable aleatoria, son grandes, es muy caro o imposible, estudiar a todos sus individuos; lo que se hace, es estudiar una muestra (una parte) de la población

En todos estos problemas que estudia la inferencia estadística juega un papel fundamental la "Teoría de la Probabilidad" (distintas formas funcionales de las distribuciones de probabilidad) y la "Teoría de Muestras" (procedimientos para tomar muestras de manera apropiada).

ESTIMACIÓN.

El material sobre teoría de la probabilidad constituye la base de la inferencia estadística, rama de la estadística que tiene que ver con el uso de los conceptos de la probabilidad para tratar con la toma de decisiones en condiciones de incertidumbre. La inferencia estadística está basada en la estimación y en la prueba de hipótesis.

Tipos de estimación.

Podemos hacer dos tipos de estimaciones concernientes a una población:

Una estimación puntual: es sólo u número que se utiliza para estimar un parámetro de población desconocido. Una estimación puntual a menudo resulta insuficiente, debido a que sólo tiene dos opciones: es correcta o está equivocada. Una estimación puntual es mucho más útil si viene acompañada por una estimación del error que podría estar implicado.

Una estimación de intervalo: es un intervalo de valores que se utiliza para estimar un parámetro de población. Esta estimación indica el error de dos maneras: por la extensión del intervalo y por la probabilidad de obtener el verdadero parámetro de la población que se encuentra dentro del intervalo.

5.2 Estimaciones puntuales y por intervalos de confianza

ESTIMACIÓN PUNTUAL

Consiste en la estimación del valor del parámetro mediante un sólo valor, obtenido de una fórmula determinada. Por ejemplo, si se pretende estimar la talla media de un determinado grupo de individuos, puede extraerse una muestra y ofrecer como estimación puntual la talla media de los individuos. Lo más importante de un estimador, es que sea un estimador eficiente. Es decir, que sea insesgado(ausencia de sesgos) y estable en el muestreo o eficiente (varianza mínima)

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

En estadística, se llama intervalo de confianza a un par de números entre los cuales se estima que estará cierto valor desconocido con una determinada probabilidad de acierto. Formalmente, estos números determinan un intervalo, que se calcula a partir de datos de una muestra, y el valor desconocido es un parámetro poblacional. La probabilidad de éxito en la estimación se representa con $1 - \alpha$ y se denomina nivel de confianza. En estas circunstancias, α es el llamado error aleatorio o nivel de significación, esto es, una medida de las posibilidades de fallar en la estimación mediante tal intervalo.

El nivel de confianza y la amplitud del intervalo varían conjuntamente, de forma que un intervalo más amplio tendrá más posibilidades de acierto (mayor nivel de

confianza), mientras que para un intervalo más pequeño, que ofrece una estimación más precisa, aumentan sus posibilidades de error.

Para la construcción de un determinado intervalo de confianza es necesario conocer la distribución teórica que sigue el parámetro a estimar, θ. Es habitual que el parámetro presente una distribución normal. También pueden construirse intervalos de confianza con ladesigualdad de Chebyshov.

En definitiva, un intervalo de confianza al 1 - α por ciento para la estimación de un parámetro poblacional θ que sigue una determinada distribución de probabilidad, es una expresión del tipo [θ1,θ2] tal que P[θ1 ≤ θ ≤ θ2] = 1 - α, donde P es la función de distribución de probabilidad de θ.

5.3 Regresión y correlación

REGRESIÓN

La regresión estadística o regresión a la media es la tendencia de una medición extrema a presentarse más cercana a la media en una segunda medición. La regresión se utiliza para predecir una medida basándonos en el conocimiento de otra.

Y_t: variable dependiente, explicada o regresando.

$X_1, X_2, \cdots, X_p$: variables explicativas, independientes o regresores.

$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$: parámetros, miden la influencia que las variables explicativas tienen sobre el regresando.

donde β_0 es la intersección o término "constante", las $\beta_i \ (i > 0)$ son los parámetros respectivos a cada variable independiente, y p es el número de parámetros independientes a tener en cuenta en la regresión. La regresión lineal puede ser contrastada con la regresión no lineal.

El modelo lineal relaciona la variable dependiente Y con K variables explicativas X_k (k = 1,...K), o cualquier transformación de éstas, que generan un hiperplano de parámetros $\beta_k$desconocidos:

$$Y = \sum \beta_k X_k + \varepsilon$$

donde ε es la perturbación aleatoria que recoge todos aquellos factores de la realidad no controlables u observables y que por tanto se asocian con el azar, y es la que confiere al modelo su carácter estocástico. En el caso más sencillo, con una sola variable explicativa, el hiperplano es una recta:

$$Y = \beta_1 + \beta_2 X_2 + \varepsilon$$