ESTADISTICA APLICADA A LA EDUCACION

Pedro Pablo Cabezas O.

I.- INTRODUCCION.

La antigüedad de la estadística se remonta a la génesis de las sociedades humanas. Empieza de una forma muy rudimentaria, lentamente va fraguando su sistematizació hasta que aparece como cátedra obligada en los albores del siglo XVII en Alemania. La sistematización conduce a la creación de tres escuelas (iniciales) debidamente caracterizadas.

La escuela Alemana que centra su interés en la información requerida para la proyección administrativa del estado. Vito de Seckendorff, Herman Conring, Godofredo de Achenwall fueron dignos representantes de la citada escuela.

La escuela Galo-Itálica considera los problemas surgidos en los juegos de azar. Trabajan ahincadamente: Pascal, Fermat, Laplace, Poisson, Bernoulli y Gauss. De ésta se inicia el estudio del Cálculo de Probabilidades.

La escuela Inglesa, se dedica a resolver situaciones de índole actuarial. Tuvo como seguidores a Petty, Halley, King y Graunt.

Las probabilidades y la estadística alcanzan un gran desarrollo con las contribuciones de Borel, Fréchet, Le-

vy, Tchebichev, Tchuprov, Markov, Kintchine, Kolmogorov, etc.

La categoría de ciencia, que tiene la estadística, se debe al impresionante empuje brindado por investigadores como R. A. Fisher, Gosset, Galton, Pearson en áreas agronómicas, biométricas, sicológicas, según sus respectivas especializaciones.

En esta brevísima introducción, es conveniente registrar los aportes de escuelas como la Escandinava, la Hindú y la Norteamericana, entre otras, que han permitido la ubicación de la estadística en un sitial de honor científico.

Hoy existen en los cinco continentes, Instituciones especializadas que propenden por el avance y buen uso de los conocimientos estadísticos.

En el caso específico de la pedagogía, Wundt, en 1789 introduce la disciplina estadística a la sicología y por ende a aquella. El proceso continúa con Francis Galton, Karl Pearson, James Cattell, Tifchener, Lee, Thorndike, Binet, Simons, Judd, entre otros.

Resumiendo se puede afirmar que la estadística y su aplicación se originan con la estructura social humana, se sistematiza a través del tiempo, posteriormente su carácter se hizo más comprensivo y actualmente se emplea en el estudio de todas las formas de vida social.

El uso de las técnicas estadísticas en la educación, tiene que ver con tres aspectos fundamentales:

1. El análisis de datos recolectados para investigaciones y desarrollo de experimentos.
2. La construcción de instrumentos de medición.
3. La búsqueda de factores causales subyacentes.

Los dos primeros aspectos posiblemente sean tratados en otros artículos de esta revista. En éste se abordará el mencionado en el punto tercero.

II.- BUSQUEDA DE FACTORES CAUSALES.

La estadística desarrolla dos metodologías para la búsqueda de esos factores, son ellos:

1. El análisis de asociación, y
2. El análisis factorial.

El análisis de asociación es la metodología estadística que permite manejar relaciones vagas, borrosas e imprecisas entre variables. La vaguedad es el fundamento diferencial de este análisis con la disciplina matemática.

El análisis factorial es la aplicación sistemática del método que sirve para obtener ciertos tipos de contrastes ortogonales entre los caracteres (tratamientos).

La desventaja del análisis de asociación frente al análisis factorial se debe al hecho de que "exista o no una relación causal entre las variables" se puede aplicar aquel, en cambio, éste, por la ortogonalidad, facilita el manejo de "la relación causal". Por lo tanto, las investigaciones pedagógicas se están elaborando bajo la óptica del análisis factorial.

Lo aconsejable sería explicar las dos teorías, pero debido a la imposiblidad temporal, se tratará de informar, de una manera suscinta, en qué consiste el marco teórico del análisis de asociación.

El análisis de asociación se divide en dos subtemas, tan distintos como se quiera, pero íntimamente relacionados. Estos son: el análisis de regresión y el de correlación. Ambos tienen sus correspondientes teorías, pero debido al nexo, una vez desarrollada la de la regresión, se puede definir la correlación y establecer las fórmulas correspondientes.

El análisis de regresión.

La palabra regresión proviene del latín (regretio, onis) y la definen los diccionarios como "retrocesión" o "acción de volver hacia atras". Su uso se ha extendido a otras ramas de la ciencia como la antropología, la geología dinámica, la historia natural, la patología, la biología, etc.

Galton, en sus estudios sobre la herencia, es quien desarrolla esta terminología en la experimentación estadística. Acerca de la "ley de Regresión Universal" dijo: "Cada peculiaridad de un hombre está compartida por su semejante, pero en promedio, en un grado menor". Experimentos posteriores llevaron a Karl Pearson a concluír que aunque padres altos tienden a tener hijos altos, sin embargo, la altura promedio de hijos de un grupo de padres altos, es menor que la altura promedio de sus padres, esto es, existe una regresión, un retroceso, un

volver hacia atrás, de la estatura de los hijos hacia la estatura de todos los padres.

Con el progreso estadístico, no se queda la regresión dentro de ese contexto, sino que evoluciona y se asocia con la palabra análisis para estudiar o establecer relaciones causales (funcionales) y permitirle a la estadística que cumpla con la misión de toda ciencia "la de predecir" igual que la de "pronosticar".

Entonces, el análisis de regresión pretende establecer modelos matemáticos, a partir de modelos estadísticos, para expresar la funcionalidad entre variables concomitantes.

Ya se ha dicho que la diferencia entre el modelo estadístico y el matemático, radica en la "vaguedad". Aquí se definirá con mayor propiedad y se dirá que es la "aleatoriedad", o sea, la expresión

$$Y = f(x_1, \ldots, x_n / \beta_1, \ldots, \beta_m)$$

donde las x_i ($i = 1, \ldots, n$) son variables aleatorias y β_i ($i = 1, \ldots, m$) son parámetros, es un modelo matemático. En cambio

$$Y^* = f_1(Y, \gamma)$$

donde Y es un modelo matemático y γ es una variable aleatoria, es un modelo estadístico.

Con el propósito de aprovechar, en parte, la infraestructura del álgebra lineal, las técnicas de computación y de ser base para los demás modelos, se ha estudiado extensamente el multilineal (caso particular el lineal), teniendo en cuenta restricciones e hipótesis (ante todo

las de normalidad para poblaciones, subpoblaciones y para la variable aleatoria de perturbación).

En un modelo de regresión lineal multivariante, una variable Y (dependiente) o "explicada", se relaciona con unas variables X_j (j = 2, ... , m), (independientes o "explicativas" por la siguiente expresión:

$$y_i = \beta_1 + \sum_{j=2}^{m} \beta_j x_{ij} + \mu_i; \quad i=1,\ldots,n \qquad (1)$$

donde las β_j (j = 2, ... , m) son desconocidas y se las denomina coeficientes de regresión (parciales) poblacionales; μ_i es el "transtorno" al azar o residual.

La expresión (1) significa que la parte de la variación total que es explicada sistemáticamente, está representada por:

$$\beta_1 + \sum_{j=2}^{m} \beta_j x_{ij}; \quad i=1,\ldots,n \qquad (1')$$

que es el correspondiente modelo matemático.

Algunos supuestos para (1) son:

1. Las X_j variables independientes (j=2,...,m) no son aleatorias.
2. Las $\mu_i \sim N(0, \sigma^2), \forall\, i \in I$, son mutuamente independientes.
3. Debe cumplirse la homoescedasticidad.
4. El número de observaciones muestrales debe superar al número de coeficientes de regresión por estimarse.

Teniendo en cuenta los supuestos, la distribución de Y es condicional ($f(y_i / x_{i2}, x_{i3}, \ldots, x_{im})$) con una ex

pectativa (esperanza) condicional igual a la parte sistemática, su notación es:

$$E(y_i / x_{i2}, x_{i3}, \dots, x_{im}) = \mu_{1,2,\dots,m}$$

$$= \beta_1 + \sum_{j=2}^{m} \beta_j x_{ij} \; ; \; i=1,\dots,n \qquad (1'')$$

A esta ecuación se la denomina de regresión multilineal poblacional y corresponde a la de un hiperplano de regresión; β_1 es la constante regresional y los β_j ($j = 2, \dots, m$) las pendientes de regresión del hiperplano o como se dijo antes, coeficientes de regresión parciales.

La variancia condicional de Y o del hiperplano regresional, es igual a la del azar, de acuerdo a la homoescedasticidad, dada la parte sistemática, o sea,

$$V(y_i / x_{i2}, \dots, x_{im}) = \sigma^2_{1,2,\dots,m}$$

$$= E(y_i - \mu_{1,2,\dots,n}) = \sigma^2$$

Se ha dicho que los β_j ($j=2,\dots,m$) son desconocidos tanto en (1) como en (1'). En consecuencia, es conveniente estimarlos a partir de muestras provenientes de poblaciones a las cuales está referida la investigación. Las técnicas del muestreo permiten la consecución de muestras representativas y confiables.

La muestra simple al azar extraida de una población de m variables, proporciona nm-tuplas

$$(x_{i2}, x_{i3}, \dots, x_{im}, y_i), \quad i = 1,\dots,n$$

y el modelo estadístico de regresión multilineal mues-

tral es:

$$y_i = b_1 + \sum_{j=2}^{m} b_j x_{ij} + e_i, \quad i = 1,\ldots,n \qquad (2)$$

el correspondiente matemático es:

$$\hat{y}_i = b_1 + \sum_{j=2}^{m} b_j x_{ij}, \quad i = 1,\ldots,n \qquad (2')$$

de (2) y (2') se deduce que

$$y_i - \hat{y}_i = e_i \quad i = 1,\ldots,n$$

donde e_i $(i = 1,\ldots,n)$ son los respectivos estimadores de μ_i $(i = 1,\ldots,n)$. Los b_j $(j = 1,\ldots,m)$ son los correspondientes estimadores de los β_j $(j = 1,\ldots,m)$. El símbolo $\wedge$ significa estimador.

Hay varios procesos para calcular "buenos" estimadores de aquellos parámetros. El método de la probabilidad máxima (máxima verosimilitud) y el de los mínimos cuadrados son dos métodos muy utilizados.

El método de mínimos cuadrados consiste en minimizar la expresión:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i)^2$$

cuyo resultado da el siguiente conjunto o sistema de ecuaciones "normales".

$$\begin{aligned}
\sum_{i=1}^{n} Y_i &= nb_1 + \sum_{j=2}^{m} b_j \sum_{i=1}^{n} x_{ij} \\
\sum_{i=1}^{n} X_{i1}Y_i &= b_1 \sum_{i=1}^{n} x_{i1} + \sum_{j=2}^{m} b_j \sum_{i=1}^{n} x_{i1}x_{ij} \\
\sum_{i=1}^{n} X_{i2}Y_i &= b_1 \sum_{i=1}^{n} x_{i2} + \sum_{j=2}^{m} b_j \sum_{i=1}^{n} x_{i2}x_{ij} \qquad (3) \\
&\vdots \qquad\qquad\qquad \vdots \\
\sum_{i=1}^{n} X_{im}Y_i &= b_1 \sum_{i=1}^{n} x_{im} + \sum_{j=2}^{m} b_j \sum_{i=1}^{n} x_{im}x_{ij}
\end{aligned}$$

El conjunto solución del sistema (3), asegurada la compatibilidad, cuyos elementos son los b_j $(j = 1,\ldots,m)$ estimadores de los β_j $(j = 1,\ldots,m)$, se pueden calcular por cualquiera de los métodos conocidos o por computadora.

De la primera ecuación de (3) se deduce que

$$b_1 = \bar{Y} - \sum_{j=2}^{m} b_j \bar{x}_j \qquad (3')$$

Por una traslación de $P(0, 0, \ldots, 0)$, (punto origen), al punto $P(\bar{x}_2, \bar{x}_3, \ldots, \bar{x}_m, \bar{y})$, el sistema (3) se reduce al sistema:

$$\begin{aligned}
\sum_{i=1}^{n} X'_{i1}Y'_i &= \sum_{j=2}^{m} b_j \sum_{i=1}^{n} x'_{i1}x'_{ij} \\
\sum_{i=1}^{n} X_{i2}Y_i &= \sum_{j=2}^{m} b_j \sum_{i=1}^{n} x'_{i2}x'_{ij} \qquad (3'') \\
&\vdots \qquad\qquad \vdots \\
\sum_{i=1}^{n} X'_{im}Y'_i &= \sum_{j=2}^{m} b_j \sum_{i=1}^{n} x'_{im}x'_{ij}
\end{aligned}$$

El conjunto solución unido al obtenido en (3') es igual al de (3). Es de advertir que los $b_j = \hat{\beta}_j$ quedan expresados en función de los valores observados en la muestra.

Calculados los estimadores b_j ($j = 1,\ldots,m$), el modelo de regresión multilineal utilizado para la predicción y la pronosticación se expresa como:

$$\hat{Y}_i = b_1 + \sum_{j=1}^{m} b_j x_{ij}$$

$$\hat{Y} = b_1 + \sum_{j=1}^{m} b_j x_j$$

Establecido el modelo, es conveniente discutir su fiabilidad para evitar equivocaciones que puedan originarse por una mala predicción o pronosticación. Lo anterior equivale a examinar si el ajuste regresional multilineal es bueno o no.

Las estimaciones puntuales o por intervalos, las pruebas de hipótesis y la construcción de bandas, son algunas de las metodologías para comprobar la bondad del ajuste.

El coeficiente de determinación poblacional (multilineal), (se lo simboliza por ρ^2 y el muestral R^2), también, aconseja o no la aplicabilidad del modelo. En efecto, éste se define como la medida que determina el grado del ajuste del hiperplano regresional a los puntos reales del espacio m-dimensional.

Para la definición analítica se tiene en cuenta la siguiente propiedad:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (5)$$

El primer miembro de (5) es la suma de cuadrados de las desviaciones totales respecto a su media (S C T). El primer término del segundo miembro es la suma de cuadrados debida a la regresión (SCR) o sea, es la parte del error total que es explicado por la regresión. El segundo término es la suma de cuadrados por error (S C E) o la parte del error total que no es explicado por la regresión, por lo tanto, queda sin explicar.

La ecuación (5) se puede escribir como:

$$S\,C\,T = S\,C\,R + S\,C\,E \qquad (5')$$

Al dividir ambos miembros de (5') entre S C T se obtiene:

$$\frac{S\,C\,R}{S\,C\,T} = 1 - \frac{S\,C\,E}{S\,C\,T} \qquad (5'')$$

El primer miembro de (5") define el coeficiente de determinación, esto es:

$$R^2 = \frac{S\,C\,R}{S\,C\,T} = 1 - \frac{S\,C\,E}{S\,C\,T} \qquad (5''')$$

Analizando (5''') se puede concluír que $0 \leq R^2 \leq 1$. Si $R^2 = 0$, SCR = 0 (entonces S C T = S C E) lo que implica que la ecuación de regresión no explica nada. Cuando $R^2 = 1$, S C R = S C T y S C E = 0, esto indica que la ecuación de regresión explica el total del error.

Para R^2 cercano a uno el ajuste es bueno y para valo-

res de R^2 por debajo de 0.5 no es aconsejable la ecuación regresional. De todas maneras lo subjetivo del concepto "cercano" no se puede eludir.

A continuación se escriben los modelos lineales bi y trivariables para efectos de concretizar la operatividad.

Modelos estadístico y matemático de regresión lineal bivariable:

$$Y_i = \beta_1 + \beta_2 x_{i2} + \mu_i \qquad i=1,\ldots,n$$

$$\hat{Y}_i = b_1 + b_2 x_{i2} \qquad i=1,\ldots,n$$

$$b_1 = \bar{y} - b_2 \bar{x} \qquad b_2 = m_{12} / m_{22}$$

$$m_{12} = \sum_{i=1}^{n} y_i x_i - n \bar{y} \bar{x}, \quad m_{22} = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

$$r^2 = b_2^2 \frac{m_{22}}{m_{11}} \qquad r = b_2 \sqrt{m_{22} / m_{11}}$$

r^2 es la notación del coeficiente de determinación muestral lineal y r el coeficiente de correlación muestral, éste se tratará posteriormente.

Modelos estadístico y matemático de regresión lineal tri-variable:

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \mu_i$$

$$Y_i = b_1 + b_2 x_{i2} + b_3 x_{i3}$$

$$b_1 = \bar{Y} - b_2 \bar{x}_2 - b_3 \bar{x}_3$$

$$b_2 = (m_{12} m_{33} - m_{13} m_{23}) / (m_{22} m_{33} - m_{23}^2)$$

$$b_3 = (m_{13}m_{22} - m_{12}m_{23})/(m_{22}m_{33} - m_{23}^2)$$

$$m_{1k} = \sum_i (Y_i - \bar{Y})(X_{ik} - \bar{X}_k) \qquad k = 2,3$$

$$m_{jk} = \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \qquad j,k = 2,3$$

$$R^2 = \frac{S\ C\ R}{S\ C\ T} = 1 - \frac{S\ C\ E}{S\ C\ T}$$

Los modelos de regresión curvilineal más conocidos son:

1. El polinómico:

$$Y = \sum_{k=0}^{m} \beta_k x^k$$

2. El exponencial simple o de interés compuesto:

$$Y = \beta_1 \beta_2^x \qquad \beta_i > 0, \quad i = 1,2$$

3. El de Gompertz:

$$\text{Ln}Y = \text{Ln}\,\beta_1 + (\text{Ln}\,\beta_2)\beta_3^x, \quad \beta_i > 0, \quad i = 1,2,3$$

4. El logístico:

$$\frac{1}{Y} = \beta_1 + \beta_2 \beta_3^x \qquad \beta_i > 0, \quad i = 1,2,3$$

5. El de Mitscherlich:

$$Y = \beta_1(1 - e^{\beta 2 - \beta_3 x}), \quad \beta_i > 0, \quad i = 1,2,3$$

6. El de la función recíproca:

$$Y = \beta_1 \pm \frac{\beta_2}{x}, \quad \beta_i, \ x > 0, \ i=1,2, \quad x \geqslant \frac{\beta_2}{\beta_1} \text{ para} -$$

Modelos que por transformaciones se los reduce al modelo de regresión multilineal o lineal, situación similar se presenta con el modelo multicurvilineal.

Todo lo tratado en esta parte constituye una introducción al análisis de regresión. Al final se incluye una bibliografía de consulta para quienes estén interesados en el tema.

El análisis de correlación.

Es la metodología desarrollada para establecer el grado de relación que existe entre variables concomitantes.

Las notaciones para los coeficientes que miden ese grado de relación son ρ , r, para el poblacional y el muestral, respectivamente. En el caso lineal bivariable y dentro de ciertos supuestos, se define ρ así:

$$\rho = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \qquad \text{y}$$

$$r = \hat{\rho} = b_2\sqrt{m_{22}/m_{11}} = \frac{m_{12}}{\sqrt{m_{11}\, m_{22}^2}}$$

Tratándose del modelo lineal trivariable los coeficientes de correlaciones parciales muestrales se los define de la siguiente manera:

$$r_{12.3} = (r_{12} - r_{13}r_{23})/\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}$$

$$r_{13.2} = (r_{13} - r_{12}r_{23})/\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}$$

$$r_{23.1} = (r_{23} - r_{12}r_{13})/\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}$$

donde

$$r_{12} = m_{12} / \sqrt{m_{11}\, m_{22}}$$

$$r_{13} = m_{13} / \sqrt{m_{11}\, m_{33}}$$

$$r_{23} = m_{23} / \sqrt{m_{22}\, m_{33}}$$

son los coeficientes de correlaciones simples. Los números que están a la derecha de los puntos indican las variables que se mantienen fijas en determinado cálculo.

De la definición de ρ y r se concluye que pueden tomar todos los valores entre -1 y 1. Cuando r = 0 no hay correlación entre x y y; r = 1 implica relación perfecta directa; r = -1 dice que la relación es perfecta inversa, presentando ambigüedad en cuanto al término "cerca" simulando al coeficiente de determinación.

Esta coincidencia no es casual, lo que ocurre es que el coeficiente de correlación no es otra cosa que la raíz cuadrada de aquel coeficiente.

La fiabilidad o no de los coeficientes de correlación parciales o simples, se pueden comprobar por inferencias, una vez conocidas sus varianzas y distribuciones.

III.- COMENTARIOS

La aplicación del análisis de asociación en el campo educativo, radica en el hecho de que, siendo el maestro uno de quienes estudian la variabilidad humana, entendida esta como "la medición de la superposición de los diversos grupos sucesivos de edad en una función determi-

nada", deseará saber, hasta lo permisible, todo lo referente a las características individuales o grupales (según sea el caso), ya que el problema de educar a la humanidad se torna complejo justamente por la heterogeneidad.

Las características individuales o de grupo se pueden representar por variables (endógenas o exógenas) y sus mediciones serían las observaciones muestrales.

Ejemplo: en la dotación intelectual (variable dependiente) influyen características como: el carácter, el temperamento, el interés, la salud, la fuerza ambiental, entre muchas, éstas serían las variables independientes.

Identificadas las variables, el análisis de asociación determinará la relación (causal) y el grado de dicha relación. Además, establecerá el modelo correspondiente para predecir características del grupo o pronosticar para determinados elementos.

El rendimiento individual o de grupo es causa primera y última del trabajo profesoral. Establecidos los objetivos (generales y específicos), es necesario asesorarse de un modelo regresional que le facilite su papel visionario en este aspecto.

No faltan quienes hacen depender el rendimiento educativo, exclusivamente, del cociente intelectual (CI), definido como la razón entre la edad mental (E.M) y la edad cronológica (E.C).

El CI puede y debe ser utilizado para predecir el futuro desenvolvimiento individual (grupal), pero no debe ser la única variable, porque, seguramente, al realizar

un análisis de regresión bivariable entre el rendimiento (v.dependiente) y el CI (v. independiente), S C R no va a explicar el error total (S C R ≠ S C T). Variables como salud, aptitudes, adaptación social, interés, ambiente familiar, dieta alimenticia, preparación profesoral, espacio físico, etc., debieran tenerse en cuenta para elaborar el modelo.

El número de variables no debe ser un factor determinante para establecer restricciones, porque aplicando regresión gradual se pueden seleccionar las más influyentes.

Estudios de esta naturaleza y más avanzados se han hecho y se están haciendo, como lo manifestaba al inicio de este artículo. Lo que está por hacerse es "la construcción de un modelo de regresión para el rendimiento académico en el departamento de Nariño, incluyendo como variables todas las ventajas y desventajas de la zona" Esta la inquietud que quiero dejar en los lectores de esta revista.

BIBLIOGRAFIA

1. Chou, Ya Lin. Análisis Estadístico, segunda edición, Interamericana, México, 1977

2. Graybill Franklin A. An Introduction to Linear Statistical Models, Vol. I, McGraw-Hill Book Company, Ind. New, York, 1961.

3. Li Ching Chun. Introducción a la Estadística Experimental, Ediciones Omega, S.A., Barcelona, 1969.

4. Snedecor George W. Métodos Estadísticos. Compañía Editorial Continental, S.A., México, 1964.

5. Casado Enrique. Estadística General. Cienes, Santiago, Chile, 1970.