LÓGICA CENTRAL DE LOS PRINCIPALES MÉTODOS ESTADÍSTICOS:

El anova, La prueba t, La correlación Y la regresión

Por: CUADROS, Jaime*

* Magíster en Docencia Universitaria de la Universidad Pedagógica Nacional. Esp. En Estadística de la Universidad Nacional de Colombia. Licenciado en Matemáticas y Estadística de la UPTC. Profesor de estadística y matemática del colegio Boyacá de Tunja, Profesor de estadística, matemática e investigación en la UPTC Tunja y Universidad Antonio Nariño, Escuela superior de Administración Pública Regional Boyacá Casanare, Fundación Universitaria Juan de Castellanos y Fundación Universitaria Monserrate. E-mail: jcuadros@telecorp.net

RESUMEN

El propósito del artículo es unificar los conocimientos acerca de los principales métodos estadísticos. El modelo lineal general equipara el valor de una variable con la suma de una constante, más la influencia parcial y ponderada de cada una de las otras variables, más el error. El coeficiente de correlación y la REG/CORR.MÚLT. (y las correspondientes pruebas de significación), la prueba t y el ANOVA, son todos casos especiales del modelo lineal general.

Palabras clave: GLM, Regresión/Correlación Múltiple, Prueba t, Anova.

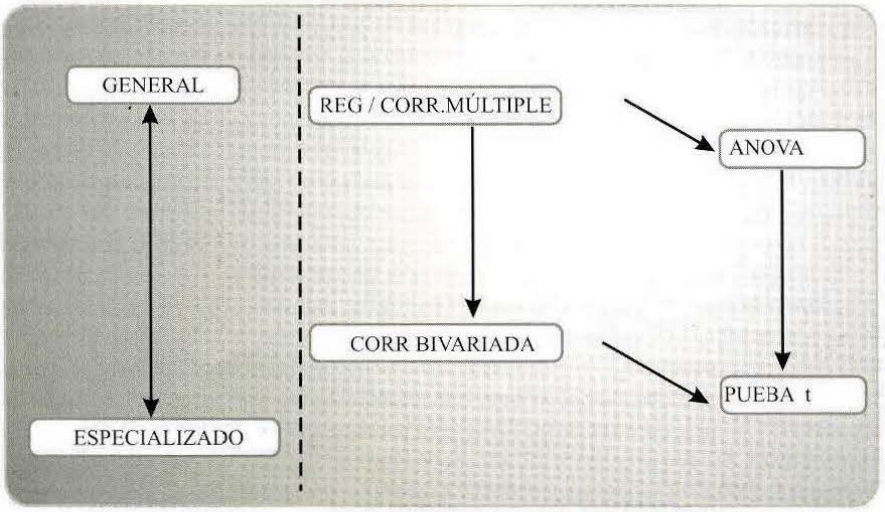
ABSTRACT

The article purpose is to unify the knowledge about the main statistical methods. The general lineal pattern put on the same level the value of a variable to the sum of a constant, plus the partial and pondered influence of each one of the other variables, plus the error. The correlation's coefficient and the REG/CORR.MÚLT (and the corresponding tests of significance), the t test and the ANOVA, they are all special cases of the general lineal pattern.

Keywords: GLM, Regression / Multiple Correlation, t Prove, Anova.

Introducción

Un alto índice de publicaciones emplean pruebas t, análisis de varianza, correlación o regresión múltiple; probablemente, se han hecho evidentes muchas semejanzas entre estos cuatro métodos. De hecho, éstos están más relacionados de lo que podría creerse: no son más que simples variaciones matemáticamente equivalentes entre sí y la mayoría tienen su origen en la misma fórmula general. Lo anterior se debe a que hay una lógica central que los sustenta y se basa en una fórmula general denominada modelo lineal general (GLM).



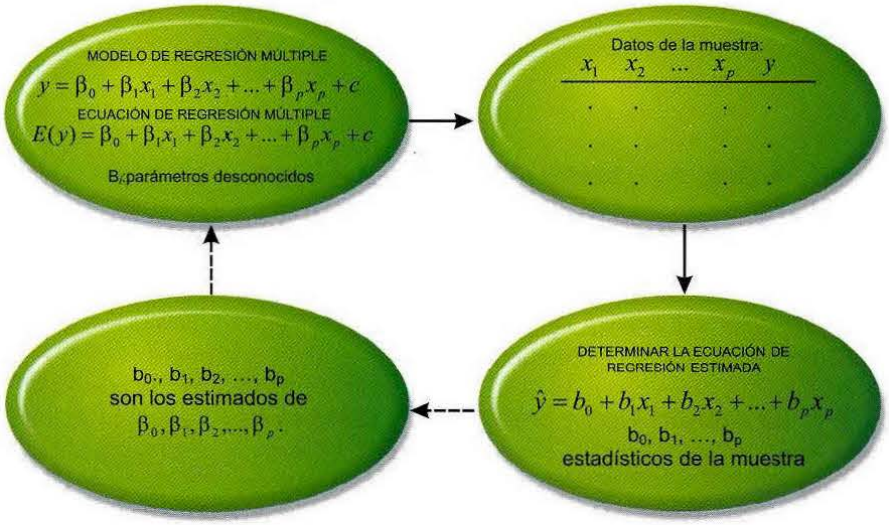
El método general es la REG/CORR.MÚLTIPLE; la correlación bivariada es un caso especial de la misma. La prueba t deriva directamente de la correlación bivariada o del ANOVA.

Cuando se dice que un procedimiento es un caso especial de otro, significa que el primero puede deducirse de la fórmula del segundo. Por eso, cuando se usan los métodos más especializados se obtiene el mismo resultado de manera general.

Un programa que realize REG/CORR.MÚLTIPLE puede lograr todo lo que se obtiene con programas más especializados de correlación bivariada, pruebas t y ANOVA.

Regresión/ correlación múltiple

La regresión múltiple es aquella situación en la que se predice el valor de una variable basándose en dos o más variables de predicción (independientes o explicativas). Se pueden crear normas de predicción para valores z y originales; estos últimos facilitan la relación con el modelo lineal general (GLM).



El método de los mínimos cuadrados $[\min \sum (y_i - \hat{y}_i)^2]^1$ usa datos de la muestra para determinar los valores de b_0, b_1, b_2...bp, que hacen que la suma de los residuales elevados al cuadrado sea mínima. En la regresión múltiple, la deducción de las fórmulas de los coeficientes b_0, b_1, b_2...,bp requiere del álgebra matricial o de paquetes estadísticos para obtener la ecuación estimada.

También es posible describir el grado general de relación entre la variable independiente (valor esperado o repuesta) y la combinación de las de predicción. Este dato se denomina COEFICIENTE DE CORRELACIÓN MULTIPLE "R", y debe ser al menos tan grande como la correlación bivariada más pequeña entre cualquiera de las variables de predicción y la variable respuesta. R2 es la reducción proporcional del error cuadrático lograda, utilizando la regla de predicción para regresión múltiple, en contraposición con la simple predicción de la variable dependiente a partir de su propia media.

Se puede probar la SIGNIFICACIÓN de una correlación múltiple (y de la correspondiente reducción proporcional del error) utilizando un procedimiento en el que la hipótesis nula establece que la correlación es cero.

Modelo lineal general (GLM)

Una forma de expresar el GLM es viéndolo como una relación matemática entre una variable respuesta y una o más variables de predicción más otras influencias no medidas, que son las que producen el error.

El principio básico establece que el valor de una variable respuesta es la consecuencia de la suma de varias influencias:

I. Cierta influencia fija β_0.

II. Influencias de otras variables $\beta_1 x_1, \beta_2 x_2, ..., \beta_p x_p$.

III. Otras influencias no medidas, que producen el error "ε "aleatorio.

Si existiera una correlación múltiple de 1,00; no existiría la influencia III.

Así, el GLM se puede expresar como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon \quad \text{donde} \varepsilon$$

es lo que queda después de tener en cuenta todos los demás elementos de predicción.

La fórmula procedente es casi idéntica a la de REG. MULTIPLE, pero con dos excepciones:

I. En lugar del valor y predicho "$\hat{y}$", tenemos el valor real y

II. Incluye el término de error "ε", debido, precisamente, a que la fórmula busca el valor real de y.

El GLM es la enunciación de las influencias que forman el valor de una respuesta en una variable determinada. Se denomina MODELO LINEAL, porque si se realiza un gráfico de la relación entre las variables respuesta y de predicción, la figura formada sería una línea recta, esto es, relación constante. La influencia que actúa como tasa de cambio (el coeficiente de regresión) de cada variable de predicción siempre es la misma.

GLM y regresión/ correlación múltiples

El vínculo entre GLM y la REG/CORR.MÚLTIPLE es muy estrecho; son prácticamente lo mismo. Tradicionalmente no se equiparon porque se consideraba que el GLM estaba implícito en otras técnicas, tales como la correlación bivariada y el ANOVA, además de la REG/CORR.MÚLTIPLE. Sin embargo, en los últimos años, los estudiosos han advertido que estas otras técnicas pueden derivar de la REG / CORR.MÚLTIPLE al igual que del GLM.

Reg/corr. bivariada como casos especiales de reg/corr.múltiple

La regresión bivariada; es decir, la predicción de una variable respuesta a partir de una variable de predicción, es un caso especial de regresión múltiple, la predicción de una variable dependiente a partir de una cantidad cualquiera de variables de predicción. Así mismo, la correlación bivariada, la relación entre una variable de predicción y una variable respuesta, es un caso especial de correlación múltiple, la relación entre una cantidad cualquiera de variables de predicción y una dependiente.

La prueba t como caso especial del anova

La relación del GLM con la CORR. y la REG. es bastante directa. El vínculo del GLM (o de la CORR y la REG) con la prueba t y el ANOVA es menos directo. Sin embargo, tanto la prueba t como el ANOVA son procedimientos para probar la diferencia entre medias de grupos. La prueba t se utiliza cuando existen sólo dos grupos. El ANOVA con

razón F, se utiliza cuando existen más de dos grupos. No hay motivo para no emplear un ANOVA sólo con dos grupos.

Las pruebas t y F son estrictamente idénticas sólo cuando se trabaja con dos grupos. Cuando existen más, no se puede realizar una prueba t ordinaria; es decir, ésta es un caso especial del ANOVA.

La idea es que la razón F del ANOVA es una medida del grado en el cual la señal (análoga a la diferencia entre los medias de grupo) excede el ruido (análogo a la variación interna de cada uno de las grupos). La misma idea se aplica a la prueba t, que también determina el grado en el cual la señal (la diferencia entre las medias de los dos grupos) excede el ruido (el desvió estándar de la distribución de diferencias de medias, que también se basa en la variación interna de los grupos).

Paralelismo entre la lógica básica de los dos métodos

El ANOVA se basa en el cálculo de una razón F (que después se compara con la F tabulada). Ésta es la estimación de la varianza poblacional centrada en la variación entre las medias de dos o más grupos y dividida por la estimación de la varianza poblacional de cada uno de éstos.

La prueba t se basa en el cálculo de un valor t (que después se compara con un punto de corte previamente definido, tomado de una tabla con una distribución t). Éste es la diferencia entre las medias de los dos grupos dividida por el desvió estándar de la distribución de diferencias de medias, el cual se calcula utilizando una estimación combinada de la varianza. En conclusión, tanto una razón F como un valor t son fracciones en las cuales el numerador se basa en las diferencias entre las medias de los grupos y el denominador en las varianzas dentro de los mismos.

Relación matemática entre los dos métodos

En los casos en los que hay sólo dos grupos, la fórmula para calcular el valor t es precisamente la raíz cuadrada de la fórmula para la razón F.

Un aspecto particular de la equivalencia matemática de t y F ayudará a comprender el modo en que dos series de cálculos, aparentemente diferentes, encierran en realidad lo mismo. Una situación con estas características es el modo en que los afecta el tamaño de la muestra. En el ANOVA, éste es parte del numerador. El numerador de la razón F es la estimación de la varianza poblacional que utiliza la diferencia entre las medias multiplicada por la cantidad de observaciones en cada grupo.

En la prueba t, el tamaño de la muestra es parte del denominador, pues utiliza la estimación combinada de la varianza poblacional dividida por la cantidad de observaciones de cada grupo. Esta aparente contradicción se resuelve, porque multiplicar el numerador de una fracción por un número tiene exactamente el mismo efecto que dividir el denominador por ese mismo número.

Otras diferencias aparentes (como la dada entre el numerador de la razón F, que se basa en una estimación de varianza, y el numerador del punto t, que es una simple diferencia entre medias) presentan una unidad subyacente similar.

ALGUNOS VÍNCULOS DE LA PRUEBA t PARA MEDIAS INDEP. Y ANOVA

PRUEBA t

- El numerador de t es la diferencia entre las medias de dos grupos.
- El denominador de t se basa en la combinación de las estimaciones de varianza poblacional calculada a partir de cada grupo.
- El denominador de t implica dividir por la cantidad de registros.
- Cuando se utilizan dos grupos: $t = \sqrt{F}$; $gl = (n_1 - 1) + (n_2 - 1)$.

ANOVA

- El numerador de F se basa en la variación entre las medias de dos o más grupos.
- El denominador de F se calcula combinando las estimaciones de varianza poblacional a partir de cada grupo.
- El numerador de F involucra la multiplicación por la cantidad de observaciones (mismo efecto t).
- Cuando se utilizan dos grupos:

  $F = t^2$; $gl_{dentro} = (n_1 - 1) + (n_2 - 1) + \ldots$

La prueba t como caso especial de la prueba de significación del coeficiente de correlación

El coeficiente de correlación es el grado de relación entre dos variables; la prueba t trata sobre la significación de la diferencia entre dos medias poblacionales ¿Cuál es la conexión posible?.

Una conexión se da en el empleo de la distribución t para determinar la significación. Analizando la lógica de las pruebas de hipótesis, se tiene:

I. La H0 establece que la población tiene una correlación igual a 0.
II. La distribución comparativa es una t con tantos gl como la cantidad de observaciones menos dos.
III. El valor en la distribución comparativa es un t_c a partir del coeficiente de correlación utilizando:

$$t = r\sqrt{n-2} / \sqrt{1-r^2}$$

Es importante señalar que la clave de todo el proceso es convertir el coeficiente de correlación en un valor t.

Un coeficiente de correlación significativo indica que la variable de predicción y la respuesta están relacionadas. Una prueba t de medias independientes, que resulta significativa, indica que la variable de predicción y la respuesta están relacionadas; ambas indican lo mismo.

La prueba t es un caso especial del coeficiente de correlación, porque ésta es sólo una instancia particular del coeficiente de correlación; es decir, es la situación en la que la variable de predicción tiene sólo dos valores.

El anova como caso especial de la prueba de significación del coeficiente de correlación múltiple

La relación entre el ANOVA y la CORR.MÚLT es paralela a la relación que se acaba de presentar entre la prueba t para medias independientes y el coeficiente de correlación (bivariado) ordinario. En ambas relaciones, uno de los dos estadísticos parece referirse a las diferencias entre medias y el otro a las asociaciones entre variables.

La resolución de esta diferencia aparente es la misma.

El ANOVA analiza si existe una diferencia, en la variable respuesta, entre las medias de los grupos que representan diferentes niveles de una variable de predicción. El método de la CORR. encara la situación como una relación entre la variable RTA. y los diferentes niveles de la variable EXPLICATIVA.

El vínculo entre el ANOVA y la CORR. es más fácil de captar si se interpreta el coeficiente de CORR. como la raíz cuadrada de la reducción proporcional del error con observaciones originales, y al ANOVA con el método del modelo estructural.

La suma de los errores cuadráticos, calculada en la correlación cuando se utiliza la regla de predicción bivariada, SCerror, es igual a la suma de desvíos cuadráticos intragrupales, SCdentro, correspondientes al ANOVA. ¿Por qué son iguales? El ANCORR. está calculando el error como la diferencia con respecto al valor predicho, y éste es la media de cada grupo; es decir, en el ANCORR. la suma de los errores cuadráticos es el resultado de elevar al cuadrado y sumar la diferencia entre cada valor y la media de su grupo (que es la predicción para cada registro en su grupo). El ANOVA está calculando la suma de los errores cuadráticos intragrupales exactamente del mismo modo, la suma de los desvíos cuadráticos de cada observación con respecto a la media de su grupo.

De otro lado, la suma de los errores cuadráticos en el ANCORR, cuando para predecir utiliza la media general de la variable RTA. (SCTotal) es igual a SCTotal en el ANOVA. Son iguales porque el ANCORR. está determinando este error como el desvío cuadrático de cada observación con respecto a la media general de todas las observaciones de la variable RTA, y el ANOVA está calculando la suma de los desvíos cuadráticos de cada observación respecto a la gran media.

Además, la reducción del error cuadrático divide la suma de cuadrados empleando la media para predecir, menos la suma de cuadrados del error, utilizando la regla de predicción bivariada, que coincide con la suma de cuadrados intergrupales (SCentre) en el ANOVA. La reducción de error en el ANCORR es equivalente a lo que agrega la regla de predicción con respecto a conocer sólo la media.

En este caso, la recta de predicción estima la media de cada grupo; por lo tanto, la reducción de error cuadrático de cada observación es la diferencia cuadrática entre la media del grupo y la general. SCentre en el ANOVA, se calcula sumando, las diferencias cuadráticas entre la media del grupo y la gran media.

Finalmente, la reducción proporcional del error (r2, también denominada proporción de varianza explicada), en el ANCORR, es exactamente igual a la proporción de varianza explicada (R2 o), una de las medidas del tamaño del efecto que se estudia en el ANOVA.

Anova para más de dos grupos como caso especial de correlación múltiple

En un ANOVA se puede codificar toda variable explicativa nominal para convertirla en una serie de variables numéricas de dos valores, la cual estará formada exactamente por una variable menor que la cantidad de niveles que tenía la nominal. (No es coincidencia que resulte el mismo número de los grados de libertad de la estimación intergrupal de varianza poblacional).

Esa capacidad para codificar una variable nominal independiente, y convertirla en una serie de variables numéricas de dos valores en el ANOVA, es una transición importante que hace posible la realización de un ANCORR múltiple.

Este procedimiento es extremadamente flexible y puede extenderse a los casos más complejos del análisis factorial de varianza. En verdad, lo importante no es que podamos realizar una codificación nominal; en la mayoría de los casos, una computadora lo hará por nosotros. Lo realmente relevante es comprender el principio que hace posible la conversión de un problema de ANOVA en un problema de REG. MÚLTIPLE.

Los supuestos y el GLM

En las diferentes técnicas basadas en el GLM, todos los procedimientos de prueba de hipótesis comparten los mismos supuestos. En el caso de la prueba t y el ANOVA, los principales se refieren a que todas las poblaciones representadas por los grupos tengan la misma varianza y sigan una distribución normal. Los supuestos de las pruebas de significación de correlación y de REG/CORR.MÚLT, son básicamente los mismos.

• ANDERSON D. SWEENEY D. y WILLIAMS T. (2001). Estadística. administración y economía. Vol 1 y 2: Thomson.

• BERENSON y LEVINE. (2000). Estadística Básica en admón. Prentice-Hall.

• CANAVOS, G. (2000). Estadística y probabilidades. Aplicaciones y métodos. McGraw- Hill.

• CHOU YA-LUN. (1984). Análisis estadístico: Interamericana. México.

• GARZO, F. y GARCÍA, F. (1993). Estadística. McGraw-Hill. España.

• GOVINDEN, L. (1991). Curso práctico de estadística. McGraw-Hill, Colombia.

• GUILFORD, S. y FRUCHTER, B. (1984). Estadística aplicada a la psicología y la educación: Graw-Hill, México.

• HABER, A. y RUNYON R. (1992). Estadística para las ciencias sociales. Addison Wesley U.S.A.

• KOROLIUK, V. (1986). Manual de la teoría de probabilidades y estadística matemática: Mir. Moscú.

• KREYSZIG, E. (1982). Introducción a la estadística matemática. Principios y Métodos: Limusa S.A. México.

• LARSON, H. (1993). Introducción a la teoría de probabilidades e Inferencia Estadística: Limusa. México.

• MENDENHALL W., BEAVER R. y BEAVER B. (2002). Introducción a la probabilidad y estadística:Thomson.

• MENDENHALL W. y SINCICH T. (1998). Probabilidad y estadística para ingeniería y ciencias: Prentice Hall.

• MEYER P. (1992) Probabilidad y aplicaciones estadísticas: Addison-Wesley.

• MILLER, FREUND y JHONSON. (1996). Probabilidad y Estadística: Prentice Hall.

• PADRON. E. (1996). Diseños experimentales: Trillas. México

• PAGANO M. y GAUVREAU K. (2001). Fundamentos de bioestadística: Thomson.

• SDHEFLER WILLIAM. (1981), Bioestadística: Fondo educativo interamericano.

• SIGEL, S. (1991). Estadística no paramétrica aplicada a las ciencias de la conducta: Trillas. México.

• STEEL, R. y TORRIE, J. (1988). Bioestadística: principios y procedimientos: McGraw-Hill.

• WALPOLE & MYERS. (1993). Probabilidad y estadística: Mc.Graw-Hill.

• WAYNE, W. (1982). Estadística con aplicaciones a las ciencias sociales y a la educación: McGraw-Hill. México.