ANÁLISIS BIVARIADO DE DATOS

Un resumen para el curso de Estadística II

UNIVERSIDAD CENTRAL DE VENEZUELA

February 20, 2012

Autor: Prof. Dimas Sulbarán

ÍNDICE

1. Análisis Bivariante

Si bien, el análisis univariado es una forma muy rica de construir conocimiento es en el análisis bivariado donde el investigador encuentra el sustrato para que el conocimiento alcance importantes niveles de comprensión, explicación y predicción de los fenómenos. El discurso cotidiano en el quehacer científico permite aceptar la siguiente afirmación: "las relaciones son la esencia del conocimiento" (Kerlinger & Lee, 2002, pág. 73). La ciencia evoluciona en la medida que se desarrollan modelos teóricos con base en la identificación de las relaciones entre variables y, por consecuente, entre constructos.

El análisis de datos bivariado es una forma evolucionada de análisis estadístico en el cual se cuantifica a nivel descriptivo e inferencial el nivel de covarianza entre dos variables y de esta forma se da cuenta de la relación entre dos variables. La cuantificación de la covarianza consiste en la construcción de coeficientes que permitan integrar en un valor estimado, información con respecto a la varianza conjunta entre dos variables y tiene como objetivo fundamental definir la magnitud y el sentido de la relación entre las variables. De este modo, el análisis conjunto de las varianzas de dos variables (regularmente definidas como X y Y) permite identificar la relación empírica entre éstas, entendiendo por relación el ajuste de los datos a una función lineal estocástica subyacente.

A partir de un referente teórico pertinente, el análisis bivariado busca someter a contrastación la tesis de asociación y hasta causalidad entre dos variables definidas. En cualquier caso, el análisis bivariado se plantea con la intención de determinar el nivel de relación entre dos variables y la función estocástica que subyace a un conjunto de observaciones (x,y). Pues si bien, la relación no es evidencia suficiente de causalidad no se puede hablar de causalidad en ausencia de relación entre las variables.

El análisis bivariante de datos involucra una familia de estadísticos cuya pertinencia está condicionada por el *nivel de medición* (Stevens, 1946) de las variables involucradas. Esta familia de estadísticos se divide en dos grandes grupos, a saber: *paramétricos* y *no paramétricos*. (Siegel & Castellan, 1995). Los paramétricos agrupan el caso de las variables con nivel de medición de intervalo o superior, distribución normal bivariada y $n > 30$. Los no paramétricos son el resto de las pruebas de correlación que no cumplen con los supuestos de las pruebas paramétricas; lo cual, les permite agrupar los estadísticos de contingencia y de correlación para variables con nivel de medición inferior a intervalo. En cualquier caso, el interés fundamental es construir un índice que permita determinar la magnitud y dirección de la relación entre las variables.

1.1. Usos de la correlación en la investigación psicológica

Si bien la descripción del comportamiento de las variables a nivel univariado resulta de gran valor para la ciencia en general; por tanto, para la psicología, la descripción de las relaciones entre variables es el insumo fundamental para la construcción y desarrollo de teorías. De esta manera, gran parte de la investigación psicológica se ha dirigido a determinar la relación entre variables estímulo y sus, teóricamente, correspondientes respuestas. Por su parte, el desarrollo de instrumentos psicométricos ha fundado de manera prominente gran parte de sus hallazgos sobre el uso de pruebas estadísticas de correlación. Elementos que han servido para la construcción de los índices de confiabilidad y validez empíricos.

1.2. El problema, lógica, de la asociación entre dos variables.

A nivel descriptivo una correlación supone un comportamiento más o menos afín entre dos variables y por tanto, suponen un conjunto de pares ordenados en los cuales los cambios en una variable figuran un reflejo en otra. (Kerlinger & Lee, 2002) Esta relación se define en un espacio bidimensional, donde cada punto o elemento muestral está determinado por su identificación con dos valores, entiéndase X_i y Y_i. en el caso de las matemáticas, una relación supone una función analítica perfectamente definida entre una variable X y una variable Y y bastan dos puntos para definir la pendiente de dicha relación, al menos, en el caso de una relación lineal. En estadística el asunto es un poco más complejo y la identificación de la relación entre dos variables estocásticas requerirá, en algunos casos más que en otros, un número mucho mayor de observaciones para reconocer la *función estocástica* y determinar así la magnitud y el sentido de la relación estadística entre las variables.

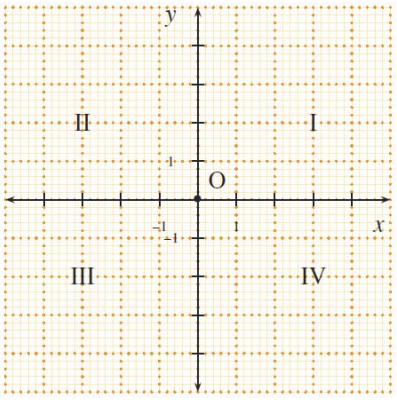
1.3. Supuestos teóricos considerados

El **primer** supuesto teórico involucrado en este proceso se refiere a la variabilidad propia de cada variable, es un sin sentido hablar de relación en ausencia de covarianza y la covarianza es inmanentemente varianza de X y de Y, en otras palabras si una variable deja de variar simplemente se convierte en una constante y si una variable es una constante cualquier diferencia

en los valores de la otra variable siempre estará asociado al mismo valor de esta primera variable. La traducción más simple, los cambios en una variable (X) no se asocian con cambios en una variable (Y) que permanece constante. El **segundo** supuesto se presenta a nivel inferencial, el contraste de hipótesis de correlación supone que la distribución de los datos bivariados se comporta de manera normal. (Pearson, 1895). Por tratarse de una distribución bivariada, la inferencia se basará en el uso de la función de distribución de los valores *t* para determinar la probabilidad de error al tomar la decisión de aceptar la hipótesis de nulidad de la relación.

1.4. Representación gráfica de una relación bivariada

La representación gráfica de las correlaciones obedece a los principios del análisis matemático de funciones. La representación de funciones parte de la noción clave del sistema de coordenadas cartesianas[1]. Este sistema se encuentra conformado fundamentalmente por dos ejes: el eje de las *abscisas* o eje de las *"x"* y el eje de las *ordenadas* o eje de las *"y"*; las cuales se cruzan de manera perpendicular, con un punto de intercepción que llamaremos *o* (origen de las coordenadas). En virtud de lo anterior y en atención a una determinada unidad de medida, se establecen con signo positivo las distancias en las semirrectas desde el origen hacia arriba y hacia la derecha, y con signo negativo desde el origen hacia abajo y hacia la izquierda. Con ello, todo el plano queda dividido en cuatro cuadrantes (I, II, III y IV), que se numeran en sentido contrario al movimiento de las agujas de un reloj. Veamos la ilustración:



[1] Este sistema de referencia se denomina sistema de ejes cartesianos o sistema cartesiano (de Cartesius, nombre latinalizado de René Descartes, filósofo y matemático francés del siglo XVII).

Una vez entendido el concepto de Plano Cartesiano, avanzamos hacia la representación gráfica de las correlaciones incorporando la noción de punto.

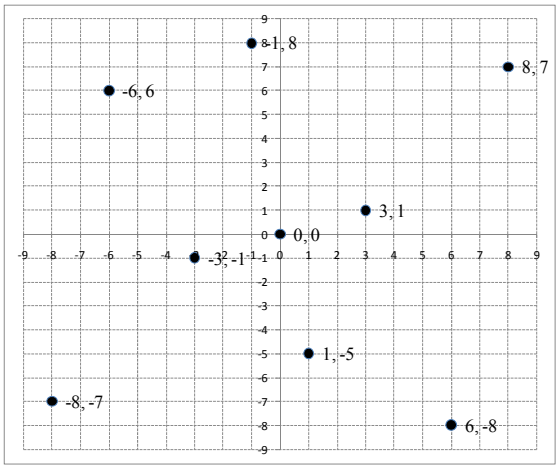
El punto posee una idea clave para nuestro asunto de las correlaciones, esta es: el punto es la mínima expresión de una línea. De acuerdo con lo estudiado, la línea que describe el comportamiento de un conjunto de puntos se define por una función. De modo que el punto puede ser definido como la mínima expresión de una "nube de puntos" y ya que toda función se define como una relación entre variables, entonces el punto define la mínima expresión de una relación entre variables, digamos: *x* y *y*.

Por cada punto (x,y) del plano pasan dos rectas perpendiculares entre sí y paralelas a cada uno de los ejes, es decir, pasa una recta paralela al eje de las *x* (abscisas) y una recta paralela al eje de las *y* (ordenadas). Estas rectas cortan los dos ejes en dos puntos. Si llamamos A y B, respectivamente, a cada uno de los puntos de corte estos definen las distancias OA y OB, por tanto, la abscisa y la ordenada del punto *P*. Por consiguiente, a cada pareja ordenada de puntos (x,y) le corresponde un punto del plano, y viceversa; a cada punto del plano le corresponde una pareja ordenada de puntos. El juego de "Batalla Naval" es un ejemplo de ello.

Veamos a continuación, de manera formal, algunos ejemplos adicionales:

Dada la siguiente tabla de datos:

X	Y	Cuadrante
3	1	I
8	7	I
-6	6	II
-1	8	II
1	-5	III
6	-8	III
-8	-7	IV
-3	-1	IV
0	0	O



Generamos el siguiente gráfico de puntos en el sistema de coordenadas

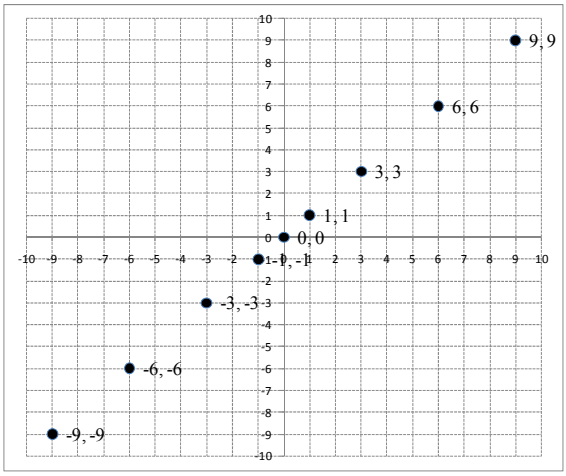
1.4.1. Gráfico de una función.

El caso anterior obedece estrictamente a una distribución de datos aleatorios. De este modo establecemos que la posición de un punto en el plano no está de ninguna manera relacionada con la aparición de los otros puntos. Dada esta situación, estamos sólo en presencia de un conjunto de puntos en el plano, más no de una función.

En el caso de una función, la presencia de un valor en y está determinada por los cambios en los valores de x bajo las condiciones de la ecuación analítica o estadística.

Veamos un ejemplo:

X	Y	Cuadrante
-9	-9	I
-6	-6	I
-3	-3	II
-1	-1	II
0	0	III
1	1	III
3	3	IV
6	6	IV
9	9	O



Hemos podido apreciar que, a diferencia del caso anterior, en este se evidencia una tendencia que puede ser perfectamente definida por una función muy sencilla y hablar así de una relación entre los valores de *x* e *y*. de modo que, *y= f(x)*; donde la *f(x)= x.*

1.5. Relaciones lineales y su cuantificación

Para abordar el tema de la cuantificación de las relaciones lineales es importante entender matemáticamente el concepto de correlación, entiéndase por correlación el cociente de la división de la covarianza entre la varianza total conjunta de X y Y, es decir, la correlación obedece fundamentalmente a la siguiente ecuación: $r_{xy} = \frac{s_{xy}}{s_x s_y}$. De esto se derivan los siguientes teoremas de la correlación: a) dada una cantidad de varianza conjunta o afín entre las variables (X, Y) igual a cero, la covarianza será igual a cero y la correlación cero; b) dada una cantidad de

varianza conjunta o afín entre las variables con un valor mayor que cero y menor a la varianza total, la covarianza será un valor mayor que cero y menor que la varianza total, por lo que la correlación estará definida en un rango entre cero y uno; c) dada una cantidad de varianza conjunta o afín entre las variables igual a la varianza total, lo cual ocurre cuando el comportamiento relativo representado por los desvíos de los puntajes en la variable X es idéntico al que reportan en la variable Y, la covarianza es igual a la varianza total y la correlación es igual a uno.

1.5.1. Coeficiente de correlación producto-momento de Pearson.

Existen diversos coeficientes que miden el grado de correlación entre dos variables. No obstante, estos han sido adaptaciones a los distintos niveles de medición de la fórmula fundamental de Karl Pearson. El coeficiente de Pearson (introducido en realidad por Francis Galton), sintetiza de forma magistral la idea fundamental de la correlación. La misma define una razón entre la cantidad de covarianza de dos variables y el producto de sus desviaciones estándar. Existen varias fórmulas para calcular el coeficiente de correlación de pearson (Glass y Stanley, 1984), como podemos ver:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{n.\sum XY - \sum X \sum Y}{\sqrt{[n.\sum X^2 - (\sum X)^2].[n.\sum Y^2 - (\sum Y)^2]}} = \frac{\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{S_x.S_y} = \frac{\sum \frac{(X_i - \bar{X})}{S_x}\frac{(Y_i - \bar{Y})}{S_y}}{n-1} = \frac{\sum Z_{x_i}.Z_{y_i}}{n-1}$$

Intentemos entender el coeficiente de correlación desde el análisis de la ecuación fundamental de pearson. El axioma fundamental es que el valor del índice de correlación varía en el intervalo [-1,1]. Así, la cuantificación de la relación se define por su magnitud y sentido. De esta manera se restringe el rango de valores a las siguientes condiciones:

La magnitud se restringe a los valores entre cero y uno [0 < r < 1], debido a que este valor hace referencia al módulo de la relación, entiéndase módulo como la distancia relativa promedio de los puntos con respecto al origen dependerá de la distancia relativa de cada punto con respecto a la media en X y Y. las medias tanto en los valores de X como en los valores de Y se convierten en el punto de origen para definir el módulo de un punto con respecto a X y Y. el máximo aporte relativo de cada punto a la construcción de la covarianza es cuando la distancia o magnitud en

términos de desviaciones típicas con relación a la media de X es idéntica a la distancia recorrida en puntuaciones tipificadas con respecto a la media de Y.

El sentido o signo de la relación (+,-) se define por la orientación que toma el punto con respecto a la media en X y Y. en la medida en que los puntos tienden a tomar el mismo sentido tanto para la distribución de los valores de X como para la distribución de los valores de Y, se contribuye a que el resultado de la covarianza presente un signo positivo. Por el contrario, si el sentido para la distancia recorrida por los puntos en el eje de las abscisas (x) es contrario al adoptado para el eje de las (y) se contribuye a la construcción de un resultado con signo negativo para la covarianza. Básicamente, el signo de la covarianza se rige por los principios algebraicos para el tratamiento de los signos.

El resultado definitivo para la covarianza que sentenciará el valor de la correlación se deriva del nivel de proporcionalidad y sentido de las distancias entre los puntajes individuales y la media de las distribuciones para los valores de X y Y, respectivamente. En resumen y he aquí una referencia para la interpretación de los resultados:

- Si $0 < r < 1$, existe evidencia de una correlación positiva o directamente proporcional. La cual es producto de un recurrente comportamiento de los valores relativos de X con relación a la media de X muy parecida a la distancia relativa y el sentido recorrido por los valores de Y con respecto a la media de Y. de esta manera cuando los valores en X se encuentran por encima de la media suelen estar asociados con valores en Y igualmente por encima de la media y viceversa.
- Si $r = 0$, no existe relación lineal. Pero esto no necesariamente implica que las variables sean independientes. Es necesario un análisis del gráfico de dispersión para descartar la existencia de relaciones no lineales entre las dos variables. Lo anterior es producto de un recurrente comportamiento de los valores relativos de X con relación a la media de X inconsistente con la distancia relativa recorrida por los valores de Y con respecto a la media de Y. de esta manera puede ocurrir que los valores en X por encima de la media estén asociados con valores en Y tanto por encima como por debajo de la media.
- Si $-1 < r < 0$, existe una correlación negativa o inversamente proporcional. La cual es producto de un recurrente comportamiento de los valores relativos de X con

relación a la media de X muy parecida a la distancia relativa recorrida por los valores de Y con respecto a la media de Y, pero en sentido contrario. De esta manera cuando los valores en X se encuentran por encima de la media suelen estar asociados con valores en Y por debajo de la media y viceversa.

1.5.2. Matriz de correlaciones.

Se define la matriz de correlación como una tabla de doble entrada para A B y C, que muestra una lista multivariable horizontalmente y la misma lista verticalmente con los correspondientes coeficientes de correlación llamado r para cada par de variables.

1.6. Combinación lineal de variables y correlación

Toda combinación lineal de variables se define como una función de x con respecto a Y. toda función describe la relación entre un conjunto de pares ordenados (x,y). Por su parte, la correlación permite cuantificar la magnitud y sentido de esta relación.

1.7. Otros coeficientes de correlación lineal

En muchos casos, los datos nos ofrecen variables que no cumplen con los requerimientos para el cálculo e interpretación del coeficiente de correlación de Pearson. Principalmente, el referente al nivel de medición de las variables involucradas, pues no se tienen observaciones de variables métricas. Los casos pueden incluir sólo una de las variables o ambas, pueden incluir variables con nivel de medición tan bajo como el nominal. Según el tipo de variables implicadas en la combinación, se cuentan: a) coeficiente phi, b) coeficiente de correlación tetracórica, c) coeficiente de asociación C de Cramer, d) coeficiente de correlación biserial y punto biserial, e) coeficiente eta, f) coeficiente de correlación tau de kendall y g) coeficiente de correlación de spearman. Los detalles se presentan a continuación.

Todos los coeficientes de correlación que se revisarán en lo sucesivo, con excepción del tau de Kendall, emplean básicamente, de una u otra forma, la teoría del producto-momento de Pearson (Glass & Stanley, 1974, pág. 176).

1.7.1. Coeficiente phi.

En estadística, el coeficiente phi φ o r_φ, también llamado coeficiente de correlación de *Mathews* es una medida de la asociación para variables binarias naturales. Esta medida es similar al coeficiente de correlación de Pearson en su interpretación. De hecho, un coeficiente de correlación de Pearson estimado para dos variables binarias nos dará el coeficiente phi (φ). El *coeficiente phi* también se relaciona con el estadístico chi-cuadrado para una tabla de contingencia de a 2×2. De forma que: $\varphi = \sqrt{\frac{\chi^2}{n}}$

Donde χ^2 corresponde al valor obtenido para chí cuadrado y *n* es el total del número de observaciones.

Un procedimiento alternativo atiende a la siguiente estructura para la tabla de datos:

	y=1	*y=0*	*total*
x=1	n_{11}	n_{10}	*n1·*
x=0	n_{01}	n_{00}	*n0·*
total	*n·1*	*n·0*	N

A partir de esta configuración de los datos en la tabla 2x2, se considera que dos variables binarias están positivamente asociadas si la mayor parte de los datos caen dentro de las celdas diagonales. Por el contrario, dos variables binarias se consideran negativamente asociadas si la mayoría de los datos se salen de la diagonal. La fórmula que cuantifica la relación es:

$$\varphi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_1.n_0.n_.{}_1n_.{}_0}}$$

Para determinar el nivel de significación estadística de estos resultados se debe establecer el valor de chí cuadrado con la formula estándar y contrastar con los valores críticos de la tabla de valores de chí, para un grado de libertad.

El cálculo del coeficiente Phí sugiere una correlación baja entre el colegio de procedencia y el ser o no admitido en la universidad, que favorece ligeramente el ingreso a la universidad de estudiantes provenientes de instituciones públicas y esta correlación es significativa ($Chí^2$= 4; gl=1; p<0.05), para un alfa de 0,05.

1.7.2. Coeficiente de correlación tetracórico

Como se ha reiterado al respecto, en ocasiones, el investigador podría estar interesado en determinar la relación entre dos variables que han sido dicotomizadas de su naturaleza continua. En este caso, debemos llevar a cabo el cálculo del coeficiente de correlación tetracórica r_t. Autores como (Glass & Stanley, 1974, pág. 166), refieren el complejo cálculo en este caso a una versión más sencilla como se presentará a continuación.

Dada la tabla de datos con forma:

	y=1	*y=0*	*total*
x=1	n_{11}	n_{10}	*n1·*
x=0	n_{01}	n_{00}	*n0·*
total	*n·1*	*n·0*	N

Se tiene una buena aproximación de los niveles de correlación entre las variables a partir de la ecuación: $r = \cos \frac{180^o}{1+\sqrt{\frac{n_{11}n_{00}}{n_{10}n_{01}}}}$

Por tratarse del coeficiente de un ángulo sus valores se limitan a un rango entre -1 y 1 y, dada la naturaleza métrica de los datos, se interpreta de forma análoga al coeficiente de correlación de Pearson.

1.7.3. Coeficiente de asociación c de crámer.

De acuerdo con Siegel (1990), el coeficiente C de Cramer es una medida del grado de asociación entre variables nominales. Con este estadístico se responde a la pregunta con relación al grado de asociación entre las variables que no era respondida por la aplicación de la prueba de independencia de Chí cuadrado, tal como ha sido estudiada hasta ahora. Tiene la bondad de que puede ser aplicada incluso para variables con un nivel de medición tan bajo como el nominal.

Una vez calculado el valor de chí cuadrado para la distribución de las contingencias, el cálculo del coeficiente Phí resulta muy sencillo. La ecuación a continuación ilustra los elementos necesarios.

$$C = \sqrt{\frac{X^2}{N(L-1)}}$$

C= Coeficiente de contingencia de Cramer.
X^2= Valor de chí cuadrado asociado con la distribución observada.
N= Tamaño de la muestra.
L= Número mínimo de filas o columnas.

1.7.4. Coeficiente de correlación Biserial y Punto Biserial

El coeficiente de correlación punto biserial Rpb permite estudiar la relación entre dos variables cuando una de ellas es una dicotomía natural y la otra tiene un nivel de medición superior a intervalo. En el caso de que la dicotomía no responda a una dicotomía natural, debemos sospechar del uso del coeficiente de correlación punto biserial y preferir el uso de la correlación biserial. El cálculo de rpb atiende a la fórmula siguiente:

$$\mathbf{r_{pb}} = \frac{\overline{\mathbf{X_1}} - \overline{\mathbf{X_0}}}{\mathbf{s_x}} \sqrt{\frac{\mathbf{n_1 n_0}}{\mathbf{n(n-1)}}}$$

Por su parte, el coeficiente de correlación biserial aplica en los casos en los que una de las variables obedece a una escala nominal dicotómica, que se distribuye de manera normal y la otra es una variable de intervalo que también se distribuye de manera normal. Su cálculo se define por la siguiente ecuación:

$$\mathbf{rb} = \frac{\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{0}}{\mathbf{s_x}} * \frac{\mathbf{n_1 n_0}}{\mathbf{un}\sqrt{(\mathbf{n^2 - n})}}$$

Donde: todos los elementos obedecen a la nomenclatura manejada para el caso de la correlación punto biserial y u es la ordenada (la altura) de la unidad de la distribución normal, en el punto a partir del cual se halla el porcentaje $100(n_1/n)$ del área bajo la curva normal.

1.7.5. Coeficiente eta

Como se mencionara, el método para la cuantificación de la relación entre dos variables que no corresponden con una función lineal sigue el procedimiento de los mínimos cuadrados. Hemos visto que el coeficiente de determinación de pearson para la relación entre dos variables

está dada por la ecuación: $r_{xy}^2 = 1 - \frac{s_e^2}{s_y^2} = \frac{s_y^2 - s_e^2}{s_y^2} = \frac{s_{\hat{y}}^2}{s_y^2}$. De forma análoga se puede determinar el nivel de varianza en *y* que es explicada por los cambios en *x*, entiéndase: *razón de correlación* o *coeficiente eta cuadrado*, si se determinan los elementos necesarios considerando algunos ajustes en su cálculo. Algunos de los ajustes que se tomarán en cuenta es que en el caso de las varianzas de error o intragrupo, se debe determinar una media de los valores de *y* por cada valor de *x* en la función.

Es importante señalar que $\eta_{x,y}^2$ y $\eta_{y,x}^2$ generalmente serán diferentes, lo cual es contrario a nuestra experiencia con r caso en el cual $r_{x,y}^2 = r_{y,x}^2$. Como sucede con r_{xy}^2, $\eta_{x,y}^2$ siempre debe ser un valor entre 0 y 1. Además, $\eta_{x,y}^2 \geq r_{x,y}^2$. La diferencia entre los dos coeficientes es la medida del grado de curvilinearidad de la línea que mejor se ajusta a la distribución de los puntos x,y.

1.7.6. Coeficiente de correlación tau de kendall.

Dado el conjunto de pares observados (x_1, y_1), (x_2, y_2), …, (x_n, y_n) para un par de variables aleatorias respectivamente, tal que todos los valores de x_i y y_i sean únicos. Se dice que un par de observaciones (x_i, y_i) y (x_j, y_j) se dice que concuerdan si el rango de ambos valores coinciden: esto es, si $x_i > x_j$ y $y_i > y_j$ o si ambos $x_i < x_j$ y $y_i < y_j$. Se dice que son discordantes, si $x_i > x_j$ y $y_i < y_j$ o si $x_i < x_j$ y $y_i > y_j$. Si el caso es $x_i = x_j$ o $y_i = y_j$, no existe concordancia ni discordancia.

De modo que el tau de kendall se define por la ecuación: $\tau = \frac{2S}{N(N-1)}$; *donde:*

$S = \sum C - \sum D$
$\sum C = suma\ de\ los\ pares\ concordantes$
$\sum D = suma\ de\ los\ pares\ discordantes$
N= número total de pares observados.

Propiedades:

El denominador es el total del número de pares de combinaciones, de modo que el coeficiente se encuentra en el rango entre -1 y 1. Como resultado:

- Si el acuerdo entre los dos rangos es perfecto, el coeficiente tiene un valor de 1.
- Si el desacuerdo entre los dos rangos es perfecto, el coeficiente tiene un valor de -1.

- Si X y Y son independientes el coeficiente τ esperado es igual a cero.

1.7.7. Coeficiente de correlación de spearman.

El coeficiente de correlación Rho de rangos ordenados de Spearman permite estudiar la relación entre dos variables que tienen un nivel de medición superior a ordinal. Aplica incluso a variables de un nivel de medición de intervalo en el caso en que estas variables no se distribuyan de manera normal. En el caso de que el número de empates no represente una cantidad importante, el cálculo de r_{ho} atiende a la fórmula siguiente:

$$r_{ho} = 1 - \frac{6 \sum d^2}{N^3 - N}$$

De acuerdo con Siegel (1995), si los sujetos constituyen una muestra aleatoria de alguna población, se puede probar si el valor observado de r_s indica una asociación entre las variables X y Y en la población. En función del tamaño de la muestra las técnicas son:

- Para N desde 4 hasta 50, los valores críticos de r_s entre los niveles de significación unidireccionales 0.25 y 0.0005, están proporcionados en la tabla de valores críticos de rs. Para una prueba bidireccional sólo se debe duplicar el nivel de significación observado.
- Para N mayor a 50 se debe hacer el ajuste a la normal como referencia.

2. Regresión Lineal

Básicamente, hablamos de regresión lineal o ajuste lineal al método estadístico matemático que permite definir una ecuación para la relación entre una variable dependiente *Y* y la(s) variable(s) independiente(s) X_i, más o menos un término aleatorio *ε*. Esta estrategia permite la construcción de modelos teóricos en un lenguaje analítico que se reduce a la forma:

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_2 \ldots + \beta_n x_n \pm \varepsilon$$

Veamos que nos dice la función de regresión lineal en términos teóricos-metodológicos. Esta función identifica los cambios o variaciones en una variable (*y*), cuya varianza total se identifica teóricamente con la variable dependiente; la cual se asume como el resultado de los cambios en una serie de variables x1, x2, … xn, que se asocian con las variables independientes o metodológicamente explicativas de la varianza sistemática. Recordemos que la varianza sistemática puede ser de dos tipos: a) aquella producto de las variables independientes en estudio y b) aquella producto de las variables independientes ajenas al estudio o variables extrañas. En el primer caso, la ecuación permite identificar por medio de los coeficientes de determinación el peso específico de cada una de las variables predictoras en estudio sobre las variaciones o cambios de la VD o predicha; mientras que en el segundo, el peso de las variables independientes no identificadas o extrañas del modelo se sopesa en el intercepto de la ecuación. (Gujarati & Porter, 2010, pág. 189). Finalmente, el valor más importante de la ecuación de regresión lineal es que permite cuantificar la varianza de error (ε) o componente estocástico en la función de estimación. De modo que, la ecuación de regresión se define metodológicamente como:

$$y = \sigma^2_{total}$$

$$\left.\begin{matrix} \beta_0 = variables\ extrañas\ no\ controladas \\ \beta_1 x_i + \beta_2 x_2 \ldots + \beta_n x_n = variables\ independientes \end{matrix}\right\} = \sigma^2_{sistemática}$$

$$\varepsilon = \sigma^2_{error}$$

Tanto en el caso de dos variables (regresión simple) como en el de más de dos variables (regresión múltiple), el análisis de regresión lineal puede utilizarse para explorar y cuantificar la relación entre una variable dependiente o criterio (*Y*) y una o más variables independientes o predictoras (x_1, x_2, …, x_n), así como para desarrollar una ecuación lineal con fines predictivos. De esta manera, el análisis de regresión inspira el principio de Max-Min-Con para la

construcción de diseños y lleva asociados una serie de procedimientos de diagnóstico (análisis de los residuos) que informan sobre la idoneidad del modelo. (Pardo & Ruíz, 2005, pág. 455).

2.1. Importancia del término de error

La diferencia fundamental entre las funciones estadísticas y las funciones matemáticas es el hecho de que las primeras atienden específicamente al estudio y control de la varianza de error. Esto es lo que permite distinguir las funciones deterministas de las funciones estocásticas. Por excelencia, la estadística es la disciplina cuyo interés fundamental es el estudio de las funciones estocásticas o de aquellas signadas tanto por elementos sistemáticos como por elementos aleatorios. Técnicamente al error de una función se le conoce como perturbación estocástica o término de error estocástico (Gujarati & Porter, 2010, pág. 40) y define la diferencia entre el valor estimado en Y y el valor observado, es decir: $\varepsilon = Y_i - E(Y|X_i)$, de este modo $Y_i = E(Y|X_i) + \varepsilon$ (Gujarati & Porter, 2010).

Hemos mencionado que la varianza de error o perturbación estocástica se define como la expresión de la influencia de las variables desconocidas y que no han sido controladas teórica ni metodológicamente. El compromiso natural del investigador es la minimización de la varianza no sistemática o de error que se traduce en desconocimiento y pérdida de precisión en la estimación. (Kerlinger & Lee, 2002). Entre los factores que se cuentan por su influencia en la construcción de la varianza de error están: a) vaguedad de la teoría, b) falta de disponibilidad de datos, c) presencia de variables periféricas, d) aleatoriedad intrínseca del comportamiento humano, e) el error de medición, f) la parsimonia del modelo, y g) errónea especificación del modelo.

2.2. Introducción a la regresión lineal en el contexto de la Psicología

Si bien, la investigación en general ha encontrado un importante apoyo en las técnicas estadísticas de análisis de regresión, en disciplinas como: Psicología, Sociología, Economía y Educación, dedicadas al complejo estudio del comportamiento humano las técnicas de regresión han sido un recurso de gran valor en la identificación y determinación de modelos teóricos que

han permitido acercarse a los ideales del conocimiento científico, a saber: la descripción, explicación, predicción y control de los fenómenos del comportamiento humano.

Comenzando con el hecho mismo de la descripción, debemos tener presente la importancia de las técnicas de regresión en la construcción de rigurosos instrumentos de medición psicométricos, bajo el esquema representacionista. Es de conocimiento general que las principales estrategias de evaluación de la validez y la confiabilidad de los instrumentos apuntan hacia la determinación de modelos de regresión que permiten definir la función de categorización que caracteriza el proceso de "atribución de valores numéricos a los objetos o eventos de acuerdo con ciertas reglas" (Stevens, 1946).

Más allá, la ciencia y, particularmente, la Psicología avanzan en la medida en que es capaz de generar teorías que no sólo permiten la descripción de los fenómenos sino que está en posibilidad de ofrecer una explicación de los fenómenos del comportamiento. En este sentido, los aportes de las técnicas de regresión estadística para identificar y determinar los modelos que describen la relación entre las variables en la forma de modelos funcionales que permiten estimar claramente la influencia de determinadas situaciones, procesos o hechos sobre los fenómenos del comportamiento.

Determinado el modelo explicativo para algún aspecto del comportamiento humano, el investigador seguirá apelando al uso de las técnicas de regresión estadística con el interés de avanzar a la predicción de la conducta. En este sentido, podríamos decir que son infinitos los casos en los cuales los modelos de regresión contribuyen a la predicción del comportamiento en los distintos campos y aplicaciones de la psicología, tales como: la clínica, el asesoramiento, el comportamiento social y organizacional y el educativo.

Finalmente, las técnicas de regresión estadística son un recurso de gran valor cuando se trata de generar conocimiento en la forma de control. La identificación de las relaciones funcionales entre las distintas variables estudiadas contribuyen a conocer cuáles y en qué medida las diferentes variables determinan las variaciones de la conducta. Este recurso le permite al investigador manipular las condiciones a fin de procurar los valores de la variable predicha en el sentido de su conveniencia.

2.3. Supuestos del modelo de regresión lineal

Los supuestos de un modelo estadístico se refieren a una serie de condiciones que deben cumplirse para garantizar la validez del modelo en cuestión. De forma que, al efectuar aplicaciones prácticas del modelo de regresión, nos veremos en la necesidad de examinar el ajuste de nuestras observaciones con estos supuestos. Entiéndase: a) linealidad, b) independencia de los errores, c) homocedasticidad, d) normalidad, e) no colinealidad y f) el número de observaciones *n* debe ser mayor que el número de parámetros por estimar. (Gujarati & Porter, 2010). De acuerdo con el teorema de Gauss-Markov, el cumplimiento de estos supuestos, son el prerrequisito para alcanzar los estimadores MELI (mejores estimadores linealmente insesgados). Los detalles con relación al concepto y las implicaciones de cada supuesto se exponen a continuación.

2.3.1. Linealidad

Por linealidad se entiende a las siguientes formas que pueden definir a una ecuación: a) linealidad en las variables y/o b) linealidad en los parámetros. Como estos términos no son independientes, las variables pueden cumpir con cualquiera de las combinaciones posibles entre la presencia y ausencia de linealidad tanto en las variables como en los parámetros, tal como se resume en el cuadro siguiente:

		Linealidad en las variables	
		Si	No
Linealidad en los parámetros	Sí	$y = \beta_0 + \beta_1 X_1 + \varepsilon$	$y = \beta_0 + \beta_1 X_1^2 + \varepsilon$
	No	$y = \beta_0 + \beta_1^2 X_1 + \varepsilon$	$y = \beta_0 + \beta_1^2 X_1^2 + \varepsilon$

De las dos interpretaciones de linealidad, se exige linealidad en los parámetros para el desarrollo de los modelos de regresión lineal que se desarrollan en este curso. Por consiguiente, cuando se hace referencia al término *regresión lineal* nos estamos refiriendo a una regresión lineal en los parámetros. Lo cual significa que los parámetros $(\beta_0 + \beta_1)$ se elevan sólo a la primera potencia. Aunque puede o no ser lineal en las variables explicativas p predictoras.

2.3.2. Independencia de los errores.

Este supuesto se resume al principio de que el error (ε) es una variable aleatoria para todo valor de x_i, por tanto la correlación entre estas dos variables debe ser igual a cero. En otras palabras, se espera que $r_{xe}=0$.

2.3.3. Homocedasticidad.

Se espera que la varianza del término de error para todos los valores de x_i sean equivalentes. Esta cualidad es necesaria, según el Teorema clásico de Gauss-Márkov, para que en un modelo los coeficientes estimados sean los mejores o eficientes, lineales e insesgados.

Cuando no se cumple esta situación, se dice que existe heterocedasticidad, que es cuando la varianza de cada término de perturbación no es un número constante

2.3.4. Normalidad.

Dado que se asume que las perturbaciones del modelo o término de error para cada valor de x_i obedecen a una distribución aleatoria, se espera que esta se distribuya de manera normal. En otras palabras, los valores de error tienden a distribuirse de la forma: $\varepsilon \sim N(\mu, \sigma)$, dado que $\mathrm{E}(v_i \mid x_i) = 0$, entonces la distribución de los errores debe cumplir con $\varepsilon \sim N(0, \sigma)$. El cumplimiento de este supuesto es fundamental para los análisis con fines inferenciales.

2.3.5. No colinealidad.

Este supuesto está relacionado, fundamentalmente, con los casos de los análisis de regresión múltiple. Atendiendo al principio fundamental de parsimonia, se espera que la incorporación de cada variable independiente o predictora al modelo se dé porque representa un valor agregado significativo en la definición del mismo para la explicación del fenómeno de interés. Para cumplir con este propósito se debe procurar que cada variable independiente sea efectivamente algo "independiente" a las otras variables explicativas que se agrega a la ecuación. En este caso, estaremos en presencia de No colinealidad. Cuando no se cumple con este supuesto se habla de la presencia de colinealidad de que las variables incluidas en el modelo están correlacionadas entre sí y resultan en un modelo redundante.

2.3.6. El número de observaciones n debe ser mayor que el número de parámetros por estimar

Se espera que el número de observaciones n sea mayor que el número de variables explicativas.

2.4. Hipótesis del modelo de regresión lineal

Como elemento clave de la construcción e modelos a partir del análisis de regresión está la estimación puntual y por intervalos, debemos considerar en este punto la introducción de las pruebas de hipótesis en el caso del estudio de los parámetros de los MRL.

En términos generales, el lenguaje estadístico ha asumido los términos hipótesis nula e hipótesis alternativa, derivadas del matrimonio forzado entre la tradición de los trabajos de Ronald Fisher y Neyman-Pearson. La hipótesis nula hace referencia, fundamentalmente, a la ausencia de relación entre las variables. Por su parte, la hipótesis alternativa apoya la tésis de relación entre las variables. Ésta última puede definirse en términos tanto simples como compuestos, en otras palabras, puede plantear el caso de un contraste puntual del tipo $H_1: \beta_2 = 1.5$ (simples) o del tipo $H_1: \beta_2 \neq 1.5$ (compuestas).

La teoría de la prueba de hipótesis se refiere al diseño de reglas o procedimientos que permitan decidir si se rechaza o no la hipótesis nula (Ho). Hay dos métodos mutuamente complementarios para diseñar tales reglas: el intervalo de confianza y las pruebas de significación. (Gujarati & Porter, 2010, pág. 113). Ambos enfoques plantean que la variable (el estadístico o estimador) en consideración sigue alguna distribución de probabilidad y que la prueba de hipótesis establece afirmaciones sobre el(los) valor(es) del(los) parámetro(s) de la distribución. La mayoría de las hipótesis estadísticas se fundan en este paradigma, conocido como el uso de técnicas paramétricas para el contraste de hipótesis.

Al igual que en el resto de las técnicas de contraste de hipótesis estadísticas, la construcción de hipótesis obedecerá a dos ámbitos discursivos, relativamente independientes, unidos por la razón del investigador. Estos dos campos son: el campo teórico o sustantivo de las

variables en estudio y el campo estadístico en el cual la relación teórica entre las variables se traduce en una función analítica con parámetros estadísticos.

2.5. Tipos de regresión lineal

Las regresiones lineales se clasifican fundamentalmente por el número de variables independientes involucradas, de este modo se dispone de dos clases de regresiones lineales: a) regresiones lineales simples, para aquellos modelos que definen la relación entre una variable predictora y una variable predicha y b) regresiones lineales múltiples, para aquellos casos en los cuales se incluyen más de una variable independiente o predictora para la estimación de los valores de una determinada variable dependiente o predicha.

2.5.1. Regresión lineal simple.

Este apartado atenderá al punto con relación al análisis de regresión en su versión más simple, es decir la regresión bivariada o función de regresión para dos variables (x,y), en la cual la variable dependiente, predicha o regresada se relaciona con una sola variable independiente, predictora, explicativa o regresora. Este caso es particularmente interesante porque permite desarrollar los fundamentos del análisis de regresión, pues el análisis de regresión múltiple sólo es una extensión lógica del análisis de regresión simple.

Autores como (Gujarati & Porter, 2010), apoyan la tesis de que los principios del análisis de regresión simple se definen en su propósito; es decir, la construcción de un estimado para la *función de regresión poblacional*. Función que denota el *valor esperado* de la distribución de Y dado un valor de X_i, es decir $E(Y|X_i)$. A estos valores medios se les llama *valores esperados condicionales*, en virtud de que sus variaciones dependen de las variaciones en la variable condicional X. de este modo se espera que la función o ecuación de regresión poblacional se defina como: $y = E(Y|X_i) = \beta_0 + \beta_1 X_i \pm \varepsilon$ la cual, típicamente se define a partir de su estimación con base en la función de regresión muestral o: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 \pm \hat{\varepsilon}$

2.5.1.1. Construcción de los parámetros de regresión lineal por el método de mínimos cuadrados ordinarios (MCO).

El método de *mínimos cuadrados ordinarios* (MCO) es uno de los mayores aportes del astrónomo y matemático Carl Friedrich Gauss, quien a finales del siglo XVIII y principios del siglo XIX desarrolló la teoría de los mínimos cuadrados para el estudio de los cuerpos celestes. A partir de los supuestos antes expuestos en concordancia con el clásico teorema de Gauss-Markov, el método en cuestión ofrece importantes propiedades estadísticas que lo han convertido en uno de los más eficaces y populares del análisis de regresión.

La aproximación por mínimos cuadrados se basa en la minimización del error cuadrático medio o, equivalentemente, en la minimización del radicando de dicho error, el llamado error cuadrático, definido como:

$$E_c(f) = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n}}$$

Para alcanzar este objetivo, se utiliza el hecho que la función *f* debe poder describirse como una *combinación lineal de una base de funciones*. Los coeficientes de la combinación lineal serán los parámetros que queremos determinar.

Veremos en este punto como se definen los estimadores para los parámetros de la ecuación de regresión poblacional, desde un discurso práctico. Los detalles con relación a la derivación analítica y cálculo de los factores de regresión lineal se presentan en el apartado con relación a la identificación del modelo y derivación de ecuaciones normales.

Dado que la ecuación de regresión lineal simple o bivariada se define como $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 \pm \hat{\varepsilon}$, por lo que se encuentra constituida por los siguientes elementos:

$$\hat{y} = valores\ estimados\ de\ Y\ para\ la\ combinación\ lineal\ dada.$$

$$\hat{\beta}_0 = intercepto\ de\ la\ ecuación.$$

$$\hat{\beta}_1 = coeficiente\ de\ determinación\ para\ la\ función\ de\ X_1.$$

$$X_1 = variable\ independiente\ o\ predictora.$$

$$\hat{\varepsilon} = término\ de\ error\ o\ perturbación\ estocástica.$$

Comencemos por $\hat{\beta}_0$, dada la función lineal con media de X y media de Y, se define la ecuación para el intercepto como: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_0\bar{X} \pm \hat{\varepsilon}$

Sin embargo, dado que $\hat{\beta}_0$ es una función lineal de X, se requiere determinar el coeficiente de determinación o pendiente $\hat{\beta}_1$ de la ecuación para poder llevar a cabo el cálculo de $\hat{\beta}_0$. De modo que, definimos $\hat{\beta}_1$ a partir de la ecuación: $\hat{\beta}_1 = \frac{S_{x,y}}{S_x^2}$, donde:

$$S_{xy} = covarianza\ de\ X\ e\ Y.$$

$$S_x^2 = varianza\ de\ X.$$

Finalmente, calculamos el término de error $\hat{\varepsilon}$ como: $\sigma_{\hat{\varepsilon}}{}^2 = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2} = S_y\sqrt{1 - r_{xy}^2}$, donde:

$$r_{xy}^2 = coeficiente\ de\ determinación\ para\ la\ relacion\ entre\ X\ e\ Y.$$

2.5.2. Regresión lineal múltiple.

Los análisis de regresión bivariados estudiados en el punto anterior, suelen ser útiles para comprender los principios del análisis de regresión, más suelen ser altamente deficientes en la práctica. La dinámica de la investigación cotidiana raramente encontrará cubiertas sus necesidades con el uso de análisis de regresión bivariados, es decir, cuando una variable predicha es explicada únicamente por la influencia de una variable independiente, más allá del error típico. Lo común es que la construcción de modelos que se acerquen de forma más fidedigna a la realidad esté dado por la inclusión de más de dos variables en la explicación de los cambios en la variable criterio. En estos casos, en los cuales el modelo está conformado por una variable dependiente a la que llamaremos Y y, por lo menos, dos variables independientes x1, x2, … xn, más un término de error estamos en presencia de un modelo de regresión múltiple.

Por extensión del modelo de regresión lineal bivariado, se define la ecuación de regresión múltiple como:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_n X_n + \varepsilon$$

Al igual que en el caso de los modelos bivariados, la ecuación está conformada por una variable dependiente, criterio o predicha (Y), el coeficiente de determinación para el intercepto (β_0), las variables independientes, regresoras o predictoras (X_1, X_2,… X_n) con sus respectivos *coeficientes de regresión parcial*, $\beta_1, \beta_2 \ldots \beta_n$ y el término de error o perturbación estocástica (ε).

A partir de los supuestos del modelo de regresión clásico, se cumple que, al tomar la esperanza condicional de Y en ambos lados de la ecuación, el resultado es:

$$E(Y_I|X_1, X_2, \dots X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_n X_N + \varepsilon$$

Como en el caso de los modelos con dos variables, el análisis de regresión múltiple es el análisis de la regresión condicional sobre los valores fijos de las variables explicativas, y lo que obtenemos es el valor promedio o la media de Y, a los valores dados a las regresoras X.

Al igual que en el caso de las ecuaciones de regresión bivariados, mostraremos en este punto como se definen los estimadores para los parámetros de la ecuación de regresión poblacional múltiple, desde un discurso práctico. Los detalles con relación a la derivación analítica y cálculo de los factores de regresión lineal se presentan como una extensión natural del apartado con relación a la identificación del modelo y derivación de ecuaciones normales.

Por consecuencia, como una extensión natural del cálculo del intercepto, la formula se convierte en la ecuación: $\bar{Y} = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \cdots + \beta_k \bar{X}_k$, despejando β_0 el resultado es la conocida ecuación:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 - \cdots - \beta_k \bar{X}_k$$

Operaciones algebraicas de factorización de términos pertinentes nos permite derivar las siguientes fórmulas para el cálculo de los coeficientes $\beta_1, \beta_2, \dots, \beta_k$ de nuestras ecuaciones normales:

$$\hat{\beta}_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2}$$

$$\hat{\beta}_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2}$$

Finalmente, el error estándar de la ecuación se calcula como: $\sigma_{\hat{\varepsilon}}{}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-k}$, donde k representa el número de variables independientes (x_1, x_2, ..., x_n) incluidas en el modelo. Los grados de libertad se definen como n-k porque para calcular el ε se requiere la estimación de los valores de $\beta_1, \beta_2 \dots \beta_n$.

2.6. Identificación del modelo: ecuaciones normales

Este punto corresponde a la deducción analítica de los estimadores lineales de mínimos cuadrados (β_0, β_1, β_2,…, β_k). No obstante, sería imposible entender el desarrollo de las ecuaciones normales sin revisar el concepto de ***derivadas parciales***, cuyo fundamento teórico se deriva del *cálculo vectorial* y se define como: una función de diversas variables es su derivada respecto a una de esas variables manteniendo las otras como constantes. Un concepto que en estadística hemos aprendido a asociar con el de coeficientes de determinación parcial. La derivada parcial de una función *f* respecto a la variable *X* se representa con cualquiera de las siguientes notaciones equivalentes:

$$\frac{\partial f}{\partial x} = \partial x f = f'x$$

En estos términos, cuando una magnitud *A* es función de diversas variables (*x,y,z*,...), es decir: $A = f(x, y, z, \dots)$, al realizar la derivada obtenemos la expresión que nos permite obtener la pendiente de la recta tangente a dicha función *A* en un punto dado. Esta recta es paralela al plano formado por el eje de la incógnita respecto a la cual se ha hecho la derivada y el eje z.

Sea $\{f_j(x)\}^m j = 1$ un conjunto de *m* funciones linealmente independientes (en un espacio vectorial de funciones), que se llamarán *funciones base*. Se desea encontrar una función *f(x)* de dicho espacio, o sea, combinación lineal de las funciones base, tomando por ello la forma:

$$f_{(x)} = c_1 f_1(x) + c_2 f_2(x) + \cdots + c_m f_m(x) = \sum_{i=1}^{m} c_j f_j(x)$$

Así, los c_j que minimizan el *error cuadrático medio* podrán ser calculados a partir del uso de derivadas parciales e igualando a cero este último, esto es:

$\frac{\partial f}{\partial x} = \frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \widehat{\beta}_i} = \sum_{i=1}^{n} 2\left(Y_i - \sum_{i=1}^{m} c_j f_j(x_i)\right)\left(-f_i(x_i)\right) = 0$, siendo i=1, 2,…,m

En términos generales, la *derivación parcial* permite construir un sistema de *m* ecuaciones con *m* incógnitas, que recibe el nombre de "Ecuaciones Normales de Gauss". Por ejemplo, en el caso de la regresión simple m=2.

2.6.1. Identificación del modelo: ecuaciones normales para regresiones bivariadas.

Se ilustrará detalladamente el caso para la ecuación de regresión lineal simple como modelo fundamental. Entendido que la práctica de mínimos cuadrados ordinarios tiene como objetivo minimizar la suma de los cuadrados de los residuos $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$, donde $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = f(\beta_0, \beta_1)$. De modo que:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1\right)^2$$

La deducción analítica de los estimadores lineales de mínimos cuadrados equivale a hallar los *k* coeficientes de regresión (β_j) del modelo, en nuestro caso particular dos: a) intercepto (β_0) y b) pendiente (β_1) de X_1. Para ello debemos derivar parcialmente con respecto a β_0 y β_1[2] e igualar estas ecuaciones a cero. De modo que derivamos para los casos de β_0 y β_1:

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_0} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)(-1) = 2 \sum_{i=1}^{n} \hat{\varepsilon}$$

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)(-X_i) = 2 \sum_{i=1}^{n} \hat{\varepsilon}\, X_i$$

E igualamos a cero las ecuaciones anteriores:

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_0} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)(-1) = 0$$

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)(-X_i) = 0$$

Al operar sobre la ecuación anterior se obtiene que:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

[2] Entiéndase una derivada parcial de una función de diversas variables X1, X2, … Xn, como la identificación de la pendiente respecto a una o cada una de esas variables, manteniendo las otras como constantes. En el discurso estadística se traduce en el cálculo de los coeficientes de determinación parcial de orden k, según el número de variables a controlar.

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$$

Después de complejas transformaciones algebraicas y simplificaciones, finalmente se obtienen los estimadores para $\beta_0 \, y \, \beta_1$, tal como se observa en las siguientes formulas, comenzando por el despeje de β_0:

Nuestra primera tarea consiste en dividir la primera ecuación normal entre *n*, de la siguiente forma:

$$\frac{\sum_{i=1}^{n} Y_i}{n} = \frac{n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X}{n}$$

Por consecuencia, la formula se convierte en la ecuación: $\bar{Y} = \beta_0 + \beta_1 \bar{X}$, despejando β_0 el resultado es la conocida ecuación: $\beta_0 = \bar{Y} - \beta_1 \bar{X}$

Por su parte, la identificación de β_1 en la segunda ecuación normal comienza con la sustitución de β_0. De este modo, la ecuación se transforma en:

$$\sum_{i=1}^{n} X_i Y_i = \left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$$

A partir de la factorización de la ecuación, se consigue que esta se traduzca en:

$$\sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i = \hat{\beta}_1 \left[\sum_{i=1}^{n} X_i^2 + \bar{X} \sum_{i=1}^{n} X_i \right]$$

Dado que:

$$\sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})$$

Y:

$$\hat{\beta}_1 \left[\sum_{i=1}^{n} X_i^2 + \bar{X} \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Entonces:

$$\hat{\beta}_1 \frac{\sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 + \bar{X}\sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Al dividir ambos términos entre *n*, el resultado es:

$$\hat{\beta}_1 = \frac{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{n}}{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}} = \frac{S_{x,y}}{S_x^2}$$

2.6.2. Identificación del modelo: ecuaciones normales para regresiones multivariadas.

Para el caso de las funciones de regresión multivariados (con más de dos variables independientes), al diferenciar parcialmente por cada una de las incógnitas de la ecuación:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \cdots - \hat{\beta}_j X_j\right)^2$$

se obtiene que:

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_0} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \ldots - \hat{\beta}_k X_{ki}\right)(-1) = 0$$

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \ldots - \hat{\beta}_k X_{ki}\right)(-X_{1i}) = 0$$

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_2} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \ldots - \hat{\beta}_k X_{ki}\right)(-X_{2i}) = 0$$

(…)

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_k} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}\right)(-X_{ki}) = 0$$

Operando según el esquema general, se obtienen las siguientes *ecuaciones normales* de Gauss:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} + \cdots + +\hat{\beta}_k \sum_{i=1}^{n} X_{ki}$$

$$\sum_{i=1}^{n} X_{1i}Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}\,X_{3i} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{2i}\,X_{ki}$$

$$\sum_{i=1}^{n} X_{2i}Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_{2i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}X_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{2i}\,X_{ki}$$

(…)

$$\sum_{i=1}^{n} X_{ki}Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_{ki} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}X_{ki} + \hat{\beta}_2 \sum_{i=1}^{n} X_{3i}X_{ki} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}^2$$

Nuestra primera tarea vuelve a ser la identificación del intercepto o $\hat{\beta}_0$, al igual que con el caso de los modelos bivariados se reduce a dividir entre n ambos lados de la ecuación:

$$\frac{\sum_{i=1}^{n} Y_i}{n} = \frac{n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} \ldots - \hat{\beta}_k X_{ki}}{n}$$

Por consecuencia, como una extensión natural del cálculo del intercepto visto, la formula se convierte en la ecuación: $\bar{Y} = \beta_0 + \beta_1\bar{X}_1 + \beta_2\bar{X}_2 + \cdots + \beta_k\bar{X}_k$, despejando β_0 el resultado es la conocida ecuación:

$$\beta_0 = \bar{Y} - \beta_1\bar{X}_1 - \beta_2\bar{X}_2 - \cdots - \beta_k\bar{X}_k$$

Operaciones algebraicas pertinentes de factorización de términos nos permite derivar las siguientes fórmulas para el cálculo de los coeficientes $\beta_1, \beta_2, \ldots, \beta_k$ de nuestras ecuaciones normales:

$$\hat{\beta}_1 = \frac{(\sum y_i x_{1i})(x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i}x_{2i})}{(x_{1i}^2)(x_{2i}^2) - (\sum x_{1i}x_{2i})^2} = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(x_{1i}^2)(x_{2i}^2) - (\sum x_{1i} x_{2i})^2} = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2}$$

2.7. Propiedades de la regresión lineal

Una vez obtenidos los estimadores de MCO de los datos de la muestra, se derivan los estimadores para los parámetros para la ecuación de regresión lineal. La función de regresión así obtenida posee las siguientes propiedades:

a) Es una línea que cruza por el punto $(\bar{X}, \bar{Y})$.
b) La suma de los residuos mínimo-cuadráticos es igual a cero, esto es: $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 0$
c) La suma de los productos cruzados entre la variable explicativa y los residuos es igual a 0, es decir $\rho_{\varepsilon X} = 0$
d) La suma de los productos cruzados entre la variable explicada y los residuos es igual a 0, es decir $\rho_{\varepsilon Y} = 0$

2.8. Bondad de ajuste y valoración del modelo

Una vez que se ha realizado el ajuste por mínimos cuadrados, conviene disponer de algún indicador que permita medir el grado de ajuste entre el modelo y los datos. En el caso de que se hayan estimado varios modelos alternativos podría utilizarse medidas de este tipo, a las que se denomina medidas de la bondad del ajuste, para seleccionar el modelo más adecuado.

La literatura estadística ofrece numerosas medidas de la bondad del ajuste. La más conocida es el coeficiente de determinación, al que se designa por r^2 o R cuadrado. (Gujarati & Porter, 2010). Como se verá en otro momento, esta medida tiene algunas limitaciones, aunque es válida para comparar modelos de regresión lineal simple.

El coeficiente de determinación se basa, fundamentalmente, en la descomposición de la varianza de la variable dependiente o predicha, a la que denominaremos varianza total, en términos de varianza explicada y residual. Vamos a ver a continuación como se obtiene esta descomposición.

Recordemos que $Y_i = \hat{Y}_i + \hat{\varepsilon}_i$

Restando a ambos miembros la media de Y $(\bar{Y})$, se tiene que: $Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + \hat{\varepsilon}_i$

Elevamos ambos miembros de la ecuación al cuadrado: $(Y_i - \bar{Y})^2 = \left[\left(\hat{Y}_i - \bar{Y}\right) + \hat{\varepsilon}_i\right]^2$

Sumando ambos miembros de la expresión anterior de 1 a *n*, y simplificando la ecuación se tiene:

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

Si esta expresión la dividimos por n en ambos términos de la ecuación se tiene:

$$\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i{}^2}{n}$$

Por lo tanto, la varianza total de la variable dependiente o predicha se descompone en dos partes: varianza explicada por la regresión o varianza de los valores ajustados y varianza residual. Es decir:

$\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n} = varianza\ total =$

$\frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{n}$ (varianza "explicada") + $\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2}{n}$ (varianza residual)

A partir de la descomposición anterior, el coeficiente de determinación se define como la proporción de la varianza total explicada por la regresión. Su expresión es la siguiente:

$$\frac{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{n}}{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}} = r^2$$

De forma equivalente, de acuerdo con la ecuación general de la varianza, el coeficiente de determinación se puede definir como 1 menos la proporción no explicada por la regresión, es decir, como:

$$1 - \left(\frac{\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2}{n}}{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}}\right) = r^2$$

Algebraicamente se puede deducir con una simple sustitución que los valores extremos del coeficiente de determinación (r^2) son: 0, cuando la varianza explicada es nula, y 1, cuando la varianza residual es nula, es decir, cuando el ajuste es perfecto.

3. Análisis Bivariado de datos cualitativos

Entenderemos por datos cualitativos el caso de datos cuantitativos con variables no métricas o categóricas. En la investigación social es muy frecuente la necesidad de lidiar con el procesamiento de datos para variables cualitativas o, en otros términos, variables categóricas o aquellas cuyo nivel de medición permite clasificarlas como: nominales. (Stevens, 1946). Ejemplo de estas son el sexo, la clase social, el lugar de procedencia, el estado civil, etc. Son variables que, en cualquier caso, sólo permiten el conteo de la cantidad de veces que aparecen los distintos valores o categorías que las constituyen como tales. En otras palabras, las variables cualitativas son, básicamente, aquellas que por su naturaleza no numérica sólo permiten el análisis de frecuencias.

Además del análisis de frecuencias, el cual es la forma más básica de análisis univariado de datos, el análisis cualitativo puede extenderse al análisis bivariado por la vía de los análisis de contingencia. Un análisis de continencia responde de forma fundamental a una dsitribución bivariada de datos, definidos en una matriz de doble entrada. Con el análisis de contingencia, las posibilidades heurísticas del investigador con relación a los datos se extiende para incluir, además de los análisis de frecuencia para cada variable por separado, los análisis de las frecuencias conjuntas o aquellos en los cuales el carácter de la unidad de análisis está definido por la combinación de dos valores (x,y).

Uno de los grandes recursos estadísticos para el análisis de datos cualitativos es la prueba de chí cuadrado. La misma es uno de los productos más importantes de la obra de Karl Pearson, con un rango de aplicaciones mucho mayor que el problema específico para el cual fue creado. (Walker, 1978). La prueba de chí cuadrado se fundamenta en el análisis de las frecuencias de los datos observados. Es por esta razón, que su aplicación es viable incluso para variables con un nivel de medición tan bajo como el nominal. Los principales contextos para el uso de la prueba de chí cuadrado se reducen a los siguientes: análisis de la bondad de ajuste y el contraste de independencia de las variables.

En cualquier caso, la hipótesis nula que da sentido al uso de la prueba de chí cuadrado es que las frecuencias observadas se comportan de la misma manera que las frecuencias determinadas para una distribución teórica de referencia. En el caso de los análisis univariados sirve de base para la inferencia del contraste asociado con la proporción. La hipótesis nula clásica, para los análisis de contingencia, sostiene que las n para las frecuencias de las distintas casillas definidas por un carácter i y j, son equivalentes más allá del error aleatorio.

3.1. Tablas de contingencia

Una distribución conjunta de dos variables Cuando se trabaja el cruce de variables categóricas se requiere, en principio, la construcción de bases de datos con por lo menos dos entradas, en las que cada entrada representa las variaciones de una determinada variable categórica. Como resultado de esta distribución se genera una presentación de los valores de cada una de las variables implicadas en filas y columnas, colocando en cada casilla el número de casos que cumple con ambos valores. De forma que las frecuencias hacen referencia a la presencia conjunta de los valores de las variables implicadas, en las distintas unidades de análisis y, por tanto, de la relación entre las variables involucradas. A estas tablas de frecuencia se les conoce como tablas de contingencia. (Pardo & Ruíz, 2005).

La tabla x, es un ejemplo de tabla de contingencia para el caso de una investigación en la cual el investigador se propuso estudiar la posible relación entre el bienestar psicológico y la disposición a fluir en el trabajo de una muestra de 200 empleados en el área de la salud.

Tabla x. *tabla de contingencia para la relación entre el bienestar psicológico y la disposición a fluir en el trabajo.*

		Disposición a Fluir en el Trabajo			
		Bajo	Medio	Alto	Total
Bienestar Psicológico General	Bajo	25	3	10	38
	Medio	65	49	22	136
	Alto	1	0	25	26
Total		91	52	57	200

Fuente: propia.

Tal como se mencionara en el párrafo anterior, en lugar de utilizar sólo dos variables o criterios de clasificación para generar una tabla de contingencia bidimensional, también se podría haber utilizado tres o más criterios, lo que llevaría a obtener tablas multidimensionales.

3.1.1. Notación de tablas de contingencia

Una vez familiarizados con lo que es una distribución conjunta y el marginal de una tabla, podemos pasar a emplear una notación para referirnos a cada uno de los elementos que la conforman. A los valores de la variable puestos en el eje de las filas se le denota como *i* y a los puestos en el eje de las columnas como *j*, por lo que a la frecuencia conjunta se le conoce como n*ij* . Al máximo de *i* se le denota como *I*, y al máximo de *j* como *J*, de forma que en este ejemplo *I*=2 y *J*=3. Para referirnos a la dimensión de la tabla, se estila multiplicar el número de filas por los de columnas del modo *I x J*, en este caso sería de 2 (filas) x 3 (columnas), por lo que nos referimos a esta como una tabla 2x3, lo que da un total de 6 casillas de frecuencias conjuntas. La Tabla 1 permite ilustrar la forma convencional de notación empleada para las tablas de contingencia. El resultado, es el siguiente:

Tabla x. *notación de una tabla 2x3.*

Variable A	**Variable B**			
	j=1	j=2	j=3	$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$
i= 1	n_{11}	n_{12}	n_{13}	$n_{1\cdot} = \sum_{i=1}^{J} n_{1j}$
i= 2	n_{21}	n_{22}	n_{23}	$n_{2\cdot} = \sum_{i=1}^{J} n_{2j}$
$n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$	$n_{\cdot 1} = \sum_{i=1}^{I} n_{i1}$	$n_{\cdot 2} = \sum_{i=1}^{I} n_{i2}$	$n_{\cdot 3} = \sum_{i=1}^{I} n_{i3}$	$N = n_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{i=1}^{j} n_{ij}$

Aclarada la notación para las frecuencias conjuntas, veamos ahora la notación para los marginales. Como ya vimos, los marginales se calculan sumando en el sentido en el que está la variable. En nuestro ejemplo, la variable B conforma las columnas, y por tanto corresponde con la letra *j*. Si estamos interesados en calcular la cantidad total de elementos en B (que se corresponde con j=1), en términos de notación debemos sumar n11+n21, es decir, el subíndice *i* cambia, pero el subíndice *j* se mantiene igual (j=1). Para expresar este cambio de la i y esta constancia de la j diremos que el marginal es n·j, en este caso más en concreto n·1=n11+n21.

Esto amerita que recordemos que el · significa que estamos sumando las filas (i) en el sentido de las columna que estamos calculando (j).

Hasta ahora se ha presentado la notación para hacer referencia a la distribución conjunta de variables, en términos de las frecuencias absolutas de cada casilla (en este caso, el número de personas). Sin embargo, en algunas ocasiones el investigador puede preferir otras opciones a las frecuencias absolutas, como es el caso de los porcentajes para poder comparar tanto poblaciones de distinto tamaño como las casillas de una misma tabla en la que los marginales son distintos. (Pardo & Ruíz, 2005). Lo habitual en ciencias sociales es que el resultado quede redondeado a un decimal, a ninguno (menos frecuente en el ámbito académico) o quizá hasta la centésima (en un prurito inútil de precisión), pero no más.

3.1.1.1. Cálculo de porcentajes

Podemos recordar que el procedimiento general para calcular cualquier tipo de porcentajes es sencillo: basta con dividir la frecuencia de la casilla que nos interesa sobre el total (marginal) que corresponda, y multiplicar por 100. A diferencia del cálculo de los porcentajes para distribuciones univariadas, en el caso de las distribuciones conjuntas se nos presentan las opciones de calcular el porcentaje de las casillas con base en los tres totales que evidenciamos en la sección anterior, a saber: a) sobre la base del total general, b) sobre la base del total de las filas, y c) sobre la base del total de las columnas.

Tomemos el caso *a* cálculo de los porcentajes con base al total general, si deseamos calcular el porcentaje sobre la base de éste total, lo que habrá que hacer es dividir la frecuencia de cada casilla (n_{ij}) sobre el total de la tabla (N). La fórmula, por tanto sería:

$$p_{ij}^{T} = \frac{n_{ij}}{N} * 100$$

Donde p_{ij}^{T} hace referencia al porcentaje para una determinada casilla *ij* con base al total (*T*), n_{ij} es el número de elementos que se encuentran en la casilla *ij* y N representa la sumatoria de todos los elementos presentes en la tabla, para todas las filas y columnas.

Para el caso *b* cálculo de los porcentajes con base al total de las filas, si deseamos calcular el porcentaje sobre la base de éste total, lo que habrá que hacer es dividir la frecuencia de cada casilla (n_{ij}) sobre el total de la fila en cuestión ($n_{i.}$). La fórmula, por tanto sería:

$$p_{ij}^{F} = \frac{n_{ij}}{n_{i.}} * 100$$

Donde p_{ij}^{F} hace referencia al porcentaje para una determinada casilla *ij* con base al total de la fila correspondiente ($n_{i.}$), n_{ij} es el número de elementos que se encuentran en la casilla *ij* y $n_{i.}$ representa la sumatoria de todos los elementos presentes en la fila *i*, para todas las columnas.

Para el caso *c* cálculo de los porcentajes con base al total de las columnas, si deseamos calcular el porcentaje sobre la base de éste total, lo que habrá que hacer es dividir la frecuencia de cada casilla (n_{ij}) sobre el total de la columna en cuestión ($n_{.j}$). La fórmula, por tanto sería:

$$p_{ij}^{C} = \frac{n_{ij}}{n_{.j}} * 100$$

Donde p_{ij}^{C} hace referencia al porcentaje para una determinada casilla *ij* con base al total de la columna correspondiente ($n_{.j}$), n_{ij} es el número de elementos que se encuentran en la casilla *ij* y $n_{.j}$ representa la sumatoria de todos los elementos presentes en la columna *j*, para todas las filas.

3.2. Lógica del contraste de *ji* cuadrado

Como hemos visto hasta ahora, cuando trabajamos con bases de datos cualitativas, la información que es capaz de proporcionarnos el estadístico estudiado, en los términos expuestos, hace referencia únicamente a la frecuencia o cantidad de individuos que cumplen con una determinada característica (simple o conjunta). En ningún caso se nos informa sobre la magnitud con la cual se presenta el carácter en cuestión. De modo que, un análisis de los datos daría lugar al tipo de pregunta siguiente: difieren de forma significativas las frecuencias observadas de las que se esperarían si se cumpliera la hipótesis de nulidad?

Por esta razón se desarrolla la prueba de ji cuadrado (χ^2) también conocida como chi-cuadrado. Algunos de los casos más emblemáticos para el uso de pruebas χ^2 son:

- La prueba χ^2 de frecuencias
- La prueba χ^2 de independencia
- La prueba χ^2 de bondad de ajuste
- La prueba χ^2 de Pearson con corrección por continuidad o corrección de Yates
- La prueba de Bartlett de homogeneidad de varianzas

En cualquier caso, la prueba de ji cuadrado permite contrastar el comportamiento de las frecuencias observadas en una distribución con la distribución de las frecuencias esperadas teóricamente cuando se cumple la hipótesis nula. (Siegel & Castellan, 1995).

Debido a las limitaciones para este espacio, se expondrán tres de las situaciones más emblemáticas en las cuales se utiliza la prueba de ji cuadrado, como lo son: el contraste de bondad de ajuste, que trata de determinar si una población cumple con una distribución específica; b) la prueba de independencia, que pretende poner de manifiesto la ausencia de relación entre las variables y c) la prueba de homogeneidad, que busca demostrar que las categorías o porciones en que se divide la población son homogéneas.

3.2.1. Bondad de ajuste: estructura de los datos y estimación de frecuencias esperadas.

Para comparar la distribución de unas frecuencias observadas contra aquellas esperadas teóricamente, debemos ser capaces de establecer qué frecuencias deben ser esperadas, cuando se cumple la hipótesis de nulidad. La hipótesis nula, en su expresión más ortodoxa, se comporta como una distribución uniforme y establece la equivalencia estadística de las frecuencias para cada una de las categorías de la variable. No obstante, otras aplicaciones del método pueden incluir valores teóricos específicos para las frecuencias esperadas en cada una de las categorías de la variable de estudio. En cualquier caso, la prueba de bondad de ajuste determinará la diferencia estadística entre la distribución observada y la teórica.

Su cálculo se resume a la siguiente ecuación:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Donde:

χ^2= valor estimado de chi cuadrado para los datos observados.

O_i= el número de casos (frecuencia) observados en la i-ésima casilla.

E_i= el número de casos (frecuencia) observados en la i-ésima casilla, cuando la hipótesis nula es verdadera.

k= el número de categorías.

La significación estadística de este valor estimado de chi cuadrado puede ser determinada con el uso de tablas que resumen los valores críticos para distintas distribuciones muestrales que se presentan como anexos en la mayoría de los textos de estadística inferencial. La distribución de probabilidad para los valores estimados de chi cuadrado se comporta como una distribución chi cuadrada con k-1 grados de libertad.

3.2.2. Igualdad de porciones e independencia: estructura de los datos y estimación de frecuencias esperadas.

En ocasiones, el investigador puede verse en la necesidad de responder a la pregunta: cuán diferentes son los comportamientos de las frecuencias observadas para una determinada variables entre una muestra de individuos agrupados en dos o más grupos? Dado que las medidas realizadas obedecen a un nivel de medición tan bajo como nominal.

La hipótesis que está siendo probada generalmente es aquella que postula que los grupos difieren respecto de alguna característica y, por tanto, respecto a la frecuencia relativa con que los miembros de los grupos caen dentro de alguna categoría. (Siegel & Castellan, 1995). Al igual que en el caso de la bondad de ajuste para una muestra, la esencia del contraste consiste en evaluar estadísticamente las diferencias entre las frecuencias observadas y las esperadas teóricamente. De hecho, su cálculo es una derivación de la fórmula para la bondad de ajuste cuando la distribución está definida por una tabla de doble entrada, tal como se presenta a continuación:

Su cálculo se resume a la siguiente ecuación:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Donde:

χ^2= valor estimado de chi cuadrado para los datos observados.

n_{ij}= el número de casos (frecuencia) que cumple con la doble condición observados en la i-ésima fila y j-ésima columna.

E_{ij}= el número de casos (frecuencia) que cumple con la doble condición observados en la i-ésima fila y j-ésima columna, cuando la hipótesis nula es verdadera.

r= el número de categorías por filas.

c= el número de categorías por columnas.

Dado que se trata de una tabla de doble entrada y bajo el supuesto de independencia para la hipótesis de nulidad, la frecuencia esperada por celda debería ser proporcional a la distribución total de filas y columnas. A diferencia del caso para la bondad de ajuste con una muestra, se calcula la frecuencia esperada en cada celda con la fórmula:

$$E_{ij} = \frac{n_{i.} n_{.j}}{N}$$

Donde:

$n_{i.} = \sum_{i=1}^{J} n_{ij} =$ el número total de casos (frecuencia) observados en la i-ésima fila, para todas las columnas.

$n_{.j} = \sum_{j=1}^{I} n_{ij}$= el número total de casos (frecuencia) observados en la j-ésima columna, para todas las filas.

$N = n_{..} = \sum_{i=1}^{I} \sum_{i=1}^{j} n_{ij}$= número total de observaciones realizadas.

La distribución de probabilidad para los valores estimados de chi cuadrado se comporta como una distribución chi cuadrada con (r-1)x(c-1) grados de libertad.

Bibliografía

Abelson, R. (1998). *La estadística razonada: reglas y principios.* Barcelona: Paidós.

American Psychological Association. (01 de Junio de 2010). *Ethical Principles of Psychologists and Code of Conduct.* Recuperado el 13 de Marzo de 2013, de http://www.apa.org/ethics/code/index.aspx#

American Psychological Association. (2010). *Manual de Publicaciones de la Amercan Psychological Assciation* (Tercera ed.). (M. Guerra Frias, Trad.) D.F., México: Manual Moderno.

Anastasi, A., & Urbina, S. (1998). *Test Psicológicos.* D.F., México: Prentice Hall.

Aparicio M., G. (1985). *Teoría subjetiva de la probabilidad: fundamentos, evolución y determinación de probabildiades.* Recuperado el 01 de Marzo de 2013, de http://eprints.ucm.es/7818/1/01.pdf

Asamblea Nacional Constituyente. (30 de Diciembre de 1999). Constitución de la República Bolivariana de Venezuela. *Gaceta Oficial Nº 36.860*. Caracas, Venezuela: Asamblea Nacional.

Asamblea Nacional de la República Bolivariana de Venezuela. (12 de Diciembre de 2010). Reforma de Ley Orgánica de Ciencia y Tecnología. *Gaceta Oficial Nº 39.575*. Caracas, Venezuela: Asamblea Nacional.

Babbie, E. (1988). *Métodos de investigació por encuesta.* (J. Utrilla, Trad.) D.F., México: Fondo de Cultura Económica.

Balluerka, N., & Vergara, A. (2002). *Diseños de Investigación Experimental en Psicología.* Madrid, España: Prentice Hall.

Birnbaum, A. (1977). The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. *Synthese, 36*(1), 19-49.

Brown, C., & Ghiselli, E. (1969). *El método científico en Psicología.* (E. Prieto, Trad.) Buenos Aires, Argentina: Paidós.

Campbell, D., & Stanley, J. (1970). *Diseños experimentales y cuasi-experimentales en la investigación social.* Buenos Aires, Argentina: Amorrortu.

Campbell, N. (1921). *What is Science?* New York: Dover Publications.

Carnap, R. (1934). On the character of philosophic problems. *Philosophy of Science, 51*(1), 5-19.

Carnap, R. (1950). *Logical Foundations of Probability.* Chicago, U.S.A.: The University of Chicago Press.

Chávez, H. (01 de Noviembre de 2001). Decreto con fuerza de Ley de la función púiblica de la Estadística. Caracas, Venezuela: Gaceta Oficial Nº 37.321.

Cochran, W. (Diciembre de 1950). The comparison of percentages in matched samples. *Biometrika, 37*(3/4), 256-266.

Cohen, J. (1992). Cosas que he aprendido (hasta ahora). *Anales de Psicología, 8*(1-2), 3-17.

Comisión de Ética, Bioética y Biodiversidad. (Diciembre de 2010). Código de ética para la vida. (T. e. Ministerio del Poder Popular para la Ciencia, Ed.) Caracas, Venezuela: Ministerio del Poder Popular para la Ciencia, Tecnología e Industrias Intermedias.

Congreso de la Republica de Venezuela. (11 de Septiembre de 1978). Ley de Ejercicio de la Psicología. *Gaceta Oficial nº 2306*. Caracas, Venezuela.

Cowles, M., & Davis, C. (1982). On The Origins of .05 Level of Significance. *American Psychologist, 37*(5), 553-558.

Diez Calzada, J. A. (1992). En torno a la lógica de la inferencia. *Enrahonar: Quaderns de filosofía*, 91-97.

Dingle, H. (1950). A theory of measurement. *The British Journal for the Philosophy of Science*, 5 - 26.

Federación de Psicólogos de Venezuela. (1981). Código de Ética Profesional del Psicólogo de Venezuela. *II Asamblea Nacional Ordinaria de la Federación de Psicólogos de Venezuela.* Barquisimeto: Autor.

Fisher, R. (Julio de 1925). Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society, 22*(5), 700-725.

Fisher, R. A. (1935). The Logic of Inductive Inference. *Journal of the Royal Statistical Society*, 39 - 82.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association, 32*(200), 675-701.

Gaito, J. (1980). Measurement scales and statistics: resurgence of an old misconception. *Psychological Bulletin*, 564 - 567.

Glass, G., & Stanley, J. (1974). *Métodos Estadísticos Aplicados a las Ciencias Sociales.* (E. Galvis, & E. Guzman, Trads.) Madrid, España: Prentice Hall.

Grande, I., & Abascal, E. (1989). *Métodos de análisis multivariante para la investigación comercial.* Barcelona: Ariel.

Gujarati, D., & Porter, D. (2010). *Econometría* (Quinta ed.). (P. Carril Villareal, Trad.) D.F., México: Mc Graw Hill.

Gutiérrez, H., & de la Vara, R. (2012). *Análisis y Diseño de Experimentos* (Tercera ed.). D.F., México: Mc Graw Hill.

Hand, D. J. (1996). Statistics and the Theory of Measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 445 - 492.

Hernández, R., Fernández, C., & Baptista, P. (2010). *Metodología de la Investigación* (Quinta ed.). D.F., México: Mc Graw Hill.

Kerlinger, F., & Lee, H. (2002). *Investigación del comportamiento. Métodos de investigación en ciencias sociales. (4a Ed).* México: Mc Graw Hill.

Knapp, T. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nurse Research*, 121 - 123.

Kruskal, W., & Wallis, W. (Diciembre de 1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association, 47*(260), 583-621.

Landau, D., & Lazarsfeld, P. (1978). Quetelet Adolphe. En W. Kruskal, & J. Tanur, *International Encyclopedia of Statistics* (Vol. 2, págs. 824-834). New York, USA: The Fee Press.

León, O., & Montero, I. (2003). *Métodos de Investigación en Psicología y Educación* (Tercera ed.). Madrid, España: Mc Graw Hill.

Lezama, L. (2011). Puntuaciones relacionadas con las normas. *Psicología*, 107-143.

López Casuso, R. (1996). *Cálculo de probabilidades e inferencia estadística* (Tercera ed.). Caracas, Venezuela: Publicaciones UCAB.

Lord, F. (1953). On the statistical treatment of football numbers. *The American Psychologist*, 750 - 751.

Luce, R. D. (1997). Quantification and symmetry: Commentary on Michell Quanttative Science and the definition of measurement in Psychology. *British Joumat of Psychology*, 395 - 398.

Magnusson, D. (1969). *Teoría de los test.* Trillas.

Magnusson, D. (1975). *Teoría de los test.* México: Biblioteca Técnica de Psicología.

Mann, H., & Whitney, D. (Marzo de 1947). On a test of wether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*(1), 50-60.

Marx, M., & Hillix, W. (1983). *Sistemas y teorías psicológicas contemporáneas* (3a ed.). Buenos Aires, Argentina: Paidós.

Mayo, D., & Cox, D. (2006). Frequentist statistics as a theory of inductive inference. *Lecture Notes-Monograph Series, 49*, 77-97.

Mayo, D., & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. *Lecture Notes-Monograph Series*, 77-97.

McGuigan, F. (1977). *Psicología Experimental: enfoque metodológico* (Segunda ed.). (A. Fabre, Trad.) D.F., México: Trillas.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157.

Michell, J. (1986). Measurement scales and statistics: a clash of paradigms. *Psychological Bulletin, 100*(3), 398 - 407.

Montgomery, D. C. (1984). *Design adn Análysis of Experiments* (Second ed.). New York, United States of América: John Wiley & Sons, Inc.

Morris, C., & Maisto, A. (2009). *Psicología* (13a ed.). D.F., México: Pearson Educación.

Muñíz, J. (1998). La medición de lo Psicológico. *Psicothema*, 1 - 21.

Narens, L., & Luce, R. D. (1986). Measurement: The Theory of Numerical Assignments. *Psychological Bulletin*, 166 - 180.

Navarro, A. (1989). *La Psicología y sus múltiples objetos de estudio.* Caracas, Venezuela: Consejo de Desarrollo Científico y Humanístico.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese, 36*, 97-131.

Neyman, J., & Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika, 20*(1/2), 175-240.

Pagano, R. (2011). *Estadística para las ciencias del comportamiento.* (M. Torres, Trad.) D.F., México: CENGAGE Learning.

Pardo, A., & Ruíz, M. (2005). *Análisis de datos con SPSS 13 Base.* Madrid, España: Mc Graw Hill.

Popper, K. (1962). *La Lógica de la Investigación Científica.* Madrid, España: Tecnos.

Recalde, L. C. (2009). Los axiomas de la cantidad de Hölder y la fundamentación del continuo lineal. *Matemáticas: Enseñanza Universitaria*, 101 - 121.

Reichenbach, H. (1949). *The Theiry of Probability.* Los Angeles, U.S.A.: University of California Press.

Restrepo, L., & Gonzalez, J. (2003). La Historia de la Probabilidad. *Revista Colombiana de Ciencias Pecuarias*, 83 - 87.

Rivadulla, A. (1991). *Probabilidad e inferencia científica.* Barcelona: Anthropos.

Saavedra, N. (2000). La axiomática de Kolmogorov: fundamentos de la teoría de la probabilidad. *Números, 43*, 185-190.

Sáchez Carrion, J. (2001). Estadística, orden natural y orden social. *Papers*, 33 - 46.

Sáiz Roca, M., de la Casa Rivas, G., Dolores Saíz, L., Ruiz, G., & Sánchez, N. (2009). Fundación y establecimiento de la Psicología Científica. En M. Sáiz Roca, *Historia de la Psicología* (págs. 55 - 150). Barcelona: UOC.

Sáiz, M. (2009). Los tiempos de reacción. La ecuación personal y el impulso nervioso. En M. Sáiz Roca, *Historia de la Psicología* (págs. 43 - 46). Barcelona: UOC.

Sáiz, M., & Sáiz, D. (2009). La Psicología Científica Británica. En M. Sáiz Roca, *Historia de la Psicología* (págs. 97 - 113). Barcelona: UOC.

Siegel, S., & Castellan, N. (1995). *Estadística No Paramétrica* (Cuarta ed.). (L. Aragón, & L. Fierros, Trads.) D.F., México: Trillas.

Sojo, V. (2004). *Ética en Investigación Psicológica con Humanos.* Manuscrito No publicado, Universidad Central de Venezuela, Escuela de Psicología, Caracas.

Stahl, S. (2006). The evolution of the normal distribution. *Mathematics Magazine*, 96 - 113.

Stevens, S. (Abril de 1935). The Operational Basis of Psychology. *The American Journal of Psychology, 47*(2), 323-330.

Stevens, S. (1935). The operational definition of psychological concepts. *Psychological Review, 42*(6), 517-527.

Stevens, S. (1946). On the theory of scales of measurement. *Science*, 677 - 680.

Student. (Marzo de 1908). The Probable Error of a Mean. *Biometrika, 6*(1), 1-25.

Thomas, H. (1982). IQ interval scales, and normal distribution. *Psychological Bulletin*, 198 - 202.

Vargas Sabadías, A. (1995). *Estadística Descriptiva e Inferencial.* Publicaciones de Universidad de Castilla-La Mancha.

Velleman, P., & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician*, 65 - 72.

Walker, H. (1978). Pearson, Karl. En W. Kruskal, & J. Tanur, *International Encyclopedia of Statistics* (Vol. 2, págs. 691-698). New York, U.S.A.: The Free Press.

Wiener, P. (1978). Peirce, Charles Sanders. En W. Kruskal, & J. Tanur, *International Encyclopedia of Statistics* (Vol. 2, págs. 698-702). New York, U.S.A.: The Free Press.

Wilcoxon, F. (Diciembre de 1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80-83.