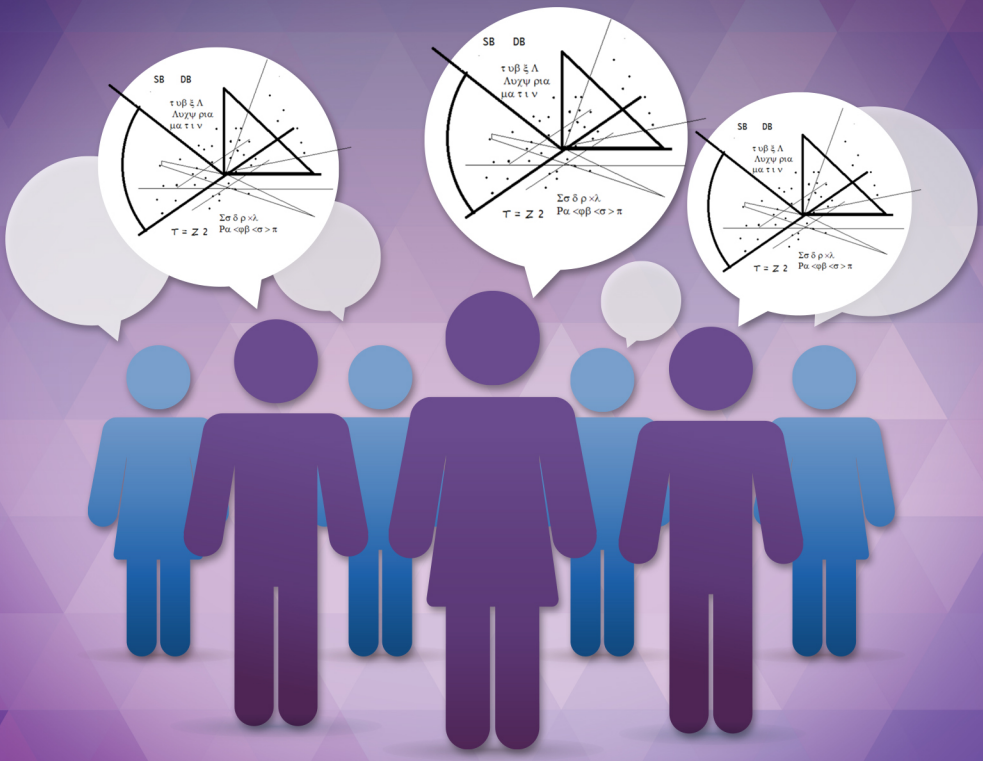
DISEÑOS MULTIVARIADOS DE INVESTIGACIÓN EN LAS CIENCIAS SOCIALES



LUCY MARÍA REIDL MARTÍNEZ
RAQUEL DEL SOCORRO GUILLÉN RIEBELING

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE ESTUDIOS SUPERIORES ZARAGOZA

DISEÑOS MULTIVARIADOS DE INVESTIGACIÓN EN LAS CIENCIAS SOCIALES

LUCY MARÍA REIDL MARTÍNEZ
RAQUEL DEL SOCORRO GUILLÉN RIEBELING

Universidad Nacional Autónoma de México
Facultad de Estudios Superiores Zaragoza

Datos para catalogación bibliográfica

Autora: Lucy María Reidl Martínez, Raquel del Socorro Guillén Riebeling.

Diseños multivariados de investigación en las ciencias sociales.

UNAM, FES Zaragoza, diciembre de 2019.

Peso: 5.1 MB.

ISBN: 978-607-30-2716-8.

Diseño de portada: Carlos Raziel Leaños Castillo.
Diseño y formación de interiores: Claudia Ahumada Ballesteros.

Diseños multivariados de investigación en las ciencias sociales.

D.R. © Universidad Nacional Autónoma de México
Av. Universidad # 3000, Col. Universidad Nacional Autónoma de México, C.U.,
Alcaldía Coyoacán, C.P. 04510, Ciudad de México, México

Facultad de Estudios Superiores Zaragoza
Av. Guelatao # 66, Col. Ejército de Oriente,
Alcaldía Iztapalapa, C.P. 09230, Ciudad de México, México

Índice

Presentación

Este libro tiene por objeto presentar a los estudiosos de las ciencias sociales en general y de la Psicología en particular, una versión simplificada y explicativa del uso de los diseños multivariados de investigación para abordar problemas y/o temas complejos en campos de la Psicología -clínica, educativa, social, salud, laboral, entre otros, y de las ciencias sociales. La comprensión de la aplicación del método, y así, poder seleccionar entre los diferentes diseños, el más adecuado al problema de investigación; la fundamentación de esa selección; cómo poder interpretar los resultados obtenidos por el paquete SPSS (Statistical Package for the Social Sciences); así como poder elaborar proyectos que requieran de este tipo de diseños. En síntesis, se pretende demostrar por medio de ejemplos desarrollados en su totalidad, la facilidad del empleo de este tipo de diseños, así como la riqueza de la información que proporcionan.

Los diseños factoriales del capítulo inicial, permiten aproximarse al análisis de datos provenientes de estudios de laboratorio o de investigaciones de campo, que tratan de asegurar una interpretación o explicación "causal" de los fenómenos estudiados.

El diseño de regresión múltiple, que se presenta en el segundo capítulo, permite aproximarse a explicaciones multicausales de algún fenómeno, así como también alienta al investigador a realizar estudios predictivos de fenómenos importantes.

El análisis factorial, presentado en el tercer capítulo, es otro diseño multivariado que ha demostrado ser muy útil en el desarrollo de instrumentos de medición o registro de variables, así como para poner a prueba hipótesis referidas a variables complejas.

En el cuarto capítulo, se presenta el análisis de discriminantes. Este diseño permite aproximarse en forma empírica al estudio de las constelaciones de variables que explican las diferencias existentes entre diversos grupos sociales. Permite también, predecir membresía a diversos grupos y ha sido empleado también, para establecer la validez concurrente de instrumentos o pruebas.

Por último, se presenta el diseño de la correlación canónica, que permite por primera vez, estudiar a un conjunto de variables en su relación a otro conjunto, pero al mismo tiempo, permitiéndole al investigador, desentrañar la compleja estructura que relaciona a ambos conjuntos.

Prólogo

La estadística es una disciplina difícil de definir. En las definiciones más comunes, no hay un acuerdo claro sobre si es una rama de las matemáticas, si es una matemática diferente o si es una ciencia objetiva. Un criterio que puede aclarar el punto es que en nuestro mundo hay eventos cuya ocurrencia se pueden anticipar de manera más o menos precisa con base en el cálculo y otras que ocurren al azar.

En el ámbito de la física, se pueden calcular con precisión la fuerza de atracción de un cuerpo, la aceleración y trayectoria de un proyectil, o la presión que cierto volumen de un gas ejerce sobre las paredes de su recipiente dependiendo de su temperatura. En este caso frecuentemente se dice que los eventos son "predecibles"(aunque este adjetivo no es muy acertado).

Por otro lado, si se conocen las condiciones ambientales de temperatura, presión, humedad, altitud, región geográfica y algunas otras más, no es posible saber con exactitud si un día determinado va a llover o no (es "impredecible"), y solamente podemos tener cierto grado de confianza en que así será, para que podamos decidir si es necesario o no llevar al trabajo un buen paraguas. La confianza en la ocurrencia o ausencia de estos eventos se ha denominado probabilidad, y en la actualidad es calculada con procedimientos estadísticos adecuados.

Así, en el mundo, los eventos "impredecibles" son más numerosos que los "predecibles". Por señalar: la ocurrencia de los sismos, de las inundaciones, el peso de un niño al nacer, la temperatura ambiente del medio día y casi cada evento de la vida cotidiana. Estos eventos se denominan aleatorios y puede considerarse que no se pueden "predecir" porque sus causas con complejas y no siempre conocidas; por lo tanto, no se puede calcular el momento en que ocurren o su magnitud. Otra posibilidad es suponer que, en realidad, no hay eventos causales, de modo que lo único que se puede hacer es registrar su ocurrencia. Otro ejemplo de eventos impredecibles son la ocurrencia de un par en una partida de cartas, la obtención de un "sol" en un volado o de un doble seis en un lanzamiento de dos dados, etcétera. Por esto, no es raro que los juegos de azar hayan servido como un modelo para calcular la probabilidad de ocurrencia de eventos aleatorios y aplicarlo en el campo de estudio de diversas ciencias, incluso de las denominadas "exactas", entre ellas la física o la química.

En Psicología, la mayoría de los eventos -o variables- abordados son aleatorios, de modo que la estadística ha servido para estudiarlos casi desde su inicio como ciencia formal a fines del

Siglo XIX y principios del Siglo XX, con el trabajo empírico y matemático de Francis Galton, Karl Pearson, William Gosset y Ronald Fisher, por mencionar a los pioneros más conocidos.

La utilidad de los métodos estadísticos para el desarrollo de otras ciencias consiste en que proporciona un procedimiento para conocer el comportamiento de variables aleatorias: cómo medirla, cómo se seleccionan los casos para obtener los datos, su concentración, su resumen, el cálculo de los estadígrafos más útiles, su representación y el cálculo de su probabilidad. Este método, aplicado a la Psicología proporciona información valiosa, como podría ser el valor esperado de la inteligencia de una persona, la actitud hacia los grupos minoritarios o el nivel de estrés en una población o las diferencias que se dan en distintos grupos con respecto a los valores de variables de nuestro interés, lo cual es compatible y complementario con el método científico.

En esta obra, la Dra. Lucy María Reidl Martínez, autora principal de la obra, en co-autoría con la Dra. Guillén Riebeling, presentan al lector, con un lenguaje y estilo claro, cómo la estadística se convierte en una herramienta eficaz para cumplir con los objetivos de la investigación científica en Psicología. Dada la dificultad de medir las variables psicológicas, las autoras abordan el proceso la medición y los procedimientos empleados en la construcción de escalas, con énfasis en la medición de su confiabilidad y validez, así como en los métodos de muestreo para la recolección de datos.

En esta obra se explica con claridad y detalle, los diseños de investigación más importantes, así como los procedimientos estadísticos pertinentes a cada caso. La lógica empleada en el análisis y la interpretación de los resultados es una parte importante en cada capítulo, con lo cual el lector alcanza una mejor comprensión de la utilidad del uso de esta herramienta en la investigación psicológica.

En resumen, esta obra se encuentra de una manera sencilla y clara el desarrollo del método científico en Psicología y cómo se utiliza una herramienta tan poderosa como lo es la estadística para extraer información importante, válida y confiable, a partir de los datos obtenidos de la investigación psicológica. El uso armonioso de ambos métodos -científico y estadístico- rara vez se expone con tanta accesibilidad por lo que la lectura de esta obra resulta sumamente interesante e instructivo.

Félix Ramos Salamanca
Profesor Carrera de Psicología
Facultad de Estudios Superiores Zaragoza, UNAM

Tema I Introducción a los Diseños Multivariados de Investigación

1.1 Necesidad de la aproximación multivariada en la investigación

Las ciencias sociales son aquellas que estudian y tratan de entender al hombre, sus instituciones, organizaciones, sus acciones y comportamientos racionales. En general, se considera que la sociología, la psicología, la antropología, la economía y las ciencias políticas, son las clasificadas dentro de este rubro. Su o sus objetos de estudio, sus unidades y niveles de análisis son de los más complejos y ambiguos que existen.

Esta complejidad y ambigüedad determinan, casi de manera automática, que su conocimiento o explicación no pueda ser simple. De esta manera, casi por principio, se debe pensar que la mayoría de los fenómenos que estas ciencias pretenden explicar, deben estar causadas, producidas, influidas o determinadas, por más de una variable. Es decir, son multicausadas. Aunque esto se ha sabido desde hace tiempo, solo hasta hace poco se ha podido estudiar empíricamente con un enfoque multivariado, que representa de manera más cercana a la realidad, la forma en que las variables se afectan entre sí.

El diseño de investigación es la disciplina de los datos. Su propósito implícito consiste en poner restricciones controladas sobre las observaciones de los fenómenos naturales. Un diseño de investigación le señala al experimentador lo que tiene que hacer y lo que no, los aspectos con los que habrá de ser cuidadoso, aquellos que deberá ignorar; es un esquema de la investigación. Si el diseño está bien concebido, el producto último del estudio tendrá más probabilidades de ser válido, desde el punto de vista empírico, y ser serio desde el punto de vista científico. La elegancia y el poder del diseño de investigación moderna y la idea de que constituye la disciplina de los datos, se hace más aparente en los diseños factoriales que se verán en esta obra.

A principios del siglo xx, la mayoría de los estudios de las disciplinas sociales empleaban una sola variable independiente, y solo dos condiciones experimentales. Este era el "diseño de investigación clásico", donde a un grupo se le llamaba experimental y al otro de control. La

idea de dos condiciones experimentales se puede ampliar a más de dos. Sin embargo, sigue habiendo una sola variable independiente: únicamente se aumenta el número de condiciones experimentales. A medio siglo, se dio un cambio trascendental en la conceptualización del diseño de investigación y en el análisis estadístico. Se introdujo más de una variable independiente y eventualmente esos diseños se llamaron factoriales. Estos últimos consisten en esencia de diseños experimentales en los que se utilizan de manera simultánea, dos o más variables independientes para estudiar sus efectos por separado o en conjunto sobre una sola variable dependiente. Estos diseños permitían estudiar problemas e hipótesis complejas de investigación. Estos diseños poseen varias ventajas, entre las más importantes están: Formular y someter a prueba teorías más complejas; Investigar problemas más realistas; y Estudiar la influencia conjunta de variables. A finales de siglo e inicio del nuevo milenio, el uso de las nuevas tecnologías permite a los investigadores elaborar estudios de manera más dinámica, contar con herramientas para los análisis de datos y obtener resultados factibles de interpretación. La aportación del conocimiento depende de los hallazgos y de la destreza del investigador para plasmarlo, habilidad insuperable por los programas de computación, por lo que el estudioso de los fenómenos sociales tendrá en sus manos la última palabra de lo obtenido y de su comunicación.

Antes de continuar, es preciso hablar de un aspecto técnico: la partición de la Varianza.

1.2 Partición de la varianza

Para poder aprovechar en forma adecuada el enfoque multivariado al estudio de los fenómenos sociales y/o psicológicos, es necesario entender una idea técnica muy simple pero muy útil también: la partición de la varianza.

La varianza tiene dos significados en la investigación: **Primero,** la varianza se emplea como un término general que expresa la variabilidad de las características de los individuos u objetos y sus diferencias. Sin embargo, también se puede asociar con diferencias entre objetos y grupos, de aquí que la idea de las diferencias individuales sea general, siempre y cuando se defina el término individual en forma amplia; **Segundo,** Este significado es más sutil y técnico. El investigador se refiere a la cantidad de varianza de una variable dependiente que

“se debe a" o “queda explicado por" una manipulación experimental o por otras variables. En la última instancia, se está haciendo una aseveración multivariada.

Por otro lado, se dirá que el enfoque es multivariado en el momento que existen dos o más variables independientes y una o más variables dependientes. Cuando se dice "se debe a", "da cuenta de", “queda explicada" o "influye", se está implicando causa o causalidad. Sin embargo, esta no es la intención: las expresiones mencionadas son en realidad un producto secundario del lenguaje.

En las Ciencias Sociales, el principio de causalidad se maneja en un extenso sentido. Se habla más bien de una dirección predominante de influencia, que queda establecida por la temporalidad de aparición de las variables, así como por su modificación. No se puede hablar de causalidad en un sentido estricto, los científicos no usan la palabra causa, especialmente porque resulta virtualmente imposible el afirmar que una cosa causa a la otra, siempre existe la posibilidad de que la supuesta causa de algo no sea la real. Lo más que el científico puede aventurarse es el establecer que existe alguna relación y que es de ésta o aquella. La cuestión es realmente académica, ya que en la ciencia no es necesario hacer afirmaciones causales; es suficiente con las llamadas afirmaciones condicionales del siguiente tipo: Si P, entonces Q, las cuales carecen de implicación causal.

Los fenómenos que presentan variabilidad o varianza, se llaman Varianza Total. Aquella que queda explicada o de la que se da cuenta, se llama Varianza Explicada o Varianza Experimental. La diferencia entre la varianza total y explicada, se conoce como Varianza Residual o Varianza de Error. De esta manera, se puede presentar a la partición de la varianza de la siguiente manera:

$$VT = VE + VR$$

Donde:

VT = Varianza Total.

VE = Varianza Explicada o Experimental.

VR = Varianza Residual o de Error.

Por otro lado, cuando son diversas las variables que "dan cuenta" de la varianza total de otra variable, las primeras pueden ser independientes, ortogonales o no relacionadas entre sí (Figura 1); o dependientes relacionadas entre sí (Figura 2). Los esquemas que corresponderían a cada situación, son los siguientes:

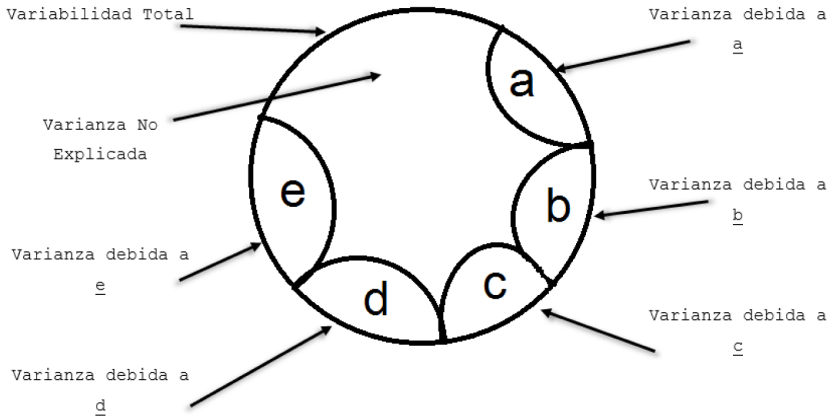


Figura 1. Variables independientes u ortogonales entre sí.

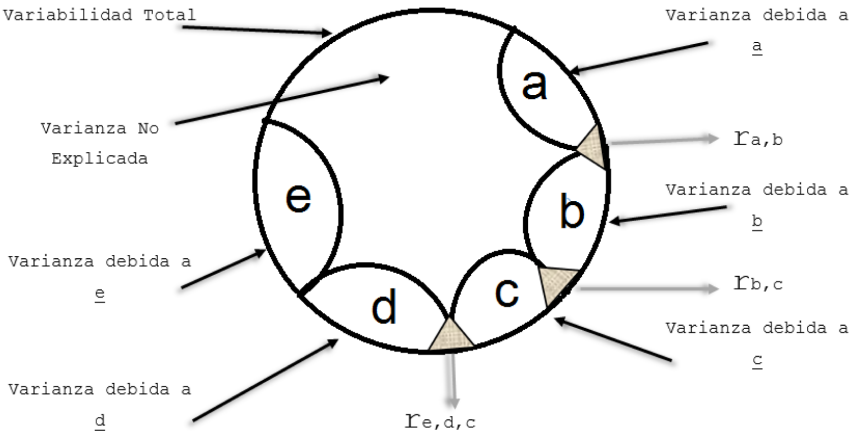


Figura 2. Variables dependientes o relacionadas entre sí.

En la Figura 1 se dice que cada variable (a,b,c,d y e) "explican" o "dan cuenta" de cierto X porcentaje de varianza total; cada una por separado. La suma de todos estos porcentajes dirían qué tanta varianza total quedó explicada por las cinco variables.

En la Figura 2 algunas variables correlacionan entre sí (ra,b; rb,c; re,d,c). En este momento existen correlaciones bivariadas (ra,b), (rb,c) y correlaciones múltiples (re,d,c.) que implican compartir varianza o sea, la covarianza. Una correlación elevada al cuadrado es una covarianza (caso bivariado). Una correlación múltiple elevada al cuadrado es un coeficiente de determinación[1] o varianza explicada por la correlación entre las variables involucradas.

Retomando, cuando la varianza explicada queda explicada por una manipulación experimental, se llama Varianza Experimental o entre los grupos; la que no quede explicada se denomina Varianza de Error o intragrupo. Lo que se pretende explicar es la Varianza Total; en ambos casos se trata de la manipulación experimental y el enfoque es multivariado.

Cuando las variables que influyen sobre la varianza total son independientes entre sí, o sea que no correlacionan entre ellas, se puede hablar de sus efectos sobre la varianza total por separado. Cuando hay relación entre ellas, se debe entonces hablar de un efecto de interacción sobre la variabilidad de la variable, atributo o característica que se pretende explicar.

1.3 Análisis de varianza

La aplicación más simple del análisis de varianza es probar la diferencia de los datos en medias entre dos grupos seleccionados al azar. Para esto también sirve la prueba "t" de Student (pero ya no se puede emplear la prueba "t" si los grupos son más de dos). En estas circunstancias se puede emplear la Prueba de Rangos de Duncan[2].

Independientemente de lo anterior, se presenta como introducción a los diseños multivariados, al análisis de varianza simple o de una variable, ya que el procedimiento estadístico que conlleva,

[1] El coeficiente de determinación es el cuadrado del coeficiente de correlación de Pearson y es la proporción de la varianza de la VD que es explicada por la varianza de la VI.

[2] La prueba de rangos de Duncan es una prueba post hoc y la validez de su aplicación está condicionada a la obtención de un resultado significativo en la prueba de análisis de varianza, que sería el procedimiento adecuado a este caso. Nota de la editora.

es el fundamento de los diseños factoriales. Los primeros diseños realmente multivariados se verán más adelante.

Por señalar, las calificaciones de la variable dependiente que resultaron de un diseño de dos grupos son los que se ilustran gráficamente en la Figura 3. La curva de la izquierda representa las calificaciones de los sujetos del grupo 1 y la distribución de frecuencias a la derecha es para el grupo 2.

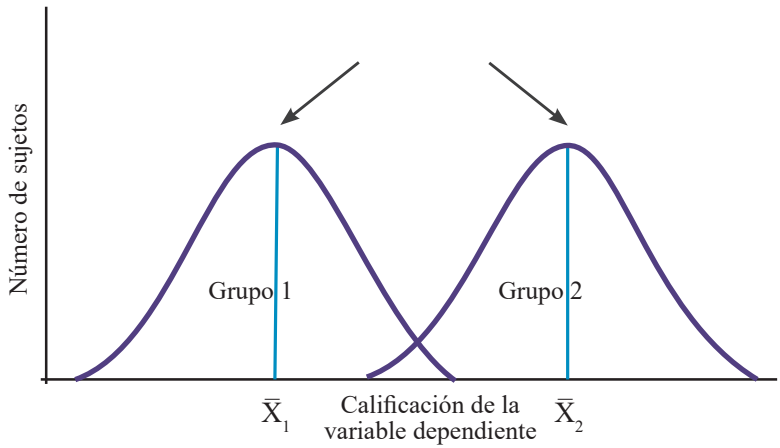


Figura 3. Naturaleza de la suma de cuadrados (SC) intra y entre, empleando solo dos grupos.

Para responder a la pregunta de si estos grupos son significativamente diferentes, se emplea el análisis de varianza. Para este es necesario determinar la suma de cuadrados (SC) total. Esta SC es el valor que resulta cuando consideran a todos los sujetos del estudio como un todo. Este se calcula con las calificaciones de la variable dependiente de todos los sujetos, ignorando el hecho de que unos están en un grupo y los otros en otro. Lo importante es señalar que la SC puede ser analizada en partes. En particular, la SC total se puede dividir en dos componentes principales: la SC inter-grupos y la SC intra-grupos. A grosso modo, la SC inter-grupos se puede considerar determinada por el grado en que las medias de dos grupos difieren. Entre más grande sea la diferencia entre las medias de los grupos, mayor es la SC inter-grupos.

Por su parte, la SC intra-grupos está determinada por el grado en que difieren los sujetos en cada grupo. Si los sujetos del grupo 1 difieren ampliamente entre sí, y/o si esto es verdad para los sujetos del grupo 2, la SC intra-grupos va a ser grande. Y entre mayor sea la SC dentro de los grupos, mayor será el "error" en el estudio o experimento. A manera de ilustración, suponga que todos los sujetos del grupo 1 han sido tratados precisamente de la misma manera; si fueron exactamente iguales cuando comenzaron el experimento, todos debían recibir la misma calificación de la variable dependiente. Si esto sucediera, la SC intra-grupos -hasta donde concierne al grupo 1-, sería cero, ya que no habría variación entre sus calificaciones. Por supuesto que la SC intra-grupos casi nunca es cero, puesto que todos los sujetos son diferentes antes del experimento y el experimentador nunca puede tratarlos exactamente de la misma manera. De lo anterior se puede deducir que cuanto mayor sea la suma de cuadrados inter-grupos y menor la suma de cuadrados intra-grupos, más posibilidades hay de que los grupos sean significativamente diferentes. Hasta aquí se ha hecho referencia al caso de un diseño con dos grupos.

El mismo razonamiento general se aplica cuando hay más de dos grupos. La SC total del estudio se analiza en dos partes: la SC intra y la Ínter o entre grupos. Si la diferencia entre varias medias es pequeña la SC entre grupos seria pequeña y si las varianzas del grupo individual son altas, las sumas de los cuadrados dentro de los grupos serán elevadas. Cuanto mayores sean las SC entre o inter-grupos y menores la SC dentro o intra-grupos, existen más posibilidades de que los grupos difieran significativamente.

A continuación se trata el cálculo de las diversas sumas de cuadrados (SC). Las ecuaciones que se emplean están basadas en el siguiente razonamiento y su cálculo automáticamente explica lo que se va a indicar a continuación.

Primero, una media (μ) se calcula basándose en todos los valores de la variable dependiente del estudio o experimento (ignorando el hecho de que los sujetos pertenecen a diferentes grupos o condiciones). Entonces, las sumas totales de los cuadrados (SCtotal), miden la desviación de todas las calificaciones de esta media general. La suma de cuadrados entre-grupos es una medida de la desviación de las medias de los diversos grupos partiendo de la media general. Y la suma de cuadrados intra-grupos es una suma combinada de cuadrados basada en la desviación de las calificaciones de cada grupo a partir de la media de dicho grupo. El propósito principal es calcular la SC total y luego analizarla, por partes.

La ecuación generalizada para calcular la SC total, es la siguiente:

$$\text{SC Total} = (\Sigma x_1^2, \Sigma x_2^2 + \cdots \Sigma x_r^2) - \frac{(\Sigma x_1, \Sigma x_2 + \cdots \Sigma x_r)^2}{N}$$

Esta indica que se continúan sumando los valores indicados (la suma de los cuadrados de las X, y la suma de las X respectivamente) para tantos grupos como se tengan.

El siguiente paso es analizar el SC en sus componentes. Los componentes principales son: la SC entre-grupos y la SC intra-grupos. Una ecuación general para calcular la SC entre-grupos, es:

$$\text{SC entre} = \frac{(\Sigma x_1)^2}{N_1} + \frac{(\Sigma x_2)^2}{N_2} + \cdots \frac{(\Sigma x_r)^2}{N_r} - \frac{(\Sigma x_1, \Sigma x_2 + \cdots \Sigma x_r)^2}{N}$$

El componente de SC intra-grupos se calcula de la siguiente manera:

$$\text{SC intra} = \text{SC total} - \text{SC entre}$$

El lector se puede preguntar ¿Dónde están las varianzas que se están analizando en el análisis de varianza? En este caso se consideran bajo un nombre diferente y se da referencia a ellos en los valores de las muestras, no como varianzas, sino como medias de los cuadrados. Es decir, se calculan los valores de las muestras como estimaciones de los valores de la población. Las medias de los cuadrados (valores de la muestra), son los cálculos de las varianzas (los valores de la población). Lo que indica que la media del cuadrado intra-grupos es una estimación de la varianza intra-grupos. La regla para calcular las medias de los cuadrados es dividir una suma dada de cuadrados entre la cantidad de grados de libertad adecuados. Las fórmulas para los grados de libertad son las siguientes:

$$\text{gl total} = N\text{-}1 \; ; \quad \text{gl entre} = r\text{-}1 \quad \text{gl intra} = N\text{-}r$$

Donde:

N: Es el número total de sujetos independientemente del grupo al que pertenezcan.

r: Es el número de grupos.

A continuación se requiere calcular para el análisis de varianza de una variable o una entrada, dos medias de cuadrados: una media de cuadrados de la fuente de variación entre-grupos y una media de cuadrados de intra-grupos. El cálculo de las medias de cuadrados consiste entonces en dividir la SC entre grupos sobre los gl entre-grupos; la SC intra-grupos sobre los gl intra-grupos.

Ahora bien, como se ha indicado anteriormente, si la media de los cuadrados entre-grupos es considerablemente mayor en relación a la media de cuadrados intra-grupos, se puede concluir que los valores de la variable dependiente para los grupos son distintos.

Sin embargo, ¿Hasta dónde es considerablemente mayor lo "considerablemente mayor"? Es decir ¿De qué magnitud debe ser el componente entre grupos a fin de que se pueda concluir que una variable independiente dada es efectiva? Para responder a esto se deberá aplicar una prueba estadística: la proporción F desarrollada por Fisher[3]. El proceso estadístico para este diseño puede definirse como sigue:

$$F = \frac{\text{media de los cuadrados entre-grupos}}{\text{media de los cuadrados intra-grupos}}$$

El numerador indica las diferencias entre-grupos (además del error experimental) y el denominador indica el error experimental, o como se le llama también la varianza de error del experimento. Más particularmente, en las aplicaciones más simples de la prueba F, el numerador contiene una estimación de la varianza de error más una estimación del efecto "real" (si es que lo hubo) de la variable independiente. El denominador es solamente una estimación de la varianza del error. Cuando se divide el numerador entre el denominador, el valor computado de F refleja el efecto de la variable independiente para producir una diferencia entre las dos medias. Supóngase que la variable independiente es totalmente inefectiva para influir sobre la variable dependiente. En este caso se esperaría que el numerador no tuviera ninguna contribución de la variable independiente (no habría ninguna media de los cuadrados

[3] Fórmula creada por el británico R. Fisher (1890-1962), La distribución F de Fisher es una distribución que depende de dos parámetros. Es una distribución que aparece, con frecuencia, como distribución de un estadístico de test, en muchos contrastes de hipótesis bajo las suposiciones de normalidad. Fuente: https:// estadisticaorquesta instrumento.wordpress.com/ 2013/01/07/la-distribucion-f-de-fisher/

entre-grupos que fuera “real"). Por lo tanto, el valor del numerador solo sería una estimación de la varianza; una estimación similar del error de la varianza está en el denominador. Así, sí se divide un valor de la varianza de error entre un valor de la varianza de error, se obtendrá una F en las vecindades de 1.00.

Por esto, cuando se obtiene una F de aproximadamente uno, se puede asegurar que la variación de la variable independiente no produjo ninguna diferencia en las medias de la variable dependiente de los grupos. Sin embargo, los numeradores pueden ser un poco mayores que el denominador, pues la media del cuadrado entre-grupos puede ser un poco mayor que la media del cuadrado intra-grupos. ¿De qué magnitud debe ser el numerador de la razón F, antes de concluir que las medias de los grupos son realmente distintas? Esto es, si el numerador es grande en relación con el denominador, F también es grande: ¿Qué tan grande debe ser F para poder rechazar la hipótesis nula?

Para responder a esta pregunta se tiene que determinar el valor de p asociado a la F obtenida. Para ello, se consulta la tabla de F (por ejemplo, la que aparece en McGuigan, 1964). En la hilera superior aparecen los valores de gl asociados al numerador y en la primera columna, aquellos gl asociados al denominador. En el cuerpo de la tabla aparecen los valores de F a diferentes p: .01, .05, .10 y .20. Si la F obtenida por el investigador es igual o mayor que la establecida por la tabla con los gl del numerador y denominador correspondientes con una p igual o menor a 0.05, se rechaza la hipótesis de nulidad, que establece la no existencia de diferencias entre los grupos estudiados. Sin embargo, se debe señalar que si la F es significativa, se sabe que existe alguna diferencia significativa entre los grupos, pero no se sabe dónde reside esa diferencia. Si los grupos eran dos, se lleva a cabo una prueba t[4]; si eran más de dos, se recomienda la prueba de Rangos de Duncan. De esta manera se sabrá dónde se encuentra la diferencia.

[4] Aunque en ese caso proporciona los mismos valores de significancia que una prueba F.

1.4 Diseño Factorial

El diseño anterior es apropiado para la investigación de una sola variable independiente que puede variar en dos o más niveles: si la variable independiente varía en dos formas o niveles, se emplea un diseño de dos grupos; si varía en tres o más niveles, se utiliza un diseño multigrupos.

Es posible estudiar más de una variable independiente en un solo experimento. El diseño que permite estudiar dos o más variables independientes, es el Diseño Factorial. Un diseño factorial completo es aquel donde se emplean todas las combinaciones posibles de los valores seleccionados de cada variable independiente

En un diseño en el que existan dos variables independientes, cada una de ellas con dos niveles o valores, existen cuatro combinaciones posibles de los valores de las variables independientes. Cada combinación posible se representa por un cuadrado o celdilla.

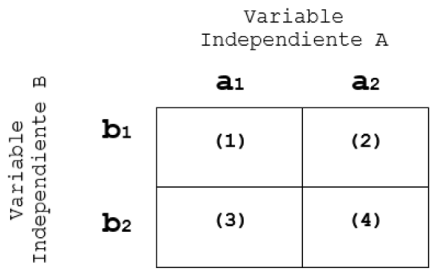


Figura 4. Diagrama de un Diseño Factorial a x b = 2 x 2 = 4.

En la Figura 4, se producen cuatro condiciones experimentales, o sea que hay cuatro grupos en el estudio. Se recomienda que el número de sujetos en cada grupo sea igual o semejante a los demás grupos.

Un análisis de las calificaciones de la variable dependiente proporciona información concerniente a las siguientes preguntas:

a) ¿Influye la variable independiente A sobre la variable dependiente?

b) ¿Influye la variable independiente B sobre la variable dependiente?

c) ¿Existe una interacción entre las variables independientes A y B?

El procedimiento para contestar a las dos primeras preguntas es directo, el tercero requiere de más consideraciones.

Para responder a la primera pregunta, se debe estudiar el efecto de la variable A (a1 y a2) sobre las calificaciones de la variable dependiente. Para esto se debe ignorar la clasificación respecto a la otra variable independiente (B: b1 y b2). Por lo tanto, se tienen dos grupos de sujetos a los cuales considerándolos como un todo, se trataron de manera similar excepto en lo que respecta a la variable A. Es decir, se tiene que hacer una comparación entre las medias de estos dos grupos; se deberá aplicar una prueba estadística para saber si la diferencia entre las medias es significativa.

Para responder a la segunda pregunta se procede de manera semejante. Se consideran dos grupos (b1 y b2) pero en esta ocasión independientemente de su clasificación en la variable A (a1 y a2). En este caso también habrán de compararse estas dos medias por medio de la prueba estadística adecuada. En cualquiera de los dos casos anteriores, cuando se considera una de las variables independientes, se puede ver el diseño factorial como si estuviera llevando a cabo un solo experimento y variando únicamente una de las variables independientes; en este caso, la otra variable independiente puede considerarse temporalmente como variable extraña cuyos efectos han sido balanceados.

Ahora bien, respecto a la tercera pregunta, antes de responderla, se explica el concepto de interacción[5]. Existe una interacción entre dos variables independientes si el valor de la variable dependiente que resulta de una variable independiente está determinada por el valor específico asumido por la otra variable independiente. La interacción también se puede plantear como sigue: Si la variable A afecta a la Variable Dependiente, subordinándose de los valores de B en los sujetos. A continuación se ilustra en forma gráfica el concepto de interacción y la falta de ella (Figura 5).

[5] Interacción se refiere a la acción que se lleva a cabo de forma recíproca entre dos o más variables, fuerzas o funciones (Diccionario https://definicion.de/interaccion/.)

La línea que une los grupos 2 y 4 representa la ejecución de los sujetos en la condición b1 de la variable independiente B; la línea que une a los grupos 1 y 3, representa la ejecución de los sujetos en la condición b2 de B.

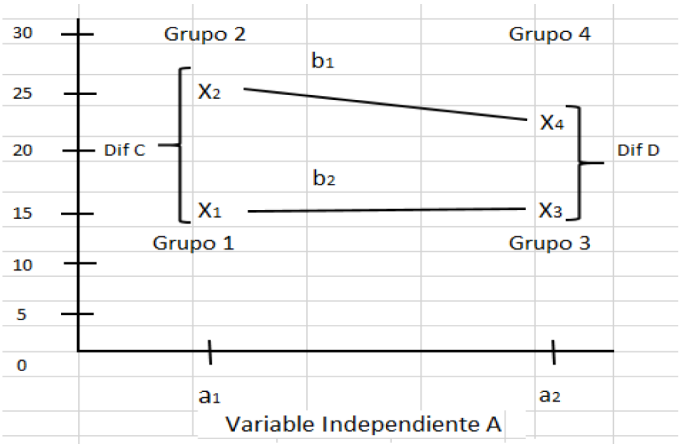


Figura 5. Concepto de interacción y su falta.

Si estos datos fueran reales ¿Cuáles serían los efectos de las variables independientes? Primero, se diría que la variación en A no afecta a la variable dependiente, porque ambas líneas son esencialmente horizontales; segundo, los sujetos de la condición b1 tienen puntajes más altos en la variable dependiente, que los de la condición b2 (la línea b1 está más alta que la línea b2). Y tercero, la diferencia entre los grupos 1 y 2 (diferencia C) es casi la misma que la diferencia entre los grupos 3 y 4 (diferencia D). Por lo tanto los puntajes en la variable dependiente de los sujetos en b2 y en b1 es esencialmente independiente de los niveles de A (a1 y a2). No existe

[6] Por supuesto se está haciendo referencia a los valores de la muestra y no a los de la población. De este modo, mientras esta afirmación es cierta para los valores de la muestra, no lo es para los valores reales de la población. Por tanto, sí las líneas de los valores de la población se encuentran aunque sea ligeramente no paralelas, existe una interacción.

interacción entre estas dos variables. Dicho de otro modo, si las líneas dibujadas en la Figura 5 son aproximadamente paralelas, es decir, si la diferencia C es aproximadamente la misma que la diferencia D, es probable que no exista interacción entre estas dos variables[6]. Sin embargo, si las líneas basadas en estas medias de las muestras no son paralelas (es decir, si la diferencia C es notoriamente distinta de la diferencia D), se presenta una interacción (Figura 6).

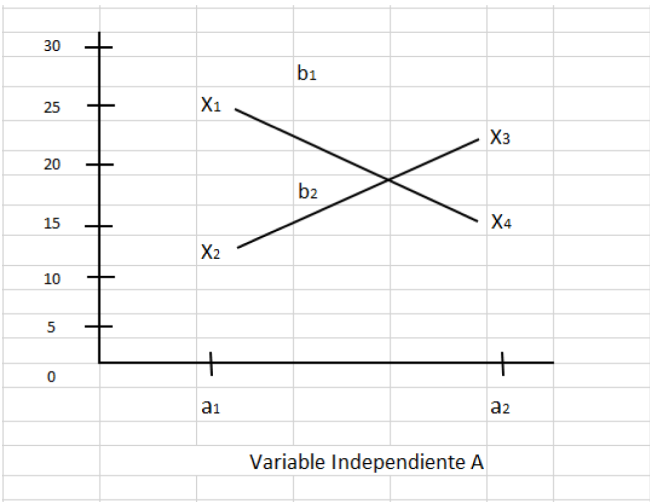


Figura 6. Existencia de Interacción.

En la Figura 6 se observa que las líneas no son paralelas, sino que se cruzan entre sí. De ahí que si estos datos fueran reales, por lo que se harían las siguientes afirmaciones: Los sujetos a b son superiores a los sujetos a1 y b2, pero los sujetos a2 y b2, son superiores a los sujetos a2 y b1. La afirmación lógica equivalente es: El efecto de b2 disminuye la ejecución de los sujetos (en la variable dependiente) a1, pero facilita la ejecución de los sujetos a2. Dicho de otra manera, la diferencia entre b1 y b2, depende de la A de los sujetos; la diferencia del grado de A depende de sí los sujetos están en b1 o b2.

1.4.1 Tipos de Diseños Factoriales

1.4.1.1. Diseños factoriales con dos variables independientes

1.4.1.1.1 Diseño factorial 2 x 2. En este diseño, se estudia el efecto de dos variables independientes, variando cada una de ellas de dos maneras. La cantidad de números empleados en la designación indica el número de variables independientes estudiadas en la investigación. Y el tamaño de los números indica la cantidad de valores de las variables independientes. Dado que el diseño 2 x 2 tiene dos números, se puede decir que hay dos variables independientes, y puesto que ambos números (son 2) se sabe que cada variable independiente asume dos valores. Del 2 x 2 también se puede decir cuántas condiciones experimentales hay, ya que el 2 multiplicado por 2 son 4.

1.4.1.1.2 Diseño factorial 2 x 3. En este diseño se estudian dos variables independientes, una que varía en tres maneras y la segunda en dos. Esquemáticamente quedaría representado en la Figura 7.

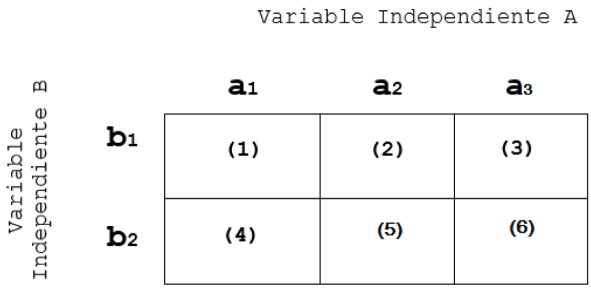


Figura 7. Diseño 2 x 3.

1.4.1.1.3 Diseño factorial 3 x 3. En este diseño se estudian dos variables independientes, ambos variando en tres maneras. Por lo tanto se asignan sujetos a nueve (3 x 3 = 9) condiciones experimentales. Esquemáticamente queda representado en la Figura 8.

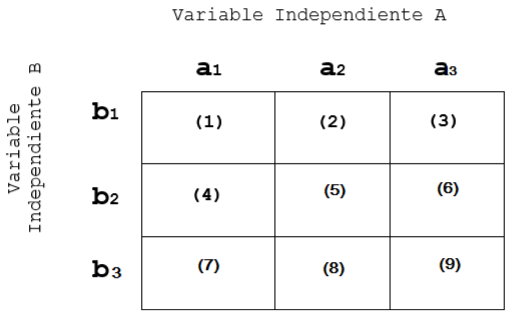


Figura 8. Diseño 3 x 3

1.4.1.1.4 Diseño factorial K x L. Cada variable independiente puede variarse en cualquier número de maneras. El diseño factorial generalizado para dos variables independientes suele llamarse diseño factorial K x L, donde K representa la primera variable independiente, y su valor indica el número de maneras en la cual es variada; en forma semejante, L denota la segunda variable dependiente. K y L pueden entonces asumir cualquier valor.

1.4.1.2 Diseños factoriales con más de dos variables independientes

1.4.1.2.1 Diseño factorial 2 x 2 x 2. Este es el diseño factorial más simple para estudiar tres variables independientes, cada una de ellas variando de dos maneras, que se muestra en la Figura 9.

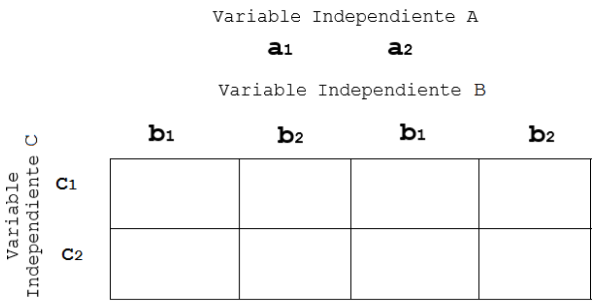


Figura 9. Diseño 2 x 2 x 2.

1.4.1.2.2 Diseño factorial K x L x M. Es evidente que cualquier variable independiente puede variarse en cualquier número de maneras. El caso general para el diseño factorial de tres variables independientes, el de K x L x M, donde K, L y M, pueden asumir cualquier valor entero positivo que el experimentador desea. Un diseño de 5 x 3 x 3, queda representado en la Figura 10.

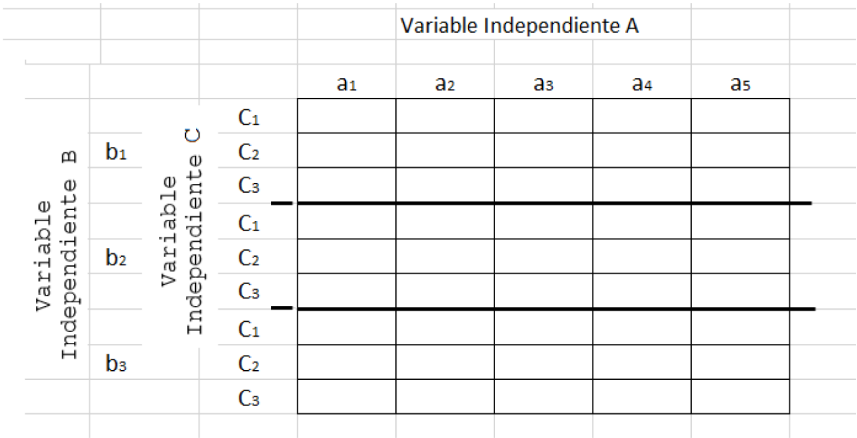


Figura 10. Diseño 5 x 3 x 3 = 45.

1.4.2 Análisis estadístico de los diseños factoriales

El análisis estadístico que se aplica más frecuentemente al diseño factorial, es el Análisis de Varianza, cuyos rudimentos se presentaron en la sección correspondiente de una variable o una entrada. Se limitará la explicación que se presenta a continuación, al diseño factorial 2 x 2. Sin embargo el procedimiento estadístico se aplica a cualquier diseño factorial; de hecho, los paquetes estadísticos de computadora han vuelto obsoleto el cálculo manual. De cualquier manera se presenta a continuación con la idea de aclarar el procedimiento, su aplicación e interpretación.

El primer paso para llevar a cabo un análisis de varianza para un diseño factorial se aproxima mucho al empleado anteriormente. Es decir, se desea calcular la suma de cuadrados (SC) total y dividirla en sus dos componentes principales: SC entre y SC intra. De aquí que los pasos y las fórmulas sean las mismas que se describieron antes. De esta manera, en este primer paso se tienen valores para:

SC Total.

SC entre-grupos.

SC intra-grupos.

La SC entre-grupos dice algo acerca de cómo difieren todos los grupos. Sin embargo, no está el interés en la comparación simultánea de los cuatro grupos, sino sólo en ciertas comparaciones. En un diseño factorial 2 x 2, se tienen dos variables independientes, cada una de ellas variada de dos maneras. En este caso interesa saber si la variación de cada variable independiente afecta o no a la variable dependiente; y si existe o no una interacción significativa entre ambas variables independientes. El primer paso, entonces, es calcular una SC entre-grupos para cada variable independiente:

SC entre la primera variable independiente (A) =

$$\frac{(\Sigma x_{1+}\ \Sigma x_3)^2}{N_1 \quad N_3} + \frac{(\Sigma x_{2+}\ \Sigma x_4)^2}{N_2 \quad N_4} - \frac{(\Sigma x_1, \Sigma x_2 + \Sigma x_3 + \Sigma x_4)^2}{N}$$

SC entre la segunda variable dependiente (B) =

$$\frac{(\Sigma x_{1+}\ \Sigma x_2)^2}{N_1 \quad N_2} + \frac{(\Sigma x_{3+}\ \Sigma x_4)^2}{N_3 \quad N_4} - \frac{(\Sigma x_1, \Sigma x_2 + \Sigma x_3 + \Sigma x_4)^2}{N}$$

La SC entre-grupos general, tiene tres partes: dos SC entre cada variable independiente y la SC de interacción. Por lo tanto la SC interacción es igual a SC entre (general) sobre SC entre para la primera variable (A) y SC entre para la segunda variable independiente (B).

Hasta ahora, se tienen los siguientes datos: Suma de cuadrados para un diseño factorial 2 x 2 y Fuente de variación, los datos se describen de la siguiente manera:

Suma de Cuadrados entre Grupos	_____
Entre Primera Variable Independiente (A)	_____
Entre Segunda Variable Independiente (B)	_____
Interacción: A x B	_____
Intra-grupos	_____
Total	_____

Los grados de libertad (gl) que habrán de emplearse para encontrar las medias de las sumas de cuadrados son los mismos de antes para la SC total, SC intra y SC entre (general). Para la SC entre la primera variable, o para la de la segunda variable independiente, la fórmula es la siguiente:

$$\text{gl entre (Primera o Segunda) variable independiente} = r - 1$$

Donde r = niveles de cada una de las variables Independientes.

Por otro lado, los gl Interacción son sencillos de calcular pues gl interacción = gl entre la primera variable (A) x gl entre la segunda variable (B).

En un diseño factorial 2 x 2, existen cuatro medias de Sumas de Cuadrados en las cuales se tiene interés acerca de las condiciones entre la primera variable Independiente (A); las condiciones entre la segunda variable independiente (B); interacción e intra-grupos. Después de haberlas calculado, se resumen los resultados en una Tabla Sumaria como la siguiente:

Tabla sumaria del análisis de varianza del diseño 2 x 2.

Fuente de Variación	Suma de Cuadrados	Gl	Media de Suma de cuadrados	F
Entre Primera Variable Independiente (A)	_____	_____	_____	_____
Entre Segunda Variable Independiente (B)	_____	_____	_____	_____
Interacción: A x B	_____	_____	_____	_____
Intra-grupos (error)	_____	_____	_____	_____
Total	_____	_____	_____	_____

Se tienen varias fuentes de sumas de cuadrados (entre) para estudiar, así como un término que representa el error experimental (la media de los cuadrados intra- grupos). Los componentes entre (Primera o Segunda variable independiente) indican la amplitud con la cual difieren las diversas condiciones experimentales: si cualquier componente dado entre la Primera o Segunda, es considerablemente grande, éste puede tomarse para indicar que la primera o segunda variable (según el caso) influye en la variable dependiente. Solo se requiere llevar a cabo las diferentes pruebas F apropiadas, para determinar si los diversos componentes de entre (la Primera o Segunda), son significativamente mayores de lo que podría esperarse por azar.

La primera F que se tiene que computar es la de entre la primera variable independiente (A). Para hacerlo simplemente se substituyen los valores correspondientes en la ecuación de F vista anteriormente. Posteriormente se calcula la F para la segunda variable y por último, la F para la interacción.

Se debe aclarar que en este caso, el investigador estaría interesado específicamente en el posible efecto de las variables independientes sobre la dependiente y en la interacción entre ellas, por lo cual, no se calcula una F para la fuente de variación entre grupos (general).

La tabla sumaria es la que se presenta en el estudio, faltaría agregar las probabilidades asociadas a cada F obtenida, con objeto de determinar las posibilidades de que las F pudiesen haber ocurrido por azar. En este tipo de diseño, las hipótesis nulas que se plantean son las siguientes:

1) No hay diferencia entre las medias de las dos condiciones de la primera variable independiente (A).

2) No hay diferencia entre las medias de las dos condiciones de la segunda variable independiente (B).

3) No hay interacción entre las dos variables independientes.

1.4.3 Elección de un término correcto de error

Uno de los problemas más importantes en el análisis estadístico es el de la elección del término correcto de error. En relación con la prueba F el problema es el de elegir denominador correcto. El término de error que se ha empleado hasta ahora es el de la media de la Suma de Cuadrados intra-grupos; generalmente este término es el correcto. Sin embargo no siempre es así, ya que depende del modelo de diseño factorial que se esté trabajando.

Los modelos del diseño factorial pueden ser tres: a) Fijo; b) Aleatorio; y c) Mixto. A continuación se describen cada uno de ellos.

a) *Modelo fijo.* Cuando los valores o niveles de cada una de las (dos) variables independientes fueron seleccionados de manera premeditada y/o arbitraria por el investigador, el modelo es fijo. En este caso la media del cuadrado intra- grupos es el término de error correcto para todas las pruebas F que se realicen. Los resultados que se encuentren, sólo serán generalizables para los valores y/o niveles estudiados en esa investigación.

b) *Modelo aleatorio.* Cuando los valores y/o niveles de las (dos) variables independientes fueron seleccionados al azar de entre todos los posibles valores y/o niveles que las variables pudieran tener.

El procedimiento es el siguiente. Se prueba la media de los cuadrados de interacción dividiéndola entre la media de los cuadrados intra. Después se prueban las medias de los cuadrados entre (variables primera y segunda) grupos, dividiéndola entre la media de los cuadrados de interacción, Los resultados que con este modelo se encuentren, serían generalizables a todos

los posibles valores y/o niveles de las variables independientes estudiados. Sin embargo, este tipo de diseños son relativamente raros en la investigación psicológica.

c) *Modelo mixto.* Este modelo se presenta cuando una variable independiente es fija, y la otra es aleatoria. En este caso, supongamos que la variable A es fija; el término de error apropiado para probar la fuente de variación de interacción es la media del cuadrado intra. La media del cuadrado de la variable aleatoria (B) se coloca en el numerador de la razón F, mientras que en el denominador de la razón F se emplea la media de los cuadrados intra-grupo. Pero para la media de la variable fija (A), se emplea la media del cuadrado de interacción en el denominador de la razón F.

De la misma manera en el modelo mixto donde B es fija, la media del cuadrado de interacción se divide entre la media del cuadrado intra; la variable aleatoria (A) se prueba empleando la media de los cuadrados intra-grupos como el término de error, pero la variable fija (B) se prueba empleando la media del cuadrado de interacción como término de error.

Tema II Regresión simple y múltiple

2.1 Introducción

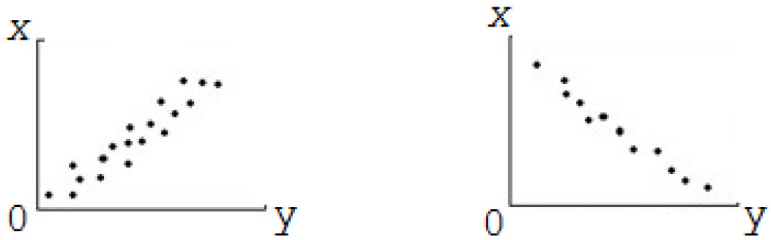
Para comprender de manera adecuada lo que es la Regresión Simple o Múltiple, se debe aclarar lo que se entiende por relación, correlación y función.

Una relación es un conjunto de pares ordenados de acuerdo con cierta regla de correspondencia. Cada uno de los miembros del par provienen de un conjunto previamente definido, y la regla de correspondencia solo señala la forma que habrán de ordenarse (formarse) esos pares. Por ejemplo, se dice que el matrimonio es una relación entre dos conjuntos de elementos: uno constituido por hombres y otro por mujeres. La regla de correspondencia en este caso sería el poseer un aval social que indique que hay un contrato matrimonial entre los miembros del par. Cuando se mide a nivel nominal, por ejemplo, cuando se asigna el número 1 al sexo masculino y un número 2 al femenino, y se clasifica a una serie de nombres de acuerdo al sexo que tienen, se está teniendo una relación también. La relación queda formada por la asignación de un 1 o un 2 a cada uno de los nombres de la lista. La regla de correspondencia señala cuándo habrá de asignarse un 1 y cuando un 2. También se forman relaciones entre un conjunto de números y un conjunto de personas cuando estos son medidos en cualquiera de los niveles de medición restantes (ordinal, intervalar y de razón). En cada nivel, la regla de correspondencia es el modelo estadístico de medición que se emplea para la asignación de números a los sujetos.

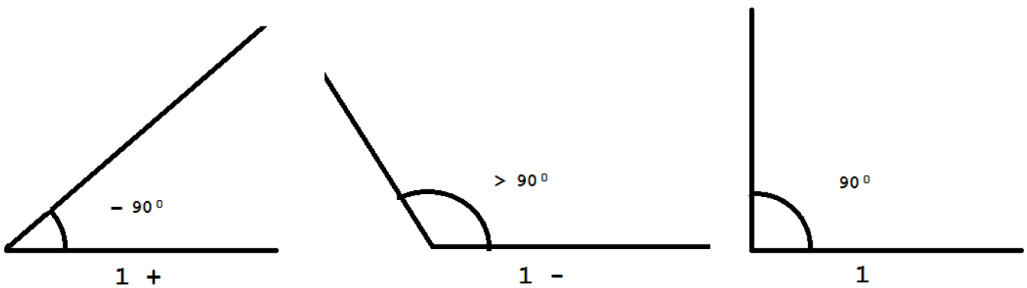
Una correlación, desde el punto de vista estadístico, representan la - cantidad de variación conjunta o covariación entre dos conjuntos, de manera que la correlación es el grado de asociación entre dos –o más- variables. No solo señala la existencia de una relación entre los elementos de los dos conjuntos, sino también la magnitud de la regla de correspondencia (covariación conjunta) que establece la relación. La magnitud de las correlaciones puede ir de **-1** a **+1** pasando por **0**.

Cuando una correlación tiene un valor cercano a -1, se dice que los dos conjuntos de elementos covarían en forma inversa. Es decir, que mientras los valores de los elementos de un conjunto se incrementan, los valores de los elementos de otro conjunto disminuyen. Cuando la correlación tiene un valor cercano a +1, la covariación entre los elementos de los conjuntos es directa. Es decir, en la medida en que incrementan los valores de los elementos miembros de un conjunto en un par, también incrementan los valores de los elementos del otro conjunto miembros del mismo par. Por último, cuando el valor de la correlación es cercano a cero, se dice que no hay correlación solo si su valor es cero. Es decir, la forma en que se comportan los elementos de un conjunto no tiene nada que ver con la forma en que se comportan los elementos del otro conjunto.

Los valores de correlación antes señalados se pueden representar gráfica^ mente por medio de dispersigramas, de la siguiente manera:



Otra manera de representar en forma gráfica los diferentes valores de la correlación es por medio de los cosenos de los ángulos que forman un sistema de coordenadas:



Así, una función señala qué tanto tiene que variar **X** para que se dé un cambio en **Y**. Una función es en realidad una ecuación, como por ejemplo, la de la línea recta:

$$Y = a + bx$$

Esta ecuación es una de las funciones más simples que existen, y de acuerdo al criterio de la parsimonia en ciencia[7], es también una de las más útiles, pues de ella se deriva el *modelo lineal*. Este modelo es el que subyace a todos los diseños multivariados que se ven en el presente libro.

2.2 Modelo Lineal

Cualquier modelo lineal de datos establece que un valor observado de la variable dependiente es igual a una suma pesada de valores asociados con una o más variables independientes, más un término de error.

Supóngase que se desea averiguar los diferentes efectos o influencias que se dan sobre una variable dependiente **Y**. Cuando se tiene un grupo de observadores de la variable **Y** se requiere que se limiten o mantengan constantes algunas cosas, las cuales suponemos determinan un nivel general que los valores u observaciones de **Y** puedan exhibir. El modelo lineal necesita reflejar esto, de tal manera que se incluye un valor constante **Xo**, que será el que representa a este nivel general, que quedó establecido al definir la muestra de observaciones en la forma en que haya sido definida. Este valor recibe un peso[8] **Ao** en cualquier observación.

Existen además, otros aspectos o cuestiones que producen las diferencias observadas en **Y**. Algunas de estas pueden estar bajo nuestro control y otras no. Todos los factores conocidos o las variables que afectan a **Y** para hacer que se den las diferencias observadas en **Y** pueden ser

[7] Con el *Principio de simplicidad: entre dos o más descripciones, explicaciones o hipótesis lógica y empíricamente posibles de un dominio de fenómenos, se elige la que explica lo que debe ser explicado con el menor número de conjeturas, es decir la descripción, explicación o hipótesis más simple*. Fuente: De Luna E. (1996). *Epistemología de la investigación taxonómica: inferencias filogenéticas y su evaluación*. Boletín de la Sociedad Botánica de México, 58, 43-53.

[8] Se utiliza *peso* como sinónimo de carga.

concebidas como los aspectos sistemáticos del estudio. Algunos de estos pueden ser controlados u observados en forma sistemática y son las XI a Xj.

Por último, existen aquellos aspectos o cuestiones que también afectan la variabilidad de **Y**, pero que no se pueden controlar o medir, o quizá ni siquiera se pueden explicar en forma adecuada. Estos factores, peculiares a un individuo particular en una situación o tiempo específico, se identifican en el modelo, como *error*.

Sin necesidad de identificar la naturaleza exacta de los aspectos constantes, sistemáticos y de error que determinan o influyen sobre él, así como las observaciones que se tengan de ésta, el modelo general lineal se representa de la siguiente manera:

$$y = ao\ xo + a_1 + Xi_1 + a_2\ Xi2 .... + aj\ xij + ei$$

Donde:

ao = El peso que el nivel general conlleva en la determinación de una observación individual de **Y**.

xo = nivel general del grupo de observaciones de **Y**.

Xij = valor de alguna variable.

a_1 = el peso o efecto de esa variable en su determinación de **Y**.

Ei= componente del valor **Y** que tiene todos los factores o aspectos que no se controlan o ni se conocen.

Es general en el sentido de que no se ha especificado la naturaleza precisa de las variables X, ni cuantas son, ni qué pesos tienen. Más bien, estas especificaciones del modelo pueden hacerse a la medida para encajar a una situación experimental o de investigación dada. Esta concepción es enormemente flexible en el sentido de que se presta a todo tipo de situaciones de colección de datos.

Los diseños presentados en el tema anterior, en éste y en los siguientes, presuponen que un modelo lineal de esta naturaleza subyace a los datos. Sin embargo, hay muchas variaciones que si se pueden aplicar a este modelo lineal general para ajustarse al tipo particular de situación o experimento que esté siendo modelado. Dos de estas variaciones tienen que ver con el hecho de si los valores de **X** en el modelo son simples "indicadores" o "marcadores" que representan a - tratamientos o grupos cualitativamente diferentes, o si cualquier **X** representa un valor numérico de la magnitud de alguna variable.

Cuando los valores **X** en un modelo lineal son indicadores o marcadores, al modelo se le llama *Modelo de Diseño Experimental* o *Modelo de Análisis de Varianza*.

Cuando las variables **X** en el modelo lineal pueden adoptar cualquier valor numérico, se le llama *Modelo de Regresión Múltiple*.

Las suposiciones del modelo lineal en su versión de modelo de análisis de varianza, efectos fijos, son las siguientes:

1. Se supone que la distribución del error (eij) es normal para cada población de tratamiento (celdilla).

2. Para cada población, la distribución del error tiene una varianza (σ^2) que se supone que es la misma para cada población de tratamiento (celdilla).

3. Los errores asociados a cualquier par de observaciones se supone que son independientes. Una consecuencia de esta suposición es que si **h** e **i** fueron un par de observaciones; **J** y **K** un par de tratamientos, entonces:

$$E(ei\ je\ hj) = 0 \quad y \quad E(ei\ je\ hk) = 0$$

En resumen, las observaciones se extraen independientemente de poblaciones de tratamiento (celdillas) normales, teniendo cada una de ellas la misma varianza, y con componentes de error independientes a través de todos los pares de observaciones.

2.2.1 Modelo de efectos fijos

La importancia de las suposiciones en el modelo de efectos - fijos. Las suposiciones anteriores proporcionan la justificación teórica para el análisis de varianza y la prueba **F** empleados tanto en los diseños factoriales de varianza (Tema I) como en la regresión simple o múltiple (Tema II). Por otro lado, en ocasiones es necesario analizar los datos cuando estas suposiciones no se cumplen claramente; en realidad, rara vez es razonable pensar que son exactamente ciertas. Se mencionarán brevemente las consecuencias de la aplicación del análisis y la prueba **F** cuando no se cumplen esas suposiciones.

Primera, cuando no se cumple la suposición de la distribución normal de los errores, se puede demostrar que, manteniendo lo demás constante, las inferencias que se hacen acerca de las medias que son válidas en el caso de las poblaciones normales, también son válidas aun cuando las formas de la distribución de las poblaciones se alejen considerablemente de la normalidad, siempre y cuando la **N** de cada muestra (celdilla) sea lo suficientemente grande.

La segunda hablaba de la homocedasticidad (varianzas iguales) de las varianzas de error en las poblaciones tratamiento (celdillas). Por lo general, y en igualdad de circunstancias, se puede violar la suposición de varianzas homogéneas sin un riesgo serio, siempre y cuando el número de casos de cada muestra (celdilla) sea el mismo. Por otro lado, cuando hay **n's** diferentes en las celdillas, la violación de varianzas homogéneas puede tener consecuencias serias en lo que se refiere a la validez de la inferencia.

La tercera, requiere independencia estadística entre los componentes de error. Si no se cumple esta suposición se pueden cometer serios errores de inferencia. Esta suposición deberá cumplirse al asegurarse el investigador de que cada observación no se relaciona con las demás observaciones, es decir, se supone independencia entre las variables de tratamiento.

2.2.2 Modelo de efectos aleatorios: Una variable

El modelo lineal de efectos aleatorios que representa la constitución de cualquier valor **Yig** en una población g, sería:

$$\mathbf{Yig} = \mu + \mathbf{aj} + \mathbf{eij}$$

Donde:

μ = Media general o de la población.

aj = El efecto de estar en la población j.

eij = el componente de error de Yij.

Las suposiciones del modelo son las siguientes:

1. Los valores posibles **aj** representan una variable aleatoria que tienen una distribución con una media de cero y una varianza σ^2A.
2. Para cualquier tratamiento (celdilla) **j**, los errores **eij** se distribuyen normalmente con una media de cero y una varianza σ^2, que es la misma para todos los posibles tratamientos **j**.
3. Los valores de la variable aleatoria **aj** que ocurren en un experimento son todos independientes unos de otros.
4. Los valores de la variable aleatoria **eij** son todos independientes.
5. Cada par de variables aleatorias **aj** y **eij** son independientes.

De acuerdo con la importancia de las suposiciones del modelo respecto a la validez de las inferencias que se llevan a cabo a partir de los resultados, se tiene, en primer lugar, que la suposición de normalidad es bastante importante en este modelo, tanto para los efectos como para el error. También en lo que se refiere a independencia del error respecto a los efectos y entre los errores mismos.

2.2.3 Modelo de efectos aleatorios: Dos variables

Las suposiciones del modelo lineal de efectos aleatorios, para dos variables (K x L), derivan de la siguiente representación matemática: la calificación del individuo **i** en la variable **Y**, en la columna **j** y la hilera **k** de una tabla es la suma de:

$$Yijk = \mu + aj + bk + cjk + eijk$$

Donde:

μ = La gran media, media general o nivel general.

aj = El efecto de la variable aleatoria en la columna j.

bk = El efecto de la variable aleatoria de la hilera k.

cjk = El efecto del error en la observación del individuo i en la situación jk.

eijk = El efecto del error interactuando en la observación del individuo i en la situación jk.

Las suposiciones son:

1. Las aj son variables aleatorias distribuidas normalmente, con una media de cero y varianzas homogéneas (σ^2_A).
2. Las bk son variables aleatorias con media cero y varianzas homogéneas (σ^2_B).
3. Cada cik tiene una distribución normal con media cero y varianza homogénea (σ^2_{AB}).
4. Los eijk se distribuyen normalmente con media de cero y varianza homogénea (σ^2_e).
5. Las aj, bk, cjk y eijk son independientes entre todos sus posibles pares.

2.2.4 Modelo de efectos mixtos

Las suposiciones del modelo lineal mixto derivan de la siguiente representación matemática:

Permítase que el factor que tiene niveles fijos se denomine **A**, y quede representado por las columnas (j) de la tabla, y que el factor muestreado al azar sea **B** y se represente en las hileras de la tabla (K). Ahora se asume que:

$$Yijk = \mu + aj + bk + cjk + eijk$$

Donde:

μ = La media general.

aj = El efecto fijo del tratamiento indicado por la columna j.

bk = Es la variable aleatoria asociada con la hilera k.

cijk = Es el efecto de interacción aleatoria que opera en la celdilla jk.

eijk = Es el error aleatorio asociado con la observación i en la celdilla jk.

Las suposiciones son las siguientes:

1. La bk y la cijk son conjuntamente normales, cada una con una media de cero y una varianza σ^2_B y σ^2_{AB} respectivamente.
2. Los eijk se distribuyen normalmente, con media cero y varianza σ^2_e.
3. Los eijk son independientes de bk y cjk.
4. Todos los términos de error eijk son independientes unos de otros.

2.3 Regresión simple

Una regresión es una función. El modelo que se usa para describir la relación entre **X** y **Y** será el lineal; en el caso de la regresión simple, el modelo es sencillo:

$$Yi = Ao + bxi. + ei$$

Donde:

Ao = Constante que se integra al valor de Yi para cualquier individuo i.

b = Un peso constante que se aplica al valor de Xi.

ei = Es el error.

Si no hubiera error, entonces: **Yi= ao + bxi**, que es estrictamente lineal en el sentido de que un dispersigrama de todos los pares (Xi, Yi) de valores caerán en una línea recta.

Obviamente, no existe ninguna ley que diga que la relación entre los valores de Xi y Yi tenga que ser lineal. Puede ser que la mejor descripción de la forma en que Xi y Yi están relacionadas en un conjunto de datos, demande o exijan una función matemática diferente.

Entonces, ¿Por qué se pone énfasis en los modelos lineales? Las razones para iniciar con reglas lineales para la predicción son varias. *En primer lugar*, las funciones lineales son las más sencillas de entender y discutir. *En segundo lugar*, estas son con frecuencia, buenas aproximaciones de otras más complicadas. *En tercer lugar*, se encuentra que en ciertas circunstancias, la única regla de predicción que se puede aplicar es la lineal.

Las suposiciones generales de la regresión, son las siguientes:

1. La muestra es aleatoria.

2. Cada conjunto de Y para una dada combinación de X's se distribuye normalmente.

3. La regresión entre X y Y es lineal.

4. Todos los conjuntos de Y tienen la misma varianza.

5. Las variables independientes (X^2) son independientes entre sí.

6. Los valores de Y están normalmente distribuidos y son estadísticamente independientes.

Estas suposiciones se refieren y se requieren, cuando se emplea la regresión como estadística inferencial y/o se desea hallar las relaciones lineales en las poblaciones. Sin embargo, cuando se desea emplear como estadística descriptiva, asociada a una muestra dada, no es necesario hacer ninguna suposición acerca de la forma de la distribución, la variabilidad de los puntajes X dentro de las columnas o "arreglos" o el nivel real de medición representado en las calificaciones para poder emplear la regresión lineal para describir a un conjunto dado de datos. En la medida en que existan N casos diferentes, cada uno teniendo dos (o más) valores numéricos (X y Y; o X1, X2 y Xn) se puede usar la estadística descriptiva de la regresión lineal.

La regresión simple es la que se da entre dos variables: una dependiente (Y) y otra independiente (X). La primera se denomina *variable criterio* y la segunda, *variable predictora*. Una regresión múltiple es la que tiene una variable dependiente o criterio, y muchas variables independientes o predictoras.

La regresión, tanto la simple como la múltiple, se pueden emplear en términos descriptivos y en términos inferenciales.

Los primeros permiten descomponer o resumir la dependencia lineal de una variable en otra(s). Entre los usos descriptivos más comunes se tienen los siguientes:

a) Encontrar la mejor ecuación lineal predictiva y evaluar su exactitud de predicción.

b) Controlar algunas variables que puedan confundir (extrañas, antecedentes, interventoras, supresoras o distorsionadoras) y evaluar la contribución de una variable especifica o de un conjunto de variables específicas.

c) Encontrar relaciones estructurales y proporcionar explicaciones para relaciones multivariadas aparentemente complejas, por medio del análisis de flujos (*path analysis*[9]).

[9] Path analysis (Análisis de Senderos o P.A.) *es un método que permite evaluar el ajuste de modelos teóricos en los que se proponen un conjunto de relaciones de dependencia entre variables* (Pérez et al, 2013: 52). También auxilia a inferir entre hipótesis causales (Batista-Foguet & Coenders-Gallart, 2000).

Cuando se use en términos inferenciales, en general, se evalúan las relaciones en la población a partir del examen de los datos de la muestra; es decir, se puede generalizar a la población.

Entre los usos inferenciales más comunes se tiene que:

a) Hacer estimaciones de la población: encontrar los parámetros de la población más probables a partir de las observaciones de la muestra.

b) Poner a prueba hipótesis como la no linealidad entre una variable dependiente y un conjunto de independientes; o el que una variable independiente particular no tiene un efecto lineal sobre la dependiente una vez que se ajustan los efectos de las otras independientes; o que los efectos de dos o más variables no son aditivos.

En la Regresión Simple, los valores de la variable dependiente se predicen a partir de una función lineal del tipo:

$$\mathbf{Y'=A+BX}$$

Donde:

Y'= Valor externado de la variable dependiente Y.

B = Constante por la cual se multiplican los valores de la variable independiente X.

A = Constante que se suma a cada caso.

La diferencia entre el valor real de **Y** y su estimado para cada caso se llama residual y corresponde al error de predicción. El residual se puede presentar como **Y-Y'**.

La estrategia de la regresión involucra la selección de valores tales de **A** y **B** que la suma de los residuales al cuadrado sea menor que cualesquiera otros valores alternativos. Es decir:

$$\Sigma\ (\mathbf{Y-Y'})^2 = \text{SS res} = \text{mínimo}$$

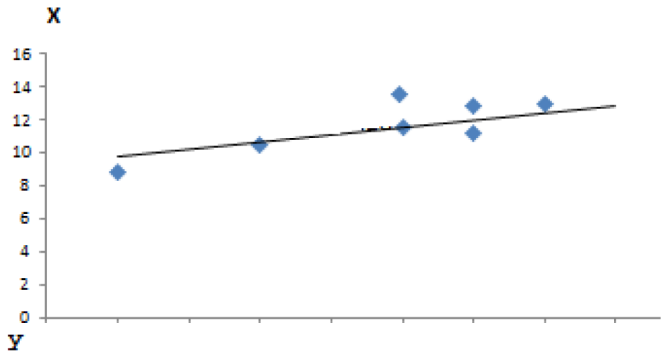
Los valores óptimos de **A** y **B** se obtienen de la siguiente manera:

$$B = \frac{\Sigma (X - \bar{X})_{(Y-Y')}}{\Sigma (X - \bar{X})^2} = SP\ xy$$

$$A = Y - BX$$

Los coeficientes **A** y **B** son asimétricos ya que **X** se toma como predictor de **Y**. Los valores son diferentes cuando **Y** se usa como predictor de **X**.

Los valores de A y B se muestran en la siguiente gráfica:



A = Intercepción en Y: el punto en el que la línea de regresión cruza al eje **Y** y representa el valor predicho de y cuando X = 0.

B = Coeficiente de regresión no estandarizado, pendiente de la línea de regresión y señala el cambio esperado en **Y** con un cambio de una unidad de **X**.

Los valores Y' predichos caen a lo largo de la línea de regresión y las distancias verticales Y-Y' de los puntos a la línea corresponden a los residuales (o errores de predicción).

Como la suma de los residuales elevada al cuadrado se minimiza, la línea de regresión se llama *Línea de Cuadrados Mínimos* o *Línea de Mejor Ajuste*.

Partición de la Suma de Cuadrados (Varianza).

Partición de la Suma de Cuadrados (Varianza). La suma de cuadrados total (SC) en **Y** (varianza de Y) se divide en componentes que: (1) Son explicados por la línea de regresión (*SS*reg) y (2) No explicados (*SS*res). Como los mínimos cuadrados garantiza que los residuales sean independientes del predictor Y', se puede decir:

$$SSy = SS\ reg + SSres$$

Por lo tanto, la prueba de exactitud de la predicción quedaría dada por:

$$Y^2_{xy} = \frac{SS\ reg}{SSy} = \frac{SSy - SSres}{SSy}$$

Donde:

Y^{r2} = proporción de varianza explicada.

r^2 = varianza no explicada.

La r^2 se llama *Coeficiente de Determinación* y la raíz cuadrada de éste, concierne a la *Correlación Producto Momento de Pearson* entre **X** y **Y**.

2.4 Regresión múltiple

La regresión múltiple es una ampliación de la regresión simple, en el sentido de que el número de variables independientes o predictoras, son más de una. La ecuación correspondiente a su modelo es:

$$Y_i = A + B, X_1 + B_2 X_2 + ...B_k X_k$$

Donde:

Yi = Valor estimado de **Y**.

A = Constante, intercepto en **Y**.

B = Coeficiente de regresión.

Entre las estadísticas importantes asociadas a la regresión, se tienen las siguientes:

a) Coeficientes estandarizados de regresión (Beta). Estos simplifican la ecuación de la regresión lineal, por que establecen la igualdad de **A** como cero. Son útiles cuando hay dos o más variables medidas con unidades diferentes (ejemplo: ingreso en pesos[10], y educación en años de escolaridad formal), se pueden comparar los efectos relativos de cada variable independiente sobre Y.

b) Coeficientes no estandarizados de regresión (B). Su ecuación es:

$$B = \left(\frac{Sy}{Sx}\right)$$

Donde:

B= Beta.

S= Desviación estándar.

Estos sirven para predecir Y' a partir de los puntajes crudos.

c) Error Estándar de Estimación (SEE). Este, es el error promedio que se comete al predecir Y' a partir de la ecuación de regresión, y su fórmula es:

$$SEE = \sqrt{\frac{SS\ reg}{N - 2}}$$

[10] Peso, moneda oficial en México.

d) Error Estándar de B. Su fórmula es la siguiente:

$$\text{Error Estándar de B} = \sqrt{\frac{\text{SS reg } (N^2)}{\text{SS x}}}$$

e) Prueba de significancia para B (regresión simple).

$$F = \frac{\text{SS reg}}{\text{SS reg / N–2}} \qquad \text{con gl} = \frac{1}{\text{N–2}}$$

f) Prueba de significancia para las F asociadas a la Regresión Múltiple, se tienen que buscar en una tabla de F con los siguientes grados de libertad.

1. Para la F general de la Regresión:

$$\text{gl} = \frac{\text{K}}{\text{N–K–1}}$$

Donde:

K = Número de variables.

N = Número de sujetos.

2. Para las F de los coeficientes de regresión:

$$\text{gl} = \frac{\text{K}}{\text{N–K–1}}$$

g) R. Múltiple: equivale en regresión múltiple, a una correlación simple entre Y y Y'; Y' se puede ver como una variable independiente construida a partir de la ecuación de regresión.

h) Análisis de Residuales: Éste se puede llevar a cabo por medio de la prueba de Durbin-Watson[11]; desgraciadamente se carece de una prueba de significancia de la misma. Se puede emplear en lugar de la Durbin-Watson, una prueba de $\mathbf{x}^2$ para una sola muestra con seis categorías, cada una de ellas correspondiendo a los puntajes de -3 a +3. Se busca rechazar la hipótesis nula que establece la no diferencia estadísticamente significativa en la distribución de frecuencia en las 6 categorías mencionadas. El objetivo del análisis de residuales es establecer la medida en que la relación entre la variable criterio o dependiente y la(s) independiente(s) o predictora(s) fue realmente lineal. La linealidad de la relación queda puesta de manifiesto al obtenerse distribuciones de frecuencia estadísticamente diferentes entre las categorías señaladas, encontrándose las frecuencias más altas en las categorías centrales (-2 a +2), de preferencia en aquellas de -1 y +1).

Uno de los problemas con los que se puede enfrentar el investigador al emplear el análisis de regresión múltiple, es el de la multicolinealidad. Es decir, la existencia de correlaciones muy altas entre algunas o todas las variables independientes o predictoras. Entre los problemas que produce la multicolinealidad se tiene que los coeficientes no se pueden determinar en forma única; que la estimación de los coeficientes de una muestra a otra fluctúa mucho; se producen efectos que confunden los resultados.

Entre las soluciones que se recomiendan cuando se presente este problema, se tienen:

a) Crear una nueva variable que sea una escala compuesta de las variables altamente interrelacionadas entre sí y usar la nueva variable compuesta en la ecuación de regresión.

b) Usar en la ecuación de regresión a sólo una de las variables del conjunto altamente correlacionado.

[11] El test Durbin-Watson es una herramienta estadística que permite detectar si los residuales de una regresión están autocorrelacionados. Fuente: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/model-assumptions/test-for-autocorrelation-by-using-the-durbin-watson-statistic/

2.5 Sugerencias generales

Se recomienda que se asegure por medio de una inspección a la matriz de intercorrelaciones que se habrá de solicitar cuando se lleve a cabo una regresión, para detectar la posible existencia de multicolinealidad entre las variables que se incluyen en el análisis. Si se observa este fenómeno, entonces será conveniente decidir sobre la realización de un análisis factorial que produzca la escala compuesta por el conjunto de variables altamente correlacionadas para incluirla en forma de calificación factorial en el análisis de regresión o decidirse sobre la otra alternativa: incluir sólo una de las variables altamente correlacionadas entre sí.

Se recomienda también que la **N** que se emplee sea grande (de 100 y más casos); que si se emplean reactivos individuales como variables independientes o predictoras, éstos cuenten con por lo menos tres opciones de respuesta, graduadas o pesadas de tal manera que la opción de respuesta que represente mayor cantidad del atributo o variable en cuestión, tenga el peso o valor más alto.

Es importante tomar en cuenta el pesaje que se da a las opciones de respuesta o la forma en que se califica la escala, instrumento, prueba o cuestionario, para simplificar la interpretación de los resultados. Debe tomarse en cuenta que los Coeficientes de Regresión (B) pueden adoptar valores positivos o negativos y por esta razón, se debe ayudar en la interpretación de los mismos por medio de la simplificación de los procedimientos de medición o registro de las variables.

Una tercera recomendación se refiere a solicitar junto con el análisis de regresión propiamente dicho, otras estadísticas que ayuden a entender e interpretar los resultados. Entre las estadísticas que es pertinente solicitar están la media y desviación estándar de las variables que se incluyen en el análisis, una matriz de intercorrelaciones entre todas las variables incluidas; y un análisis de residuales donde aparezcan las diferencias Y-Y' en forma gráfica, junto con la prueba de Durbin-Watson.

El análisis de regresión se puede solicitar en el programa de análisis estadístico (SPSS) en dos formas: D1RECT y STEPWISE. La primera proporciona información general sobre la significatividad de la regresión, así como los coeficientes de regresión (B) pero no presenta la prueba F asociada a cada coeficiente. La forma STEPWISE, efectúa el análisis paso a paso, incluyendo en cada paso aquella variable que más varianza explica de la variable dependiente.

Es decir, este procedimiento, ingresa al análisis en el primer paso, aquella variable o predictor que mayor varianza explica la variable criterio o variable dependiente. En segundo lugar se incluye a la siguiente variable o predictor que más varianza explica, y así sucesivamente. Antes de incluir a la variable en un paso, se calcula su F para ver si ésta pasa el nivel de tolerancia establecido por el programa. Si lo pasa, se incluye; si no lo pasa, no se incluye. Cuando entre las variables o predictores aún no incluidas en el análisis ya no se encuentran variables con niveles de tolerancia adecuados para ser incluidos, se detiene el proceso y aparece una leyenda señalando que los predictores restantes carecen de un nivel adecuado para ser incluidos. Se recomienda preferentemente emplear el procedimiento paso a paso. Algunas de las razones son:

a) Este procedimiento proporciona información sobre la significatividad de la regresión en su totalidad (solo variables incluidas).

b) Proporciona una prueba de significancia para cada coeficiente de regresión (B) por separado.

c) Proporciona una estimación del error estándar para cada B.

d) Detiene el análisis cuando las variables restantes ya no cubren el mínimo de tolerancia establecido por el programa para ser incluida.

2.6 Interpretación de resultados

La interpretación de los resultados se llevará a cabo después de los análisis estadísticos realizados a los datos sometidos. Se pueden obtener las medias de los puntajes crudos para cada una de las variables estudiadas, junto con su desviación estándar y el número de sujetos que contestaron a cada variable o pregunta. Posteriormente se obtiene una matriz de intercorrelaciones entre las variables sometidas a análisis. Esta matriz puede usarse para que el investigador se ayude en la interpretación de los resultados de la regresión, que sirve también para detectar la existencia de multicolinealidad entre las variables. Se toma en cuenta el tamaño de ***F** de la Regresión*, cuyo número se consulta en una tabla de **F**, con grados de libertad en el numerador y número en el denominador (este dato aparece en la columna denominada DF). Ejemplo: Una F (10.251) = 5.24628, tiene una probabilidad (p.) asociada menor a 0.01. Esto significa que la regresión

es estadísticamente significativa a ese nivel (0.01). A continuación se observará qué tanto correlacionan **Y** y **Y'**, la segunda derivada de la ecuación de regresión obtenida. Es decir, cuál es el valor de *R. Múltiple*. Después la *R. Cuadrada* que proporciona información de la proporción de varianza de la variable dependiente -en este caso que cubrió el programa- y que queda explicada por las variables incluidas en el análisis. El error estándar, respecto a la $\overline{Y}$. La Tabla de regresión (Análisis de Varianza: Regresión, Residual; DF; Suma de Cuadrados y Medias de Cuadrados) son los elementos necesarios para conformar el valor de **F** general, que es la primera que se analiza. Se ven ahora cuáles son los valores que obtuvieron las variables que entraron a formar parte de la ecuación de regresión.

2.7 Reporte de un análisis de regresión

Un reporte de un análisis de regresión deberá establecer las características de la muestra estudiada, las características de los instrumentos empleados (formato, aplicación, calificación, variables que pretendió medir, confiabilidad y validez); el tamaño de la muestra (N). Se pueden presentar, aunque no es necesario, dos tablas, una de medias y desviaciones de los puntajes crudos de las variables analizadas, y otra con las intercorrelaciones entre ellas.

Es indispensable presentar una Tabla sumario del análisis de regresión que contenga los siguientes elementos: Valores de R múltiple. R cuadrada; R cuadrada ajustada; error estándar. Los resultados del análisis de varianza con sus grados de libertad, suma de cuadrados y media de cuadrados asociados a la regresión y al residual. El valor F obtenida y su **p** asociada.

Las variables incluidas en la ecuación, son sus valores **B** y la constante; sus valores Beta los valores del error estándar de B. sus **F's** asociados y sus correspondientes. En caso de que alguna variable o varias no quedarán incluidas, será suficiente con solo mencionarlo.

Síntesis

1. Como alternativa del análisis de varianza [no causa efecto].

2. Más flexible que el diseño factorial (análisis factorial de varianza = experimento causa-efecto rígido).

3. El análisis de regresión es un método de correlación con distribución normal.

4. Cuadro: Análisis de varianza, gl, Suma de cuadrados (SC) Raíz cuadrada de la suma de cuadrados F p.

R Múltiple, R al cuadrado, R al cuadrado estándar, Error estándar, Variables, B, Beta, Error estándar de B, F y p.

Referencias

Batista-Foguet, J. M. B. & Gallart, G. C. (2000). *Modelos de Ecuaciones Estructurales*. Madrid: La Muralla, S.A.

De Luna E. (1996). Epistemología de la investigación taxonómica: inferencias filogenéticas y su evaluación. *Boletín de la Sociedad Botánica de México, 58*, 43-53.

Miles, J. & Shevlin, M.(2001). *Applying Regression Correlation*. Sage Publications, London.

Pérez, E., Medrano, L.A. & Sánchez-Rosas, J. (2013). El Path Analysis: conceptos básicos y ejemplos de aplicación. *Revista Argentina de Ciencias del Comportamiento, 5(1)*, 52-66.

Tema III Análisis factorial

3.1 Introducción

El análisis factorial se distingue de los demás diseños multivariados por su capacidad de reducir a un número menor un número mayor o muy grande de datos o variables. En este caso las variables pueden conceptualizarse como todas independientes o todas dependientes. Este diseño puede responder a la interrogante ¿Es posible reducir a este conjunto de variables o datos a un conjunto menor de ellos pero que sin embargo explique la mayor parte de la variabilidad o varianza que manifiesta o demuestra el conjunto mayor?

El análisis de los datos parte de una matriz de intercorrelaciones en la que se establecen las correlaciones de todas y cada una de las variables contra todas y cada una de las mismas. Esto da como resultado una matriz cuadrada en cuya diagonal se encuentran 1.00 que corresponden al coeficiente de correlación que se obtendría al correlacionar esa variable consigo misma. Por encima y por debajo de la diagonal se encuentran los coeficientes de correlación de cada variable de cada columna contra todas las variables de todas las hileras. Se dice que esta matriz es simétrica puesto que los coeficientes de correlación de las celdillas correspondientes por encima y por debajo de la diagonal, son iguales.

El investigador supone que es posible que exista algún patrón subyacente a las intercorrelaciones de tal manera que los datos se pueden re-arreglar o reducir a un conjunto menor de factores o componentes. Estos pueden entonces considerarse como variables originadoras o variables fuente, que explican o dan cuenta de las intercorrelaciones observadas entre los datos o variables.

Entre las aplicaciones más usuales del análisis factorial se tienen el uso exploratorio y el uso confirmatorio. El primero, el exploratorio, se refiere a la exploración o detección de patrones de variables con vista al descubrimiento de nuevos conceptos y/o a responder a la interrogante de la posibilidad de reducir los datos. El segundo se refiere a poner a prueba estructuras de variables en términos del número de factores, del significado de los mismos y de sus cargas factoriales.

Un tercer uso se refiere a emplear, al análisis factorial en la construcción de instrumentos de medición o índices que se puedan usar como nuevas variables en análisis posteriores.

3.2 Tipos de análisis factorial

El análisis factorial no es un concepto unitario, sino que depende de la. Alternativas que se pueden tomar en cada uno de los tres pasos principales que se siguen: a) preparación de la matriz de intercorrelaciones; b) la extracción de los factores iniciales (lo que habla de la posibilidad de reducción de los datos) c) rotación a una solución terminal (que produzca factores simples e interpretables).

A continuación se presentará cada uno de los tres pasos en forma más de tallada.

3.2.1. Preparación de la matriz de intercorrelaciones

Los pasos que se siguen para la preparación de la matriz de correlaciones son los siguientes:

1. Definir el universo de contenido relevante de variables. Si el estudio es exploratorio, se deberá decidir cuál conjunto de variables se va a explorar en el sentido de averiguar la posible existencia de una estructura factorial subyacente; si el estudio es confirmatorio, las hipótesis planteadas señalan o delimitan al conjunto inicial de variables que serán sometidas a análisis Por último, si el procedimiento se va a emplear para la construcción de un instrumento se procede a definir conceptualmente la variable; se determinarán las dimensiones constitutivas; se derivan los indicadores pertinentes y de éstos se elaboran los reactivos que constituirán el instrumento que se pretende elaborar.

2. Este conjunto de variables se aplican a una muestra representativa de la población de interés del investigador. La muestra debe ser de un tamaño suficientemente grande, de 100 o más sujetos. Una vez recogida la información, se someten a análisis que produzcan medidas de la asociación existente entre el conjunto de variables.

3. Con los coeficientes de correlación producto momento de Pearson se elabora la matriz de intercorrelaciones. Debe de recordarse que antes de someter a análisis de correlación, se debe de haber establecido para todas las variables, la normalidad de las mismas. Esto se puede determinar sometiendo a los datos a una FRECUENCIAS en el que se pidan todas las estadísticas. Entre las estadísticas que se obtienen con este programa se encuentran la ***Kurtosis*** y el ***Skewnes***. Estos valores deben de estar cercanos a cero, lo que indica que la variable se distribuye en forma normal. En caso de que los valores de Kurtosis o de Skewnes sean extremos (en relación a los valores del conjunto total de las variables en cuestión) se deben descartar del análisis, pues una de las suposiciones del modelo lineal, que en caso de no cumplirse produce resultados espurios[12], es el de la distribución normal de las variables.

Si las correlaciones que se calculan son entre variables o atributos, el análisis se conoce como del **tipo R**. Por otro lado, si las correlaciones se calculan para individuos (entre sujetos) se le llama **tipo Q**. Por último si estos se calcularon entre unidades, como por ejemplo objetos o comunidades se conoce como **tipo S**.

3.2.2. Extracción de factores iniciales

La extracción inicial de factores, explora la posibilidad de reducción de los datos. Es decir, determina si se puede encontrar un conjunto de nuevas variables en base a las intercorrelaciones observadas.

Las nuevas variables se pueden definir de dos maneras: a) como transformaciones matemáticas exactas a partir de los datos originales: y a este procedimiento se le llama análisis de componentes principales; b) como suposiciones inferenciales acerca de la estructuración de las variables y su fuente de variación; a este procedimiento se le conoce como análisis factorial clásico o solución de factores comunes.

Los factores iniciales se extraen de tal manera que cada factor es independiente de los demás. Se debe señalar que la solución factorial inicial, no cumple con los requisitos fundamentales del análisis factorial que son el de la estructura simple y el desarrollo positivo.

[12] Relación o correlación espuria, en estadística, situación en donde dos o más variables de medidas están estadísticamente relacionadas pero no tienen relación de causalidad entre ella. (Definiciones, 2018)

El requisito de estructura simple se refiere al hecho de que cada variable cargue en un factor alto, y cerca de cero en los demás. El desarrollo positivo se refiere al hecho de que las cargas factoriales sean o tengan signo positivo.

El hecho de que estos requisitos no se cumplan, hace que los factores iniciales que se obtienen sean muy difíciles de interpretar, ya que muchas variables aparecen con cargas más o menos altas en varios factores al mismo tiempo por un lado; y por el otro, aparecen una gran cantidad de cargas factoriales asociadas a signos negativos.

De hecho, esta extracción inicial solo permite responder a la interrogante de la posibilidad de reducir un conjunto de datos a un número menor de variables o factores que expliquen la mayor cantidad de varianza posible, originalmente contenida en la matriz de intercorrelaciones. En virtud de lo anteriormente expuesto y con objeto de poder interpretar los resultados obtenidos, así como poder cumplir en la medida de lo posible con los criterios de estructura simple y desarrollo positivo, se procede a pasar a la tercera etapa del análisis; la rotación. Sin embargo, antes de pasar a esta etapa, veremos los tipos de factores por los que se puede optar en el proceso de extracción inicial.

3.2.2.1 Factores definidos

El método de los componentes principales transforma a un conjunto de variables en un nuevo conjunto de variables compuestas o componentes principales que son independientes entre sí, o sea, ortogonales.

Cuando se opta por este tipo de solución, no se requiere hacer ninguna suposición acerca de la estructura subyacente de las variables. El investigador tan solo se pregunta; ¿Cuál es la mejor combinación lineal de variable que explica la mayor cantidad de varianza en los datos como conjunto, que cualquier otra? De esta manera, el primer componente principal es el mejor resumen de la combinación lineal exhibida en los datos; el segundo componente es el segundo mejor que sigue, obtenido del residual después de haber extraído el primero y así sucesivamente, hasta agotar la cantidad de varianza total existente.

El segundo componente es independiente (ortogonal) del primero; explica la varianza no explicada por el primero, y así con los componentes sucesivos. El modelo se puede expresar como sigue:

$$Zj = aj_1 F_1 + aj_2 F_2 + ...+ aj_n F_n$$

Dónde cada una de las **n** observadas se describen linealmente en términos de los componentes nuevos no correlacionados, F_1 F_2 F_n. De esta manera, los primeros **m** componentes, mucho menos que las n variables originales, explican la mayor parte de la varianza de los datos.

3.2.2.2 Factores inferidos

El análisis factorial clásico supone que las correlaciones observadas resultan de una regularidad subyacente en los datos. Se supone que la variabilidad observada está influida por diferentes determinantes, algunos compartidos con otras variables del conjunto y otros no compartidas con ninguna otra. A los determinantes compartidos se les llama comunes y a los idiosincráticos se les llama *Factores Únicos.*

La parte única de una variable no contribuye a la relación entre las variables; las correlaciones son resultado de los factores comunes; éstos explican todas las relaciones observadas y son menores en número que las variables ori¬ginales. Este modelo se puede expresar como sigue:

$$Zj = aj_1 F_1 = aj_2 F_2 + ...+ aj_m + d_j u_j$$

Dónde:

Zj = Variable j en forma estandarizada

F_1 = Factores hipotéticos

u_j = Factor único para la variable j

aj_1 = Coeficiente estandarizado de regresión múltiple de la variable j en el factor i (carga factorial),

d_j = Coeficiente de regresión estandarizado de la variable j en el factor único j.

Este modelo tiene además las siguientes suposiciones:

1. La correlación entre los factores comunes y los únicos es igual a cero:

$$r(Fj, ui) = 0$$

2. La correlación entre los factores únicos es igual a cero:

$$r(uj, uk) = 0$$

3. El factor único **u** es independiente (ortogonal) de todos los factores comunes y de los factores únicos asociados a otras variables, por lo tanto, si hay correlación entre dos variables, ésta se debe a los factores comunes.

4. Al complemento de la varianza única (**u**) se le llama *Comunalidad* (**h^2**). La comunalidad se estima a partir de los datos:

$$1 - u = h^2$$

El que se usen factores definidos o inferidos depende de que se suponga la existencia de varianza única. Otra consideración se refiere a si el estudio que se lleva a cabo es exploratorio o confirmatorio. En caso de ser exploratorio se recomienda que el análisis sea de componentes principales; si el estudio fuera confirmatorio, se recomienda emplear el modelo de factores comunes.

3.2.3 Rotación a Factores Terminales

La configuración exacta de la estructura factorial no es única. Una solución factorial puede transformarse en otra sin violar las suposiciones básicas. Existen muchas formas estadísticas equivalentes de definir las dimensiones subyacentes del mismo conjunto de datos.

Algunas soluciones son más parsimoniosas y simples; otras más informativas. Cada una dice algo ligeramente diferente acerca de la estructura de los datos. Cada investigador escoge la que más le conviene de acuerdo con sus fines teóricos y/o prácticos. Las opciones más importantes son: la ortogonal, donde se supone independencia (no correlación) entre los factores extraídos; y la oblicua, que supone correlación entre los factores.

3.3 Procedimiento general

El análisis factorial completo proporciona las siguientes seis matrices:

a) Una matriz de correlaciones de las variables analizadas.

b) Cargas factoriales iniciales.

c) Pesos para estimar las variables a partir de los factores (factor pattern matrix o matriz del patrón factorial).

d) Pesos para estimar factores a partir de las variables a) factor estimate (estimación de factores); o b) factor score coefficient matrix (matriz de coeficientes de calificaciones factoriales).

e) Correlación entre los factores y las variables o cargas factoriales (factor structure matrix (matriz de la estructura factorial).

f) Matriz de intercorrelaciones de los factores terminales.

Las variables pueden introducirse por medio de los puntajes crudos que los sujetos obtuvieron en cada una de ellas, o por medio de una matriz cuadrada (igual número de columnas (**k**) e hileras (**r**)).

Someter a análisis factorial a un conjunto de variables significa en el nivel más general, expresar a una variable como la confinación lineal de ciertas variables (factores) independientes, ya sea definidos o inferidos.

La matriz del patrón factorial (factor pattern matrix) contiene los pesos o coeficientes de regresión de los factores comunes y por lo tanto señala la composición de una variable en términos de los factores hipotéticos.

La estimación de factores (factor estimate) o matriz de coeficientes de calificaciones factoriales (factor score coefficient matrix) proporciona un medio de estimar puntajes factoriales a partir de variables observadas. Es decir, son los pesos o coeficientes de regresión que se emplean para estimar las calificaciones factoriales a partir de las variables observadas expresadas en unidades o puntajes **Z**.

La matriz de la estructura factorial rotada (rotated factor structure matrix) está constituida por los coeficientes de correlación (o cargas factoriales) entre cada variable y cada factor. Esta es la matriz que se emplea para interpretar (nombrar) a los factores extraídos u obtenidos.

Cuando se emplea un método ortogonal de rotación, la matriz de patrón factorial (*factor pattem matrix*) y la matriz de la estructura factorial (*factor structure matrix*) son iguales, por lo que aparece únicamente la segunda.

La matriz de Intercorrelaciones entre los factores terminados solo se obtiene cuando la solución terminal tuvo una rotación oblicua. Está constituida por las intercorrelaciones de las dimensiones (o factores) subyacentes, y puede servir para análisis factoriales de un orden más alto (someter a análisis factorial los puntajes que los sujetos obtienen en los factores obtenidos en el primer nivel o primer paso).

La importancia de los factores está dada primero, por el valor **Eigen** que obtiene cada uno de los factores extraídos. El valor Eigen es la raíz de la ecuación (polinomio) que explica la matriz reducida de varianza (la que se trabaja a partir de la extracción de los factores iniciales). Un segundo aspecto es el que se habla de la importancia de los factores es el porcentaje de varianza de la matriz reducida que cada factor explica; así son más importantes aquellos factores que tienen un valor Eigen y un porcentaje de varianza explicada mayores.

3.4. Métodos de análisis factorial

El SPSS (Paquete estadístico para las Ciencias Sociales) cuenta con diferentes métodos de análisis factorial. Entre los más usados, se tienen cuatro:

a) El método de componentes principales sin iteración (PAI);

b) El método de componentes principales con iteración (PA2);

c) Factorización canónica de Rao (RAO); y

d) Método alfa (ALPHA)(α).

A continuación se presenta en forma breve cada uno de ellos.

3.4.1 Componentes principales sin iteración (PA1)

Este método de componentes principales sin iteracción, consta de dos procesos:

En el primero, la diagonal de la matriz de intercorrelaciones no se altera (en la diagonal aparece la unidad -1,00) y, se extraen componentes principales. El interés primordial es obtener el número menor de componentes que expliquen la mayor cantidad de varianza de los datos. Se retienen para la solución final rotada aquellos componentes cuyos valores eigen son iguales o mayores que 1.00. De esta manera se asegura que solo los componentes que explican por lo menos la varianza total de una variable serán significativos.

En el segundo procedimiento, se reemplazan los valores de la diagonal con la estimación de la comunalidad (h^2) y también se extraen factores o componentes principales. Las estimaciones de la comunalidad más usadas son:

Primero: emplear la correlación múltiple al cuadrado de cada variable con el resto de las variables del conjunto.

Segundo, consiste en utilizar el valor absoluto de la correlación más alta en cada columna de la matriz de intercorrelaciones, como comunalidad estimada (h^2) de esa variable (la de esa columna) para todas y cada una de las variables (columnas).

Este método se recomienda en estudios de tipo exploratorio, y en aquellos en los que se encuentre multicolinealidad entre las variables. Es decir, cuando las variables que se someten a análisis están muy correlacionadas entre sí.

3.4.2 Componentes principales con iteración (PA2)

En este método, como se señaló anteriormente, se reemplazan los valores de la diagonal principal de la matriz de intercorrelaciones con estimados de la comunalidad (h^2). El procedimiento de iteración sería para mejorar los estimados de la comunalidad.

En primer lugar, el programa determina el número de factores a extraerse a partir de la matriz no reducida de datos. En segundo lugar reemplaza la diagonal principal con estimaciones de la comunalidad. En tercer lugar, extrae el mismo número de factores de la matriz reducida, y las varianzas que explican estos factores son los nuevos estimados de la comunalidad. Este proceso se continúa hasta que la diferencia entre dos estimados de comunalidad sucesivos es mínima (una milésima o 0.001). Este método también se recomienda para estudios de tipo exploratorio. Es el más empleado por los investigadores como una primera aproximación al estudio del fenómeno en cuestión.

3.4.3 Factorización canónica de RAO

Este método proporciona una solución en la que la correlación entre el conjunto de factores hipotético (los que el investigador planteo como posiblemente existentes) y el conjunto de variables se maximiza. Es un ejemplo típico del análisis factorial clásico o de factores comunes y únicos.

Este método supone que la matriz de correlaciones está basada en una muestra de casos y se pregunta acerca de los parámetros de la población. Busca el número mínimo de factores que expliquen la matriz de correlaciones original. En este caso el investigador se plantea la siguiente interrogante: ¿Cuántos factores se requieren para que el ajuste entre los datos y los factores hipotéticos no se desvíen significativamente del azar? En este caso el investigador está sometiendo a prueba implícitamente la hipótesis nula que señala la no existencia de diferencias entre la cantidad de varianza que sus factores explican y aquella que se encuentre en la matriz original de intercorrelaciones. Es un ejemplo del caso donde lo importante es aceptar la hipótesis nula.

Este método extrae el número mínimo de factores que permite explicar la mayor cantidad de varianza respecto a la varianza de la matriz original no reducida. Este método se recomienda para estudios confirmatorios.

3.4.4 Análisis factorial tipo Alfa (α ALPHA)

Este método también corresponde al análisis factorial clásico o de factores comunes y únicos.

En este caso se parte del supuesto de que las variables incluidas en el análisis se consideran una muestra representativa del universo o población de variables. En este sentido, este método se deriva del *modelo dominio - muestra* de medición en ciencias sociales, y se recomienda por lo tanto, para ser empleado específicamente cuando el objetivo primordial del estudio en el de construir un instrumento, y no el de reducir a un número menor un conjunto dado de datos.

La razón por la cual reciba el nombre alpha es que uno de los objetivos primordiales del análisis es calcular o extraer factores que sean internamente consistentes. La prueba de consistencia Interna a la que se hace referencia en este procedimiento es la del coeficiente alpha de Cronbach. Se recordará que este coeficiente se aplica a instrumentos cuyos reactivos tengan más de dos opciones de respuesta. De esta manera, los factores obtenidos por medio de este método, si fueran analizados con el coeficiente alpha de Cronbach, obtendrían valores que señalan la existencia de consistencia Interna entre sus variables o reactivos constituyentes.

Las siguientes recomendaciones son particularmente importantes si se desea emplear este método de análisis factorial; aunque también son adecuadas para cuando se emplea alguno de los otros métodos ya mencionados.

En primer lugar, procúrese no emplear más de 100 variables o reactivos referidos a una variable o concepto complejo. Esto se debe a que aunque el programa SPSS cuenta con la posibilidad de manejar un número mayor de 100 variables en un análisis factorial, en la medida en que se incrementa el número de variables más allá de 100, el valor que la determinante de la matriz original del intercorrelaciones puede adquirir puede ser tal que impida invertirla. Esto representa, que no podrá producir una matriz de coeficiente de calificaciones factoriales. Esto a

su vez expresa que el investigador tendría que emplear la matriz de la estructura factorial como coeficientes de calificación factorial, lo que implica que sus factores terminales deberán quedar constituidos por variables o reactivos con cargas factoriales mínima, de valor absoluto de 0.40 y mayores[13].

En segundo lugar, se recomienda tener una muestra cuya **N** sea por lo menos de 5 veces el número de reactivos que contenga el instrumento inicial. Es decir, el tamaño mínimo de la muestra deberá ser: N = 5K, (dónde K es igual al número de reactivos). Lo ideal es que N sea igual a 10K. Esta recomendación se dirige particularmente a la situación en la que el investigador tiene como objetivo específico la elaboración de un instrumento. Si este no fuera el objetivo primordial del estudio y el método de análisis factorial que se empleara fuera otro, sigue implicando la recomendación de una N grande, entendiéndose por grande, más de 100 casos como mínimo. Se recordará que los diseños correlaciónales multivariados son diseños de muestras grandes[14]. En tercer lugar, se recomienda que todos los reactivos o variables sean de opción múltiple, en alguna de sus acepciones para garantizar una distribución normal de las variables. Es decir, por ningún motivo deberán los reactivos o variables registrar respuestas dicotómicas; las opciones de respuesta deben ser por lo menos tres hasta cinco o siete; de preferencia cuatro o cinco. Entre los tipos de reactivos de opción múltiple se tienen, por ejemplo; tres o más opciones de respuesta excluyentes; ordenar o clasificar a lo largo de un continuo de 3 a 7 intervalos, donde los extremos estén definidos, así como el intervalo intermedio. Las escalas de preferencia, los de grados de acuerdo, los de frecuencia de ocurrencia, etc. son, siempre y cuando tengan más de dos posibilidades de respuesta, variedades de reactivos de opción múltiple.

Cuarto y último, si no se pudiera tener un número de cien variables para ser sometidas a análisis, se recomienda llevar a cabo análisis factoriales parcializados. Los criterios que determinan cómo se pueden dividir las variables para ser sometidos a análisis factorial son principalmente dos: un criterio cualitativo, y otro cuantitativo.

[13] El texto de la primera edición señalaba como recomendación para este método: *Otra razón que subyace a esta recomendación, es el incremento del tiempo de máquina y del espacio del trabajo requerido por la computadora.*
Actualmente con el desarrollo de las nuevas tecnologías, los análisis por el programa SPSS es muy breve, lo que facilita las codificaciones.

[14] Con objeto de anular resultados espurios, producto del artefacto del proceso de computación.

Criterio Cualitativo. Hace referencia a la fundamentación teórica que subyace a la elaboración de los reactivos o índices de las variables. Es decir se someten a análisis factorial a aquellos reactivos que se supone están midiendo o la misma variable o la misma dimensión de una variable dada.

Criterio Cuantitativo. Puede cumplirse de dos maneras:

a) Analizando una matriz de intercorrelaciones inicial que contenga todas las variables o reactivos del estudio. El análisis consiste en agrupar las variables por sus magnitudes de correlación; formar un grupo con aquellas que tengan las correlaciones más altas; un segundo grupo con las que sigan y así sucesivamente, formando cuantos grupos sean necesarios o pertinentes. Pueden ser, por ejemplo dos: en uno todas aquellas variables que tengan correlaciones entre sí que excedan un cierto valor establecido por el investigador y un segundo grupo con las variables cuyas intercorrelaciones no excedan el valor estipulado.

b) Sometiendo a la matriz inicial de intercorrelaciones a un análisis de conglomerados, como por ejemplo aplicando el coeficiente de pertenencia. Se somete a toda la matriz de correlaciones a un análisis de pertenencia y entonces se someten a factorización los diferentes conjuntos de variables que pertenecieron a los diversos conglomerados detectados en la matriz de intercorrelaciones inicial. En realidad, el primer caso es una forma no muy estricta "a ojo de buen cubero[15]", de hacer lo que se hace, en el segundo caso, de manera matemáticamente más formal.

3.5. Métodos de Rotación

Se mencionó en la sección de Rotación a factores terminales, que las opciones de solución terminal más importantes eran la Ortogonal y la Oblicua. En esta sección se ampliarán más de cada una de ellas.

[15] Expresión que hace referencia aproximada de una cantidad cualquiera sin tener la absoluta certeza, la medida es el conocimiento adquirido por ensayo y error.

3.5.1 Métodos ortogonales de rotación

Entre los métodos ortogonales de rotación, que suponen independencia (no correlación) entre los factores terminales, se tienen tres: a) QUARTIMAZ; b) VARIMAX; y c) EQUIMAX. El investigador seleccionará el que más convenga a sus objetivos, tomando en cuenta lo que a continuación se presenta.

3.5.1.1 Quartimax. Este método tiene por objeto rotar los ejes de los factores para maximizar el principio de la estructura simple. Es decir, asegurarse que una variable cargue alto en un factor, y cero o cerca de cero en los demás. Este método es recomendable cuando las características mismas de los reactivos tengan la mayor probabilidad de exclusividad en un factor con respecto a los demás. Se recomienda para estudios confirmatorios; es decir aquellos que ponen a prueba hipótesis.

3.**5.1.2 Varimax.** El método Varimax produce soluciones factoriales que maximiza la cantidad de varianza explicada. En el que se emplea más comúnmente y se recomienda en estudios exploratorios y cuyo objetivo primordial sea la reducción de datos.

3.5.1.3 Equimax. El método Equimax es una combinación de los dos primeros; el cual, busca obtener factores que maximicen la varianza explicada y que al mismo tiempo, los factores queden constituidos por variables o reactivos que carguen alto en un factor y cero, o cerca de cero en los demás.

Las soluciones terminales rotadas que producen estos tres métodos difieren entre sí. Sin embargo, dentro de cada uno de ellos, la solución obtenida es la óptima. Cada uno de ellos produce la mejor solución factorial matemáticamente posible. Esta situación permite al investigador, por ejemplo, comparar soluciones cuando se tienen el mismo conjunto de reactivos y diferentes muestras; o diferentes conjuntos de reactivos (o variables) y la misma muestra, para escoger aquella que mejor se conforme o sus hipótesis o al conocimiento acumulado en esa área particular de interés. O le permite también, detectar, la estabilidad de la estructura factorial (o la falta de la misma), en muestras diferentes en el tiempo o espacio. En cualquier caso, el investigador está seguro, por lo menos, de que las soluciones encontradas son las óptimas desde el punto de vista matemático.

3.5.2 Método Oblicuo de rotación

Se recordará que este método se emplea cuando no se supone independencia entre los factores; por el contrario, se supone la existencia de correlación entre los factores obtenidos. El grado de relación u oblicuidad de los ejes de referencia está determinado por el ángulo que éstos forman entre sí.

Se recordará que existe una manera en la que se puede representar en forma gráfica una correlación. Esta representación se hace con un sistema de coordenadas. Cuando estas son perpendiculares entre sí, el ángulo que separa a la ordena da de la abscisa es de 90º. El coseno de un ángulo de 90° es igual a cero. Por esto se dice que los ejes de referencia ortogonales o perpendiculares representan una correlación de cero, o sea, absoluta independencia entre los ejes. Cuando se habla de relación o dependencia entre los factores, se refiere a la *Oblicuidad Gráfica* entre los ejes de referencia. Se recordará que cuando el ángulo que se forma entre dos líneas es de 0o, el coseno de ese ángulo es 1.00 (máxima correlación) y conforme el ángulo se va acercando a 90º, el coseno de ese ángulo (correlación) va siendo menor hasta llegar a 0.

La oblicuidad de los ejes de referencia, que son los que determinan los valores de las cargas factoriales de las variables que constituyen a los diferentes factores, queda establecida por δ(delta). Cuando el valor de δ es positivo, menor o igual a 1, se supone que los factores están extremadamente correlacionados y los ejes de referencia son muy oblicuos. Si el valor de δ es igual a 0, los ejes son bastantes oblicuos; este es el valor de default que tiene el programa del SPSS; si el valor de δ va de -0.5 a -5, se supone una oblicuidad menor. Cuando adquiere un valor menor a -5, los ejes son casi ortogonales; o sea, que se supone la casi independencia entre los factores.

El investigador puede determinar el grado de oblicuidad (o correlación) que supone entre sus factores, dependiendo de la literatura sobre el tema estudiado, estableciendo un valor para δ.

El problema de esta aproximación es el hecho de que no existen soluciones óptimas b únicas. Existen tantas soluciones como oblicuidades establezca el investigador. Debido a esto, se recomienda que el investigador decida de antemano, el grado de relación que supone existe entre los factores que se extraerán, y en virtud de este punto, escoja por medio del valor δ, la oblicuidad

o correlación que supone tienen los factores. Si no se está seguro de qué tan correlacionados puedan estar los factores obtenidos, es recomendable solicitar dos o tres soluciones que correspondan a diferentes grados de oblicuidad, para que el investigador seleccione a posteriori aquella que mejor concuerde ya sea con sus hipótesis, o con el conocimiento acumulado en esa área de estudio.

En cualquier caso, es más difícil establecer comparaciones entre los hallazgos obtenidos en diferentes ocasiones o con diferentes muestras para el mismo conjunto de datos, pues las soluciones serán diferentes, dependiendo del grado de relación que el investigador haya escogido como el adecuado en cada ocasión.

3.6. Opciones adicionales del Programa de Análisis Factorial del paquete estadístico (Statistical Package for the Social Sciences) (SPSS)

Entre las opciones adicionales para interpretar los resultados de la aplicación de un análisis factorial a un conjunto de datos, las más empleados por los investigadores y por lo tanto las que se recomiendan son las que dependen de los objetivos del estudio y se presentan a continuación.

Es pertinente solicitar siempre Medias y Desviaciones Estándar para cada una de las variables que se someten a análisis. Esto proporciona información sobre los puntajes crudos obtenidos por los sujetos en las diferentes variables. En ocasiones, pueden ayudar a la interpretación de los factores obtenidos.

Se recomienda también solicitar que se imprima la Matriz de Coeficientes de Calificaciones Factoriales (Factor Score Coefficient Matrix). Las razones para esto serían:

- Si el objetivo del estudio es la elaboración de un instrumento; el instrumento final deberá poder ser empleado por otros investigadores en forma directa (sin tener que realizar otro análisis factorial a partir de las **ji** variables o reactivos originales) y deberá permitir la obtención de calificaciones factoriales de sus sujetos.

- Si el objetivo del estudio era simplemente la reducción de un número grande de variables en una etapa exploratoria o preliminar de un estudio o proyecto mayor, el investigador deberá

poder contar con un sistema de calificación factorial del número reducido de variables de interés en la muestra final de su investigación.

El investigador puede solicitar al programa que le produzca e imprima calificaciones factoriales de los sujetos empleados como muestra para la reducción de variables o elaboración del instrumento, pues tiene interés en poner a prueba hipótesis ulteriores referidas a los factores obtenidos. Para esto solicita FACSCORE. Es decir, las calificaciones que los sujetos obtuvieron en los factores extraídos pueden constituir un nuevo conjunto de datos que se someterán a otros análisis estadísticos con fines ya sea exploratorios, descriptivos o confirmatorios.

Cuando el estudio tiene como principal objetivo someter a prueba alguna hipótesis el investigador puede modificar los siguientes parámetros, dependiendo de sus hipótesis específicas:

a) FACTORS, que señala cuántos factores deberán extraerse del espacio reducido de variabilidad;

b) EIGEN, que señala el valor Eigen mínimo que el investigador desee tengan los factores que se obtengan;

c) ITERATE, que señala cuántas iteraciones habrán de llevarse a cabo: esta situación es común cuando los datos no alcanzan la convergencia (diferencia de 0.001 entre los cálculos sucesivos de la estimación de la varianza en el método PA2) con veinticinco iteraciones (que es el valor de default del programa);

d) STOPFACT, cuando, se desea que se detenga el proceso de extracción de factores en aquella iteración en la que la comunalidad (h^2) varía de una estimación a la sucesiva en una cantidad diferente a 0.001 (valor de default).

3.7. Interpretación de resultados

Para ejemplificar la interpretación de los resultados, se muestra a continuación los pasos llevados a cabo de un conjunto de datos computados, resultantes de un análisis factorial con un grupo de estudiantes universitarios voluntarios quienes respondieron a los cuestionarios.

Las variables que se sometieron al análisis factorial fueron doce que se referían al profesor: cubrió el programa, claro, organizado, flexible, responsable, puntual, cumplido, ejemplos adecuados, señala objetivos de la exposición, sistemático, integra la información, relaciona conceptos. Para el programa fueron ocho variables: da una visión general del área, cubre puntos esenciales, proporciona información actualizada, secuencia pedagógica, interesante. La otra variable con siete preguntas, fue el método de enseñanza empleado por el profesor: si permitía la aplicación del conocimiento a situaciones prácticas, creativo, entretenido, sistematizado, logra retención del conocimiento, requiere más horas de estudio, es reforzante.

El instrumento consistió de escalas bipolares, un extremo definido como se menciona en el párrafo anterior, y el extremo contrario con lo opuesto a lo señalado anteriormente. Se disponía de siete intervalos (u opciones) de respuesta, tipo diferencial semántico. El instrumento fue aplicado a una muestra de 262 sujetos (alumnos). Se discutirán los resultados de un análisis factorial de componentes principales con interacción (PA2), con rotación Ortogonal Varimax y los mismos datos con PA2 y Rotación Oblicua, con una δ =-0.00 (valor de default del programa).

Los resultados que aparecen en el listado de computadora son los siguientes:

En ***primer lugar:*** aparece la lista de variables que se sometieron a análisis.

En ***segundo lugar:*** las medias y desviaciones estándar, así como el número de casos que respondieron a cada variable.

En ***tercer lugar:*** la matriz de intercorrelaciones entre todas las variables. Se observa que la matriz es cuadrada; es decir, tiene igual número de columnas e hileras; también es simétrica: las correlaciones por encima y por debajo de la diagonal son iguales en las celdillas correspondientes. En la diagonal se encuentran unos (l.00): el coeficiente de correlación de cada variable consigo mismo. Al final de la matriz de correlaciones aparece el valor de la determinante de la matriz. Este valor señala si será posible invertir la matriz; cuando éste sea el caso, inmediatamente aparece la matriz de correlaciones invertida.

En ***cuarto lugar:*** aparecen datos referidos la estimación de la varianza, el número de factores extraídos inicialmente, con su valor Eigen asociado y el porcentaje de varianza que cada uno explica; inmediatamente aparece la varianza acumulada que explican los factores sucesivos. En

esta parte, se deberá poner especial atención a los valores Eigen. La magnitud de los valores Eigen es la que le señala al investigador si el factor obtenido es válido o bueno. Por acuerdo general entre los estudiosos del campo, se aceptan como factores adecuados, aquellos que obtienen valores Eigen de 1.00 o mayores. Esto significa que dicho factor explica la varianza total de por lo menos una variable o reactivo. En este caso, el factor cumple la función reductora de datos que el análisis conlleva. Por lo tanto, el investigador pondrá especial atención a los valores Eigen obtenidos por los factores resultantes de la extracción inicial de los mismos.

En el ejemplo del estudio en cuestión, se observó que solo los primeros cinco factores iniciales obtuvieron valores Eigen de 1.00 o mayores. Esto señala que la matriz original de correlaciones K^2 (24^2) o Kr, se puede reducir a una matriz más pequeña 5K (5 X 24) que se conforma por 5 columnas (factores) y 24 hileras (r) o variables (reactivos). Estos cinco factores explicaron el 64.6% de la varianza de la matriz original de intercorrelaciones y se denomina *Matriz de Varianza Reducida*. Esta última responde en primer lugar, a la interrogante sobre la posibilidad de reducción de los datos originales; en segundo lugar, es la que se empleará en el resto del procedimiento de análisis.

En ***quinto lugar:*** aparece la Matriz Factorial de Componentes Principales con Iteraciones. Inmediatamente antes de ella, aparece una leyenda que indica el número de iteraciones que se requirieron para alcanzar la convergencia en las estimaciones de la comunalidad, que son los valores que se colocan en la diagonal de la matriz de intercorrelaciones para continuar el procedimiento de análisis. Esta matriz, factorial reducida, no ha sido rotada aún por lo que presenta dificultades en su interpretación. Estas dificultades se deben a que aún no se han rotado los ejes de referencia para cumplir, en la medida de lo posible, con los criterios de estructura simple y desarrollo positivo. Por esta razón, esta matriz se pasa por alto y se continúa con el procedimiento de rotación.

En ***sexto lugar:*** aparecen las estimaciones de la comunalidad después de haberse dado 10 iteraciones; los factores válidos que se extrajeron de la matriz de varianza reducida, junto con sus valores Eigen y los porcentajes de varianza explicada individual y acumulada.

Como se puede observar, el espacio reducido arroja sólo dos factores cuyos valores Eigen son iguales o mayores a la unidad. El primero obtuvo un valor Eigen de 9.5538, explicando un

71.6% de la varianza de la matriz reducida; el segundo, adquiere un valor Eigen de 1.65674, y explica el 12.4% de la varianza. Entre ambos, explican el 84.1% de la varianza.

Lo anterior significa que, en última instancia, las 24 variables (reactivos) originales pueden reducirse a dos nuevas variables (factores) complejas que explican el 84% de la varianza reducida de la matriz de intercorrelaciones inicial u original.

Hasta aquí, los resultados son iguales si se analizan con rotación oblicua u ortogonal. De aquí en adelante, la interpretación y los datos que se emplearon para llevarla a cabo, difieren cada uno de los métodos. A continuación se verán los datos analizados con Rotación Ortogonal Varimax y después con Rotación Oblicua (A =0.000).

3.7.1 Rotación Ortogonal Varimax

Para el presente ejemplo, se optó por la Rotación Ortogonal Varimax, pues era el interés primordial, el explicar la mayor cantidad de varianza posible.

El ***séptimo lugar:*** corresponde a la interpretación de los factores obtenidos, empleando para este efecto, la matriz factorial varimax rotada (que corresponde a la matriz de la estructura factorial). De esta matriz que consiste de cinco columnas y 24 hileras, tan solo se estudian las dos primeras columnas, que son los dos únicos factores que obtuvieron valores Eigen iguales o mayores a la unidad.

En cada columna o factor se deberá buscar a aquellas variables o reactivos que tengan una carga factorial igual o mayor a 0.40, ya sea positiva o negativa. En el factor 1, se observan 12 variables con cargas factoriales mayores a 0.40. En el factor 2 se encuentran seis variables con pesos mayores a 0.40 y una con un peso de 0.38.

Observando el contenido de las variables que cargan en el primer factor, se puede decir que éste se refiere básicamente al programa y al sistema de enseñanza empleada para impartirlo, mientras que las variables que cargan alto en el segundo factor, se refieren en su gran mayoría (6 de 7) a cualidades del profesor como docente. De esta forma, se pueden interpretar los

resultados obtenidos diciendo que el Factor I es una nueva variable, compleja que se podría llamar características estructurales de la Enseñanza (Programa y sistema de impartición), y el Factor II, características personales de la Enseñanza (Cualidades del Profesor).

De lo anteriormente expuesto se puede deducir que interpretar una matriz estructural factorial significa:

a) Detectar cuáles variables cargan alto en cada factor;

b) Analizar el contenido de esas variables y/o los procesos subyacentes a los mismos; y

c) Darle(s) un nombre o etiquetarlo(s).

Antes de continuar, se debe señalar que solo se consideran como factores, aquellos que además de tener valor Eigen mayor a la unidad, incluyan por lo menos tres variables o reactivos. Cuando el número de factores es grande, se observa que los últimos, ya sea que son opuestos a los primeros o que con dificultad, reúnen la característica de las tres variables o reactivos con carga factorial alta.

Por estas razones, y porque además la cantidad de varianza que explican son cada vez menores, se recomienda llevar a cabo una prueba Scree[16], que consiste en desechar aquellos factores cuyos valores Eigen muestren cambios mínimos en su magnitud de uno al que sigue.

Para denominar o interpretar un factor, las consideraciones primordiales que deberán tomarse en cuenta son: a) Las magnitudes de las cargas factoriales; b) Los signos de las mismas; y c) El hecho de que las variables carguen alto en un factor y bajo en los restantes.

[16] El test de sedimentación o Scree-test (Cattell, 1988) es un procedimiento gráfico bivariado donde se representan puntos cuyas coordenadas son los valores propios de la matriz de correlación original, las proporciones de varianza total explicada, en el eje de ordenadas, y el número de componentes en el de abscisas… a partir de cierto punto la función se hace prácticamente horizontal y es este punto el que indica el número más adecuado de factores. La lógica es que, a partir de este número, los sucesivos factores son triviales y sólo explican la varianza residual (Ferrando & Anguiano-Carrasco, 2010: 26).

En ocasiones se da el caso de que una variable cargue alto en más de un factor. Cuando sucede esto, se puede optar por una o una combinación de los siguientes criterios, con objeto de decidir en cuál o cuáles factores quedarán esas variables o reactivos.

Primero, se puede establecer un criterio numérico que exprese: *la variable permanecerá en aquel factor en el que cargue más alto*.

Segundo, se puede tomar la decisión con base en cuál o cuáles factores es más congruente el contenido de esa variable y las demás que lo constituyen.

En tercer lugar, se puede decidir dejarlos en todos los factores en los que aparezcan.

3.7.2 Rotación Oblicua

Para el ejemplo que se está presentando, se optó por rotar los ejes de referencia, con un alto grado de Oblicuidad o Correlación supuesta entre los factores obtenidos, ya que se empleó una $\delta = 0.00$, que es el valor de default del programa estadístico empleado.

Como ***octavo lugar:*** en el caso de una rotación oblicua, se encuentra la Matriz del Patrón Factorial (Factor Pattern) inmediatamente después de las estimaciones de la comunalidad posteriores a la iteración y de los valores Eigen, porcentajes de varianza individual y acumulada de los factores significativos obtenidos de la matriz de varianza reducida (igual que en el caso anterior). Esta matriz (del patrón factorial) señala la forma en que cada variable o reactivo de cada hilera está constituido en términos de los factores encontrados. Recuérdese que los coeficientes en este caso son Coeficientes Estandarizados (β beta) de Regresión que indican la composición de cada variable en términos de los factores obtenidos.

En ***noveno lugar:*** aparece una matriz de intercorrelaciones entre los factores obtenidos. Ésta le señala al investigador si el grado de correlación, y por lo tanto de oblicuidad supuesta por él, correspondió al que solicitó en el análisis por medio del valor de δ. Si en esa matriz se observan correlaciones substanciales (de 0.30 y más) entre los factores obtenidos, se debería de haber trabajado con una δ positiva y pequeña; si, por el contrario, las correlaciones son pequeñas o cercanas a cero, se debería haber establecido un valor negativo mayor para δ.

En el ejemplo en cuestión, se observan la mayoría de las correlaciones que van de 0.25 a 0.57, por lo que el grado de oblicuidad solicitado (0.00) fue el correcto.

Por último, en ***décimo lugar:*** aparece la matriz de la Estructura Factorial (Factor Structure). Con ésta, se procede a hacer la interpretación de los factores.

Se debe recordar que sólo fueron dos factores los que obtuvieron Eigen iguales o mayores a la unidad y por lo tanto en ellas se concentran los contenidos siguientes.

El Factor I tiene 18 variables con cargas factoriales mayores a 0.40. Esto lo hace aparecer como un factor muy general de evaluación docente, ya que solo 6 variables no cargaron alto en este factor. El Factor 2, tiene 10 variables con carga factorial mayor a 0.40 y una con una carga de 0.38. En este caso, la mayoría de las variables (9-11) se refieren a los aspectos personales del profesor en su función docente. Lo anterior nos permitirá interpretar (o nombrar) a los factores como sigue: Factor I: Evaluación Docente General; Factor II: Evaluación Docente Personal o del profesor.

Por otro lado, es conveniente señalar que de acuerdo con la matriz de intercorrelaciones entre los factores, se observa que la correlación entre ambos factores es de 0.30446.

3.8. Instrumento factorial final, o reducción final de datos

Cuando el objetivo del investigador es elaborar un instrumento factorial que sirva como índice de una variable compleja, quedan aún dos pasos a seguir, sin tomar en cuenta los procedimientos requeridos para determinar la confiabilidad y la validez del instrumento.

El primer paso consiste en reestructurar el instrumento original, de manera que en el instrumento final sólo queden incluidos los reactivos que constituyen los diferentes factores obtenidos. Cuando este es el caso, el investigador tendrá interés en proporcionar a los futuros usuarios los siguientes elementos: a) La escala factorial final; b) El instructivo de aplicación; y c) El instructivo de calificación factorial.

El segundo paso, el instructivo de calificación factorial deberá señalar que los puntajes crudos obtenidos por los sujetos en cada uno de los reactivos de cada factor, habrán de ser transformados

en puntajes Z[17], empleando para esto, las medias y desviaciones estándar obtenidos en la muestra con la que se elaboró el instrumento. Esto significa que el investigador habrá de proporcionar una tabla que contenga las medias y desviaciones estándar de los puntajes crudos para todos los reactivos que forman la escala factorial final.

Deberá proporcionar además, una tabla que contenga los coeficientes de calificación factorial de su instrumento, con objeto de que los futuros usuarios del instrumento puedan derivar puntajes o calificaciones factoriales de los sujetos a los que se les aplique, a partir de los puntajes crudos que obtengan en la escala factorial final.

El procedimiento de calificación factorial consiste en calificar por separado los reactivos que constituyen a cada factor. Los pasos a seguir son: a) Obtener puntajes crudos para los reactivos de cada factor; b) Transformarlos en puntaje "Z" empleando la Tabla de Medias y Desviaciones Estándar[18]; c) Multiplicar cada puntaje "Z" por el Coeficiente de Calificación Factorial correspondiente e ir sumando o restando (dependiendo del signo del coeficiente) los reactivos restantes de cada factor, hasta agotar los que correspondan. Este puntaje que se obtiene está dado en unidades de Desviación Estándar o "Z".

De esta manera, el investigador puede obtener calificaciones factoriales para los sujetos medidos con el instrumento elaborado.

Los coeficientes de calificación factorial se obtienen de la Matriz de Coeficientes de Calificación Factorial (Factor Score Coefficient Matrix), la última que aparece en el listado de computadora.

También se puede solicitar que los factores finales queden integrados por variables o reactivos cuyas cargas factoriales sean de 0.40 y más, ya que cuando se imposibilita la inversión de la

[17] Un puntaje Z indica la dirección y grado en que un valor individual obtenido se aleja de la media, en una escala de unidades de desviación estándar. Fórmula: $Z = \frac{x - \bar{x}}{\sigma} = \frac{x}{\sigma}$ (Reidl-Martínez & Gómez-Perezmitre, 2010: 100).

[18] De acuerdo con Hernández et al. (2010) las fórmulas son:

Media: $\bar{X} = \frac{X_1 + X_2 + X_3 + X_k}{N}$ Desviación estándar: $s = \sqrt{\frac{\text{sumatoria } (X - \bar{X}^2)}{N}}$

matriz de intercorrelaciones, se tendrá que emplear como matriz de coeficientes de calificaciones factoriales, a la matriz de la estructura factorial.

Se ha visto que la correlación entre los puntajes obtenidos con los coeficientes de la matriz de calificaciones factoriales, y los obtenidos con las cargas factoriales, es alta y significativa.

Por otro lado, cuando el objetivo primordial del investigador es el de reducir el número de datos, con objeto de emplear los factores obtenidos como nuevas variables en análisis subsecuentes que someten a prueba a diferentes hipótesis, la situación es diferente.

Si las hipótesis van a someterse a prueba con diferentes muestras, el investigador deberá emplear la Matriz de Coeficientes Factoriales de Calificación (Factor Score Coefficient Matrix) de la manera como fue explicada anteriormente. Lo mismo vale decir si la intención del investigador es replicar el estudio en otras ocasiones y con otras muestras. Los procedimientos a seguir serán los mismos que los señalados.

3.9. Reporte de un análisis factorial

Como última sección de este apartado se señalan algunas indicaciones y/o recomendaciones sobre la forma más conveniente para reportar los resultados de un análisis factorial por el investigador.

Es muy importante señalar qué tipo de análisis factorial se empleó (PA1, PA2, ALFA, Rao, Imagen); qué tipo y clase de rotación se seleccionó: Ortogonal (especificando si fue Varimax, Equimax, o Quartimax) u Oblicuo (especificando el valor de δ empleado). Se debe señalar el número inicial de variables, y el tamaño de la muestra empleada (N), así como sus características más sobresalientes.

Así mismo, se debe describir el tipo de reactivos empleados (opciones, cuántos, calificación original, entre otros) o el tipo de variables (pruebas que midieron "x" o "y", índices, entre otros).

Se debe presentar una Tabla que muestre los factores extraídos; esta tabla deberá contener la siguiente información:

a) Número de la variable o reactivo;

b) Nombre[19] de la variable, o el reactivo en su totalidad;

c) La carga factorial de cada variable o reactivo en cada factor;

d) La media y desviación estándar de los puntajes crudos de cada variable o reactivo;

e) El coeficiente de calificación factorial de cada reactivo o variable (si se tiene la información, la matriz de coeficientes factoriales de calificación);

f) El valor Eigen de cada factor;

g) El porcentaje de varianza explicada por cada factor;

h) La interpretación o nombre (o etiqueta) dado a cada factor.

i) Señalar cuánta varianza acumulada explicaron los factores obtenidos.

j) Si el método seleccionado recurrió a iteraciones (como es el caso del PA2), se deberá señalar cuántas se requirieron para alcanzar la convergencia.

[19] McDaniel & Gates (1999) señalan que esto es algo subjetivo y requiere de una combinación de intuición y conocimiento de las variables.

Tema IV Análisis de discriminantes

4.1 Introducción

El análisis de discriminantes se emplea cuando se desea diferenciar estadísticamente entre dos o más grupos de casos. Los grupos quedan definidos por la situación de la investigación particular. El inicio del procedimiento llevado a cabo por el investigador es seleccionar un conjunto de variables denominadas Discriminantes, que miden características en las que se espera o conocen de alguna manera que difieran en los grupos. Con ello el investigador, en base a la literatura pertinente revisada, supone que existen ciertas variables que distinguen a los grupos que está separando, algunas características que diferencian a los grupos en cuestión y ahora desea saber si esas mismas variables en conjunto, distinguen a los grupos, es decir, se desea discriminar entre los grupos de manera que se pueda diferenciarlos

El objetivo matemático del análisis de discriminantes es pesar y combinar linealmente a las variables discriminantes, en forma tal que los grupos sean tan diferentes estadísticamente hablando como sea posible. El análisis de discriminantes trata de hacer esto formando una o más combinaciones lineales de las variables discriminantes. A estas combinaciones lineales se les conoce como funciones discriminantes. Una función discriminante se puede representar de la siguiente manera.

$$Di = dij\ Zj + dj2 + \ldots + dipZp$$

Donde:

Di= Calificación en la función discriminante

d = Coeficientes pesados

Z = Puntajes estandarizados de las p variables discriminantes empleadas en el análisis.

El máximo número de funciones que se puede obtener en un análisis de discriminantes es:

a) Una menos que el número de grupos.

b) Igual al número de variables discriminantes si hay más grupos que variables.

4.2 Objetivos de la investigación

Los objetivos de la investigación que se puede tener cuando se emplea el análisis de discriminantes son de *Análisis* y de *Clasificación*.

Análisis. Este procedimiento proporciona pruebas estadísticas para medir el éxito con el que las variables discriminantes realmente discriminan cuando se combinan en las funciones discriminantes obtenidas. Cuando hay más de dos grupos, es posible obtener una discriminación satisfactoria con un número menor del máximo número de funciones posibles. En esta situación de análisis, los coeficientes pesados o estandarizados, se interpretan en forma semejante a los *coeficientes de regresión múltiple* y a las *cargas factoriales del análisis factorial*. Sirven para identificar a aquellas variables que contribuyen más a la diferenciación entre los grupos estudiados.

Clasificación. Este objetivo señala que una vez que se encuentran las funciones de las variables que proporcionan discriminación satisfactoria, se derivan funciones de clasificación que permiten la clasificación de casos nuevos cuya membrecía se desconoce. Como prueba de lo adecuado de las funciones discriminantes, se clasifica a los datos originales para ver qué tanto acuerdo se encuentra entre la predicción de membresía derivada de la o las funciones obtenidas y la membresía real de cada caso. El procedimiento que se emplea para llevar a cabo la clasificación es el siguiente: Se usa una combinación lineal diferente y separada para cada grupo. Cada una de ellas arroja como resultado una probabilidad de membresía para cada grupo. Posteriormente el caso se asigna o clasifica en el grupo que tiene la probabilidad más alta.

4.3 Métodos de análisis de discriminantes

En términos generales existen dos métodos para llevar a cabo un análisis de discriminantes: el *directo* y el *paso a paso* (Stepwise).

En el método *directo*, todas las variables independientes o discriminantes entran juntas y se crean las funciones discriminantes a partir del conjunto total de variables, independientemente del poder de discriminación de cada una de ellas.

En el método *paso a paso*, se emplea a la mejor variable discriminante de acuerdo con un criterio determinado; posteriormente a la segunda mejor variable y así sucesivamente. En cada paso, se pueden quitar las variables ya seleccionadas, si se observa que reducen las discriminaciones cuando se combinan con variables más recientes. Cuando ya se seleccionaron todas las variables, y las que siguen ya no contribuyen a una mayor discriminabilidad[20], se detiene el proceso. En otras palabras, cada variable independiente se selecciona sobre la base de su poder discriminativo. El conjunto reducido de variables es casi tan bueno para discriminar entre los grupos y a veces es mejor al conjunto total que se somete a análisis.

Algunas variables seleccionadas previamente pueden perder su poder discriminativo debido a que alguna combinación de otras tiene o proporciona la información discriminante que tenía la variable que se fue. Cuando éste es el caso, es bueno quitarla para evitar redundancia en la información.

Al inicio de este procedimiento, cada variable se pone a prueba previamente para ver si aún hace una contribución suficiente a la función discriminante. Si no lo hace, o si algunas o una se puede excluir, se quita la menos útil.

4.4 Importancia de las funciones discriminantes

Se cuenta con dos conjuntos de medidas para estimar la importancia de las funciones discriminantes que se constituyen durante el proceso de análisis.

[20] La condición de ser discriminables (Diccionario Universal, 2018).

Dentro del primer conjunto están los valores Eigen, que señalan la importancia relativa de cada función. El porcentaje de varianza explicada del espacio reducido también es un indicador de la importancia de la función. La *correlación canónica* es un indicador de la asociación entre la función discriminante y la membresía a los diferentes grupos. Si se eleva al cuadrado la correlación canónica, se indica la proporción de varianza de la función discriminante que queda explicada por los grupos estudiados.

Dentro del segundo conjunto, se encuentran varias pruebas de significancia relacionados con la información discriminante no explicada por funciones anteriores. La principal es la *Lambda de Wilk*, que es una prueba de significancia de la diferencia total entre varios centroides. Es la proporción o razón de las determinantes de las matrices entre grupos respecto de las matrices de sumas de cuadrados totales. En realidad, la Lambda es una extensión de la prueba F.

$$\Lambda = \frac{(W)}{(T)}$$

Donde:

T = Suma de cuadrados de la matriz de la muestra total.

W = Suma de cuadrados de la matriz intragrupos y es igual a la suma de las diferentes sumas de cuadrados de los K grupos.

Esta prueba tiene una distribución muestral semejante a la distribución de Chi-cuadrada, con grados de libertad igual a P (K-1), donde p es igual al número de variables, y K es igual al número de grupos. La Lambda se transforma a Ji cuadrada y ésta se interpreta como siempre.

4.5 Interpretación de coeficientes

En un análisis de discriminantes, existen tres tipos principales de coeficientes: a) Estandarizados de la función discriminante; b) Los no estandarizados de la función discriminante; y c) Los de la función de clasificación.

a) *Coeficientes estandarizados de la función discriminante* son los **dij** de la fórmula arriba presentada. Se usan para computar calificaciones discriminativas si las variables discriminantes originales se encuentran en unidades **Z**. También se emplean para establecer los valores de los centroides o medias del grupo en la función. Es decir, si un centroide es una media de calificaciones discriminantes, para poder calcular esta media, se emplean los coeficientes estandarizados. Cuando se establecen comparaciones entre medias o centroides, se está indagando qué tan lejos están los grupos en esa dimensión (función). La importancia analítica de los coeficientes estandarizados de la función discriminante está dada por la magnitud de los mismos, ignorando el signo (+,-); y se refiere a la contribución relativa de esa variable a la función. El signo asociado al coeficiente establece si su contribución es positiva o negativa. Su interpretación se realiza de la misma forma que los pesos o coeficientes beta en regresión. Esto también quiere decir que permite inferir el comportamiento de las variables de la función en la población de donde proviene la muestra del estudio.

b) *Los coeficientes no estandarizados de la función discriminante* se emplean para calcular calificaciones discriminantes (Di) con puntajes crudos o para conocer el comportamiento de las variables constituyentes de la función, en la muestra. Se interpretan de manera semejante a los coeficientes B de regresión múltiple.

c) *Los coeficientes de la función de clasificación* sirven para clasificar los sujetos a partir de los puntajes crudos en el grupo en el que obtengan mayor puntaje. Esto significa que habrá tantas funciones clasificadoras como grupos existan; y que se calcula un puntaje de clasificación (Ci) para cada sujeto en cada uno de los grupos estudiados. En aquel grupo o con aquella función de clasificación con la que obtenga un porcentaje mayor, será el grupo en el que se le clasifique. La ecuación de la (s) función(es) de clasificación se puede representar como sigue:

La ecuación de la (s) función(es) de clasificación se puede representar como sigue:

$$Ci = Ci_1V_1 + ci_2 V_2 + ... + cipV_p + cio$$

4.6 Tipo de análisis de discriminantes

Se recordará que se señaló anteriormente que existe el método directo y el paso a paso de análisis. Entre los métodos paso a paso, existen cinco tipos: 1) WILK'S; 2) MAHAL; 3) MAXMIF; 4) MINRESID; y 5) RAO.

4.6.1 Wilk. El tipo Wilk's emplea como criterio de inclusión de las variables en cada paso, una F multivariada total para poner a prueba las diferencias entre los centroides, tomando en cuenta la diferencia entre ellos, así como su cohesión interna (Homogeneidad de varianza u Homocedasticidad) de los grupos.

4.6.2 Mahal. Este método tiene como principal objetivo maximizar la distancia entre los dos grupos más cercanos. La prueba que emplea como criterio de inclusión de las sucesivas variables es la D cuadrada de Mahalanobis[21] (una medida de distancia).

4.6.3 Maxminf. El tipo Maxminf busca maximizar el valor más pequeño de F entre pares de grupos, y el criterio de entrada es una F.

4.6.4 Minresid. El tipo Minresid separa a los grupos que se encuentran más cercanos, y su objetivo primordial es el de disminuir la varianza residual.

4.6.5 Rao. El tipo Rao usa la V de Rao como criterio de elección de variables sucesivas. Escoge aquella que contribuye al mayor incremento en V y por lo tanto incrementa la separación más grande posible, general entre los grupos.

[21] *La distancia de Mahalanobis es una medida de distancia introducida por Mahalanobis en 1936. Su utilidad radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales. Se diferencia de la distancia euclídea en que tiene en cuenta la correlación entre las variables aleatorias*. Fuente: Mahalanobis, P.C. (1936/2018)., página 49.

4.7 Interpretación de resultados

A continuación se muestra un ejemplo, donde se procede a interpretar un listado de computadora resultante de un análisis de discriminantes.

Los datos que se emplearon para el presente ejemplo son similares a las respuestas a un instrumento aplicado a 100 niños de 4 a 5 años de edad, que mide relaciones afectivas con la madre (Cruz-Velasco & Galindo Morales, 1989). Cincuenta niños asistieron a un Centro de Desarrollo Infantil de tiempo completo, los otros 50 niños asistían por medio tiempo. Se supone que existen diferencias en el desarrollo o expresión de sus relaciones afectivas con sus madres, entre estos dos grupos (tiempo completo vs. medio tiempo). El instrumento aplicado consta de 40 reactivos, de tres opciones de respuesta, donde la calificación de las opciones era tal que se obtuviera un puntaje más alto (3) en la opción que reflejara relaciones afectivas más adecuadas con la madre. Un puntaje intermedio (2) y de (1) en aquella que reflejara las relaciones afectivas más inadecuadas con la madre.

Estos 40 reactivos fueron sometidos a una prueba de diferencia de medias, de t de Student, para determinar, en una primera instancia, cuáles reactivos eran respondidos por separado de manera diferente por los niños de los grupos estudiados. De estos cuarenta reactivos, doce arrojaron t's con probabilidades asociadas de 0.05 o menores. Estos doce reactivos fueron los que se sometieron a análisis de discriminantes.

Los reactivos son los siguientes:

Preguntas	Número del reactivo
1. ¿Con quién te gustaría platicar más?	04
2. ¿Quién te habla más bonito?	06
3. ¿Con quién te gustaría ir al cine?	08
4. ¿Quién prefieres que te platique un cuento?	14
5. ¿Quién cumple lo que te promete?	15
6. ¿Quién es más gritona?	16
¿Quién piensa que eres malo(a)?	17

Preguntas	Número del reactivo
¿Con quién estás más tiempo?	24
¿Quién te escucha más?	25
¿Quién no te hace caso?	34
¿Quién está siempre contigo?	35
¿ A quién extrañas más cuando no la(o) ves	38

Sabiendo que estos reactivos discriminaban por separado a los grupos estudiados, se sometieron al análisis de discriminantes con el objeto de averiguar si se conforman en una función discriminante que distingue a ambos grupos; la función puede quedar constituida por todas o algunas de las variables o reactivos sometidos a análisis.

A continuación era encontrar si la función extraída, permitía derivar una función de clasificación que produjera predicciones de membresía de grupo adecuados.

El método de análisis empleado fue el *paso a paso*, tipo Mahal. Se estableció una probabilidad de 0.5, ya que el número de sujetos en cada grupo estudiado es igual.

Si el número de sujetos pertenecientes a cada grupo son muy diferentes, como 25 en uno y 75 en otro, se puede establecer una probabilidad *a priori* con la instrucción PRIOR, para que se tome esta situación en cuenta en el momento de predecir membresía de los grupos. Como se observará, el listado consta de diversas partes. Se va a explicar cada una de ellas por separado.

En *primer lugar*, aparece el número de casos en cada uno de los grupos. En *segundo lugar*, aparecen las medias de los grupos para cada uno de los reactivos incluidos en el análisis. Aparece una media para cada grupo y la media total de ambos grupos reunidos. En *tercer lugar*, aparecen las desviaciones estándar de cada reactivo para cada grupo por separado y en conjunto (ambos grupos). En *cuarto lugar,* aparece la matriz de covarianza colapsada entre grupos, e inmediatamente después, la matriz de correlación colapsada entre grupos. En la primera se observan los coeficientes de covarianza entre cada posible par de reactivos en ambos grupos juntos; y en la segunda, los coeficientes de correlación entre todos los posibles pares de reactivos; hasta ahora, no se ha iniciado el análisis propiamente dicho. Estos datos son

generados y servirán al investigador para interpretar los resultados del análisis propiamente dicho, como se verá más adelante.

A partir de aquí se inicia el análisis de discriminantes.

En primer lugar aparecen diversas leyendas que son concomitantes al tipo de análisis que se haya seleccionado realizar. En una primera instancia se señalan los aspectos que se tomarán en cuenta para la selección de cada variable en cada paso sucesivo del análisis. De esta manera se puede leer lo siguiente:

a) Regla de selección: Maximización de la distancia (D cuadrada (D^{232})) mínima de Mahalnobis[22] entre los grupos.

b) Número máximo de pasos: 24 (el doble del número de variables incluidos en el análisis).

c) Nivel mínimo de tolerancia: 0.001 (una milésima entre una D^2 significativa y la siguiente).

d) F mínima para entrar: 1.000 (el valor mínimo que F tiene que tener para que entre la variable al análisis).

e) F máxima para desplazar: 1.000.

En una segunda instancia aparecen leyendas relacionadas a las funciones discriminantes canónicas. La primera de ellas se refiere al máximo número posible de funciones, que en este caso es de uno, ya que sólo hay dos grupos. Enseguida se señala el porcentaje de la varianza acumulada que se explicará: 100.00. Por último, la máxima significancia que puede obtener la Lambda de Wilk: 1.000[23].

En este apartado también aparece la probabilidad *a priori* establecida para cada grupo: 0.50000, que es el valor de default (en caso de no modificarse con PRIOR) y que para el presente ejemplo es la adecuada, pues existe igual número de sujetos en cada uno de los dos grupos (Nj = 50; N2 = 50).

[22] Véase Cuadras, 1989, páginas 306-308. Disponible en: www.ine.es/ss/Satellite?

[23] Que así se maneja con esta prueba. Nota de la autora.

En segundo lugar aparecen las variables no incluidas en el análisis después del paso 0, o sea antes del primer paso, Aquí aparecen todas las variables que se someten a análisis. En esta sección se puede ver cuál será la primera variable que se incluirá en el primer paso del análisis: aquella que tiene el valor F, o D cuadrada más alto. En este caso, la variable es el Reactivo número 25.

En tercer lugar aparece el Paso 1, donde se señala que la variable que se incluyó fue efectivamente el Reactivo número 25. Inmediatamente a continuación se observa el valor de la Lambda de Wild, su F equivalente; los grados de libertad asociados a éstos y su nivel de significancia. También se observa el valor de la D cuadrada mínima, y su F equivalente, así como sus grados de libertad asociados y su nivel de significancia. Dentro de este primer paso también aparecen las variables incluidas en este primer paso y aquellas no incluidas. De entre estas últimas estudiando cuidadosamente la columna de F o de D cuadrada, se observa que la siguiente variable que se incluirá en el análisis será el Reactivo #14.

Esta secuencia de datos aparece en todos y cada uno de los pasos hasta llegado el momento en el que los valores F o D cuadrada ya no alcancen los niveles de tolerancia pertinentes o sea insuficientes para continuar el procesamiento de datos.

En el presente ejemplo, el análisis se detuvo en el paso 9. En ese paso aparece una leyenda que señala que ya no se continuará con el análisis.

Inmediatamente después aparece una Tabla Sumaria. En está aparecen los reactivos que se incluyeron en el análisis (nueve variables), los valores Lambda de Wilk que cada una obtuvo, con su significancia; sus valores mínimos de D cuadrada y su significancia.

Después de analizar la Tabla Sumaria, se debe buscar la sección intitulada *Funciones Canónicas Discriminantes*. En esta sección se observan los siguientes da¬tos importantes:

En la primera parte:

a) El número de funciones extraídos de los datos (en este caso, uno).

b) El valor Eigen obtenido por la misma, que señala qué tan importante es. En este caso fue 0.87477.

c) A continuación, aparece el porcentaje de varianza explicada del espacio re-ducido, así como el porcentaje de varianza acumulado que fue 100%, en ambos casos. El espacio reducido se refiere a la variabilidad que existe en las variables discriminantes que formaron parte de 1a función canónica discriminante encontrada (sólo nueve variables de las doce originalmente incluidas).

d) La correlación canónica que de 0.6830830. Esto significa que las nueve variables correlacionan 0.68 con los grupos.

En la segunda parte, aparece información referida a la significancia de la función que contiene la información existente en los datos antes de que se constituyera la misma. Los datos importantes que aparecen son los siguientes:

a) El valor de la Lambda de Wilk, 0.5333976.

b) Su transformación a Ji cuadrada: 58.76, con sus 9 grados de libertad y su nivel de significancia al 0.0000.

Lo anterior significa que la función canónica discriminante conformada por las nueve variables incluidas es significativa al 0,0000.

En seguida se deberán estudiar cuidadosamente los Coeficientes Estandarizados de la Función Canónica Discriminante. Esto se hace con objeto de averiguar cuáles variables son las que más contribuyeron a diferenciar a los grupos. En el presente caso, las variables se podrían ordenar agrupados de la siguiente manera: Las variables Reactivos 24, 25 y 34 (*Con quién estás más tiempo*; *Quién te escucha más*; y *Quién no te hace caso*), son las que más contribuyeron a la función ya que sus coeficientes están todos arriba de 0.40. Posteriormente se podrían agrupar las variables Reactivos 6 y 14 (*Quién habla más bonito*; *Quién prefieres que te platique un cuento*) que obtuvieron coeficientes mayores a 0.30. Finalmente, un último grupo de variables podría quedar constituido por las variables Reactivos 8, 15, 16, y 17 (*Con quién te gustaría ir al cine*; *Quién cumple lo que te promete*; *Quién es más gritona*; *Quién piensa que eres malo*) que tienen coeficientes que van de 0.20 a 0.29.

Como se puede observar, las variables que discriminan entre los grupos estudiados hacen referencia a: Atención y cuidado en el primer grupo; Interacción verbal entre el niño y el adulto; y Relación afectiva positiva-negativa entre el niño y el adulto.

Lo anterior ejemplifica la información que proporciona un análisis de discriminantes. Estos resultados se comparan con la literatura referida al tema de investigación en cuestión y se puede así, poner a prueba diversas hipótesis de investigación.

A continuación se pueden ver los coeficientes de las funciones de clasificación, que juntos con las constantes, sirven para predecir la membresía de los sujetos a los diferentes grupos; multiplicando los puntajes obtenidos por ellos en las variables discriminantes incluidas en la función. En aquella función donde el sujeto hubiese obtenido el puntaje más alto (la del grupo 1 o 2) y se predice a qué grupo pertenece.

Inmediatamente después se consultan los centroides de los grupos. En el presente ejemplos se obtuvieron centroides de 0.92582 para el grupo 1 (medio tiempo: 4 horas) y de -092582 para el grupo 2 (Tiempo Completo: 8 horas).

Esto significa que la distancia que los separa es de casi dos unidades Z, o de desviación estándar. Se deberá recordar que los centroides son resultado de obtener calificaciones de discriminación empleando los coeficientes estandarizados de la función canónica discriminante, que está en unidades Z. Por lo anterior, los centroides también están en Z.

Se puede consultar, si así se desea, la *prueba de igualdad de matrices de covarianza, la M de Box*[24]. En este caso, M = 72.422, que equivale en forma aproximada a una F de 1.4527, que con 45 y 31550.9 grados de libertad (gl), dan una significancia de $p < 0.001$. Esto significa que las matrices de covarianza no son iguales, lo que contraviene a uno de los supuestos de este modelo: La igualdad de las matrices de covarianza. Sin embargo, se ha visto que el modelo es tan fuerte, que el hecho de no cumplir con es¬te supuesto, no afecta los resultados de manera notoria.

[24] La prueba Box M es una extensión del test de Barttlet para escenarios multivariantes y permite contrastar la igualdad de matrices entre grupos. Dada la sensibilidad de este test se recomienda emplear un límite de significancia de 0.001 (Tabachnick & Fidell, 2001; Amat, 2016).

Por último, se solicitó en este ejemplo el análisis para cada sujeto los siguientes datos: Número del sujeto, grupo al que pertenece; probabilidad más alta de pertenecer al grupo (P (X/h) P (h/X); segunda probabilidad más alta P(h/X); calificaciones discriminativas. Las probabilidades se calculan con PRIOR y con los coeficientes estandarizados de la función canónica discriminante. Si se suman estas calificaciones discriminantes para cada grupo por separado y se dividen entre el número de sujetos en cada grupo, se obtiene los centroides.

Al final de este listado, aparece el resumen de la clasificación en un cuadro de doble entrada donde aparece la siguiente información: Grupo, Número de casos y membresía del grupo predicha. En cada celdilla aparece la frecuencia de los casos predichos para cada grupo, así como su expresión porcentual. En la parte inferior del cuadro aparece el porcentaje de casos "agrupados" -definidos como pertenecientes a alguno de los grupos estudiados-, correctamente clasificado: 84.00%. Esto significa que, conociendo la información de los sujetos en las nueve variables, el investigador podría predecir la membresía de sus sujetos investigados con una exactitud o corrección del 84 porciento.

La información de calificaciones discriminativas y los centroides pueden ser representados en forma gráfica, obteniéndose un dispersigrama para cada grupo y otro en conjunto, donde aparecen claramente señalados los centroides de cada grupo.

Hasta ahora, se puede decir que un análisis de discriminantes señala cuántas y cuáles variables son suficientes y necesarias para explicar la diferencia entre dos o más grupos y qué tan bien lo hacen. En este caso, se puede interpretar a las variables discriminantes como independientes y a los grupos como variables dependientes.

4.8 Reporte de un análisis de discriminantes

A continuación se presentan algunas recomendaciones generales respecto de la forma de reportar los resultados de un análisis de discriminantes.

En primer lugar, se debe señalar si el método empleado fue directo o paso a paso; en caso de haber sido paso a paso, se deberá señalar qué tipo se usó (Mahal, Rao, Wilk's, Maxminf o Minresid).

También, cuántas variables se incluyeron inicialmente, describiéndolas someramente. Se deberá señalar si estas variables fueron incluidas a partir de la literatura revisada o si son el resultado de algún análisis estadístico previo. Se deberá definir los grupos con claridad, estableciendo también el tamaño de los mismos, si se modificó el valor por default de PRIOR (0.50), en virtud de los tamaños de los grupos. También se deberá dar los valores de los centroides correspondientes a los grupos estudiados.

Es conveniente seleccionar algunas estadísticas adicionales al mero análisis de discriminantes. Se recomienda que se soliciten las medias y desviaciones estándar de las variables discriminantes. También es pertinente solicitar algunas de las opciones como son: listado de calificaciones discriminantes y probabilidades de clasificación; resumen de predicciones de clasificación, dispersigrama separados y conjunto de los grupos estudiados.

Lo siguiente que se tiene que reEs conveniente seleccionar algunas estadísticas adicionales al solo análisis de discriminantes. Se recomienda que se solicite las medias y desviaciones estándar de las variables discriminantes. También es pertinente solicitar algunas de las opciones como son: listado de calificaciones discriminantes y probabilidades de clasificación; resumen de predicciones de clasificación, dispersigrama separados y conjunto de los grupos estudiados.

Lo siguiente que se tiene que reportar es la función canónica discriminante, con sus valores Eigen, porcentaje absoluto y acumulado(s) de varianza explicada, la correlación canónica; la Lambda de Wilk, la Ji cuadrada, con los grados de libertad asociados y su nivel de significancia. Si este último está por debajo de la λ planteada por el investigador, éste podría rechazar la hipótesis nula que subyace al análisis.

Enseguida se deberá presentar la Tabla Sumaria, que incluye a las variables que formaron la función canónica discriminante, junto con sus valores individuales de Lambda de Wilk, y sus significancias asociadas: a la D cuadrada mínima (Si es tipo Mahal); la V de Rao (Si es Tipo Rao); la F (Si es maxminf); el residual (Si es minresid) y sus significancias asociadas.

Los siguientes elementos de resultados que se habrán de reportar pueden conformar una Tabla donde se incluyan los siguientes datos: número de la variable, coeficientes estandarizados y coeficientes no estandarizados.

También los elementos que se integran, que son: Los coeficientes de clasificación, las medias y la desviación estándar. Donde las medias y desviaciones estándar corresponden a los obtenidos por los reactivos o variables en forma de puntajes crudos.

Por último, se puede presentar el cuadro que resumen la clasificación de los sujetos por medio de las funciones de clasificación derivados de la función canónica discriminante encontrada. Si el investigador desea hacer más explícita la diferencia entre los grupos estudiados, puede ilustrarse empleando los dispersigramas correspondientes.

Los elementos anteriores presentan al lector del reporte de investigación una idea muy clara de los resultados encontrados. Obviamente, el investigador deberá discutir sus hallazgos a la luz del marco teórico o conceptual que dirigió su investigación, así como en relación con descubrimientos previos relacionados a estudios semejantes al actualmente realizado.

Tema V Correlación canónica

5.1 Introducción

El modelo de la correlación canónica parte de una serie de supuestos que tienen que ver, por un lado, con las variables que se incluirán en el estudio, y por el otro con el objetivo estadístico primordial del mismo. Así, supone que las variables están medidas todas a nivel intervalar, o en forma tal que cumplan las características de este nivel de medición (distribución normal alrededor de un punto de origen arbitrario); supone también que las relaciones entre las variables son lineales. Por último, supone como objetivo estadístico principal, el de explicar la mayor cantidad de varianza posible.

5.2 Fundamentación teórica

Deseando esclarecer el momento en el que es deseable emplear el diseño multivariable de investigación denominado *Correlación Canónica*, se recordará algunos aspectos importantes de los diferentes diseños contemplados hasta ahora.

Cuando al investigador le interesa explicar una sola variable dependiente a partir de un conjunto limitado de otras variables (independientes), la elección adecuada es emplear una regresión múltiple. A medida en que el número de variables en el conjunto de datos se incrementa, la regresión múltiple deja de ser la opción más adecuada. Eventualmente se llega al momento en que no se puede hacer una interpretación significativa de los resultados a menos que se emplee algún procedimiento de reducción de datos. Cuando éste es el caso, la opción adecuada es la de emplear el análisis factorial; sin embargo, este procedimiento no distingue entre variables dependientes e independientes.

Una solución posible en estas circunstancias es la de dividir al conjunto de datos en variables dependientes por un lado, independientes por el otro y someter a cada conjunto por separado al análisis factorial. En caso de optarse por esta estrategia, tendrán que tomar en cuenta tres circunstancias: *La primera* se refiere a la fortaleza de la justificación teorética que avala la división de las variables en cuestión en los dos conjuntos; *La segunda* se refiere al hecho de que si el objetivo es establecer o explicar las variables dependientes(ahora factores) por medio de las independientes (también factores), se podría emplear una regresión múltiple tomando a cada factor de las variables dependientes y por otro lado, todos los factores o variables independientes. Esto se debería hacer una y otra vez, hasta agotar los factores que corresponden a las variables dependientes. Sin embargo al existir más de un factor que funcione como variable dependiente, y si el investigador desea tratarlos a todos en forma simultánea, y no por separado, este conjunto de procedimientos ya es inútil; *La tercera* circunstancia se refiere a la naturaleza del análisis factorial en sí mismo. El análisis factorial selecciona factores sobre la base de las interrelaciones de las variables disponibles y trata de maximizar la varianza explicada sólo en esas variables. De esta manera, los factores derivados de un análisis factorial de un conjunto de variables no necesariamente son las mejores variables compuestas que explican o dan cuenta del otro conjunto de varianza, la estrategia adecuada es el *Diseño de Correlación Canónica.*

El análisis de correlación canónica se lleva a cabo a partir de dos conjuntos de variables, donde a cada uno de ellos se les puede dar un significado teórico como conjunto. Este procedimiento busca derivar una combinación lineal de cada conjunto de variables de tal forma que se maximice la correlación entre las dos combinaciones lineales. Se pueden derivar muchos pares de combinaciones lineales. Estas variantes canónicas son esencialmente equivalentes a los *componentes principales* producidos por el análisis del mismo nombre, con la diferencia de que el criterio de solución es otro. Ambas técnicas producen combinaciones lineales a partir de las variables originales, pero el análisis de la correlación canónica no lo hace con el objeto de explicar la mayor cantidad de varianza posible dentro de un conjunto de variables, sino tratando de dar cuenta de la máxima cantidad de relación existente entre los dos conjuntos de variables.

El análisis de componentes principales y el de correlación canónica son análogos en diferentes aspectos. El análisis de componentes principales selecciona un primer componente principal que explica la mayor cantidad de varianza de un conjunto dado de variables, y después computa un

segundo componente principal que explica la mayor cantidad de varianza que no quedó explicada por el primer componente, y así sucesivamente. El análisis de correlación canónica sigue un procedimiento semejante, se selecciona el primer par de variantes canónicas de tal manera que tengan la intercorrelación más alta posible, dadas las variables particulares involucradas. Después se selecciona el segundo conjunto, (par) de variantes canónicas. Es decir, aquel par que explica la mayor cantidad de relación existente entre los dos conjuntos de variables y que no fue explicado por el primer par de variantes canónicas, y así sucesivamente.

En virtud de que ambos procedimientos explican la varianza residual (no explicada por variantes o componentes precedentes o anteriores), ambos producen combinaciones lineales de variables que son independientes o no están correlacionadas unas con otras. La diferencia entre estos dos procedimientos -análisis de componentes principales y análisis de correlación canónica-, es que, en el primero, los componentes se pueden rotar a una solución terminal mientras que los segundos no.

5.3 Información producida por el análisis de correlación canónica

Los tipos de información más importantes producidos por el análisis de correlación canónica son de: a) *Las variantes canónicas* y b) *Las correlaciones canónicas entre ellas*.

a) *Las variantes canónicas* vienen en dos conjuntos: uno para cada uno de los subconjuntos de variables introducidos al análisis. Estas variantes están compuestas por coeficientes que reflejan la importancia de las variables originales en el conjunto, al formar las variantes. El punto esencial del análisis de correlación canónica es que las variantes canónicas de cada conjunto deben corresponder. Es decir, la primera variante canónica del primer conjunto de variables y la primera variante canónica del segundo conjunto de variables se eligen para que correlacionen en forma máxima una con otra; y así para las segundas y sucesivos pares de variantes canónicas.

b) La cantidad de correlación entre cada par de variantes canónicas correspondientes es la *correlación canónica entre ellas*. Su cuadrado, que es equivalente al valor Eigen, represente la cantidad de varianza de una variante canónica que queda explicada por la

otra variante canónica. Como un coeficiente de correlación canónica o de otro tipo, una correlación entre A y B siempre es equivalente a la correlación que hay entre B y A, desde un punto de vista matemático, no importa cuál subconjunto de variables es considerado por el investigador como variables dependientes y cuál como independientes; el resultado es siempre el mismo.

5.4 Resultados impresos del análisis de correlación canónica

El programa SPSS produce en forma automática dos tipos de información analizada: La tabla sumaria del análisis de correlación canónica y dos matrices de variantes canónicas.

La tabla sumaria contiene los siguientes elementos:

1. Las magnitudes de los valores Eigen. Se presentan en la primera columna de la tabla en orden descendente e igual en número al número de variables del conjunto más pequeño. Los valores Eigen pueden interpretarse como la proporción de la varianza compartida por el par de variantes canónicas al que corresponden.

2. Las correlaciones canónicas. Estas son simplemente las raíces cuadradas de los valores Eigen y su significado es equivalente al de un simple coeficiente de correlación producto momento de Pearson.

3. Las siguientes cuatro columnas proporcionan información relativa a la significancia estadística de las correlaciones canónicas. La primera estadística que se proporciona es la Lambda de Wilk, que pone a prueba la hipótesis nula que establece que no hay asociación lineal residual entre los dos conjuntos de variables, después de haberse extraído las variantes canónicas (si es que las hubo) precedentes. Para una prueba real de significancia se dan los valores de Ji-cuadrada en la siguiente columna, seguida por los grados de libertad, y los valores de P (niveles de significancia) asociados a la misma en la última columna.

Inmediatamente después de la Tabla Sumaria se presentan dos matrices de coeficientes estandarizados de las variantes canónicas; una para cada uno de los dos conjuntos de variables

incluidas en el análisis. Estos proporcionan la información exacta de los pares de variantes canónicas correspondientes que producen las correlaciones canónicas presentadas en la Tabla Sumaria. El tamaño de los coeficientes indica la contribución relativa de las variables originales en la composición de las variantes canónicas. Estos coeficientes se calculan sólo para los pares de variantes canónicas cuyas correlaciones canónicas fueron significativas de acuerdo con los parámetros por default del programa.

Se debe tomar en cuenta que cuando los datos contienen relaciones del tipo que el análisis de correlación canónica puede descifrar, el procedimiento los puede identificar correctamente. Es decir, no se puede asegurar que el procedimiento de análisis de correlación canónica encuentre siempre algún patrón comprensible de relaciones. Puede suceder también, que el análisis produzca variantes canónicas, aun claramente definidas, pero que no tengan sentido para el investigador. La función de este procedimiento es tan sólo el manipular las intercorrelaciones entre las variables para ver si existe algún tipo particular de patrón en los datos. Es asunto del investigador explicar por qué existe ese patrón y el sentido teórico que éste tenga.

Se recomienda que el investigador solicite siempre las estadísticas que arrojan medias y desviaciones estándar de las variables incluidas en el análisis, así como la matriz de intercorrelaciones entre las mismas, para ayudarse en la interpretación de los resultados.

5.5 Interpretación de resultados

El estudio presentado en esta obra ejemplifica los análisis y la interpretación de los resultados. Se llevó a cabo una evaluación de los profesores que impartían la materia Psicología Social Introductoria. Se les pidió a 262 alumnos que evaluaran al profesor, al programa de la materia y al sistema de enseñanza. La evaluación se llevaba a cabo a través de la aplicación de un conjunto de 24 escalas bipolares, tipo diferencial semántico. Las escalas fueron sometidas al análisis factorial con rotación ortogonal varimax que dieron dos factores. El Factor I hizo referencia a las características estructurales del proceso de enseñanza-aprendizaje y el Factor II se refirió a las características personales del profesor. El primer factor abarcó a las siguientes escalas: el programa da una visión general del área, cubre puntos esenciales, contiene información actualizada, tiene secuencia pedagógica, es interesante, el sistema de enseñanza permite la

aplicación del conocimiento a situaciones prácticas, es creativo, entretenido, está sistematizado, logra la retención del conocimiento y es reforzante. El segundo factor, abarcó a las escalas del profesor: es claro, organizado, responsable, puntual, cumplido, sistemático y emplea ejemplos adecuados.

Ahora, fue interés del investigador el averiguar si las características de personalidad del profesor (Factor II) se combinan linealmente de manera que correlacionen o expliquen a las características estructurales del proceso de enseñanza aprendizaje (Factor I): sistema de enseñanza y programa de la materia constituidas en alguna combinación lineal específica. Para esta situación, fue conveniente emplear la correlación canónica.

De esta manera, se introdujeron como un subconjunto de variables a las escalas que constituyeron al Factor I y como otro subconjunto de variables a las constituyentes del Factor II, y se sometieron a un análisis de correlación canónica. El objetivo en esta ocasión fue únicamente didáctico, aunque se podría suponer que las características personales de un profesor pueden explicar el sistema de enseñanza y el manejo de un programa de una materia.

Los resultados muestran en primer lugar, las medias y desviaciones estándar de las diferentes escalas bipolares que constituyeron los factores, señalándose, también, el número de sujetos o casos que respondieron a las mencionadas escalas.

A continuación, aparecen los resultados de la correlación canónica. Estos se dan en dos partes. En la parte superior aparece la Tabla Sumaria de la correlación canónica.

En la primera columna aparece el número de variantes canónicas encontradas (siete). En la siguiente columna aparecen los valores Eigen asociados a esas variantes canónicas. En esta columna se observa que la primera variante canónica comparte el 59% de la varianza en el par de variantes correspondientes; el segundo par de variantes canónicas comparte el 20% de varianza aproximadamente; el tercer par de variantes canónicas tan sólo comparte el 9%. El resto de los pares de variantes canónicas comparten menos del 7% de la varianza.

En la tercera columna aparecen los valores de la correlación canónica entre los pares correspondientes de las variantes canónicas. Se puede observar que la primera correlación es de 0.77, lo que significa que está es la correlación entre los miembros del primer para de

variantes canónicas; la segunda es de 0.44 y la tercera de 0.30. Las restantes correlaciones son menores de 0.30.

En la cuarta columna aparecen los valores de Lambda de Wilk, a continuación, los de Ji-cuadrada, seguidos por sus grados de libertad y niveles de significancia asociados. Estos cuatro valores señalan la significatividad estadística de los pares de variantes canónicos encontrados. Se puede ver entonces, que sólo tres de los siete pares de variantes canónicas tienen niveles de significancia iguales o menores a 0.05. Los cuatro pares de variantes canónicos restantes tienen niveles de significancia mayores a 0.05. Esto significa que se encontraron tres pares de combinaciones lineales con alta correlación entre sí.

Se procede entonces a ver los resultados y se observa que aparecen tres pares de variantes canónicas, con sus coeficientes correspondientes al primer y al segundo conjunto de variables.

Si se observa el primer par de variantes canónicas, con sus coeficientes correspondientes al primero y segundo conjunto de variables, se señala que:

a) El programa sea entretenido, esté sistematizado y logre la retención del conocimiento se relaciona con que el profesor sea claro, sistemático y emplee ejemplos adecuados.

En el segundo par de variantes canónicas se puede observar:

a) El que el programa sea entretenido, esté sistematizado y logre la retención del conocimiento se relaciona con que el profesor sea claro, sistemático y emplee ejemplos adecuados.

b) El que la información del programa sea actualizada y éste tenga secuencia pedagógica tiene que ver con la responsabilidad del profesor.

Por último, en el tercer par de variantes canónicas se observa:

a) Que si el sistema de enseñanza o programa es creativo y entretenido, esto tiene que ver con la puntualidad del profesor, y

b) El que el programa cubra puntos esenciales y el sistema de enseñanza logre la retención del conocimiento, se relaciona con la organización y responsabilidad del profesor, así como con el hecho de que emplee ejemplos adecuados.

El investigador deberá señalar en la discusión de resultados la pertinencia teórica de los hallazgos encontrados. En el presente ejemplo, la intención fue didáctica, y no teórica.

5.6 Reporte de un análisis de correlación canónica

Cuando se reporta un análisis de correlación canónica, se deberá incluir la siguiente información.

1. Descripción breve, conceptual y de medición respecto a todas y cada una de las variables que se incluyen en el análisis.

2. Fundamentación teórico metodológica referida a la clasificación de las variables incluidas en los dos subconjuntos requeridos.

3. Características de la muestra empleada en el estudio, así como su forma de selección y tamaño.

4. Tabla de medias y desviaciones estándar de las variables estudiadas.

5. Tabla Sumaria del Análisis de correlación canónica.

6. Matrices de coeficientes de variantes canónicas significativas.

7. Análisis, interpretación y discusión de los resultados encontrados. En esta sección se requiere que se haga una descripción breve y somera de los resultados de la Tabla Sumario y los pares de variantes canónicas extraídos (semejante a la antes expuesta). También se deberá relacionar los resultados encontrados, desde el punto de vista teórico, con otros estudios que se puedan relacionar al presente, sus resultados y recomendaciones. Por último, se deberá discutir los hallazgos, a la luz de los modelos o del modelo teórico que le pueden dar un significado congruente con el conocimiento existente en el área de estudio.

Referencias bibliográficas

Amat, J. (2016). *Análisis discriminante lineal (LDA) y Análisis discriminante cuadrático (QDA).* Disponible en: https://rpubs.com/Joaquin_AR/233932.

Cuadras, C.M. (1989). Distancias Estadísticas - Instituto Nacional de Estadística. *Estadística Española, 30(119)*, 295 – 378. Disponible en: www.ine.es/ss/Satellite?

Definiciones (2016). *Definición de espurio*. Disponible en: www.definiciones-de.com/Definicion/de/espurio.php.

Diccionario Universal (2018). *Discriminibilidad*. Disponible en: http://diccionario-universal.com/definitions/?spanish_word= discriminability

Ferrando, P.J. & Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación en psicología. *Papeles del Psicólogo, 31(1)*, 18 – 33. Disponible en: http://www.redalyc.org/articulo.oa? id=77812441003.

Hernández-Sampieri, R., Fernández-Collado, C. & Baptista-Lucio, M.P. (2010). *Metodología de la investigación*. México: Mcgraw-HILL / Interamericana Editores.

McDaniel, C.J. & Gates, R. (1999). *Investigación de Mercados Contemporánea*. México: Thomson Learning.

Tabachnick, B. G. & Fidell, L. S. (2001). *Using multivariate statistics*. Needham, MA: Allyn & Bacon.

DISEÑOS MULTIVARIADOS DE INVESTIGACIÓN EN LAS CIENCIAS SOCIALES

LUCY MARÍA REIDL MARTÍNEZ
RAQUEL DEL SOCORRO GUILLÉN RIEBELING

Este libro tiene por objeto presentar a los estudiosos de las ciencias sociales en general y de la Psicología en particular, una versión simplificada y explicativa del uso de los diseños multivariados de investigación para abordar problemas y/o temas complejos en campos de la Psicología -clínica, educativa, social, salud, laboral, entre otros, y de las ciencias sociales. La comprensión de la aplicación del método, y así, poder seleccionar entre los diferentes diseños, el más adecuado al problema de investigación; la fundamentación de esa selección; cómo poder interpretar los resultados obtenidos por el paquete SPSS (Statistical Package for the Social Sciences); así como poder elaborar proyectos que requieran de este tipo de diseños. En síntesis, se pretende demostrar por medio de ejemplos desarrollados en su totalidad, la facilidad del empleo de este tipo de diseños, así como la riqueza de la información que proporcionan.

Facultad de Estudios Superiores Zaragoza,
Campus I. Av. Guelatao No. 66 Col. Ejército de Oriente,
Campus II. Batalla 5 de Mayo s/n Esq. Fuerte de Loreto.
Col. Ejército de Oriente.
Iztapalapa, C.P. 09230 Ciudad de México.
Campus III. Ex fábrica de San Manuel s/n,
Col. San Manuel entre Corregidora y Camino a Zautla,
San Miguel Contla, Santa Cruz Tlaxcala.

http://www.zaragoza.unam.mx

9786073027168