

# **MODELO ACL Y MODELOS CON ERRORES DE MEDICIÓN: TÉCNICAS PARA EL ANÁLISIS DE ESCALABILIDAD DE VARIABLES CATEGÓRICAS**

*Dr. Marcelino Sánchez Rivero*

Departamento de Economía Aplicada y Organización de Empresas  
Universidad de Extremadura

## **RESUMEN:**

Una de las técnicas que permite clasificar individuos a partir de criterios subjetivos ( como opiniones, actitudes, preferencias, etc. ) es el modelo de Análisis de Clases Latentes, en el que tanto la variable latente analizada como las variables indicadoras de la misma son de naturaleza categórica. En principio, el modelo ACL no plantea ninguna hipótesis sobre las variables observadas que definen a la variable latente. Sin embargo, estas variables pueden guardar un cierto orden en su “grado de dificultad”, lo que habilitaría un análisis de escalabilidad de dichas variables. En estas circunstancias, será preciso imponer determinadas restricciones sobre los parámetros del modelo ACL, lo que provocará la aparición de nuevos modelos como el modelo de Proctor, el modelo de porcentaje de error específico de cada variable observada, el modelo de porcentaje de error específico de cada posición de escala y el modelo de Distancia Latente de Lazarsfeld.

En la presente comunicación se pretende demostrar la versatilidad de estos modelos con errores de medición para analizar la escalabilidad de cuatro variables relacionadas con un fenómeno turístico cada vez más habitual: el fraccionamiento de las vacaciones en dos o más períodos de tiempo.

## **PALABRAS CLAVE:**

Análisis de Clases Latentes, reproducibilidad, escalabilidad, restricciones de igualdad.

## 1. INTRODUCCIÓN.

El análisis de escala se ha erigido tradicionalmente en una eficaz herramienta para describir la estructura que subyace bajo un conjunto de datos. El empleo de escalas unidimensionales desde una óptica eminentemente exploratoria ha permitido en múltiples ocasiones clasificar individuos en función de la posición que ocupan en la citada escala. En otras ocasiones, el análisis de escala se ha empleado con un carácter confirmatorio para verificar la conformidad existente entre un conjunto de datos y una hipótesis previamente formulada.

A pesar del mayor poder explicativo y de la mayor flexibilidad de las escalas multidimensionales, las escalas unidimensionales han sido utilizadas con gran profusión en la investigación de fenómenos sociales. Las razones de este frecuente uso se encuentran en su mayor facilidad de interpretación, en su carácter previo a las escalas multidimensionales y en su mejor adaptación a las teorías sociales. Las escalas unidimensionales se construyen a partir de varios *items* ( o variables ) que representen correctamente un determinado concepto teórico y que eliminen parte de los errores aleatorios obtenidos como consecuencia de la utilización de un único ítem para construir la escala.

La construcción de escalas de medidas puede llevarse a cabo mediante modelos determinísticos o mediante modelos probabilísticos. En un modelo determinístico, la probabilidad de una respuesta determinada a una variable de la escala sólo puede ser 0 o 1. Por el contrario, si el modelo es probabilístico, dicha probabilidad puede variar entre 0 y 1, sin que esté restringida a uno de estos dos valores extremos.

Han sido muchas las escalas unidimensionales propuestas para analizar el comportamiento de los individuos ( escala de Thurstone, escala de Likert, ... ), pero posiblemente sea la escala de Guttman (1950) la más difundida en la investigación social. Guttman considera que, a partir de un conjunto de  $k$  variables o ítems de naturaleza dicotómica ( 1 = respuesta afirmativa; 2 = respuesta negativa ), es posible construir una escala con  $k+1$  posiciones, de forma que la primera posición de escala representa una respuesta afirmativa (1,1,1, ..., 1) a todas las variables, mientras que la última posición de escala representa una respuesta negativa (2,2,2, ..., 2) a la totalidad de variables. Las posiciones intermedias reflejan, gradualmente, respuestas negativas a las variables empleadas para construir la escala, las cuales están ordenadas de menor a mayor dificultad de observación de una respuesta positiva.

En la escala de Guttman, se puede asignar un valor a cada posición de escala. Así, a la primera posición de escala se le puede asignar el valor  $k$  (  $k$  respuestas afirmativas a las variables de la escala); a la segunda posición de escala el valor  $k-1$  ( esto es,  $k-1$  respuestas afirmativas a las  $k-1$  primeras variables y 1 respuesta negativa a la última variable ); a la tercera posición el valor  $k-2$  ( 2 respuestas negativas a las dos últimas variables y  $k-2$  respuestas afirmativas a las  $k-2$  primeras variables ); y así sucesivamente hasta llegar a la última posición de escala, a la que se le asignará el valor 0 ( 0 respuestas afirmativas a las variables ).

A pesar de sus buenas intenciones, la escala de Guttman presenta graves inconvenientes, como su incapacidad para trabajar con grandes conjuntos de ítems, la asunción de inexistencia de datos perdidos o su limitación a ítems únicamente dicotómicos. Pero es, sin lugar a dudas, su carácter determinista el que ha llevado a varios autores a proponer modelos alternativos, en los que la existencia de error de medición se convierte en parte esencial de dichos modelos, convirtiéndose de esta forma

en modelos con enfoque probabilístico. El objetivo de esta comunicación es presentar algunas de estas alternativas y demostrar empíricamente su validez en el análisis de escala.

## 2. EL MODELO DE ANÁLISIS DE CLASES LATENTES.

La investigación en las ciencias sociales se caracteriza, entre otras cosas, por la gran cantidad de conceptos, especialmente los relacionados con teorías sociales y de comportamiento, que no pueden ser cuantificados de una forma directa. Así, el status socioeconómico de un individuo, su ideología política o su actitud ante cualquier fenómeno son algunos ejemplos de las muchas variables que, en el ámbito social, no pueden ser objeto de una medición directa. La imposibilidad de observación de estas variables determina que su conocimiento empírico sólo pueda adquirirse a través de otras variables, susceptibles de observación y medición, que contengan información relevante sobre aquellas.

Surge de esta forma la llamada *observación indirecta*, en la que variables teóricas que no son directamente observables ( o **variables latentes** ) son analizadas mediante un conjunto de variables observables ( o **variables manifiestas** o **indicadores** ) que cuantifican la intensidad con que aquellas se presentan en la población objeto de estudio.

La asociación entre las variables latentes y sus indicadores es posible gracias a los llamados **modelos de variables latentes**. De esta amplia familia de modelos, en la presente comunicación se abordará el caso particular en el que todas las variables que intervienen en el análisis ( tanto latentes como observadas ) son de naturaleza categórica. El modelo adecuado para analizar este caso se conoce con el nombre de **modelo de Análisis de Clases Latentes** o modelo ACL. Este modelo no sólo identifica variables inobservadas a partir de un conjunto de indicadores de las mismas, sino que también clasifica a la población objeto de estudio en diferentes segmentos ( clases latentes ). Estos segmentos, que son excluyentes y exhaustivos, agrupan a los individuos de la población que poseen las mismas características y definen las diferencias existentes entre las clases identificadas.

Por otra parte, todo el planteamiento del modelo ACL gira en torno a una **hipótesis inicial** fundamental. Esta hipótesis establece que los indicadores de una variable latente son condicionalmente independientes dada una clase de dicha variable. En otros términos, las variables manifiestas están correlacionadas entre sí, pero esta correlación desaparece cuando se fija una categoría de la variable latente que aquellas pretenden medir. Esta hipótesis se conoce con el nombre de **independencia local**, y es básica para entender el modelo ACL, ya que significa que la asociación existente entre los indicadores está motivada únicamente por la asociación directa existente entre cada uno de estos indicadores por separado y la variable latente objeto de estudio.

En cualquier caso, las implicaciones de la independencia local van más allá de lo expuesto hasta ahora. De hecho, la independencia local explica también el hecho de que cada categoría ( clase ) de la variable latente posea probabilidades condicionadas de respuesta a los indicadores que son muy diferentes a las probabilidades condicionadas asociadas a otra categoría de la misma variable latente, y que los individuos encuadrados en una misma categoría tengan la misma probabilidad de respuesta a los indicadores. Por tanto, la independencia local es la responsable de que exista un gran *homogeneidad intragrupal* y, al mismo tiempo, una apreciable *heterogeneidad intergrupala*.

Sobre la base de esta hipótesis inicial, la probabilidad de que un individuo se encuadre en la clase  $t$  de la variable latente  $X$  y en las categorías  $i, j$  y  $k$ , respectivamente, de las variables observadas  $A, B$  y  $C$  ( y que se representará por  $\Pi_{ijkt}^{ABCX}$  ) se puede expresar como el producto de las probabilidades de responder a las variables observadas en una categoría dada de la misma, condicionadas a la pertenencia del individuo a la clase  $t$  de la variable latente  $X$ , es decir:

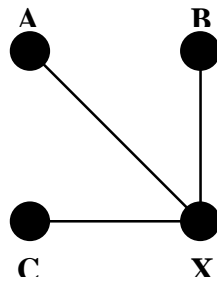
$$\Pi_{ijkt}^{ABCX} = \Pi_{it}^{\bar{A}X} \times \Pi_{jt}^{\bar{B}X} \times \Pi_{kt}^{\bar{C}X} \times \Pi_t^X \quad (2.1.)$$

siendo  $\Pi_{it}^{\bar{A}X}$  ( $\Pi_{jt}^{\bar{B}X}$ ,  $\Pi_{kt}^{\bar{C}X}$ , respectivamente) la probabilidad condicionada de que un individuo que pertenece a la clase  $t$  de la variable latente  $X$  responda en el nivel  $i$  ( $j, k$ , respectivamente) de la variable  $A$  ( $B, C$ , respectivamente).

Aunque el modelo (2.1.) está expresado en términos de probabilidades condicionadas, podría reformularse en términos log-lineales, ya que un modelo ACL es, en realidad, un modelo logarítmico-lineal en el que una de las variables categóricas ( la variable  $X$  ) es latente. Por consiguiente, el logaritmo de las probabilidades  $\Pi_{ijkt}^{ABCX}$  puede descomponerse en un efecto global, en los efectos primarios de las variables observadas y de la variable latente y en los efectos de interacción de segundo orden entre cada indicador y la variable latente a la que se asocia, es decir:

$$\log \Pi_{ijkt}^{ABCX} = \tau_0 + \tau_i^A + \tau_j^B + \tau_k^C + \tau_t^X + \tau_{it}^{AX} + \tau_{jt}^{BX} + \tau_{kt}^{CX} \quad (2.2.)$$

y cuya representación gráfica ( que se muestra en la *Figura 1* ) permite apreciar con claridad la hipótesis de independencia local de la que parte el modelo.



*Figura 1*

De la expresión (2.1.) se desprende que los parámetros de un modelo ACL son las probabilidades de clases latentes ( $\Pi_t^X$ ) y las probabilidades condicionadas correspondientes a las variables observadas ( $\Pi_{it}^{\bar{A}X}$ ,  $\Pi_{jt}^{\bar{B}X}$  y  $\Pi_{kt}^{\bar{C}X}$ ).

Las probabilidades de clases latentes, que dado el carácter excluyente y exhaustivo de las categorías de la variable  $X$  deben verificar que  $\sum_{t=1}^T \Pi_t^X = 1$ ,

determinarán el tamaño relativo de cada una de las  $T$  clases de la variable latente. El valor mínimo de  $T$  debe ser 2, ya que una variable con una sola clase latente es sinónimo de independencia mutua entre todas las variables observadas. Estas probabilidades de clases latentes proporcionan información muy valiosa sobre la población objeto de estudio, ya que si todas arrojan valores similares, se estará ante una población homogéneamente distribuida, mientras que si dichas probabilidades toman valores muy dispares, la población analizada poseerá una distribución heterogénea, en la que se conjugan segmentos altamente representativos con otros escasamente significativos.

Por su parte, las probabilidades condicionadas definen las características propias de cada una de las clases de la variable latente, ya que cuantifican la proximidad o lejanía de los individuos encuadrados en cada grupo respecto a cada categoría concreta de las variables observadas. Cuanto mayor disparidad exista entre las probabilidades condicionadas ( referidas a un mismo indicador ) de cada clase latente, mayor será la heterogeneidad existente entre los grupos que define la variable latente.

Dentro de cada clase latente, se verifica que la suma de las probabilidades condicionadas asociadas a cada variable observada debe ser igual a 1. Para la expresión (2.1.) se tendrá que:

$$\sum_{i=1}^I \Pi_{it}^{\bar{A}X} = 1 \qquad \sum_{j=1}^J \Pi_{jt}^{\bar{B}X} = 1 \qquad \sum_{k=1}^K \Pi_{kt}^{\bar{C}X} = 1$$

Si una variable latente posee T clases, y dado que existirá siempre una probabilidad de clases latentes que podrá calcularse a partir de las demás ( como consecuencia de que la suma de las T probabilidades debe ser igual a 1 ), será necesario estimar un total de (T-1) probabilidades de clases latentes. De forma similar, si las variables observadas A, B y C poseen I, J y K categorías, respectivamente, y teniendo presente que existirá una probabilidad condicionada para cada variable observada y en cada clase latente que podrá calcularse a partir de las demás, el número total de probabilidades condicionadas a estimar será igual a (I-1)T + (J-1)T + (K-1)T.

La estimación de las (I-1)T + (J-1)T + (K-1)T probabilidades condicionadas y de las (T-1) probabilidades de clases latentes se lleva a cabo mediante procedimientos iterativos de máxima verosimilitud que presentan grandes similitudes con los empleados para la estimación de modelos log-lineales. Una técnica muy empleada es el llamado **algoritmo EM**, propuesto por Goodman en 1974 y posteriormente desarrollado por Dempster, Laird y Rubin en 1977. Mediante este algoritmo, y a través de un proceso iterativo, las estimaciones sucesivas de los parámetros del modelo ACL se obtienen resolviendo un sistema de ecuaciones de verosimilitud. El proceso iterativo comienza con unos valores iniciales para las probabilidades de clases latentes y para las probabilidades condicionadas, que se designarán por  $\Pi_t^{X(0)}$ ,  $\Pi_{it}^{\bar{A}X(0)}$ ,  $\Pi_{jt}^{\bar{B}X(0)}$  y  $\Pi_{kt}^{\bar{C}X(0)}$ . Estos valores iniciales posibilitan el cálculo de un valor inicial para las probabilidades estimadas  $\hat{\Pi}_{ijkt}^{ABCX}$ ,  $\hat{\Pi}_{ijk}$  y  $\hat{\Pi}_{ijkt}^{ABC\bar{X}}$  ( $\Pi_{ijkt}^{ABCX(0)}$ ,  $\Pi_{ijk}^{(0)}$  y  $\Pi_{ijkt}^{ABC\bar{X}(0)}$ ). A su vez, estos últimos valores permiten obtener un nuevo valor para las probabilidades de clases latentes y para las probabilidades condicionadas ( $\Pi_t^{X(1)}$ ,  $\Pi_{it}^{\bar{A}X(1)}$ ,  $\Pi_{jt}^{\bar{B}X(1)}$  y  $\Pi_{kt}^{\bar{C}X(1)}$ ), y así sucesivamente. Este proceso iterativo continuaría de forma indefinida hasta que se alcance un número predeterminado de iteraciones o hasta que la diferencia entre dos estimaciones sucesivas sea inferior a una cantidad previamente fijada y llamada **nivel de convergencia**.

Para que las estimaciones máximo-verosímiles de los parámetros del modelo ACL se puedan obtener es necesario que el modelo esté **localmente identificado**, lo que significa que los parámetros del modelo deben estar unívocamente determinados por las probabilidades estimadas mediante el algoritmo EM. Una condición necesaria, aunque no suficiente, para que el modelo ACL esté localmente identificado es que el número de grados de libertad del test empleado para contrastarlo no sea negativo. Por su parte, Goodman ( 1974a ) propuso un método que es condición necesaria y suficiente para

verificar la identificabilidad local de un modelo ACL. Este método consiste en obtener las derivadas parciales de las funciones  $\Pi_{ijk}$  con respecto a todos y cada uno de los parámetros del modelo y construir una matriz con estas derivadas parciales, que tendrá  $IJK-1$  filas y  $(I+J+K-2)T-1$  columnas. De esta forma, la condición necesaria y suficiente para que el modelo esté localmente identificado es que el rango de dicha matriz sea igual al número de columnas de la misma. En el caso de que el modelo no estuviese localmente identificado, será preciso imponer determinadas restricciones sobre los parámetros del modelo ya que, de esta forma, el número total de parámetros a estimar disminuirá.

En otro orden de cosas, la bondad de ajuste del modelo ACL a los datos observados puede verificarse mediante el test estadístico de Pearson ( $X^2$ ) o mediante el test estadístico de la razón de verosimilitud ( $G^2$ ), aunque éste último será preferible al primero, dada su propiedad de particionabilidad (propiedad muy útil para comparar diferentes modelos). En consecuencia, el modelo ACL se contrastará utilizando el siguiente test, en el que se comparan las frecuencias esperadas estimadas,  $\hat{m}_{ijk}$ , con las frecuencias observadas de la tabla de contingencia,  $n_{ijk}$ :

$$G^2 = 2 \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K n_{ijk} \log \frac{n_{ijk}}{\hat{m}_{ijk}} \quad (2.3.)$$

Los grados de libertad de este test se calculan como la diferencia entre el número de casillas de la tabla de contingencia menos 1 y el número total de parámetros a estimar, de forma que para una tabla de contingencia de 3 variables con I, J, y K categorías, respectivamente, y dada una variable latente con T clases, los grados de libertad necesarios para contrastar el modelo ACL serían:

$$(IJK-1) - [(T-1)+T(I-1)+T(J-1)+T(K-1)] = IJK - (I+J+K-2)T$$

Una vez estimado y contrastado el modelo, deberá procederse a ubicar a los individuos de la muestra en las diferentes clases de la variable latente, de forma que, dado que los individuos que poseen la misma modalidad de respuesta a las variables observadas pertenecerán a una misma clase latente, dicha asignación de clases se realizará para cada modalidad de respuesta. Para ello, la probabilidad condicionada de que un individuo pertenezca a la clase  $t$  de la variable X, dadas las respuestas a las variables categóricas observadas A, B y C en los niveles  $i, j$  y  $k$ , respectivamente, se calculará de la siguiente forma:

$$\Pi_{ijk}^{ABC\bar{X}} = \frac{\Pi_{ijkt}^{ABCX}}{\sum_{t=1}^T \Pi_{ijkt}^{ABCX}} \quad (2.4.)$$

La anterior probabilidad debe calcularse para la totalidad de clases de la variable latente X, de forma que a los individuos cuya modalidad de respuesta sea  $i,j,k$  se les asignará aquella clase que arroje un mayor valor de  $\Pi_{ijk}^{ABC\bar{X}}$ .

Por consiguiente, la asignación de los individuos a las clases latentes tiene un carácter probabilístico, lo que supone la existencia de un cierto error en dicha asignación. Para cuantificar la magnitud de este error, Clogg (1979) propone la “proporción correctamente clasificada” y el coeficiente  $\lambda$ . La primera de estas medidas se calcula a través de la siguiente expresión:

$$E_1 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \left( \Pi_{ijkt}^{ABC\bar{X}} \times p_{ijk} \right)$$

donde  $p_{ijk}$  es la proporción observada de individuos en la casilla  $(i,j,k)$  de la tabla.  $E_1$  tomará valores comprendidos entre 0 y 1, de forma que cuanto mayor sea su valor más fuerte será la relación existente entre la variable latente y las variables observadas y, en consecuencia, más correcta será la asignación de los individuos a las clases latentes.

El coeficiente  $\lambda$ , por su parte, se define de la siguiente forma:

$$\lambda = \frac{E_2 - E_3}{E_2}$$

donde  $E_2 = 1 - \Pi_{t'}^X$  es el porcentaje de error que resultaría de asignar todos los individuos a la clase latente  $(t')$  cuya probabilidad sea más alta, mientras que  $E_3$  es la proporción de individuos incorrectamente clasificados, es decir,  $E_3 = 1 - E_1$ . El coeficiente  $\lambda$  también toma valores comprendidos entre 0 y 1, de forma que puede afirmarse que cuanto más se aproxime a este último valor, menor será el error global cometido al realizar la asignación de clases.

Aunque las dos medidas anteriores dependen de parámetros desconocidos, se pueden utilizar las estimaciones máximo-verosímiles de dichos parámetros para obtener estimaciones muestrales de las citadas medidas.

Aunque el modelo ACL constituye una técnica estadística muy útil para realizar análisis exploratorios, uno de sus valores añadidos es la posibilidad de contrastar hipótesis sobre la naturaleza de la variable latente y, por extensión, sobre la teoría que subyace en la misma. Este análisis de carácter **confirmatorio** se efectúa imponiendo restricciones sobre las probabilidades de clases latentes y/o sobre las probabilidades condicionadas del modelo.

Las restricciones que pueden imponerse sobre los parámetros del modelo ACL son de dos tipos: restricciones de **igualdad** y restricciones de **valor**. Las restricciones de igualdad se utilizan para contrastar si dos o más probabilidades condicionadas o de clases latentes toman el mismo valor. Aunque estas restricciones de igualdad pueden imponerse sobre los tamaños relativos de las clases latentes, nos interesa especialmente las restricciones de igualdad que se imponen sobre las probabilidades condicionadas del modelo, como las siguientes:

$$H_0^{(2.1)}: \Pi_{i1}^{\bar{A}X} = \Pi_{i2}^{\bar{A}X} = \dots = \Pi_{iT}^{\bar{A}X}$$

$$H_0^{(2.2)}: \Pi_{1t}^{\bar{A}X} = \Pi_{1t}^{\bar{B}X} = \dots = \Pi_{1t}^{\bar{Z}X}$$

El primer conjunto de restricciones establece que la probabilidad de que un individuo responda a la variable A en la categoría  $i$  es siempre la misma sea cual sea la clase de la variable latente X a la que pertenezca dicho individuo. Por su parte, el segundo conjunto de restricciones indica que la probabilidad de que los individuos pertenecientes a la clase  $t$  de la variable X se sitúen en la categoría 1 de las variables observadas será la misma para todas estas variables observadas.

Las restricciones de valor permiten contrastar hipótesis relacionadas, por ejemplo, con la importancia de una variable observada sobre las clases latentes o con el

tamaño relativo de una clase latente determinada. Centrando nuevamente la atención en las restricciones de valor sobre las probabilidades condicionadas y obviando las que se pueden imponer sobre las probabilidades de clases latentes ( que son especialmente útiles para contrastar la equiprobabilidad de las clases latentes o la mayor o menor heterogeneidad de la población objeto de estudio ), debemos señalar que, aunque se puede contrastar si una probabilidad condicionada toma un valor  $p$  (  $0 \leq p \leq 1$  ), generalmente las restricciones de valor sobre las probabilidades condicionadas suelen formularse de una de las dos formas siguientes:

$$H_0^{(2.3.)}: \Pi_{ii}^{\bar{A}X} = 0$$

$$H_0^{(2.4.)}: \Pi_{ii}^{\bar{A}X} = 1$$

La diferencia esencial entre ambas restricciones es que, en el primer caso, se está imponiendo una única restricción sobre una probabilidad condicionada de la variable observada A, mientras que en el segundo caso se está imponiendo una restricción a la totalidad de probabilidades condicionadas de la variable A, ya que el hecho de que se restrinja al valor 1 la probabilidad condicionada asociada a la categoría  $i$  de la citada variable determina que las probabilidades condicionadas asociadas a las  $I-1$  categorías restantes de la variable A deben ser iguales a 0.

La imposición de restricciones, tanto de igualdad como de valor, sobre los parámetros del modelo ACL afecta al proceso de estimación de los mismos, ya que en dicho proceso se liberarán tantos grados de libertad como parámetros se restrinjan. Estos grados de libertad se pueden emplear para determinar si las restricciones han mejorado sustancialmente el ajuste del modelo ACL. Para ello, bastará comparar la bondad de ajuste del modelo restringido con la del modelo no restringido, de forma que si el incremento en el valor de  $G^2$  es pequeño en relación con el incremento de los grados de libertad, se podrá admitir que las restricciones impuestas al modelo han mejorado su ajuste a los datos observados.

### 3. MODELOS CON ERRORES DE MEDICIÓN.

Las aplicaciones del modelo ACL no se limitan exclusivamente a la identificación de variables latentes y a la segmentación de la población objeto de estudio, sino que se extienden también al análisis de escalabilidad de un conjunto de variables observadas.

Esta idea de escalabilidad tiene su origen en la llamada **escala de Guttman**, que considera que múltiples variables dicotómicas pueden ser ordenadas en función de la “dificultad” para construir una escala unidimensional de un fenómeno latente. Esta ordenación, que es única, presupone que si un individuo responde negativamente a una variable determinada, deberá responder también negativamente a las restantes variables de la escala.

Este enfoque de Guttman ha sido muy criticado por su carácter determinista y por la naturaleza *ad hoc* de los criterios de evaluación de la escala, ya que los mismos se basan en reglas empíricas que poseen una escasa justificación estadística. Para superar estas críticas se han propuesto algunas alternativas probabilísticas que incorporan nociones de errores de medición y que proporcionan criterios de evaluación basados en conceptos probabilísticos.



El análisis de escalabilidad de un conjunto de variables dicotómicas debe iniciarse estableciendo la ordenación de la mismas mediante la teoría o las investigaciones previas, o bien mediante la inspección de la distribución marginal de una respuesta afirmativa a cada una de las citadas variables. Así, aunque un conjunto de  $k$  variables puede ordenarse de  $k!$  formas distintas, sólo una de estas posibles ordenaciones implicará una graduación ( de menor a mayor ) en la dificultad de observar una respuesta afirmativa. Según lo expuesto, cuando se trabaja con 4 variables categóricas dicotómicas, y 1 denota una respuesta afirmativa y 2 una respuesta negativa, el planteamiento de Guttman admite como válidas, una vez ordenadas las 4 variables, únicamente 5 modalidades de respuesta (  $k+1$  con carácter general ), que serían (1,1,1,1), (1,1,1,2), (1,1,2,2), (1,2,2,2) y (2,2,2,2), mientras que las 11 modalidades de respuesta restantes son consideradas como respuestas incorrectas.

Partiendo de este planteamiento inicial, se puede determinar la escalabilidad de un conjunto de variables observadas a través de los coeficientes de reproducibilidad y de escalabilidad. El primero de estos coeficientes se define de la siguiente forma:

$$R = 1 - \frac{E_1}{k \times n} \quad (3.1.)$$

donde  $E_1$  es la suma de las frecuencias observadas de las modalidades de respuesta que se consideran incorrectas ( es decir, aquellas con errores de escala );  $k$  es el número de variables observadas empleadas en la escala; y  $n$  es el tamaño muestral total. A partir del anterior coeficiente, Guttman determinó que un conjunto de variables podría considerarse como escalable si  $R$  es superior a 0,9, esto es, si los errores de medición no superan el 10 % del total de respuestas.

Por su parte, el coeficiente de escalabilidad se obtiene mediante la siguiente expresión:

$$S = 1 - \frac{E_1}{E_2} \quad (3.2.)$$

donde  $E_2$  es la suma de las frecuencias marginales no modales de todas y cada una de las variables observadas. Como puede apreciarse, el mínimo valor del coeficiente de escalabilidad será 0, cuando los errores de medición coincidan con los errores marginales, mientras que su valor máximo será 1, en el caso de que no existan errores de medición. Guttman consideró que valores de  $S$  superiores a 0,6 determinan que las variables observadas pueden ser ordenadas en una escala unidimensional y acumulativa.

Para salvar los inconvenientes derivados de la naturaleza determinista de la escala de Guttman, han surgido en los últimos años varias propuestas alternativas que incluyen el error de respuesta como parte integral del modelo. Estas alternativas asumen que existe un determinado error de medición para cada una de las variables observadas y que, por tanto, ninguna variable puede considerarse una medida perfecta, puesto que las respuestas asociadas a dichas variables no tienen un carácter determinístico, sino probabilístico. Como podrá apreciarse a continuación, estos modelos alternativos son simplemente casos particulares del modelo de Análisis de Clases Latentes.

Las principales alternativas probabilísticas a la escala de Guttman han sido las siguientes:

### 3.1. Modelo de Proctor ( Proctor, 1970 ):

Proctor considera que la escala posee porcentajes de error homogéneos para todas las variables observadas y para todas las posiciones de escala. El modelo propuesto por Proctor supone que las  $k+1$  posiciones de escala están representadas por las  $k+1$  clases latentes de un modelo ACL restringido. Así, si el porcentaje de error para un conjunto de 4 variables observadas es de un 10 %, todo individuo situado en la posición de escala 2 (1,1,1,2) tendrá una probabilidad de un 90 % de responder correctamente a las tres primeras variables y una probabilidad de un 10 % de responder de forma incorrecta a la última variable. En consecuencia, a diferencia de la escala propuesta por Guttman ( que considera que, en la posición de escala anterior, todo individuo tiene una probabilidad de un 100 % de responder correctamente a las tres primeras variables y una probabilidad del 0 % de responder de forma incorrecta a la última variable ), el modelo de Proctor admite que las respuestas dadas en las distintas posiciones de escala están sujetas a error.

Si se consideran 4 variables observadas ( A, B, C y D ) y una variable latente X, para obtener el modelo de Proctor, bastará imponer al modelo ACL las siguientes restricciones de igualdad sobre sus probabilidades condicionadas:

$$\begin{aligned}\pi_{11}^{\bar{A}X} &= \pi_{11}^{\bar{B}X} = \pi_{11}^{\bar{C}X} = \pi_{11}^{\bar{D}X} = \\ &= \pi_{12}^{\bar{A}X} = \pi_{12}^{\bar{B}X} = \pi_{12}^{\bar{C}X} = \pi_{22}^{\bar{D}X} = \\ &= \pi_{13}^{\bar{A}X} = \pi_{13}^{\bar{B}X} = \pi_{23}^{\bar{C}X} = \pi_{23}^{\bar{D}X} = \\ &= \pi_{14}^{\bar{A}X} = \pi_{24}^{\bar{B}X} = \pi_{24}^{\bar{C}X} = \pi_{24}^{\bar{D}X} = \\ &= \pi_{25}^{\bar{A}X} = \pi_{25}^{\bar{B}X} = \pi_{25}^{\bar{C}X} = \pi_{25}^{\bar{D}X}\end{aligned}$$

La estimación de este modelo requiere únicamente 1 grado de libertad para estimar las probabilidades condicionadas y  $k$  grados de libertad para estimar las probabilidades de clases latentes. Mediante el test  $G^2$  se podrá determinar si el modelo de Proctor se ajusta satisfactoriamente a los valores observados. En el caso de que el ajuste no sea aceptable, Clogg y Sawyer ( 1981 ) apuntan algunos motivos que podrían explicar esta falta de ajuste, como una ordenación incorrecta de las variables observadas, unos porcentajes de error diferentes para cada variable, la existencia de una subpoblación en la que la ordenación de las variables sea completamente diferente a la propuesta o la posibilidad de que las posiciones de escala no representen una única dimensión latente.

### 3.2. Modelo de porcentaje de error específico de cada variable observada

( Clogg y Sawyer, 1981 ):

Ante la posibilidad de que no todas las variables posean el mismo porcentaje de error, este modelo asume que cada una de las  $k$  variables observadas que forman la escala posee un porcentaje de error diferente. Este modelo es también un caso particular del modelo ACL en el que se han impuesto las siguientes restricciones sobre las probabilidades condicionadas:

$$\begin{aligned}\pi_{11}^{\bar{A}X} &= \pi_{12}^{\bar{A}X} = \pi_{13}^{\bar{A}X} = \pi_{14}^{\bar{A}X} = \pi_{25}^{\bar{A}X} \\ \pi_{11}^{\bar{B}X} &= \pi_{12}^{\bar{B}X} = \pi_{13}^{\bar{B}X} = \pi_{24}^{\bar{B}X} = \pi_{25}^{\bar{B}X} \\ \pi_{11}^{\bar{C}X} &= \pi_{12}^{\bar{C}X} = \pi_{23}^{\bar{C}X} = \pi_{24}^{\bar{C}X} = \pi_{25}^{\bar{C}X} \\ \pi_{11}^{\bar{D}X} &= \pi_{22}^{\bar{D}X} = \pi_{23}^{\bar{D}X} = \pi_{24}^{\bar{D}X} = \pi_{25}^{\bar{D}X}\end{aligned}$$

Esta alternativa probabilística a la escala de Guttman precisa  $k$  grados de libertad para estimar los  $k$  porcentajes de error diferentes y otros  $k$  grados de libertad para estimar las probabilidades de clases latentes.

### 3.3. Modelo de porcentaje de error específico de cada posición de escala

( Clogg y Sawyer, 1981 ):

Este modelo conjetura que dentro de cada clase latente ( o posición de escala ) el porcentaje de error es el mismo para todas las variables observadas. En otros términos, para los individuos encuadrados en una posición de escala concreta, la probabilidad de responder de forma incorrecta a cada una de las variables observadas es la misma para la totalidad de variables. La forma de recoger estas conjeturas es mediante la imposición de las siguientes restricciones de igualdad sobre las probabilidades condicionadas del modelo ACL:

$$\begin{aligned}\Pi_{11}^{\bar{A}X} &= \Pi_{11}^{\bar{B}X} = \Pi_{11}^{\bar{C}X} = \Pi_{11}^{\bar{D}X} \\ \Pi_{12}^{\bar{A}X} &= \Pi_{12}^{\bar{B}X} = \Pi_{12}^{\bar{C}X} = \Pi_{22}^{\bar{D}X} \\ \Pi_{13}^{\bar{A}X} &= \Pi_{13}^{\bar{B}X} = \Pi_{23}^{\bar{C}X} = \Pi_{23}^{\bar{D}X} \\ \Pi_{14}^{\bar{A}X} &= \Pi_{24}^{\bar{B}X} = \Pi_{24}^{\bar{C}X} = \Pi_{24}^{\bar{D}X} \\ \Pi_{25}^{\bar{A}X} &= \Pi_{25}^{\bar{B}X} = \Pi_{25}^{\bar{C}X} = \Pi_{25}^{\bar{D}X}\end{aligned}$$

La estimación de un porcentaje de error para cada clase latente justifica la necesidad de emplear  $k+1$  grados de libertad, a los que habrá que sumar los  $k$  grados de libertad necesarios para estimar las probabilidades de clases latentes del modelo.

### 3.4. Modelo de Distancia Latente de Lazarsfeld ( Lazarsfeld, 1950a y 1950b ):

Posiblemente este modelo sea la alternativa probabilística a la escala de Guttman más elaborada ya que, a diferencia de los modelos anteriores ( que asumían que el porcentaje de error es el mismo para respuestas afirmativas y para respuestas negativas ), considera que el porcentaje de error para respuestas afirmativas difiere del porcentaje de error para respuestas negativas. Esta distinción entre porcentajes de error es válida para todas las variables observadas a excepción de las dos variables extremas, ya que Lazarsfeld y Henry ( 1968 ) demuestran que no es posible estimar más que un porcentaje de error para cada una de las variables extremas ( de aquí la necesidad de asegurarse previamente cuál es la variable más “fácil” de todas y cuál es la más “difícil” ). En consecuencia, las variables observadas intermedias poseerán dos porcentajes de error, uno para las respuestas afirmativas y otro para las respuestas negativas, mientras que las variables extremas tendrán un único porcentaje de error, que será válido tanto para las respuestas afirmativas como para las negativas. Además de establecer esta distinción entre los porcentajes de error, el modelo de Lazarsfeld considera también que dichos porcentajes son iguales para todas las clases latentes, pero diferentes para cada variable observada. En estas condiciones, las restricciones que habría que imponer sobre las probabilidades condicionadas del modelo ACL serían las siguientes:

$$\begin{aligned}\Pi_{11}^{\bar{A}X} &= \Pi_{12}^{\bar{A}X} = \Pi_{13}^{\bar{A}X} = \Pi_{14}^{\bar{A}X} = \Pi_{25}^{\bar{A}X} \\ \Pi_{11}^{\bar{B}X} &= \Pi_{12}^{\bar{B}X} = \Pi_{13}^{\bar{B}X} \\ \Pi_{24}^{\bar{B}X} &= \Pi_{25}^{\bar{B}X} \\ \Pi_{11}^{\bar{C}X} &= \Pi_{12}^{\bar{C}X}\end{aligned}$$

$$\Pi_{23}^{\bar{C}X} = \Pi_{24}^{\bar{C}X} = \Pi_{25}^{\bar{C}X}$$

$$\Pi_{11}^{\bar{D}X} = \Pi_{22}^{\bar{D}X} = \Pi_{23}^{\bar{D}X} = \Pi_{24}^{\bar{D}X} = \Pi_{25}^{\bar{D}X}$$

Según se deduce de las restricciones anteriores, el modelo de Distancia Latente requiere la estimación de  $2(k-2)$  probabilidades condicionadas y de  $k$  probabilidades de clases latentes. En consecuencia, serán necesarios  $3k-2$  grados de libertad para estimar los parámetros del modelo.

Cuando únicamente uno de los cuatro modelos presentados se ajuste de forma satisfactoria a los datos, es evidente que ese modelo será el elegido para representar la escalabilidad de un conjunto de variables observadas. El problema surge cuando dos o más modelos arrojan un ajuste satisfactorio que dificulta la elección del mejor modelo. En estos casos, Clogg y Sawyer ( 1981 ) aconsejan complementar el criterio de la bondad de ajuste con la comparación de la “proporción correctamente clasificada” y del coeficiente  $\lambda$  de los modelos implicados para facilitar la elección del modelo óptimo.

#### 4. ANÁLISIS DE ESCALABILIDAD APLICADO AL FRACCIONAMIENTO VACACIONAL.

Al objeto de constatar la utilidad de los modelos con errores de medición en el análisis de escalabilidad se ha utilizado el estudio 2.193 del Centro de Investigaciones Sociológicas sobre el “comportamiento de los españoles ante las vacaciones”, de Octubre de 1995. De toda la información incluida en este estudio, hemos seleccionado las cuatro preguntas siguientes ( cuyas respuestas posibles aparecen a continuación ):

- a) *¿ Ha tomado, o tomará, usted sus vacaciones principales en dos o más períodos de tiempo diferentes ?.*
  - Sí, he fraccionado, o fraccionaré, mis vacaciones principales en dos o más períodos de tiempo.
  - No, he disfrutado, o disfrutaré, de mis vacaciones principales de forma continuada.
- b) *¿ Se despreocupa de sus gastos durante las vacaciones ?.*
  - Sí, no me preocupo de los gastos y hago lo que me apetece.
  - No, cuido mucho mis gastos durante las vacaciones, tratando de no pasarme de lo previsto.
- c) *¿ Suele salir, al menos tres o cuatro días consecutivos, de su lugar de residencia por motivos que no sean de trabajo, durante la Semana Santa, las fiestas de Navidad o algún puente o fin de semana largo ?.*
  - Sí, siempre que puedo.
  - No, nunca o casi nunca.
- d) *¿ Es usted de la opinión de que todos los años hay que salir de vacaciones, por lo menos una vez, aunque ello suponga sacrificios considerables en la economía familiar o, por el contrario, cree que salir de vacaciones no es tan importante ?.*
  - Sí, creo que hay que salir de vacaciones todos los años, aunque ello suponga sacrificarse.
  - No, salir de vacaciones no es tan importante.

Estas cuatro preguntas serán designadas en lo sucesivo como variable **F**, variable **D**, variable **S** y variable **C**, respectivamente. Las frecuencias observadas correspondientes a las 16 modalidades de respuesta asociadas a estas cuatro preguntas se recogen en la *Tabla 4.1*.

Con carácter previo al análisis de escalabilidad propiamente dicho, será necesario determinar el grado de dificultad ( de menor a mayor dificultad de observación de una respuesta afirmativa ) inherente a las cuatro variables anteriores, al objeto de establecer entre las mismas un orden determinado en la escala. Para ello, se han calculado las distribuciones marginales de las dos posibles respuestas a cada una de las preguntas. Los datos de la *Tabla 4.2*. revelan que la variable C es la que registra una mayor proporción de respuestas afirmativas, y que la variable F es la que posee menor número de respuestas afirmativas. Por tanto, las variables C y F serán las variables extremas de la escala, mientras que las variables S y D poseen niveles intermedios de dificultad en respuestas afirmativas. Todo sugiere, por consiguiente, que la ordenación lógica de estas cuatro variables sea (C,S,D,F), donde la variable C es la menos difícil de todas y la variable F es la más difícil de todas.

*Tabla 4.1.*  
Frecuencias observadas de las variables C, S, D y F

| VARIABLE F | VARIABLE D | VARIABLE S | VARIABLE C |     |
|------------|------------|------------|------------|-----|
|            |            |            | Sí         | No  |
| Sí         | Sí         | Sí         | 75         | 18  |
|            |            | No         | 43         | 13  |
|            | No         | Sí         | 121        | 35  |
|            |            | No         | 86         | 26  |
| No         | Sí         | Sí         | 271        | 65  |
|            |            | No         | 197        | 60  |
|            | No         | Sí         | 449        | 150 |
|            |            | No         | 579        | 230 |

*Fuente:* elaboración propia a partir de datos del CIS.

*Tabla 4.2.*  
Distribución marginal de respuestas de las variables C, S, D y F

|            | Sí                        | No                        |
|------------|---------------------------|---------------------------|
| VARIABLE C | 1821<br>( <b>75,3 %</b> ) | 597<br>( <b>24,7 %</b> )  |
| VARIABLE S | 1184<br>( <b>48,9 %</b> ) | 1234<br>( <b>51,1 %</b> ) |
| VARIABLE D | 742<br>( <b>30,6 %</b> )  | 1676<br>( <b>69,4 %</b> ) |
| VARIABLE F | 417<br>( <b>17,2 %</b> )  | 2001<br>( <b>82,8 %</b> ) |

*Fuente:* elaboración propia a partir de datos del CIS.

Si un 1 representa una respuesta afirmativa y un 2 una respuesta negativa, existirán sólo 5 modalidades de respuesta que se considerarán correctas en la anterior escala, que serían (1,1,1,1), (1,1,1,2), (1,1,2,2), (1,2,2,2) y (2,2,2,2), y cuyas frecuencias observadas son 75, 271, 449, 579 y 230, respectivamente. Las 11 modalidades de respuesta restantes se considerarán incorrectas.

La escalabilidad de estas cuatro variables puede determinarse mediante los coeficientes de reproducibilidad y de escalabilidad. En concreto, el coeficiente de reproducibilidad es:

$$R = 1 - \frac{814}{4 \times 2.418} = 0,9158$$

el cual supera el mínimo comúnmente aceptado de 0,9, lo que nos lleva a concluir que las cuatro variables analizadas pueden ser representadas en una escala. Por su parte, el coeficiente de escalabilidad es:

$$S = 1 - \frac{814}{2.940} = 0,7231$$

que también supera el valor mínimo de 0,6. Por tanto, las variables C, S, D y F pueden ser ordenadas en una escala unidimensional y acumulativa.

Una vez confirmada la escalabilidad de las variables, el empleo del modelo ACL para el análisis de la misma requiere contrastar, en primer lugar, el modelo de independencia completa, ya que la aceptación de este modelo supondría la inexistencia de una variable latente y, por consiguiente, la imposibilidad de formular modelos con errores de medición para la escala construida. Pues bien, como puede apreciarse en la *Tabla 4.3.*, el modelo de independencia es claramente rechazado tanto a un 1 % como a un 5 % de nivel de significación, lo que confirma la presencia de una variable latente cuyas 5 clases representarán las posiciones de escala señaladas con anterioridad.

Confirmado el modelo ACL con 5 posiciones de escala, se procederá a valorar la idoneidad de cada una de las alternativas probabilísticas a la escala determinística de Guttman.

La primera de estas alternativas es el **modelo de Proctor**, el cual estima un único porcentaje de error para todas las variables de la *Tabla 4.1.* y para todas las posiciones de escala de 0,1838. La distribución de la población en las cinco posiciones de escala, según este modelo, sería la siguiente:

- Posición de escala (1,1,1,1): 1,4101 %.
- Posición de escala (1,1,1,2): 16,5930 %.
- Posición de escala (1,1,2,2): 30,6311 %.
- Posición de escala (1,2,2,2): 42,3035 %.
- Posición de escala (2,2,2,2): 9,0623 %.

En cualquier caso, el ajuste del modelo de Proctor ( *Tabla 4.3.* ) a los datos de la *Tabla 4.1.* es bastante insatisfactorio, lo que determina claramente su rechazo tanto a un 1 % como a un 5 % de nivel de significación.

Seguidamente se ha estimado el **modelo de porcentaje de error específico de cada variable observada**. Según este modelo, el porcentaje de error es diferente para cada una de las variables C, S, D y F, cifrándose en los valores siguientes:

- Variable C: 0,2469.
- Variable S: 0,5581.
- Variable D: 0,8518
- Variable F: 0,1701

Por su parte, los tamaños relativos de las 5 clases latentes del modelo ACL son estimados por este modelo en los valores siguientes:

- Posición de escala (1,1,1,1): 0,3583 %.
- Posición de escala (1,1,1,2): 77,0068 %.
- Posición de escala (1,1,2,2): 2,2631 %.

- Posición de escala (1,2,2,2): 20,3717 %.
- Posición de escala (2,2,2,2): 0,0000 %.

Sin embargo, el modelo de porcentaje de error específico de cada variable observada es el modelo menos adecuado para analizar la escalabilidad de las variables C, S, D y F, ya que proporciona un ajuste incluso peor que el del modelo de Proctor, lo que nos confirma claramente su rechazo ( véase la *Tabla 4.3.* ).

*Tabla 4.3.*  
Bondad de ajuste del modelo de independencia completa y  
de otros modelos con errores de medición

|                                                                     | g.l. | $G^2$     | $p$    | $\chi^2$  | $p$    |
|---------------------------------------------------------------------|------|-----------|--------|-----------|--------|
| Independencia completa                                              | 11   | 77,175780 | 0,0000 | 79,958650 | 0,0000 |
| Modelo de Proctor                                                   | 10   | 43,145440 | 0,0000 | 39,801750 | 0,0001 |
| Modelo de porcentaje de error específico de cada variable observada | 7    | 59,510500 | 0,0000 | 60,498470 | 0,0000 |
| Modelo de porcentaje de error específico de cada posición de escala | 7    | 27,900220 | 0,0002 | 27,312620 | 0,0003 |
| Modelo de Distancia Latente de Lazarsfeld                           | 5    | 3,281725  | 0,6566 | 3,278738  | 0,6571 |

*Fuente:* elaboración propia a partir de los cálculos realizados por los programas MLLSA y SHAZAM.

El tercer modelo analizado es el **modelo de porcentaje de error específico de cada posición de escala**. Según este modelo, las 5 clases latentes del modelo ACL poseen diferentes porcentajes de error. La estimación de estos porcentajes y del tamaño relativo de cada clase latente se muestra a continuación:

|                                 | <u>Porcentaje de error</u> | <u>Tamaño relativo</u> |
|---------------------------------|----------------------------|------------------------|
| - Posición de escala (1,1,1,1): | 0,9009                     | 6,7754 %               |
| - Posición de escala (1,1,1,2): | 0,2430                     | 21,1199 %              |
| - Posición de escala (1,1,2,2): | 0,2118                     | 34,5003 %              |
| - Posición de escala (1,2,2,2): | 0,1524                     | 36,8474 %              |
| - Posición de escala (2,2,2,2): | 0,1231                     | 0,7569 %               |

Aunque este modelo proporciona una mejora sustancial del ajuste a los datos observados respecto al modelo de Proctor (  $G^2 = 43,145440 - 27,902220 = 15,24322$  con  $10 - 7 = 3$  grados de libertad ), lo cierto es que dicho modelo tampoco proporciona un ajuste satisfactorio a los datos ( *Tabla 4.3.* ), lo que nos obliga a rechazarlo tanto a un 1 % como a un 5 % de nivel de significación.

Finalmente, se ha estimado el **modelo de Distancia Latente de Lazarsfeld**, cuyo ajuste a los datos de la *Tabla 4.1.* es bastante satisfactorio, ya que arroja un valor del test de la razón de verosimilitud de  $G^2 = 3,281725$  con 5 grados de libertad, lo que significa que dicho modelo puede ser aceptado para representar la escalabilidad de las variables C, S, D y F. La “proporción correctamente clasificada” de este modelo es igual a un

67,21130 % y el coeficiente  $\lambda$  asciende a 0,34649. Las estimaciones de los porcentajes de error de las variables objeto de análisis y de los tamaños relativos de las 5 clases latentes se muestra en la *Tabla 4.4.*, de cuyo análisis se desprenden los siguientes comentarios finales:

1º) Los tamaños de las diferentes posiciones de escala son bastante dispares, ya que frente al 49,82 % de la posición 2 y al 28,25 % de la posición 4, los tamaños relativos de las tres posiciones de escala restantes son considerablemente menores, especialmente los de las posiciones 5 y 3 ( 7,00 % y 2,32 %, respectivamente ).

2º) Algo más del 12 % de los entrevistados manifiesta una actitud muy interesada hacia el fenómeno vacacional, en general, y hacia su fraccionamiento, en particular. Estos individuos se sitúan en la posición inferior de la escala ( posición 1 ), catalogan de necesarias a las vacaciones, aprovechan la Semana Santa, la Navidad y los puentes para disfrutar de su tiempo de ocio en lugares diferentes al de su residencia habitual, se despreocupan de sus gastos y fraccionan sus vacaciones principales. Sin embargo, en esta posición de escala el porcentaje de error correspondiente a la variable D ( es decir, la probabilidad de que muestren una cierta preocupación por sus gastos vacacionales ) es bastante elevado, ya que se sitúa en un 59,08 %.

*Tabla 4.4.*

Estimaciones de los porcentajes de error y de los tamaños relativos de las posiciones de escala bajo el modelo de Distancia Latente de Lazarsfeld

| Clase latente | Posición de escala | Porcentajes de error |            |            |            | Tamaño de la posición de escala |
|---------------|--------------------|----------------------|------------|------------|------------|---------------------------------|
|               |                    | Variable C           | Variable S | Variable D | Variable F |                                 |
| 1             | (1,1,1,1)          | 0,2057               | 0,3513     | 0,5908     | 0,0622     | 0,125941                        |
| 2             | (1,1,1,2)          | 0,2057               | 0,3513     | 0,5908     | 0,0622     | 0,498267                        |
| 3             | (1,1,2,2)          | 0,2057               | 0,3513     | 0,1368     | 0,0622     | 0,023221                        |
| 4             | (1,2,2,2)          | 0,2057               | 0,1976     | 0,1368     | 0,0622     | 0,282563                        |
| 5             | (2,2,2,2)          | 0,2057               | 0,1976     | 0,1368     | 0,0622     | 0,070007                        |

*Fuente:* elaboración propia a partir de los cálculos realizados por el programa MLLSA.

3º) La posición de escala más frecuente es la 2, en la que los individuos también apuntan el carácter necesario de las vacaciones, salen en Semana Santa, Navidad o puentes y se despreocupan de los gastos, aunque, por unos motivos u otros, no fraccionan sus vacaciones principales.

4º) En el extremo opuesto de la escala se encuentran los individuos menos interesados por las vacaciones y por su posible fraccionamiento. Así, la posición superior de la escala ( posición 5 ) agrupa a los sujetos que consideran que las vacaciones no son tan importantes y que, en coherencia con esta forma de pensar, no salen en Semana Santa, Navidad o puentes, se preocupan mucho de sus gastos vacacionales y, obviamente, no fraccionan sus vacaciones. En la posición de escala inmediatamente inferior ( posición 4 ) se encuentran los individuos que, si bien creen en la importancia de las vacaciones, por unos u otros motivos no salen en los cortos períodos de descanso laboral existentes a lo largo del año ni tampoco fraccionan sus vacaciones principales.



5º) El único porcentaje de error de la variable “más fácil” ( variable C ) se estima en 0,2057 que es bastante más elevado que el porcentaje de error estimado de la variable “más difícil” ( variable F ), que se sitúa en 0,0622.

6º) La variable intermedia S posee un porcentaje de error estimado para las respuestas afirmativas ( 0,3513 ) considerablemente mayor que para las respuestas negativas ( 0,1976 ).

7º) De todas las variables incluidas en la escala, es la variable D la que posee el porcentaje de error más elevado para las respuestas afirmativas ( 0,5908 ), circunstancia que, a pesar de que la estimación del porcentaje de error para las respuestas negativas de esta variable alcanza un valor bastante más pequeño ( 0,1368 ), debería alertarnos sobre la pertenencia de esta variable a la escala construida o, al menos, tener una cierta precaución a la hora de interpretar el significado de la misma en el conjunto de la escala.

## 5. BIBLIOGRAFÍA.

- C.I.S. (1995): *El comportamiento de los españoles ante las vacaciones (III)*. Estudio 2.193, Octubre de 1995. Centro de Investigaciones Sociológicas. Madrid.

- CLOGG, C.C. (1977): *Unrestricted and restricted maximum likelihood latent structure analysis: a manual for users*. Working Paper 1977-09. University Park, PA: Population Issues Research Office.

- CLOGG, C.C. (1979): “Some latent structure models for the analysis of Likert-type data”. *Social Science Research*, nº 8, pp. 287-301.

- CLOGG, C.C. y SAWYER, D.O. (1981): *A comparison of alternative models for analyzing the scalability of response patterns*. En S. Leinhardt (ed.) *Sociological Methodology*. San Francisco: Josey-Bass.

- DEMPSTER, A.P., LAIRD, N.M. y RUBIN, D.B. (1977): “Maximum likelihood from incomplete data via the EM algorithm ( with discussion )”. *Journal of the Royal Statistical Society*, serie B, nº 39, pp. 1-38.

- GOODMAN, L.A. (1974a): “Exploratory latent structure analysis using both identifiable and unidentifiable models”. *Biometrika*, nº 61, pp. 215-231.

- GOODMAN, L.A. (1974b) : “The analysis of system of qualitative variables when some of the variables are unobservable. Part I-A: Modified latent structure approach”. *American Journal of Sociology*, nº 79, pp. 1179-1259.

- GOODMAN, L.A. (1979): “On the estimation of parameters in latent structure analysis”. *Psychometrika*, nº 44, pp. 123-128.

- GUTTMAN, L.L. (1944): “A basis for scaling qualitative data”, *American Sociological Review*, nº 9, pp. 139-150.

- GUTTMAN, L.L. (1950): “The basis for scalogram analysis”, pp. 60-90. En S.A. Stouffer et al. (eds.) *Measurement and Prediction*, Princeton, NJ: Princeton University Press.

- HENRY, N.W. (1983): *Latent structure analysis*. En S. Kotz y N.L. Johnson (eds.). *Encyclopedia of Statistical Sciences* ( pp. 497-504 ). Ed. John Wiley & Sons, New York.

- LAZARSFELD, P.F. (1950a): “The logical and mathematical foundations of latent structure analysis”. En S.A. Stouffer et al. (eds.) *Measurement and Prediction*, Princeton, NJ: Princeton University Press.

- LAZARSELD, P.F. (1950b): “The interpretation and computation of some latent structures”. En S.A. Stouffer et al. (eds.) *Measurement and Prediction*, Princeton, NJ: Princeton University Press.
- LAZARSELD, P.F. y HENRY, N.W. (1968): *Latent Structure Analysis*. Boston: Houghton Mifflin.
- McCUTHEON, A.L. (1987): *Latent class analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-064. Beverly Hills: Sage Pubns.
- PROCTOR, C.H. (1970): “A probabilistic formulation and statistical analysis of Guttman scaling”, *Psychometrika*, nº 35, pp. 73-78.
- TORGERSON, W.S. (1962): *Theory and Methods of Scaling*. Ed. John Wiley & Sons, New York.