Doble Grado en
Matemáticas y Estadística

TRABAJO FIN DE GRADO

Medidas de Asociación para Variables Nominales y Ordinales

Alumna: Gloria Vizcaíno Castaño
Tutor: Juan Manuel Muñoz Pichardo

Sevilla, Junio de 2022

Índice general

Resumen

La evaluación de la relación y la intensidad de la misma entre dos o más variables aleatorias es un objetivo presente en cualquier estudio estadístico. Para ello se han propuesto una amplia gama de medidas estadísticas adaptadas a la naturaleza de las variables analizadas.

El objetivo de este trabajo se centra en recopilar las medidas de asociación entre variables nominales y ordinales, incluyendo la definición teórica e inferencia estadística sobre las mismas.

Asimismo, el trabajo incluye su implementación y/o el uso de librerías de R, con una ilustración sobre datos reales. Además, con objeto de ilustrar la aplicabilidad de las mismas, incluye referencias sobre trabajos científicos recientes en los que se han utilizado estas medidas.

Abstract

The evaluation of the relationship and its intensity between two or more random variables is an objective present in any statistical study. For this purpose, a wide range of statistical measures adapted to the nature of the variables analyzed have been proposed.

The aim of this project is to compile measures of association between nominal and ordinal variables, including their theoretical definition and statistical inference.

Also, the report includes their implementation in R and/or the use of R libraries, with an illustration on real data. In addition, in order to illustrate their applicability, it includes references to recent scientific studies in which these measures have been used.

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Aunque existen una gran cantidad de métodos para medir la magnitud de la asociación entre dos variables, hay grandes dificultades para interpretar y comparar las distintas medidas, ya que a menudo difieren en su estructura, lógica e interpretación.

Así pues, las distintas medidas de asociación desarrolladas a lo largo de los años constituyen una mezcla de enfoques lógicos, estructurales y de interpretación.

Es conveniente clasificar las distintas medidas de asociación por el nivel de medición para el que fueron diseñadas originalmente y para el que son más apropiadas, reconociendo que algunas medidas pueden ser adecuadas para más de un nivel de medición, especialmente las numerosas medidas originalmente diseñadas para el análisis de tablas de contingencia 2×2, en las que el nivel de medición es a veces irrelevante.

Además de la consideración de la estructura, la lógica y la interpretación, un inconveniente importante de las medidas de asociación es la determinación del p-valor de la medida obtenida bajo la hipótesis nula.

Existen dos enfoques principales para determinar los p-valores de las medidas de asociación: el modelo poblacional de Neyman-Pearson y el modelo de permutación de Fisher-Pitman.

El modelo poblacional está plagado de suposiciones que rara vez se cumplen en la práctica y que, algunas veces, son inapropiadas; por ejemplo, independencia, normalidad, homogeneidad de la varianza…

De aquí en adelante se usará casi exclusivamente el modelo de permutación, ya que está libre de cualquier suposición de distribución, no requiere un muestreo aleatorio, es completamente dependiente de los datos, proporciona valores de probabilidad exactos y es ideal para el análisis de muestras pequeñas.

Por tanto, en este Trabajo Fin de Grado, se usará el enfoque de permutación para la medición de la asociación estadística, definida ampliamente para incluir medidas de correlación, asociación y concordancia.

1.1. Definiciones

Dado que el título de este Trabajo de Fin de Grado es “Medidas de Asociación para Variables Nominales y Ordinales”, es conveniente, en primer lugar, definir “Medición” y

"Asociación".

1.1.1. Definición de Medición

Según Cowles, la medición es la mejor manera de describir con precisión los acontecimientos y las relaciones entre ellos [Cowles, 2001]. La medición ha sido una característica fundamental de la civilización humana desde sus inicios. Así, la medición es la aplicación de las matemáticas a los acontecimientos, el uso de números para designar objetos y acontecimientos, y sus relaciones.

Más formalmente, la medición es el proceso de la asignación a los fenómenos empíricos de un sistema numérico.

Se pueden distinguir cuatro niveles o escalas de medición: nominal, ordinal, de intervalo y de razón:

- El nivel **nominal** de medición no mide cantidades, simplemente clasifica los acontecimientos en una serie de categorías no ordenadas y se agrupan los acontecimientos que tienen características comunes. Ejemplos de clasificaciones nominales son el género, el tipo de sangre o el estado civil.
- La esencia del nivel **ordinal** de medición es que emplea las características de "mayor que" ($>$) o "menor que" ($<$). Las relaciones ($>$) y ($<$) no son reflexivas ni simétricas, pero sí transitivas. Ejemplos de escalas ordinales son el orden de nacimiento, el rango académico o las escalas Likert (Muy de acuerdo, De acuerdo, Neutral, En desacuerdo, Totalmente en desacuerdo).
- Las escalas de nivel de **intervalo** introducen otra dimensión en el proceso de medición y ordenan los eventos en intervalos de igual apariencia. En las escalas de intervalo, no hay un punto cero absoluto: si hay un valor de cero, éste se define arbitrariamente. Las temperaturas medidas en grados Fahrenheit o Centígrados son ejemplos tradicionales de medición de intervalos.
- Las escalas de **razón** son escalas que no sólo incorporan todas las características de una escala de intervalo, sino que tienen puntos cero absolutos, lo que permite la construcción de relaciones significativas. Ejemplos de escalas de intervalo son el tiempo, la edad, la altura o los grados Kelvins (0 Kelvins es el cero absoluto, definido como la ausencia de movimiento molecular).

Desde el punto de vista estadístico, las mediciones a nivel de intervalo y de razón suelen tratarse juntas y, en general, se denominan simplemente mediciones de nivel de intervalo.

1.1.2. Definición de Asociación

Aunque hay muchas formas de definir la asociación, quizás la más sencilla y útil sea: *se dice que dos variables están asociadas cuando la distribución de los valores de una variable difiere para diferentes valores de la otra variable.*

Además, si un cambio en la distribución de los valores de una variable no provoca un cambio en la distribución de los valores de la otra variable, se dice que las variables son independientes.

1.2. Dimensiones de la asociación

Hay que tener en cuenta varias dimensiones a la hora de medir la asociación:

1.2.1. Nivel de Medición

Como hemos visto anteriormente, pueden ser: variables de nivel nominal (categóricas), de nivel ordinal (clasificadas) y de nivel de intervalo. Además, en algunos casos, se consideran mezclas de los tres niveles de medición: variables de nivel nominal y ordinal, de nivel nominal y de intervalo, y de nivel ordinal e intervalo.

1.2.2. Simetría y Asimetría

Una medida de asociación puede ser asimétrica, con variables independientes y dependientes bien definidas, dando lugar a dos índices que miden la fuerza de la asociación dependiendo de la variable que se considere dependiente; o simétrica, dando lugar a un único índice de fuerza de asociación.

1.2.3. Asociación unidireccional y bidireccional

Las medidas de asociación pueden cuantificar la asociación unidireccional entre variables basándose en la medida en que una variable implica a la otra, pero no a la inversa. Por otro lado, la asociación bidireccional o mutua se refiere a la medida en que las dos variables se implican mutuamente. Todas las medidas asimétricas son medidas de asociación unidireccional, y algunas medidas simétricas son medidas de asociación unidireccional.

1.2.4. Clasificación cruzada

Las medidas de asociación se han construido históricamente para datos clasificados en tablas de contingencia de doble entrada o, alternativamente, en simples listas bivariadas de medidas de respuesta. Además, algunas medidas suelen calcularse para ambos casos.

1.2.5. Correlación, asociación y concordancia

Las medidas de **asociación** pueden medir de diversas maneras la correlación, la asociación o la concordancia. Muchos autores han tratado de distinguir entre los conceptos de correlación y asociación. Hay dos ámbitos correspondientes al término “asociación”:

- El más general incluye todos los tipos de medidas de asociación entre dos variables en todos los niveles de medición.
- El más restrictivo está reservado a las medidas diseñadas específicamente para medir el grado de relación entre dos variables en los niveles de medición nominal y ordinal.

Así pues, en este trabajo, la asociación se usará de dos maneras. En primer lugar, como un concepto global que incluye medidas de correlación, asociación y concordancia; y en segundo lugar, se utilizará más específicamente como una medida de relación entre dos variables de nivel nominal, dos variables de nivel ordinal o alguna combinación de ambas.

En general, la **correlación** suele referirse a las medidas de covariación derivadas de las ecuaciones de regresión basadas en el método de mínimos cuadrados ordinarios. A menudo, pero no siempre, la correlación simple mide la relación entre dos variables a nivel de intervalo de medida, donde las dos variables se etiquetan normalmente como X e Y. La medida de correlación más usada es el coeficiente de correlación de Pearson al cuadrado.

Las medidas de **concordancia** intentan determinar la identidad de dos variables en cualquier nivel de medición, es decir, $X_i = Y_i$ para todo i.

La correlación y la concordancia se suelen confurdir, a continuación se muestra un ejemplo para ententender las diferencias:

Supongamos que un investigador desea establecer la relación entre los valores observados y los predichos por la regresión, y e $\hat{y}$, respectivamente. La concordancia implica que la relación funcional entre y e $\hat{y}$ puede describirse mediante una la recta $x = y$. Si por ejemplo obtenemos los pares (1,1), (3,3), (8,8), el coeficiente de correlación de Pearson al cuadrado es $r^2_{y,\hat{y}} = 1$ y el porcentaje de concordancia es del 100 %, es decir, los elementos de los tres pares $(y, \hat{y})$ son iguales.

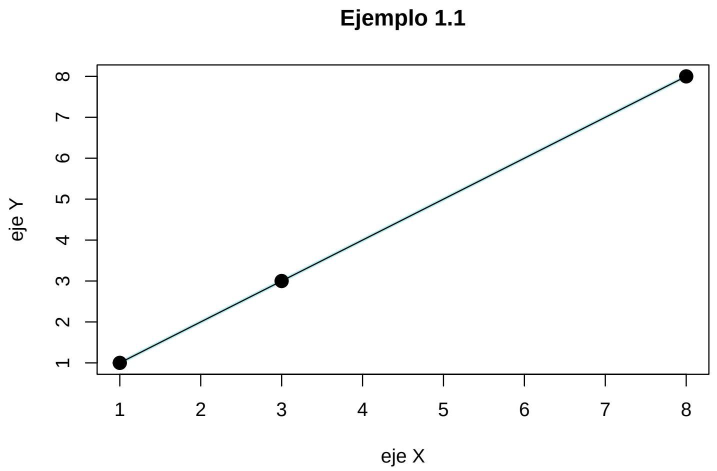

Figura 1.1: Representación gráfica Ejemplo 1.1

En este contexto, el coeficiente de correlación de Pearson al cuadrado, $r^2_{y,\hat{y}}$, también se ha utilizado como medida de concordancia. Sin embargo, $r^2_{y,\hat{y}} = 1,00$ implica una relación lineal entre y e $\hat{y}$, donde tanto la el corte con el eje de ordenadas como la pendiente son arbitrarias. Así, aunque la concordancia perfecta se describe con un valor de 1,00, también es cierto que $r^2_{y,\hat{y}} = 1,00$ describe una relación lineal que puede o no reflejar una concordancia perfecta, por ejemplo para los valores $(y, \hat{y})$: (2, 4), (4, 5), (6, 6), (8, 7), y (10, 8), el coeficiente de correlación de Pearson es $r^2_{y,\hat{y}} = 1,00$, y el porcentaje de concordancia es del 20 %, es decir, sólo un par valores coinciden [Berry et al., 2010].

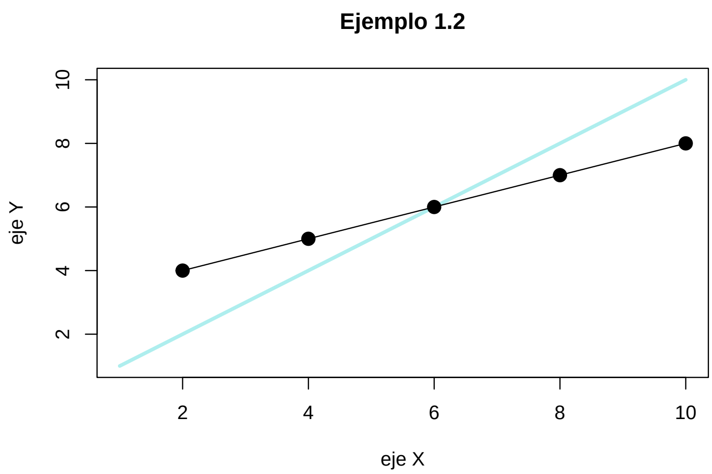

Figura 1.2: Representación gráfica Ejemplo 1.2

1.3. Criterios para las medidas de asociación

Varios investigadores han escrito sobre criterios importantes para las medidas de asociación, como Costner [Costner, 1965] y Goodman y Kruskal [Goodman and Kruskal, 1954]. Sin embargo, esta sección se basa principalmente en los criterios que Weiss [Weiss, 1968] consideraba más importantes.

Los criterios más importantes para las medidas de asociación incluyen la normalización adecuada, la interpretación, la independencia de las frecuencias marginales y la magnitud (grado o fuerza) de la asociación:

- Normalización: Idealmente, los valores de una medida de asociación deberían cubrir el mismo rango que los valores de probabilidad, es decir, de 0 a 1. Además, la medida de asociación debe ser cero cuando las variables son independientes y uno cuando hay una asociación perfecta. Cuando sea conveniente considerar la asociación inversa, entonces menos uno debe representar la asociación negativa perfecta.

- Interpretación: Una medida de asociación debe tener una interpretación significativa, como la reducción proporcional del error probable, la proporción de la varianza explicada o la proporción por encima de lo que cabría esperar por azar. Muchas medidas de asociación carecen notablemente de este aspecto. De hecho, muchas medidas no permiten ninguna interpretación, excepto que un valor más alto indica más asociación que un valor más bajo, e incluso eso es a menudo cuestionable.

- Independencia de las frecuencias marginales: Idealmente, una medida de asociación no debería cambiar con un aumento (disminución) de los totales de frecuencia de filas o columnas; es decir, la medida de asociación debería ser independiente de los totales de frecuencia marginal. Algunas medidas de asociación tienen esta propiedad, como las diferencias porcentuales y los odds ratio, pero muchas otras no.

- Grado de asociación: Los valores de una medida de asociación deben aumentar (disminuir) con el aumento (disminución) de los grados de asociación. Así, cuando las frecuencias de las celdas de una tabla de contingencia indican cambios en la asociación, la medida de asociación debería cambiar de forma acorde.

1.4. Grado de Asociación

Las diferentes medidas de asociación evalúan el grado de asociación de diversas maneras. Entre las diversas formas de medir la fuerza de la asociación se encuentran la desviación de la independencia, las comparaciones por pares, la correspondencia incremental y la concordancia entre variables:

- Desviación de la independencia: Las medidas de asociación que se basan en la desviación de la independencia plantean cómo serían los datos si las dos variables fueran independientes, es decir, que no hubiera asociación, y luego miden el grado en que los datos observados se apartan de la independencia.
- Comparaciones por pares: Algunas medidas de asociación se basan en comparaciones por pares donde las diferencias entre las medidas de respuesta se calculan entre todos los pares de mediciones posibles y se dividen en pares concordantes y discordantes. Un par concordante es aquel en el que la dirección de la diferencia con una variable coincide con la dirección de la diferencia con la segunda variable. Un par discordante es aquel en el que la dirección de la diferencia con una variable no es igual a la dirección de la diferencia con la segunda variable.
- Correspondencia incremental: El grado de asociación se basa en la medida en que un aumento (disminución) incremental en una variable va acompañado de un aumento (disminución) en la otra variable. Este enfoque se denomina convencionalmente "correlación" en lugar de "asociación".
- Concordancia entre variables: El grado de asociación se mide por el grado en que los valores de una variable discrepan de los valores de la otra variable, por encima de lo esperado por el mero azar.

1.5. Modelo estadístico de Permutación

La Medición de la Asociación adopta un enfoque de permutación para generar valores de probabilidad exactos y de remuestreo para varias medidas de asociación.

Las medidas estadísticas de permutación tienen varias ventajas sobre los métodos estadísticos clásicos de la estadística clásica, ya que son óptimas para muestras pequeñas, pueden utilizarse para analizar muestras no aleatorias, son completamente dependientes de los datos, no necesitan hipótesis distribucionales y proporcionan valores de probabilidad exactos.

Hoy en día, las pruebas estadísticas de permutación son consideradas por muchos como la herramienta más importante para evaluar y validar las pruebas estadísticas convencionales.

Un inconveniente obvio de los métodos estadísticos de permutación es la cantidad de cálculos necesarios, pero es cierto que la llegada de la informática hizo posible que los métodos de permutación fueran viables para muchos problemas. Aunque, actualmente, los potentes algoritmos computacionales y los ordenadores modernos hacen que los análisis de permutación sean prácticos para muchas aplicaciones de investigación.

A continuación, se compara el modelo poblacional convencional y el modelo estadístico de permutación:

- El modelo poblacional supone un muestreo aleatorio de una o más poblaciones específicas. Según el modelo poblacional, el nivel de significación estadística que resulta de la aplicación de una prueba estadística a los resultados de un experimento o encuesta corresponde a la frecuencia con la que se rechazaría la hipótesis nula en muestreos aleatorios repetidos de una población especificada. Dado que el muestreo repetido de la población especificada no es práctico, se supone que la distribución de muestreo de los estadísticos de prueba generados en el muestreo aleatorio repetido se ajusta a una distribución teórica aproximada, como la distribución normal. El tamaño de una prueba estadística es la probabilidad bajo la hipótesis nula de que los resultados repetidos basados en muestras aleatorias de muestras aleatorias del mismo tamaño sean iguales o más extremos que el resultado observado.
- El modelo de permutación no supone ni requiere un muestreo aleatorio de una población determinada. Para el modelo de permutación exacta, se calcula un estadístico de prueba para los datos observados. A continuación, las observaciones se permutan sobre todos los posibles arreglos de los datos observados, y el estadístico de prueba seleccionado se calcula para cada uno de los posibles arreglos. La proporción de arreglos con valores del estadístico de prueba iguales o más extremos que el estadístico de prueba observado da la probabilidad exacta del valor del estadístico de prueba observado. Cuando el número de disposiciones posibles de los datos observados es muy grande, los métodos de permutación exacta son poco prácticos y se hacen necesarios los métodos de permutación por remuestreo de Monte Carlo. Los métodos de remuestreo generan una muestra aleatoria de todas las posibles ordenaciones de los datos observados, y el valor de la probabilidad de remuestreo es la proporción de ordenaciones con valores del estadístico de prueba iguales o más extremos que el valor del estadístico de prueba observada.

1.5.1. Distribución de probabilidad hipergeométrica

Los métodos estadísticos de permutación exacta, especialmente cuando se aplican a las tablas de contingencia, dependen en gran medida de los valores de probabilidad hipergeométricos. En esta sección, se hace una breve introducción a la distribución de probabilidad hipergeométrica ilustrando su cálculo e interpretación. Para las tablas de contingencia de 2×2, el cálculo de las probabilidades hipergeométricas es fácil de demostrar. Consideremos la siguiente tabla de contingencia 2×2:

Tabla 1.1: Clasificación cruzada de dos variables categóricas, A y B, con 2 categorías cada una

	A_1	A_2	Total
B_1	n_{11}	n_{12}	R_1
B_2	n_{21}	n_{22}	R_2
Total	C_1	C_2	N

donde $n_{11}, \ldots, n_{22}$ denotan las cuatro frecuencias absolutas, R_1 y R_2 denotan los totales de las frecuencias marginales de cada fila, C_1 y C_2 denotan los totales de las frecuencias marginales de cada columna y

$$N = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \ .$$

Dado que la tabla de contingencia que aparece en la *Tabla 2.1* es una tabla 2×2 y, en consecuencia tiene sólo un grado de libertad, la probabilidad de cualquier frecuencia de celda constituye la probabilidad de toda la tabla de contingencia. Por lo tanto, la probabilidad del punto hipergeométrico para la celda que contiene n_{11} viene dada por:

$$p(n_{11}|R_1, C_1, N) = \left(\begin{array}{c} C_1 \\ n_{11} \end{array}\right)\left(\begin{array}{c} C_2 \\ n_{12} \end{array}\right)\left(\begin{array}{c} N \\ R_1 \end{array}\right)^{-1} =$$

$$= \left(\begin{array}{c} R_1 \\ n_{11} \end{array}\right)\left(\begin{array}{c} R_2 \\ n21 \end{array}\right)\left(\begin{array}{c} N \\ C_1 \end{array}\right)^{-1} = \frac{R_1!R_2!C_1!C_2!}{N!n_{11}!n_{12}!n_{21}!n_{22}!}$$

El cálculo de las probabilidades hipergeométricas para las tablas de contingencia $r \times c$ es más complejo que para las tablas de contingencia simples 2×2. Consideremos la tabla de contingencia 4×3:

Tabla 1.2: Clasificación cruzada de dos variables categóricas, A y B, con 3 y 4 categorías respectivamente

	A_1	A_2	A_3	Total
B_1	n_{11}	n_{12}	n_{13}	R_1
B_2	n_{21}	n_{22}	n_{23}	R_2
B_3	n_{31}	n_{32}	n_{33}	R_3
B_4	n_{41}	n_{42}	n_{43}	R_4
Total	C_1	C_2	C_3	N

donde $n_{11}, \ldots, n_{43}$ denotan las 12 frecuencias absolutas de las celdas, $R_1, ..., R_4$ denotan los totales de frecuencia marginal de las cuatro filas, C_1, C_2 y C_3 denotan los totales de las frecuencias marginales de las tres columna y

$$N = \sum_{i=1}^{4}\sum_{j=1}^{3} n_{ij} \ .$$

Cuando sólo hay dos filas, como en el ejemplo anterior con la tabla 2×2, cada probabilidad de columna es binomial, pero con cuatro filas cada probabilidad de columna es multinomial. Es bien sabido que una probabilidad multinomial puede obtenerse de una serie interconectada de expresiones binomiales. Por ejemplo, para la columna A_1 de la *Tabla 2.2*:

$$\left(\begin{array}{c} C_1 \\ n_{11} \end{array}\right)\left(\begin{array}{c} C_1 - n_{11} \\ n_{21} \end{array}\right)\left(\begin{array}{c} C_1 - n_{11} - n_{21} \\ n_{31} \end{array}\right) =$$

$$= \frac{C_1!}{n_{11}!(C_1 - n_{11})!} \times \frac{(C_1 - n_{11})!}{n_{21}!(C_1 - n_{11} - n_{21})!} \times \frac{(C_1 - n_{11} - n_{21})!}{n_{31}(C_1 - n_{11} - n_{21} - n_{31})!} = \frac{C_1!}{n_{11}!n_{21}!n_{31}!n_{42}!}$$

Para la columna A_2:

$$\binom{C_2}{n_{12}} \binom{C_2 - n_{12}}{n_{22}} \binom{C_2 - n_{12} - n_{22}}{n_{32}} =$$

$$= \frac{C_2!}{n_{12}!(C_2 - n_{12})!} \times \frac{(C_2 - n_{12})!}{n_{22}!(C_2 - n_{12} - n_{22})!} \times \frac{(C_2 - n_{12} - n_{22})!}{n_{32}(C_2 - n_{12} - n_{22} - n_{32})!} = \frac{C_2!}{n_{12}!n_{22}!n_{32}!n_{42}!}$$

y para la distribución de frecuencias marginales de las filas:

$$\binom{N}{R_1} \binom{N - R_1}{R_2} \binom{N - R_1 - R_2}{R_3} =$$

$$= \frac{N!}{R_1!(N - R_1)!} \times \frac{(N - R_1)!}{R_2!(N - R_1 - R_2)!} \times \frac{(N - R_1 - R_2)!}{R_3!(N - R_1 - R_2 - R_3)!} = \frac{N!}{R_1!R_2!R_3!R_4!}$$

Por consiguiente, para tablas de contingencia $r \times c$ se tiene:

$$p(n_{ij}|R_i, C_j, N) = \frac{(\prod_{i=1}^{r} R_i!)(\prod_{j=1}^{c} C_j!)}{N! \prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!}$$

De esta forma, la ecuación anterior puede generalizarse fácilmente a tablas de contingencia multidireccionales más complejas [Mielke and Berry, 1988a].

1.6. Métodos de Predicción

Las medidas de asociación se basan en diferentes modelos, los más usados, debido a su fácil interpretación, son el de corrección por azar y el de reducción proporcional del error.

Medidas de Corrección por Azar

Corregida por el azar significa “quitar del acuerdo global”, es decir, el acuerdo que se obtendría si las valoraciones se hicieran aleatoriamente. Este tipo de medidas se contruyen de la siguiente forma:

$$medida_{CA} = \frac{p_o - p_e}{1 - p_e}$$

Donde, p_o es la proporción observada de observaciones en las que los jueces están de acuerdo, p_e es la proporción de observaciones para las que se espera una concordancia por azar, $p_o - p_e$ es la proporción de concordancia más allá de la esperada por el azar, $1 - p_e$ es la máxima proporción posible de concordancia más allá de la esperada por el azar.

Su interpretación es bastante sencilla, los valores positivos indican una concordancia superior a la esperada por el azar y los valores negativos indican una concordancia inferior.

Medidas de Reducción Proporcional del Error

La *Reducción Proporcional del Error* es un criterio estadístico que cuantifica hasta qué punto el conocimiento sobre una variable puede ayudar a predecir otra variable.

Es decir, ayuda a comprender hasta qué punto conocer una variable X puede ayudar a predecir otra variable Y. Si la relación entre las variables es cero, saber X no te ayudar a predecir Y. Y si existe una "correlación" perfecta , conocer X permite predecir Y con un 100 % de confianza.

La reducción proporcional del error es, por tanto, la ganancia de precisión en la predicción de la variable dependiente Y a partir del conocimiento de la variable independiente X (o un conjunto de múltiples variables). Este tipo de medidas es especialmente útil y popular cuando se utiliza para evaluar modelos en los que la variable dependiente es binaria.

Para la construcción de estas medidas es necesario definir los siguientes errores:

- El primer tipo es el error de predicción basado únicamente en el conocimiento de la distribución de la variable dependiente, denominado "error del primer tipo" (E_1) y que consiste en el número esperado de errores al predecir las c categorías de la variable dependiente $(X_1, ..., X_c)$ a partir de la distribución observada de las marginales de la variable dependiente $(n_{.1}, ..., n_{.c})$.
- El segundo tipo es el error de predicción basado en el conocimiento de las distribuciones de las de la variable independiente y de la dependiente, denominado "error del segundo tipo" (E_2) y que consiste en el número de errores esperados al predecir las c categorías de la variable dependiente $(X_1, ..., X_c)$ a partir del conocimiento de las r categorías de la variable independiente $(Y_1, ..., Y_r)$.

Así, las medidas de Reddución Proporcional del Error se contruyen como sigue:

$$medida_{RPE} = \frac{E_1 - E_2}{E_1}$$

Y su interpretación es la siguiente, los valores positivos indican una mejora proporcional de los errores de predicción con el conocimiento de ambas variables, en comparación de los errores de predicción con el conocimiento de una sola variable.

Variando el <u>tipo de variables</u>, el <u>método de predicción</u> y la <u>medida del error</u> es como se obtienen las diferentes medidas de asociación.

Para ilustrar el proceso, se contruye a continuación la medida de asociación más conocida y usada: el coeficiente de correlación lineal de Pearson, $\rho_{X,Y}$.

El coeficiente de correlación lineal es una medida de reducción proporcional del error considerando dos variables (X e Y) continuas, como criterio de predicción se usa la esperanza matemática, es decir, $E[Y] = \mu$ y $[Y|X = x] = \mu(x)$, y como medida del error el error cuadrático medio, $E[(Y-\mu)^2] = \sigma_Y^2$ y $E[(Y-\mu(x))^2] = \sigma^2_{Y|X=x}$. Por tanto, la medida es:

$$\rho_{X,Y} = \frac{\sigma_Y^2 - \sigma^2_{Y|X=x}}{\sigma_Y^2} \; .$$

Capítulo 2

Medidas para Variables Nominales

Este capítulo se centra en las medidas de asociación diseñadas para las variables de nivel nominal, pero indagando en los métodos estadísticos de permutación exactos y de Monte Carlo para las medidas de asociación nominal que se basan en criterios distintos del estadístico de prueba chi-cuadrado de Pearson.

En primer lugar, se describen dos medidas asimétricas de asociación de nivel nominal propuestas por Goodman y Kruskal en 1954, λ y t. A continuación, el coeficiente kappa no ponderado de Cohen, κ, que proporciona una introducción a la medición de la concordancia, en contraste con las medidas de asociación. También se incluye en el capítulo la medida d_N^c de Leik y Gove de asociación nominal.

Algunas medidas diseñadas para variables de nivel ordinal también sirven como medidas de asociación para variables de nivel nominal cuando r (número de filas) $= 2$ y c (número de columnas) $= 2$, es decir, una tabla de contingencia 2×2. Otras medidas se diseñaron originalmente para tablas de contingencia 2×2 con variables de nivel nominal, entre estas medidas de asociación están las diferencias porcentuales, las medidas Q e Y de Yule, los odds ratio y las medidas asimétricas de Somers, d_{yx} y d_{xy}.

2.1. Medidas λ_a y λ_b de Goodman y Kruskal

Un problema común al que se enfrentan muchos investigadores es el análisis de una tabla de clasificación cruzada en la que ambas variables son categóricas, ya que las variables categóricas no suelen contener tanta información como las variables de nivel ordinal o intervalo (y por supuesto, las numéricas). Las medidas basadas en la chi-cuadrado, como la ϕ^2 de Pearson, la T^2 de Tschuprov, la V^2 de Cramér y la C de Pearson, son poco útiles en la práctica debido a las dificultades de interpretación.

En 1954, Leo Goodman y William Kruskal propusieron varias medidas nuevas de asociación [Goodman and Kruskal, 1954]. Entre las medidas se encontraban dos medidas de predicción asimétricas de reducción proporcional del error para los análisis de una muestra aleatoria de dos variables categóricas: λ_a, para cuando se considera que X es la variable dependiente, y λ_b, para cuando se considera que Y es la variable dependiente.

Estas medidas de reducción proporcional del error consideran como criterio de predicción la moda, es decir, la categoría más frecuente, y como medida del error la tasa de error de clasificación entre las categorías de la variable dependiente.

Sea una tabla de contingencia $r \times c$ como la representada a continuación:

Tabla 2.1: Clasificación cruzada de dos variables categóricas, X e Y, con c y r categorías respectivamente

	X_1	X_2	$\dots$	X_c	Total
Y_1	n_{11}	n_{12}	$\dots$	n_{1c}	$n_{1.}$
Y_2	n_{21}	n_{22}	$\dots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Y_r	n_{r1}	n_{r2}	$\dots$	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.c}$	N

donde X_j con $j = 1, ..., c$ representan las c categorías de la variable dependiente X, Y_i con $i = 1, ..., r$ denotan las r categorías para la variable independiente Y, n_{ij} la frecuencia absoluta de celda para $i = 1, ...r$ y $j = 1, ...c$, y N es el total observaciones. Denotamos por un punto $(.)$ la suma parcial de todas las filas o todas las columnas, según la posición del $(.)$ en la lista de subíndices. Por lo tanto, $n_{i.}$ denota el total de la frecuencia marginal de la i-ésima fila, $i = 1, ...r$, sumada en todas las columnas, y $n_{.j}$ indica la frecuencia marginal de la j-ésima columna, $j = 1, ...c$, sumada en todas las filas.

Teniendo en cuenta la notación de la *tabla* 2.1, se definen

$$W = \sum_{i=1}^{r} max(n_{i1}, n_{i2}, ..., n_{ic}) \qquad y \qquad A = max(n_{.1}, n_{.2}, ..., n_{.c})$$

Entonces, λ_a (siendo X la variable dependiente) viene dado por:

$$\lambda_a = \frac{W - A}{N - A}$$

De la misma manera, se definen

$$Z = \sum_{j=1}^{c} max(n_{1j}, n_{2j}, ..., n_{rj}) \qquad y \qquad B = max(n_{1.}, n_{2.}, ..., n_{r.})$$

E igualmente, λ_b (siendo Y la variable dependiente) viene dado por:

$$\lambda_b = \frac{Z - B}{N - B}$$

Tanto λ_a como λ_b son medidas de reducción proporcional del error. Consideremos λ_a y dos posibles casos:

- Caso 1: Sólo son conocidas las categorías de la variable dependiente X.
- Caso 2: Son conocidas tanto las categorías de la variable X como las categorías de la variable independiente Y.

En el caso 1, se predice con la categoría de la variable dependiente X que tiene la mayor frecuencia marginal total (moda), que en este caso es $A = max(n_{.1}, ..., n_{.c})$. Entonces, la probabilidad de error es $N - A$; definimos esto como "errores del primer tipo" o E_1. En el caso 2, se predice con la categoría de la variable dependiente X que tiene la mayor frecuencia absoluta (moda) en cada categoría de la variable independiente Y, que en este caso es, para $Y = Y_i$, $i = 1, \ldots, r$:

$$W_i = max(n_{i1}, n_{i2}, ..., n_{ic}).$$

Por tanto,

$$E_2^{(i)} = n_{i.} - W_i \ , \quad i = 1, ..., r$$

y como consecuencia, la probabilidad de error es

$$E_2 = \sum_{i=1}^{r} (n_{i.} - W_i) = N - W \ .$$

A estos errores los definimos como "errores del segundo tipo" o E_2. Entonces, λ_a puede expresarse como

$$\lambda_a = \frac{E_1 - E_2}{E_1} = \frac{N - A - (N - W)}{N - A} = \frac{W - A}{N - A}$$

Como señalaron Goodman y Kruskal en 1954, se observó inmediatamente un problema con las interpretaciones tanto de λ_a como de λ_b. Dado que ambas medidas estaban basadas en los valores modales de las categorías de la variable independiente, cuando los valores modales ocurrían todos en la misma categoría de la variable dependiente, λ_a y λ_b serían cero. Así, λ_a y λ_b son iguales a cero bajo independencia, pero también podían ser iguales a cero para casos distintos de independencia. Esto hace que tanto λ_a como λ_b sean difíciles de interpretar.

De esta manera, como explicaron Goodman y Kruskal en 1954:

1. λ_a es indeterminada si y sólo si la población se encuentra en una columna; es decir, aparece en una categoría de la variable X.

2. En caso contrario, el valor de λ_a se encuentra entre 0 y 1.

3. λ_a es 0 si y sólo si el conocimiento de la clasificación Y no ayuda a predecir la clasificación X.

4. λ_a es 1 si y sólo si el conocimiento de un objeto de la categoría Y especifica completamente su categoría X, es decir, si cada fila de la tabla de clasificación cruzada contiene como máximo un valor distinto de cero.

5. En el caso de la independencia estadística, λ_a, cuando está determinada, es cero. Lo contrario no tiene por qué ser cierto: λ_a puede ser cero sin que haya independencia estadística.

6. λ_a no cambia con ninguna permutación de filas o columnas.

2.2. Medidas t_a y t_b de Goodman y Kruskal

Se consideran dos variables categóricas, X e Y, con la variable X como variable dependiente y la variable Y como variable independiente. La *tabla* 2.1 proporciona la notación para la clasificación cruzada.

El estadístico t_a de Goodman y Kruskal también es una medida de la reducción proporcional del error de predicción.

Sin ninguna información de Y, se predice la categoría de X aleatoriamente, según las distribuciones de frecuencias marginales. Así, si se predice en la categoría X_1, la tasa de error es:

$$\frac{N - n_{.1}}{N} \ .$$

Por tanto, el error esperado es

$$P[error] = \sum_{j=1}^{c} P[error|X_j]P[X_j] = \sum_{j=1}^{c} \frac{N - n_{.j}}{N}\frac{n_{.j}}{N} = \frac{1}{N^2}\sum_{j=1}^{c}(N - n_{.j})n_{.j} \ .$$

Y por consiguiente, el número de errores esperados del primer tipo viene dado por:

$$E_1 = \frac{1}{N}\sum_{j=1}^{c} n_{.j}(N - n_{.j}) \ .$$

Asimismo, conociendo que la variable independiente Y toma la categoría Y_i, el número de errores esperados será:

$$\frac{1}{n_{i.}}\sum_{j=1}^{c} n_{ij}(n_{i.} - n_{ij}) \ .$$

En consecuencia, el número de errores esperados del segundo tipo viene dado por:

$$E_2 = \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{n_{ij}(n_{i.} - n_{ij})}{n_{i.}} \ .$$

El estadístico t_a de Goodman y Kruskal se define, por tanto, como:

$$t_a = \frac{E_1 - E_2}{E_1}$$

Una forma de cálculo eficiente para la t_a de Goodman y Kruskal viene dada por:

$$t_a = \frac{N\displaystyle\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i.}} - \sum_{j=1}^{c} n_{.j}^2}{N^2 - \displaystyle\sum_{j=1}^{c} n_{.j}^2}$$

La medida t_a indica la reducción proporcional del error de predicción dado el conocimiento de la distribución de la variable independiente, Y, conociendo la distribución de la variable dependiente, X. Como se define, t_a es un estimador puntual del parámetro poblacional τ_a de Goodman y Kruskal para la población de la que se obtuvo la muestra de N casos. Si la variable Y se considera la variable dependiente y la variable X la variable independiente, entonces el estadístico de Goodman y Kruskal t_b y el parámetro poblacional asociado τ_b se definen análogamente.

Como t_a toma valores de 0 a 1, posee una interpretación clara y significativa de reducción proporcional del error [Costner, 1965], y se caracteriza por una alta validez intuitiva [Hunter, 1973]; el estadístico de prueba t_a plantea dificultades si la hipótesis nula es que $H_0 : \tau_a = 0$ [Margolin and Light, 1974]. El problema es que la distribución muestral de t_a no es asintóticamente normal bajo la hipótesis nula. En consecuencia, la aplicabilidad de t_a de Goodman y Kruskal a las pruebas típicas de hipótesis nulas se ha visto muy limitada.

Aunque la t_a fue desarrollada por Goodman y Kruskal en 1954, no fue hasta 1963 cuando se estableció la normalidad asintótica para t_a, y hasta 1972 no se obtuvo la varianza asintótica correcta para t_a, pero sólo para $0 < \tau_a < 1$ [Goodman and Kruskal, 1963].

2.3. Una prueba asimétrica de homogeneidad

A veces, se necesita determinar si las proporciones de elementos en un conjunto de categorías mutuamente excluyentes son las mismas para dos o más grupos. Cuando se extraen muestras aleatorias independientes de cada uno de los $g \geq 2$ grupos y luego se clasifican en $r \geq 2$ categorías mutuamente excluyentes, la prueba adecuada es una prueba de homogeneidad de las g distribuciones. Por tanto, una de las distribuciones marginales se conoce, es decir, los totales de frecuencia marginal de fila o columna que indican el número de elementos de cada uno de los g grupos. Esto se denomina muestreo multinomial *producto*, ya que la distribución de muestreo es el producto de g distribuciones multinomiales y la hipótesis nula es que las g distribuciones multinomiales son idénticas.

Una prueba de homogeneidad es diferente de una prueba de independencia, en la que se extrae una única muestra y se clasifica en ambas variables. En una prueba de independencia, ambos conjuntos de totales de frecuencias marginales se conocen sólo después de que se hayan recogido los datos. Esto se denomina muestreo multinomial simple, ya que la distribución del muestreo es una distribución multinomial [Böhning and Holling, 1989]. El test de homogeneidad más utilizado es el test chi-cuadrado de Pearson con $gl = (r-1)(g-1)$ grados de libertad.

El test de homogeneidad chi-cuadrado de Pearson es un test simétrico, que produce un solo valor para una tabla de contingencia $r \times g$. Por el contrario, un test asimétrico produce dos valores dependiendo de la variable que se considere dependiente. Como señala Berkson, si las diferencias son todas en una dirección, una prueba simétrica como la chi-cuadrado es insensible a este hecho [Berkson, 1938].

Una prueba simétrica de homogeneidad, por su naturaleza, excluye la información conocida sobre los datos: qué variable es la independiente y qué variable es la dependiente.

Se consideran las variables A y B, con B la variable dependiente. Sean $B_1, ..., B_r$ las $r \geq 2$ categorías de la variable dependiente, $A_1, ..., A_g$ las $g \geq 2$ categorías de la variable independiente, n_{ij} la frecuencia de celda en la i-ésima fila y j-ésima columna, $i = 1, ., r$ y $j = 1, ..., g$, y N el tamaño total de la muestra. Sean $n_{1.}, ...n_{r.}$ las frecuencias marginales de la variable B y $n_{.1}, \ldots, n_{.g}$ las frecuencias marginales totales de la variable A. La clasificación cruzada de las variables A y B se muestra en la tabla siguiente:

Tabla 2.2: Clasificación cruzada de dos variables categóricas, A y B, con g y r categorías respectivamente

	A_1	A_2	$\ldots$	A_g	Total
B_1	n_{11}	n_{12}	$\ldots$	n_{1g}	$n_{1.}$
B_2	n_{21}	n_{22}	$\ldots$	n_{2g}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_r	n_{r1}	n_{r2}	$\ldots$	n_{rg}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.g}$	N

Aunque nunca se ha propuesto como prueba de homogeneidad, la prueba asimétrica t_b, introducida por primera vez por Goodman y Kruskal en 1954 [Goodman and Kruskal, 1954], es una alternativa atractiva a la prueba simétrica de homogeneidad, la chi-cuadrado.

Si el parámetro poblacional se denota por τ_b, se puede suponer que si $\tau_b = 0$, entonces las distribuciones de B en las subpoblaciones definidas por las categorías de A son homogéneas. Así, si se desea realizar el contraste $\begin{cases} H_o : \tau_b = 0 \\ H_1 : \tau_b \neq 0 \end{cases}$, el estadístico de la prueba viene dado por la versión muestral:

$$t_b = \frac{N \sum_{j=1}^{g} \sum_{i=1}^{r} \frac{n_{ij}^2}{n_{.j}} - \sum_{i=1}^{r} n_{i.}^2}{N^2 - \sum_{i=1}^{r} n_{i.}^2}$$

Si se considera que la variable dependiente es A, el estadístico de la prueba $\begin{cases} H_o : \tau_b = 0 \\ H_1 : \tau_b \neq 0 \end{cases}$ viene dado por dado por:

$$t_a = \frac{N \sum_{i=1}^{r} \sum_{j=1}^{g} \frac{n_{ij}^2}{n_{i.}} - \sum_{j=1}^{g} n_{.j}^2}{N^2 - \sum_{j=1}^{g} n_{.j}^2} .$$

El estadístico de prueba t_b toma valores entre 0 y 1; t_b es 0 si y sólo si hay homogeneidad sobre las r categorías de la variable dependiente, B, para todos los g grupos, y t_b es 1 si y sólo si el conocimiento de la variable A_j, con $j = 1, ..., g$, determina completamente el conocimiento de la variable B_i para $i = 1, ..., r$. Del mismo modo, el estadístico de prueba t_a es 0 si y sólo si existe homogeneidad sobre las g categorías de la variable dependiente, A, para todos los r grupos, y t_a es 1 si y sólo si el conocimiento de la variable B_i, con $i = 1, ..., r$, determina completamente el conocimiento de la variable A_j para $j = 1, ..., g$.

Aunque no existe una equivalencia general para los estadísticos de prueba t_b o t_a con χ^2, existen ciertas relaciones en condiciones especiales: Si $g = 2$, entonces $\chi^2 = N t_b$, y si $g > 2$ y $n_{.j} = N/g$ para $j = 1, ..., g$, entonces $\chi^2 = N(g-1)t_b$. Del mismo modo, si $r = 2$, $\chi^2 = Nt_a$, y si $r > 2$ y $n_{i.} = N/r$ para $i = 1, ..., r$, entonces $\chi^2 = N(r-1)t_a$. De lo anterior, se deduce que si $r = g = 2$, entoces $t_b = t_a = \chi^2/N$, que es el coeficiente de contingencia al cuadrado de la media de Pearson, ϕ^2. Por último, cuando $N \to \infty$, se tiene que $t_b(N-1)(r-1)$ y $t_a(N-1)(g-1)$ se distribuyen según una chi-cuadrado con $(r-1)(g-1)$ grados de libertad.

Existen tres métodos para determinar la distribución de probabilidades del estadístico de prueba t_b o t_a: procedimientos exactos, de remuestreo de Montecarlo y asintóticos. A continuación se recogen algunos comentarios sobre estos métodos, considerando sólo t_b (los métodos son análogos para t_a):

- Valores exactos de probabilidad: Bajo la hipótesis nula, $H_0 : \tau_b = 0$, cada uno de los M posibles ordenamientos de los N elementos sobre las rg categorías de la tabla de contingencia es igualmente probable con distribuciones marginales fijas. Para cada ordenamiento de los datos observados en el conjunto de referencia de todos los ordenamientos posibles, se calcula el estadístico de prueba deseado. El valor exacto de la probabilidad de un estadístico t_b observado es la suma de las probabilidades puntuales hipergeométricas asociadas a valores mayores o iguales a t_b.

- Métodos de remuestreo: Una prueba exacta no es práctica desde el punto de vista computacional, excepto para muestras bastante pequeñas. Un método alternativo que evita las exigencias computacionales de una prueba exacta es una aproximación a través de remuestreo sobre permutaciones. Bajo la hipótesis nula, $H_0 : \tau_b = 0$, se generan y estudian un subconjunto aleatorio de Monte Carlo de todos los posibles ordenamientos igualmente probables de los datos observados. Para cada ordenamiento seleccionado al azar de los datos observados, se calcula el estadístico del test. La probabilidad de remuestreo de Montecarlo de un estadístico t_b observado es simplemente la proporción de los valores seleccionados aleatoriamente de t_b iguales o mayores que el valor observado de t_b.

- Valores asintóticos de probabilidad: Bajo la hipótesis nula, $H_0 : \tau_b = 0$, cuando $N \to \infty$, $t_b(N-1)(g-1)$ se distribuye según una chi-cuadrado con $(r-1)(g-1)$ grados de libertad. El valor de la probabilidad asintótica es la proporción de la chi-cuadrado apropiada igual o mayor que el valor observado de $t_b(N-1)(g-1)$.

2.4. Medidas de concordancia

La medición de la concordancia es un caso especial de medición de la asociación entre dos o más variables. Muchos problemas de investigación estadística requieren medir la concordancia, en lugar de la asociación o la correlación. Los índices de concordancia miden el grado en que un conjunto de medidas de respuesta son idénticas a otro conjunto, es decir, concuerdan.

En 1957, Robinson [Robinson, 1957] explicó que la concordancia estadística requiere que los valores emparejados sean idénticos, mientras que la correlación sólo requiere que los valores emparejados estén descritos por alguna función matemática. Por tanto, la concordancia es una medida más restrictiva que la correlación.

Según Berry, una medida de acuerdo entre evaluadores debería, como mínimo, incorporar siete atributos básicos [Berry and Mielke, 1988]:

- Una medida de concordancia debe ser corregida por el azar, es decir, cualquier coeficiente de concordancia debe reflejar la cantidad de concordancia que excede lo que se esperaría por el azar.
- Una medida de concordancia entre evaluadores posee una ventaja añadida si es directamente aplicable a la evaluación de la fiabilidad.
- Varios investigadores han comentado la simplicidad de la distancia Euclídea para las medidas de concordancia entre evaluadores, señalando que la elevación al cuadrado de las diferencias entre los valores de las escalas es, en el mejor de los casos, cuestionable, aunque reconocen que las diferencias al cuadrado permiten interpretaciones más claras de los coeficientes [Fleiss, 1973][Krippendorff, 1970].
- Toda medida de acuerdo debe tener una base estadística. Una medida de concordancia sin una prueba de significación adecuada está muy limitada en su aplicación a situaciones prácticas de investigación. Los análisis asintóticos son interesantes y útiles, en condiciones de muestras grandes, pero suelen tener una utilidad práctica limitada cuando el tamaño de las muestras es pequeño.
- Una medida de concordancia que sirva para datos multivariantes tiene una ventaja decisiva sobre las medidas de acuerdo univariantes. Así, si un observador localiza un conjunto de datos en un espacio r-dimensional, una medida de concordancia multivariante puede determinar el grado en que un segundo observador localiza el mismo conjunto de datos en el espacio r-dimensional definido.
- Una medida de concordancia debe ser capaz de analizar los datos en cualquier nivel de medición. La medida kappa de Cohen para la concordancia entre evaluadores es, actualmente, la medida de concordancia más utilizada. Se han establecido extensiones de la kappa de Cohen a datos clasificados de forma incompleta por Iachan [Iachan, 1984], a datos ordinales totalmente clasificados y a datos de intervalo por Berry y Mielke en 1988 [Berry and Mielke, 1988].
- Una medida de concordancia debe ser capaz de evaluar la información de más de dos calificadores o jueces. Fleiss propuso una medida de concordancia para múltiples calificadores en una escala nominal [Fleiss, 1971]. Landis y Koch consideraron la concordancia entre varios calificadores en términos de una opinión mayoritaria [Landis and Koch, 1977]. Light se centró en una extensión de la medida kappa de Cohen [Cohen, 1960] de concordancia entre calificadores a múltiples calificadores que se basaba en la media de todos los valores kappa por pares [Light, 1971].

2.4.1. Medida de concordancia A de Robinson

Una de las primeras medidas de concordancia máxima corregida fue desarrollada por W.S. Robinson en 1957 [Robinson, 1957]. Supongamos que hay $k = 2$ jueces y califican independientemente N objetos. Robinson propuso una nueva medida de concordancia basada en el coeficiente de correlación intraclase que denominó A. Sean dos conjuntos de valoraciones con N pares de valores, (X_{1i}, X_{2i}) con $i = 1, \ldots, N$. Robinson definió A como

$$A = 1 - \frac{D}{D_{max}},$$

donde D (Desacuerdo) viene dado por:

$$D = \sum_{i=1}^{N}(X_{1i} - \bar{X}_i)^2 + \sum_{i=1}^{N}(X_{2i} - \bar{X}_i)^2$$

y X_{1i} = valor de X_1 para el i-ésimo par de valoraciones, X_{2i} = valor de X_2 para el i-ésimo par de valoraciones, $\bar{X}_i$ = la media de X_1 y X_2 para el i-ésimo par de valoraciones, es decir, $\bar{X}_i = \frac{1}{2}(X_{1i} + X_{2i})$.

Para encontrar una medida de concordancia relativa, más que absoluta, Robinson estandarizó D por su rango de variación posible, dado por:

$$D_{max} = \sum_{i=1}^{N}(X_{1i} - \bar{X})^2 + \sum_{i=1}^{N}(X_{2i} - \bar{X})^2,$$

donde la media viene dada por:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_{1i} + \sum_{i=1}^{N} X_{2i}}{2N}$$

Coeficiente de correlación intraclase

El coeficiente de correlación intraclase (r_I) entre N pares de observaciones sobre dos variables es, por definición, el momento producto ordinario de Pearson (interclase) entre $2N$ pares de observaciones, de los cuales los primeros N son las observaciones originales, y los segundos N las observaciones originales con X_{1i} sustituyendo a X_{2i} y viceversa, con $i = 1, ...N$ [Fisher, 1934]:

$$r_I = \frac{N\sum_{i=1}^{N} X_{1i}X_{2i} - \sum_{i=1}^{N} X_{1i}\sum_{i=1}^{N} X_{2i}}{\sqrt{[N\sum_{i=1}^{N} X_{1i}^2 - (\sum_{i=1}^{N} X_{1i})^2][N\sum_{i=1}^{N} X_{2i}^2 - (\sum_{i=1}^{N} X_{2i})^2]}}$$

Para el caso de dos variables, las relaciones entre el coeficiente de concordancia de Robinson y el coeficiente de correlación intraclase vienen dadas por:

$$r_I = 2A - 1; A = \frac{r_I + 1}{2}$$

Por tanto, en el caso de dos variables, la correlación intraclase es una función lineal simple del coeficiente de concordancia.

Para $k > 2$ conjuntos de valoraciones, las relaciones entre el coeficiente de correlación intraclase y la A de Robinson no son tan simples, pero sigue siendo una función lineal del coeficiente de concordancia, y vienen dadas por:

$$r_I = \frac{kA - 1}{k - 1}; A = \frac{r_I(k - 1) + 1}{k}.$$

De las expresiones anteriores se observa que el valor del coeficiente intraclase no depende sólo de A sino también de k, el número variables. El rango de A de Robinson siempre incluye los valores desde 0 hasta 1, independientemente del número de observaciones. Por lo tanto, las comparaciones entre los coeficientes de concordancia basados en diferentes números de variables son equiparables. El límite superior del coeficiente de correlación intraclase es siempre la unidad, pero su límite inferior es $-1/(k-1)$. Para $k=2$ variables, el límite inferior de r_I es -1, pero para $k=3$ variables es $-1/2$, para $k=4$ es $-1/3$, para $k=5$ el límite inferior es $-1/4$, y así sucesivamente.

2.4.2. Medida de concordancia π de Scott

Una de las primeras medidas de concordancia corregida por el azar[1] fue introducida por William Scott en 1955 [Scott, 1955]. Supongamos que dos jueces o calificadores clasifican independientemente cada una de las N observaciones en una de las c categorías. Las clasificaciones resultantes pueden mostrarse en una tabla de contingencia $c \times c$, con las frecuencias absolutas en cada celda. Sea $n_{i.}$ la frecuencia marginal de la i-ésima fila, $i=1,...r$; sea $n_{.j}$ la frecuencia marginal de la j-ésima columna, $j=1,....c$ y sea

$$N = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$$

la frecuencia total de la tabla. El coeficiente de concordancia de Scott para los datos de nivel nominal viene dado por:

$$\pi = \frac{p_o - p_e}{1 - p_e}$$

donde

$$p_o = \frac{1}{N}\sum_{i=1}^{c} n_{ii} \qquad y \qquad p_e = \frac{1}{4N^2}\sum_{k=1}^{c}(n_{.k} + n_{k.})^2 \ .$$

Aunque la π de Scott es interesante desde una perspectiva histórica, esta medida ha caído en desuso. Basada en proporciones conjuntas, la π de Scott asume que los dos jueces tienen la misma distribución de respuestas. La medida κ de Cohen no hace esta suposición y, en consecuencia, ha surgido como la medida preferida de concordancia entre evaluadores corregida por el azar para dos jueces/calificadores.

2.4.3. Medida de concordancia κ de Cohen

Actualmente, la medida más popular de concordancia entre dos jueces o calificadores es la medida corregida por el azar, propuesta por primera vez por Jacob Cohen en 1960 y denominada kappa [Cohen, 1960]. La kappa de Cohen mide la magnitud de la concordancia entre $b=2$ observadores en la asignación de N objetos a un conjunto de c categorías disjuntas y desordenadas. En 1968, Cohen propuso una versión de kappa que permitía ponderar las c categorías [Cohen, 1968]. Mientras que el kappa original (no ponderado) no distinguía entre magnitudes de desacuerdo, el kappa ponderado incorporaba la magnitud de cada desacuerdo y proporcionaba un crédito parcial para las discordancias cuando la

[1] https://www.cis.es/cis/opencm/ES/2_bancodatos/estudios/listaEstudiosYear.jsp?year=2021

concordancia no era completa. El enfoque habitual consiste en asignar pesos a cada par de desacuerdos, con pesos mayores que indican un mayor desacuerdo.

Tanto en el caso no ponderado como en el ponderado, kappa es igual a +1 cuando se produce una concordancia perfecta entre dos o más jueces, 0 cuando la concordancia es igual a la esperada en condiciones de independencia, y negativo cuando la concordancia es inferior a la esperada por azar. En esta sección se estudiará únicamente la kappa no ponderada, ya que es la que se usa normalmente para datos categóricos no ordenados.

Supongamos que dos jueces o calificadores clasifican independientemente cada una de las N observaciones en una de las c categorías desordenadas, exhaustivas y mutuamente excluyentes. Las clasificaciones resultantes pueden mostrarse en una clasificación cruzada $c \times c$ con proporciones para las entradas de las celdas:

Tabla 2.3: Clasificación cruzada $c \times c$ con proporciones en celdas

Fila\Columna	**1**	**2**	$\dots$	**c**	**Total**
1	p_{11}	p_{12}	$\cdots$	p_{1c}	$p_{1.}$
2	p_{21}	p_{22}	$\cdots$	p_{2c}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c	p_{c1}	p_{c2}	$\cdots$	p_{cc}	$p_{c.}$
Total	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.c}$	$p_{..}$

donde $p_{i.}$ denota la proporción marginal de la i-ésima fila, $i = 1, ...c$; $p_{.j}$ denota la proporción marginal de la j-ésima columna, $j = 1, ...c$ y $p_{..} = 1,00$. Usando esta notación, el coeficiente kappa no ponderado de Cohen para datos de nivel nominal viene dado por:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

donde

$$p_o = \sum_{i=1}^{c} p_{ii} \qquad y \qquad p_e = \sum_{i=1}^{c} p_{i.}p_{.i} \ .$$

Considerando los pesos

$$w_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases},$$

se tiene que

$$p_o = 1 - \sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}p_{ij} \qquad y \qquad p_e = 1 - \sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}p_{i.}p_{.j} \ .$$

Por tanto:

$$\kappa = 1 - \frac{1 - p_o}{1 - p_e} = 1 - \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}p_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}p_{i.}p_{.j}} \ .$$

El estimador de este coeficiente basado en frecuencias, el coeficiente kappa de Cohen muestral, puede definirse en términos de valores de frecuencia absoluta como

$$\kappa = \frac{\sum_{i=1}^{c} n_{ii} - \sum_{i=1}^{c} E_{ii}}{N - \sum_{i=1}^{c} E_{ii}},$$

donde n_{ii} denota el valor de frecuencia de celda observado en la diagonal principal de la tabla de concordancia $c \times c$ y E_{ii} denota un valor de frecuencia de celda esperado en la diagonal principal, es decir,

$$E_{ii} = \frac{n_{i.}n_{.i}}{N}, \qquad i = 1, .., c.$$

En la configuración de la tabla anterior, p_o es la proporción observada de observaciones en las que los jueces están de acuerdo, p_e es la proporción de observaciones para las que se espera una concordancia es la proporción de concordancia esperada por el azar, $p_o - p_e$ es la proporción de concordancia más allá de la esperada por azar, $1 - p_e$ es la proporción máxima posible de concordancia más allá de lo esperado por el azar, y el estadístico de la prueba kappa de Cohen es la proporción de concordancia entre los dos jueces, una vez eliminada la concordancia por azar.

2.4.4. Aplicación con varios jueces

Originalmente, la medida κ de Cohen de la concordancia entre evaluadores corregida por el azar fue diseñada sólo para $b = 2$ jueces. En esta sección, se introduce un procedimiento para calcular κ (no ponderado) con múltiples jueces. Aunque el procedimiento es apropiado para cualquier número de $c \geq 2$ categorías disjuntas y desordenadas y $b \geq 2$ jueces, la descripción del procedimiento se limita a $b = 3$ jueces independientes.

Sean $b = 3$ jueces que clasifican independientemente N objetos en c categorías disjuntas y desordenadas. La clasificación puede conceptualizarse como una tabla de contingencia $c \times c \times c$ con c filas, c columnas y c niveles. Sea n_{ijk} el número de objetos clasificados en la i-ésima categoría por el juez 1, en la j-ésima categoría por el juez 2 y en la k-ésima categoría por el juez 3 y sean R_i , C_j y S_k las frecuencias marginales de las filas, las columnas y los niveles para $i, j, k = 1, ..., c$. La frecuencia total viene dada por:

$$N = \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} n_{ijk} \ .$$

Extendiendo la expresión obtenida para 2 jueces, el estadístico kappa no ponderado de Cohen para una tabla de contingencia de tres dimensiones está dado por:

$$\kappa = 1 - \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} n_{ijk}}{\sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} R_i C_j S_k} \qquad (2.1),$$

donde w_{ijk} son los "pesos" de desacuerdo asignados a cada celda para $i, j, k = 1, ., c$. Para el kappa no ponderado, los pesos de desacuerdo vienen dados por:

$$w_{ijk} = \begin{cases} 0 & i = j = k \\ 1 & c.c \end{cases} .$$

Obviamente, para el estimador del coeficiente kappa de Cohen poblacional, también se puede plantear el estudio de la significación de la medida, es decir, $\begin{cases} H_o : \kappa = 0 \\ H_1 : \kappa \neq 0 \end{cases}$.

Dada una tabla de contingencia $c \times c \times c$ con N objetos clasificados de forma cruzada por $b = 3$ jueces independientes, una prueba de permutación exacta consiste en generar todos los ordenamientos posibles, igualmente probables, de los N objetos en las c^3 celdas, preservando las distribuciones marginales de frecuencia. Para cada ordenamiento de las frecuencias de las celdas, se calcula el estadístico kappa no ponderado, κ, y la probabilidad exacta puntual hipergeométrica bajo la hipótesis nula, $p(n_{ijk}|R_i, C_j, S_k, N)$, donde

$$p(n_{ijk}|R_i, C_j, S_k, N) = \frac{(\prod_{i=1}^{c} R_i!)(\prod_{j=1}^{c} C_j!)(\prod_{k=1}^{c} S_k!)}{(N!)^{b-1} \prod_{i=1}^{c} \prod_{j=1}^{c} \prod_{k=1}^{c} n_{ijk}!} .$$

Si κ_o denota el valor del estadístico kappa no ponderado observado, el valor exacto de la cola de probabilidad asociada al valor κ_o bajo la hipótesis nula viene dado por:

$$P(\kappa_o) = \sum_{l=1}^{M} \Psi_l(n_{ijk}|R_i, C_j, S_k, N) \ ,$$

donde

$$\Psi_l(n_{ijk}|R_i, C_j, S_k, N) = \begin{cases} p(n_{ijk}|R_i, C_j, S_k) & si \quad \kappa \geq \kappa_o \\ 0 & c.c \end{cases}$$

y M denota el número total de posibles ordenamientos de frecuencia de celdas igualmente probables en el conjunto de referencia de todos los ordenamientos posibles de las frecuencias de las celdas, dado las distribuciones de frecuencias marginales observadas. Cuando M es muy grande, como es típico de contingencia multidireccional, las pruebas exactas no son prácticas y es necesario recurrir a los procedimientos de remuestreo de Monte Carlo. En estos casos, una muestra aleatoria de los M posibles ordenamientos igualmente probables de las frecuencias de las celdas proporciona una comparación de los estadísticos de la prueba κ calculados en L tablas aleatorias multidireccionales con el estadístico de la prueba κ calculado en la tabla de contingencia multidireccional observada.

Un algoritmo eficiente de remuestreo de Monte Carlo para generar ordenaciones aleatorias de frecuencias de celdas para tablas de contingencia multidireccionales con distribuciones de frecuencias marginales fijas fue desarrollado por Mielke, Berry y Johnston en 2007 [Mielke et al., 2007]. Para una tabla de contingencia de tres dimensiones con r filas, c columnas y s niveles, el algoritmo de remuestreo se presenta en 12 sencillos pasos:

PASO 1. Construir una tabla de contingencia $r \times c \times s$ a partir de los datos observados.

PASO 2. Obtener los totales de frecuencia marginal fija $R_1, ..., R_r, C_1, ..., C_c, S_1, ...S_s$, y el total de frecuencias N. Establecer un contador de remuestreo $JL = 0$, y fijar L igual al número de muestras deseado.

PASO 3. Establecer el contador de remuestreo $JL = JL + 1$.

PASO 4. Establecer los contadores de frecuencia marginal $JR_i = R_i$ para $i = 1, ., r$; $JC_j = C_j$ para $j = 1, ..., c$; $JS_k = S_k$ para $k = 1, ...s$, y $M = N$.

PASO 5. Establecer $n_{ijk} = 0$ para $i = 1, ..., r, j = 1, ...c$, y $k = 1, ..., s$, y establecer los contadores de filas, columnas y niveles IR, IC e IS iguales a cero.

PASO 6. Crear las distribuciones de probabilidad acumulada PR_i , PC_j y PS_k a partir de los totales de frecuencia marginal ajustados JR_i , JC_j y JS_k para $i = 1, ., r, j = 1, ...c$, y $k = 1, ..., s$, donde $PR_1 = JR_1/M$ y $PR_i = PR_{i-1} + JR_i/M$ para $i = 1, ..., r$; $PC_1 = JC_1/M$ y $PC_j = PC_{j-1} + JC_j/M$ para $j = 1, ..., c$ y $PS_1 = JS_1/M$ y $PS_k = PS_{k-1} + JS_k/M$ para $k = 1, ..., s$.

PASO 7. Generar tres números pseudoaleatorios uniformes U_r , U_c y U_s sobre $[0, 1)$ y establecer los índices de fila, columna y nivel $i = j = k = 1$, respectivamente.

PASO 8. Si $U_r \leq PR_i$, entonces $IR = i$, $JR_i = JR_i - 1$, e ir al PASO 9; en caso contrario $i = i + 1$ y se repite el PASO 8.

PASO 9. Si $U_c \leq PC_j$, entonces $IC = j$, $JC_j = JC_j - 1$, e ir al PASO 10; en caso contrario, $j = j + 1$ y se repite el PASO 9.

PASO 10. Si $U_s \leq PS_k$, entonces $IS = k$, $JS_k = JS_k - 1$, e ir al PASO 11; en caso contrario, $k = k + 1$ y se repite el PASO 10.

PASO 11. Establecer $M = M - 1$ y $n_{IR,IC,IS} = n_{IR,IC,IS} + 1$. Si $M > 0$, ir al PASO 4; en caso contrario, obtener el estadístico de prueba requerido.

PASO 12. Si $JL < L$, ir al PASO 3; en caso contrario, STOP.

Al terminar el procedimiento de remuestreo, se obtiene la κ de Cohen, como se indica en la ecuación (2.1), para cada una de las L tablas de contingencia aleatorias de tres dimensiones, dadas las distribuciones de frecuencia marginal fijas. Si κ_o denota el valor observado de κ, entonces bajo la hipótesis nula, el valor de la probabilidad aproximada del remuestreo para κ_o viene dado por:

$$P(\kappa_o) = \frac{1}{L} \sum_{l=1}^{L} \Psi_l(\kappa),$$

donde

$$\Psi_l(\kappa) = \begin{cases} 1 & si \quad \kappa \geq \kappa_o \\ 0 & c.c \end{cases} .$$

2.5. Medida de asociación d_N^c de Leik y Gove

En 1971, Robert Leik y Walter Gove propusieron una nueva medida de asociación nominal basada en comparaciones por pares de las diferencias entre las observaciones

[Leik and Gove, 1971]. Sugirieron una medida de asociación de reducción proporcional en el error que fuese corregida para la cantidad máxima real de asociación, dadas las distribuciones de frecuencias marginales observadas.

Consideremos dos variables de nivel nominal que se han clasificado de forma cruzada en una tabla de contingencia $r \times c$, donde r y c denotan el número de filas y columnas, respectivamente (véase *tabla* 2.1).

Si Y y X representan las variables de fila y columna, respectivamente, hay $N(N-1)/2$ pares de objetos en la tabla que pueden dividirse en cinco tipos de pares exhaustivos y mutuamente excluyentes:

- Pares concordantes, es decir, pares de objetos que están clasificados en el mismo orden tanto en la variable X como en la variable Y.
- Pares discordantes, es decir, pares de objetos que se clasifican en un orden en la variable X y en el orden inverso en la variable Y.
- Pares empatados en la variable Y pero que difieren en la variable X.
- Pares empatados en la variable X pero que difieren en la variable Y.
- Pares empatados en ambas variables X e Y.

Para una tabla de contingencia $r \times c$, los pares concordantes vienen dados por:

$$C = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \Big(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} n_{kl} \Big) \ ,$$

los pares discordantes vienen dados por:

$$D = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \Big(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} n_{kl} \Big) \ ,$$

los pares de objetos empatados en la variable X pero que difieren en la variable Y vienen dados por:

$$T_x = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} \Big(\sum_{k=j+1}^{c} n_{ik} \Big) \ ,$$

los pares de objetos empatados en la variable Y pero que difieren en la variable X vienen dados por:

$$T_y = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} \Big(\sum_{k=i+1}^{r} n_{kj} \Big) \ ,$$

y los pares de objetos empatados en la variable X y en la variable Y vienen dados por:

$$T_{xy} = \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}(n_{ij} - 1) \ .$$

Entonces,

$$C + D + T_x + T_y + T_{xy} = \frac{N(N-1)}{2} \ .$$

Usando la tabla de contingencia observada, se contruye la tabla de contingencia esperada bajo la hipótesis de independencia entre ambas variables de la siguiente forma:

$$E_{ij} = \frac{n_{i.}n_{.j}}{N} \qquad con \quad i = 1, ..., r \quad y \quad j = 1, .., c \ .$$

A partir de la tabla de contingencia esperada se obtienen los parámetros C', D', T'_x, T'_y y T'_{xy} (véase [Berry et al., 2010]):

$$T'_{xy} = \frac{1}{2N^2}(\sum_{i=1}^{r} n_{i.}^2)(\sum_{j=1}^{c} n_{.j}^2) - \frac{N}{2} \ ; \qquad T'_y = \frac{1}{2}\sum_{i=1}^{r} n_{i.}^2 - \frac{N}{2} - T'_{xy} \ ;$$

$$T'_x = \frac{1}{2}\sum_{j=1}^{c} n_{.j}^2 - \frac{N}{2} - T'_{xy} \ ; \qquad C' = D' = \frac{1}{2}[\frac{N(N-1)}{2} - T'_x - T'_y - T'_{xy}] \ .$$

Hay que tener en cuenta que C', D', T'_x, T'_y y T'_{xy} se calculan sobre los totales de frecuencia marginal de la tabla de contingencia observada, que son invariantes bajo permutaciones.

El estadístico de prueba d_N^c se basa en tres tablas de contingencia: la tabla de valores observados, la tabla de valores esperados y una tabla de valores máximos que se describirá a continuación.

Un algoritmo para generar un ordenamiento de las frecuencias de las celdas en una tabla de contingencia $r \times c$ que proporciona el valor máximo de un conjunto estadístico es:

PASO 1: Mantener las frecuencias marginales observadas de una tabla de contingencia $r \times c$ y eliminar los valores de frecuencias de celdas (n_{ij}).

PASO 2: Si algún par de frecuencias marginales, uno de cada conjunto de marginales, son iguales entre sí, introducir ese valor en la tabla como n_{ij} y restar el valor de los dos totales de frecuencia marginal. Repetir el PASO 2 hasta que no haya dos totales de frecuencia marginal iguales. Si todos los totales de frecuencia marginal se han reducido a cero, ir al PASO 5; de lo contrario, ir al PASO 3.

PASO 3: Observar la frecuencia marginal más grande que queda en cada conjunto e introducir el menor de los dos valores en n_{ij} . A continuación, restar ese valor (más pequeño) de los dos totales de frecuencias marginales.

PASO 4: Si todos los totales de frecuencia marginal se han reducido a cero, ir al PASO 5; En caso contrario, ir al PASO 2.

PASO 5: Establecer los valores n_{ij} restantes como 0, $i = 1, ..., r$ y $j = 1, ..., c$.

Denotamos la doble comilla ($''$) como una suma de pares calculada sobre los valores de frecuencias de las celdas maximizadas. Entonces, se pueden obtener los valores asociados a esta tabla:

$$C'' = \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} n_{ij}(\sum_{k=i+1}^{r}\sum_{l=j+1}^{c} n_{kl}) \ ; \qquad D'' = \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} n_{i,c-j+1}(\sum_{k=i+1}^{r}\sum_{l=1}^{c-j} n_{kl}) \ ;$$

$$T''_x = \sum_{i=1}^{r}\sum_{j=1}^{c-1} n_{ij}(\sum_{k=j+1}^{c} n_{ik}) \; ; \qquad T''_y = \sum_{j=1}^{c}\sum_{i=1}^{r-1} n_{ij}(\sum_{k=i+1}^{r} n_{kj}) \; ;$$

$$T''_{xy} = \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(n_{ij}-1) \; .$$

Entonces, de la misma forma que antes,

$$C'' + D'' + T''_x + T''_y + T''_{xy} = \frac{N(N-1)}{2} \; .$$

Dados los valores observados, esperados y maximizados de C, D, T_x , T_y y T_{xy}, los errores del primer tipo (E_1) -la variación entre la independencia y la máxima asociación- vienen dados por:

$$E_1 = T'_y - T''_y$$

y los errores del segundo tipo (E_2) -la variación entre la tabla observada y la tabla de máxima asociación- vienen dados por:

$$E_2 = T_y - T''_y \; .$$

Entonces, siguiendo el patrón de las medidas de asociación de reducción proporcional del error,

$$d_N^c = \frac{E_1 - E_2}{E_1} = \frac{(T'_y - T''_y) - (T_y - T''_y)}{T'_y - T''_y} = \frac{T'_y - T_y}{T'_y - T''_y} \; .$$

Dado que d_N^c es una medida simétrica, el número de valores empatados en la variable X puede utilizarse en lugar del número de valores empatados en la variable Y. Así pues,

$$d_N^c = \frac{T'_x - T_x}{T'_x - T''_x} \; .$$

Alternativamente, d_N^c puede definirse en términos del número de valores empatados tanto en X como en Y. Así,

$$d_N^c = \frac{T'_{xy} - T_{xy}}{T'_{xy} - T''_{xy}} \; .$$

Por último, como los datos son categóricos, C y D pueden considerarse agrupados. Por lo tanto,

$$d_N^c = \frac{(C' + D') - (C + D)}{(C' - D') - (C'' + D'')} \; .$$

Como señalan Leik y Gove, para ayudar a interpretar la relación entre las variables X e Y, sería preferible determinar explícitamente el número de pares perdidos por los requisitos marginales de la tabla de contingencia. La asociación puede entonces ser definida dentro de esos límites, permitiendo que el índice alcance la unidad si las frecuencias de las celdas están tan cerca de un patrón perfecto como lo permitan las distribuciones marginales. Así pues, la proporción de casos que se considera es

$$1 - \frac{2(T''_x + T''_y)}{N(N-1)} \; .$$

Para obtener un test de la significación de la medida de asociación, se puede aplicar la estrategia de los test de permutaciones, bien exactos o bien aproximados a través de la aplicación del método de Montecarlo. Consideremos la expresión

$$d_N^c = \frac{T'_y - T_y}{T'_y - T''_y} \ .$$

Es evidente que T'_y y T''_y son invariantes bajo permutación. Por lo tanto, la probabilidad de d_N^c bajo la hipótesis nula puede determinarse sólo por la distribución discreta de permutación de T_y, que se obtiene fácilmente de la tabla de contingencia observada. Los métodos estadísticos de permutación exacta son muy eficaces cuando sólo se calcula la parte variable del estadístico de prueba definido en cada uno de los M ordenamientos posibles de los datos observados; en este caso, T_y.

Capítulo 3

Medidas para Variables Ordinales (Parte 1)

Las medidas de las relaciones entre dos variables de nivel ordinal suelen ser más informativas que las medidas entre simples variables de nivel nominal (categóricas), ya que las categorías disjuntas y ordenadas suelen contener más información que las categorías disjuntas y desordenadas. Las medidas de asociación para dos variables de nivel ordinal suelen ser de dos tipos: las que se basan en las diferencias entre pares, como las medidas τ_a y τ_b de Kendall y la medida γ de Goodman y Kruskal, y las que se basan otros criterios, como la medida kappa ponderada de Cohen de concordancia entre evaluadores y el análisis ridit de Bross.

En este capítulo se incluyen las medidas de asociación ordinal τ_a y τ_b de Kendall, τ_c de Stuart, las medidas asimétricas d_{yx} y d_{xy} de Somers, las medidas $d_{y.x}$ y $d_{x.y}$ de Kim, la medida e de Wilson o el coeficiente de correlación rango-biserial de Cureton [AÑADIR LAS DEL OTRO TEMA CUANDO LOS JUNTE].

3.1. Principales medidas de Asociación Ordinal por pares

Una serie de medidas de asociación para dos variables de nivel ordinal se basan en comparaciones por pares de las diferencias entre los rangos. El estadístico de prueba S, definido por Maurice Kendall en 1938 [Kendall, 1938], desempeña un papel importante; este estadístico se expresa a menudo como $S = C - D$, donde C y D indican el número de pares concordantes y discordantes, respectivamente.

Sean dos variables ordinales clasificadas de forma cruzada en una tabla de contingencia $r \times c$, donde r y c denotan el número de filas y columnas, respectivamente. Sean $n_{i.}$, $n_{.j}$, y n_{ij} los totales de frecuencia marginal de las filas, los totales de las frecuencias marginales de las columnas y el número de individuos en la celda ij, respectivamente, para $i = 1, ..., r$ y $j = 1, ...c$, y sea N el número total de individuos en la tabla de contingencia $r \times c$, es decir

$$n_{i.} = \sum_{j=1}^{c} n_{ij} \ , \qquad n_{.j} = \sum_{i=1}^{r} n_{ij} \qquad y \qquad N = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \ .$$

La siguiente tabla muestra una notación convencional para una tabla de contingencia $r \times c$ para dos variables categóricas, X_i para $i = 1, ..., r$ e Y_j para $j = 1, ..., c$, con el orden determinado para X por $X_1 < X_2 < \ldots < X_r$ y para Y por $Y_1 < Y_2 < \ldots < Y_c$:

Tabla 3.1: Clasificación cruzada de dos variables ordinales, X e Y.

X\Y	Y_1	Y_2	$\ldots$	Y_c	**Total**
X_1	n_{11}	n_{12}	$\ldots$	n_{1c}	$n_{1.}$
X_2	n_{21}	n_{22}	$\ldots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.c}$	N

Si X e Y representan las variables de fila y columna, respectivamente, hay $N(N-1)/2$ pares de individuos en la tabla que pueden dividirse en cinco tipos de pares mutuamente exhaustivos y exclusivos: pares concordantes, pares discordantes, pares empatados en la variable X pero no en la variable Y, pares empatados en la variable Y pero no en la variable X, y pares empatados en ambas variables.

Para una tabla de contingencia $r \times c$, los pares concordantes (pares de objetos que están clasificados en el mismo orden tanto en la variable X como en la variable Y) vienen dados por:

$$C = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} n_{kl} \right) ,$$

los pares discordantes (pares de objetos que se clasifican en un orden en la variable X y en el orden inverso en la variable Y) vienen dados por:

$$D = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \left(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} n_{kl} \right) ,$$

los pares de objetos empatados en la variable X pero que difieren en la variable Y vienen dados por:

$$T_x = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=j+1}^{c} n_{ik} \right) ,$$

los pares de objetos empatados en la variable Y pero que difieren en la variable X vienen dados por:

$$T_y = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} \left(\sum_{k=i+1}^{r} n_{kj} \right) ,$$

y los pares de objetos empatados en la variable X y en la variable Y vienen dados por:

$$T_{xy} = \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(n_{ij}-1) \ .$$

Entonces,

$$C + D + T_x + T_y + T_{xy} = \frac{N(N-1)}{2} \ .$$

Dados C, D, T_x, T_y y N, se suelen definir seis medidas de asociación ordinal, cada una de las cuales tiene el mismo numerador, $S = C - D$, pero diferentes denominadores[1][2].

3.1.1. Medida τ_a de Kendall de Asociación Ordinal

La primera de estas medidas de asociación por pares propuesta fue la τ_a de Kendall [Kendall, 1948], que es una medida simétrica de asociación ordinal y se diseñó originalmente para medir la asociación entre dos conjuntos de valores de rangos no empatados, donde los dos conjuntos de puntuaciones de rango se etiquetan habitualmente como X e Y, aunque los rangos también pueden representarse en una tabla de contingencia $r \times c$ donde $n_{i.} = n_{.j} = 1$ para $i = 1, ..., r$ y $j = 1, ..., c$.

La τ_a de Kendall se define simplemente como la diferencia entre las proporciones de pares concordantes y discordantes, está dada por

$$\tau_a = \frac{C}{\frac{N(N-1)}{2}} - \frac{D}{\frac{N(N-1)}{2}} = \frac{C-D}{\frac{N(N-1)}{2}} = \frac{2S}{N(N-1)} \ .$$

La τ_a de Kendall se presenta como una alternativa al coeficiente de correlación de rango de Spearman [Kraft and van Eeden, 1968].

3.1.2. Medida τ_b de Kendall de Asociación Ordinal

Cuando existen valores empatados, la medida de asociación ordinal τ_a de Kendall no es la mejor opción, ya que ignora los dos conjuntos de valores empatados, T_x y T_y. Por esta razón, Kendall desarrolló τ_b [Kendall, 1948], una alternativa a τ_a, dada por

$$\tau_b = \frac{S}{\sqrt{(C + D + T_x)(C + D + T_y)}} \ .$$

La τ_b de Kendall es una medida estríctamente monótona de asociación ordinal, es decir, para cada aumento de categoría en la variable X, se espera que haya un aumento de categoría en la variable Y. Por consiguiente, τ_b sólo puede alcanzar límites de ± 1 para tablas de contingencia en las que $r = c$ y las distribuciones de frecuencias marginales de filas y columnas sean idénticas. Más concretamente, τ_b no puede alcanzar generalmente valores de ± 1 debido a la desigualdad de Cauchy:

[1] https://www.cis.es/cis/opencm/ES/2_bancodatos/estudios/listaEstudiosYear.jsp?year=2021

[2] En realidad hay más de seis medidas de asociación ordinal basadas en comparaciones por pares; aquí sólo se tratan las seis medidas más comunes.

El cuadrado de la suma de los productos de dos conjuntos será igual o menor que el producto de las sumas al cuadrado de dos conjuntos. Formalmente, para las variables X e Y,

$$(\sum_{i=1}^{N} x_i y_i))^2 \leq \sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2 \ .$$

En consecuencia, el numerador de τ_b será igual o menor que el denominador, permitiendo que τ_b alcance ± 1 sólo cuando todas las observaciones se concentren en una de las dos diagonales principales de la tabla de contingencia. Si ninguna frecuencia marginal es cero, esto significa que τ_b sólo puede alcanzar ± 1 para una tabla de contingencia cuadrada con distribuciones de frecuencias marginales idénticas. Es importante señalar que, dado que las categorías están ordenadas, las distribuciones de frecuencias marginales deben ser idénticas, no simplemente equivalentes.

Por lo tanto, la τ_b de Kendall no es la medida más apropiada de asociación ordinal para la tabla de contingencia $r \times c$, con $r \neq c$.

3.1.3. Medida τ_c de Stuart de Asiciación Ordinal

Alan Stuart propuso la τ_c [Stuart, 1953], que modifica la τ_b de Kendall para las tablas de contingencia en las que $r \neq c$, la medida viene dada por

$$\tau_c = \frac{2mS}{N^2(m-1)} \ ,$$

donde $m = min(r, c)$.

Si N es un múltiplo de m y $r = c$ con distribuciones marginales de frecuencia idénticas, de modo que todas las observaciones caen en la diagonal de la tabla de contingencia y todas las frecuencias de las celdas son iguales, el valor máximo de la S de Kendall viene dado por

$$S_{max} = \frac{N^2(m-1)}{2m} \ .$$

Entonces, si $N = m$,

$$\frac{N^2(m-1)}{2m} = \frac{N^2(N-1)}{2N} = \frac{N(N-1)}{2} \ .$$

Sin embargo, si N no es un múltiplo de m, la expresión de S_{max} sigue siendo un límite superior que no se puede alcanzar. De ello se concluye que la τ_c de Stuart puede alcanzar a veces (y para N grande, puede alcanzar casi siempre) ± 1.

3.1.4. Medida γ de Goodman y Kruskal

En 1954 Goodman y Kruskal desarrollaron una nueva medida de asociación simétrica para dos variables de nivel ordinal que denominaron gamma, γ [Goodman and Kruskal, 1954]. Esta medida de asociación ordinal es una medida de reducción proporcional al error que se basa únicamente en los pares no empatados, C y D, y viene dada por

$$\gamma = \frac{S}{C+D} = \frac{C-D}{C+D} = \frac{C}{C+D} - \frac{D}{C+D} \ .$$

Por tanto, de la expresión se deduce que γ es simplemente la diferencia entre las proporciones de los pares iguales y no iguales, ignorando todos los pares empatados, es decir, T_x, T_y y T_{xy}.

Según Goodman y Kruskal, existe un problema potencial con γ, y es que es inestable en varios "puntos de corte", es decir, γ tiende a aumentar a medida que las categorías de una tabla de contingencia se colapsan porque no tiene en cuenta los pares empatados (y el número de pares empatados aumenta a medida que la tabla se colapsa). Además, γ es una medida de asociación ordinal (no estrictamente) monótona, es decir, para cada aumento (disminución) de la categoría ordenada en la variable X, la variable Y aumenta (disminuye) o permanece igual.

3.1.5. Medidas d_{yx} y d_{xy} de Somers

En 1962 el sociólogo Robert Somers [Somers, 1962] propuso dos alternativas asimétricas a la medida γ de Goodman y Kruskal, dadas por

$$d_{yx} = \frac{C - D}{C + D + T_y} = \frac{S}{C + D + T_y} \ ,$$

donde T_y denota el número de pares empatados en la variable Y pero no empatados en la variable X, y

$$d_{xy} = \frac{C - D}{C + D + T_x} = \frac{S}{C + D + T_x} \ ,$$

donde T_x denota el número de pares empatados en la variable X pero no empatados en la variable Y.

A diferencia de las cuatro medidas simétricas, τ_a, τ_b, τ_c y γ, las medidas d_{yx} y d_{xy} de Somers dependen de qué variable, Y o X, se considera que es la dependiente. Observando las ecuaciones anteriores, se ve que Somers incluyó en los denominadores de d_{yx} y d_{xy} el número de valores empatados en la variable dependiente: T_y para d_{yx} y T_x para d_{xy}. La razón para incluir los valores empatados es simplemente que cuando la variable Y es la variable dependiente (d_{yx}), si dos valores de la variable independiente, X, difieren pero los dos valores correspondientes de la variable dependiente, Y, no difieren (están empatados), hay evidencia de una falta de asociación y los empates en la variable Y (T_y) deben ser incluidos en el denominador donde actúan para disminuir el valor de d_{yx}. El mismo razonamiento es válido para la medida d_{xy} de Somers.

Por último, es evidente que que la medida τ_b de Kendall de asociación ordinal es simplemente la media geométrica de las medidas d_{yx} y d_{xy} de Somers, dadas por

$$\tau_b = \sqrt{d_{yx} d_{xy}} \ .$$

3.2. Métodos estadísticos de permutación

Para aplicar las técnicas exactas de permutaciones en una tabla de contingencia $r \times c$, es necesario calcular la medida de asociación ordinal seleccionada para las frecuencias de celdas observadas y enumerar exhaustivamente todas las M posibles ordenaciones igualmente probables de los N individuos en las rc celdas, fijadas las distribuciones de frecuencias marginales observadas. Para cada ordenación en el conjunto de referencia de todas las permutaciones, se calcula una medida de asociación ordinal, digamos T , y la probabilidad hipergeométrica exacta puntual bajo la hipótesis de no asociación ordinal, $p(n_{ij}|n_{i.}, n_{.j}, N)$, donde

$$p(n_{ij}|n_{i.}, n_{.j}, N) = \frac{(\prod_{i=1}^{r} n_{i.}!)(\prod_{j=1}^{c} n_{.j}!)}{N! \prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \ .$$

Si T_o denota el valor del estadístico observado, los p-valores a una cola (P) de T_o son las sumas de los valores $p(n_{ij}|n_{i.}, n_{.j}, N)$ asociados a los valores T calculados en las posibles ordenaciones igualmente probables de las frecuencias de las celdas que son iguales o mayores que T_o cuando T_o es positivo, $Pr[T \geq T_o|no \ \ asociación \ \ ordinal]$, e iguales o menores que T_o cuando T_o es negativo, $Pr[T \leq T_o|no \ \ asociación \ \ ordinal]$. Así, la probabilidad exacta hipergeométrica de T_o cuando T es positivo viene dado por

$$P = \sum_{k=1}^{M} \Psi_p(T_k) p(n_{ij}|n_{i.}, n_{.j}, N) \ ,$$

donde

$$\Psi_p(T_k) = \begin{cases} 1 & T_k \geq T_o \\ 0 & c.c. \end{cases}$$

y la probabilidad exacta hipergeométrica de T_o cuando T es negativo viene dado por

$$P = \sum_{k=1}^{M} \Psi_n(T_k) p(n_{ij}|n_{i.}, n_{.j}, N) \ ,$$

donde

$$\Psi_n(T_k) = \begin{cases} 1 & T_k \leq T_o \\ 0 & c.c. \end{cases} \ .$$

Cuando el número de posibles ordenaciones de las frecuencias de las celdas es muy grande, los tests exactos de permutaciones son poco prácticos y son necesarios los métodos de Monte Carlo. Los métodos estadísticos de permutación de Monte Carlo generan una muestra aleatoria de todas las posibles ordenaciones de las frecuencias de las celdas, extraídos con reemplazo, dadas las distribuciones de frecuencias marginales observadas. Los p-valores (de cola superior e inferior) del estadístico T son simplemente las proporciones de los valores T calculados en las ordenaciones de frecuencias de celdas seleccionadas aleatoriamente que son iguales o mayores que T_o cuando T_o es positivo e iguales o menores que T_o cuando T_o es negativo. Así, el p-valor de remuestreo de Monte Carlo de T_o cuando T es positivo viene dado por

$$P(T \geq T_o|No \ \ asociación \ \ ordinal) = \frac{número \ \ de \ \ veces \ \ que \ \ T \geq T_o}{L} \ ,$$

donde L denota el número de ordenaciones aleatorias de los datos observados.

3.3. Otras medidas de Asociación Ordinal por pares

3.3.1. Medidas $d_{y.x}$ y $d_{x.y}$ de Kim

En 1971, Kim propuso medidas asimétricas proporcionales de reducción del error de asociación ordinal dadas por

$$d_{y.x} = \frac{C - D}{C + D + T_x} \qquad y \qquad d_{x.y} = \frac{C - D}{C + D + T_y}$$

[Kim, 1971]. A diferencia de las medidas d_{yx} y d_{xy} de Somers de asociación ordinal, que ajustan los empates en la variable dependiente; las medidas $d_{y.x}$ y $d_{x.y}$ de Kim ajustan los empates en la variable independiente. Es evidente que las medidas $d_{y.x}$ y $d_{x.y}$ de Kim son equivalentes a las medidas d_{xy} y d_{yx} de Somers, respectivamente.

3.3.2. Medida e de Asociación Ordinal de Wilson

En 1974, Wilson propuso otra medida de asociación ordinal que denominó e [Wilson, 1974]. Argumentando que una medida de asociación debería ajustarse para los valores empatados tanto en la variable X como en la variable Y, Wilson sugirió una medida simétrica de asociación ordinal dada por

$$e = \frac{C - D}{C + D + T_x + T_y} = \frac{S}{C + D + T_x + T_y} \; .$$

Como observó Wilson, e toma los valores de ± 1 si y sólo si los datos son estríctamente monótonos. Es obvio, a partir de la ecuación anterior, que la e de Wilson es equivalente a la d_{yx} de Somers cuando $T_x = 0$ y es equivalente a la d_{xy} de Somers cuando $T_y = 0$. Además, si $T_x = 0$ como $T_y = 0$, entonces $e = d_{yx} = d_{xy} = \gamma = \tau_a = \tau_b$.

3.3.3. Medida S de Whitfield para una Variable Binaria y una Variable Ordinal

En 1947 John Whitfield propuso una medida de correlación entre dos variables en la que una variable estaba compuesta por N rangos y la otra variable era dicotómica [Whitfield, 1947]. Un ejemplo servirá para ilustrar el procedimiento de Whitfield:

Considere las puntuaciones de rango que figuran en la siguiente tabla, donde las categorías de la variable dicotómica vienen dadas por "0" y "1" y las puntuaciones de rango van de 1 a 6:

Tabla 3.2: Ejemplo para la medida S de Whitfield

Casos	1	2	3	4	5	6
Binaria	0	1	0	0	0	1

Sea $n_0 = 4$ el número de casos en la categoría "0", sea $n_1 = 2$ el número de casos en la categoría "1" y sea $N = n_0 + n_1$.

Whitfield diseñó un procedimiento para calcular un estadístico que denominó S: teniendo en cuenta las $N = 6$ puntuaciones de la tabla, se consideran los $n_0 = 4$ casos en la categoría identificada por "0": 1, 3, 4 y 5. Empezando por el caso 1 con la categoría "0", no hay puntuaciones de rango con la categoría "1" a la izquierda de "0" y hay dos casos con la categoría "1" a la derecha de "0", por lo que $0 - 2 = -2$. Para el caso 3 con la categoría "0", hay una puntuación de rango a la izquierda de "0" con la categoría "1" (rango 2) y una puntuación de rango a la derecha de "0" con la categoría "1"; por tanto, $1 - 1 = 0$. Igualmente, para el caso 4: $1-1 = 0$ y para el caso 5: $1-1 = 0$. La suma de las diferencias entre las variables "0" y "1" es $S = -2 + 0 + 0 + 0 = -2$. De este modo, el enfoque de Whitfield se adapta a muestras con $n_0 = n_1$, así como a cualquier número de puntuaciones de rango empatadas.

Como el número de pares posibles de N enteros consecutivos viene dado por

$$\frac{N(N-1)}{2}$$

Whitfield definió y calculó la medida de asociación de rangos entre las variables "0" y "1" como

$$\tau = \frac{2S}{N(N-1)} \ .$$

La S de Whitfield es idéntica a la S de Kendall y está directamente relacionada con el estadístico U de suma de rangos de dos muestras de Mann y Whitney y con el estadístico W de suma de rangos de dos muestras de Wilcoxon. Las relaciones entre los estadísticos S de Whitfield y U de Mann y Whitney [Mann and Whitney, 1947] vienen dadas por

$$S = 2U - n_0 n_1 \qquad y \qquad U = \frac{S + n_0 n_1}{2}$$

y las relaciones entre la S de Whitfield y la W de Wilcoxon [Wilcoxon, 1945] vienen dadas por

$$S = n_1(N+1) - 2W \qquad y \qquad W = \frac{n_1(N+1) - S}{2} \ .$$

3.3.4. Coeficiente de correlación rango-biserial de Cureton

Sean dos variables, una representada por rangos y la otra por una dicotomía. En 1956, el psicólogo Edward Cureton propuso una nueva medida de correlación para una variable de rangos y una variable dicotómica denominada r_{rb} para la correlación rango-biserial [Cureton, 1956].

Cureton afirmó que el coeficiente de correlación debería tipificarse adecuadamente entre ± 1 y debería ser estrictamente no paramétrico, definido únicamente en términos de inversiones y concordancias entre pares de rangos, sin el uso de medias, varianzas, covarianzas o regresión. El coeficiente de correlación rango-biserial se define como:

$$r_{rb} = \frac{S}{S_{max}} \ ,$$

donde $S = C - D$ es el estadístico del test de Kendall [Kendall, 1938] y Whitfield [Whitfield, 1947], con C el número de pares concordantes y D el número de pares discordantes; y $S_{max} = n_0 n_1$, con n_0 el número de puntuaciones de rango en la categoría "0" y n_1 el número de puntuaciones de rango en la categoría "1".

En 1966, Glass [Glass, 1966] dedujo una fórmula simplificada para la r_{rb}, suponiendo que no hay puntuaciones de rango empatadas, dada por

$$r_{rb} = \frac{2}{N}(\bar{y}_1 - \bar{y}_0) \ ,$$

donde $\bar{y}_0$ y $\bar{y}_1$ son las medias aritméticas de los valores de la variable dicotómica codificados como "0" y "1", respectivamente.

Glass proporcionó dos fórmulas de cálculo alternativas dadas por

$$r_{rb} = \frac{2}{n_0}(\bar{y}_1 - \frac{N+1}{2}) \qquad \text{ó} \qquad r_{rb} = \frac{2}{n_1}(\frac{N+1}{2} - \bar{y}_0) \ .$$

Capítulo 4

Medidas para Variables Ordinales (Parte 2)

Este capítulo continúa la descripción de las medidas de asociación para dos variables de nivel ordinal iniciada en el capítulo 3, pero se centra las medidas que se basan en criterios distintos a las comparaciones por pares entre las puntuaciones de los rangos, aunque es inevitable cierto solapamiento.

Se incluyen métodos estadísticos de permutación exactos y de Monte Carlo para el coeficiente de correlación de Spearman, la medida de concordancia de la regla del pie de Spearman, el coeficiente de concordancia de Kendall y Babington Smith, la medida kappa ponderada de Cohen para la concordancia, y el análisis ridit de Bross.

4.1. Coeficiente de correlación de rango de Spearman

Se consideran dos clasificaciones de N objetos consistentes en los N primeros enteros y sean X_i e Y_i para $i = 1, ..., N$ la primera y la segunda clasificación, respectivamente. La correlación de rangos no está exenta de críticas, ya que la derivación de la fórmula para esta correlación implica la suposición de que las diferencias entre los rangos sucesivos son iguales.

Una medida popular de la correlación entre las dos clasificaciones es el coeficiente de correlación de orden de rango de Spearman, dado por

$$\rho = 1 - \frac{\sum_{i=1}^{N} d_i^2}{\frac{N(N^2-1)}{6}} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \,, \tag{4.1}$$

donde $d_i = X_i - Y_i$ para $i = 1, ..., N$. Charles Spearman desarrolló ρ en el primero de los dos artículos sobre la medición de la asociación y la correlación en 1904 y 1906 que aparecieron en el *American Journal of Psychology* [Spearman, 1904] y en el *British Journal of Psychology* [Spearman, 1906], respectivamente. Con dos conjuntos de puntuaciones de rango sin valores empatados, X_i e Y_i para $i = 1, ..., N$, se tiene que

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i = \frac{N(N+1)}{2}$$

y

$$\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i^2 = \frac{N(N+1)(2N+1)}{6} \ ,$$

Spearman simplemente sustituyó en la fórmula de Pearson del coeficiente de correlación producto-momento dado por

$$r_{xy} = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2}\sqrt{N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}}$$

y simplificó la ecuación, dando lugar a la Ecuación (4.1).

Obsérvese que el denominador del coeficiente de correlación de orden de rango de Spearman, $N(N^2-1)/6$, como se indica en la Ec. (4.1), representa la mitad del valor máximo de $\sum_{i=1}^{N} d_i^2$ cuando X_i e Y_i , $i = 1, ..., N$, son puntuaciones de rango sin valores empatados y las puntuaciones de rango de Y_i son la inversa exacta de las puntuaciones de rango de X_i, es decir, $Y_i = N - X_i + 1$ para $i = 1, ..., N$. Por lo tanto, la ρ de Spearman es una medida de correlación de rango corregida al máximo y toma valores entre ± 1, donde $+1$ indica una asociación positiva perfecta y -1 indica una asociación negativa perfecta.

Se demuestra fácilmente que el denominador de la ecuación (4.1), $N(N^2-1)/6$, es la mitad del valor máximo de $\sum_{i=1}^{N} d_i^2$ cuando X_i e Y_i para $i = 1, ..., N$ son ambas son puntuaciones de rango no empatadas y las puntuaciones de rango Y_i son la inversa de las puntuaciones de rango x_i. Para el valor máximo de $\sum_{i=1}^{N} d_i^2$, se define:

$$\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (X_i - Y_i)^2 = \sum_{i=1}^{N} X_i^2 + \sum_{i=1}^{N} Y_i^2 - 2\sum_{i=1}^{N} X_i Y_i \ .$$

Ya que, para N valores de rango no empatados,

$$\sum_{i=1}^{N} X_i^2 = \sum_{i=1}^{N} Y_i^2 = \frac{N(N+1)(2N+1)}{6}$$

y, para $X_i = 1, ..., N$ e $Y_i = N - x_i + 1$, $i = 1, ..., N$,

$$\sum_{i=1}^{N} X_i Y_i = \frac{N(N+1)(N+2)}{6} \ ,$$

entonces, al sustituir en la Ec. (4.1) se obtiene:

$$\sum_{i=1}^{N} d_i^2 = \frac{2N(N+1)(2N+1)}{6} - \frac{2N(N+1)(N+2)}{6} =$$

$$= \frac{2N(N+1)(N-1)}{6} = \frac{N(N^2-1)}{3} \ ,$$

que es el doble del valor de $N(N^2-1)/6$.

Kendall y Babington Smith observaron que para interpretar la significación de un valor de ρ, es necesario considerar sólo la distribución de valores obtenida a partir de las clasificaciones observadas con todas las demás permutaciones posibles de los números enteros de 1 a N, y señalaron además que en la práctica suele ser más conveniente considerar sólo la distribución de $\sum_{i=1}^{N} d_i^2$ ya que $N(N^2-1)/6$ es invariable bajo permutación [Kendall et al., 1939].

4.2. Medida de concordancia de la regla del pie de Spearman

Los artículos de 1904 [Spearman, 1904] y 1906 [Spearman, 1906] de Charles Spearman contenían dos nuevas medidas de correlación de rango: el conocido coeficiente de correlación de rango de Spearman, ρ, y un segundo coeficiente de correlación menos conocido que Spearman denominó "la regla del pie". Se consideran dos clasificaciones de N objetos que consisten en los primeros N enteros y sean X_i e Y_i para $i = 1, ..., N$ la primera y la segunda clasificación, respectivamente. Entonces, la regla de Spearman viene dada por:

$$\mathcal{R} = 1 - \frac{\sum_{i=1}^{N} |X_i - Y_i|}{\frac{N^2-1}{3}} = \frac{3\sum_{i=1}^{N} |X_i - Y_i|}{N^2 - 1} \ . \tag{4.2}$$

A diferencia del coeficiente de correlación de orden de rango de Spearman, el denominador del coeficiente de la regla del pie de Spearman, $\sim(N^2-1)/3$, como se da en la Ec. (4.2), no representa la mitad del valor máximo de $\sum_{i=1}^{N} |X_i - Y_i|$ cuando X_i e Y_i para $i = 1, ..., N$ son ambas puntuaciones de rango no empatadad y las puntuaciones de rango Y_i son la inversa exacta de las puntuaciones de rango X_i, es decir, $Y_i = N - X_i + 1$ para $i = 1, ..., N$. Por lo tanto, la $\mathcal{R}$ de Spearman no es una medida máximo-corregida de correlación de rangos y es, en cambio, una medida de concordancia corregida por el azar.

Se puede demostrar fácilmente que la $\mathcal{R}$ de Spearman es una medida de concordancia corregida por el azar y no es, de hecho, una medida convencional de correlación, lo que explica por qué $\mathcal{R}$ puede, en ocasiones, dar valores negativos y sólo puede alcanzar un valor de -1 cuando $N = 2$. Para demostrar que el valor esperado de $\sum_{i=1}^{N} |d_i|$ viene dado por $(N^2-1)/3$, se considera que:

$$\sum_{i=1}^{N} |d_i| = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} |i-j| = \frac{2}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (j-i) =$$

$$= \frac{1}{N} \sum_{i=1}^{N-1} [N(N+1) + i^2 - i(2N+1)] =$$

$$= \frac{N(N+1)}{6N} [6(N+1) + (2N-1) - 3(2N+1)] =$$

$$= \frac{N^2 - 1}{3} \ .$$

Por lo tanto, el coeficiente de la regla del pie de Spearman, dado por

$$\mathcal{R} = a - \frac{\sum_{i=1}^{N} |d_i|}{\frac{N^2-1}{3}}$$

es una medida de concordancia corregida por el azar cuando el valor esperado de $\sum_{i=1}^{N} |d_i|$ viene dado por $(N^2 - 1)/3$, ya que adopta la forma clásica de las medidas de concordancia corregidas por el azar dada por

$$concordancia = \frac{discordancia \ \ observada}{discordancia \ \ esperada}$$

[Krippendorff, 1970].

Existen tres limitaciones de la regla del pie de Spearman que contribuyen a su falta de uso en la investigación contemporánea [Stuart, 1977]: En primer lugar, a diferencia de otras medidas de correlación de rangos, $\mathcal{R}$ no norma adecuadamente entre los límites de ± 1; en segundo lugar, al igual que la ρ de Spearman, $\mathcal{R}$ se limita a los datos totalmente clasificados y no se adapta a las puntuaciones de rango empatadas; y en tercer lugar, debido a la suma de las diferencias absolutas entre las puntuaciones de rango, tradicionalmente ha sido algo engorroso establecer el valor de probabilidad de un valor observado de $\mathcal{R}$, especialmente cuando N es pequeño.

La $\mathcal{R}$ de Spearman alcanza un valor máximo de $+1$ cuando X_i es idéntica a Y_i para $i = 1, ..., N$ y no hay valores empatados. Sin embargo, si $Y_i = N - x_i + 1$ para $i = 1, ..., N$, entonces $\mathcal{R} = -0,5$ cuando N es impar y

$$\mathcal{R} = -0.5(1 + \frac{3}{N^2 - 1})$$

cuando N es par [Kendall, 1962]. En consecuencia, $\mathcal{R}$ no puede alcanzar un valor mínimo de -1, excepto cuando $N = 2$. Spearman, aparentemente sin saber que R era una medida corregida por el azar y reconociendo que los valores negativos de $\mathcal{R}$ no representaban una correlación inversa, sugirió ingenuamente que “es mejor tratar toda correlación como positiva” [Spearman, 1904]. Maurice Kendall señaló explícitamente esta aparente falta tipificación como un defecto de la regla del pie y sugirió una corrección dada por

$$\mathcal{R}' = 1 - \frac{\sum_{i=1}^{N} |X_i - Y_i|}{N^2} \ ,$$

que aseguraba un límite adecuado de $+1$ cuando las dos clasificaciones estaban en completa concordancia y -1 cuando las dos clasificaciones eran inversas entre sí [Kendall, 1962]. Sin embargo la corrección, aunque bien intencionada, destruyó por completo la interpretación de la regla del pie de Spearman.

4.2.1. Probabilidad de la regla del pie de Spearman

Cuando las variables X e Y consisten en su totalidad en puntuaciones de rango no empatadas de 1 a N y la variable Y es una permutación de las observaciones de rango en la variable X, existen métodos para determinar la probabilidad de una $\mathcal{R}$ observada bajo la hipótesis nula de que cualquiera de los $N!$ ordenamientos de los valores de X o Y es igualmente probable. Si

$$D = \sum_{i=1}^{N} |X_i - Y_i|$$

entonces, dado que $\mathcal{R}$ es simplemente una transformación lineal de D, la probabilidad de un valor observado de D es la probabilidad de un valor observado de $\mathcal{R}$. Las tablas de la función de distribución acumulativa exacta de D para $2 \leq N \leq 10$ y los valores de probabilidad aproximados basados en métodos deMonte Carlo para $11 \leq N \leq 15$ fueron publicados por Ury y Kleinecke en 1979 [Ury and Kleinecke, 1979]. En 1988 Franklin amplió el trabajo de Ury y Kleinecke, informando de la función de distribución acumulativa exacta de D para $11 \leq N \leq 18$, y discutió la tasa de convergencia a una distribución normal aproximada [Franklin, 1988]. En 1990 Salama y Quade utilizaron las propiedades de la cadena de Markov para obtener la exacta función de distribución acumulativa de D para $4 \leq N \leq 40$ y siguieron investigando aproximaciones a la distribución discreta de D [Salama and Quade, 1990]. Si la variable X o la variable Y contiene valores empatados, entonces el cálculo de un valor de probabilidad exacto es más complejo.

4.2.2. Rangos Múltiples

La regla del pie de Spearman, tal como se presentó originalmente en sus artículos de 1904 y 1906 en el *American Journal of Psychology* [Spearman, 1904] y el *British Journal of Psychology* [Spearman, 1906], respectivamente, se limitaba a $N \geq 2$ puntuaciones de rango no empatadas y $b = 2$ jueces. Sin embargo, como Berry y Mielke demostraron en 1998, la regla de Spearman puede generalizarse para incluir tanto puntuaciones de rango empatadas como no empatadas y $b \geq 2$ conjuntos de clasificaciones [Berry and Mielke, 1998]. Sea

$$\delta = [N \binom{b}{2}]^{-1} \sum_{i=1}^{N} \sum_{r<s} |X_{ri} - X_{si}|$$

denota una función de distancia media basada en todas las $\binom{b}{2}$ posibles diferencias absolutas emparejadas entre los valores de las clasificaciones por b jueces y sea

$$\mu_\delta = [N^2 \binom{b}{2}]^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{r<s} |X_{ri} - X_{sj}|$$

denota el valor esperado de δ donde b es el número de jueces, N es el número de objetos, y $\sum_{r<s}$ es la suma sobre todos los r y s tales que $1 \leq r < s \leq N$. Entonces, la generalización de la medida de la regla del pie de Spearman viene dada por

$$\mathfrak{R} = 1 - \frac{\delta}{\mu_\delta} \ ,$$

donde $\mathfrak{R}$ es una medida de concordancia corregida por el azar entre los b jueces que no se limita a las puntuaciones de rango no empatadas. Nótese que en el caso de $b = 2$ jueces, la ecuación anterior se reduce a la regla del pie de Spearman de 1906 para $b = 2$ jueces.

4.3. Coeficiente de Concordancia

Mientras que el coeficiente de correlación de rango de Spearman y las medidas τ_a y τ_b de Kendall expresan el grado de asociación entre dos variables medidas en puntuaciones de rango o transformadas en ellas, el coeficiente de concordancia expresa el grado de asociación entre múltiples conjuntos de puntuaciones de rango.

En 1939, Maurice Kendall y Bernard Babington Smith publicaron un artículo en *The Annals of Mathematical Statistics* sobre "The problem of m rankings" en el que desarrollaron el conocido coeficiente de concordancia [Kendall and Smith, 1939]. Sean N y m el número de puntuaciones de rango y el número de jueces, respectivamente, entonces Kendall y Babington Smith definieron el coeficiente de concordancia como

$$W = \frac{12S}{m^2(N^3 - N)} \ ,$$

donde S es la suma observada de los cuadrados de las desviaciones de las sumas de los rangos respecto al valor medio $m(N + 1)/2$.

Dado que $m^2(N^3 - N)$ en el denominador de la ecuación anterior es invariable en todas las permutaciones de los datos observados, Kendall y Babington Smith demostraron que para para comprobar si un valor observado de W es estadísticamente significativo sólo es necesario considerar la distribución de S permutando los N rangos de todas las formas posibles e igualmente posibles. Si uno de los rangos es fijo, hay $(N!)^{m-1}$ valores posibles de S. Basándose en este procedimiento de permutación, Kendall y Babington Smith crearon cuatro tablas que proporcionan valores de probabilidad exactos para $N = 3$ y $m = 2, ..., 10$, $N = 4$ y $m = 2, ..., 6$, y $N = 5$ y $m = 3$.

W también puede definirse como

$$W = \frac{12 \sum_{i=1}^{N} R_i^2 - 3m^2 N(N + 1)}{m^2 N(N^2 - 1)} \ ,$$

donde R_i para $i = 1, ..., N$ es la suma de las puntuaciones de rango para el i-ésimo de los N objetos y no hay puntuaciones de rango empatadas. También se sabe que W puede definirse como una función del valor medio de todos los coeficientes de correlación de orden de rango de Spearman por pares, dado por

$$\bar{\rho} = \frac{2}{m(m - 1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij} \ .$$

A este respecto, Kendall y Babington Smith demostraron que $\bar{\rho}$ es simplemente el coeficiente intraclase, r_I, para los m conjuntos de clasificaciones. Las relaciones entre W y $\bar{\rho}$ vienen dadas por

$$\bar{\rho} = \frac{mW - 1}{m - 1} \qquad y \qquad W = \frac{\bar{\rho}(m-1) + 1}{m} \ .$$

Si todos los ordenamientos de los conjuntos de puntuaciones de rango observadas ocurren con igual probabilidad, el valor exacto de la probabilidad del valor observado de W calculado sobre M posibles ordenamientos igualmente probables de las puntuaciones de rango observadas bajo la hipótesis nula es

$$P(W \geq W_o | H_0) = \frac{número \ \ de \ \ valores \ \ W \geq W_o}{M} \ ,$$

donde W_o indica el valor observado de W.

4.3.1. Procedimientos relacionados

Desde hace tiempo se conoce que la estructura de datos para el coeficiente de concordancia de Kendall y Babington Smith [Kendall and Smith, 1939] es la misma que la del análisis de varianza de dos vías de Friedman para rangos [Friedman, 1937] y la misma que la de la relación de correlación de Wallis para datos de puntuación de rangos [Wallis, 1939]. Mientras que la prueba de Friedman, por ejemplo, proporciona un valor de probabilidad global de las diferencias globales entre rangos, existe un procedimiento relacionado que proporciona un valor de probabilidad exacto para la suma de rangos de un solo objeto, respondiendo a la pregunta: ¿cuándo el total de un solo objeto no se debe al azar bajo la hipótesis nula de asignación aleatoria?

Supongamos que cada uno de los N jueces asigna de forma independiente K rangos distintos no empatados a $K \geq 2$ objetos. Si S denota la suma de los N rangos para un objeto determinado bajo la hipótesis nula de que cada uno de los N jueces asigna los K rangos a los K objetos al azar, es decir, que cada objeto ocurre con probabilidad $1/K$, entonces la probabilidad puntual exacta de S viene dada por

$$p_S = K^{-N} C_{S-N}$$

para $S = N, N+1, ..., NK$. Sea $m = S - N$, entonces

$$C_m = \sum_{j=0}^{v} (-1)^j \left(\begin{array}{c} N \\ j \end{array} \right) \left(\begin{array}{c} m - jK + N - 1 \\ N - 1 \end{array} \right)$$

y $v = min(N, m/K)$, es decir, el mayor número entero no negativo menor o igual que N o m/K. El valor exacto de la probabilidad unilateral de S viene dado por

$$P_1 = \sum_{j=N}^{w} p_j \ ,$$

donde $w = min(S, NK + N - S)$ y el valor exacto de la probabilidad bilateral de S es dado por

$$P_2 = min(2P_1, 1) \ ,$$

ya que la distribución de S es simétrica respecto a $N(K+1)/2$ bajo la hipótesis nula.

4.4. Medida de acuerdo u de Kendall

A veces, en lugar de pedir a un grupo de jueces que clasifiquen un conjunto de objetos, se les presenta una serie de pares de objetos y se les pide que indiquen su preferencia por uno de los dos objetos de cada par. Este procedimiento se denomina comparación por pares. Cuando se recogen datos mediante el método de las comparaciones por pares, es posible calcular el grado de concordancia entre los jueces. En 1940, Kendall y Babington Smith [Kendall and Smith, 1940] propusieron un coeficiente de concordancia para evaluar las comparaciones pareadas entre k jueces para N clasificaciones, dado por

$$u = \frac{2S}{\binom{k}{2}\binom{N}{2}} - 1$$

donde

$$S = \sum_{i=1}^{N}\sum_{j=1}^{N}\binom{a_{ij}}{2} = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}^2 - k\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} + \binom{k}{2}\binom{N}{2} \ ,$$

a_{ij} es el número de veces que un objeto asociado a la fila i de una matriz de preferencias es preferido al objeto asociado a la fila j, y $a_{ij} \geq 2$ para $i, j = 1, ..., N$. El número máximo de concordancias, que se produce cuando $\binom{N}{2}$ celdas de la matriz de preferencias contienen k cada una, es $\binom{N}{2}\binom{k}{2}$ y, por tanto, sólo en el caso de concordancia completa, $u = +1$ [Kendall and Smith, 1940].

Mientras que el valor máximo de u es $+1$ cuando hay concordancia completa entre los k jueces, el número mínimo de concordancias se produce cuando cada celda de la matriz de preferencias contiene $k/2$ si k es par o $(k \pm 1)/2$ si k es impar. Así, cuando k es par, el valor mínimo de u es $-1/(k-1)$ y cuando k es impar, el valor mínimo de u es $-1/k$. Como el valor esperado de u es cero [@44] y los valores mínimos de u son $-1/(k-1)$ cuando k es par y $-1/k$ cuando k es impar, u es claramente una medida de concordancia corregida por el azar, aunque aparentemente esto no fue reconocido por Kendall y Babington Smith cuando desarrollaron u en 1940.

4.5. Medida Kappa de Cohen

En 1960, Jacob Cohen desarrolló el estadístico kappa, una medida corregida por el azar de la concordancia entre dos jueces para un conjunto de c categorías disjuntas y no ordenadas [Cohen, 1960]. En 1968, Cohen amplió el kappa para medir la concordancia entre dos jueces para un conjunto de c categorías ordenadas y disjuntas [Cohen, 1968]. El kappa original para c categorías disjuntas y no ordenadas se conoció como kappa "no ponderado", o κ, y el kappa para c categorías disjuntas y ordenadas se conoció como kappa "ponderado", o κ_w. Mientras que la kappa no ponderada no distinguía entre magnitudes de desacuerdo, la kappa ponderada incorporaba la magnitud de cada desacuerdo y otorgaba un crédito parcial a las discrepancias cuando la concordancia no era completa [Maclure and Willett, 1987]. El enfoque habitual consiste en asignar pesos a cada par de desacuerdos (pesos mayores indican un mayor desacuerdo). El kappa no ponderado para c categorías disjuntas y no ordenadas se trata en el capítulo 2. A continuación se presenta el kappa ponderado para c categorías disjuntas y ordenadas:

La medición de la concordancia es un caso especial de medición de la asociación entre dos variables de nivel ordinal. Los índices de concordancia miden el grado en que un conjunto de medidas de respuesta son idénticas a otro conjunto, es decir, concuerdan, en lugar del grado en que un conjunto de medidas de respuesta es una función lineal de otro conjunto de medidas de respuesta, es decir, correlacionan. Al igual que la regla de Spearman, la medida kappa ponderada de Cohen de la concordancia es una medida corregida por el azar, que refleja la cantidad de concordancia por encima de lo que cabría esperar por azar. Así, la kappa ponderada es igual a uno cuando se produce una concordancia perfecta, es igual a cero en caso de independencia y puede ser ligeramente negativo cuando la concordancia es inferior a la esperada por el azar [Fleiss et al., 2003].

Para simplificar, consideremos $N \geq 2$ objetos clasificados de forma cruzada por $b = 2$ jueces independientes en una tabla de contingencia $c \times c$ con c categorías disjuntas y ordenadas denotadas por $a_1, ..., a_c$, como en la siguiente tabla:

Tabla 4.1: Notación para una tabla de validación cruzada de N objetos por 2 jueces en c categorías disjuntas y ordenadas

Juez 1 \ Juez 2	a_1	a_2	$\dots$	a_c	Total
a_1	n_{11}	n_{12}	$\dots$	n_{1c}	$n_{1.}$
a_2	n_{21}	n_{22}	$\dots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_c	n_{c1}	n_{c2}	$\dots$	n_{cc}	$n_{c.}$
Total	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.c}$	N

Donde $n_{i.}$ es el total de la frecuencia marginal de la i-ésima fila, $i = 1, ..., c$, $n_{.j}$ es la frecuencia marginal total de la j-ésima columna, $j = 1, ..., c$, n_{ij} son las frecuencias de las celdas y w_{ij} denotan los pesos de las celdas. Cuando las c categorías para los $b = 2$ jueces están ordenadas de forma similar, entonces n_{ii} , $i = 1, \dots, c$, y n_{ij} , $i \neq j$, denotan las frecuencias de las celdas de concordancia y discordancia, respectivamente.

Aunque se han propuesto diversos esquemas de ponderación para la kappa ponderada de Cohen, el más popular es la ponderación cuadrática dada por $w_{ij} = (i-j)^2$ para $i, j = 1, ..., c$, donde las ponderaciones de discordancia de la categoría progresan geométricamente hacia fuera desde la diagonal de concordancia, es decir, 0^2, 1^2, 2^2, 3^2, etc. Sin embargo, la ponderación lineal en la que $w_{ij} = |i-j|$ para $i, j = 1, ..., c$, donde las ponderaciones de la categoría de desacuerdo progresan linealmente hacia fuera desde la diagonal de concordancia, es decir, 0, 1, 2, 3, y así sucesivamente.

Una fórmula sencilla para el cálculo del estadístico de la prueba kappa ponderada de Cohen con $b = 2$ jueces viene dada por

$$\kappa_w = 1 - \frac{\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}n_{ij}}{\frac{1}{N^2}\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}n_{i.}n_{.j}} \ .$$

Dada una tabla de concordancia $c \times c$ con N objetos clasificados por las valoraciones de dos jueces independientes en c categorías ordenadas y disjuntas, una prueba de permutación exacta genera un conjunto de referencia de todos los M posibles ordenamientos igualmente probables de los N objetos en las c^2 celdas, preservando el número total de objetos en cada categoría, es decir, las distribuciones de frecuencia marginal. Para cada ordenamiento de frecuencias de celdas con distribuciones de frecuencias marginales fijas, se calculan el estadístico kappa ponderado, κ_w, y el valor de probabilidad exacto, $p(n_{ij}|n_{i.}, n_{.j}, N)$, donde

$$p(n_{ij}|n_{i.}, n_{.j}, N) = \frac{(\prod_{i=1}^{c} n_{i.}!)(\prod_{j=1}^{c} n_{.j}!)}{N! \prod_{j=1}^{c}\prod_{j=1}^{c} n_{ij}!}$$

es el valor de probabilidad hipergeométrico convencional de una tabla de contingencia $c \times c$.

Sea κ_o el valor del estadístico kappa ponderado observado y M el número total de ordenamientos distintos de la frecuencia de las celdas de los N objetos en la tabla de clasificación $c \times c$, dados los totales de frecuencia marginales fijos. Entonces el valor exacto de la probabilidad de κ_o bajo la hipótesis nula viene dado por

$$P(\kappa_o|H_0) = \sum_{k=1}^{M} \Psi(\kappa_k) p(n_{ij}|n_{i.}, n_{.j}, N) \ ,$$

donde

$$\Psi(\kappa_k) = \begin{cases} 1 & \kappa_k \geq \kappa_o \\ 0 & c.c. \end{cases} \ .$$

Cuando el conjunto de referencia de todos los M ordenamientos posibles es muy grande, los análisis exactos de permutación son poco prácticos y se hacen necesarias las aproximaciones de remuestreo de Monte Carlo. Sea L una muestra aleatoria de todos los M valores posibles de κ_w. Entonces, bajo la hipótesis nula, el valor de la probabilidad aproximada de remuestreo para el valor observado de κ_w, κ_o, viene dado por

$$P(\kappa_o) = \frac{1}{L}\sum_{l=1}^{L} \Psi_l(\kappa_w) \ .$$

4.5.1. Comparación de la ponderación lineal y cuadrática

Existe una considerable controversia sobre qué pesos deben utilizarse con el estadístico kappa ponderado de Cohen, κ_w. La elección de los pesos es completamente arbitraria y se puede utilizar cualquier peso de celda de desacuerdo. La ponderación lineal es quizás la más útil, porque como las ponderaciones de las celdas de desacuerdo progresan linealmente hacia fuera desde la diagonal de concordancia, su interpretación es más facil.

Es evidente que la ponderación lineal y la ponderación cuadrática producen los mismos resultados para las tablas de contingencia 2×2. También está muy claro que los valores kappa ponderados linealmente y los ponderados cuadráticamente suelen diferir muy poco para las tablas de contingencia 3×3. La ponderación lineal y cuadrática generalmente produce mayores diferencias con tablas de contingencia más grandes. Brenner y Kliebsch demostraron que la forma lineal del coeficiente kappa ponderado es menos sensible al número de categorías que la forma cuadrática; en consecuencia, recomendaron usar la forma lineal siempre que el número de categorías de la escala ordinal sea grande [Brenner and Kliebsch, 1996].

4.5.2. Kappa ponderado con múltiples jueces

Aunque la kappa ponderada de Cohen fue diseñada originalmente para $b = 2$ jueces independientes y se limita a ellos, la kappa ponderada puede generalizarse y ampliarse para medir la concordancia entre múltiples jueces. La generalización de la kappa de Cohen a múltiples jueces ha sido controvertida durante mucho tiempo, con muchos pasos en falso y callejones sin salida en el camino. En 1988 Berry y Mielke generalizaron la medida de concordancia kappa de Cohen para dar cabida a múltiples jueces [Berry and Mielke, 1988] y en 2008 Mielke, Berry y Johnston proporcionaron un eficiente algoritmo de remuestreo de Monte Carlo para analizar los datos de concordancia con múltiples jueces [Mielke et al., 2008].

En esta sección se presenta un procedimiento algorítmico para calcular el kappa ponderado y no ponderado con múltiples calificadores. Aunque el procedimiento es apropiado para cualquier número de $c \geq 2$ categorías disjuntas y ordenadas y $b \geq 2$ jueces, la descripción del procedimiento se limitan a $b = 3$ jueces independientes para simplificar la presentación, sin pérdida de generalidad.

Consideremos $b = 3$ jueces que clasifican independientemente N objetos en c categorías ordenadas y disjuntas. La clasificación puede conceptualizarse como una tabla de contingencia $c \times c \times c$ con c filas, c columnas y c cortes. Sean n_{ijk}, R_i, C_j y S_k las frecuencias de las celdas y los totales de las frecuencias marginales de las filas, columnas y cortes para $i, j, k = 1, ..., c$ y sea el total de frecuencias dado por

$$N = \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} n_{ijk} \ .$$

La estadística de la prueba kappa ponderada de Cohen para una tabla de contingencia de tres vías viene dada por

$$\kappa_w = \frac{N^2 \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} n_{ijk}}{\sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} R_i C_j S_k} \ ,$$

donde w_{ijk} son los pesos de desacuerdo asignados a cada celda para $i, j, k = 1, ..., c$. Bajo la hipótesis nula de que los jueces clasifican los N objetos de forma independiente con totales de frecuencia marginal fijos, $E[\kappa_w] = 0$.

Como se ha comentado anteriormente, se han propuesto diversas funciones de ponderación para el kappa ponderado para dos jueces, donde las ponderaciones de celdas arbitrarias se denotan como w_{ij} e i y j designan las c categorías para cada juez, $i, j = 1, ..., c$. Normalmente, los pesos de las celdas se definen de forma que $w_{ii} = 0$ para $i = 1, ..., c$ y los pesos son simétricos, es decir, $w_{ij} = w_{ji}$ para $i, j = 1, ..., c$. Algunos ejemplos de sistemas de ponderación para dos jueces son la ponderación lineal, en la que $w_{ij} = |i - j|$, ponderación cuadrática donde $w_{ij} = (i - j)^2$, y kappa no ponderado donde

$$w_{ij} = \begin{cases} 0 & i = j \\ 1 & c.c. \end{cases} .$$

Para tres jueces, las ponderaciones de desacuerdo de las celdas vienen dadas por w_{ijk}, donde i, j, y k designan las c categorías de cada juez. De forma análoga a w_{ij}, w_{ijk} puede ser definida de forma que $w_{iii} = 0$ para $i = 1, ..., c$ y las ponderaciones son simétricas, es decir, $w_{ijk} = w_{ikj} = w_{jik} = w_{jki} = w_{kij} = w_{kji}$ para i, j, $k = 1, ..., c$. Ejemplos de sistemas de ponderación para tres jueces incluyen la ponderación lineal donde

$$w_{ijk} = |i - j| + |i - k| + |j - k|$$

y la ponderación cuadrática donde

$$w_{ijk} = (i - j)^2 + (i - k)^2 + (j - k)^2$$

para i, j, $k = 1, ..., c$.

El kappa ponderado para tres jueces se reduce al kappa no ponderado cuando

$$w_{ijk} = \begin{cases} 0 & i = j = k \\ 1 & c.c. \end{cases} .$$

Dada una tabla de contingencia $c \times c \times c$ con N objetos clasificados de forma cruzada por tres jueces independientes, una prueba de permutación exacta implica generar todos los ordenamientos posibles, igualmente probables, de los N objetos en las c^3 celdas, preservando las distribuciones de frecuencia marginal observadas. Para cada ordenamiento en el conjunto de referencia de todas las permutaciones de las frecuencias de las celdas, se calcula el estadístico kappa ponderado, κ_w, y el valor de la probabilidad puntual hipergeométrica exacta bajo la hipótesis nula, $p(n_{ijk}|R_i, C_j, S_k, N)$, donde

$$p(n_{ij}|n_{i.}, n_{.j}, N) = \frac{(\prod_{i=1}^{c} R_i!)(\prod_{j=1}^{c} C_j!)(\prod_{j=1}^{c} S_k!)}{(N!)^2 \prod_{j=1}^{c} \prod_{j=1}^{c} \prod_{k=1}^{c} n_{ijk}!}$$

[Mielke and Berry, 1988b].

Si κ_o denota el valor del estadístico kappa ponderado observado, el valor de probabilidad exacto de κ_o bajo la hipótesis nula viene dado por

$$P(\kappa_o|H_0) = \sum_{l=1}^{M} \Psi_l(n_{ijk}|R_i, C_j, S_k, N) \ ,$$

donde

$$\Psi_l(n_{ijk}|R_i, C_j, S_k, N) = \begin{cases} p(n_{ijk}|R_i, C_j, S_k, N) & \kappa_w \geq \kappa_o \\ 0 & c.c. \end{cases}$$

y M denota el número total de posibles ordenamientos de frecuencia de celdas igualmente probables dados los totales de frecuencia marginal observados. Cuando el conjunto de referencia de M ordenamientos posibles es muy grande, como es típico en las tablas de contingencia multidireccionales, las pruebas exactas son poco prácticas y se hacen necesarios los procedimientos de remuestreo de Monte Carlo. En el remuestreo, una muestra aleatoria de tamaño L extraída de los M posibles ordenamientos de frecuencias de celdas permite comparar los estadísticos de prueba κ_w calculados en las L tablas aleatorias con el estadístico de prueba κ_w calculado en la tabla observada.

4.6. Análisis Ridit

En 1958, I.D.J. Bross introdujo la puntuación ridit para el análisis de datos categóricos ordenados, donde "ridit" es un acrónimo de *Relative to an I dentified Distribution* (relativo a una distribución identificada) y la "it" representa un tipo de transformación similar a logit y probit [Bross, 1958]. Son comunes dos aplicaciones del análisis ridit:

- La primera compara los grupos de tratamiento y de control, donde el grupo de control observado sirve como grupo de referencia y los ridits se calculan para las c categorías disjuntas y ordenadas del grupo de control y se aplican a las c categorías disjuntas y ordenadas del grupo de tratamiento.

  En la primera aplicación, el grupo de control y los ridits correspondientes se tratan como una población infinita y parámetros de población, respectivamente.

- La segunda aplicación compara dos grupos de tratamiento independientes en los que ninguno de los grupos de tratamiento se considera un grupo de referencia y se calculan los ridits para las c frecuencias de categorías ordenadas y disjuntas de cada grupo de tratamiento; y se aplican a las c categorías ordenadas y disjuntas del otro grupo de tratamiento. En esta aplicación, los $k = 2$ grupos de tratamiento se consideran muestras finitas independientes, sin que ninguno se identifique como grupo de referencia. El supuesto de la segunda aplicación de que los grupos son finitos es, en realidad, más realista. En 2009 Mielke, Long, Berry y Johnston generalizaron el análisis ridit para $k \geq 2$ grupos de tratamiento independientes [Mielke et al., 2009].

Se considera una tabla de contingencia de clasificación cruzada $c \times k$ con c categorías de respuesta disjuntas y ordenadas y k grupos de tratamiento desordenados. Siguiendo la notación de Bross, m_{ij} denota la frecuencia de celda observada de la i-ésima fila y j-ésima columna para $i = 1, ..., c$ y $j = 1, ..., k$, sean

$$M_j = \sum_{i=1}^{c} m_{ij}$$

los totales de frecuencia de tratamiento no ordenados para $j = 1, ..., k$, y sea

$$N = \sum_{i=1}^{c} \sum_{j=1}^{k} m_{ij}$$

el total de la frecuencia de la tabla para todas las ck celdas. Las puntuaciones ridit para el j-ésimo tratamiento observado, $j = 1, ..., k$, vienen dadas por

$$R_{1j} = \frac{m_{1j}}{2M_j} \ , \qquad R_{2j} = \frac{m_{1j} + \frac{m_{2j}}{2}}{M_j} \ , \ \ ... \ , \qquad R_{cj} = \frac{m_{1j} + ... + m_{c-1,j} + \frac{m_{cj}}{2}}{M_j} \ .$$

Así, la puntuación ridit R_{ij} para la i-ésima de las c categorías en el j-ésimo de los k tratamientos es la proporción de observaciones en las categorías inferiores a la i-ésima categoría en el j-ésimo tratamiento mas la mitad de la proporción de observaciones en la i-ésima categoría del j-ésimo tratamiento.

4.6.1. Cálculo

Se defina el estadístico de prueba T como

$$T = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |x_{ij} - x_{ji}| \ ,$$

donde

$$x_{ij} = \sum_{k=1}^{c} \frac{R_{ki} m_{kj}}{M_j}$$

para $i, j = 1, ..., k$.

En el contexto de un análisis ridit de k tratamientos, los procedimientos de permutación exacta examinan todas las posibles asignaciones igualmente probables de los N sujetos a las c categorías ordenadas y disjuntas. Alternativamente, los procedimientos de permutación de remuestreo de Monte Carlo examinan un subconjunto aleatorio seleccionado de entre todas las posibles asignaciones de los N sujetos a las c categorías ordenadas y separadas. La hipótesis nula de una prueba de permutación especifica que todos los resultados posibles del análisis ridit son igualmente probables.

4.6.1.1. Procedimientos de permutación exacta

Los M_j sujetos del j-ésimo grupo de tratamiento, $j = 1, ..., k$, se clasifican en c categorías disjuntas y ordenadas. Entre las c^N configuraciones de asignación igualmente probables bajo la hipótesis nula, hay

$$W = \prod_{j=1}^{k} \binom{M_j + c - 1}{c - 1}$$

particiones distintas de las c^N configuraciones de asignación de los k grupos de tratamiento. En una aplicación típica, W y c^N suelen ser muy grandes.

Por lo tanto, un análisis de permutación exacto no suele ser práctico para los análisis ridit con $k > 2$ tratamientos y se recomiendan los procedimientos de permutación de Monte Carlo.

4.6.1.2. Procedimientos de permutación de remuestreo

Un procedimiento de permutación de remuestreo de Monte Carlo genera L conjuntos de \$N\$ asignaciones aleatorias seleccionadas con reemplazo de las c^N configuraciones de asignación igualmente probables de los k grupos de tratamiento. En general, $L = 1.000.000$

es suficiente para garantizar una precisión de tres decimales [Johnston et al., 2007]. Para cada uno de los L conjuntos, los contadores de las c categorías disjuntas y ordenadas indexadas por $i = 1, ..., c$ se ponen a cero y se generan j variables aleatorias uniformes independientes, U_j, sobre $[0, 1)$, para $j = 1, ..., N$. Si U_j pertenece a $[\frac{i-1}{c}, \frac{i}{c})$, el i-ésimo de los contadores c se incrementa en 1. A continuación se calcula el estadístico T de la prueba ridit para cada uno de los L conjuntos de N asignaciones aleatorias de las frecuencias de las categorías ordenadas. Sea T_o el valor observado de T. Entonces, dados los estadísticos ridit del remuestreo $T_1, ..., T_L$, el valor de la probabilidad de la cola superior del remuestreo de T_o bajo la hipótesis nula viene dado por

$$P = \frac{1}{L} \sum_{i=1}^{L} \Psi(T_i) \ ,$$

donde

$$\Psi(T_i) = \begin{cases} 1 & T_i \geq T_o \\ 0 & c.c. \end{cases} \ .$$

Capítulo 5

Ilustración sobre datos reales

A la hora de seleccionar una medida de asociación adecuada, hay que tener en cuenta varios criterios. Liebetrau proporciona algunas directrices para seleccionar la medida más apropiada [Liebetrau, 1983]:

En primer lugar, nos planteamos si las variables son nominales, ordinales, de intervalo o alguna combinación de las tres. Una medida de asociación para variables de nivel nominal (categóricas) no debería depender de categorías ordenadas. Por otro lado, una medida de asociación para variables de nivel ordinal debería depender de categorías ordenadas, ya que si se ignora el orden de las categorías, se pierde información.

En segundo lugar, ¿se medirá la correlación, la asociación o la concordancia? En general, las medidas asimétricas son adecuadas para la predicción, mientras que las medidas simétricas son apropiadas para la asociación o la correlación, dependiendo del nivel de medición.

En tercer lugar, ¿se va a utilizar la medida de asociación para hacer inferencia? Bajo el modelo poblacional de Neyman-Pearson, esto requiere conocer el error estándar del estimador y a menudo requiere hacer suposiciones sobre la naturaleza de la población, así como un muestreo aleatorio. Bajo el modelo de permutación de Fisher-Pitman, no se requiere conocer el error estándar, no es necesario el muestreo aleatorio, los supuestos de distribución son irrelevantes y las pruebas de permutación dependen completamente de los datos. Las probabilidad basadas en permutaciones pueden ser exactas, basadas en el conjunto de referencia de todas las permutaciones posibles de los datos observados, o aproximadas, basados en una gran muestra aleatoria extraída del conjunto de referencia con el método Monte Carlo.

En cuarto lugar, ¿es la medida de asociación sensible a los cambios de frecuencias marginales? Que la medida de asociación no se altere al multiplicar o dividir una o ambas columnas o filas de la tabla de contingencia por cualquier factor arbitrario es una propiedad muy importante de la medida. En general, los valores de una medida de asociación calculados en dos muestras diferentes no pueden compararse si la medida depende de las frecuencias marginales. Prácticamente todas las medidas de asociación para las variables de nivel nominal y ordinal son sensibles a los cambios de frecuencias marginales. Algunas excepciones notables son la odds ratio, las diferencias porcentuales y la medida de asociación Q de Yule.

En quinto lugar, ¿es la medida de asociación estable ante cambios en el número de categorías? Una medida de asociación estable es aquella en la que el valor no cambia cuando, por ejemplo, el número de categorías se cambia de cinco a cuatro. La medida gamma de asociación ordinal de Goodman y Kruskal es especialmente inestable, mientras que la medida τ_b de asociación ordinal de Kendall es relativamente estable.

En sexto lugar, ¿es el valor de la medida de asociación fácilmente interpretable? Algunas medidas de asociación tienen interpretaciones claras y significativas, como las medidas de reducción proporcional del error y las medidas corregidas por el azar. Otras medidas no tienen una interpretación significativa, excepto cuando poseen los valores extremos de 0, 1 o -1. Por ejemplo, la medida simétrica gamma de Goodman y Kruskal de asociación ordinal y las medidas asimétricas d_{yx} y d_{xy} de Somers de asociación ordinal poseen interpretaciones de reducción proporcional en el error, donde los valores positivos indican una mejora proporcional de los errores de predicción con el conocimiento de ambas variables, en comparación de los errores de predicción con el conocimiento de una sola variable. Las medidas kappa no ponderada y kappa ponderada de Cohen para la concordancia y la regla del pie de Spearman poseen interpretaciones corregidas por el azar, en las que los valores positivos indican una concordancia superior a la esperada por el azar y los valores negativos indican una concordancia inferior.

En séptimo lugar, ¿la medida de asociación tiene en cuenta los valores empatados? Algunas medidas de asociación requieren ajustes complicados para usar los valores empatados, mientras que otras incorporan los valores empatados sin ningún ajuste. Por ejemplo, el coeficiente de correlación de rango de Spearman requiere complejos ajustes para los valores empatados.

En octavo lugar, ¿se generalizará fácilmente la medida de asociación para datos multivariantes? El análisis multivariante es cada vez más importante en la investigación, por lo que las medidas de asociación que se adaptan a los datos multivariantes son extremadamente útiles.

En noveno lugar, ¿para qué tipo de asociación asume la medida de asociación su valor extremo? Algunas medidas de asociación asumen sus valores extremos, por ejemplo, $+1$ en casos de asociación débil, como es el caso del estadístico gamma de Goodman y Kruskal. Otras medidas de asociación asumen sus valores extremos sólo en el caso de una asociación perfecta. Por ejemplo, la medida τ_b de Kendall de asociación ordinal asume un valor de $+1$ sólo cuando existe una asociación extremadamente fuerte.

En décimo lugar, ¿es la medida de asociación fácil de calcular? Algunas medidas de asociación son notoriamente difíciles de calcular. La medida de correlación tetracórica de Pearson para tablas de contingencia 2×2 o la medida de asociación nominal d_N^c de Leik y Gove son de las más dificiles de calcular. Por otro lado, la medida de la regla del pie de Spearman se diseñó específicamente para facilitar su cálculo.

Una vez desarrolladas cada una de las medidas de asociación, tanto nominales como ordinales, y describir los criterios que se usan para seleccionar la más adecuada, se va a proceder a mostrar un caso práctico donde aplicar las diferentes medidas.

5.1. Conjunto de datos

Los datos que se usarán son los datos reales correspondientes a la encuesta sobre tendencias sociales realizada por el Centro de Investigación Sociológica[1] (C.I.S.) en España del 24 al 30 de noviembre de 2021.

Los datos corresponden a las respuestas de una muestra de 2691 personas procedentes de 888 municipios y 50 provincias. El procedimiento de muestreo fue el siguiente: se seleccionaron aleatoriamente teléfonos fijos y móviles con un porcentaje del 22,0 % y del 78,0 %, respectivamente. La selección de los individuos se llevó a cabo mediante la aplicación de cuotas de sexo y edad. Los estratos se formaron de forma independiente considerando las 17 comunidades autónomas y las dos ciudades autónomas y el tamaño de hábitat, dividido en 7 categorías: menor o igual a 2.000 habitantes; de 2.001 a 10.000; de 10.001 a 50.000; de 50.001 a 100.000; de 100.001 a 400.000; de 400.001 a 1.000.000; y más de 1.000.000 de habitantes. Los cuestionarios aplicaron mediante entrevista telefónica asistida por ordenador (CATI).

La encuesta está formada por 33 preguntas (véase Apéndice A) y de ella se han extraído las variables más interesantes para estudiar el grado de asociación entre las preguntas. Además, las variables se han limpiado y agrupado para que tengan el menor número de categorías posibles (véase Apéndice B) para facilitar la interpretación de las medidas.

Las variables ecogidas han sido las siguientes:

Variables Nominales

- SEXO: Sexo de la persona que realiza la encuesta.
- CCAA: Comunidad Autónoma a la que está empadronada.
- INESTADO: Respuesta de la persona ante la pregunta ¿debería intervenir el Estado en la vida económica?
- INFLUPOLITICA: Mayor influencia en las decisiones políticas de los últimos dos años.
- SISTEMAUTOM1: Cómo afectará al número de parados la utilización de robots y sistemas automáticos de trabajo en los próximos diez años.
- SIMBOLO: Objeto (artefacto, aparato, máquina) que identifica y simboliza en mayor grado la época actual.
- IDENTPERSONAS: Grupo de personas con el que piensa el encuestado que tiene más intereses comunes.
- PARTIDOSIMP: Partido político con el que simpatiza más la persona encuestada.
- SITLABORAL: Situación laboral en la que se encuentra actualmente.
- RELIGION: Religión en la que cree la persona encuestada.

Variables Ordinales

- EDAD: Rango de edad de la persona encuestada (intervalos de amplitud 10 años).
- INTPOLITICO: Grado de interes de la persona por las cuestiones políticas.
- FUTURO1: Cantidas de cambios sociales y económicos de aquí a diez años según las personas que han realizado la encuesta.
- FUTURO2: Grado de positividad ante los cambios sociales y económicos producidos de aquí a diez años.

[1] https://www.cis.es/cis/opencm/ES/2_bancodatos/estudios/listaEstudiosYear.jsp?year=2021

- SISTEMAUTOM2: Aumento del número de robots y sistemas automáticos en las empresas en los próximos diez años (mucho, bastante, algo, poco, nada o casi nada).
- ESCIDEOL: Posición que la persona tiene en una escala del 1 al 10 respecto a los ideales políticos, en la que 1 significa "lo más a la izquierda" y 10 "lo más a la derecha".
- ESTUDIOS: Grado de estudios básicos (No sabe leer ni escribir; no ha ido a la escuela, pero sabe leer y escribir; ha ido a la escuela).
- OPINION: ¿Las nuevas generaciones vivirán mejor, peor o igual de lo que se ha vivido hasta ahora?
- CLASESOCIAL: Clase social que mejor define al encuestado.

5.2. Medidas de Asociación para Variables Nominales

Se realizará un análisis descriptivo y se aplicarán las diferentes medidas de asociación, creadas para variables categóricas, a los siguientes pares de variables:

- SEXO - IDENTPERSONAS
- SEXO - INESTADO
- CCAA - SIMBOLO
- RELIGION - PARTIDOSIMP
- SITLABORAL - SISTEMAUTOM1
- INFLUPOLITICA - PARTIDOSIMP
- INESTADO - PARTIDOSIMP

Para así, conocer el grado de asociación entre ellas y poder comparar las distintas medidas.

5.2.1. SEXO - IDENTPERSONAS

En primer lugar, se estudia si a partir del sexo de la persona se puede predecir el grupo de personas con el que tiene más intereses comunes.

Un resumen de los datos muestrales se observa en la siguiente tabla, donde aparece el total de personas para cada categoría de la variable IDENTPERSONAS diferenciado por la variable SEXO. Además, también aparece el porcentaje para ambos sexos de cada una de las categorías:

Tabla 5.1: Variables IDENTPERSONAS y SEXO

IDENTPERSONAS	SEXO	Total	Porcentaje
Aficiones y Gustos	Hombre	292	0.5378
Aficiones y Gustos	Mujer	251	0.4622
Clase Social	Hombre	95	0.6786
Clase Social	Mujer	45	0.3214
Generación	Hombre	396	0.5388
Generación	Mujer	339	0.4612
Género	Hombre	24	0.1633

Tabla 5.1: Variables IDENTPERSONAS y SEXO (continúa)

IDENTPERSONAS	SEXO	Total	Porcentaje
Género	Mujer	123	0.8367
Ideas Políticas	Hombre	112	0.5209
Ideas Políticas	Mujer	103	0.4791
Ideas Religiosas	Hombre	27	0.4655
Ideas Religiosas	Mujer	31	0.5345
Municipio	Hombre	81	0.6328
Municipio	Mujer	47	0.3672
N.S/N.C	Hombre	164	0.4838
N.S/N.C	Mujer	175	0.5162
Profesión/Trabajo	Hombre	168	0.5526
Profesión/Trabajo	Mujer	136	0.4474
Religión/Nacionalidad	Hombre	41	0.5000
Religión/Nacionalidad	Mujer	41	0.5000

A continuación, se representan dichos datos en un gráfico de barras para facilitar su interpretación:

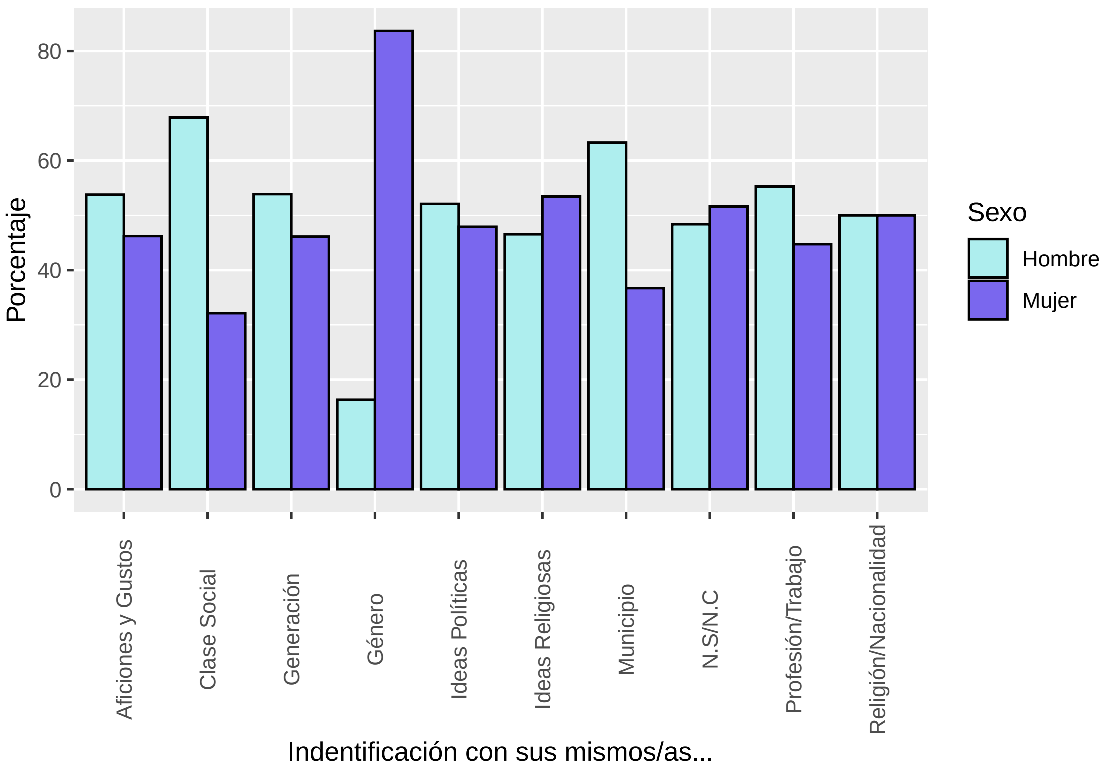

Figura 5.1: Gráfico de barras SEXO-IDENTPERSONAS

A través del gráfico, se intuye que no habrá demasiada asociación entre las dos variables, ya que parece que la variable SEXO no influye demasiado en la categoría de la variable IDENTPERSONAS. Las categorías que están más diferenciadas por el sexo son: Género, Clase Social y Municipio.

Para calcular las diferentes medidas de asociación se considera que la variable dependiente es IDENTPERSONAS.

Las diferentes medidas obtenidas se observan en la siguiente tabla:

Tabla 5.2: Medidas de variables IDENTPERSONAS y SEXO

	Estadístico	Límite Inferior IC	Límite Superior IC
λ de Goodman y Krukal	0.0923	0.0683	0.1163
t de Goodman y Kruskal	0.0422	0.0281	0.0563

Se observa que ambas medidas, tanto la λ como la t de Goodman y Kruskal, presentan valores cercanos a 0, es decir, la asociación entre ambas variables no es demasiado significativa. Aunque cabe destacar que como el 0 no pertenece al intervalo de confianza, la asociación tampoco se consideraría nula.

Como se observó en el gráfico 5.1, las variables en las que el sexo toma un papel más importante son Género, Municipio y Clase Social. A continuación, se calculan de nuevos estas medidas usando únicamente esas categorías de la variable IDENTPERSONAS:

Tabla 5.3: Medidas de variables IDENTPERSONAS y SEXO (2)

	Estadístico	Límite Inferior IC	Límite Superior IC
λ de Goodman y Krukal	0.4200	0.2978	0.5422
t de Goodman y Kruskal	0.2244	0.1488	0.3001

Se puede ver cómo ambas medidas han aumentado considerablemente sus valores, y por tanto, restringiénsonos a estas categorías, se obtiene que el conocimiento de los valores de la variable SEXO permite reducir la incertidumbre en la predicción del comportamiento de la variable IDENTPERSONAS en un 42 % y un 22,44 %, respectivamente.

5.2.2. SEXO - INESTADO

Igualmente, se plantea ahora si la variable sexo influye en la respuesta de cada persona a la pregunta ¿debería intervenir el Estado en la vida económica? y viceversa.

Realizando el mismo proceso, se obtiene la tabla:

Tabla 5.4: Variables INESTADO y SEXO

INESTADO	SEXO	Total	Porcentaje
N.S/N.C	Hombre	118	0.4591
N.S/N.C	Mujer	139	0.5409
No	Hombre	350	0.6119
No	Mujer	222	0.3881
Sí	Hombre	932	0.5005
Sí	Mujer	930	0.4995

Y el gráfico:

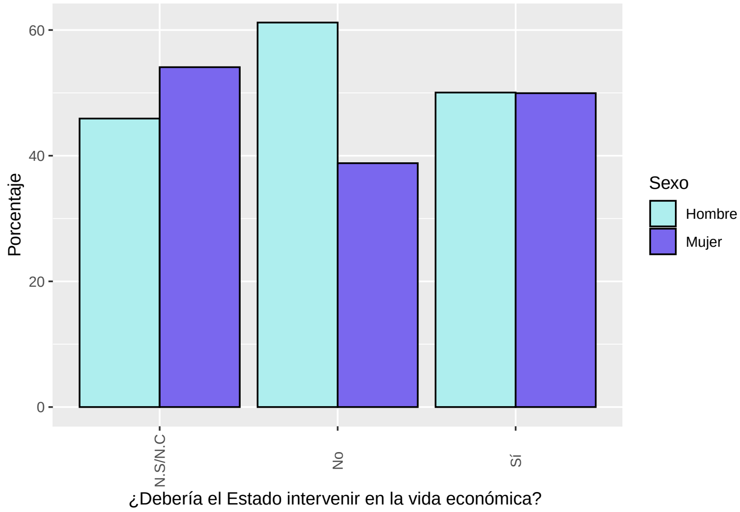

Figura 5.2: Gráfico de barras SEXO-INESTADO

Mediante el análisis descriptivo no hay signos de que haya asociación entre las variables. Esto se comprueba con los valores que obtenemos al calcular las medidas de asociación siguientes:

Tabla 5.5: Medidas de variables INESTADO y SEXO

	Estadístico	Límite Inferior IC	Límite Superior IC
$\lambda_{INESTADO}$ de Goodman y Krukal	0.0000	0.0000	0.0000
λ_{SEXO} de Goodman y Krukal	0.0000	0.0000	0.0000
$t_{INESTADO}$ de Goodman y Kruskal	0.0089	0.0016	0.0163
t_{SEXO} de Goodman y Kruskal	0.0089	0.0016	0.0163

Mediante este ejemplo, se puede observar la particularidad de que λ puede tomar el valor de 0 para casos distintos a la independencia. En este caso se ha obtenido la tabla de contingencia:

	Hombre	Mujer
No	350	222
Sí	932	930

En esta situación, la moda de la categorías de INESTADO coinciden que son la misma categoría de la variable SEXO (Hombre), por tanto,

$$W = \sum_{i=1}^{2} max(n_{i1}, n_{i2}) = 350 + 932 = 1282; \qquad A = max(n_{.1}, n_{.2}) = 1282$$

y así,

$$\lambda_{INESTADO} = \frac{W - A}{N - A} = 0 \ .$$

Análogamente, la moda de la categorías de SEXO coinciden que son la misma categoría de la variable INESTADO (Sí), por tanto,

$$Z = \sum_{j=1}^{2} max(n_{1j}, n_{2j}) = 932 + 930 = 1862; \qquad B = max(n_{1.}, n_{2.}) = 1862$$

y así,

$$\lambda_{SEXO} = \frac{Z - B}{N - B} = 0 \ .$$

5.2.3. CCAA - SIMBOLO

A continuación, se plantea la cuestión de si hay relación entre la elección del símbolo que mejor define la época actual y la comunidad autónoma a la que pertenecen los encuestados. Gráficamente:

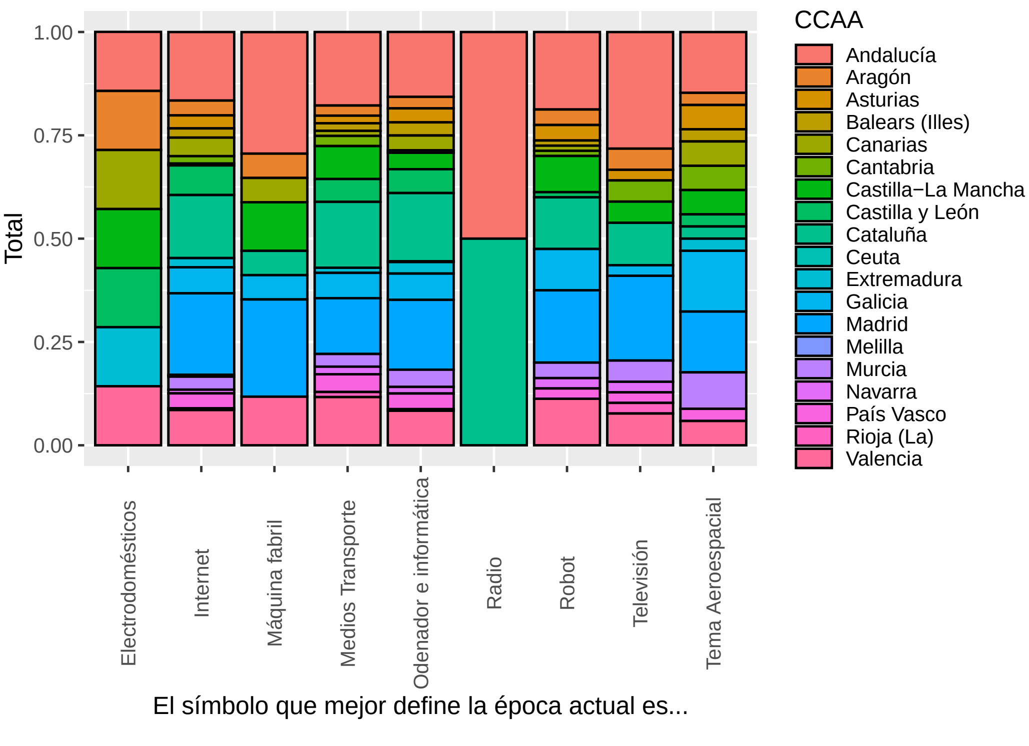

Figura 5.3: Gráfico de barras CCAA-SÍMBOLO

Las medidas se asociación aparecen recogidas en la siguiente tabla:

Tabla 5.7: Medidas de variables SIMBOLO y CCAA

	Estadístico	Límite Inferior IC	Límite Superior IC
$\lambda_{SIMBOLO}$ de Goodman y Krukal	0.0147	0.0000	0.0491
λ_{CCAA} de Goodman y Krukal	0.0035	0.0000	0.0155
$t_{SIMBOLO}$ de Goodman y Kruskal	0.0068	0.0039	0.0097
t_{CCAA} de Goodman y Kruskal	0.0163	0.0090	0.0235

Sin importar qué variable se tome como dependiente, no hay evidencias para rechazar la hipótesis nula de que $\lambda = 0$, y por tanto, con esta medida las variables SIMBOLO y CCAA no presentan asociación. Sin embargo, con la medida t de Goodman y Kruskal, la asociación sí es significativa aunque con el conocimiento de las variables CCAA y SIMBOLO sólo permite reducir la incertidumbre en la predicción del comportamiento de la otra variable en un 0,68 % y un 1,62 %, respectivamente.

5.2.4. RELIGION - PARTIDOSIMP

Seguidamente se cuantifica la asociación entre las creencias religiosas y los partidos políticos más afines a los entrevistados.

Análisis descriptivo:

Tabla 5.8: Variables PARTIDOSIMP y RELIGIÓN

PARTIDOSIMP	RELIGION	Total	Porcentaje
BNG	Agnóstico/a, indiferente, ateo/a	9	0.8182
BNG	Católico/a no practicante	1	0.0909
BNG	Creyente de otra religión	1	0.0909
CCa-PNC-NC	Católico/a no practicante	2	0.6667
CCa-PNC-NC	Católico/a practicante	1	0.3333
Ciudadanos	Agnóstico/a, indiferente, ateo/a	30	0.2752
Ciudadanos	Católico/a no practicante	64	0.5872
Ciudadanos	Católico/a practicante	14	0.1284
CUP	Agnóstico/a, indiferente, ateo/a	16	0.8000
CUP	Católico/a no practicante	2	0.1000
CUP	Católico/a practicante	1	0.0500
CUP	Creyente de otra religión	1	0.0500
EAJ-PNV	Agnóstico/a, indiferente, ateo/a	6	0.3158
EAJ-PNV	Católico/a no practicante	9	0.4737
EAJ-PNV	Católico/a practicante	3	0.1579
EH Bildu	Agnóstico/a, indiferente, ateo/a	14	0.7778
EH Bildu	Católico/a no practicante	3	0.1667
EH Bildu	Católico/a practicante	1	0.0556
ERC	Agnóstico/a, indiferente, ateo/a	35	0.6034
ERC	Católico/a no practicante	16	0.2759
ERC	Católico/a practicante	2	0.0345
ERC	Creyente de otra religión	4	0.0690
JxCat	Agnóstico/a, indiferente, ateo/a	18	0.5000
JxCat	Católico/a no practicante	13	0.3611
JxCat	Católico/a practicante	5	0.1389
Más País	Agnóstico/a, indiferente, ateo/a	35	0.7143
Más País	Católico/a no practicante	11	0.2245
Más País	Católico/a practicante	2	0.0408
Más País	Creyente de otra religión	1	0.0204
Més Compromís	Agnóstico/a, indiferente, ateo/a	18	0.7200
Més Compromís	Católico/a no practicante	6	0.2400

Tabla 5.8: Variables PARTIDOSIMP y RELIGIÓN (continúa)

PARTIDOSIMP	**RELIGION**	**Total**	**Porcentaje**
Més Compromís	Católico/a practicante	1	0.0400
Navarra Suma	Católico/a no practicante	2	0.5000
Navarra Suma	Católico/a practicante	2	0.5000
PACMA	Agnóstico/a, indiferente, ateo/a	12	0.5217
PACMA	Católico/a no practicante	8	0.3478
PACMA	Creyente de otra religión	2	0.0870
PP	Agnóstico/a, indiferente, ateo/a	45	0.1372
PP	Católico/a no practicante	172	0.5244
PP	Católico/a practicante	108	0.3293
PP	Creyente de otra religión	3	0.0091
PRC	Agnóstico/a, indiferente, ateo/a	1	1.0000
PSOE	Agnóstico/a, indiferente, ateo/a	219	0.4179
PSOE	Católico/a no practicante	206	0.3931
PSOE	Católico/a practicante	75	0.1431
PSOE	Creyente de otra religión	16	0.0305
Teruel Existe	Católico/a no practicante	1	1.0000
Unidas Podemos	Agnóstico/a, indiferente, ateo/a	190	0.7600
Unidas Podemos	Católico/a no practicante	53	0.2120
Unidas Podemos	Católico/a practicante	1	0.0040
Unidas Podemos	Creyente de otra religión	5	0.0200
VOX	Agnóstico/a, indiferente, ateo/a	33	0.2062
VOX	Católico/a no practicante	70	0.4375
VOX	Católico/a practicante	52	0.3250
VOX	Creyente de otra religión	4	0.0250

Esta vez, no se representan los porcentajes, sino las frecuencias absolutas. Así se ve también cuáles son los partidos mayoritarios:

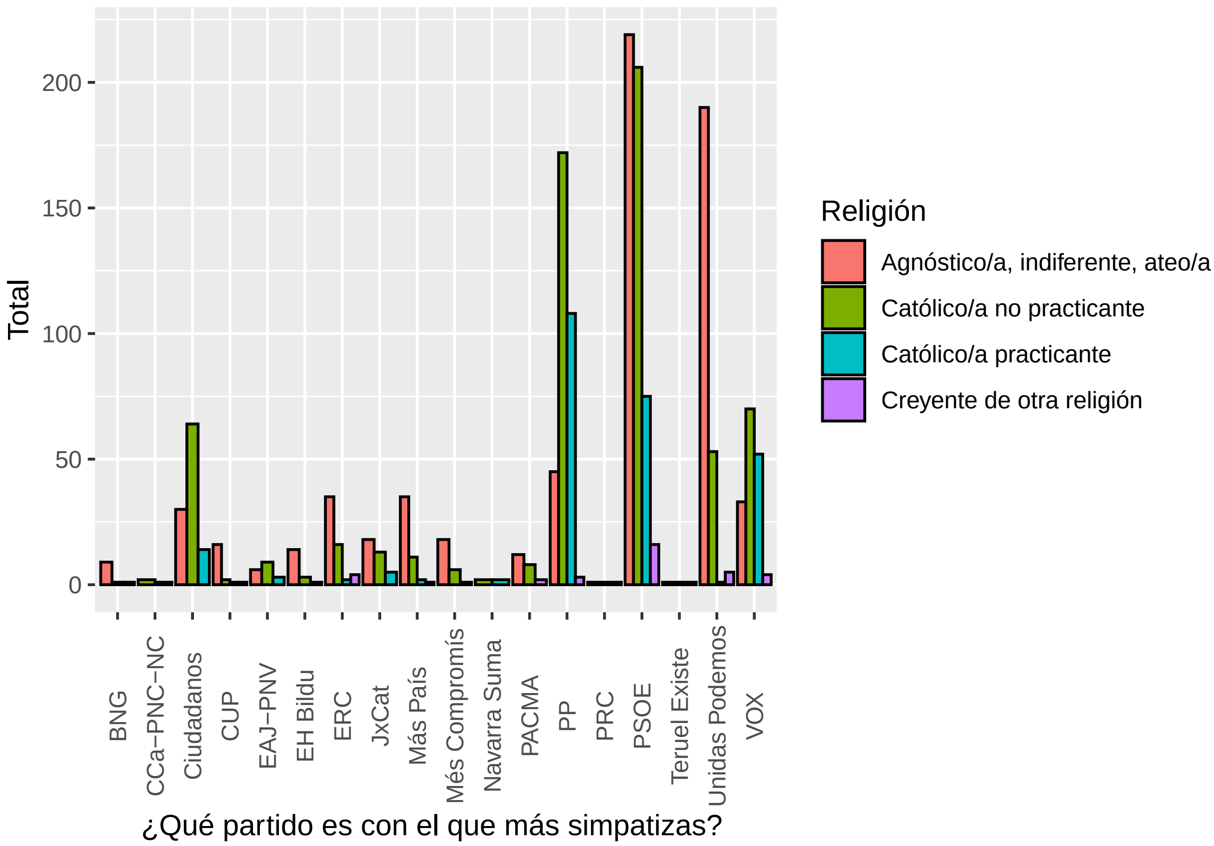

Figura 5.4: Gráfico de barras RELIGIÓN-PARTIDOSIMP

Considerando la variable PARTIDOSIMP como la variable dependiente, se obtienen las siguientes medidas:

Tabla 5.9: Medidas de variables PARIDOSIMP y RELIGION

	Estadístico	Límite Inferior IC	Límite Superior IC
λ de Goodman y Krukal	0.2182	0.1800	0.2565
t de Goodman y Kruskal	0.1276	0.1058	0.1495

Se puede considerar que hay asociación entre las variables RELIGIÓN y PARTIDOSIMP. Más concretamente, el conocimiento de los valores de la variable RELIGION permite reducir la incertidumbre en la predicción de la variable PARTIDOSIMP en un 21,82 % y un 12,76 %, considerando las medidas λ y t de Goodman y Kruskal respectivamente.

5.2.5. SITLABORAL - SISTEMAUTOM1

Se plantea ahora la cuestión de si a través de la situación laboral se puede predecir cómo afectará al número de parados la utilización de robots y sistemas automáticos de trabajo en los próximos diez años.

Los datos muestrales son los siguientes:

Tabla 5.10: Variables SISTEMAUTOM1 y SITLABORAL

SISTEMAUTOM1	**SITLABORAL**	**Total**	**Porcentaje**
Aumento paro	Estudiante	43	0.0303
Aumento paro	Jubilado/a o pensionista (anteriormente ha trabajado)	302	0.2131
Aumento paro	N.C.	24	0.0169
Aumento paro	Paro y busca primer empleo	8	0.0056
Aumento paro	Paro y ha trabajado	163	0.1150
Aumento paro	Pensionista (no ha trabajado)	27	0.0191
Aumento paro	Trabaja	788	0.5561
Aumento paro	Trabajo doméstico no remunerado	62	0.0438
Más trabajo	Estudiante	17	0.0445
Más trabajo	Jubilado/a o pensionista (anteriormente ha trabajado)	63	0.1649
Más trabajo	N.C.	4	0.0105
Más trabajo	Paro y busca primer empleo	1	0.0026
Más trabajo	Paro y ha trabajado	22	0.0576
Más trabajo	Pensionista (no ha trabajado)	2	0.0052
Más trabajo	Trabaja	261	0.6832
Más trabajo	Trabajo doméstico no remunerado	12	0.0314
Misma situación	Estudiante	20	0.0257
Misma situación	Jubilado/a o pensionista (anteriormente ha trabajado)	149	0.1913
Misma situación	N.C.	9	0.0116
Misma situación	Paro y busca primer empleo	3	0.0039
Misma situación	Paro y ha trabajado	40	0.0513
Misma situación	Pensionista (no ha trabajado)	7	0.0090
Misma situación	Trabaja	529	0.6791
Misma situación	Trabajo doméstico no remunerado	22	0.0282
N.S/N.C.	Estudiante	1	0.0088
N.S/N.C.	Jubilado/a o pensionista (anteriormente ha trabajado)	33	0.2920
N.S/N.C.	N.C.	1	0.0088
N.S/N.C.	Paro y ha trabajado	10	0.0885
N.S/N.C.	Trabaja	61	0.5398
N.S/N.C.	Trabajo doméstico no remunerado	7	0.0619

A través de ellos, se obtiene el gráfico:

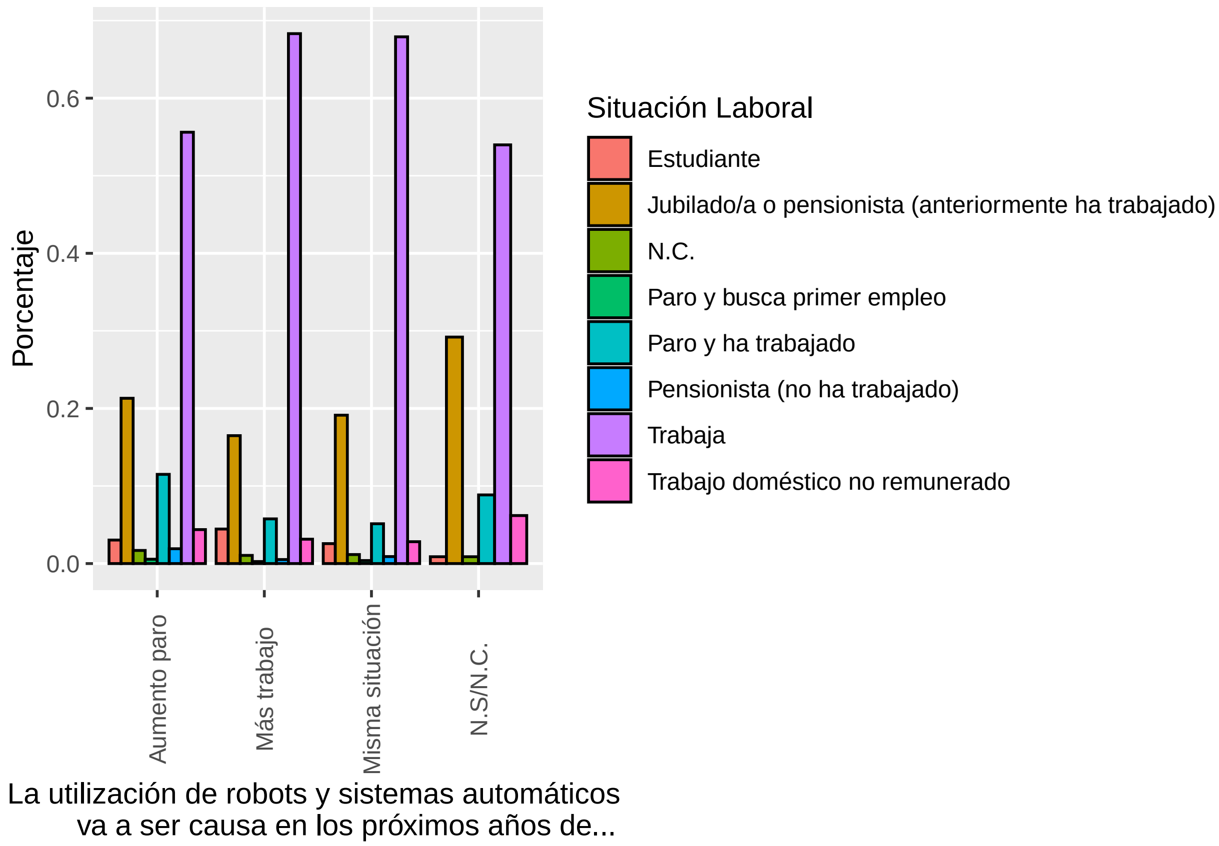

Figura 5.5: Gráfico de barras SITLABORAL-SISTEMAUTOM1

Las medidas de asociación que se obtienen, considerando la variable SISTEMAUTOM1, son:

Tabla 5.11: Medidas de variables SISTEMAUTOM1 y SITLABORAL

	Estadístico	Límite Inferior IC	Límite Superior IC
λ de Goodman y Krukal	0.0000	0.000	0.0000
t de Goodman y Kruskal	0.0092	0.004	0.0143

Como $\lambda = 0$, indica que el conocimiento de los valores de una variable no permite predecir el comportamiento de la otra variable.

El estadístico t de Goodman toma el valor de 0.0092, este resultado confirma las conclusiones comentadas justo antes.

Se obtienen las respectivas asociaciones correspondientes a varios aspectos políticos:

5.2.6. INFLUPOLITICA - PARTIDOSIMP

Se consideran únicamente los partidos que obtuvieron más votos, es decir, PSOE, PP, Unidas Podemos, VOX y Ciudadanos; y se observa si existe relación con cuál ha sido la mayor influencia en las decisiones políticas de los últimos dos años.

Tabla 5.12: Variables PARTIDOSIMP y INFLUPOLITICA

PARTIDOSIMP	INFLUPOLITICA	Total	Porcentaje
Ciudadanos	Candidatos	32	0.2936
Ciudadanos	Conocido Concreto	7	0.0642
Ciudadanos	Internet	8	0.0734
Ciudadanos	N.S/N.C	29	0.2661
Ciudadanos	Prensa	8	0.0734
Ciudadanos	Prog. Electorales	2	0.0183
Ciudadanos	Radio	9	0.0826
Ciudadanos	Televisión	14	0.1284
PP	Candidatos	101	0.3079
PP	Conocido Concreto	38	0.1159
PP	Internet	20	0.0610
PP	N.S/N.C	89	0.2713
PP	Prensa	17	0.0518
PP	Prog. Electorales	11	0.0335
PP	Radio	25	0.0762
PP	Televisión	27	0.0823
PSOE	Candidatos	159	0.3034
PSOE	Conocido Concreto	47	0.0897
PSOE	Internet	34	0.0649
PSOE	N.S/N.C	167	0.3187
PSOE	Prensa	36	0.0687
PSOE	Prog. Electorales	11	0.0210
PSOE	Radio	29	0.0553
PSOE	Televisión	41	0.0782
Unidas Podemos	Candidatos	87	0.3480
Unidas Podemos	Conocido Concreto	27	0.1080
Unidas Podemos	Internet	28	0.1120
Unidas Podemos	N.S/N.C	66	0.2640
Unidas Podemos	Prensa	16	0.0640
Unidas Podemos	Prog. Electorales	6	0.0240
Unidas Podemos	Radio	12	0.0480
Unidas Podemos	Televisión	8	0.0320
VOX	Candidatos	61	0.3812
VOX	Conocido Concreto	13	0.0813
VOX	Internet	16	0.1000
VOX	N.S/N.C	43	0.2688
VOX	Prensa	10	0.0625
VOX	Prog. Electorales	5	0.0312
VOX	Radio	7	0.0437
VOX	Televisión	5	0.0312

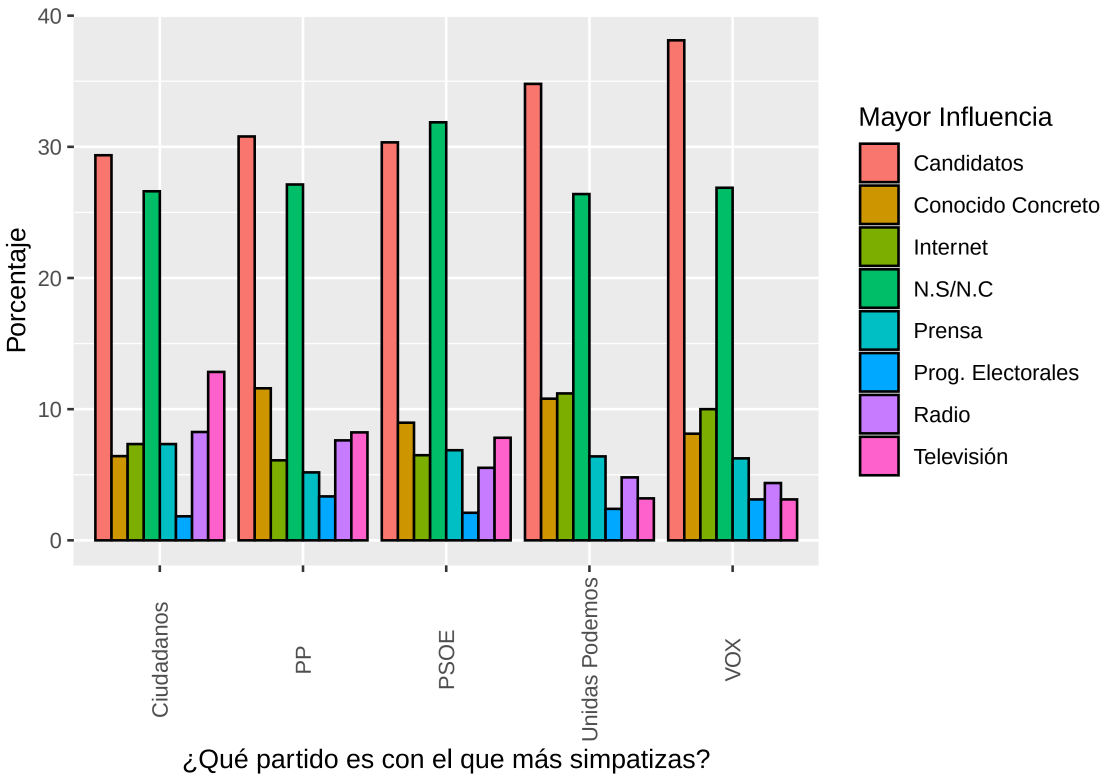

Figura 5.6: Gráfico de barras INFLUPOLITICA-PARTIDOSIMP

Observando el gráfico, parece que para los principales partidos, la influencia en las decisiones políticas es similar; por lo que debería traducirse en que que no hay (o hay poca asociación) entre ambas variables. Esto se verifica a través de las medidas de asociación:

Tabla 5.13: Medidas de variables PARTIDOSIMP e INFLUPOLITICA

	Estadístico	Límite Inferior IC	Límite Superior IC
λ de Goodman y Krukal	0.0086	0e+00	0.0464
t de Goodman y Kruskal	0.0039	8e-04	0.0070

5.2.7. INESTADO - PARTIDOSIMP

Para finalizar con las medidas de asociación nominal, se estudia la asociación entre la elección del martido más afín a la persona y su opición sobre si el Estado debería intervenir en la vida económica.

Un resumen se los datos se muestra en la siguiente tabla y gráfico:

Tabla 5.14: Variables PARTIDOSIMP y INESTADO

PARTIDOSIMP	INESTADO	Total	Porcentaje
Ciudadanos	No	33	0.3028
Ciudadanos	Sí	63	0.5780
PP	No	140	0.4268

Tabla 5.14: Variables PARTIDOSIMP y INESTADO (continúa)

PARTIDOSIMP	INESTADO	Total	Porcentaje
PP	Sí	152	0.4634
PSOE	No	57	0.1088
PSOE	Sí	437	0.8340
Unidas Podemos	No	13	0.0520
Unidas Podemos	Sí	227	0.9080
VOX	No	76	0.4750
VOX	Sí	69	0.4312

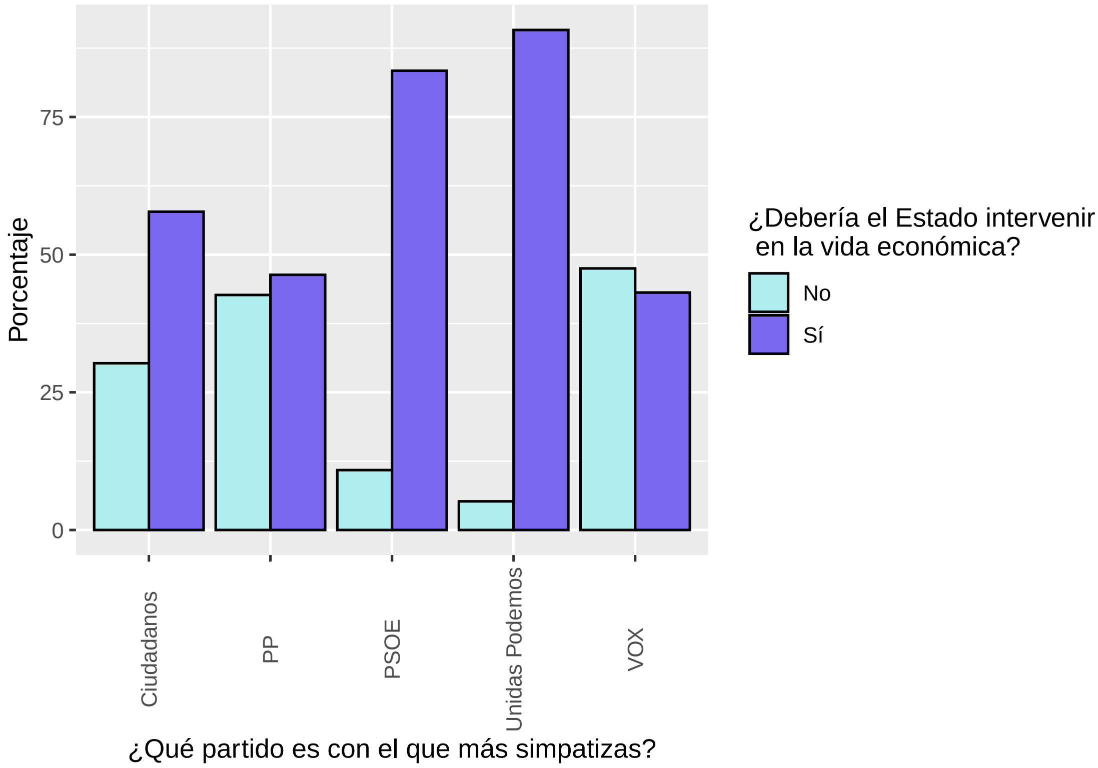

Figura 5.7: Gráfico de barras INESTADO-PARTIDOSIMP

Las medidas de asociación que se obtienen son:

Tabla 5.15: Medidas de variables PARTIDOSIMP e INESTADO

	Estadístico	Límite Inferior IC	Límite Superior IC
$\lambda_{PARIDOSIMP}$ de Goodman y Krukal	0.0219	0.0000	0.0951
$\lambda_{INESTADO}$ de Goodman y Krukal	0.1074	0.0738	0.1410
$t_{PARIDOSIMP}$ de Goodman y Kruskal	0.1896	0.1469	0.2323
$t_{INESTADO}$ de Goodman y Kruskal	0.0574	0.0431	0.0717

Tomando la variable PARIDOSIMP como dependiente , $\lambda = 0.0219$. Así pues, el conocimiento de los valores de la variable PARIDOSIMP permite reducir la incertidumbre en la predicción del comportamiento de la variable INESTADO en un 2.19 %, pero como el 0 pertenece al intervalo de confianza, se puede concluir que la asociación no es significativa usando esta medida. Por el contrario, usando la medida t de Goodman y Kruskal, la asociación sí que es evidente y en este caso, el conocimiento de los valores de la variable PARIDOSIMP permite reducir la incertidumbre en la predicción del comportamiento de la variable INESTADO en un 18.96 %.

Tomando la variable INESTADO como dependiente, ambas medidas son significativas. $\lambda = 0.1074$ y $t = 0.0574$ así, el conocimiento de los valores de la variable INESTADO permite reducir la incertidumbre en la predicción del comportamiento de la variable PARIDOSIMP en un 10.74 % y un 5,74 % respectivamente.

\ \

Es de destacar que las medidas obtenidas no describen del todo bien el comportamiento del gráfico, ya que se observa gran relación entre estas variables. Por tanto, se da un paso más y se agrupan los partidos en los bloques de izquierda (PSOE y Unidas Podemos) y derecha (PP, Vox y Ciudadanos), para así calcular de nuevo dichas medidas:

Tabla 5.16: Medidas de variables PARTIDOSIMP e INESTADO (2)

	Estadístico	Límite Inferior IC	Límite Superior IC
$\lambda_{PARTIDOSIMP}$ de Goodman y Krukal	0.0000	0.0000	0.0000
$\lambda_{INESTADO}$ de Goodman y Krukal	0.3358	0.2823	0.3894
$t_{PARTIDOSIMP}$ de Goodman y Kruskal	0.1788	0.1371	0.2205
$t_{INESTADO}$ de Goodman y Kruskal	0.1788	0.1371	0.2205

En este caso, que $\lambda_{PARTIDOSIMP}$ sea 0 no es porque no exista asociación, se debe a la problemática ya vista anteriormente de que la moda de los dos bloques ideológicos coinciden que son la misma categoría de la variable INESTADO (Sí).

En las demás medidas el grado de asociación sí que ha aumentado considerablemente.

5.3. Medidas de Asociación para Variables Ordinales

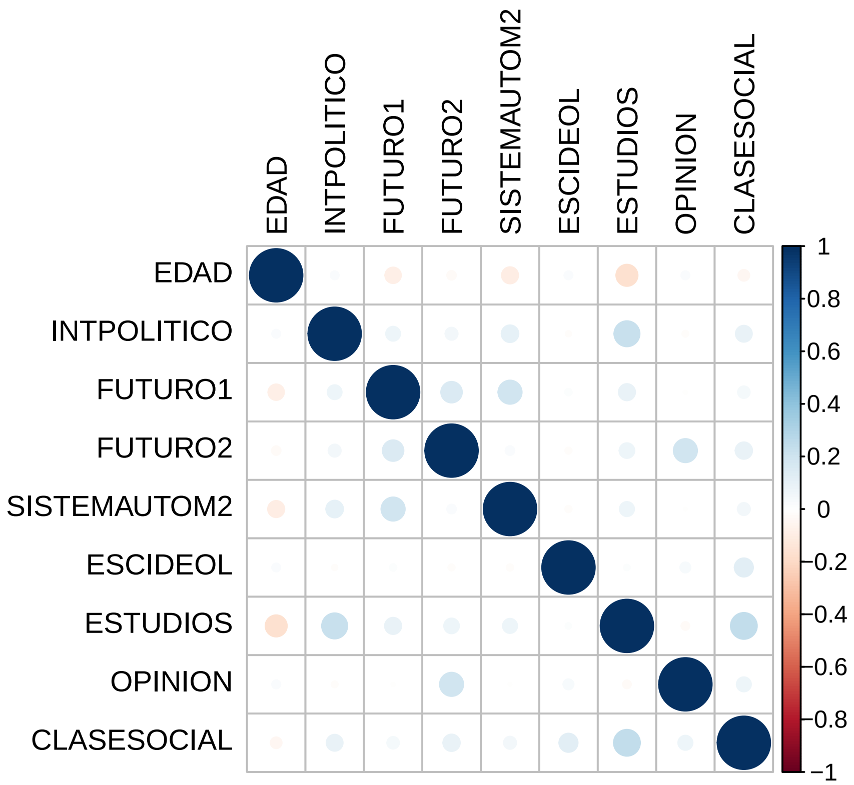

Figura 5.8: Gráfico de correlaciones método Kendall

5.3.1. EDAD - INTPOLITICO

Tabla 5.17: Variables EDAD e INTPOLITICO

EDAD	INTPOLITICO	Total	Porcentaje
18-24	1-Nada	11	0.0791
18-24	2-Poco	30	0.2158
18-24	3-Regular	49	0.3525
18-24	4-Bastante	33	0.2374
18-24	5-Mucho	16	0.1151
25-34	0-N.S/N.C	1	0.0031
25-34	1-Nada	32	0.0985
25-34	2-Poco	45	0.1385
25-34	3-Regular	80	0.2462
25-34	4-Bastante	131	0.4031
25-34	5-Mucho	36	0.1108
35-44	0-N.S/N.C	1	0.0020
35-44	1-Nada	29	0.0572
35-44	2-Poco	75	0.1479
35-44	3-Regular	138	0.2722
35-44	4-Bastante	192	0.3787
35-44	5-Mucho	72	0.1420

Tabla 5.17: Variables EDAD e INTPOLITICO (continúa)

EDAD	INTPOLITICO	Total	Porcentaje
45-54	0-N.S/N.C	1	0.0015
45-54	1-Nada	61	0.0915
45-54	2-Poco	75	0.1124
45-54	3-Regular	180	0.2699
45-54	4-Bastante	246	0.3688
45-54	5-Mucho	104	0.1559
55-64	0-N.S/N.C	1	0.0019
55-64	1-Nada	48	0.0892
55-64	2-Poco	70	0.1301
55-64	3-Regular	129	0.2398
55-64	4-Bastante	204	0.3792
55-64	5-Mucho	86	0.1599
65-99	0-N.S/N.C	4	0.0078
65-99	1-Nada	58	0.1126
65-99	2-Poco	54	0.1049
65-99	3-Regular	133	0.2583
65-99	4-Bastante	191	0.3709
65-99	5-Mucho	75	0.1456

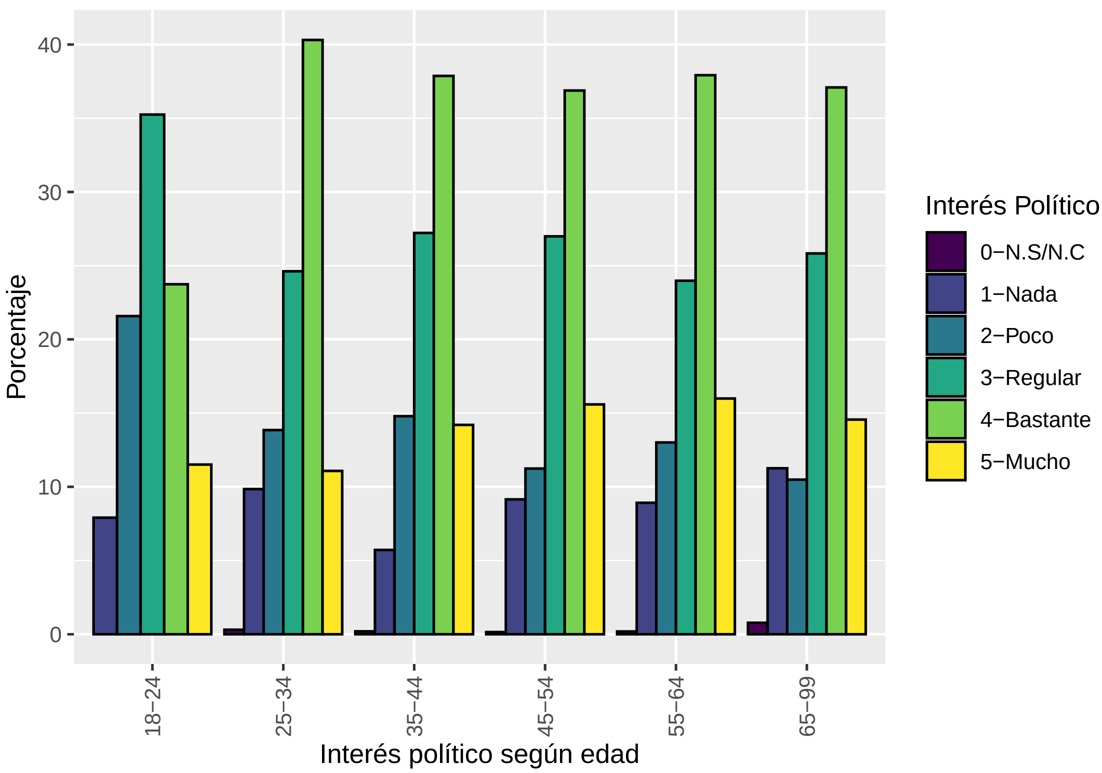

Figura 5.9: Gráfico de barras EDAD-INTPOLITICO

5.3.2. ESTUDIOS - INTPOLITICO

Tabla 5.18: Medidas de variables EDAD e INTPOLITICO

	Estadístico	Límite Inferior IC	Límite Superior IC
τ_a de Kendall	0.0203	0.0167	0.0240
τ_b de Kendall	0.0261	-0.0045	0.0567
τ_c de Stuart	0.0254	-0.0044	0.0551
γ de Goodman y Kruskal	0.0336	-0.0058	0.0730
d_{EDAD} de Sommers	0.0251	-0.0043	0.0545
$d_{INTPOLITICO}$ de Sommers	0.0272	-0.0048	0.0592

Tabla 5.19: Variables ESTUDIOS e INTPOLITICO

ESTUDIOS	**INTPOLITICO**	**Total**	**Porcentaje**
0-N.C	1-Nada	1	0.5000
0-N.C	2-Poco	1	0.5000
1-Sin estudios	0-N.S/N.C	2	0.0357
1-Sin estudios	1-Nada	15	0.2679
1-Sin estudios	2-Poco	10	0.1786
1-Sin estudios	3-Regular	15	0.2679
1-Sin estudios	4-Bastante	11	0.1964
1-Sin estudios	5-Mucho	3	0.0536
2-Primaria	0-N.S/N.C	2	0.0167
2-Primaria	1-Nada	25	0.2083
2-Primaria	2-Poco	23	0.1917
2-Primaria	3-Regular	41	0.3417
2-Primaria	4-Bastante	22	0.1833
2-Primaria	5-Mucho	7	0.0583
3-Secundaria 1ª etapa	0-N.S/N.C	2	0.0057
3-Secundaria 1ª etapa	1-Nada	65	0.1862
3-Secundaria 1ª etapa	2-Poco	60	0.1719
3-Secundaria 1ª etapa	3-Regular	104	0.2980
3-Secundaria 1ª etapa	4-Bastante	90	0.2579
3-Secundaria 1ª etapa	5-Mucho	28	0.0802
4-Secundaria 2ª etapa	1-Nada	30	0.0714
4-Secundaria 2ª etapa	2-Poco	44	0.1048
4-Secundaria 2ª etapa	3-Regular	132	0.3143
4-Secundaria 2ª etapa	4-Bastante	159	0.3786
4-Secundaria 2ª etapa	5-Mucho	55	0.1310
5-F.P. y otros no superiores	0-N.S/N.C	1	0.0019
5-F.P. y otros no superiores	1-Nada	48	0.0889
5-F.P. y otros no superiores	2-Poco	104	0.1926
5-F.P. y otros no superiores	3-Regular	158	0.2926
5-F.P. y otros no superiores	4-Bastante	177	0.3278
5-F.P. y otros no superiores	5-Mucho	52	0.0963
6-Superiores	0-N.S/N.C	1	0.0008

Tabla 5.19: Variables ESTUDIOS e INTPOLITICO (continúa)

ESTUDIOS	INTPOLITICO	Total	Porcentaje
6-Superiores	1-Nada	55	0.0457
6-Superiores	2-Poco	107	0.0889
6-Superiores	3-Regular	259	0.2151
6-Superiores	4-Bastante	538	0.4468
6-Superiores	5-Mucho	244	0.2027

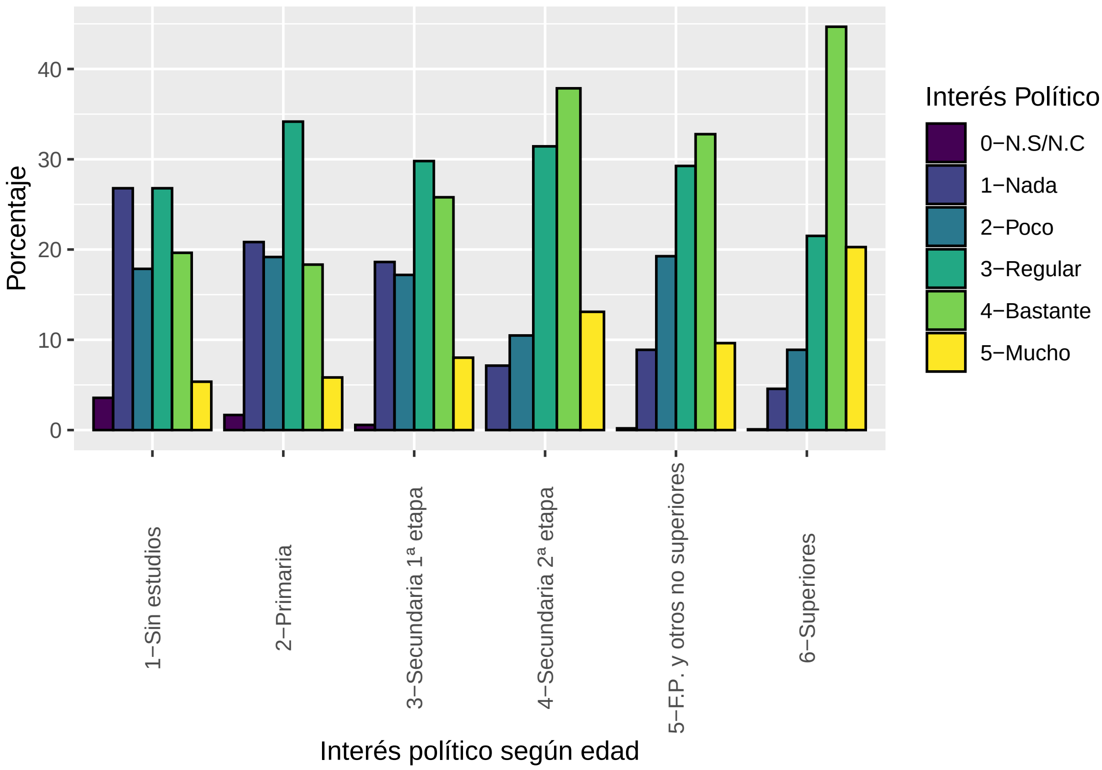

Figura 5.10: Gráfico de barras ESTUDIOS-INTPOLITICO

Tabla 5.20: Medidas de variables ESTUDIOS e INTPOLITICO

	Estadístico	Límite Inferior IC	Límite Superior IC
τ_a de Kendall	0.0219	0.0186	0.0251
τ_b de Kendall	0.0294	-0.0040	0.0627
τ_c de Stuart	0.0291	-0.0039	0.0622
γ de Goodman y Kruskal	0.0394	-0.0053	0.0842
$d_{ESTUDIOS}$ de Sommers	0.0276	-0.0037	0.0590
$d_{INTPOLITICO}$ de Sommers	0.0312	-0.0043	0.0667

5.3.3. OPINION - FUTURO2

Tabla 5.21: Variables OPINION y FUTURO2

FUTURO2	OPINION	Total	Porcentaje
0-N.S/N.C	0-N.S/Duda/N.C	34	0.0761
0-N.S/N.C	1-Peor	233	0.5213
0-N.S/N.C	2-Igual	81	0.1812
0-N.S/N.C	3-Mejor	99	0.2215
1-Muy negativos	0-N.S/Duda/N.C	1	0.0081
1-Muy negativos	1-Peor	110	0.8943
1-Muy negativos	2-Igual	4	0.0325
1-Muy negativos	3-Mejor	8	0.0650
2-Negativos	0-N.S/Duda/N.C	11	0.0186
2-Negativos	1-Peor	427	0.7213
2-Negativos	2-Igual	73	0.1233
2-Negativos	3-Mejor	81	0.1368
3-Indiferentes	0-N.S/Duda/N.C	9	0.0308
3-Indiferentes	1-Peor	156	0.5342
3-Indiferentes	2-Igual	58	0.1986
3-Indiferentes	3-Mejor	69	0.2363
4-Positivos	0-N.S/Duda/N.C	33	0.0278
4-Positivos	1-Peor	498	0.4199
4-Positivos	2-Igual	230	0.1939
4-Positivos	3-Mejor	425	0.3583
5-Muy positivos	1-Peor	21	0.4118
5-Muy positivos	2-Igual	5	0.0980
5-Muy positivos	3-Mejor	25	0.4902

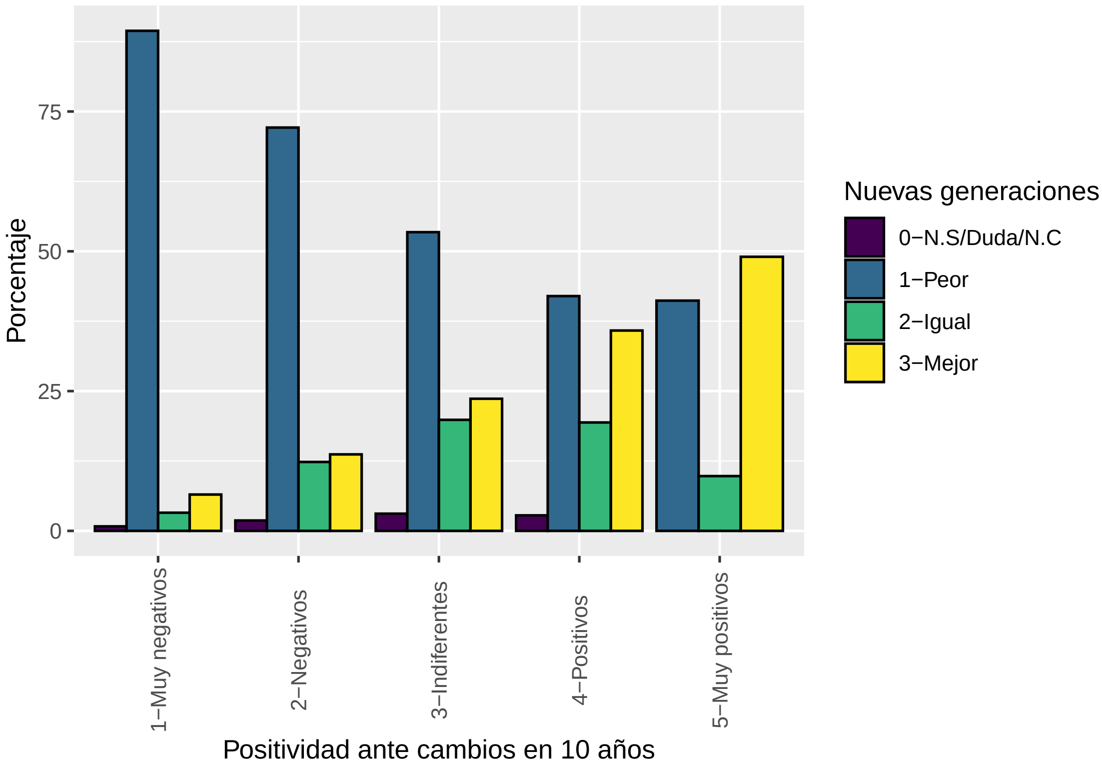

Figura 5.11: Gráfico de barras OPINION-FUTURO2

Tabla 5.22: Medidas de variables OPINION y FUTURO2

	Estadístico	Límite Inferior IC	Límite Superior IC
τ_a de Kendall	0.0203	0.0167	0.0240
τ_b de Kendall	0.0261	-0.0045	0.0567
τ_c de Stuart	0.0254	-0.0044	0.0551
γ de Goodman y Kruskal	0.0336	-0.0058	0.0730
d_{EDAD} de Sommers	0.0251	-0.0043	0.0545
$d_{INTPOLITICO}$ de Sommers	0.0272	-0.0048	0.0592

5.3.4. CLASESOCIAL - ESTUDIOS

Tabla 5.23: Variables CLASE SOCIAL y ESTUDIOS

ESTUDIOS	CLASESOCIAL	Total	Porcentaje
0-N.C	0-N.S/N.C	1	0.5000
0-N.C	2-Clase trabajadora/obrera/proletariado	1	0.5000
1-Sin estudios	0-N.S/N.C	18	0.3214
1-Sin estudios	1-Clase baja/pobre	8	0.1429
1-Sin estudios	2-Clase trabajadora/obrera/proletariado	6	0.1071

Tabla 5.23: Variables CLASE SOCIAL y ESTUDIOS (continúa)

ESTUDIOS	CLASESOCIAL	Total	Porcentaje
1-Sin estudios	3-Clase media-baja	10	0.1786
1-Sin estudios	4-Clase media-media	14	0.2500
2-Primaria	0-N.S/N.C	18	0.1500
2-Primaria	1-Clase baja/pobre	24	0.2000
2-Primaria	2-Clase trabajado-ra/obrera/proletariado	21	0.1750
2-Primaria	3-Clase media-baja	20	0.1667
2-Primaria	4-Clase media-media	36	0.3000
2-Primaria	5-Clase alta y media alta	1	0.0083
3-Secundaria 1ª etapa	0-N.S/N.C	48	0.1375
3-Secundaria 1ª etapa	1-Clase baja/pobre	47	0.1347
3-Secundaria 1ª etapa	2-Clase trabajado-ra/obrera/proletariado	42	0.1203
3-Secundaria 1ª etapa	3-Clase media-baja	63	0.1805
3-Secundaria 1ª etapa	4-Clase media-media	142	0.4069
3-Secundaria 1ª etapa	5-Clase alta y media alta	7	0.0201
4-Secundaria 2ª etapa	0-N.S/N.C	30	0.0714
4-Secundaria 2ª etapa	1-Clase baja/pobre	23	0.0548
4-Secundaria 2ª etapa	2-Clase trabajado-ra/obrera/proletariado	39	0.0929
4-Secundaria 2ª etapa	3-Clase media-baja	70	0.1667
4-Secundaria 2ª etapa	4-Clase media-media	234	0.5571
4-Secundaria 2ª etapa	5-Clase alta y media alta	24	0.0571
5-F.P. y otros no superiores	0-N.S/N.C	37	0.0685
5-F.P. y otros no superiores	1-Clase baja/pobre	28	0.0519
5-F.P. y otros no superiores	2-Clase trabajado-ra/obrera/proletariado	70	0.1296
5-F.P. y otros no superiores	3-Clase media-baja	95	0.1759
5-F.P. y otros no superiores	4-Clase media-media	292	0.5407
5-F.P. y otros no superiores	5-Clase alta y media alta	18	0.0333
6-Superiores	0-N.S/N.C	73	0.0606
6-Superiores	1-Clase baja/pobre	31	0.0257
6-Superiores	2-Clase trabajado-ra/obrera/proletariado	88	0.0731
6-Superiores	3-Clase media-baja	125	0.1038
6-Superiores	4-Clase media-media	742	0.6163
6-Superiores	5-Clase alta y media alta	145	0.1204

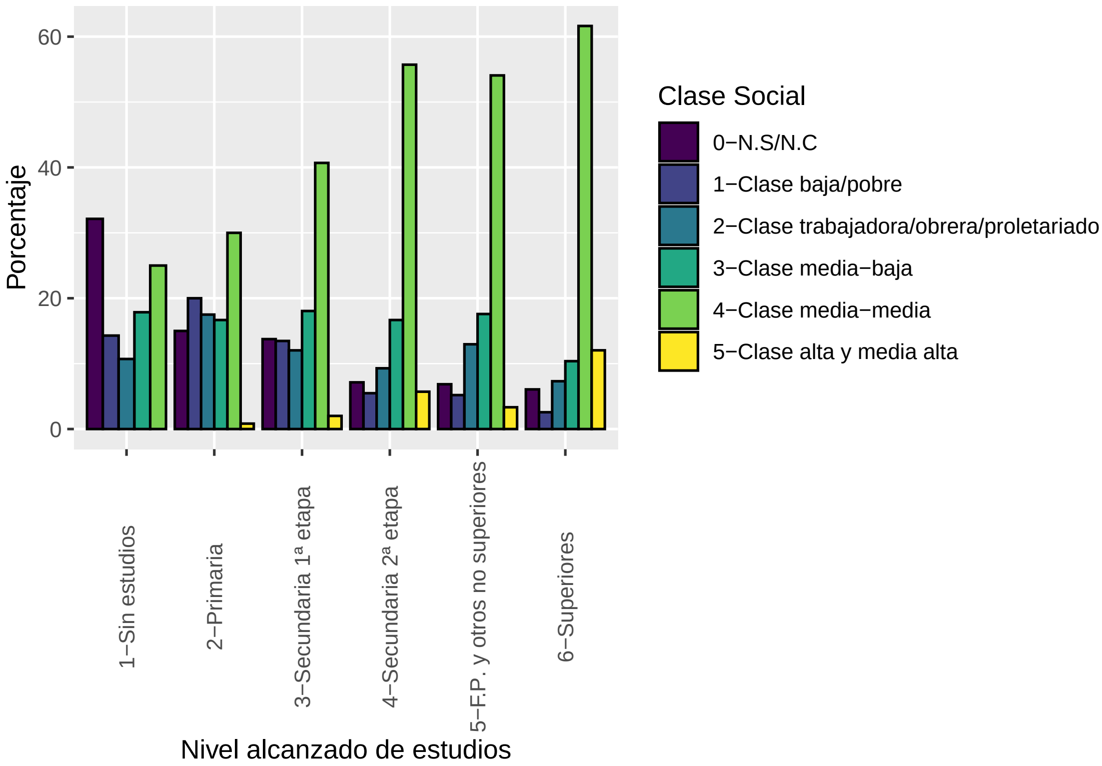

Figura 5.12: Gráfico de barras CLASE SOCIAL-ESTUDIOS

Tabla 5.24: Medidas de variables CLASESOCIAL y ESTUDIOS

	Estadístico	Límite Inferior IC	Límite Superior IC
τ_a de Kendall	0.1582	0.1537	0.1626
τ_b de Kendall	0.2423	0.2099	0.2748
τ_c de Stuart	0.1976	0.1705	0.2248
γ de Goodman y Kruskal	0.3603	0.3145	0.4061
$d_{CLASESOCIAL}$ de Sommers	0.2622	0.2273	0.2971
$d_{ESTUDIOS}$ de Sommers	0.2240	0.1914	0.2565

5.3.5. ESCIDEOL - EDAD

Tabla 5.25: Variables EDAD y ESCIDEOL

ESCIDEOL	EDAD	Total	Porcentaje
0-N.S./N.C.	18-24	10	0.0549
0-N.S./N.C.	25-34	17	0.0934
0-N.S./N.C.	35-44	30	0.1648
0-N.S./N.C.	45-54	48	0.2637
0-N.S./N.C.	55-64	33	0.1813
0-N.S./N.C.	65-99	44	0.2418

Tabla 5.25: Variables EDAD y ESCIDEOL (continúa)

ESCIDEOL	EDAD	Total	Porcentaje
01 Izquierda	18-24	8	0.0374
01 Izquierda	25-34	26	0.1215
01 Izquierda	35-44	40	0.1869
01 Izquierda	45-54	48	0.2243
01 Izquierda	55-64	51	0.2383
01 Izquierda	65-99	41	0.1916
02	18-24	8	0.0447
02	25-34	30	0.1676
02	35-44	36	0.2011
02	45-54	32	0.1788
02	55-64	42	0.2346
02	65-99	31	0.1732
03	18-24	24	0.0625
03	25-34	47	0.1224
03	35-44	96	0.2500
03	45-54	89	0.2318
03	55-64	72	0.1875
03	65-99	56	0.1458
04	18-24	19	0.0635
04	25-34	52	0.1739
04	35-44	49	0.1639
04	45-54	70	0.2341
04	55-64	59	0.1973
04	65-99	50	0.1672
05	18-24	32	0.0462
05	25-34	71	0.1026
05	35-44	123	0.1777
05	45-54	185	0.2673
05	55-64	146	0.2110
05	65-99	135	0.1951
06	18-24	12	0.0435
06	25-34	32	0.1159
06	35-44	57	0.2065
06	45-54	76	0.2754
06	55-64	55	0.1993
06	65-99	44	0.1594
07	18-24	10	0.0472
07	25-34	27	0.1274
07	35-44	39	0.1840
07	45-54	54	0.2547
07	55-64	37	0.1745
07	65-99	45	0.2123
08	18-24	9	0.0647
08	25-34	14	0.1007

Tabla 5.25: Variables EDAD y ESCIDEOL (continúa)

ESCIDEOL	EDAD	Total	Porcentaje
08	35-44	20	0.1439
08	45-54	39	0.2806
08	55-64	23	0.1655
08	65-99	34	0.2446
09	18-24	4	0.1379
09	25-34	1	0.0345
09	35-44	5	0.1724
09	45-54	6	0.2069
09	55-64	4	0.1379
09	65-99	9	0.3103
10 Derecha	18-24	3	0.0353
10 Derecha	25-34	8	0.0941
10 Derecha	35-44	12	0.1412
10 Derecha	45-54	20	0.2353
10 Derecha	55-64	16	0.1882
10 Derecha	65-99	26	0.3059

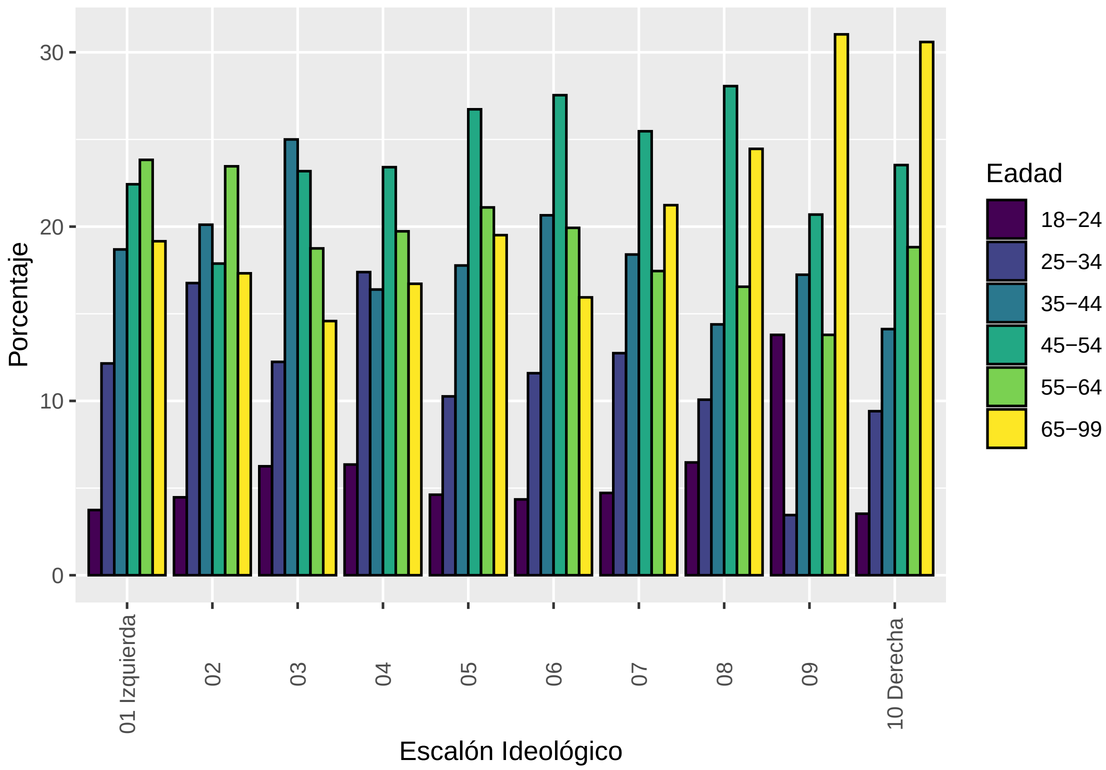

Figura 5.13: Gráfico de barras EDAD-ESCIDEOL

5.3.6. FUTURO1 - FUTURO2

Tabla 5.26: Medidas de variables ESCIDEOL y EDAD

	Estadístico	Límite Inferior IC	Límite Superior IC
τ_a de Kendall	0.0324	0.0266	0.0381
τ_b de Kendall	0.0390	0.0081	0.0699
τ_c de Stuart	0.0388	0.0081	0.0696
γ de Goodman y Kruskal	0.0470	0.0098	0.0841
$d_{ESCIDEOL}$ de Sommers	0.0381	0.0079	0.0682
d_{EDAD} de Sommers	0.0400	0.0083	0.0716

Tabla 5.27: Variables FUTURO1 y FUTURO2

FUTURO1	**FUTURO2**	**Total**	**Porcentaje**
0-N.S/N.C	0-N.S/N.C	94	0.6667
0-N.S/N.C	1-Muy negativos	1	0.0071
0-N.S/N.C	2-Negativos	7	0.0496
0-N.S/N.C	3-Indiferentes	16	0.1135
0-N.S/N.C	4-Positivos	23	0.1631
1-Ningún o casi ningún cambio	0-N.S/N.C	149	1.0000
2-Pocos cambios	0-N.S/N.C	31	0.0609
2-Pocos cambios	1-Muy negativos	10	0.0196
2-Pocos cambios	2-Negativos	96	0.1886
2-Pocos cambios	3-Indiferentes	136	0.2672
2-Pocos cambios	4-Positivos	234	0.4597
2-Pocos cambios	5-Muy positivos	2	0.0039
4-Bastantes cambios	0-N.S/N.C	92	0.0915
4-Bastantes cambios	1-Muy negativos	38	0.0378
4-Bastantes cambios	2-Negativos	234	0.2328
4-Bastantes cambios	3-Indiferentes	74	0.0736
4-Bastantes cambios	4-Positivos	550	0.5473
4-Bastantes cambios	5-Muy positivos	17	0.0169
5-Muchos cambios	0-N.S/N.C	81	0.0913
5-Muchos cambios	1-Muy negativos	74	0.0834
5-Muchos cambios	2-Negativos	255	0.2875
5-Muchos cambios	3-Indiferentes	66	0.0744
5-Muchos cambios	4-Positivos	379	0.4273
5-Muchos cambios	5-Muy positivos	32	0.0361

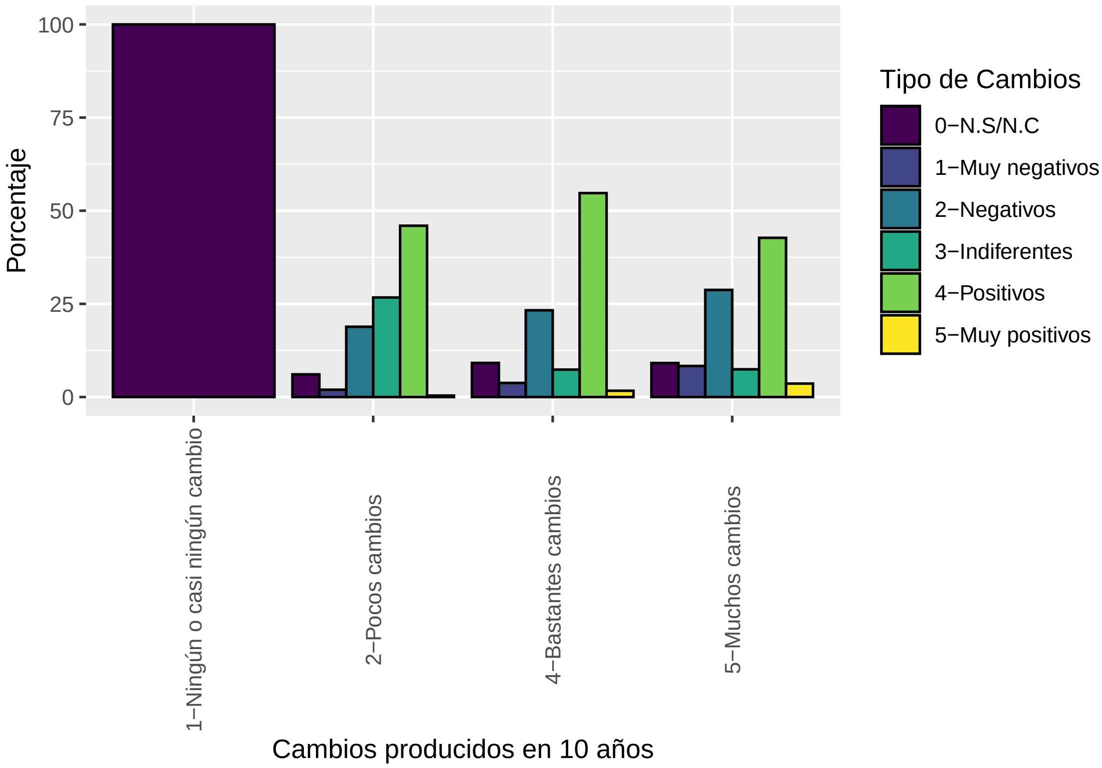

Figura 5.14: Gráfico de barras FUTURO1-FUTURO2

Tabla 5.28: Medidas de variables FUTURO1 y FUTURO2

	Estadístico	Límite Inferior IC	Límite Superior IC
τ_a de Kendall	-0.0356	-0.0388	-0.0325
τ_b de Kendall	-0.0559	-0.0925	-0.0193
τ_c de Stuart	-0.0475	-0.0786	-0.0164
γ de Goodman y Kruskal	-0.0863	-0.1427	-0.0299
$d_{FUTURO1}$ de Sommers	-0.0552	-0.0914	-0.0190
$d_{FUTURO2}$ de Sommers	-0.0566	-0.0944	-0.0188

5.3.7. SISTEMAUTOM2 - FUTURO1

Tabla 5.29: Variables SISTEMAUTOM2 y FUTURO1

FUTURO1	SISTEMAUTOM2	Total	Porcentaje
0-N.S/N.C	0-N.S/N.C	16	0.1134752
0-N.S/N.C	1-No aumentarán nada o casi nada	5	0.0354610
0-N.S/N.C	2-Aumentarán poco	8	0.0567376
0-N.S/N.C	3-Aumentarán algo	21	0.1489362
0-N.S/N.C	4-Aumentarán bastante	52	0.3687943
0-N.S/N.C	5-Aumentarán muchos	39	0.2765957

Tabla 5.29: Variables SISTEMAUTOM2 y FUTURO1 (continúa)

FUTURO1	SISTEMAUTOM2	Total	Porcentaje
1-Ningún o casi ningún cambio	0-N.S/N.C	3	0.0201342
1-Ningún o casi ningún cambio	1-No aumentarán nada o casi nada	9	0.0604027
1-Ningún o casi ningún cambio	2-Aumentarán poco	8	0.0536913
1-Ningún o casi ningún cambio	3-Aumentarán algo	21	0.1409396
1-Ningún o casi ningún cambio	4-Aumentarán bastante	65	0.4362416
1-Ningún o casi ningún cambio	5-Aumentarán muchos	43	0.2885906
2-Pocos cambios	0-N.S/N.C	1	0.0019646
2-Pocos cambios	1-No aumentarán nada o casi nada	6	0.0117878
2-Pocos cambios	2-Aumentarán poco	38	0.0746562
2-Pocos cambios	3-Aumentarán algo	87	0.1709234
2-Pocos cambios	4-Aumentarán bastante	231	0.4538310
2-Pocos cambios	5-Aumentarán muchos	146	0.2868369
4-Bastantes cambios	0-N.S/N.C	8	0.0079602
4-Bastantes cambios	1-No aumentarán nada o casi nada	15	0.0149254
4-Bastantes cambios	2-Aumentarán poco	40	0.0398010
4-Bastantes cambios	3-Aumentarán algo	102	0.1014925
4-Bastantes cambios	4-Aumentarán bastante	487	0.4845771
4-Bastantes cambios	5-Aumentarán muchos	353	0.3512438
5-Muchos cambios	0-N.S/N.C	6	0.0067644
5-Muchos cambios	1-No aumentarán nada o casi nada	6	0.0067644
5-Muchos cambios	2-Aumentarán poco	26	0.0293123
5-Muchos cambios	3-Aumentarán algo	84	0.0947012
5-Muchos cambios	4-Aumentarán bastante	283	0.3190530
5-Muchos cambios	5-Aumentarán muchos	482	0.5434047

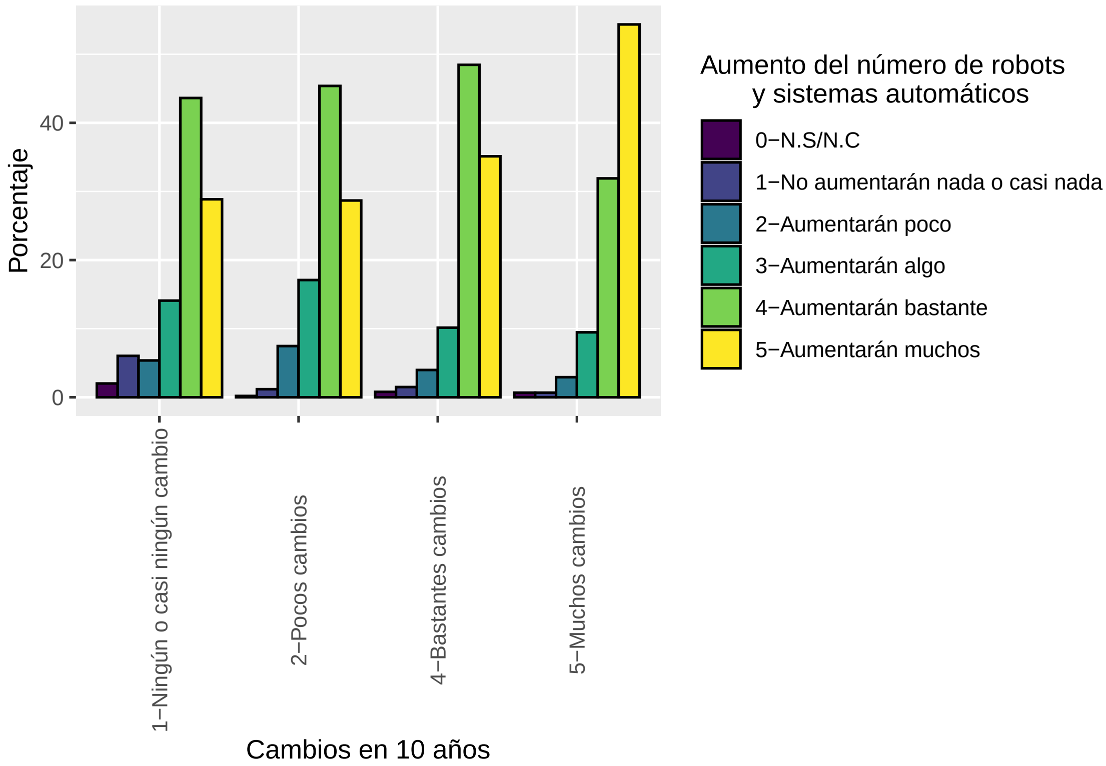

Figura 5.15: Gráfico de barras FUTURO1-SISTEMAUTOM2

Tabla 5.30: Medidas de variables SISTEMAUTOM2 y FUTURO1

	Estadístico	Límite Inferior IC	Límite Superior IC
τ_a de Kendall	0.1255	0.1218	0.1292
τ_b de Kendall	0.1895	0.1552	0.2239
τ_c de Stuart	0.1672	0.1368	0.1976
γ de Goodman y Kruskal	0.2811	0.2313	0.3308
$d_{SISTEMAUTOM2}$ de Sommers	0.1949	0.1595	0.2303
$d_{FUTURO1}$ de Sommers	0.1844	0.1502	0.2185

5.3.8. ESCIDEOL - CLASESOCIAL

Tabla 5.31: Variables ESCIDEL e CLASE SOCIAL

ESCIDEOL	CLASESOCIAL	Total	Porcentaje
0-N.S./N.C.	0-N.S/N.C	37	0.2033
0-N.S./N.C.	1-Clase baja/pobre	20	0.1099
0-N.S./N.C.	2-Clase trabajadora/obrera/proletariado	14	0.0769
0-N.S./N.C.	3-Clase media-baja	24	0.1319
0-N.S./N.C.	4-Clase media-media	78	0.4286
0-N.S./N.C.	5-Clase alta y media alta	9	0.0495

Tabla 5.31: Variables ESCIDEL e CLASE SOCIAL (continúa)

ESCIDEOL	CLASESOCIAL	Total	Porcentaje
01 Izquierda	0-N.S/N.C	16	0.0748
01 Izquierda	1-Clase baja/pobre	27	0.1262
01 Izquierda	2-Clase trabajadora/obrera/proletariado	47	0.2196
01 Izquierda	3-Clase media-baja	39	0.1822
01 Izquierda	4-Clase media-media	79	0.3692
01 Izquierda	5-Clase alta y media alta	6	0.0280
02	0-N.S/N.C	10	0.0559
02	1-Clase baja/pobre	6	0.0335
02	2-Clase trabajadora/obrera/proletariado	39	0.2179
02	3-Clase media-baja	40	0.2235
02	4-Clase media-media	70	0.3911
02	5-Clase alta y media alta	14	0.0782
03	0-N.S/N.C	25	0.0651
03	1-Clase baja/pobre	18	0.0469
03	2-Clase trabajadora/obrera/proletariado	47	0.1224
03	3-Clase media-baja	51	0.1328
03	4-Clase media-media	216	0.5625
03	5-Clase alta y media alta	27	0.0703
04	0-N.S/N.C	17	0.0569
04	1-Clase baja/pobre	13	0.0435
04	2-Clase trabajadora/obrera/proletariado	23	0.0769
04	3-Clase media-baja	44	0.1472
04	4-Clase media-media	166	0.5552
04	5-Clase alta y media alta	36	0.1204
05	0-N.S/N.C	63	0.0910
05	1-Clase baja/pobre	45	0.0650
05	2-Clase trabajadora/obrera/proletariado	63	0.0910
05	3-Clase media-baja	104	0.1503
05	4-Clase media-media	378	0.5462
05	5-Clase alta y media alta	39	0.0564
06	0-N.S/N.C	9	0.0326
06	1-Clase baja/pobre	13	0.0471
06	2-Clase trabajadora/obrera/proletariado	14	0.0507
06	3-Clase media-baja	29	0.1051
06	4-Clase media-media	184	0.6667
06	5-Clase alta y media alta	27	0.0978
07	0-N.S/N.C	20	0.0943
07	1-Clase baja/pobre	6	0.0283
07	2-Clase trabajadora/obrera/proletariado	9	0.0425
07	3-Clase media-baja	22	0.1038
07	4-Clase media-media	136	0.6415
07	5-Clase alta y media alta	19	0.0896
08	0-N.S/N.C	13	0.0935
08	1-Clase baja/pobre	5	0.0360

Tabla 5.31: Variables ESCIDEL e CLASE SOCIAL (continúa)

ESCIDEOL	CLASESOCIAL	Total	Porcentaje
08	2-Clase trabajadora/obrera/proletariado	5	0.0360
08	3-Clase media-baja	17	0.1223
08	4-Clase media-media	91	0.6547
08	5-Clase alta y media alta	8	0.0576
09	0-N.S/N.C	3	0.1034
09	1-Clase baja/pobre	1	0.0345
09	2-Clase trabajadora/obrera/proletariado	1	0.0345
09	3-Clase media-baja	2	0.0690
09	4-Clase media-media	18	0.6207
09	5-Clase alta y media alta	4	0.1379
10 Derecha	0-N.S/N.C	12	0.1412
10 Derecha	1-Clase baja/pobre	7	0.0824
10 Derecha	2-Clase trabajadora/obrera/proletariado	5	0.0588
10 Derecha	3-Clase media-baja	11	0.1294
10 Derecha	4-Clase media-media	44	0.5176
10 Derecha	5-Clase alta y media alta	6	0.0706

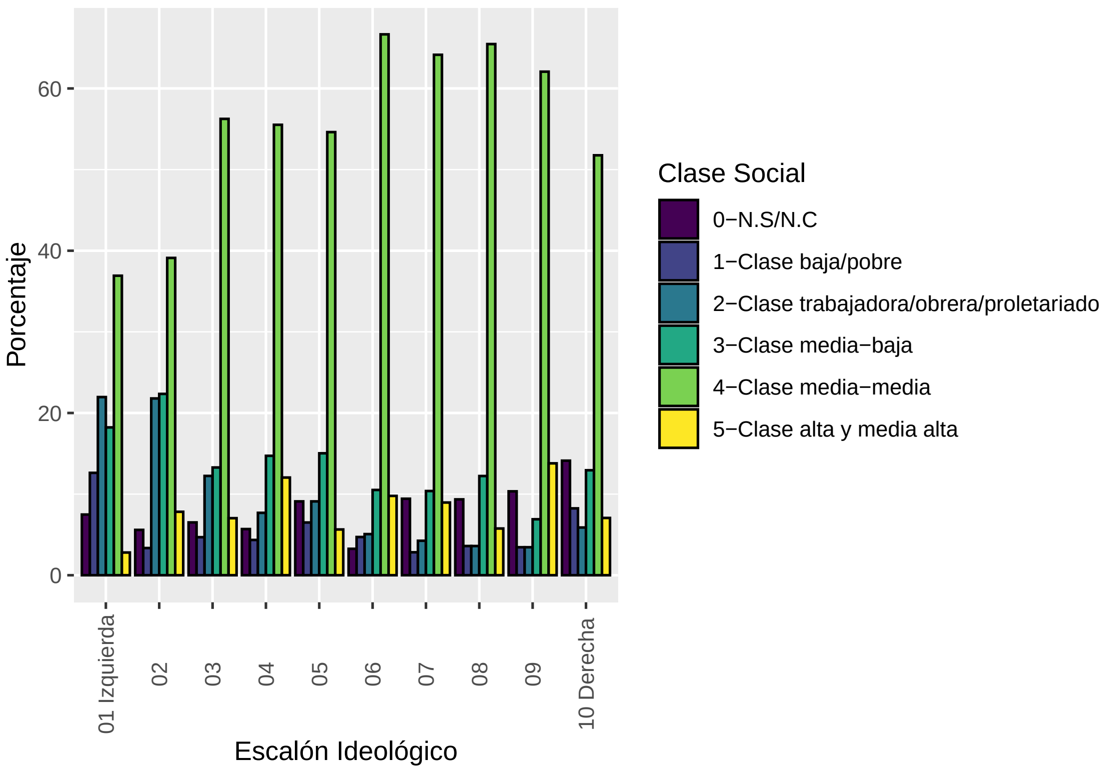

Figura 5.16: Gráfico de barras ESCIDEOL-CLASESOCIAL

Tabla 5.32: Medidas de variables ESCIDEOL y CLASESOCIAL

	Estadístico	Límite Inferior IC	Límite Superior IC
τ_a de Kendall	0.1003	0.0940	0.1065
τ_b de Kendall	0.1388	0.1071	0.1705
τ_c de Stuart	0.1253	0.0965	0.1541
γ de Goodman y Kruskal	0.1915	0.1481	0.2349
$d_{ESCIDEOL}$ de Sommers	0.1159	0.0894	0.1425
$d_{CLASESOCIAL}$ de Sommers	0.1662	0.1277	0.2047

5.4. Medidas de Asociación Nominal (dicotómica) - Ordinal

Aquí va Cureton y S de Withfield si la encuentro

Apéndice A

Encuesta sobre Tendencias Sociales: Cuestionario

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

«Información sujeta a secreto estadístico (Ley 12/89, de 9 de mayo, de la Función Estadística Pública) y al Reglamento General de Protección de Datos y la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales.» Plan Estadístico Nacional 2021-2024. RD 1110/2020, de 15 de diciembre.

«Buenos días/ tardes, soy (nombre propio) y estoy realizando una encuesta para el Centro de Investigaciones Sociológicas que tiene por finalidad estudiar cuestiones de interés público relacionadas con cambios y tendencias sociales. Su teléfono ha sido obtenido al azar. La entrevista será grabada y se borrará en un plazo inferior a un mes. Tiene usted la posibilidad de ejercer sus derechos de acceso, rectificación, supresión y limitación, recogidos en la normativa de Protección de Datos a través de nuestra página web www.cis.es y le garantizamos el absoluto anonimato de sus respuestas de acuerdo con la normativa de secreto estadístico. Si desea ampliar esta información puede consultar nuestra página web. ¿Sería tan amable de contestar a nuestras preguntas? La encuesta dura unos 20 minutos. Muchas gracias.»

PC1. Pregunta contacto 1. ¿Me puede decir a qué provincia y municipio estoy llamando...?

[TIPO_TEL]
FIJO....... 1
MÓVIL 2

[CCAA]
____

[PROVINCIA]
____

[MUNICIPIO]
____

[BESAREA]
____

[ENTREV]
____

ENTREVISTADOR/A: SI LA PERSONA QUE CONTESTA ES DIFERENTE DE LA QUE COGIÓ EL TELÉFONO PRESENTARSE:

Buenos días/tardes, mi nombre es… y le llamo del Centro de Investigaciones Sociológicas porque estamos realizando una encuesta de opinión sobre cambios y tendencias sociales. Dura unos 20 minutos. ¿Sería tan amable de colaborar con nosotros?

[SEXO]
Hombre 1
Mujer 2

[EDADEXACTA]
____

[EDAD]
de 18 a 24....... 1
de 25 a 34....... 2
de 35 a 44....... 3
de 45 a 54....... 4
de 55 a 64....... 5
65 y más....... 6

P.0 En primer lugar quisiera preguntarle si tiene Ud....

Controles:
Si P0=3. Advertir. Lo sentimos pero no está Ud. incluido/a en la población para entrevistar. Muchas gracias.

[P0]
La nacionalidad española 1
La nacionalidad española y otra 2
Otra nacionalidad....... 3

Saltos:
Si P0=3 ir a fin cuestionario.

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

P.1 En términos generales, ¿es Ud. una persona a la que le interesan mucho las cuestiones políticas, le interesan bastante, regular, poco, o no le interesan nada?
[INTPOL]
Mucho 1
Bastante........ 2
Regular 3
Poco 4
Nada 5
N.S........ 8
N.C........ 9

P.2 ¿A qué le parece Ud. que debiera darse más importancia en nuestra sociedad, a favorecer la igualdad y solidaridad entre las personas, o a hacer posible que cada cual llegue lo más alto que pueda con su esfuerzo y su trabajo?
[IMPIDEAS]
A favorecer la igualdad y solidaridad entre las personas 1
A hacer posible que cada cual llegue lo más alto que pueda con su esfuerzo y su trabajo 2
(NO LEER) Un equilibrio entre los dos aspectos 3
(NO LEER) Está en duda 4
N.S........ 8
N.C........ 9

P.3 Hay quienes piensan que el Estado no debe intervenir en la vida económica dejándolo todo en manos de la iniciativa privada. Por el contrario, hay quienes consideran que el Estado sí debe intervenir en la economía. ¿Con cuál de estas posturas está Ud. más de acuerdo?
[INESTADO1]
El Estado no debe intervenir en la vida económica 1
El Estado sí debe intervenir en la economía........ 2
(NO LEER) Depende de las condiciones económicas 3
(NO LEER) Está en duda 4
N.S. 8
N.C. 9

P.3ª ¿Y cómo le parece a Ud. que debe intervenir el Estado? ¿Dirigiendo y planificando toda la actividad económica, o interviniendo sólo en determinados sectores de interés público y fijando orientaciones generales?

Filtros:
Si NO (INESTADO1=2 O INESTADO1=3) ir a la siguiente.

[INESTADO2]
El Estado debe dirigir y planificar toda la actividad económica...... 1
El Estado debe intervenir sólo en determinados sectores de interés público y fijando orientaciones generales 2
(NO LEER) Está en duda 3
Otras posturas (especificar) 96
N.S. 98
N.C. 99

Filtros:
Si NO INESTADO2=(96) ir a la siguiente.

[INESTADO2_COD]
____

P.4 En general, ¿podría valorar de 1 a 10 la confianza que Ud. tiene en estos momentos en cada uno de estas organizaciones políticas e instituciones, entendiendo que el 10 representaría la máxima confianza y el 1 la mínima confianza?
[CONFIANZA]

	1 Mínima	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10 Máxima*	*N.S.*	*N.C.*
En los partidos políticos	1	2	3	4	5	6	7	8	9	10	98	99
En los sindicatos	1	2	3	4	5	6	7	8	9	10	98	99
En el Gobierno de España	1	2	3	4	5	6	7	8	9	10	98	99
En el Parlamento español	1	2	3	4	5	6	7	8	9	10	98	99
En los medios de comunicación social (prensa, radio, televisiones)	1	2	3	4	5	6	7	8	9	10	98	99
En la Justicia	1	2	3	4	5	6	7	8	9	10	98	99
En la Constitución de 1978	1	2	3	4	5	6	7	8	9	10	98	99

P.5 Y hace cinco años, en general, ¿tenía Ud. más, menos o igual confianza que ahora en...?
[CONFIANZA_RETR]

	Más que ahora	*Menos que ahora*	*Igual que ahora*	*(NO LEER) No sabe, duda*	*N.C.*
En los partidos políticos	1	2	3	8	9
En los sindicatos	1	2	3	8	9
En el Gobierno de España	1	2	3	8	9
En el Parlamento español	1	2	3	8	9
En los medios de comunicación social (prensa, radio, televisiones)	1	2	3	8	9
En la Justicia	1	2	3	8	9
En la Constitución de 1978	1	2	3	8	9

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

P.6 Y según sus impresiones, y tal como van las cosas, ¿piensa Ud. que dentro de cinco años tendrá Ud. más, menos o igual confianza que ahora en...?
[CONFIANZA_PROS]

	Más que ahora	*Menos que ahora*	*Igual que ahora*	*(NO LEER) No sabe, duda*	*N.C.*
En los partidos políticos	1	2	3	8	9
En los sindicatos	1	2	3	8	9
En el Gobierno de España	1	2	3	8	9
En el Parlamento español	1	2	3	8	9
En los medios de comunicación social (prensa, radio, televisiones)	1	2	3	8	9
En la Justicia	1	2	3	8	9
En la Constitución de 1978	1	2	3	8	9

P.7 De las principales decisiones políticas que ha tomado usted en los últimos dos años (votar o no votar, manifestar una u otra opinión, etc.), en general, ¿qué le ha influido a usted más, o qué ha tenido más en cuenta en primer lugar en sus opiniones políticas? ¿Y en segundo lugar? (Entrevistador/a: LEER LAS OPCIONES DE RESPUESTA)

En primer lugar
[INFLUPOL1]
Alguna persona concreta (de sus familiares, amigos/as, conocidos/as…) *1*
Un/a candidato/a *2*
La radio *3*
La televisión *4*
Las redes sociales, Internet *5*
Los periódicos *6*
Las revistas *7*
Otra respuesta (especificar) *96*
(NO LEER) Ninguna, nada *97*
N.S. *98*
N.C. *99*

Saltos:
Si INFLUPOL1=97 ir a [INFLUMOV_1] - Percepción respecto al presente de la influencia de organizaciones y movimientos sociales (dentro de cinco años) | Partidos políticos
Si INFLUPOL1=98 ir a [INFLUMOV_1] - Percepción respecto al presente de la influencia de organizaciones y movimientos sociales (dentro de cinco años) | Partidos políticos
Si INFLUPOL1=99 ir a [INFLUMOV_1] - Percepción respecto al presente de la influencia de organizaciones y movimientos sociales (dentro de cinco años) | Partidos políticos

Filtros:
Si NO INFLUPOL1=(96) ir a la siguiente.
[INFLUPOL1_COD]
____

Saltos:
Si INFLUPOL1=97 ir a [INFLUMOV_1] - Percepción respecto al presente de la influencia de organizaciones y movimientos sociales (dentro de cinco años) | Partidos políticos
Si INFLUPOL1=98 ir a [INFLUMOV_1] - Percepción respecto al presente de la influencia de organizaciones y movimientos sociales (dentro de cinco años) | Partidos políticos
Si INFLUPOL1=99 ir a [INFLUMOV_1] - Percepción respecto al presente de la influencia de organizaciones y movimientos sociales (dentro de cinco años) | Partidos políticos

En segundo lugar
[INFLUPOL2]
Alguna persona concreta (de sus familiares, amigos/as, conocidos/as…) *1*
Un/a candidato/a *2*
La radio *3*
La televisión *4*
Las redes sociales, Internet *5*
Los periódicos *6*
Las revistas *7*
Otra respuesta (especificar) *96*
Ninguna más *97*
N.S. *98*
N.C. *99*

Filtros:
Si NO INFLUPOL2=(96) ir a la siguiente.
[INFLUPOL2_COD]
____

P.8 Por lo que Ud. sabe, dentro de cinco años, ¿qué influencia y poder tendrán en la sociedad esta serie de movimientos y organizaciones sociales? ¿Más, menos o igual que ahora?
[INFLUMOV]

	Más que ahora	*Menos que ahora*	*Igual que ahora*	*(NO LEER) No sabe, duda*	*N.C.*
Partidos políticos	1	2	3	8	9
Sindicatos	1	2	3	8	9
Movimientos ecologistas	1	2	3	8	9
Movimientos pacifistas	1	2	3	8	9
Movimientos feministas	1	2	3	8	9
Movimientos antiglobalización y altermundistas (anti liberalismo económico)	1	2	3	8	9
Organizaciones de consumidores	1	2	3	8	9

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

P.9 Como Ud. sabe, hay mucha gente que está hablando del futuro y de los cambios sociales y económicos que se van a producir. ¿Cree Ud. que, de aquí a diez años, en general, habrá muchos cambios, bastantes cambios, pocos cambios o ningún o casi ningún cambio?

[FUTUR1]

Muchos cambios 1
Bastantes cambios 2
(NO LEER) Regular, ni muchos ni pocos 3
Pocos cambios 4
Ningún o casi ningún cambio 5
N.S., duda 8
N.C. 9

P.10 Y, en general, ¿cree Ud. que estos cambios en su conjunto van a ser muy positivos, positivos, negativos, muy negativos o indiferentes?

Filtros:
Si FUTUR1=5 O FUTUR1=8 O FUTUR1=9 ir a la siguiente.

[FUTUR2]

Muy positivos 1
Positivos 2
Negativos 3
Muy negativos 4
Indiferentes 5
(NO LEER) En parte positivos y en parte negativos 6
N.S., duda 8
N.C. 9

P.11 Por lo que Ud. sabe, ¿cree Ud. que dentro de diez años la influencia mundial de la Unión Europea será mayor, igual o menor que ahora?
¿Y la de Estados Unidos?
¿Y la de Japón?
¿Y la de Rusia?
¿Y la de China?
¿Y la de los países árabes?
¿Y la de los países latinoamericanos?

[INFLUPAISES]

	Mayor que ahora	*Menor que ahora*	*Igual que ahora*	*N.S.*	*N.C.*
Unión Europea	1	2	3	8	9
Estados Unidos	1	2	3	8	9
Japón	1	2	3	8	9
Rusia	1	2	3	8	9
China	1	2	3	8	9
Países árabes	1	2	3	8	9
Países latinoamericanos	1	2	3	8	9

P.12 Como Ud. sabe, en algunas empresas se están utilizando robots industriales y sistemas automáticos de trabajo. ¿Cree Ud. que el número de robots y sistemas automáticos en las empresas en los próximos diez años va a aumentar mucho, bastante, algo, poco, nada o casi nada?

[SAT1]

Aumentarán mucho 1
Aumentarán bastante 2
Aumentarán algo 3
Aumentarán poco 4
No aumentarán nada o casi nada 5
N.S., duda 8
N.C. 9

P.13 ¿Le parece a Ud. que la utilización de robots y sistemas automáticos de trabajo en general va a ser causa en los próximos diez años de un aumento del paro, o por el contrario cree que dará lugar a la creación de más puestos de trabajo, o cree que todo seguirá prácticamente igual, que no influirá en el desempleo?

[SAT2]

Dará lugar a un aumento del paro 1
Dará lugar a la creación de más puestos de trabajo 2
Cree que todo seguirá prácticamente igual, que no influirá en el desempleo 3
(NO LEER) Está en duda 4
Otras respuestas (especificar) 96
N.S. 98
N.C. 99

Filtros:
Si NO SAT2=(96) ir a la siguiente.

[SAT2_COD]

____

P.14 De las siguientes innovaciones o adelantos tecnológicos que le voy a mencionar, ¿le parece a Ud. que van a desarrollarse mucho, bastante, regular, poco, nada o casi nada en los próximos diez años?

[DESIN]

	Mucho	*Bastante*	*Regular*	*Poco*	*Nada o casi nada*	*(NO LEER) No sabe, duda*	*N.C.*
Los robots industriales y los sistemas automáticos de trabajo	1	2	3	4	5	8	9
La biotecnología	1	2	3	4	5	8	9
Las nuevas fuentes de energía (p. ej. solar, eólica)	1	2	3	4	5	8	9
La microelectrónica, los ordenadores	1	2	3	4	5	8	9
La ingeniería genética	1	2	3	4	5	8	9
Las nuevas tecnologías del ocio y del esparcimiento (p. ej. los videojuegos)	1	2	3	4	5	8	9
Las nuevas tecnologías del hogar	1	2	3	4	5	8	9
Los transportes	1	2	3	4	5	8	9

P.15 ¿Cuál le parece a Ud. que es el objeto (artefacto, aparato, máquina) que identifica y simboliza en mayor grado la época actual en primer lugar?, ¿y en segundo lugar? (Entrevistador/a: NO LEER, RESPUESTA ESPONTÁNEA. ATENCIÓN: SI LA PERSONA SE REFIERE A CUESTIONES INMATERIALES, VOLVER A LEER "OBJETO").

En primer lugar

[SIMBOL1]

La máquina fabril 1
La televisión 2
El coche 3
La radio 4
El ordenador 5
El robot 6
El avión 7
El teléfono móvil 8
Internet 9
Los trenes de alta velocidad 10
Los satélites espaciales 11
Otro (especificar) 96
Ninguno 97
N.S. 98
N.C. 99

Filtros:
Si NO SIMBOL1=(96) ir a la siguiente.

[SIMBOL1_COD]

____

En segundo lugar

Filtros:
Si SIMBOL1=97 ir a la siguiente.
Si SIMBOL1=98 ir a la siguiente.
Si SIMBOL1=99 ir a la siguiente.

[SIMBOL2]

La máquina fabril 1
La televisión 2
El coche 3
La radio 4
El ordenador 5
El robot 6
El avión 7
El teléfono móvil 8
Internet 9
Los trenes de alta velocidad 10
Los satélites espaciales 11
Otro (especificar) 96
Ninguno más 97
N.S. 98
N.C. 99

Filtros:
Si NO SIMBOL2=(96) ir a la siguiente.

[SIMBOL2_COD]

____

P.16 ¿En cuál o cuáles de los siguientes campos le gustaría a Ud. que hubiera más avances científicos y tecnológicos en los próximos años? (Entrevistador/a: LEER LAS OPCIONES DE RESPUESTA) (MÁXIMO TRES)

[CIENCIA1_1]

La tecnología de la información 1
La ingeniería genética 2
La exploración del espacio 3
La biotecnología 4
La energía solar 5
Internet 6
Las telecomunicaciones 7
La fecundación in vitro 8
La energía nuclear 9
Los trasplantes de órganos 10
Otra respuesta (especificar) 96
(NO LEER) Ninguno 97
N.S. 98
N.C. 99

Saltos:
Si CIENCIA1_1=97 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)
Si CIENCIA1_1=98 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)
Si CIENCIA1_1=99 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)

Filtros:
Si NO CIENCIA1_1=(96) ir a la siguiente.

[CIENCIA1_1_COD]

____

Saltos:
Si CIENCIA1_1=97 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)
Si CIENCIA1_1=98 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)
Si CIENCIA1_1=99 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)

[CIENCIA1_2]

La tecnología de la información 1
La ingeniería genética 2
La exploración del espacio 3
La biotecnología 4
La energía solar 5
Internet 6
Las telecomunicaciones 7
La fecundación in vitro 8
La energía nuclear 9
Los trasplantes de órganos 10
Otra respuesta (especificar) 96
(NO LEER) Ninguno más 97
N.S. 98
N.C. 99

Saltos:
Si CIENCIA1_2=97 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)
Si CIENCIA1_2=98 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)
Si CIENCIA1_2=99 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

Filtros:
Si NO CIENCIA1_2=(96) ir a la siguiente.
[CIENCIA1_2_COD]

Saltos:
Si CIENCIA1_2=97 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)
Si CIENCIA1_2=98 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)
Si CIENCIA1_2=99 ir a [CIENCIA2_1] - Preferencias en la limitación del desarrollo científico y tecnológico en los próximos años (1)

[CIENCIA1_3]
La tecnología de la información............ 1
La ingeniería genética............ 2
La exploración del espacio............ 3
La biotecnología............ 4
La energía solar............ 5
Internet............ 6
Las telecomunicaciones............ 7
La fecundación in vitro............ 8
La energía nuclear............ 9
Los trasplantes de órganos............ 10
Otra respuesta (especificar)............ 96
(NO LEER) Ninguno más............ 97
N.S.............. 98
N.C.............. 99

Filtros:
Si NO CIENCIA1_3=(96) ir a la siguiente.
[CIENCIA1_3_COD]

P.17 ¿Y en cuál o cuáles de estos campos cree que habría que limitar el desarrollo científico y tecnológico? (Entrevistador/a: LEER OTRA VEZ SI ES NECESARIO) (MÁXIMO TRES RESPUESTAS)
[CIENCIA2_1]
La tecnología de la información............ 1
La ingeniería genética............ 2
La exploración del espacio............ 3
La biotecnología............ 4
La energía solar............ 5
Internet............ 6
Las telecomunicaciones............ 7
La fecundación in vitro............ 8
La energía nuclear............ 9
Los trasplantes de órganos............ 10
Otra respuesta (especificar)............ 96
(NO LEER) Ninguno............ 97
N.S.............. 98
N.C.............. 99

Saltos:
Si CIENCIA2_1=97 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)
Si CIENCIA2_1=98 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)
Si CIENCIA2_1=99 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)

Filtros:
Si NO CIENCIA2_1=(96) ir a la siguiente.
[CIENCIA2_1_COD]

Saltos:
Si CIENCIA2_1=97 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)
Si CIENCIA2_1=98 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)
Si CIENCIA2_1=99 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)

[CIENCIA2_2]
La tecnología de la información............ 1
La ingeniería genética............ 2
La exploración del espacio............ 3
La biotecnología............ 4
La energía solar............ 5
Internet............ 6
Las telecomunicaciones............ 7
La fecundación in vitro............ 8
La energía nuclear............ 9
Los trasplantes de órganos............ 10
Otra respuesta (especificar)............ 96
(NO LEER) Ninguno más............ 97
N.S.............. 98
N.C.............. 99

Saltos:
Si CIENCIA2_2=97 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)
Si CIENCIA2_2=98 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)
Si CIENCIA2_2=99 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)

Filtros:
Si NO CIENCIA2_2=(96) ir a la siguiente.
[CIENCIA2_2_COD]

Saltos:
Si CIENCIA2_2=97 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)
Si CIENCIA2_2=98 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)
Si CIENCIA2_2=99 ir a [PROBLMUN_1] - Principales problemas mundiales dentro de diez años (1)

[CIENCIA2_3]
La tecnología de la información............ 1
La ingeniería genética............ 2
La exploración del espacio............ 3
La biotecnología............ 4
La energía solar............ 5
Internet............ 6
Las telecomunicaciones............ 7
La fecundación in vitro............ 8
La energía nuclear............ 9
Los trasplantes de órganos............ 10
Otra respuesta (especificar)............ 96
(NO LEER) Ninguno más............ 97
N.S.............. 98
N.C.............. 99

Filtros:
Si NO CIENCIA2_3=(96) ir a la siguiente.
[CIENCIA2_3_COD]

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

P.18 ¿Cuáles cree Ud. que serán los tres principales problemas del mundo dentro de diez años? (Entrevistador/a: NO LEER, RESPUESTA ESPONTÁNEA) (MÁXIMO TRES RESPUESTAS)

Problema 1
Anotar si no está en la lista y pulsar "intro"
[PROBLMUN_1]

El paro, la falta de trabajo *1*
Las guerras *2*
El hambre, la falta de alimentos *3*
La inmigración *4*
El terrorismo *5*
La delincuencia y la inseguridad ciudadana *6*
La pobreza, desigualdades y problemas sociales *7*
La vivienda *8*
La contaminación *9*
Las drogas *10*
La destrucción de la naturaleza *11*
El agua, las sequías, la desertificación *12*
Las enfermedades víricas *13*
La superpoblación *14*
La crisis económica *15*
La violencia *16*
La falta de solidaridad *17*
El racismo *18*
La crisis de valores *19*
La deshumanización *20*
La falta de recursos naturales *21*
Los residuos nucleares *22*
La crisis política *23*
El armamento nuclear *24*
La crisis religiosa *25*
Los nacionalismos *26*
La corrupción *27*
Las carencias del Estado de bienestar (sanidad, educación) *28*
El funcionamiento de la democracia *29*
Las epidemias *30*
El cambio climático, el calentamiento global *31*
Otro *96*
Ninguno *97*
N.S. *98*
N.C. *99*

Saltos:
Si PROBLMUN_1=97 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas
Si PROBLMUN_1=98 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas
Si PROBLMUN_1=99 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas

Filtros:
Si NO PROBLMUN_1=(96) ir a la siguiente.
[PROBLMUN_1_COD]

Saltos:
Si PROBLMUN_1=97 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas
Si PROBLMUN_1=98 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas
Si PROBLMUN_1=99 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas

Problema 2
Anotar si no está en la lista y pulsar "intro"
[PROBLMUN_2]

El paro, la falta de trabajo *1*
Las guerras *2*
El hambre, la falta de alimentos *3*
La inmigración *4*
El terrorismo *5*
La delincuencia y la inseguridad ciudadana *6*
La pobreza, desigualdades y problemas sociales *7*
La vivienda *8*
La contaminación *9*
Las drogas *10*
La destrucción de la naturaleza *11*
El agua, las sequías, la desertificación *12*
Las enfermedades víricas *13*
La superpoblación *14*
La crisis económica *15*
La violencia *16*
La falta de solidaridad *17*
El racismo *18*
La crisis de valores *19*
La deshumanización *20*
La falta de recursos naturales *21*
Los residuos nucleares *22*
La crisis política *23*
El armamento nuclear *24*
La crisis religiosa *25*
Los nacionalismos *26*
La corrupción *27*
Las carencias del Estado de bienestar (sanidad, educación) *28*
El funcionamiento de la democracia *29*
Las epidemias *30*
El cambio climático, el calentamiento global *31*
Otro *96*
Ninguno más *97*
N.S. *98*
N.C. *99*

Saltos:
Si PROBLMUN_2=97 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas
Si PROBLMUN_2=98 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas
Si PROBLMUN_2=99 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

Filtros:
Si NO PROBLMUN_2=(96) ir a la siguiente.
[PROBLMUN_2_COD]

____

Saltos:
Si PROBLMUN_2=97 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas
Si PROBLMUN_2=98 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas
Si PROBLMUN_2=99 ir a [IDEASPROS1_1] - Percepción respecto al presente del aumento de las creencias religiosas y otras cuestiones sociales en nuestro país (dentro de diez años) | Creencias religiosas

Problema 3
Anotar si no está en la lista y pulsar "intro"
[PROBLMUN_3]

El paro, la falta de trabajo *1*
Las guerras *2*
El hambre, la falta de alimentos *3*
La inmigración *4*
El terrorismo *5*
La delincuencia y la inseguridad ciudadana *6*
La pobreza, desigualdades y problemas sociales *7*
La vivienda *8*
La contaminación *9*
Las drogas *10*
La destrucción de la naturaleza *11*
El agua, las sequías, la desertificación *12*
Las enfermedades víricas *13*
La superpoblación *14*
La crisis económica *15*
La violencia *16*
La falta de solidaridad *17*
El racismo *18*
La crisis de valores *19*
La deshumanización *20*
La falta de recursos naturales *21*
Los residuos nucleares *22*
La crisis política *23*
El armamento nuclear *24*
La crisis religiosa *25*
Los nacionalismos *26*
La corrupción *27*
Las carencias del Estado de bienestar (sanidad, educación) *28*
El funcionamiento de la democracia *29*
Las epidemias *30*
El cambio climático, el calentamiento global *31*
Otro *96*
Ninguno más *97*
N.S. *98*
N.C. *99*

Filtros:
Si NO PROBLMUN_3=(96) ir a la siguiente.
[PROBLMUN_3_COD]

____

P.19 Por otra parte, a unos diez años vista, ¿cree Ud. que en un país como el nuestro habrá más creencias religiosas que ahora, menos que ahora o igual que ahora...? (Entrevistador/a: PREGUNTAR DE LA MISMA FORMA PARA TODAS LAS FRASES)
[IDEASPROS1]

	Más que ahora	*Menos que ahora*	*Igual que ahora*	*N.S.*	*N.C.*
Creencias religiosas	1	2	3	8	9
Interés en las cosas materiales (propiedades)	1	2	3	8	9
Interés por el propio éxito	1	2	3	8	9
Libertad sexual	1	2	3	8	9
Violencia	1	2	3	8	9
Mujeres en puestos de responsabilidad	1	2	3	8	9
Integración de las personas con discapacidad	1	2	3	8	9

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

P.20 Y siguiendo con otros temas, dentro de diez años, ¿cree Ud. que habrá más relación con la familia o lazos familiares que ahora, menos que ahora o igual que ahora...? (Entrevistador/a: PREGUNTAR DE LA MISMA FORMA PARA TODAS LAS FRASES)
[IDEASPROS2]

	Más que ahora	*Menos que ahora*	*Igual que ahora*	*N.S.*	*N.C.*
Relación con la familia o lazos familiares	1	2	3	8	9
Separaciones y divorcios	1	2	3	8	9
Nacimientos, natalidad	1	2	3	8	9
Soledad o aislamiento	1	2	3	8	9
Cuidados familiares a los mayores	1	2	3	8	9
Tiempo libre para el ocio	1	2	3	8	9
Personas realizadas o que disfruten con su trabajo	1	2	3	8	9
Interés por la salud y el cuidado personal	1	2	3	8	9

P.21 Y ¿cree Ud. que habrá más paro o falta de trabajo que ahora, menos que ahora o igual que ahora...? (Entrevistador/a: PREGUNTAR DE LA MISMA FORMA PARA TODAS LAS FRASES)
[IDEASPROS3]

	Más que ahora	*Menos que ahora*	*Igual que ahora*	*N.S.*	*N.C.*
Paro o falta de trabajo	1	2	3	8	9
Diferencias sociales y económicas	1	2	3	8	9
Consumo de drogas	1	2	3	8	9
SIDA y otras enfermedades víricas	1	2	3	8	9
Delincuencia	1	2	3	8	9
Protección social del Estado	1	2	3	8	9
Igualdad de oportunidades	1	2	3	8	9

P.22 Y para terminar con estas preguntas sobre España, dentro de diez años, ¿cree Ud. que habrá más deterioro del medio ambiente que ahora, menos que ahora o igual que ahora..? (Entrevistador/a: PREGUNTAR DE LA MISMA FORMA PARA TODAS LAS FRASES)
[IDEASPROS4]

	Más que ahora	*Menos que ahora*	*Igual que ahora*	*N.S.*	*N.C.*
Deterioro del medio ambiente	1	2	3	8	9
Muertes por cáncer	1	2	3	8	9
Robots que sustituyen a trabajadores	1	2	3	8	9
Racismo y xenofobia	1	2	3	8	9
Personas teletrabajando en casa	1	2	3	8	9
Epidemias graves como la de COVID-19	1	2	3	8	9

P.23 Pasando a hablar de cuestiones mundiales, dentro de diez años, ¿cree Ud. que las desigualdades entre los países pobres y los países ricos, en general, serán mayores, menores o iguales que ahora? (Ent.: SEÑALA LA RESPUESTA Y PASA A LA SIGUIENTE PREGUNTA. HAZ LO MISMO PARA LAS SIGUIENTES)
¿Y cree que habrá más, menos o igual número de guerras y conflictos internacionales que ahora?
¿Y cree que habrá más, menos o igual número de personas que emigrarán desde los países pobres hacia los países ricos que ahora?
¿Cree Ud. que dentro de diez años la ONU tendrá un papel internacional más activo y eficaz, menos, o igual que ahora?
[IDEASPROS5]

	Mayores/Más que ahora	*Menores/Menos que ahora*	*Igual que ahora*	*N.S.*	*N.C.*
Desigualdades entre países pobres y ricos	1	2	3	8	9
Guerras y conflictos internacionales	1	2	3	8	9
Personas que emigren de los países pobres a los ricos	1	2	3	8	9
Eficacia de la ONU (Organización de Naciones Unidas)	1	2	3	8	9

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

P.24 De los siguientes grupos de personas que le voy a leer, ¿me puede decir con cuál se identifica Ud. más en primer lugar, es decir, con cuál piensa Ud. que tiene más intereses comunes? ¿Y en segundo lugar? (Entrevistador/a: LEER LAS OPCIONES DE RESPUESTA)

En primer lugar

[IDENTIFIC1]

- *Con las del mismo género* *1*
- *Con las de la misma edad y generación* *2*
- *Con las de la misma región o nacionalidad* *3*
- *Con las de la misma clase social* *4*
- *Con las del mismo municipio* *5*
- *Con las que tienen sus mismas ideas políticas* *6*
- *Con las que tienen sus mismas aficiones, gustos, modas* *7*
- *Con las que tienen sus mismas ideas religiosas* *8*
- *Con las que tienen su misma profesión y trabajo* *9*
- *(NO LEER) Con todos/as* *10*
- *Otras (especificar)* *96*
- *(NO LEER) Con ningún grupo en concreto* *97*
- *N.S.* *98*
- *N.C.* *99*

Filtros:
Si NO IDENTIFIC1=(96) ir a la siguiente.

[IDENTIFIC1_COD]

____

En segundo lugar

Filtros:
Si IDENTIFIC1=97 ir a la siguiente.
Si IDENTIFIC1=98 ir a la siguiente.
Si IDENTIFIC1=99 ir a la siguiente.

[IDENTIFIC2]

- *Con las del mismo género* *1*
- *Con las de la misma edad y generación* *2*
- *Con las de la misma región o nacionalidad* *3*
- *Con las de la misma clase social* *4*
- *Con las del mismo municipio* *5*
- *Con las que tienen sus mismas ideas políticas* *6*
- *Con las que tienen sus mismas aficiones, gustos, modas* *7*
- *Con las que tienen sus mismas ideas religiosas* *8*
- *Con las que tienen su misma profesión y trabajo* *9*
- *Otras (especificar)* *96*
- *Ninguno más* *97*
- *N.S.* *98*
- *N.C.* *99*

Filtros:
Si NO IDENTIFIC2=(96) ir a la siguiente.

[IDENTIFIC2_COD]

____

P.25 Cuando se habla de política se utilizan normalmente las expresiones izquierda y derecha. Situándonos en una escala de 10 casillas, como un termómetro, que van del 1 al 10, en la que 1 significa "lo más a la izquierda" y 10 "lo más a la derecha", ¿en qué casilla se colocaría Ud.?

[ESCIDEOL]

- *1 Izda.* *1*
- *2* *2*
- *3* *3*
- *4* *4*
- *5* *5*
- *6* *6*
- *7* *7*
- *8* *8*
- *9* *9*
- *10 Dcha.* *10*
- *N.S.* *98*
- *N.C.* *99*

P.26 ¿Me podría decir si en las elecciones generales del 10 de noviembre de 2019...? (Entrevistador/a: LEER LAS OPCIONES DE RESPUESTA).

Controles:
Si PARTICIPACIONG=(2) Y (EDADEXACTA>20). Advertir. Según la edad registrada, la persona entrevistada podía votar.
Si PARTICIPACIONG=(1;7) Y (EDADEXACTA<20). Advertir. El 10 de noviembre de 2019, ¿Ud. tenía más de 18 años?, para rectificar la respuesta marcar "Cancelar"

[PARTICIPACIONG]

- *Fue a votar y votó* *1*
- *Votó por correo* *7*
- *No tenía edad para votar* *2*
- *Fue a votar pero no pudo hacerlo* *3*
- *No fue a votar porque no pudo* *4*
- *Prefirió no votar* *5*
- *No tenía derecho a voto* *6*
- *No recuerda* *8*
- *N.C.* *9*

P.26a ¿Y podría decirme a qué partido o coalición votó (para el Congreso)? (Entrevistador/a: NO LEER, RESPUESTA ESPONTÁNEA)

Filtros:
Si NO PARTICIPACIONG=(1;7) ir a la siguiente.
Controles:
Si SCODCOMUNID=(9) Y RECUVOTOG=(21). Advertir. En Cataluña UP es En Comú Podem, para rectificar la respuesta, marcar "CANCELAR"
Si SCODCOMUNID=(15) Y RECUVOTOG=(1). Advertir. En Navarra PP es UPN (Navarra Suma), para rectificar la respuesta, marcar "CANCELAR"
Si SCODCOMUNID=(15) Y RECUVOTOG=(4). Advertir. En Navarra Cs es UPN (Navarra Suma), para rectificar la respuesta, marcar "CANCELAR"
Si SCODCOMUNID=(12) Y RECUVOTOG=(21). Advertir. En Galicia UP es En Común-Unidas Podemos, para rectificar la respuesta, marcar "CANCELAR"
Si SCODCOMUNID=(12) Y RECUVOTOG=(6). Advertir. En Galicia es En Común-Unidas Podemos, para rectificar la respuesta, marcar "CANCELAR"

[RECUVOTOG]

- *PSOE* *2*
- *PP* *1*
- *VOX* *18*
- *Unidas Podemos* *21*
- *En Comú Podem* *6*
- *En Común - Unidas Podemos* *67*
- *Ciudadanos* *4*
- *Más País* *50*
- *ERC* *8*
- *JxCat* *9*
- *CUP* *19*
- *EAJ-PNV* *11*
- *EH Bildu* *12*
- *CCa-PNC-NC* *13*
- *Navarra Suma (UPN)* *14*
- *Més Compromís* *7*
- *BNG (Bloque Nacionalista Galego)* *24*
- *PRC (Partido Regionalista de Cantabria)* *43*
- *Teruel Existe* *68*
- *PACMA (Partido Animalista)* *17*
- *Otros partidos* *95*
- *En blanco* *96*
- *Voto nulo* *77*
- *No recuerda* *98*
- *N.C.* *99*

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

P.27 ¿Y con qué partido simpatiza más? (Entrevistador/a: NO LEER, RESPUESTA ESPONTÁNEA)

Controles:
Si SCODCOMUNID=(9) Y PARTSIMP=(21). Advertir. En Cataluña UP es En Comú Podem, para rectificar la respuesta, marcar "CANCELAR"
Si SCODCOMUNID=(15) Y PARTSIMP=(1). Advertir. En Navarra PP es UPN (Navarra Suma), para rectificar la respuesta, marcar "CANCELAR"
Si SCODCOMUNID=(15) Y PARTSIMP=(4). Advertir. En Navarra Cs es UPN (Navarra Suma), para rectificar la respuesta, marcar "CANCELAR"
Si SCODCOMUNID=(12) Y PARTSIMP=(21). Advertir. En Galicia UP es En Común-Unidas Podemos, para rectificar la respuesta, marcar "CANCELAR"
Si SCODCOMUNID=(12) Y PARTSIMP=(6). Advertir. En Galicia es En Común-Unidas Podemos, para rectificar la respuesta, marcar "CANCELAR"

[PARTSIMP]

PSOE ... 2
PP ... 1
VOX ... 18
Unidas Podemos ... 21
En Comú Podem ... 6
En Común - Unidas Podemos ... 67
Ciudadanos ... 4
Más País ... 50
ERC ... 8
JxCat ... 9
CUP ... 19
EAJ-PNV ... 11
EH Bildu ... 12
CCa-PNC-NC ... 13
Navarra Suma (UPN) ... 14
Més Compromís ... 7
BNG (Bloque Nacionalista Galego) ... 24
PRC (Partido Regionalista de Cantabria) ... 43
Teruel Existe ... 68
PACMA (Partido Animalista) ... 17
Otro ... 95
Ninguno ... 97
N.C. ... 99

P.28 ¿Ha ido Ud. a la escuela o cursado algún tipo de estudios? (Entrevistador/a: en caso negativo, preguntar si sabe leer y escribir).

Controles:
Si EDADEXACTA<41 Y ESCUELA<3. Advertir. ENTREVISTADOR/A: Asegúrate de que la persona entrevistada no haya ido a la escuela NUNCA, ya que es extraño que una persona menor de 40 años no haya ido a la escuela

[ESCUELA]

No, no sabe leer y escribir ... 1
No, pero sabe leer y escribir ... 2
Sí, ha ido a la escuela ... 3
N.C. ... 9

Saltos:
Si NO ESCUELA=3 ir a [SITLAB] - Situación laboral de la persona entrevistada

P.28a ¿Cuáles son los estudios de más alto nivel oficial que Ud. ha cursado (con independencia de que los haya terminado o no)? Por favor, especifique lo más posible, diciéndome el curso en que estaba cuando los terminó (o los interrumpió) y también el nombre que tenían entonces esos estudios (ej: 3 años de estudios primarios, primaria, 5º de bachillerato, Maestría Industrial, preuniversitario, 4º de EGB, licenciatura, doctorado, FP1, etc.). (ENTREVISTADOR/A: si aún está estudiando, anotar el último curso que haya completado y el ciclo correcto en las opciones de respuesta. Si no ha completado la primaria, anotar nº de años que asistió a la escuela, diferenciando entre menos de 5 y más de 5).

[CURSOENTREV]

CURSO ________ ***N.S. - N.R. = 98*** ***N.C. = 99***

[NOMBREESTENTREV]

NOMBRE DE ESTUDIOS ________ ***N.S. - N.R. = 98*** ***N.C. = 99***

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

Controles:
Si NIVELESTENTREV=10 Y EDADEXACTA<20. Advertir. ¿Seguro que es el nivel de estudios cursado? Para rectificar, CANCELAR. Para confirmar, ACEPTAR.
Si NIVELESTENTREV=11 Y EDADEXACTA<20. Advertir. ¿Seguro que es el nivel de estudios cursado? Para rectificar, CANCELAR. Para confirmar, ACEPTAR.
Si NIVELESTENTREV=12 Y EDADEXACTA<20. Advertir. ¿Seguro que es el nivel de estudios cursado? Para rectificar, CANCELAR. Para confirmar, ACEPTAR.
Si NIVELESTENTREV=13 Y EDADEXACTA<20. Advertir. ¿Seguro que es el nivel de estudios cursado? Para rectificar, CANCELAR. Para confirmar, ACEPTAR.
Si NIVELESTENTREV=14 Y EDADEXACTA<20. Advertir. ¿Seguro que es el nivel de estudios cursado? Para rectificar, CANCELAR. Para confirmar, ACEPTAR.

[NIVELESTENTREV]

01. Menos de 5 años de escolarización o primaria incompleta... 1
02. Educación primaria (Educación primaria de LOGSE, 5º Curso de EGB, Enseñanza primaria antigua) 2
03. Cualificación profesional grado inicial (FP grado inicial). PCPI (Programas de Cualificación Profesional Inicial, que no precisan de titulación académica de la primera etapa de secundaria para su realización). Programas de garantía social.. 3
04. Educación secundaria (ESO, EGB. Graduado Escolar. Certificado de Escolaridad, Bachillerato Elemental) 4
05. FP de grado medio (Ciclo/módulo formativo de FP (grado medio), de Artes Plásticas y Diseño, Música y danza, Enseñanzas deportivas, FP I, Bachiller laboral elemental. Oficialía Industrial; Bachillerato Comercial). 5
06. Bachillerato (Bachillerato LOGSE, BUP, Bachillerato superior (6º), Bachillerato universitario (7º), Incluidos COU y PREU)... 6
07. FP de grado superior (Ciclo/módulo formativo de FP (grado superior) de Artes Plásticas, Diseño, Música y danza, Deporte, FP II, Bach. Laboral Sup., Maestría industrial, Perito Mercantil; Secretariado de 2º grado; Grado Medio conservatorio).. 7
08. Arquitectura-ingeniería técnica (Arquitectura/ingeniería técnica, Aparejador; Peritos).. 8
09. Diplomatura (ATENCIÓN: solo Diplomaturas oficiales, no codificar aquí los tres primeros años de una licenciatura o grado con mayor duración). 9
10. Grado (Estudios de grado, Enseñanzas Artísticas equivalentes (desde 2006))... 10
11. Licenciatura (Titulaciones con equivalencia oficial: 2º ciclo INEF; Danza y arte dramático (desde 1992); Grado superior de música) 11
12. Arquitectura/ingeniería 12
13. Máster oficial universitario (Especialidades médicas o equivalente).. 13
14. Doctorado.. 14
15. Títulos propios de posgrado (máster no oficial, etc.)............... 15
16. Otros estudios 16
N.S./No recuerda.. 98
N.C.. 99

P.29 ¿En qué situación laboral se encuentra Ud. actualmente? (ENTREVISTADOR/A: LEER OPCIONES DE RESPUESTA).

[SITLAB]

Trabaja 1
Jubilado/a o pensionista (anteriormente ha trabajado) 2
Pensionista (anteriormente no ha trabajado) 3
En paro y ha trabajado antes 4
En paro y busca su primer empleo... 5
Estudiante 6
Trabajo doméstico no remunerado 7
Otra situación 8
N.C. 9

P.30 ¿Me puede decir cuál es su ocupación actual? (ENTREVISTADOR/A: LEER OPCIONES DE RESPUESTA).

Filtros:
Si NO SITLAB=1 ir a la siguiente.

[CNO11]

Directores/as y gerentes.. 1
Profesionales y científicos/as e intelectuales 2
Técnicos/as y profesionales de nivel medio.................................... 3
Personal de apoyo administrativo 4
Trabajadores/as de los servicios y vendedores/as de comercios y mercados... 5
Agricultores/as y trabajadores/as cualificados/as agropecuarios/as, forestales y pesqueros/as.................................. 6
Oficiales/as, operarios/as y artesanos/as de artes mecánicas y de otros oficios 7
Operadores/as de instalaciones y máquinas y ensambladores/as 8
Ocupaciones elementales 9
Ocupaciones militares y cuerpos policiales.................................... 10
Otra/o 11
N.C. 99

P.31 ¿Cómo se define Ud. en materia religiosa: católico/a practicante, católico/a no practicante, creyente de otra religión, agnóstico/a, indiferente o no creyente, o ateo/a?

[RELIGION]

Católico/a practicante 1
Católico/a no practicante... 2
Creyente de otra religión 3
Agnóstico/a (no niegan la existencia de Dios pero tampoco la descartan) 4
Indiferente, no creyente.. 5
Ateo/a (niegan la existencia de Dios) 6
N.C. 9

P.32 Para terminar, en líneas generales, ¿cree Ud. que las nuevas generaciones vivirán mejor, peor o igual de lo que se ha vivido hasta ahora?

[OPINIION]

Mejor 1
Peor.. 2
Igual 3
N.S., duda 8
N.C. 9

Estudio: TENDENCIAS SOCIALES
Clave: ECIS3343

P.33 ¿A qué clase social diría Ud. que pertenece?
(Entrevistador/a: NO LEER, RESPUESTA ESPONTÁNEA)
[CLASESOCIAL]
Clase alta ***1***
Clase media-alta ***2***
Clase media-media ***3***
Clase media-baja ***4***
Clase trabajadora/obrera ***5***
Clase baja ***12***
Clase pobre ***6***
Infraclase ***7***
Proletariado ***8***
A los/as de abajo ***9***
Excluidos/as ***10***
A la gente común ***11***
Otra (especificar) ***96***
No cree en las clases ***97***
No sabe, duda ***98***
N.C. ***99***

Filtros:
Si NO CLASESOCIAL=(96) ir a la siguiente.
[CLASESOCIAL_COD]
____

FIN DE LA ENTREVISTA.

MUCHAS GRACIAS POR SU AMABILIDAD Y POR EL TIEMPO QUE NOS HA DEDICADO.

Apéndice B

Código R

B.1. Preparación de datos

```
library(readr)
X3343 <- read_csv2("MD3343/3343.csv")

# se quitan variables que no interesan
datos = X3343[,c(6,11:14,16,39,48,49,57,58,67,110,112,117,120,121,125,
                 126,128)]

# se agrupan las variables nominales:

for (i in 1:length(datos$IDENTIFIC1)){
if (datos$IDENTIFIC1[i]=="Con las de la misma clase social"){
    datos$IDENTIFIC1[i]="Clase Social"}
else {if (datos$IDENTIFIC1[i]=="Con las de la misma edad y generación") {
      datos$IDENTIFIC1[i]="Generación"}
      else{if (datos$IDENTIFIC1[i]=="Con las de la misma región o nacionalidad" ){
      datos$IDENTIFIC1[i]="Religión/Nacionalidad"}
      else {if (datos$IDENTIFIC1[i]=="Con las del mismo género") {
      datos$IDENTIFIC1[i]="Género"}
      else {if (datos$IDENTIFIC1[i]=="Con las del mismo municipio") {
      datos$IDENTIFIC1[i]="Municipio"}
      else {if (datos$IDENTIFIC1[i]=="Con las que tienen su misma profesión y trabajo
      datos$IDENTIFIC1[i]="Profesión/Trabajo"}
      else {if (datos$IDENTIFIC1[i]=="Con las que tienen su misma profesión y trabajo
      datos$IDENTIFIC1[i]="Profesión/Trabajo"}
      else {if (datos$IDENTIFIC1[i]=="Con las que tienen sus mismas aficiones, gustos
      datos$IDENTIFIC1[i]="Aficiones y Gustos"}
      else {if (datos$IDENTIFIC1[i]=="Con las que tienen sus mismas ideas políticas")
      datos$IDENTIFIC1[i]="Ideas Políticas"}
      else {if (datos$IDENTIFIC1[i]=="Con las que tienen sus mismas ideas religiosas"
      datos$IDENTIFIC1[i]="Ideas Religiosas"}
        else{datos$IDENTIFIC1[i]="N.S/N.C"}
```

```
      }}}}}}}}
}}

for (i in 1:length(datos$INESTADO1)){
if (datos$INESTADO1[i]=="El Estado no debe intervenir en la vida económica"){
    datos$INESTADO1[i]="No"}
else {if (datos$INESTADO1[i]=="El Estado sí debe intervenir en la economía" ) {
      datos$INESTADO1[i]="Sí"}
          else{datos$INESTADO1[i]="N.S/N.C"}
}
  }

for (i in 1:length(datos$CCAA)){
if (datos$CCAA[i]=="Ceuta (Ciudad Autónoma de)" ){
    datos$CCAA[i]="Ceuta"}
else {if (datos$CCAA[i]=="Murcia (Región de)") {
      datos$CCAA[i]="Murcia"}
      else{if (datos$CCAA[i]=="Comunitat Valenciana"){
      datos$CCAA[i]="Valencia"}
      else {if (datos$CCAA[i]=="Madrid (Comunidad de)") {
      datos$CCAA[i]="Madrid"}
      else {if (datos$CCAA[i]=="Navarra (Comunidad Foral de)") {
      datos$CCAA[i]="Navarra"}
      else {if (datos$CCAA[i]=="Asturias (Principado de)"){
      datos$CCAA[i]="Asturias"}
      else {if (datos$CCAA[i]=="Melilla (Ciudad Autónoma de)"){
      datos$CCAA[i]="Melilla"}
        else{datos$CCAA[i]=datos$CCAA[i]}
      }}}}}
}}

for (i in 1:length(datos$INFLUPOL1)){
if (datos$INFLUPOL1[i]=="Alguna persona concreta (de sus familiares, amigos/as, conoc
    datos$INFLUPOL1[i]="Conocido Concreto"}
else {if (datos$INFLUPOL1[i]=="La radio") {
      datos$INFLUPOL1[i]="Radio"}
      else{if (datos$INFLUPOL1[i]=="La televisión" ){
      datos$INFLUPOL1[i]="Televisión"}
      else {if (datos$INFLUPOL1[i]=="Las redes sociales, Internet") {
      datos$INFLUPOL1[i]="Internet"}
      else {if (datos$INFLUPOL1[i]=="Los periódicos" |
                datos$INFLUPOL1[i]=="Las revistas" ) {
      datos$INFLUPOL1[i]="Prensa"}
      else {if (datos$INFLUPOL1[i]=="Un/a candidato/a, las actuaciones de los/as cand
      datos$INFLUPOL1[i]="Candidatos"}
      else {if (datos$INFLUPOL1[i]=="Los programas electorales"){
      datos$INFLUPOL1[i]="Prog. Electorales"}
```

```
        else{datos$INFLUPOL1[i]="N.S/N.C"}
      }}}}}
}}

for (i in 1:length(datos$SIMBOL1)){
if (datos$SIMBOL1[i]=="El ordenador, el portátil" |
    datos$SIMBOL1[i]=="El software y la informática"){
    datos$SIMBOL1[i]="Odenador e informática"}
      else {if (datos$SIMBOL1[i]=="El robot aspirador y el robot de cocina" | datos$S
      datos$SIMBOL1[i]="Robot"}
      else {if (datos$SIMBOL1[i]=="Internet, wifi, rúter, fibra, conectividad") {
      datos$SIMBOL1[i]="Internet"}
      else {if (datos$SIMBOL1[i]=="La máquina fabril, la maquinaria industrial y agrí
      datos$SIMBOL1[i]="Máquina fabril"}
      else {if (datos$SIMBOL1[i]=="La radio"){
      datos$SIMBOL1[i]="Radio"}
      else {if (datos$SIMBOL1[i]=="La televisión y la smart TV"){
      datos$SIMBOL1[i]="Televisión"}
      else {if (datos$SIMBOL1[i]=="Los trenes de alta velocidad" |
                datos$SIMBOL1[i]=="Los medios de transporte en general" |
                datos$SIMBOL1[i]=="El coche, la automoción, el coche eléctrico" |
                datos$SIMBOL1[i]=="El avión"){
      datos$SIMBOL1[i]="Medios Transporte"}
      else {if (datos$SIMBOL1[i]=="Los satélites espaciales, las naves, los viajes es
      datos$SIMBOL1[i]="Tema Aeroespacial"}
      else {if (datos$SIMBOL1[i]=="Otros electrodomésticos (freidora, microondas)" |
                datos$SIMBOL1[i]=="Los electrodomésticos para la climatización (aire 
      datos$SIMBOL1[i]="Electrodomésticos"}
        else{datos$SIMBOL1[i]="N.S/N.C"}
      }}}}}}}
}}

for (i in 1:length(datos$SIMPATIA)){
if (datos$SIMPATIA[i]=="BNG (Bloque Nacionalista Galego)"){
    datos$SIMPATIA[i]="BNG"}
else {if (datos$SIMPATIA[i]=="En Comú Podem" |
          datos$SIMPATIA[i]=="En Común - Unidas Podemos" |
          datos$SIMPATIA[i]=="Unidas Podemos") {
      datos$SIMPATIA[i]="Unidas Podemos"}
      else{if (datos$SIMPATIA[i]=="PACMA (Partido Animalista)"){
      datos$SIMPATIA[i]="PACMA"}
      else {if (datos$SIMPATIA[i]=="PRC (Partido Regionalista de Cantabria)") {
      datos$SIMPATIA[i]="PRC"}
      else {if (datos$SIMPATIA[i]=="Navarra Suma (UPN)") {
      datos$SIMPATIA[i]="Navarra Suma"}
      else {if (datos$SIMPATIA[i]=="Otro" | datos$SIMPATIA[i]=="N.C." |
                datos$SIMPATIA[i]=="Ninguno"){
```

```
      datos$SIMPATIA[i]="N.S/N.C"}
        else{datos$SIMPATIA[i]=datos$SIMPATIA[i]}
      }}}}}
}

for (i in 1:length(datos$SITLAB)){
if (datos$SITLAB[i]=="En paro y busca su primer empleo" ){
    datos$SITLAB[i]="Paro y busca primer empleo"}
else {if (datos$SITLAB[i]=="En paro y ha trabajado antes") {
      datos$SITLAB[i]="Paro y ha trabajado"}
      else{if (datos$SITLAB[i]=="Jubilado/a o pensionista
               (anteriormente ha trabajado)"){
      datos$SITLAB[i]="Jubilado/pensionista (ha trabajado)"}
      else {if (datos$SITLAB[i]=="Pensionista (anteriormente no ha trabajado)" ) {
      datos$SITLAB[i]="Pensionista (no ha trabajado)"}
      else {if (datos$SITLAB[i]=="Navarra (Comunidad Foral de)") {
      datos$SITLAB[i]="Navarra"}
      else {if (datos$SITLAB[i]=="N.C." | datos$SITLAB[i]=="Otra situación"){
      datos$SITLAB[i]="N.C."}
      else {if (datos$SITLAB[i]=="Melilla (Ciudad Autónoma de)"){
      datos$SITLAB[i]="Melilla"}
        else{datos$SITLAB[i]=datos$SITLAB[i]}
      }}}}}
}}

for (i in 1:length(datos$SAT2)){
if (datos$SAT2[i]=="Cree que todo seguirá prácticamente igual, que no influirá en el
    datos$SAT2[i]="Misma situación"}
else {if (datos$SAT2[i]=="Dará lugar a la creación de más puestos de trabajo" ) {
      datos$SAT2[i]="Más trabajo"}
      else{if (datos$SAT2[i]=="Dará lugar a un aumento del paro"){
      datos$SAT2[i]="Aumento paro"}
        else{datos$SAT2[i]="N.S/N.C."}
      }}
}

for (i in 1:length(datos$P0)){
  if (datos$P0[i]=="La nacionalidad española" ){
      datos$P0[i]="Española"}
  else{datos$P0[i]="Española y otra"}
}

# nueva variable que se estudia en caso concreto:
datos$PARTIDOBLOQUES=rep(1,length(datos$SIMPATIA))

for (i in 1:length(datos$SIMPATIA)){
if (datos$SIMPATIA[i]=="PSOE" | datos$SIMPATIA[i]=="Unidas Podemos"){
```

```
    datos$PARTIDOBLOQUES[i]="Bloque Izquierda"}
else {if (datos$SIMPATIA[i]=="PP" | datos$SIMPATIA[i]=="VOX" |
          datos$SIMPATIA[i]=="Ciudadanos") {
      datos$PARTIDOBLOQUES[i]="Bloque Derecha"}
        else{datos$PARTIDOBLOQUES[i]="Otro"}
      }}

# se redefinen las variables ordinales para que sigan un "orden"

for (i in 1:length(datos$EDAD)){
if (datos$EDAD[i]<=24){
    datos$EDAD[i]="18-24"}
else {if (datos$EDAD[i]<=34) {
      datos$EDAD[i]="25-34"}
      else{if (datos$EDAD[i]<=44) {
      datos$EDAD[i]="35-44"}
      else {if (datos$EDAD[i]<=54) {
      datos$EDAD[i]="45-54"}
      else {if (datos$EDAD[i]<=64) {
      datos$EDAD[i]="55-64"}
      else {if (datos$EDAD[i]>64) {
      datos$EDAD[i]="65-99"}
          else{datos$EDAD[i]="N.C"}
      }}}}
}}

for (i in 1:length(datos$INTPOL)){
if (datos$INTPOL[i]=="Nada"){
    datos$INTPOL[i]="1-Nada"}
else {if (datos$INTPOL[i]=="Poco") {
      datos$INTPOL[i]="2-Poco"}
      else{if (datos$INTPOL[i]=="Regular") {
      datos$INTPOL[i]="3-Regular"}
      else {if (datos$INTPOL[i]=="Bastante") {
      datos$INTPOL[i]="4-Bastante"}
      else {if (datos$INTPOL[i]=="Mucho") {
      datos$INTPOL[i]="5-Mucho"}
          else{datos$INTPOL[i]="0-N.S/N.C"}
      }}}
}}

for (i in 1:length(datos$FUTUR1)){
if (datos$FUTUR1[i]=="Ningún o casi ningún cambio"){
    datos$FUTUR1[i]="1-Ningún o casi ningún cambio"}
else {if (datos$FUTUR1[i]=="Pocos cambios") {
      datos$FUTUR1[i]="2-Pocos cambios"}
      else{if (datos$FUTUR1[i]=="Regular, ni muchos ni pocos") {
```

```
      datos$FUTUR1[i]="3-Regular, ni muchos ni pocos"}
      else {if (datos$FUTUR1[i]=="Bastantes cambios") {
      datos$FUTUR1[i]="4-Bastantes cambios"}
      else {if (datos$FUTUR1[i]=="Muchos cambios") {
      datos$FUTUR1[i]="5-Muchos cambios"}
          else{datos$FUTUR1[i]="0-N.S/N.C"}
      }}}
}}

for (i in 1:length(datos$FUTUR2)){
if (datos$FUTUR2[i]=="Muy negativos"){
    datos$FUTUR2[i]="1-Muy negativos"}
else {if (datos$FUTUR2[i]=="Negativos") {
      datos$FUTUR2[i]="2-Negativos"}
      else{if (datos$FUTUR2[i]=="En parte positivos y en parte negativos" |
               datos$FUTUR2[i]=="Indiferentes"){
      datos$FUTUR2[i]="3-Indiferentes"}
      else {if (datos$FUTUR2[i]=="Positivos") {
      datos$FUTUR2[i]="4-Positivos"}
      else {if (datos$FUTUR2[i]=="Muy positivos") {
      datos$FUTUR2[i]="5-Muy positivos"}
        else{datos$FUTUR2[i]="0-N.S/N.C"}
      }}}
}}

for (i in 1:length(datos$CLASESUB)){
if (datos$CLASESUB[i]=="Clase baja/pobre"){
    datos$CLASESUB[i]="1-Clase baja/pobre"}
else {if (datos$CLASESUB[i]=="Clase trabajadora/obrera/proletariado") {
      datos$CLASESUB[i]="2-Clase trabajadora/obrera/proletariado"}
      else{if (datos$CLASESUB[i]=="Clase media-baja"){
      datos$CLASESUB[i]="3-Clase media-baja"}
      else {if (datos$CLASESUB[i]=="Clase media-media") {
      datos$CLASESUB[i]="4-Clase media-media"}
      else {if (datos$CLASESUB[i]=="Clase alta y media alta") {
      datos$CLASESUB[i]="5-Clase alta y media alta"}
        else{datos$CLASESUB[i]="0-N.S/N.C"}
      }}}
}}

for (i in 1:length(datos$ESCIDEOL)){
if (datos$ESCIDEOL[i]=="1 Izquierda") {
    datos$ESCIDEOL[i]="01 Izquierda"}
else {if (datos$ESCIDEOL[i]=="2"){
  datos$ESCIDEOL[i]="02"}
      else{if (datos$ESCIDEOL[i]=="3"){
      datos$ESCIDEOL[i]="03"}
```

```
      else {if (datos$ESCIDEOL[i]=="4") {
      datos$ESCIDEOL[i]="04"}
      else {if (datos$ESCIDEOL[i]=="5") {
      datos$ESCIDEOL[i]="05"}
      else {if (datos$ESCIDEOL[i]=="6"){
      datos$ESCIDEOL[i]="06"}
      else {if (datos$ESCIDEOL[i]=="7") {
      datos$ESCIDEOL[i]="07"}
      else {if (datos$ESCIDEOL[i]=="8") {
      datos$ESCIDEOL[i]="08"}
      else {if (datos$ESCIDEOL[i]=="9"){
      datos$ESCIDEOL[i]="09"}
      else {if (datos$ESCIDEOL[i]=="10 Derecha"){
      datos$ESCIDEOL[i]="10 Derecha"}
else {datos$ESCIDEOL[i]="0-N.S./N.C."}}
      }}}}}}}}


for (i in 1:length(datos$ESTUDIOS)){
if (datos$ESTUDIOS[i]=="Sin estudios"){
    datos$ESTUDIOS[i]="1-Sin estudios"}
else {if (datos$ESTUDIOS[i]=="Primaria") {
      datos$ESTUDIOS[i]="2-Primaria"}
      else{if (datos$ESTUDIOS[i]=="Secundaria 1ª etapa"){
      datos$ESTUDIOS[i]="3-Secundaria 1ª etapa"}
      else {if (datos$ESTUDIOS[i]=="Secundaria 2ª etapa") {
      datos$ESTUDIOS[i]="4-Secundaria 2ª etapa"}
      else {if (datos$ESTUDIOS[i]=="F.P." | datos$ESTUDIOS[i]=="Otros") {
      datos$ESTUDIOS[i]="5-F.P. y otros no superiores"}
      else {if (datos$ESTUDIOS[i]=="Superiores"){
      datos$ESTUDIOS[i]="6-Superiores"}
        else{datos$ESTUDIOS[i]="0-N.C"}
      }}}}
}}

for (i in 1:length(datos$OPINION)){
if (datos$OPINION[i]=="Peor"){
    datos$OPINION[i]="1-Peor"}
else {if (datos$OPINION[i]=="Igual") {
      datos$OPINION[i]="2-Igual"}
      else{if (datos$OPINION[i]=="Mejor") {
            datos$OPINION[i]="3-Mejor"}
            else {datos$OPINION[i]="0-N.S/Duda/N.C"}
      }
}}

for (i in 1:length(datos$SAT1)){
```

```
if (datos$SAT1[i]=="No aumentarán nada o casi nada"){
    datos$SAT1[i]="1-No aumentarán nada o casi nada"}
else {if (datos$SAT1[i]=="Aumentarán poco") {
      datos$SAT1[i]="2-Aumentarán poco"}
      else{if (datos$SAT1[i]=="Aumentarán algo"){
      datos$SAT1[i]="3-Aumentarán algo"}
      else {if (datos$SAT1[i]=="Aumentarán bastante") {
      datos$SAT1[i]="4-Aumentarán bastante"}
      else {if (datos$SAT1[i]=="Aumentarán mucho") {
      datos$SAT1[i]="5-Aumentarán muchos"}
        else{datos$SAT1[i]="0-N.S/N.C"}
      }}}
}}

names(datos)=c("CCAA","SEXO","EDAD","NACIONALIDAD","INTPOLITICO",
               "INESTADO","INFLUPOLITICA","FUTURO1","FUTURO2",
               "SISTEMAUTOM2","SISTEMAUTOM1","SIMBOLO","IDENTPERSONAS",
               "ESCIDEOL","PARTIDOSIMP","ESTUDIOS","SITLABORAL",
               "RELIGION","OPINION","CLASESOCIAL","PARTIDOBLOQUES")

# Ponemos las variables categoricas como factores y las ordinales como
# factores ordenados:
datos$CCAA=as.factor(datos$CCAA)
datos$SEXO=as.factor(datos$SEXO)
datos$EDAD=as.ordered(datos$EDAD)
datos$NACIONALIDAD=as.factor(datos$NACIONALIDAD)
datos$INTPOLITICO=as.ordered(datos$INTPOLITICO)
datos$INESTADO=as.factor(datos$INESTADO)
datos$INFLUPOLITICA=as.factor(datos$INFLUPOLITICA)
datos$FUTURO1=as.ordered(datos$FUTURO1)
datos$FUTURO2=as.ordered(datos$FUTURO2)
datos$SISTEMAUTOM2=as.ordered(datos$SISTEMAUTOM2)
datos$SISTEMAUTOM1=as.factor(datos$SISTEMAUTOM1)
datos$SIMBOLO=as.factor(datos$SIMBOLO)
datos$IDENTPERSONAS=as.factor(datos$IDENTPERSONAS)
datos$ESCIDEOL=as.ordered(datos$ESCIDEOL)
datos$PARTIDOSIMP=as.factor(datos$PARTIDOSIMP)
datos$ESTUDIOS=as.ordered(datos$ESTUDIOS)
datos$SITLABORAL=as.factor(datos$SITLABORAL)
datos$RELIGION=as.factor(datos$RELIGION)
datos$CLASESOCIAL=as.ordered(datos$CLASESOCIAL)
datos$OPINION=as.ordered(datos$OPINION)
datos$PARTIDOBLOQUES=as.factor(datos$PARTIDOBLOQUES)

# lo guardamos:

save(datos,file="C:/Users/glori/OneDrive - UNIVERSIDAD DE SEVILLA/Escritorio/MemoriaT
```

```
# save(datos,file="C:/Users/Escritorio/MemoriaTFEGloriaVizcaino/
#                  datos.Rdata")

load("datos.Rdata")
```

B.2. Análisis Descriptivo y Medidas

```
library(dplyr)

d2 <- datos %>%
group_by(IDENTPERSONAS,SEXO) %>% # agrupa por variables
summarise(Total=n()) %>% # resume calculando nuevas variables
mutate(Porcentaje=round(Total/sum(Total),4))

library(knitr)
library(kableExtra)

kable(d2,caption ="Variables IDENTPERSONAS y SEXO",longtable=TRUE) %>%
  row_spec(0, bold=T, background='gray!6',align='c') %>%
  kable_styling(latex_options = c("repeat_header",
                                  "condensed","hold_position"),
                repeat_header_text="(contin\\'ua)",
                position = "center",full_width = FALSE)
library(ggplot2)

ggplot(data = d2,
       mapping = aes(x = factor(IDENTPERSONAS),y = Porcentaje*100,
                     fill = factor(SEXO))) +
  geom_bar(position = 'dodge', colour="black", stat = 'identity') +
  labs(x = "Indentificación con sus mismos/as...", y = "Porcentaje",
       fill = "Sexo") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5)) +
  scale_fill_manual(values=c("#AEEEEE","#7A67EE"))
tabla1=table(datos$IDENTPERSONAS,datos$SEXO)
tabla1=tabla1[-8,] #quitar los N.S/N.C

library(DescTools)
a1=Lambda(tabla1, direction="column", conf.level = 0.95)
#en dirección va la independiente
b1=GoodmanKruskalTau(tabla1, direction="column", conf.level = 0.95)

c1=rbind.data.frame(a1,b1)
rownames(c1)=c("$\\lambda$ de Goodman y Krukal","$t$ de Goodman y Kruskal")
colnames(c1)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(c1,caption ="Medidas de variables IDENTPERSONAS y SEXO",
```

```
      row.names = T,escape = FALSE, digits = rep(4,3))%>%
  kable_styling(latex_options = c("striped", "condensed","hold_position"),
position = "center",
full_width = T)
tabla1a=tabla1[c(2,4,7),]
a2=Lambda(tabla1a, direction="column", conf.level = 0.95)
b2=GoodmanKruskalTau(tabla1a, direction="column", conf.level = 0.95)

c2=rbind.data.frame(a2,b2)
rownames(c2)=c("$\\lambda$ de Goodman y Krukal","$t$ de Goodman y Kruskal")
colnames(c2)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(c2,caption ="Medidas de variables IDENTPERSONAS y SEXO (2)",
      row.names = T,escape = FALSE, digits = rep(4,3))%>%
  kable_styling(latex_options = c("striped", "condensed","hold_position"),
position = "center",
full_width = T)
d3 <- datos  %>%
group_by(INESTADO,SEXO)  %>% # agrupa por variables
summarise(Total=n())  %>% # resume calculando nuevas variables
mutate(Porcentaje=round(Total/sum(Total),4))

kable(d3,caption ="Variables INESTADO y SEXO",longtable=TRUE)  %>%
  row_spec(0, bold=T, background='gray!6',align='c')  %>%
  kable_styling(latex_options = c("repeat_header","condensed","hold_position"),
position = "center",
repeat_header_text="(contin\\'ua)",
full_width = FALSE)


ggplot(data = d3,
       mapping = aes(x = factor(INESTADO),y = Porcentaje*100,
                     fill = factor(SEXO))) +
  geom_bar(position = 'dodge', colour="black", stat = 'identity') +
  labs(x = "¿Debería el Estado intervenir en la vida económica? ",
       y = "Porcentaje", fill = "Sexo") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5)) +
  scale_fill_manual(values=c("#AEEEEE","#7A67EE"))
tabla2=table(datos$INESTADO,datos$SEXO)
tabla2=tabla2[-1,] #quitar los N.S/N.C

a3.1=Lambda(tabla2, direction="column", conf.level = 0.95)
a3.2=Lambda(tabla2, direction="row", conf.level = 0.95)
b3.1=GoodmanKruskalTau(tabla2, direction="column", conf.level = 0.95)
b3.2=GoodmanKruskalTau(tabla2, direction="row", conf.level = 0.95)

c3=rbind.data.frame(a3.1,a3.2,b3.1,b3.2)
```

```
rownames(c3)=c("$\\lambda_{INESTADO}$ de Goodman y Krukal","$\\lambda_{SEXO}$ de Good
colnames(c3)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(c3,caption ="Medidas de variables INESTADO y SEXO",
      row.names = T,escape = FALSE, digits = rep(4,3))%>%
  kable_styling(latex_options = c("striped", "condensed","hold_position"),
position = "center",
full_width = T)
kable(tabla2,longtable=TRUE) %>%
   kable_styling(position = "center")
tabla4=table(datos$SIMBOLO,datos$CCAA)
tabla4=tabla4[-5,] #quitar los N.S/N.C

# kable(tabla4[,1:6],longtable=TRUE)%>%
#   row_spec(0, bold=T, background='gray!6',align='c')%>%
#   kable_styling(position = "center",full_width = T)
# kable(tabla4[,7:12],longtable=TRUE)%>%
#   row_spec(0, bold=T, background='gray!6',align='c')%>%
#   kable_styling(position = "center",full_width = T)
# kable(tabla4[,13:19],longtable=TRUE)%>%
#   row_spec(0, bold=T, background='gray!6',align='c')%>%
#   kable_styling(position = "center",full_width = T)
d5 <- datos %>%
group_by(SIMBOLO,CCAA) %>% # agrupa por variables
summarise(Total=n()) %>%# resume calculando nuevas variables
mutate(Porcentaje=round(Total/sum(Total),4))

# Esta tabla es demasiado grande, mejor no mostrarla

# kable(d5,caption ="Variables SÍMBOLO y CCAA")%>%
#   kable_styling(latex_options = c("striped", "condensed","hold_position"),
# position = "center",
# full_width = FALSE)


d5=d5[-(51:69),]
ggplot(data = d5,
       mapping = aes(x = factor(SIMBOLO),y = Porcentaje,
                    fill = factor(CCAA))) +
  geom_bar(colour="black", stat = 'identity') +
  labs(x = "El símbolo que mejor define la época actual es...",
       y = "Total", fill = "CCAA") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5),
        legend.key.height = unit (0.35, 'cm'))

a5.1=Lambda(tabla4, direction="column", conf.level = 0.95)
b5.1=GoodmanKruskalTau(tabla4, direction="column", conf.level = 0.95)
```

```
a5.2=Lambda(tabla4, direction="row", conf.level = 0.95)
b5.2=GoodmanKruskalTau(tabla4, direction="row", conf.level = 0.95)

c5=rbind.data.frame(a5.1,a5.2,b5.1,b5.2)
rownames(c5)=c("$\\lambda_{SIMBOLO}$ de Goodman y Krukal","$\\lambda_{CCAA}$ de Goodm
colnames(c5)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(c5,caption ="Medidas de variables SIMBOLO y CCAA",
      row.names = T,escape = FALSE, digits = rep(4,3), longtable=T) %>%
  kable_styling(latex_options = c("repeat_header","condensed","hold_position"),
                position = "center",repeat_header_text="(contin\\'ua)",
                full_width = FALSE)
d7 <- datos %>%
group_by(PARTIDOSIMP,RELIGION) %>% # agrupa por variables
summarise(Total=n()) %>%# resume calculando nuevas variables
mutate(Porcentaje=round(Total/sum(Total),4))

d7=d7[-c(9,17,25,36:40,46,56,62,67),]

kable(d7,caption ="Variables PARTIDOSIMP y RELIGIÓN",longtable=TRUE) %>%
  row_spec(0, bold=T, background='gray!6',align='c') %>%
  kable_styling(latex_options = c("repeat_header","condensed","hold_position"),
position = "center",
repeat_header_text="(contin\\'ua)", full_width = FALSE)

ggplot(data = d7,
       mapping = aes(x = factor(PARTIDOSIMP),y = Total,
                     fill = factor(RELIGION))) +
  geom_bar(position = 'dodge', colour="black", stat = 'identity') +
  labs(x = "¿Qué partido es con el que más simpatizas? ",
       y = "Total", fill = "Religión") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5))
tabla6=table(datos$PARTIDOSIMP,datos$RELIGION)
tabla6=tabla6[-11,-5] #quitar los N.S/N.C

a8=Lambda(tabla6, direction="column", conf.level = 0.95)
b8=GoodmanKruskalTau(tabla6, direction="column", conf.level = 0.95)

c8=rbind.data.frame(a8,b8)
rownames(c8)=c("$\\lambda$ de Goodman y Krukal","$t$ de Goodman y Kruskal")
colnames(c8)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(c8,caption ="Medidas de variables PARIDOSIMP y RELIGION",
      row.names = T,escape = FALSE, digits = rep(4,3))%>%
  kable_styling(latex_options = c("striped", "condensed","hold_position"),
position = "center",
full_width = T)
```

```
d8 <- datos %>%
group_by(SISTEMAUTOM1,SITLABORAL) %>% # agrupa por variables
summarise(Total=n()) %>% # resume calculando nuevas variables
mutate(Porcentaje=round(Total/sum(Total),4))

kable(d8,caption ="Variables SISTEMAUTOM1 y SITLABORAL",longtable=TRUE) %>%
  row_spec(0, bold=T, background='gray!6',align='c') %>%
  kable_styling(latex_options = c("repeat_header","condensed","hold_position"),
position = "center",
repeat_header_text="(contin\\'ua)",
full_width = FALSE) %>%
column_spec(2,width = "7cm")


ggplot(data = d8,
       mapping = aes(x = factor(SISTEMAUTOM1),y = Porcentaje,
                     fill = factor(SITLABORAL))) +
  geom_bar(position = 'dodge', colour="black", stat = 'identity') +
  labs(x = "La utilización de robots y sistemas automáticos
       va a ser causa en los próximos años de...",
       y = "Porcentaje", fill = "Situación Laboral") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5))
tabla7=table(datos$SISTEMAUTOM1,datos$SITLABORAL)
tabla7=tabla7[-4,-3] #quitar los N.S/N.C

a9=Lambda(tabla7, direction="column", conf.level = 0.95)
b9=GoodmanKruskalTau(tabla7, direction="column", conf.level = 0.95)

c9=rbind.data.frame(a9,b9)
rownames(c9)=c("$\\lambda$ de Goodman y Krukal","$t$ de Goodman y Kruskal")
colnames(c9)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(c9,caption ="Medidas de variables SISTEMAUTOM1 y SITLABORAL",
      row.names = T,escape = FALSE, digits = rep(4,3))%>%
  kable_styling(latex_options = c("striped", "condensed","hold_position"),
position = "center",
full_width = T)
d10 <- datos %>%
group_by(PARTIDOSIMP,INFLUPOLITICA) %>% # agrupa por variables
summarise(Total=n()) %>% # resume calculando nuevas variables
mutate(Porcentaje=round(Total/sum(Total),4))
d10=d10[c(8:15,81:88,90:97,99:114),]
kable(d10,caption ="Variables PARTIDOSIMP y INFLUPOLITICA",longtable=TRUE) %>%
  row_spec(0, bold=T, background='gray!6',align='c') %>%
  kable_styling(latex_options = c("repeat_header","condensed","hold_position"),
position = "center",
repeat_header_text="(contin\\'ua)",
```

```
full_width = FALSE)

ggplot(data = d10,
       mapping = aes(x = factor(PARTIDOSIMP),y = Porcentaje*100,
                     fill = factor(INFLUPOLITICA))) +
  geom_bar(position = 'dodge', colour="black", stat = 'identity') +
  labs(x = "¿Qué partido es con el que más simpatizas?",
       y = "Porcentaje", fill = "Mayor Influencia") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5))
tabla9=table(datos$PARTIDOSIMP,datos$INFLUPOLITICA)
tabla9=tabla9[c(3,14,16,18,19),] #escoger los partidos con más votos

a11=Lambda(tabla9, direction="column", conf.level = 0.95)
b11=GoodmanKruskalTau(tabla9, direction="column", conf.level = 0.95)

c11=rbind.data.frame(a11,b11)
rownames(c11)=c("$\\lambda$ de Goodman y Krukal","$t$ de Goodman y Kruskal")
colnames(c11)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(c11,caption ="Medidas de variables PARTIDOSIMP e INFLUPOLITICA",
      row.names = T,escape = FALSE, digits = rep(4,3))%>%
  kable_styling(latex_options = c("striped", "condensed","hold_position"),
position = "center",
full_width = T)
d11 <- datos %>%
group_by(PARTIDOSIMP,INESTADO) %>% # agrupa por variables
summarise(Total=n()) %>% # resume calculando nuevas variables
mutate(Porcentaje=round(Total/sum(Total),4))
d11=d11[c(5,6,31,32,35,36,39,40,42,43),]
kable(d11,caption ="Variables PARTIDOSIMP y INESTADO",longtable=TRUE) %>%
  row_spec(0, bold=T, background='gray!6',align='c') %>%
  kable_styling(latex_options = c("repeat_header","condensed","hold_position"),
position = "center",
repeat_header_text="(contin\\'ua)",
full_width = FALSE)

ggplot(data = d11,
       mapping = aes(x = factor(PARTIDOSIMP),y = Porcentaje*100,
                     fill = factor(INESTADO))) +
  geom_bar(position = 'dodge', colour="black", stat = 'identity') +
  labs(x = "¿Qué partido es con el que más simpatizas?",
       y = "Porcentaje", fill = "¿Debería el Estado intervenir \n en la vida económic
  theme(axis.text.x  = element_text(angle=90, vjust=0.5)) +
  scale_fill_manual(values=c("#AEEEEE","#7A67EE"))
tabla11=table(datos$PARTIDOSIMP,datos$INESTADO)
tabla11=tabla11[c(3,14,16,18,19),-1] #escoger los partidos con más votos
```

```
a12.1=Lambda(tabla11, direction="column", conf.level = 0.95)
a12.2=Lambda(tabla11, direction="row", conf.level = 0.95)
b12.1=GoodmanKruskalTau(tabla11, direction="column", conf.level = 0.95)
b12.2=GoodmanKruskalTau(tabla11, direction="row", conf.level = 0.95)

c12=rbind.data.frame(a12.1,a12.2,b12.1,b12.2)
rownames(c12)=c("$\\lambda_{PARIDOSIMP}$ de Goodman y Krukal","$\\lambda_{INESTADO}$ d
colnames(c12)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(c12,caption ="Medidas de variables PARTIDOSIMP e INESTADO",
      row.names = T,escape = FALSE, digits = rep(4,3),longtable=T) %>%
  kable_styling(latex_options = c("repeat_header","condensed","hold_position"),
                position = "center",repeat_header_text="(contin\\'ua)",
                full_width = T)
tabla11.=table(datos$PARTIDOBLOQUES,datos$INESTADO)
tabla11.=tabla11.[-3,-1]

a13.1=Lambda(tabla11., direction="column", conf.level = 0.95)
a13.2=Lambda(tabla11., direction="row", conf.level = 0.95)
b13.1=GoodmanKruskalTau(tabla11., direction="column", conf.level = 0.95)
b13.2=GoodmanKruskalTau(tabla11., direction="row", conf.level = 0.95)

c13=rbind.data.frame(a13.1,a13.2,b13.1,b13.2)
rownames(c13)=c("$\\lambda_{PARTIDOSIMP}$ de Goodman y Krukal","$\\lambda_{INESTADO}$
colnames(c13)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(c13,caption ="Medidas de variables PARTIDOSIMP e INESTADO (2)",
      row.names = T,escape = FALSE, digits = rep(4,3), longtable=T) %>%
  kable_styling(latex_options = c("repeat_header","condensed","hold_position"),
                position = "center",repeat_header_text="(contin\\'ua)",
                full_width = T)
cor=cor(apply(datos[,c(3,5,8:10,14,16,19,20)],2,xtfrm),method = "kendall")

library(corrplot)
corrplot(cor,title = "", tl.col = "black")

d1_ <- datos %>%
group_by(EDAD,INTPOLITICO) %>% # agrupa por variables
summarise(Total=n()) %>% # resume calculando nuevas variables
mutate(Porcentaje=round(Total/sum(Total),4))

kable(d1_,caption ="Variables EDAD e INTPOLITICO",longtable=TRUE) %>%
  row_spec(0, bold=T, background='gray!6',align='c') %>%
  kable_styling(latex_options = c("repeat_header","condensed","hold_position"),
position = "center",
repeat_header_text="(contin\\'ua)",
full_width = FALSE)
```

```
ggplot(data = d1_,
       mapping = aes(x = factor(EDAD),y = Porcentaje*100,
                     fill = factor(INTPOLITICO))) +
  geom_bar(position = 'dodge', colour="black", stat = 'identity') +
  labs(x = "Interés político según edad", y = "Porcentaje",
       fill = "Interés Político") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5))
tabla1_=table(datos$EDAD,datos$INTPOLITICO)
tabla1_=tabla1_[,-1]

e1=KendallTauA(tabla1_, conf.level = 0.95)
f1=KendallTauB(tabla1_, conf.level = 0.95)
g1=StuartTauC(tabla1_, conf.level = 0.95)
h1=GoodmanKruskalGamma(tabla1_, conf.level = 0.95)
i1=SomersDelta(tabla1_, direction="column", conf.level = 0.95)
j1=SomersDelta(tabla1_, direction="row", conf.level = 0.95)

k1=rbind.data.frame(e1,f1,g1,h1,i1,j1)
rownames(k1)=c("$\\tau_a$ de Kendall","$\\tau_b$ de Kendall",
               "$\\tau_c$ de Stuart","$\\gamma$ de Goodman y Kruskal",
               "$d_{EDAD}$ de Sommers","$d_{INTPOLITICO}$ de Sommers")
colnames(k1)=c("Estadístico","Límite Inferior IC","Límite Superior IC")

kable(k1,caption ="Medidas de variables EDAD e INTPOLITICO",
      row.names = T,escape = FALSE, digits = rep(4,3)) %>%
  kable_styling(latex_options = c("striped", "condensed","hold_position"),
position = "center",
full_width = T)
```

Bibliografía

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2021. URL https://CRAN.R-project.org/package=rmarkdown. R package version 2.11.

J. Berkson. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Am. Stat. Assoc.*, 33, 1938.

Kenneth J. Berry, Janis Johnston, and Paul W. Jr. Mielke. *The Measurement of Association: A Permutation Statistical Approach*. 2010.

K.J. Berry and P.W. Mielke. A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educ. Psychol. Meas.*, 48, 1988.

K.J. Berry and P.W. Mielke. Extension of spearman's footrule to multiple rankings. *Psychol. Rep.*, 82, 1998.

D. Böhning and H. Holling. A monte carlo study on minimizing chi-square distances under the hypothesis of homogeneity or independence for a two-way contingency table. *Statistics*, 20:55–70, 1989.

H. Brenner and U. Kliebsch. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, 7, 1996.

I.D.J. Bross. How to use ridit analysis. 14, 1958.

J. Cohen. *A coefficient of agreement for nominal scales*, volume 20. 1960.

J. Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, 70, 1968.

Herbert L. Costner. *Criteria for Measures of Association*. American Sociological Review, 1965.

M. Cowles. *Statistics in Psychology: An Historical Perspective*. 2 edition, 2001.

E.E. Cureton. Rank-biserial correlation. *Psychometrika*, 21, 1956.

R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 5th ed edition, 1934.

J.L. Fleiss. Measuring nominal scale agreement among many raters. *Psych. Bull.*, 76, 1971.

J.L. Fleiss. The equivalence of weighted kappa and the intraclass coefficient as measures of reliability. *Educ. Psychol. Meas.*, 33, 1973.

J.L. Fleiss, B. Levin, and M.C. Paik. *Statistical Methods for Rates and Proportions.* 5 edition, 2003.

L.A. Franklin. Exact tables of spearman's footrule for $n = 11(1)18$ with estimate of convergence and errors for the normal approximation. *Stat. Probab. Lett.*, 6, 1988.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, 32, 1937.

G.V Glass. Note on rank-biserial correlation. *Educ. Psychol. Meas.*, 26, 1966.

L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications. *J. Am. Stat. Assoc.*, 49, 1954.

L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications, iii: Approximate sampling theory. *J. Am. Stat. Assoc.*, 58, 1963.

A.A. Hunter. *On the validity of measures of association: The nominal-nominal two-by-two case.* 1973.

R. Iachan. Measures of agreement for incompletely ranked data. *Educ. Psychol. Meas.*, 44, 1984.

J.E. Johnston, K.J. Berry, and P.W. Mielke. Permutation tests: Precision in estimating probability values. *Percept. Motor Skill*, 105, 2007.

M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30, 1938.

M.G. Kendall. *Rank Correlation Methods.* 1948.

M.G. Kendall. *Rank Correlation Methods, 3rd edn.* 1962.

M.G. Kendall and B. Babington Smith. The problem of m rankings. *Ann.Math. Stat.*, 10, 1939.

M.G. Kendall and B. Babington Smith. On the method of paired comparisons. *Biometrika*, 31, 1940.

M.G. Kendall, S.F.H. Kendall, and B. Babington Smith. The distribution of spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika*, 30, 1939.

J.-O. Kim. Predictive measures of ordinal association. *Am. J. Soc.*, 76, 1971.

C.A. Kraft and C. van Eeden. *A Nonparametric Introduction to Statistics.* Macmillan, 1968.

K. Krippendorff. *Bivariate agreement coefficients for reliability of data.* Borgatta, E.F. (ed.) Sociological Methodology, 1970.

J.R. Landis and G.G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33, 1977.

R.K. Leik and W.R. Gove. *Integrated approach to measuring association.* Costner, H.L. (ed.) Sociological Methodology, 1971.

A.M. Liebetrau. *Measures of Association.* 1983.

R.J. Light. *Measures of response agreement for qualitative data: Some generalizations and alternatives*, volume 76. Psychol. Bull, 1971.

M. Maclure and W.C. Willett. Misinterpretation and misuse of the kappa statistic. *Am. J. Epidemiol.*, 126, 1987.

H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, 18, 1947.

B.H. Margolin and R.J. Light. An analysis of variance for categorical data, ii: Small sample comparisons with chi square and other competitors. *J. Am. Stat. Assoc*, 69, 1974.

P.W. Mielke and K.J. Berry. Cumulant methods for analyzing independence ofr-way contingency tables and goodness-of-fit frequency data. *Biometrika*, 1988a.

P.W. Mielke and K.J. Berry. Cumulant methods for analyzing independence of r-way contingency tables and goodness-of-fit frequency data. *Biometrika*, 75, 1988b.

P.W. Mielke, K.J. Berry, and J.E. Johnston. Resampling programs for multiway contingency tables with fixed marginal frequency totals. *Psychol. Rep.*, 101, 2007.

P.W. Mielke, K.J. Berry, and J.E. Johnston. Resampling probability values for weighted kappa with multiple raters. *Psychol. Rep.*, 102, 2008.

P.W. Mielke, M.A. Long, K.J. Berry, and J.E. Johnston. g-treatment ridit analysis: Resampling permutation methods. *Stat. Methodol.*, 6, 2009.

W.S. Robinson. The statistical measurement of agreement. *Am. Sociol. Rev.*, 22, 1957.

I.A. Salama and D. Quade. A note on spearman's footrule. *Commun. Stat. Simul. C*, 19, 1990.

W.A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opin. Quart.*, 19, 1955.

R.H. Somers. A new asymmetric measure of association for ordinal variables. *Am. Sociol. Rev.*, 27, 1962.

C.E. Spearman. The proof and measurement of association between two things. *Am. J. Psychol.*, 15, 1904.

C.E. Spearman. 'footrule' for measuring correlation. *Brit. J. Psychol.*, 2, 1906.

A. Stuart. The estimation and comparison of strengths of association in contingency tables. *Biometrika*, 40, 1953.

A. Stuart. Spearman-like computation of kendall's tau. *Brit. J. Math. Stat. Psy.*, 30, 1977.

H.K. Ury and D.C. Kleinecke. Tables of the distribution of spearman's footrule. *J. R. Stat. Soc. C Appl.*, 28, 1979.

W.A. Wallis. The correlation ratio for ranked data. *J. Am. Stat. Assoc.*, 34, 1939.

R.S. Weiss. *Statistics in Social Research: An Introduction.* Wiley, 1968.

J.W. Whitfield. Rank correlation between two variables, one of which is ranked, the other dichotomous. *Biometrika*, 34, 1947.

Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2021a. URL https://CRAN.R-project.org/package=ggplot2. R package version 3.3.5.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021b. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.7.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bull.*, 1, 1945.

T.P. Wilson. *Measures of association for bivariate ordinal hypotheses.* Blalock, H.M., 1974.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2021. URL https://yihui.org/knitr/. R package version 1.34.