UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

DEPARTAMENTO DE ESTADÍSTICA

TRABAJO DE GRADUACIÓN:

CARACTERIZACIÓN DE LA INSUFICIENCIA RENAL CRÓNICA EN LAS COMUNIDADES DEL BAJO LEMPA DE EL SALVADOR, A TRAVÉS DE TÉCNICAS MULTIVARIANTES EN EL PERIODO COMPRENDIDO DESDE AGOSTO HASTA DICIEMBRE DE 2009

PRESENTADO POR:

MIGUEL ANGEL GÓMEZ ANGEL

PARA OPTAR EL GRADO DE:

LICENCIATURA EN ESTADÍSTICA

CIUDAD UNIVERSITARIA, 21 DE OCTUBRE DE 2016

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

DEPARTAMENTO DE ESTADÍSTICA

TRABAJO DE GRADUACIÓN:

CARACTERIZACIÓN DE LA INSUFICIENCIA RENAL CRÓNICA EN LAS COMUNIDADES DEL BAJO LEMPA DE EL SALVADOR, A TRAVÉS DE TÉCNICAS MULTIVARIANTES EN EL PERIODO COMPRENDIDO DESDE AGOSTO HASTA DICIEMBRE DE 2009

PRESENTADO POR:

MIGUEL ANGEL GÓMEZ ANGEL

ASESORES:

MSc. PORFIRIO ARMANDO RODRÍGUEZ

DR. CARLOS MANUEL ORANTES NAVARRO

CIUDAD UNIVERSITARIA, 21 DE OCTUBRE DE 2016

AUTORIDADES

UNIVERSIDAD DE EL SALVADOR

RECTOR UNIVERSITARIO:

LIC. LUIS ARGUETA ANTILLÓN (INTERINO)

SECRETARIA GENERAL:

DRA. ANA LETICIA ZAVALETA DE AMAYA

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

DECANO:

LIC. MAURICIO HERNÁN LOVO

SECRETARIA:

LICDA. DAMARIS MELANY HERRERA TURCIOS

ESCUELA DE MATEMÁTICA

DIRECTOR:

DR. JOSÉ NERYS FUNES TORRES

SECRETARIA:

MSc. ALBA IDALIA CÓRDOVA CUÉLLAR

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

ASESORES:

_____

MSc. PORFIRIO ARMANDO RODRÍGUEZ

DOCENTE ESCUELA DE MATEMÁTICA

_____

DR. CARLOS MANUEL ORANTES NAVARRO

NEFRÓLOGO DEL INSTITUTO NACIONAL DE SALUD DE EL SALVADOR

CIUDAD UNIVERSITARIA, 21 DE OCTUBRE DE 2016

AGRADECIEMIENTOS

Miguel Angel Gómez Angel:

Gracias Amado Dios Padre Celestial Divino Jesús por amarme tanto, por haber dado la vida por mí, por estar conmigo siempre a cada instante de mi vida amándome, escuchándome, ayudándome, orientándome, dirigiéndome, protegiéndome, corrigiéndome.

Gracias Señor Jesús por amarme tanto, por darme infinidad de bendiciones en mi vida, entre de las cuales esta, el haber podido llegar a ser Licenciado en Estadística. Por lo tanto bendito seas Amado Jesús por los siglos de los siglos. Amén, así sea y así es.

Gracias Amada Madre Virgen María por amarme tanto, por llevarme siempre de tu mano, por protegerme, por envolverme con tu manto a cada instante de mi vida. Bendita seas Madre Virgen María, por siempre y para siempre. Amén, así sea y así es.

Gracias Amados Arcángeles, Amados Ángeles, Amado Ángel de mi Guarda, Por amarme tanto, por estar conmigo a cada instante de mi vida escuchándome, ayudándome, orientándome, dirigiéndome, protegiéndome, corrigiéndome. Benditos sean Arcángeles, Ángeles, y Ángel de mi Guarda, por siempre y para siempre. Amén, así sea y así es.

Gracias Amado San Judas Tadeo por ser mi especial y poderoso protector, Bendito seas por siempre y para siempre. Amén, así sea y así es.

Le doy gracias a mis padres Miguel Angel y María Luz, por apoyarme en todo momento, por los principios y valores cristianos, espirituales y morales que me han inculcado, y en particular por apoyarme a lo largo del estudio de mi carrera de Licenciatura en Estadística.

Les agradezco a mis asesores MSc. Porfirio Armando Rodríguez y el Dr. Carlos Manuel Orantes Navarro, por haber compartido sus conocimientos, sus amistad, sus confianza, sus apoyo, y haber dedicado sus tiempo.

Les agradezco a mis amigos o hermanos con quienes hasta el día de hoy he convivido y que han formado parte de mis experiencias de vida, de las cuales he aprendido mucho acerca del valor que tiene la vida.

DEDICATORIA

A Dios nuestro Señor y a mis padres Miguel Angel y María Luz. Bendito y alabado seas Señor Dios del Universo por los siglos de los siglos. Amén, así sea y así es. Y que Dios siga derramando infinidad de bendiciones sobre mis padres.

Miguel Angel Gómez Angel

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. RESUMEN

La Insuficiencia Renal Crónica (IRC) constituye un problema de salud poblacional en El Salvador [1]. Los residentes de las comunidades del Bajo Lempa, en el departamento de Usulután, municipio de Jiquilisco, El Salvador, percibieron la existencia de una alta prevalencia de IRC en la región. De ahí que el Ministerio de Salud decidiera diseñar el estudio, conocido como Nefrolempa 2009, dirigido a estudiar la Enfermedad Renal Crónica (ERC) en todos sus estadios en la región [2]. El presente análisis titulado "CARACTERIZACIÓN DE LA INSUFICIENCIA RENAL CRÓNICA EN LAS COMUNIDADES DEL BAJO LEMPA DE EL SALVADOR, A TRAVÉS DE TECNICAS MULTIVARIANTES EN EL PERIODO COMPRENDIDO DESDE AGOSTO HASTA DICIEMBRE DE 2009" es un estudio secundario descriptivo y analítico, basado en los datos obtenidos del estudio Nefrolempa, el cual se realizó a través de una pesquisa activa de la ERC y los factores de riesgo asociados en la región. Dado que el estudio Nefrolempa fue dirigido a estudiar la ERC en todos sus estadios, en el presente análisis se plantea la pregunta: ***¿cuál es el comportamiento epidemiológico y clínico de la IRC y su asociación con los factores de riesgo?*** Por lo tanto, el objetivo general del estudio es: "Realizar una caracterización multivariante que discrimine entre aquellos factores que influyen para el padecimiento de la IRC en las comunidades del Bajo Lempa". Y Además, el propósito es presentar evidencia estadística y científica, que pueda servir como modelo en estudios similares. Se tiene a disposición la base de datos del estudio Nefrolempa, la cual cuenta con características sociales, epidemiológicas y clínicas. Dicha base de datos se adaptó y transformó de manera que permita desarrollar los análisis propuestos. En seguida, se realizan procedimientos estadísticos que permiten obtener una caracterización más adecuada de las variables, se determinan variables asociadas con la IRC a través de un abordaje descriptivo multivariante y también se determinan si los datos cumplen los criterios e hipótesis que se requieren para poder abordar la Regresión Logística Binaria. Posteriormente, se procede a aplicar la Regresión Logística Binaria, para determinar factores de riesgo asociados con la IRC. En conjunto con médicos del INS, se realiza el análisis e interpretación de resultados. Finalmente se plantean las conclusiones del estudio.

PALABRAS CLAVE. Insuficiencia Renal Crónica/epidemiología, factores de riesgo, prevalencia, plaguicidas, agroquímicos, El Salvador.

2. INTRODUCCIÓN

La Enfermedad Renal Crónica (ERC) se clasifica en 5 estadios. Para los estadios 1 y 2 se le diagnostica a través de la determinación de la función renal y presencia de marcadores de daño renal persistentes durante al menos 3 meses. Para los estadios 3, 4 y 5 se le conoce como Insuficiencia Renal Crónica (IRC) [3]. La ERC Actualmente se reconoce como un importante problema global de salud de la población [4, 5].

En los países desarrollados, el progresivo aumento en el número de pacientes con ERC (Enfermedad Renal Crónica) y consecuentemente aquellos que requieren Tratamiento de Reemplazo Renal (TRR), ya sea por diálisis o transplante renal, está alcanzando niveles de epidemia, aumentando de 5 a 8% anualmente [4, 5]. Aunque los datos de los países en desarrollo son escasos, se estima que en el 2030, un 70% de los pacientes con Enfermedad Renal Terminal (ERT, estadio 5 de la ERC), se encontrarán en los países en desarrollo, en los cuales esta demanda creciente sobrepasará la capacidad presupuestaria de los sistemas de atención de salud [6,7].

En Centroamérica y al sur de México se ha reportado un aumento de ERC en la última década. Los resultados de los estudios epidemiológicos varían y refieren la alta prevalencia en áreas costeras principalmente en agricultores hombres, sobre todo menores de 60 años, que están expuestos a productos agroquímicos en combinación con la presencia de otros factores de riesgo [8-11].

Haciendo una descripción de la región de estudio, el rio Lempa es el más largo en Centroamérica y desagua en el Océano Pacifico. El lecho serpentea a través de Guatemala, Honduras y El Salvador. Los principales ríos que fluyen a través de las ciudades desaguan en el Lempa, llevando consigo los desechos sólidos y líquidos de las industrias, los asentamientos urbanos y marginales [12].

En el sur de El Salvador, a lo largo de las riberas del Lempa hasta su desembocadura se encuentran distribuidas comunidades pobladas por personas de escasos recursos económicos, que trabajan principalmente en la agricultura. Esta región se conoce como el Bajo Lempa [13]. Los residentes del Bajo Lempa percibieron la existencia de una alta prevalencia de IRC en dichas comunidades. Los funcionarios del Ministerio de Salud decidieron estudiar y enfocar este problema de manera integral, percatándose de que los pacientes de la región del Bajo Lempa que comenzaban diálisis eran pre-dominantemente trabajadores

agrícolas hombres menores de 60 años que, entre otros factores, estaban expuestos a agroquímicos [2].

Por lo anterior, el diseño del estudio, conocido como Nefrolempa, está dirigido a profundizar más en la ERC en todos sus estadios, investigando la prevalencia de la enfermedad en la región, así como sus factores de riesgo, tradicionales y no tradicionales. Los objetivos fueron identificar los factores de riesgo de la ERC y los marcadores de daño renal y vascular en la orina, medir el funcionamiento renal y describir la prevalencia de la ERC en la población de 18 o más años de edad residente en tres comunidades rurales de la región del Bajo Lempa en el municipio de Jiquilisco, El Salvador: Nueva Esperanza, Ciudad Romero, y la Canoa [2].

El Instituto Nacional de Salud (INS) de El Salvador, cuenta con información recolectada del estudio Nefrolempa, organizada en una base de datos en forma de matriz, donde los registros o unidades de estudio son las personas, a las cuales se les ha medido una serie de indicadores o variables con el objeto de medir el nivel de influencia de la enfermedad en la región.

Consecuentemente, a partir de la base de datos se toma una muestra de individuos con variables tanto sociales, epidemiológicas y clínicas las cuales se suponen que podrían tener algún tipo de asociación plausible con el fenómeno de la IRC, por lo que se realizan procedimientos estadísticos, teniendo por objeto principal una caracterización multivariante orientada a la determinación de variables y factores de riesgo asociados con la IRC, en las comunidades del Bajo Lempa.

Como primer paso del tratamiento de la información, se realiza una depuración y validación de la base de datos. Luego, se realizan análisis de tipo univariado y bivariado con el fin de obtener una descripción adecuada de los datos que permite descubrir características de interés vinculadas con la IRC. Así mismo dentro del análisis bivariado se encuentran variables asociadas de manera significativa con la variable dependiente y denominada “Estado”, que clasifica a personas con IRC y sin ERC. Seguidamente se aplican técnicas multivariantes, para caracterizar o buscar una tipología de individuos representados por distintas variables, dentro de la cual se contempla la determinación de variables y factores de riesgo asociados con la IRC. Finalmente en conjunto con médicos del INS se proporciona el análisis e interpretación de los resultados y así se plantean las conclusiones del estudio.

3. PLANTEAMIENTO DEL PROBLEMA, ANTECEDENTES, JUSTIFICACIÓN Y OBJETIVOS

3.1. PLANTEAMIENTO DEL PROBLEMA

Como resultado de la evolución demográfica e epidemiológica, en la actualidad, se considera que en todo el mundo la IRC se está comportando como una epidemia. Se ha llegado a estimar que el número de personas afectadas a nivel mundial es superior a los 500 millones [14].

La IRC está caracterizada por una creciente incidencia y prevalencia de los pacientes en Tratamiento de Reemplazo Renal (TRR), los cuales están asociados a una prematura mortalidad, discapacidad, disminución de la calidad de vida y un elevado y creciente costo de los servicios de salud, constituyendo una de las principales causas de muerte en el mundo [15].

En El Salvador, la IRC constituye un serio problema de salud poblacional. Esta enfermedad es la principal causa de muerte hospitalaria en la población adulta, la segunda causa de mortalidad en toda la población masculina y la quinta causa de muerte en personas con edades mayores o iguales a 18 años. Se considera que el conocimiento epidemiológico es incompleto. Dada esta realidad se plantea la necesidad de generar un análisis detallado en torno a esta problemática [1].

El estudio Nefrolempa fue dirigido al abordaje integral de la Enfermedad Renal Crónica en todos sus estadios, dentro de los cuales se encuentran los estadios 3 al 5, que corresponden a la Insuficiencia Renal Crónica. Sin embargo en el presente estudio, se plantea la siguiente pregunta de investigación: **¿cuál es el comportamiento epidemiológico y clínico de la IRC y su asociación con los factores de riesgo?**

Para darle solución a la pregunta de investigación, se considera pertinente conocer de forma apropiada este problema de salud a través de un estudio por medio de análisis estadísticos multivariantes.

3.2. ANTECEDENTES Y JUSTIFICACION

3.2.1. ANTECEDENTES

En El Salvador, en los últimos años, se ha generado una serie de estudios acerca de la ERC, entre los cuales es apropiado mencionar: “”Enfermedad Renal Crónica y Factores de Riesgo Asociados en El Bajo Lempa, El Salvador. Estudio Nefrolempa, 2009”. Este estudio fue realizado por un equipo de investigación apoyado por el Ministerio de Salud de El Salvador y profesores del Instituto de Nefrología del Ministerio de Salud Pública de Cuba como asesores, con el auspicio de la OPS y la participación activa de médicos salvadoreños y estudiantes de la Escuela Latinoamericana de Medicina de Cuba, la Escuela de Medicina de la Universidad de El Salvador y la Asociación de Comunidades Unidas del Bajo Lempa [2].

Se realizó un estudio transversal, descriptivo y analítico desde agosto hasta Diciembre del 2009, combinando métodos epidemiológicos y clínicos a través de una pesquisa activa de ERC y los factores de riesgo en la población de 18 o más años de edad residente en tres comunidades rurales del área del Bajo Lempa en el municipio de Jiquilisco, El Salvador: Nueva Esperanza, Ciudad Romero y La Canoa [2].

El estudio se desarrolló en tres fases [2]:

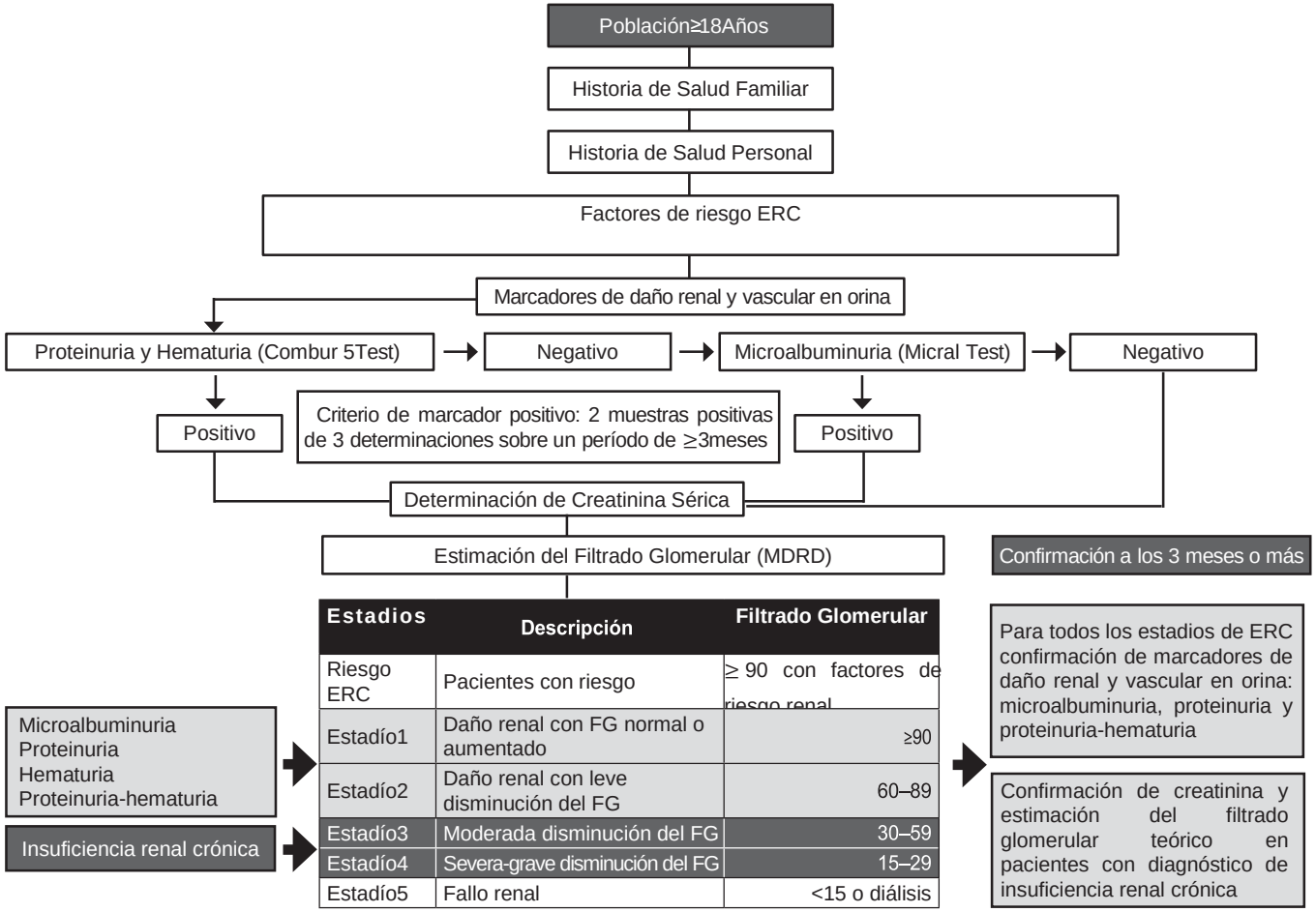
- Pesquisa activa de factores de riesgo de ERC y marcadores de daño renal y vascular, y evaluación de la función renal en la población objeto del estudio.

- Tres meses después se confirmaron los marcadores de daño renal y vascular en la orina, la evaluación del funcionamiento renal y los casos de ERC, su clasificación por estadio, y la asociación con los factores de riesgo presentes.

- La evaluación clínica de los casos individuales de ERC y la propuesta de una Unidad de Salud Renal para el seguimiento de los pacientes.

El algoritmo para el diagnóstico de la ERC se resume en la Figura 3.1.Figura 3.1. Algoritmo diagnóstico de la enfermedad renal crónica, estudio Nefrolempa.

Figura 3.1. Algoritmo diagnóstico de la enfermedad renal crónica, estudio Nefrolempa.



Universo del estudio Todas las personas con edades mayores o iguales que 18 años de edad en las tres comunidades, identificados por un censo de población realizado de casa en casa, un total de 878 personas, reunían los requisitos para participar [2].

Criterios de Inclusión Los residentes permanentes de 18 años de edad en las comunidades de Nueva Esperanza, Ciudad Romero y la Canoa, que después del consentimiento informado expresaron su aceptación a participar en el estudio. El estudio incluyó el 88.3% de la población enumerada, 775 personas (343 hombres, 432 mujeres) [2].

En la Tabla 3.1 se presentan las variables estudiadas en Nefrolempa:

Tabla 3.1. Variables del estudio Nefrolempa.

<table>
<tr><th>Variable</th><th>Descripción</th></tr>
<tr><td>Edad</td><td>Mayor o igual que 18 años</td></tr>
<tr><td>Sexo</td><td>Masculino y femenino</td></tr>
<tr><td>Estado civil</td><td>Soltero(a), unión consensual (o libre), casado(a), viudo(a), divorciado(a)</td></tr>
<tr><td rowspan="8">Factores de riesgo referidos en la entrevista</td><td>Historia familiar de enfermedad renal crónica, diabetes, hipertensión arterial</td></tr>
<tr><td>Historia personal de enfermedad renal crónica, diabetes, hipertensión arterial</td></tr>
<tr><td>Hábito de fumar actual o pasado</td></tr>
<tr><td>Consumo de alcohol actual o pasado</td></tr>
<tr><td>Uso de medicamentos antiinflamatorios no esteroides (AINES), plantas medicinales y antibióticos</td></tr>
<tr><td>Ocupación laboral</td></tr>
<tr><td>Contacto con agroquímicos</td></tr>
<tr><td>Historia de enfermedades infecciosas</td></tr>
<tr><td>Tensión arterial (mmHg [1])
Clasificación JNC7–2003 [2][16]</td><td>
<table>
<tr><td></td><td>Sistólica</td><td>Diastólica</td></tr>
<tr><td>Normal</td><td>Menor a 120</td><td>y menor a 80</td></tr>
<tr><td>Pre-hipertensión</td><td>De 120 a 139</td><td>o de 80 a 89</td></tr>
<tr><td>Estadio 1 HTA</td><td>De 140 a 159</td><td>o de 90 a 99</td></tr>
<tr><td>Estadio 2 HTA</td><td>Mayor o igual que 60</td><td>o mayor o igual que100</td></tr>
</table>
</td></tr>
<tr><td>Hipertensión arterial</td><td>Hipertensos conocidos (antes del diagnóstico médico) Hipertensos diagnosticados en el estudio</td></tr>
<tr><td>Diabetes mellitus</td><td>Diabéticos conocidos (antes del diagnóstico médico) Personas aparentemente sanas con hiperglucemia Mayor o igual a 7 mmol/L [3] (126 mg/dL [4]) [17] detectada durante el estudio</td></tr>
<tr><td>Glucosuria</td><td>Prueba positiva: nivel Mayor o igual a 1+ (50 mg/dL o 2.8 mmol/L) [18] Combur 5 Test (Roche, Alemania)</td></tr>
<tr><td>Glucosa en ayunas alterada (GAA o pre-diabetes)</td><td>Participantes aparentemente sanos con glucosa en ayunas de 100 a 125 mg/dL (De 5.6 a 6.9 mmol/L) [17]</td></tr>
<tr><td>Estado nutricional</td><td>Peso bajo menor a 18.5 kg/m^2 [5]</td></tr>
</table>

[1] Milímetros de mercurio
[2] Clasificación de la hipertensión arterial según Joint National Committee on the Prevention, Detection, Evaluation, and Treatment of High Blood Pressure-2003
[3] Milimol por litro
[4] Miligramos por decilitro
[5] Kilogramo por metro al cuadrado

Índice Quetelet [19] o Índice de Masa Corporal (IMC)	Peso normal de 18.5 a 24.9 kg/m^2 Sobrepeso de 25 a 29.9 kg/m^2 Obesidad Mayor o igual que 30 kg/m^2
Obesidad abdominal Circunferencia abdominal (cm) [19]	Masculina Mayor a 102 cm [6] Femenina Mayor a 88 cm
Dislipidemia	Colesterol total Mayor a 240 mg/dL y/o LDL Mayor a 160 mg/dL y/o HDL Menor a 35 mg/dL(hombres) y menor a 39 mg/dL(mujeres) y/o triglicéridos plasmáticos Mayor a 150 mg/dL [20]
Síndrome metabólico (SM)	Presencia de 3 o más normas (criterios) de la OMS [20]: obesidad central, disminución de HDL, hipertrigliceridemia, HTA, glucosa en ayunas alterada
Marcadores de daño renal y vascular en la Orina	Proteinuria, hematuria, proteinuria-hematuria, microalbuminuria Detectados en la fase confirmatoria y que se mantienen positivos en 2 de 3 pruebas, con al menos 3 meses de intervalo
Proteinuria	Prueba positiva: mayor o igual que 1+(30 mg/dL o 0.3 g/L [7]) [18] Combur 5 Test (Roche, Alemania)
Hematuria	Prueba positiva: mayor o igual que 1+(de 5 a 10 eritrocitos/SL) [18] Combur 5 Test (Roche, alemania)
Proteinuria con hematuria	Prueba positiva: ambas pruebas mayor o igual que 1+
Microalbuminuria	Prueba positiva: mayor o igual que 1+(20 mg/L) [18] Micral Test (Roche, Alemania)
Enfermedad renal crónica Clasificación KDOQI 2002 [3]	Tasa de filtración glomerular (TFG) menor a 60 mL/min/1.73 m^2 [8] o Tasa de filtración glomerular mayor o igual que 60 mL/min/1,73 m^2 y marcadores de daño renal Estadios 1 y 2: persistencia demarcadores de daño renal y vascular durante al menos 3 meses; ensayo de creatinina enzimática (Roche, Alemania) y estimado de la TFG utilizando la fórmula MDRD [21]. Estadios 3, 4, and 5: (IRC) diagnóstico por la TFG según los niveles promedio de creatinina obtenidos durante al menos 3 meses con o sin presencia de marcadores de daño
Enfermedad renal crónica no diabética	ERC no asociada con diabetes o con diabetes, pero sin microalbuminuria ni proteinuria [22]
Enfermedad renal crónica diabética	ERC asociada con diabetes con microalbuminuria urinaria que perdura≥3 meses [22]

Procedimientos Registro y codificación. A cada paciente se le asignó un número de registro y un código para el seguimiento clínico posterior [2].

Historia y examen clínicos. Éstos se realizaron para elaborar la historia clínica personal (datos personales, antecedentes personales y familiares de la enfermedad, factores de riesgo ambientales, ocupacionales y de conducta) y mediciones físicas (estatura, peso, presión arterial, circunferencia abdominal) [2].

Exámenes de Laboratorio Clínico. La primera muestra de orina de la mañana se analizó por medio de tiras Combur 5 Test y Micral Test y el lector de tiras para

[6] Centímetros
[7] Gramos por litro
[8] Mililitros de sangre sobre minuto dada la superficie corporal del cuerpo de 1.73 por metro cuadrado

muestras de orina URISYS (Roche Diagnostics, Alemania). Se extrajo sangre al paciente en condiciones de ayuno para medir creatinina, glucosa, colesterol, LDL, HDL y triglicéridos en suero. Las muestras se procesaron en un laboratorio clínico instalado en la comunidad y equipado con un espectrofotómetro Cobas c111 (Roche, Alemania) y sus respectivos reactivos [2].

Control de la calidad, estandarización de los procedimientos y validación de los datos. Todos los instrumentos y herramientas de medición se calibraron para garantizar la calidad y confiabilidad. Los exámenes de laboratorio se realizaron siguiendo las indicaciones de los fabricantes utilizando los controles apropiados. Las mediciones y análisis fueron realizados por personal entrenado y certificado [2].

Consideraciones éticas Se obtuvo el consentimiento informado por escrito de todos los participantes. Los pacientes estuvieron de acuerdo con la publicación de los resultados del estudio siempre que no se revelaran sus identidades. Todos los pacientes recibieron atención médica y seguimiento a través de los servicios de salud pública [2].

A continuación se presentan resultados del estudio Nefrolempa en la Tabla 3.2 y en la Tabla 3.3:

Tabla 3.2. Prevalencia de factores de riesgo de ERC en adultos en el Bajo Lempa, El Salvador (n=775; sexo masculino: 343, sexo femenino: 432).

Factor de riesgo	Prevalencia (%)		
	Masculino	Femenino	Total
Edad mayor o igual a 60años	17.3	10.2	13.3
Historia familiar de enfermedad renal crónica	21.9	21.2	21.6
Historia familiar de diabetes mellitus	22.7	27.8	22.9
Historia familiar de hipertensión arterial	35.3	44.2	40.3
Historia personal de enfermedades renales	33.8	51.9	43.9
Hábito de fumar	29.4	1.4	13.8
Hábito de fumar anterior	31.8	5.8	17.3
Consumo de alcohol	40.5	4.6	20.5
Uso de plantas medicinales	58.9	68.1	64.0
Uso de medicamentos antiinflamatorios no esteroides	72.1	76.9	74.8
Trabajador agrícola	80.6	6.8	40.6
Contacto con agroquímicos	82.5	24.8	50.3
Historia personal de enfermedades infecciosas	80.8	91.9	86.9
Pre-hipertensión	33.5	24.5	28.5
Hipertensión arterial	23.3	11.8	16.9
Diabetes mellitus	9.9	10.6	10.3
Bajo peso	2.6	2.3	2.5
Sobrepeso	32.3	35.2	34.0
Obesidad	13.6	29.3	22.4
Obesidad abdominal	5.2	29.2	18.6
Dislipidemia	64.7	61.8	63.1
Hipercolesterolemia	19.8	25.9	23.2
LDL elevado	14.0	16.0	15.0

HDL disminuido	24.2	15.7	19.6
Hipertrigliceridemia	51.3	48.4	49.7
Síndrome metabólico	22.2	34.0	28.8

Tabla 3.3. Regresión logística múltiple de la ERC y sus factores de riesgo en adultos, Bajo Lempa, El Salvador (n=775; 343 sexo masculino, 432 sexo femenino).

VARIABLES	Coeficiente B_i	Odds ratios (OR)	Valor de p de significancia	Intervalo de Confianza del 95%	
				Límite inferior	Límite Superior
Edad	0.043	1.044	0.000	1.028	1.060
Sexo masculino	0.812	2.253	0.050	1.001	5.058
Historia familiar de ERC	0.546	1.726	0.031	1.050	2.837
Contacto con agroquímicos	0.210	1.234	0.510	0.660	2.308
Agricultor	0.301	1.352	1.352	0.634	2.883
Hipertensión arterial	1.004	2.730	0.003	1.423	5.239
Diabetes	0.226	1.254	0.509	0.641	2.450
Obesidad/peso normal	0.038	0.962	0.908	0.501	1.847
Sobrepeso/peso normal	0.272	0.762	0.408	0.399	1.453
Dislipidemia	0.023	0.977	0.939	0.542	1.761
Síndrome metabólico	0.022	1.022	0.951	0.501	2.087

3.2.2. JUSTIFICACIÓN

Se señala primeramente que en El Salvador se gastan anualmente alrededor de 15 millones de dólares para costear los gastos de TRR, la cual cubre solo a una pequeña proporción de todos los casos existentes en el país. Se estima que solo uno de cada cuatro pacientes con Enfermedad Renal Crónica Terminal (ERCT) llega a los hospitales. La enfermedad renal rara vez se identifica en su etapa de aparecimiento, puesto que los pacientes reciben atención médica en la etapa avanzada de la enfermedad y en la mayoría de casos no se hacen biopsias. Usualmente, la necesidad de TRR es evidente a los pocos meses del diagnóstico de IRC [23].

A diferencia del estudio Nefrolempa el cual fue dirigido al abordaje integral de la ERC en todos sus estadios, el presente estudio se enfoca en la IRC (estadios 3, 4 y 5), partiendo de la base de datos muestreada que se supone exhibe una realidad compleja relacionada de manera coherente con el fenómeno de esta enfermedad, dado que cuenta con cierta cantidad de variables involucradas. En tales circunstancias, se realiza una caracterización de IRC, a través de técnicas multivariantes, en donde se tiene por objeto la determinación de variables y factores de riesgo asociados con la IRC.

Por lo tanto, se espera que un análisis de este tipo genere un aporte importante, mediante evidencia científica en las soluciones buscadas en torno a esta problemática. De esta manera, con los resultados obtenidos, los investigadores tendrán evidencia científica que sirva como base para otras investigaciones y que permita a futuro diseñar estrategias efectivas para el tratamiento de la IRC. A consecuencia, podría ser posible tener evidencia para implementar medidas de prevención, detección, evaluación e intervención de la enfermedad, con lo que se garantizaría la reducción de costos en planificaciones estratégicas de proyectos futuros y así revertir la tendencia del número de personas afectadas a nivel nacional.

3.3. OBJETIVOS

3.3.1. OBJETIVO GENERAL

Realizar una caracterización de la IRC en las comunidades del Bajo Lempa de El Salvador en el periodo comprendido desde agosto hasta diciembre de 2009, mediante técnicas multivariantes.

3.3.2. OBJETIVOS ESPECIFICOS

1. Identificar asociaciones entre variables sociales, epidemiológicas y clínicas con la IRC.

2. Analizar el comportamiento de la IRC, a través de la exploración de variables sociales, epidemiológicas y clínicas.

3. Determinar factores de riesgo y asociaciones entre éstos relacionados con la IRC, según sexo y grupos de edad.

4. Presentar evidencia estadística científica, que pueda servir como modelo en otros proyectos de investigación.

4. METODOLOGÍA

Procedimiento para la recolección de información

Para desarrollar el presente análisis secundario de la IRC, el Instituto Nacional de Salud (INS) de El Salvador proporciono la base de datos del estudio Nefrolempa. Posteriormente el conjunto de datos es sometido a un proceso de selección de variables e individuos, que permite desarrollar un estudio orientado a la determinación de variables y factores de riesgo asociados con la IRC.

Tipo y diseño general del estudio

Tipo de estudio

Descriptivo: a través de análisis univariado y bivariado dirigido a la descripción de características que podrían influir en la aparición de la IRC Así como también multivariante orientado a la determinación de variables asociadas con la IRC.

Analítico: a través de la creación de un modelo que determina factores de riesgo asociados con la IRC.

Diseño general del análisis

El análisis se desarrollara en fases:

Fase 1: Preparación de la base de datos

En primer lugar, se realiza la adecuación de la base de datos para el análisis. Esto consiste en seleccionar individuos y variables que permitan desarrollar los análisis para la determinación de variables o factores de riesgo asociados con la IRC (Individuos: personas con IRC y sin ERC; Variables: las que se suponen podrían tener algún tipo de asociación coherente con el fenómeno la IRC).

1. Procedimiento para la selección de la muestra:

La base de datos del estudio Nefrolempa consta de 775 personas y 192 variables, entre las cuales se encuentran dos que permiten crear la variable dependiente, y, que se le denomina “Estado”, la cual clasifica a 76 personas con IRC y 624 sin ERC. De esta manera se determina una muestra de 700 individuos para el estudio.

2. Selección de posibles variables asociadas con la IRC:

A excepción de las dos variables con las que se forma la variable dependiente y, se tiene que de entre las restantes 190 variables disponibles se seleccionan las posibles variables explicativas (independientes), es decir, los posibles factores de riesgo o variables que se suponen podrían tener algún tipo de asociación razonable y coherente con la IRC, o en otras palabras las que se suponen podrían estar dentro del contexto o marco de estudio de la IRC.

Fase 2: Análisis Univariado de la base de datos de muestra

a) Identificación de valores ausentes en la base de datos seleccionada:

Se examinan las variables seleccionadas para el análisis de la IRC, con el objeto de identificar valores ausentes, lo cual permite si es necesario excluir de los análisis las variables que presentan una inadecuada cantidad de datos.

b) Se revisa la base de datos de muestra para cerciorarse de la presencia de variables cuantitativas, y de ser así se realiza el siguiente análisis univariado:

Mediante técnicas estadísticas univariadas se presenta una revisión descriptiva, por medio de medidas de tendencia central, como son: valor mínimo, valor máximo, media, mediana, varianza, desviación estándar, coeficiente de kurtosis, y coeficiente de asimetría, los cuales ayudan al entendimiento de las distribuciones de cada una de las variables que son mostradas mediante diagramas de caja, utilizados con la finalidad de manifestar el comportamiento y la calidad de los datos.

Verificación de datos atípicos: se presenta una revisión que consiste en identificar la presencia de datos atípicos para poder realizar una adecuada descripción de datos, en donde se tiene por objeto identificar variables con valores atípicos que pueden distorsionar de manera significativa correlaciones existentes entre variables.

c) Análisis Univariado de variables cualitativas:

Se presenta la distribución de recuentos y porcentajes de personas de cada variable cualitativa a través de tablas y gráficos de barra. En particular se muestra la prevalencia de personas con IRC y sin ERC, dada la muestra de 700 personas.

Fase 3: Análisis Bivariado

1. Si se tienen variables cuantitativas incluidas en la base de datos de muestra se realiza el siguiente análisis bivariado:

Se muestran el análisis de relación lineal por cada par de variables cuantitativas, el cual está dirigido a conocer la existencia de relaciones lineales entre más de dos variables cuantitativas.

2. Análisis Bivariado de variables cualitativas:

Se presentan recuentos y porcentajes o prevalencias de variables por sexo y aparición de IRC dada la muestra de 700 personas considerada, por medio de tablas de contingencia. Para visualizar de mejor manera las características de algunas variables se muestran gráficos de barra. Además se averigua la existencia de relaciones lineales entre la variable dependiente y denominada "Estado", que clasifica a personas con IRC y sin ERC y el resto de variables cualitativas.

Fase 4: Aplicación de un análisis descriptivo multivariante

A través de técnicas multivariantes se presenta un abordaje descriptivo que ayuda a determinar variables asociadas con la IRC.

Fase 5: Revisión de supuestos

Se revisa los supuestos y criterios básicos que se deben cumplir para la elaboración de un Modelo Regresión Logística (MRL), a través de los tipos de técnicas bivariantes y multivariantes realizadas en el presente análisis secundario de la IRC.

Fase 6: Aplicación de la Regresión Logística Binaria Múltiple

Seguidamente se presenta la aplicación de la Regresión Logística Binaria Múltiple, que permite encontrar factores de riesgo asociados con la IRC.

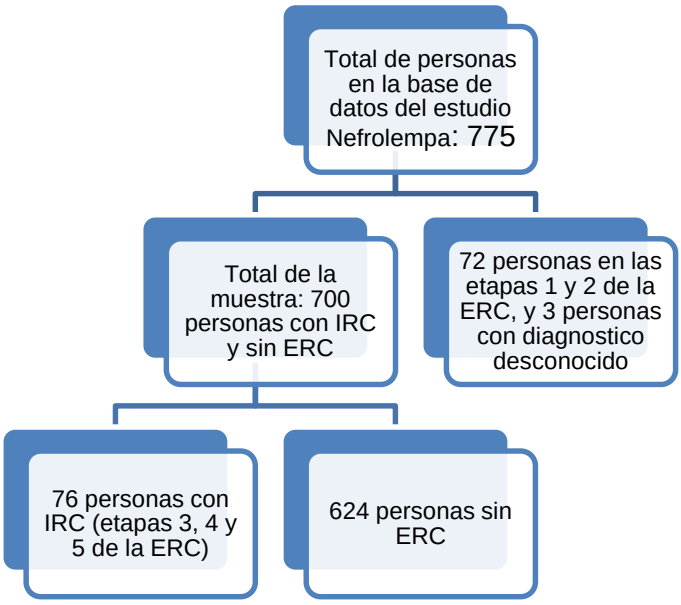
Universo de estudio

El universo de estudio está comprendido por 700 personas con edades mayores o iguales a 18 años, residentes en tres comunidades rurales del Bajo Lempa, en el municipio de Jiquilisco, El Salvador: Nueva Esperanza. Ciudad Romero y La Canoa, los cuales fueron seleccionados de la base de datos del estudio Nefrolempa (2009).

Selección y tamaño de la muestra

La base de datos del estudio Nefrolempa (2009), consta de un total de 775 registros de personas con edades mayores o iguales a 18 años, de las cuales según los resultados del estudio, 76 resultaron con diagnóstico de IRC, y 624 fueron determinadas con no padecer ERC, resultando una muestra de 700 personas para realizar los análisis. Los restantes registros no se seleccionan por que 72 de éstos se identifican con las etapas 1 y 2 de la ERC, lo cual están fuera del marco de estudio de la IRC y los otros 3 registros tienen un diagnostico desconocido. En la Figura 4.1 se logra identificar el tamaño de la muestra.

Figura 4.1. Identificación del tamaño de la muestra



Definiciones operacionales

Las variables son identificadas por los nombres de los campos de la base de datos, las cuales se les puede conocer como sociales, epidemiológicas y clínicas, que a su vez en términos estadísticos se dividen en cualitativas y cuantitativas. Las variables cualitativas son el resultado de un diseño de encuesta dirigido a observar lo social, aspectos epidemiológicos, información clínica proveniente de exámenes de laboratorio, historia de salud personal y familiar de los individuos.

Las variables cuantitativas son el resultado de la recolección y observación del historial clínico de los individuos, a través de exámenes de laboratorio. En la Tabla 4.1 se muestra la operacionalización de variables seleccionadas de la base de datos del estudio Nefrolempa.

Tabla 4.1. Operacionalización de variables para los análisis.

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
Estado	IRC_no_ERC	Estado de salud de la persona que indica si tiene IRC o no padece ERC **(Variable dependiente y)**	1=persona con IRC (IRC) 2=persona sin ERC (No ERC)
Sexo	Sexo	Sexo de la persona	1=masculino (M) 2=femenino (F)
Grupo de edades	Grupo_edades	Grupos de edades para las personas	1=18 a 29 años (18 a 29) 2=30 a 39 años (30 a 39) 3=40 a 49 años (40 a 49) 4=50 a 59 años (50 a 59) 5=60 a 69 años (60 a 69) 6=mayor o igual a 70 años (>=70)
Estado civil	Estado_civil	Estado civil de la persona	1=soltero (sol) 2=acompañado (acom) 3=casado (cas) 4=divorciado (divor) 5=viudo (viu)
Ocupación laboral	Ocupación	Ocupación, labor que realiza la persona	1=agricultor (agr) 2=agricultor y ama de casa (agrama) 3=agricultor y fumigador (agrfum) 4=ama de casa (amacas) 5=desempleado (desem) 6=estudiante (estu) 7=Enfermera (enfe) 8=Laboratorio Clínico (labcli) 9=Médico (medi) 10=Odontóloga (odon) 11=Psicóloga (psico) 12=Promotor de salud (promo) 13=otros (otros)
Nivel de escolaridad	Escolaridad	Nivel de escolaridad terminada por la persona	1=no estudia (noestu) 2=parvularia (parvu) 3=primer ciclo (pricic) 4=segundo ciclo (segcic) 5=tercer ciclo (tercic) 6=bachillerato (bachi) 7=universitario (unive)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
Fumador actual	Fumador_actual	Si la persona fuma o no, actualmente	1=si fuma actualmente (si) 2=no fuma actualmente (no)
Exfumador	Ex_fumador	Si la persona es o no es exfumador	1=si es exfumador (si) 2=no es exfumador (no)
Nunca fumo	Nunca_fumo	Si la persona nunca ha fumado o si ha fumado	1=nunca ha fumado (si) 2=si ha fumado (no)
Tiempo de fumar	Tiempo_fuma	Tiempo de fumar de una persona	1=menor de 1 año (<1año) 2=1 a 10 años (1-10años) 3=mayor de 10 años (>10años)
Alcohólico actual	Alcohólico_actual	Si la persona consume o no, alcohol actualmente	1=si consume alcohol actualmente (si) 2=no consume alcohol actualmente (no)
Exalcohólico	Ex_alcohólico	Si la persona es o no es exalcohólico	1=si es exalcohólico (si) 2=no es exalcohólico (no)
Nunca alcohol	Nunca_alcohol	Si la persona nunca ha consumido alcohol o si ha consumido alcohol	1=nunca ha consumido alcohol (si) 2=si ha consumido alcohol (no)
Tiempo de alcohólico	Tiempo_alcohol	Tiempo de consumir alcohol una persona	1=1 vez al mes (1almes) 2=1 vez por semana (1alsem) 3=1 vez al día (1aldia)
Actividad física	Actividad_física	Si la persona realiza o no, actividad física	1=si realiza actividad física (si) 2=no realiza actividad física (no)
Tiempo de actividad física	Tiempo_física	Tiempo de realizar actividad física por parte de la persona	1=menor de 30 minutos al día (<30mind) 2=30 a 60 minutos al día (30-60mind) 3=mayor de 60 minutos al día (>60mind)
Consumo de frutas y vegetales	Consumo_frutas_vegetales	Si la persona consume o no, frutas y vegetales	1=si consume frutas y vegetales (si) 2=no consume frutas ni vegetales (no)
Frecuencia de	Frecuencia_frutas	Frecuencia con la que	1=1 vez a la semana

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
consumo de frutas y vegetales		consume frutas y vegetales la persona	(1vezsem) 2= 3 a 5 veces a la semana (3-5vezsem) 3= todos los días (todosdias)
Utilización del Sistema Nacional de salud (Unidad de Salud, Hospital)	Sistema_nacional_salud	Si la persona utiliza o no, el Sistema Nacional de Salud (Unidad de Salud, Hospital)	1=si utiliza el Sistema Nacional de Salud (Unidad de Salud, Hospital) (si) 2=no utiliza el Sistema Nacional de Salud (Unidad de Salud, Hospital) (no)
Utilización del ISSS	ISSS	Si la persona utiliza o no, el ISSS	1=si utiliza el ISSS (si) 2=no utiliza el ISSS (no)
Utilización de clínica privada lucrativa	Clínica_privada	Si la persona utiliza o no, la clínica privada lucrativa	1=si utiliza clínica privada lucrativa (si) 2=no utiliza clínica privada lucrativa (no)
Utilización de clínica privada no lucrativa (ONG)	Clínica_ONG	Si la persona utiliza o no, la clínica privada no lucrativa (ONG, parroquial)	1=si utiliza clínica privada no lucrativa (si) 2=no utiliza clínica privada no lucrativa (no)
Utilización de curandero	Curandero	Si la persona utiliza o no, el curandero	1=si utiliza curandero (si) 2=no utiliza curandero (no)
Antecedentes familiares de diabetes mellitus	AFDM	Con o sin antecedentes familiares de diabetes mellitus	1=con antecedentes familiares de diabetes mellitus (si) 2=sin antecedentes familiares de diabetes mellitus (no) 3=no sabe si existen antecedentes familiares de diabetes mellitus (no sabe)
Antecedentes familiares de hipertensión arterial	AFHA	Con o sin antecedentes familiares de hipertensión arterial	1=con antecedentes familiares de hipertensión arterial (si) 2=sin antecedentes familiares de hipertensión arterial (no) 3=no sabe si existen antecedentes familiares de hipertensión arterial (no sabe)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
Antecedentes familiares de enfermedad en los riñones	AFER	Con o sin antecedentes familiares de enfermedad en los riñones	1=con antecedentes familiares de enfermedad en los riñones (si) 2=sin antecedentes familiares de enfermedad en los riñones (no) 3=no sabe si existen antecedentes familiares de enfermedad en los riñones (no sabe)
Antecedentes personales de diabetes mellitus	APDM	Con o sin antecedentes personales de diabetes mellitus	1=con antecedentes personales de diabetes mellitus (si) 2=sin antecedentes personales de diabetes mellitus (no) 3=no sabe si tiene antecedentes personales de diabetes mellitus (no sabe)
Antecedentes personales de hipertensión arterial	APHA	Con o sin antecedentes personales de hipertensión arterial	1=con antecedentes personales de hipertensión arterial (si) 2=sin antecedentes personales de hipertensión arterial (no) 3=no sabe si tiene antecedentes personales de hipertensión arterial (no sabe)
Antecedentes personales de enfermedad cerebrovascular	APEC	Con o sin antecedentes personales de enfermedad cerebrovascular	1=con antecedentes personales de enfermedad cerebrovascular (si) 2=sin antecedentes personales de enfermedad cerebrovascular (no) 3=no sabe si tiene antecedentes personales de enfermedad cerebrovascular (no sabe)
Antecedentes personales de enfermedad renal	APER	Con o sin antecedentes personales de enfermedad renal	1=con antecedentes personales de enfermedad renal (si) 2=sin antecedentes personales de enfermedad renal (no)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
Antecedentes personales de alguna enfermedad en la próstata	APEprostata	Si la persona ha padecido o padece actualmente de alguna enfermedad en la próstata, o nunca ha padecido una enfermedad de este tipo	1=con antecedentes personales de alguna enfermedad en la próstata (si) 2=sin antecedentes personales de alguna enfermedad en la próstata (no) 3=no sabe si tiene antecedentes personales de alguna enfermedad en la próstata (no sabe)
Antecedentes personales de hipertrofia prostática	APhipertrofia	Con o sin antecedentes personales de hipertrofia prostática	1=con antecedentes personales de hipertrofia prostática 2=sin antecedentes personales de hipertrofia prostática
Antecedentes personales de cáncer de próstata	APCprostata	Con o sin antecedentes personales de cáncer de próstata en la persona	1=con antecedentes personales de cáncer de próstata 2=sin antecedentes personales de cáncer de próstata
Antecedente personal de bajo peso al nacer	APBPN	Con o sin antecedente personal de bajo peso al nacer (¿Al nacer su peso fue menor de 5 libras y 3 onzas (< 2.5 kilogramos o < de 2500 grs)?)	1=con antecedente personal de bajo peso al nacer (si) 2=sin antecedente personal de bajo peso al nacer (no) 3=no sabe si tiene antecedente personal de bajo peso al nacer (no sabe)
Antecedentes de diabetes mellitus en la madre	ADMmadre	Con o sin antecedentes de diabetes en la madre	1=con antecedentes de diabetes en la madre (si) 2=sin antecedentes de diabetes en la madre (no)
Antecedentes de diabetes mellitus en el padre	ADMpadre	Con o sin antecedentes de diabetes en el padre	1=con antecedentes de diabetes en el padre (si) 2=sin antecedentes de diabetes en el padre (no)
Antecedentes de diabetes mellitus en el hermano(a)	ADMhermano	Con o sin antecedentes de diabetes en el hermano(a)	1=con antecedentes de diabetes en el hermano(a) (si) 2=sin antecedentes de diabetes en el

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
			hermano(a) (no)
Antecedentes de diabetes mellitus en el hijo(a)	ADMhijo	Con o sin antecedentes de diabetes en el hijo(a)	1=con antecedentes de diabetes en el hijo(a) (si) 2=sin antecedentes de diabetes en el hijo(a) (no)
Antecedentes de hipertensión arterial en la madre	AHAmadre	Con o sin antecedentes de hipertensión en la madre	1=con antecedentes de hipertensión en la madre (si) 2=sin antecedentes de hipertensión en la madre (no)
Antecedentes de hipertensión arterial en el padre	AHApadre	Con o sin antecedentes de hipertensión en el padre	1=con antecedentes de hipertensión en el padre (si) 2=sin antecedentes de hipertensión en el padre (no)
Antecedentes de hipertensión arterial en el hermano(a)	AHAhermano	Con o sin antecedentes de hipertensión en el hermano(a)	1=con antecedentes de hipertensión en el hermano(a) (si) 2=sin antecedentes de hipertensión en el hermano(a) (no)
Antecedentes de hipertensión arterial en el hijo(a)	AHAhijo	Con o sin antecedentes de hipertensión en el hijo(a)	1=con antecedentes de hipertensión en el hijo(a) (si) 2=sin antecedentes de hipertensión en el hijo(a) (no)
Antecedentes de enfermedad en los riñones en la madre	AERmadre	Con o sin antecedentes de enfermedad en los riñones en la madre	1=con antecedentes de enfermedad en los riñones en la madre (si) 2=sin antecedentes de enfermedad en los riñones en la madre (no)
Antecedentes de enfermedad en los riñones en el padre	AERpadre	Con o sin antecedentes de enfermedad en los riñones en el padre	1=con antecedentes de enfermedad en los riñones en el padre (si) 2=sin antecedentes de enfermedad en los riñones en el padre (no)
Antecedentes de enfermedad en los riñones en el hermano(a)	AERhermano	Con o sin antecedentes de enfermedad en los riñones en el hermano(a)	1=con antecedentes de enfermedad en los riñones en el hermano(a) (si) 2=sin antecedentes de

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
			enfermedad en los riñones en el hermano(a) (no)
Antecedentes de enfermedad en los riñones en el hijo(a)	AERhijo	Con o sin antecedentes de enfermedad en los riñones en el hijo(a)	1=con antecedentes de enfermedad en los riñones en el hijo(a) (si) 2=sin antecedentes de enfermedad en los riñones en el hijo(a) (no)
Uso de plantas medicinales	Uso_plantas_medicinales	Si la persona Ingiere o no, habitualmente algún cocimiento o infusión de yerbas o plantas medicinales	1=si usa plantas medicinales (si) 2=no usa plantas medicinales (no)
Contacto con agroquímicos	Contacto_agroquímicos	Si la persona ha tenido o tiene en la actualidad algún nivel de exposición con productos agroquímicos, o nunca ha estado expuesto a estos productos	1=si ha tenido o tiene contacto con agroquímicos (si) 2=no tiene contacto con agroquímicos (no)
Contacto con Terfos 48 EC (Clorpirifos)	Terfos	Si la persona ha tenido o tiene algún nivel de exposición con Terfos 48 EC, o nunca ha estado expuesto a este producto	1=con exposición a Terfos (si) 2=sin exposición a Terfos (no)
Contacto con Folidol (Metilparation)	Folidol	Si la persona ha tenido o tiene algún nivel de exposición con Folidol, o nunca ha estado expuesto a este producto	1=con exposición a Folidol (si) 2=sin exposición a Folidol (no)
Contacto con Tamaron 60 SL (Metamidofos-Acaricida)	Tamaron	Si la persona ha tenido o tiene algún nivel de exposición con Tamaron 60 SL, o nunca ha estado expuesto a este producto	1=con exposición a Tamaron (si) 2=sin exposición a Tamaron (no)
Contacto con Volaton (Phoxim)	Volaton	Si la persona ha tenido o tiene algún nivel de exposición con Volaton, o nunca ha estado expuesto a este producto	1=con exposición a Volaton (si) 1=sin exposición a Volaton (no)
Contacto con Terbufos 10 Gr (Terbufos-Nema)	Terbufos	Si la persona ha tenido o tiene algún nivel de exposición con Terbufos 10 Gr, o nunca ha estado expuesto a este producto	1=con exposición a Terbufos (si) 2=sin exposición a Terbufos (no)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
Contacto con Counter 10G (Terbufos)	Counter	Si la persona ha tenido o tiene algún nivel de exposición con Counter 10G, o nunca ha estado expuesto a este producto	1=con exposición a Counter (si) 2=sin exposición a Counter (no)
Contacto con Mocap 10 GR (Etroprofos)	Mocap	Si la persona ha tenido o tiene algún nivel de exposición con Mocap 10 GR, o nunca ha estado expuesto a este producto	1=con exposición Mocap (si) 2=sin exposición Mocap (no)
Contacto con Carbofurano	Carbofurano	Si la persona ha tenido o tiene algún nivel de exposición con Carbofurano, o nunca ha estado expuesto a este producto	1=con exposición a Carbofurano (si) 2=sin exposición a Carbofurano (no)
Contacto con Marshal (Carbosulfan)	Marshal	Si la persona ha tenido o tiene algún nivel de exposición con Marshal, o nunca ha estado expuesto a este producto	1=con exposición a Marshal (si) 2=sin exposición a Marshal (no)
Contacto con Semevin (Thiodicarb)	Semevin	Si la persona ha tenido o tiene algún nivel de exposición con Semevin, o nunca ha estado expuesto a este producto	1=con exposición a Semevin (si) 2=sin exposición a Semevin (no)
Contacto con Lannate (Methomyl)	Lannate	Si la persona ha tenido o tiene algún nivel de exposición con Methomyl, o nunca ha estado expuesto a este producto	1=con exposición a Lannate (si) 2=sin exposición a Lannate (no)
Contacto con Karate	Karate	Si la persona ha tenido o tiene algún nivel de exposición con Karate, o nunca ha estado expuesto a este producto	1=con exposición a karate (si) 2=sin exposición a karate (no)
Contacto con Gramoxone (Paraquat)	Gramoxone	Si la persona ha tenido o tiene algún nivel de exposición con Gramoxone, o nunca ha estado expuesto a este producto	1=con exposición a Gramoxone (si) 2=sin exposición a Gramoxone (no)
Contacto con Ranger	Ranger	Si la persona ha tenido o tiene algún nivel de exposición con Ranger, o nunca ha estado expuesto a este producto	1=con exposición a Ranger (si) 2=sin exposición a Ranger (no)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
Contacto con Roundup 35,6 SL (Roundup 35,6 SL)	Roundup	Si la persona ha tenido o tiene algún nivel de exposición con Roundup 35,6 SL, o nunca ha estado expuesto a este producto	1=con exposición a Roundup (si) 2=sin exposición a Roundup (no)
Contacto con Batalla (Glifosato)	Batalla	Si la persona ha tenido o tiene algún nivel de exposición con Batalla, o nunca ha estado expuesto a este producto	1=con exposición a Batalla (si) 2=sin exposición a Batalla (no)
Contacto con Basta (Glufosinato de amonio)	Basta	Si la persona ha tenido o tiene algún nivel de exposición con Basta, o nunca ha estado expuesto a este producto	1=con exposición a Basta (si) 2=sin exposición a Basta (no)
Contacto con Diuron	Diuron	Si la persona ha tenido o tiene algún nivel de exposición con Diuron, o nunca ha estado expuesto a este producto	1=con exposición a Diuron (si) 2=sin exposición a Diuron (no)
Contacto con Ametrina	Ametrina	Si la persona ha tenido o tiene algún nivel de exposición con Ametrina, o nunca ha estado expuesto a este producto	1=con exposición a Ametrina (si) 2=sin exposición a Ametrina (no)
Contacto con Terbutrina	Terbutrina	Si la persona ha tenido o tiene algún nivel de exposición con Terbutrina, o nunca ha estado expuesto a este producto	1=con exposición a Terbutrina (si) 2=sin exposición a Terbutrina (no)
Contacto con Gesaprin (Atrazina)	Gesaprin	Si la persona ha tenido o tiene algún nivel de exposición con Gesaprin, o nunca ha estado expuesto a este producto	1=con exposición a Gesaprin (si) 2=sin exposición a Gesaprin (no)
Contacto con Hedonal (2,4 D)	Hedonal	Si la persona ha tenido o tiene algún nivel de exposición con Hedonal, o nunca ha estado expuesto a este producto	1=con exposición a Hedonal (si) 2=con exposición a Hedonal (no)
Contacto con DDT	DDT	Si la persona ha tenido o tiene algún nivel de exposición con DDT, o nunca ha estado expuesto a este producto	1=con exposición a DDT (si) 2=sin exposición a DDT (no)
Contacto con DDD	DDD	Si la persona ha tenido o tiene algún nivel de	1=con exposición a DDD (si)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
		exposición con DDD, o nunca ha estado expuesto a este producto	2=sin exposición a DDD (no)
Contacto con Endrin	Endrin	Si la persona ha tenido o tiene algún nivel de exposición con Endrin, o nunca ha estado expuesto a este producto	1=con exposición a Endrin (si) 2=sin exposición a Endrin (no)
Contacto con Dieldrin	Dieldrin	Si la persona ha tenido o tiene algún nivel de exposición con Dieldrin, o nunca ha estado expuesto a este producto	1=con exposición a Dieldrin (si) 2=sin exposición a Dieldrin (no)
Contacto con Lindano	Lindano	Si la persona ha tenido o tiene algún nivel de exposición con Lindano, o nunca ha estado expuesto a este producto	1=con exposición a Lindano (si) 2=sin exposición a Lindano (no)
Antecedentes personales de enfermedades infecciosas, diagnosticadas por un médico	Antecedentes_infecciosas	Si la persona ha padecido o padece en la actualidad enfermedades infecciosas, diagnosticada por un médico	1=si ha padecido o padece en la actualidad (si) 2=nunca ha padecido (no)
Antecedentes personales de Parasitismo	Parasitismo	Si la persona ha padecido o padece en la actualidad Parasitismo, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Parasitismo (si) 2=nunca ha padecido Parasitismo (no)
Antecedentes personales de Filariasis	Filariasis	Si la persona ha padecido o padece en la actualidad Filariasis, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Filariasis (si) 2=nunca ha padecido Filariasis (no)
Antecedentes personales de Meningitis	Meningitis	Si la persona ha padecido o padece en la actualidad Meningitis, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Meningitis (si) 2=nunca ha padecido Meningitis (no)
Antecedentes personales de Varicela	Varicela	Si la persona ha padecido o padece en la actualidad Varicela, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Varicela (si) 2=nunca ha padecido Varicela (no)
Antecedentes personales de enfermedad de Chagas	Chagas	Si la persona ha padecido o padece en la actualidad enfermedad de Chagas, o nunca ha padecido esta	1=si ha padecido o padece en la actualidad enfermedad de Chagas (si)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
		enfermedad	2=nunca ha padecido enfermedad de Chagas (no)
Antecedentes personales de Hepatitis B	Hepatitis B	Si la persona ha padecido o padece en la actualidad Hepatitis B, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Hepatitis B (si) 2=nunca ha padecido Hepatitis B (no)
Antecedentes personales de Hepatitis C	Hepatitis C	Si la persona ha padecido o padece en la actualidad Hepatitis C, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Hepatitis C (si) 2=nunca ha padecido Hepatitis C (no)
Antecedentes personales de Amigdalitis	Amigdalitis	Si la persona ha padecido o padece en la actualidad Amigdalitis, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Amigdalitis (si) 2=nunca ha padecido Amigdalitis (no)
Antecedentes personales de Piodermitis	Piodermitis	Si la persona ha padecido o padece en la actualidad Piodermitis, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Piodermitis (si) 2=nunca ha padecido Piodermitis (no)
Antecedentes personales de Paludismo	Paludismo	Si la persona ha padecido o padece en la actualidad Paludismo, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Paludismo (si) 2=nunca ha padecido Paludismo (no)
Antecedentes personales de Tuberculosis	Tuberculosis	Si la persona ha padecido o padece en la actualidad Tuberculosis, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Tuberculosis (si) 2=nunca ha padecido Tuberculosis (no)
Antecedentes personales de Sifilis	Sifilis	Si la persona ha padecido o padece en la actualidad Sifilis, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Sifilis (si) 2=nunca ha padecido Sifilis (no)
Antecedentes personales de VIH	VIH	Si la persona ha padecido o padece en la actualidad VIH, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad VIH (si) 2=nunca ha padecido VIH (no)
Antecedentes personales de Fiebre Tifoidea	Tifoidea	Si la persona ha padecido o padece en la actualidad Fiebre Tifoidea, o nunca ha padecido esta enfermedad	1=si ha padecido o padece en la actualidad Fiebre Tifoidea (si) 2=nunca ha padecido Fiebre Tifoidea (no)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
Uso de medicamentos antiinflamatorios no esteroides (AINES)	Uso_AINES	Si la persona ingiere o no, habitualmente medicamentos para los dolores, fiebre o como antiinflamatorio	1=si usa AINES (si) 2=no usa AINES (no)
Uso de Aspirina	Aspirina	Si la persona ingiere o no, Aspirina	1=si usa Aspirina (si) 2=no usa Aspirina (no)
Uso de Flurbiprofeno	Flurbiprofeno	Si la persona ingiere o no, Flurbiprofeno	1=si usa Flurbiprofeno (si) 2=no usa Flurbiprofeno (no)
Diclofenaco (cataflán, voltarén)	Diclofenaco	Si la persona ingiere o no, Diclofenaco (cataflán, voltarén)	1=si usa Diclofenaco (cataflán, voltarén) (si) 2=no usa Diclofenaco (cataflán, voltarén) (no)
Uso de Naproxeno	Naproxeno	Si la persona ingiere o no, Naproxeno	1=si usa Naproxeno (si) 2=no usa Naproxeno (no)
Uso de Piroxicam	Piroxicam	Si la persona ingiere o no, Piroxicam	1=si usa Piroxicam (si) 2=no usa Piroxicam (no)
Uso de Ketoprofeno	Ketoprofeno	Si la persona ingiere o no, Ketoprofeno	1=si usa Ketoprofeno (si) 2=no usa Ketoprofeno (no)
Uso de Meloxicam	Meloxicam	Si la persona ingiere o no, Meloxicam	1=si usa Meloxicam (si) 2=no usa Meloxicam (no)
Uso de Ibuprofeno (advil, motrin, dorival, ..)	Ibuprofeno	Si la persona ingiere o no, Ibuprofeno(advil, motrin, dorival, …)	1=si usa Ibuprofeno (advil, motrin, dorival, …) (si) 2=no usa Ibuprofeno (advil, motrin, dorival, …) (no)
Uso de Indometacina	Indometacina	Si la persona ingiere o no, Indometacina	1=si usa Indometacina (si) 2=no usa Indometacina (no)
Uso de Acetominofen	Acetominofen	Si la persona ingiere o no, Acetominofen	1=si usa Acetominofen (si) 2=no usa Acetominofen (no)
Uso de antibióticos	Uso_antibióticos	Si la persona ha tenido o no, en los últimos doce meses tratamiento con	1=si usa antibióticos (si) 2=no usa antibióticos (no)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
		algún antibiótico	
Uso de Kanamicina	Kanamicina	Si la persona ha tenido o no, en los últimos doce meses tratamiento con Kanamicina	1=si usa Kanamicina (si) 2=no usa Kanamicina (no)
Uso de Gentamicina	Gentamicina	Si la persona ha tenido o no, en los últimos doce meses tratamiento con Gentamicina	1=si usa Gentamicina (si) 2=no usa Gentamicina (no)
Uso de Amikacina	Amikacina	Si la persona ha tenido o no, en los últimos doce meses tratamiento con Amikacina	1=si usa Amikacina (si) 2=no usa Amikacina (no)
Uso de Neomicina	Neomicina	Si la persona ha tenido o no, en los últimos doce meses tratamiento con Neomicina	1=si usa Neomicina (si) 2=no usa Neomicina (no)
Uso de Sulfadiazina	Sulfadiazina	Si la persona ha tenido o no, en los últimos doce meses tratamiento con Sulfadiazina	1=si usa Sulfadiazina (si) 2=no usa Sulfadiazina (no)
Uso de TMP-SMX	TMP-SMX	Si la persona ha tenido o no, en los últimos doce meses tratamiento con TMP-SMX	1=si usa TMP-SMX (si) 2=no usa TMP-SMX (no)
Uso de Cefaloridina	Cefaloridina	Si la persona ha tenido o no, en los últimos doce meses tratamiento con Cefaloridina	1=si usa Cefaloridina (si) 2=no usa Cefaloridina (no)
Uso de Anfotericina	Anfotericina	Si la persona ha tenido o no, en los últimos doce meses tratamiento con Anfotericina	1=si usa Anfotericina (si) 2=no usa Anfotericina (no)
Clasificación de glucosa	Clasifica_glucosa	Clasificación de glucosa como causa de la diabetes mellitus en la persona	1=glucosa normal (norm) 2=pre-diabetes (preD) 3=con diabetes mellitus (D)
Clasificación de creatinina diferenciada por sexo	Clasifica_creatinina_sexo	Clasificación de creatinina en una persona, tomando en cuenta el sexo	1=creatinina normal en femenino (normf) 2=creatinina normal en masculino (normm) 3=creatinina elevada en femenino (elevf) 4=creatinina elevada en masculino (elevm)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
Clasificación de colesterol	Clasifica_colesterol	Clasificación de colesterol en la persona	1=colesterol normal (norm) 2=limítrofe (limit) 3=colesterol alto (alto)
Clasificación de triglicéridos	Clasifica_triglicéridos	Clasificación de triglicéridos en la persona	1=triglicéridos normal (norm) 2=triglicéridos levemente alto (levalto) 3=triglicéridos elevado (elev) 4=triglicéridos muy elevado (melev)
Clasificación de HDL	Clasifica_HDL	Clasificación de HDL en la persona	1=HDL alto (alto) 2=HDL dudoso (dudo) 3=HDL bajo (bajo)
Clasificación de LDL	Clasifica_LDL	Clasificación de LDL en la persona	1=LDL óptimo (opti) 2=limítrofe bajo (libajo) 3=limítrofe alto (lialto) 4=LDL alto (alto) 5=LDL muy alto (malto)
Ocurrencia de Dislipidemia	Ocurre_dislipidemia	Si la persona tiene o no, dislipidemia	1=con dislipidemia (si) 2=sin dislipidemia (no)
Dislipidemia debida a colesterol	Dislipidemia_colesterol	Si la persona tiene dislipidemia debida a colesterol, o no tiene problemas de colesterol	1=con dislipidemia debida a colesterol (si) 2=sin problemas de colesterol (no)
Dislipidemia debida a triglicéridos	Dislipidemia_triglicéridos	Si la persona tiene dislipidemia debida triglicéridos, o no tiene problemas de triglicéridos.	1=con dislipidemia debida a triglicéridos (si) 2=sin problemas de triglicéridos (no)
Dislipidemia debida a HDL	Dislipidemia_HDL	Si la persona tiene dislipidemia debida a HDL, o no tiene problemas de HDL	1=con dislipidemia debida a HDL (si) 2=sin problemas de HDL (no)
Dislipidemia debida a LDL	Dislipidemia_LDL	Si la persona tiene dislipidemia debida a LDL, o no tiene problemas de LDL	1=con dislipidemia debida a LDL (si) 2=sin problemas de LDL (no)
Clasificación de diabetes mellitus	Clasifica_DM	Clasificación de diabetes mellitus en la persona	1=no diabetes mellitus (noD) 2=pre-diabetes (preD) 3=con diabetes mellitus (D)
Ocurrencia de diabetes mellitus	Ocurre_DM	Presencia o ausencia de diabetes mellitus en la persona	1=con diabetes mellitus (si) 2=sin diabetes mellitus

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
			(no)
Conocimiento de existencia de diabetes mellitus	Conocer_DM	Conocimiento de la existencia de diabetes mellitus en la persona	1=no diabetes mellitus (noD) 2=diabetes mellitus conocida (Dco) 3=diabetes mellitus no conocida (Dnoco)
IRC con diabetes mellitus	IRC_DM	Si la persona con IRC tiene o no diabetes mellitus, o si la persona no tiene ERC	1=no padece ERC (noERC) 2=IRC no diabética (IRCnoD) 3=IRC diabética (IRCD)
Tratamiento de diabetes mellitus	Tratamiento_DM	Si la persona diabética tiene tratamiento o no, para la diabetes mellitus, o si la persona no tiene diabetes	1=no diabetes mellitus (noD) 2=con tratamiento de diabetes mellitus (traD) 3=sin tratamiento de diabetes mellitus (notraD)
Uso de insulina para diabetes mellitus	DM_insulina	Si la persona usa insulina para la diabetes mellitus, o no usa insulina	1=usa insulina para la diabetes mellitus (si) 2=no usa insulina (no)
Uso de hipogporal para diabetes mellitus	DM_hipogporal	Si la persona usa hipogporal para la diabetes mellitus, o no usa hipogporal	1=usa hipogporal para la diabetes mellitus (si) 2=no usa hipogporal (no)
Tratamiento con dieta para diabéticos	DM_dieta	Si la persona tiene tratamiento con dieta para la diabetes mellitus, o no tiene tratamiento con dieta	1=si tiene tratamiento con dieta para la diabetes mellitus(si) 2=sin tratamiento con dieta (no)
Tratamiento combinado para diabetes mellitus	DM_combinado	Si la persona tiene tratamiento combinado para la diabetes mellitus, o no tiene tratamiento combinado	1=con tratamiento combinado para la diabetes mellitus (si) 2=sin tratamiento combinado (no)
Clasificación de hipertensión arterial (JNC7 - 2003)	HAJNC7	Clasificación de Tensión arterial en la persona, por medio de la clasificación JNC7–2003	1=tensión arterial normal (norm) 2=pre-hipertensión (preHA) 3=Estadio 1 de hipertensión arterial (HA1) 4=Estadio 2 de hipertensión arterial (HA2)
Hipertensión arterial sistólica	HA_sistólica	Si la persona tiene o no, hipertensión arterial	1=con hipertensión arterial sistólica (si)

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
		sistólica	2=sin hipertensión arterial sistólica (no)
Hipertensión arterial diastólica	HA_diastólica	Si la persona tiene o no, hipertensión arterial diastólica,	1=con hipertensión arterial diastólica (si) 2=sin hipertensión arterial diastólica (no)
Hipertensión arterial sistólica aislada	HA_sistólica_aislada	Si la persona tiene o no, hipertensión arterial sistólica aislada	1=con hipertensión arterial sistólica aislada (si) 2=sin hipertensión arterial sistólica aislada (no)
Obesidad central	Obesidad_central	Presencia o ausencia de obesidad central en la persona	1=con obesidad central (si) 2=sin obesidad central (no)
Síndrome metabólico (SM)	SM	Si la persona presenta 3 o más normas (criterios) de la OMS: obesidad central, disminución de HDL, hipertrigliceridemia, hipertensión arterial, glucosa en ayunas alterada. Se dice que sufre de SM, de lo contrario está exento de este padecimiento	1=con síndrome metabólico (si) 2=ausencia de síndrome metabólico (no)
Clasificación de IMC	Clasifica_IMC	Clasificación de IMC en la persona	1=peso bajo (pesob) 2=peso normal (norm) 3=sobrepeso (speso) 4=obesidad (obes)
Presencia de marcadores de daño renal	Marcadores	Si la persona tiene o no, marcadores de daño renal, o si tiene anuria	1=con marcadores de daño renal (mar) 2=sin marcadores de daño renal (nomar) 3=con anuria (anur)
Ocurrencia de hemoglobina (Hb)	Ocurre_Hb	Si la persona tiene o no, anemia o hemoglobina (Hb<11g/dl)	1=con hemoglobina (si) 2=sin hemoglobina (no)
Clasificación KDOQUI de la IRC	KDOQUI	Clasificación según KDOQUI de estadios de la IRC en la persona	1=estadio 3a (est3a) 2=estadio 3b (est3b) 3=estadio 4 (est4) 4=estadio 5 (est5)
Clasificación de dislipidemia	Clasifica_dislipidemia	Clasificación de dislipidemia en la persona	1=no dislipidemia (nodis) 2=dislipidemia debida a colesterol (col) 3=dislipidemia debida a

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
			colesterol y HDL (col+hdl) 4=dislipidemia debida a colesterol, HDL y LDL (col+hdl+ldl) 5=dislipidemia debida a colesterol y LDL (col+ldl) 6=dislipidemia debida a colesterol, LDL y triglicéridos (col+ldl+tri) 7=dislipidemia debida a colesterol, LDL, triglicéridos y HDL (col+ldl+tri+hdl) 8=dislipidemia debida a colesterol y triglicéridos (col+tri) 9=dislipidemia debida a colesterol, triglicéridos y HDL (col+tri+hdl) 10=dislipidemia debida a HDL (hdl) 11=dislipidemia debida a LDL (ldl) 12=dislipidemia debida a triglicéridos (tri) 13= dislipidemia debida a triglicéridos y HDL (tri+hdl) 14= dislipidemia debida a triglicéridos y LDL (tri+ldl) 15= dislipidemia con triglicéridos, LDL y HDL (tri+ldl+hdl)
Comunidad	Comunidad	Comunidad de procedencia de la persona	1=Nueva Esperanza (Espe) 2=Ciudad Romero (Rome) 3=La Canoa (Cano)
Nivel de glucosa	Glucosa	Nivel de Glucosa en miligramos sobre decilitros en la persona	Numérica: miligramos sobre decilitros (mg/dL)
Nivel de hemoglobina	Hb	Nivel de Hemoglobina en gramos sobre decilitros en la persona	Numérica: gramos sobre decilitros (g/dL)
Nivel de creatinina	Creatinina	Nivel de Creatinina en miligramos sobre decilitros en la persona	Numérica: miligramos sobre decilitros (mg/dL)
Estimación de filtrado glomerular	Filtrado	Estimación de Filtrado Glomerular en mililitros de sangre sobre minuto dada la superficie corporal del	Numérica: mililitros de sangre sobre minuto dada la superficie corporal del cuerpo de

Nombre de variable	Código de variable	Descripción de variable	Valores de variable
		cuerpo de 1.73 metros cuadrados, en la persona	1.73 metros cuadrados (ml/min/1.73m^2)
Nivel de colesterol	Colesterol	Niveles de Colesterol en miligramos sobre decilitros en la persona	Numérica: miligramos sobre decilitros (mg/dL)
Nivel de triglicéridos	Triglicéridos	Nivel de Triglicéridos en miligramos sobre decilitros en la persona	Numérica: miligramos sobre decilitros (mg/dL)
Nivel de HDL	HDL	Nivel de HDL en miligramos sobre decilitros en la persona	Numérica: miligramos sobre decilitros (mg/dL)
Nivel de LDL	LDL	Nivel de HDL en miligramos sobre decilitros en la persona	Numérica: miligramos sobre decilitros (mg/dL
Peso	Peso_kg	Peso en kilogramos de la persona	Numérica: kilogramos (kg)
Talla	Talla_cm	Talla o estatura en centímetros de la persona	Numérica: centímetros (cm)
Medida de cintura	Cintura_cm	Medida de la cintura en centímetros de la persona	Numérica: centímetros (cm)
Medida de cadera	Cadera_cm	Medida de la cadera en centímetros de la persona	Numérica: centímetros (cm)
Promedio de sistólica	Media_sistólica	Promedio de sistólica 1 y sistólica 2 en milímetros de mercurio en la persona	Numérica: milímetros de mercurio (mmHg)
Promedio de diastólica	Media_diastólica	Promedio de diastólica 1 y diastólica 2 en milímetros de mercurio en la persona	Numérica: milímetros de mercurio (mmHg)
Índice de masa corporal (IMC)	IMC	Índice de masa corporal en kilogramos sobre metros cuadrados de la persona	Numérica: kilogramos sobre metros cuadrados (kg/m^2)
Número de agroquímicos	Número_químicos	Número de agroquímicos utilizados por la persona	Numérica: cantidad de agroquímicos
Número de AINES	Número_AINES	Número de AINES utilizados por la persona	Numérica: cantidad de AINES
Número de criterios según OMS	Criterios	Número de criterios (normas) según la OMS concernientes al SM, presentes en la persona	Numérica: número de criterios presentes en la persona.

5. MARCO TEÓRICO

5.1. MARCO TEÓRICO DE LA INSUFICIENCIA RENAL CRÓNICA

En el año 2002 se publicó la definición y clasificación de la enfermedad renal crónica independiente de la causa de la enfermedad. Esta clasificación en cinco estadios facilita la puesta en marcha de planes de acción en cuanto al cuidado de la enfermedad renal crónica, con el desarrollo de guías diagnósticas y recomendaciones terapéuticas [3].

Definición de la ERC: Se define la ERC como la disminución de la función renal, en función de un Filtrado Glomerular (FG) mayor a 60 ml/min/1,73 m^2, o también como la presencia de daño estructural y/o funcional del riñón, diagnosticada por un método directo (alteraciones anatomo-patológicas en la biopsia renal) o de forma indirecta mediante la presencia de marcadores en orina: albúminuria o proteinuria, hematuria; o en las pruebas de imagen: hidronefrosis, riñones pequeños; o en sangre: creatinina sérica elevada, alteraciones acido-base entre otras, de forma persistente durante al menos 3 meses [3].

Evaluación de la función renal: La evaluación de la función renal debe ser realizada de forma ideal por medio del Filtrado Glomerular. Múltiples métodos son utilizados para eso, desde la medición del aclaramiento de inulina, que se considera la regla de oro, hasta métodos radioisotópicos como el 125 I-iodotalamato, 99m Tc-DTPA y el 51 Cr-EDTA, que son métodos precisos, pero caros, de difícil realización técnica para ser aplicados en la práctica clínica diaria frente a la masividad de esta enfermedad. La cistatina C, se ha utilizado, pero, de acuerdo con sus resultados, no existe un consenso para su aplicación en la práctica clínica. La medición de la creatinina sérica no es un parámetro que de forma aislada sirva para cuantificar la función renal, sino a partir de ésta estimarla por aclaramiento de esa sustancia, recolectando orina de 24 horas, o a partir de ésta aplicando ecuaciones matemáticas.

Por estar influenciado el metabolismo de la creatinina por factores como la masa muscular, la edad, el sexo y la raza, entre otros, para una misma cifra de creatinina, distintas personas pueden tener diferente función renal. La recolección de la orina de 24 horas es engorrosa e insegura y no aporta mayor fidelidad que las fórmulas matemáticas, quedando lo primero limitado a las condiciones extremas de edad o superficie corporal, malnutrición extrema u obesidad, amputaciones, cuadriplejías o paraplejías, enfermedades musculoesqueléticas y dieta vegetariana, entre otras. [24]

En la actualidad, dentro de múltiples ecuaciones, el Filtrado Glomerular es calculado usando frecuentemente la formula Modification Diet Renal Disease (MDRD): [21, 25]

$$MDRD\ abreviada$$
$$= \begin{cases} 186 \times creatinina\ sérica(mg/dl)^{-1.154} \times edad^{-0.203} \times 0.742 \times 1.212, si\ es\ mujer\ y\ de\ raza\ negra \\ 186 \times creatinina\ sérica(mg/dl)^{-1.154} \times edad^{-0.203} \times 0.742, si\ es\ mujer\ y\ no\ es\ de\ raza\ negra \\ 186 \times creatinina\ sérica(mg/dl)^{-1.154} \times edad^{-0.203} \times 1.212, si\ es\ hombre\ y\ es\ de\ raza\ negra \\ 186 \times creatinina\ sérica(mg/dl)^{-1.154} \times edad^{-0.203}, si\ solamente\ es\ hombre \end{cases}$$

Definición de la IRC: La insuficiencia renal crónica está definida por una disminución de la función renal expresada por una Tasa de Filtración Glomerular (TFG) mayor a 60mL/min/1.73 m^2 de superficie corporal de forma persistente durante al menos 3 meses [3]. Corresponden a los estadios 3, 4 y 5 de ERC.

Clasificación de la ERC: La ERC se clasifica en 5 estadios, según el valor del FG. Para los estadios 1 y 2 se requiere la presencia de marcadores de daño renal persistentes durante al menos 3 meses. Para la IRC (estadios 3,4 y 5) un FG mayor a 60 ml/min/1,73 m^2 es criterio de diagnóstico suficiente [3].

Esta clasificación es útil para determinar el grado de severidad del daño del riñón, definir las acciones de intervención apropiadas para cada estadio y evaluar la efectividad de las mismas y además para evaluar la progresión de la enfermedad [24].

Actualmente, se ha propuesto modificar la clasificación de la ERC basada en 3 aspectos: subdivisión del estadio 3 en estadio 3b (FG: de 30 a 44 ml/min/1,73 m^2) y estadio 3a (FG de 45 a 59 mL/min/1,73 m^2), en la Tabla 5.1 se muestra esta clasificación [26].

Tabla 5.1. Descripción de clasificación de la ERC.

Población con riesgos y Estadios de la ERC	Filtrado Glomerular (mL/min/1.73 m^2)	Definición
Población aparentemente sana	Mayor de 90 sin factores de riesgo renal	No ERC
Individuos con riesgo incrementado	Mayor de 90 con factores de riesgo renal	No ERC
Estadio 1	Mayor de 90	ERC (estadios 1 y 2): presencia de marcadores de daño renal: Proteinuria, Hematuria, Microalbuminuria, Proteinuria-Hematuria. Persistentes durante al menos 3 meses.
Estadio 2	De 60 a 89	
Estadio 3ª	De 45 a 59	IRC (estadios 3, 4 y 5): definida por una disminución de la función
Estadio 3b	De 30 a 44	

Estadio 4	De 15 a 29 (severa-grave disminución del FG)	renal expresada por una Tasa de Filtración Glomerular (TFG) menor de 60 ml/min/1.73 m^2 de forma persistente durante al menos 3 meses.
Estadio 5	Menor de 15 (diálisis)	

Factores de riesgo para el desarrollo de la ERC: Entre los factores de riesgo que pueden contribuir a la ERC se distingue entre aquellos que incrementan la susceptibilidad, los que inician directamente la enfermedad, y los que causan empeoramiento del daño renal y aceleran la declinación de la función renal, tal como se muestra en la Tabla 5.2 [27].

Tabla 5.2. Descripción de factores de riesgo para el desarrollo de la ERC.

Tipo	Mecanismos	Factores de riesgo
Factores de susceptibilidad	Susceptibilidad incrementada al daño renal.	Edad avanzada
		Historia familiar de enfermedad ERC
		Reducción de la masa nefronal
		Bajo peso al nacer
		Factores raciales
		Bajo ingreso económico
		Bajo nivel educacional.
Factores de iniciación	Inician directamente el daño renal.	Diabetes Mellitus
		Hipertensión Arterial
		Enfermedades autoinmunes
		Infecciones Sistémicas
		Infecciones de las vías urinarias
		litiasis renal
		Obstrucción del tracto urinario bajo
		Toxicidad por drogas
		Metales pesados
		Sustancias químicas del medio ambiente
Factores de progresión	Causan empeoramiento del daño renal y aceleran la declinación de la función renal.	Enfermedades hereditarias
		Proteinuria (proteínas en orina)
		Presión arterial alta
		Hiperglucemia
		Dislipidemia (colesterol alto)

Posibles factores de riesgo asociados con la IRC, según el estudio Nefrolempa: [28]

- ✓ Sexo
- ✓ Edad categorizada
- ✓ Antecedentes familiares de ERC

- ✓ Diabetes mellitus
- ✓ Hipertensión arterial
- ✓ Síndrome metabólico
- ✓ Dislipidemia
- ✓ Hábito de fumar
- ✓ Contacto con plantas medicinales
- ✓ Enfermedades infecciosas
- ✓ Consumo de antiinflamatorios no asteroideos
- ✓ Contacto con agroquímicos
- ✓ Consumo de alcohol
- ✓ Ocupación
- ✓ Obesidad
- ✓ Antecedentes familiares de HTA
- ✓ Antecedentes familiares de DM

Definiciones de variables del estudio Nefrolempa concernientes con los análisis de IRC:

Insuficiencia Renal Crónica No Diabética (IRCND): es la ERCND con disminución de la función renal expresada por una TFG mayor a 60 mL/min/1.73 m^2 de superficie corporal de forma persistente durante al menos 3 meses.

Insuficiencia Renal Crónica Diabética (IRCD): es la ERCD con disminución de la función renal expresada por una TFG mayor a 60 mL/min/1.73 m^2 de superficie corporal de forma persistente durante al menos 3 meses.

Antecedente familiar de ERC, Diabetes Mellitus e Hipertensión arterial: referido por la persona entrevistada.

Bajo peso al nacer: referido por el paciente previamente informado por un médico, cuando el peso es menor de 2500 gramos.

Escolaridad terminada: parvularia, primer ciclo, segundo ciclo, tercer ciclo, bachillerato, universitario, otros estudios, no estudia.

Pacientes con Diabetes Mellitus (DM): diabéticos conocidos (pacientes con diagnóstico previo realizado por un médico); y personas aparentemente sanas con hiperglucemia mayor o igual a 7 mili mol sobre litros (mmol/L) (126 mg/dL), detectadas en el estudio [17].

Alteración de la Glucosa en Ayunas (AGA o pre-diabetes): pacientes aparentemente sanos que durante el estudio presenten cifras de glucemia en ayunas de 100 a 125 mg/dL (de 5.6 a 6.9 mmol/L).

Tabla 5.3. Criterios de diagnósticos de Diabetes Mellitus, utilizando muestras de sangre en ayuno (18 mg/dl = 1 mmol/L). ADA. Diabetes Care 27:S5-S10, 2004.

Glucemia en ayunas	Valor mmol/L(mg/dL)
Normal	Mayor a 5.6 (100)
Alteración de la Glucosa en Ayunas (AGA o pre-diabetes)	De 5.6 a 6.9 (de 100 a 125)
Diagnóstico provisional DM	Mayor o igual a 7 (126)

Pacientes con Hipertensión Arterial: hipertensos conocidos (pacientes con diagnóstico previo por un médico); e hipertensos diagnosticados en el estudio.

Tabla 5.4. Clasificación de la hipertensión arterial según Joint National Committee on the Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC7).

Categoría	Presión arterial Sistólica en milímetros de mercurio (mmHg)	Presión arterial Diastólica (mmHg)
Normal	Mayor a 120	Mayor a 80
Pre-hipertensión	De 120 a 139	De 80 a 89
HTA: estadio 1	De 140 a 159	De 90 a 99
HTA: estadio 2	Mayor o igual a 160	Mayor o igual a 110

Antecedente personal de enfermedad cardiovascular: pacientes diagnosticados previamente por un médico.

Antecedentes de enfermedad cerebrovascular: pacientes diagnosticados previamente por un médico.

Enfermedad prostática: referida por el paciente, previamente diagnosticada por un médico: Prostatitis, hiperplasia prostática benigna, cáncer de próstata.

Enfermedades Infecciosas: referidas por el paciente durante la entrevista médica.

Diagnóstico presuntivo de Infección del Tracto urinario (ITU): personas de cualquier edad o sexo con o sin antecedentes personales de ITU con nitritos en orina detectados durante el estudio.

Consumo de medicamentos analgésicos antiinflamatorios no esteroideos (AINES) y antibióticos: referidos por el paciente durante la entrevista médica.

Consumo de Plantas medicinales: referidas por el paciente durante la entrevista médica.

Contacto con agroquímicos: referido durante la entrevista médica.

Condición nutricional: mediante el índice de masa corporal (IMC) = kg/m^2 . La población fue evaluada de acuerdo a los siguientes parámetros: bajo peso: menor a 18.5 kg/m^2; peso normal: de 18.5 a 24.9 kg/m^2; sobrepeso: de 25 a 29.9 kg/m^2; obesidad: mayor o igual que 30 kg/m^2.

Obesidad central: circunferencia abdominal, mayor a 102 cm en hombres y mayor a 88 cm en mujeres.

Síndrome Metabólico (SM): el diagnóstico se estableció cuando están presentes 3 o más de los siguientes factores descritos en Tabla 5.5: [29]

Tabla 5.5. Diagnóstico del Síndrome Metabólico según el National Cholesterol Education Program – Adult Treatment Panel III (JAMA 2001; 285:2486-97).

DIAGNÓSTICO DEL SÍNDROME METABÓLICO, *3 o más de los siguientes factores*	
Obesidad abdominal (Circunferencia Abdominal)	Hombres, mayor a 102 cm
	Mujeres, mayor a 88 cm
HDL	Hombres, mayor a 40 mg/dl
	Mujeres, mayor a 50 mg/dl
Triglicéridos	Mayor o igual a 150 mg/dl
Presión Arterial	Mayor o igual a 130 / Mayor o igual a 85 mm Hg
Glicemia	Mayor o igual a 110 mg/dl

Dislipidemia: Colesterol Total mayor a 240 mg/dL, y/o LDL mayor a 160 mg/dL y/o HDL (menor de 35 mg/dL en hombres y 39 mg/dL en mujeres) y/o Triglicéridos plasmáticos (superior a 150 mg/dL).

Hábito de fumar: fumadores actuales y por tiempo de exposición, ex fumadores y tiempos de exposición.

Determinación de Hemoglobina (Hb): a pacientes con diagnóstico de IRC confirmado. Valores Normal: Hombres: de 14.0 a 17.7 g/dl; Mujeres: de 12.3 a 15.3 g/dl. Se consideró anemia cuando en la concentración de Hb, sea menor de 11 g/L en hombres adultos y mujeres postmenopáusicas y menor de 10 en mujeres premenopausicas [30].

Tabla 5.6. Niveles séricos del lipidograma según la National Cholesterol Education Program Adult Treatment Panel III.

TIPO DE LÍPIDO	NIVEL SÉRICO (mg/dL)	
Colesterol total	Mayor a 200	Deseable
	De 200 a 239	Limítrofe
	Mayor a 240	Alto
LDL colesterol	Mayor a 100	Óptimo
	De 100 a 129	Limítrofe bajo
	De 130 a 159	Limítrofe alto
	De 160 a 189	Alto
	Mayor a 190	Muy alto
HDL colesterol	Mayor a 40	Bajo
	Mayor a 60	Alto
Triglicéridos	Mayor a 150	Normal
	De 150 a 199	Levemente
	De 200 a 499	Elevados
	Mayor a 500	Muy elevados

5.2. SÍNTESIS TEÓRICA DE ESTADÍSTICA UNIVARIANTE

Si se trata de describir datos multivariantes se supone estudiar anticipadamente cada variable aisladamente [31].

5.2.1. VARIABLES CUANTITATIVAS

El estudio univariante de la variable escalar (no multivariante o no vectorial) cuantitativa x_j implica calcular su media: [31]

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

donde $i = 1, 2, \ldots, n$ son los elementos o datos en la variable, j es el número que identifica a una variable en particular, que para una **variable binaria** es la frecuencia relativa de aparición del atributo y para una numérica es el centro de gravedad o geométrico de los datos. Se calcula una medida de variabilidad con relación a la media, promediando las desviaciones entre los datos y su media.

Si definimos las desviaciones mediante $d_{ij} = (x_{ij} - \bar{x}_j)^2$, donde el cuadrado se toma para prescindir del signo, se define la desviación típica por: [31]

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} d_{ij}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n}}$$

y su cuadrado es la varianza:

$$s_j^2 = \sum_{i=1}^{n} \frac{d_{ij}}{n}.$$

Para comparar la variabilidad de distintas variables conviene construir medidas de variabilidad relativa que no dependan de las unidades de medida. Una de estas medidas es el *coeficiente de variación*: para una variable x_j seria,

$$CV_j = \sqrt{\frac{s_j^2}{\bar{x}_j^2}}$$

donde de nuevo se toman los cuadrados para prescindir del signo y suponemos que $\bar{x}_j$ es distinto de cero. En tercer lugar, conviene calcular los *coeficientes de asimetría*, que miden la simetría de los datos respecto a su centro, y que se calculan como: [31]

$$A_j = \frac{1}{ns_j^3}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^3$$

Este coeficiente es cero para una variable simétrica. Cuando el valor absoluto del coeficiente es aproximadamente mayor que uno podemos concluir que los datos tienen una distribución claramente asimétrica [31].

Una característica importante de un conjunto de datos es su homogeneidad. Si las desviaciones d_{ij} son muy distintas esto sugiere que hay datos que se separan mucho de la media y que tenemos por tanto alta heterogeneidad. Una posible medida de homogeneidad es la varianza de las d_{ij}, dada por: [31]

$$\frac{1}{n}\sum_{i=1}^{n}(d_{ij} - s_j^2)^2$$

Ya que las medias de las desviaciones $\bar{d}_j = s_j^2$. Se calcula una medida adimensional análoga al coeficiente de variación dividiendo la varianza de las desviaciones por el cuadrado de la media de las desviaciones s_j^4, con lo que tenemos el *coeficiente de homogeneidad*, que puede escribirse,

$$H_j = \frac{\frac{1}{n}\sum_{i=1}^{n}(d_{ij} - s_j^2)^2}{s_j^4}$$

Este coeficiente es siempre mayor o igual que cero. Desarrollando el cuadrado del numerador como $\sum_{i=1}^{n}(d_{ij} - s_j^2)^2 = \sum_{i=1}^{n} d_{ij}^2 + ns_j^4 - 2s_j^2\sum_{i=1}^{n} d_{ij}$ este coeficiente puede escribirse también como: [31]

$$H_j = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^4}{ns_j^4} - 1 = K_j - 1$$

El primer miembro de esta expresión K_j, es una forma alternativa de medir la homogeneidad y se conoce como *coeficiente de kurtosis.* Como $H_j > 0$, el coeficiente de kurtosis será igual o mayor que uno. Ambos coeficientes miden la relación entre la variabilidad de las desviaciones y la desviación media. Es fácil comprobar que: [31]

1. Si hay unos pocos datos atípicos muy alejados del resto, la variabilidad de las desviaciones será grande, debido a estos valores y los coeficientes de kurtosis o de homogeneidad serán altos.

2. Si los datos se separan en dos mitades correspondientes a dos distribuciones muy alejadas entre sí, es decir, tenemos dos conjuntos separados de datos distintos, la media de los datos estará equidistante de los dos grupos de datos y las desviaciones de todos los datos serán similares, con lo que el coeficiente H_j será muy pequeño (cero en el caso extremo en que la mitad de los datos son iguales a cualquier número -α, y la otra mitad igual a α).

Un objetivo central de la descripción de datos es decidir si los datos son una muestra homogénea de una población o corresponden a una mezcla de poblaciones distintas que deben estudiarse separadamente. Un caso especialmente importante de heterogeneidad es la presencia de una pequeña proporción de observaciones atípicas (outliers), que corresponden a datos heterogéneos con el resto. La detección de estas observaciones es fundamental para una correcta descripción de la mayoría de los datos, ya que estos valores extremos distorsionan los valores descriptivos del conjunto. El coeficiente de kurtosis puede ayudar en este objetivo, ya que tomará un valor alto, mayor que 7 u 8 [31].

En el análisis inicial de los datos conviene siempre calcular la media y la mediana de cada variable. Si ambas son similares, la media es un buen indicador del centro de los datos. Sin embargo, si difieren mucho, la media puede no ser una buena medida del centro de los datos debido a: [31]

1. Una distribución asimétrica,
2. La presencia de valores atípicos (que afectaran mucho a la media y poco a la mediana)
3. Heterogeneidad en los datos.

Siempre resulta útil representar gráficamente las variables continuas mediante un histograma o un diagrama de caja. Estas representaciones ayudaran a detectar asimetrías, heterogeneidad, datos atípicos, etc [31].

5.2.2. VARIABLES CUALITATIVAS O CATEGÓRICAS (ATRIBUTOS)

Las variables o características a analizar presentan k modalidades o categorías no numéricas, exhaustivas y mutuamente excluyentes. Por ejemplo, el sexo, el estado civil o el nivel de escolaridad de un individuo. Según que las modalidades admitan, o no, una ordenación, se hablará de atributos **ordinales** (nivel de escolaridad, etapas de la tensión arterial, etc.) o atributos **nominales** (sexo, ocupación laboral, etc.) [32].

1. **Atributo nominal**

Característica cuyas modalidades no admiten ordenación. Ejemplos: sexo, estado civil, ocupación laboral, contacto con agroquímicos, etc. Para analizar estas variables se utilizan tablas estadísticas o distribuciones de frecuencias y representaciones gráficas.

Tablas de frecuencias. Tabla que recoge las modalidades o categorías de la característica y sus frecuencias, tanto en términos absolutos como relativos (porcentajes). Véase la Tabla 5.7: [32]

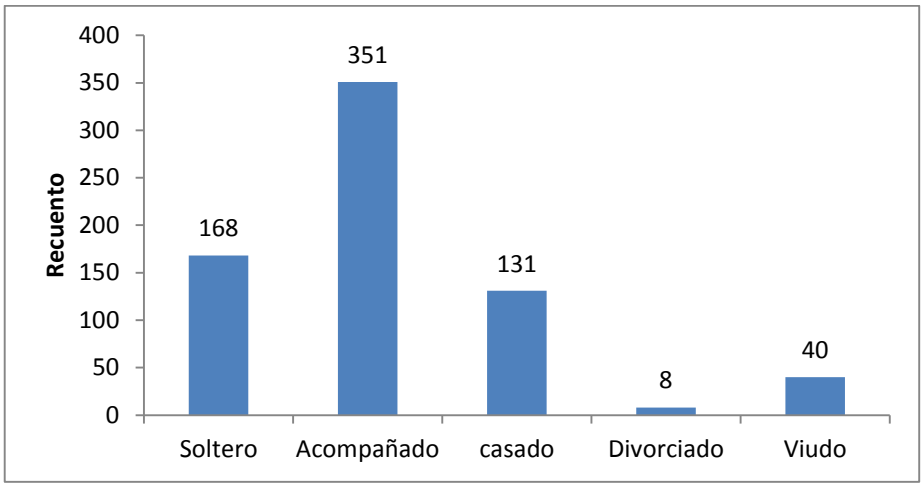
Tabla 5.7. Ejemplo de distribución de frecuencias de estado civil.

Variable: Estado civil	Recuento o frecuencia absoluta	Porcentaje
Soltero	168	24.1
Acompañado	351	50.3
Casado	131	18.8
Divorciado	8	1.1
Viudo	40	5.7
Total	698	100.0
Valores ausentes	2	0.3

Representación gráfica

Diagrama o gráfico de barras. Cada modalidad de la característica está representada por un rectángulo cuya altura corresponde a su frecuencia (absoluta o relativa). Véase Figura 5.1: [32]

Figura 5.1. Ejemplo de gráfico de barras de estado civil (frecuencias absolutas).



Resúmenes estadísticos. En este tipo de características únicamente tiene sentido la moda: modalidad que se presenta un mayor número de veces [32].

2. Atributo ordinal

Característica cuyas modalidades admiten ordenación. Ejemplos: grupos de edades, nivel de escolaridad, etapas de la tensión arterial, etc.

Tablas de frecuencias. Tabla que recoge las modalidades o categorías de la característica y sus frecuencias, tanto en términos absolutos como relativos [32].

Representación gráfica

Diagrama o gráfico de barras. Cada modalidad de la característica está representada por un rectángulo cuya altura corresponde a su frecuencia (absoluta o relativa) [32].

Resúmenes estadísticos. Aunque, en principio, la única medida con un sentido claro es la moda, en ciertas ocasiones (por ejemplo, cuando el atributo refleja actitudes, valoraciones, etc.) puede tener interés otro tipo de medidas relacionadas con el orden de los datos: mínimo, máximo, mediana, cuartiles [32].

5.3. SÍNTESIS TEÓRICA DE ESTADÍSTICA BIVARIANTE

Las técnicas estadísticas bivariantes permiten el análisis conjunto de dos características de los individuos de una población con el propósito de detectar posibles relaciones entre ellas. La naturaleza (nominal, ordinal o numérica) de las

características objeto de estudio determinará las herramientas más adecuadas para su análisis [33].

5.3.1. VARIABLES CUANTITATIVAS

Covarianza entre dos variables

Para variables escalares la variabilidad respecto a la media se mide habitualmente por la varianza, o la desviación típica. La relación lineal o dependencia lineal entre dos variables se mide por la covarianza. La *covarianza* entre dos variables (x_j, x_k) se calcula con: [31]

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Correlación entre dos variables

Se refiere a la dependencia lineal entre dos variables estudiada mediante el coeficiente de correlación lineal o simple. Este coeficiente para las variables x_j, x_k es: [31]

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

y tiene las propiedades siguientes:

(1) $0 \leq |r_{jk}| \leq 1$;

(2) si existe una relación lineal exacta entre las variables, $x_{ij} = a + bx_{ik}$, entonces $|r_{jk}| = 1$;

(3) r_{jk} es invariante ante transformaciones lineales de las variables [31].

Representaciones gráficas

Conviene construir los diagramas de dispersión de las variables por pares. Con p variables existen $p(p-1)/2$ gráficos posibles que pueden disponerse en forma de matriz y son muy útiles para entender el tipo de relación existente entre pares de variables, e identificar puntos atípicos en la relación bivariante. En particular, estos gráficos son importantes para apreciar si existen relaciones no lineales, en cuyo caso las covarianzas pueden no ser un buen resumen de la dependencia entre dos variables [31].

Transformación lineal: estandarización univariante

Muchas propiedades importantes de los datos son independientes de las unidades de medida de las variables y no variarán si se hace un cambio de escala. Una transformación lineal importante es la estandarización univariante de las variables [31].

Llamando $\boldsymbol{x} = (x_1, \dots, x_p)$ al vector $p \times 1$ de la variable vectorial, que contiene p variables escalares, la transformación lineal se expresa como: [31]

$$\boldsymbol{y} = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{x} - \bar{\boldsymbol{x}})$$

donde la matriz $\boldsymbol{D}^{-1/2}$ es cuadrada y diagonal; es decir el número de filas es igual al número de columnas y contiene ceros en sus casillas a excepción de su diagonal principal. Entonces, la matriz con los términos: [31]

$$D^{-\frac{1}{2}} = \begin{pmatrix} s_1^{-1} & 0 & 0 \\ 0 & s_2^{-1} & \ddots & 0 \\ 0 & 0 & & s_p^{-1} \end{pmatrix}$$

convierte las variables originales $\boldsymbol{x}$, en otras nuevas variables $\boldsymbol{y}$, de media cero y varianza unidad. Cada componente del vector $\boldsymbol{x}$, x_j para $j = 1, \dots, p$, se transforma con $y_j = (x_j - \bar{x}_j)/s_j$. La matriz de varianzas y covarianzas de las nuevas variables será la matriz de correlación de las variables primitivas. Esta transformación es la estandarización univariante de las variables [31].

5.3.2. VARIABLES CUALITATIVAS

Para analizar dos variables cualitativas como, por ejemplo: estado civil y grupos de edades se utiliza una tabla de contingencia o de doble entrada, la cual reúne la distribución conjunta de frecuencias de las dos variables [33].

Tablas de contingencia. Recoge, en términos absolutos o relativos, la distribución conjunta de las dos variables de análisis [33]. Véase la Tabla 5.8.

Tabla 5.8. Ejemplo de distribución de recuentos de estado civil y grupos de edades.

Estado civil	Grupos de edades						Total
	18 a 29	30 a 39	40 a 49	50 a 59	60 a 69	>=70	
Soltero	106	23	16	15	6	2	168
Acompañado	143	89	60	34	14	11	351
Casado	17	20	38	32	16	8	131
Divorciado	1	1	3	1	2	0	8
Viudo	0	1	5	9	7	18	40
Total	267	134	122	91	45	39	698

En la Tabla 5.8, la última fila recoge lo que se conoce como distribución marginal de la variable estado civil y la última columna la distribución marginal de la variable grupos de edades [33].

Luego, dividiendo cada casilla por el total de datos, se obtiene la distribución conjunta en términos relativos, como se muestra en la Tabla 5.9 [33].

Tabla 5.9. Ejemplo de distribución de porcentajes de estado civil y grupos de edades.

Estado civil	Grupos de edades						Total
	18 a 29	30 a 39	40 a 49	50 a 59	60 a 69	>=70	
Soltero	15.2%	3.3%	2.3%	2.1%	.9%	.3%	24.1%
Acompañado	20.5%	12.8%	8.6%	4.9%	2.0%	1.6%	50.3%
cas	2.4%	2.9%	5.4%	4.6%	2.3%	1.1%	18.8%
Divorciado	0.1%	0.1%	0.4%	0.1%	0.3%	0%	1.1%
Viudo	0%	0.1%	0.7%	1.3%	1.0%	2.6%	5.7%
Total	38.3%	19.2%	17.5%	13.0%	6.4%	5.6%	100.0%

En la Tabla 5.8, al ignorar los valores de la última fila y dividir el valor de cada casilla por el total de su correspondiente fila que se ubica en la última columna, se obtiene la distribución de frecuencias relativas de la variable columna condicionada a la variable fila, o en otras palabras se obtiene la distribución de porcentajes de la variable grupos de edades dentro de la variable estado civil, tal como se muestra en la Tabla 5.10 [33].

Tabla 5.10. Ejemplo de distribución de porcentajes de grupos de edades condicionados por estado civil.

Estado civil	Grupos de edades						Total
	18 a 29	30 a 39	40 a 49	50 a 59	60 a 69	>=70	
Soltero	63.1%	13.7%	9.5%	8.9%	3.6%	1.2%	100.0%
Acompañado	40.7%	25.4%	17.1%	9.7%	4.0%	3.1%	100.0%
Casado	13.0%	15.3%	29.0%	24.4%	12.2%	6.1%	100.0%
Divorciado	12.5%	12.5%	37.5%	12.5%	25.0%	0%	100.0%
Viudo	0%	2.5%	12.5%	22.5%	17.5%	45.0%	100.0%

En la Tabla 5.8, al hacer caso omiso de los valores de la última columna y dividir el valor de cada casilla por el total de su correspondiente columna que se ubica en la última columna, se obtiene la distribución de frecuencias relativas de la variable fila condicionada a la variable columna, o en otras palabras se obtiene la distribución de porcentajes de la variable estado civil por la variable grupos de edades, tal como en la Tabla 5.11 [33].

Tabla 5.11. Ejemplo de distribución de porcentajes de estado civil condicionados por grupos de edades.

Estado civil	Grupos de edades					
	18 a 29	30 a 39	40 a 49	50 a 59	60 a 69	>=70
Soltero	39.7%	17.2%	13.1%	16.5%	13.3%	5.1%
Acompañado	53.6%	66.4%	49.2%	37.4%	31.1%	28.2%
Casado	6.4%	14.9%	31.1%	35.2%	35.6%	20.5%
Divorciado	.4%	.7%	2.5%	1.1%	4.4%	0%
Viudo	0%	.7%	4.1%	9.9%	15.6%	46.2%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

El interés del análisis de dos variables cualitativas se concreta en saber si existe asociación o dependencia entre dos características cualitativas y, de ser así, cuál es el grado y el sentido de la asociación [33]. Este escenario se expone en la teoría de análisis de correspondencias, la cual tiene un enfoque multivariante.

5.4. FUNDAMENTO TEÓRICO DE ANÁLISIS MULTIVARIANTE

5.4.1. INTRODUCCIÓN

Describir cualquier situación real, por ejemplo, las características físicas de una persona, la situación política en un país, las propiedades de una imagen, el rendimiento de un proceso, la calidad de una obra de arte o las motivaciones del comprador de un producto, los factores de riesgo que más influyen en una enfermedad, requiere tener en cuenta simultáneamente muchas variables. Para describir las características físicas de una persona podemos utilizar variables como su estatura, su peso, la longitud de sus brazos y de sus piernas, etc. [31].

Para describir la situación política de un país, variables como la existencia o no de un régimen democrático, el grado de participación política de los ciudadanos, el número de partidos y sus afiliados, etc. En el caso de una caracterización de factores de riesgo relacionados con una enfermedad como la IRC se requiere utilizar variables con datos recolectados a través de métodos epidemiológicos y clínicos. El análisis de datos multivariantes tiene por objeto el estudio estadístico de muchas variables medidas en elementos de una población. Pretende los siguientes objetivos: [31]

1. Resumir el conjunto de variables en unas pocas nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información.
2. Encontrar grupos en los datos si existen.
3. Clasificar nuevas observaciones en grupos definidos.
4. Relacionar dos conjuntos de variables.

El primer objetivo trata de la descripción de una realidad compleja donde existen muchas variables, se simplifica mediante la construcción de uno o varios índices o indicadores que la resumen. Por ejemplo, la calidad de una universidad o de un departamento se resume en unos pocos indicadores, en el caso del presente estudio, la determinación de características sociales, epidemiológicas y clínicas asociadas con la IRC se podría determinar a través de unas pocas variables.

Disponer de estos indicadores tiene varias ventajas:

1. si son pocas podemos representarlas gráficamente y comparar distintos conjuntos de datos o instantes en el tiempo;
2. simplifican el análisis al permitir trabajar con un número menor de variables;

3. si las variables indicadoras pueden interpretarse, podemos mejorar nuestro conocimiento de la realidad estudiada.

El análisis multivariante de datos proporciona métodos objetivos para conocer cuántas variables indicadoras, que a veces se denomina factores, son necesarias para describir una realidad compleja y determinar su estructura.

El segundo objetivo es identificar grupos si existen. Si observamos un conjunto de variables en empresas, esperamos que los datos indiquen una división de las empresas en grupos en función de su rentabilidad, su eficacia comercial o su estructura productiva. En particular para este estudio de la IRC, se podría esperar una separación de dos grupos de personas, uno que no tenga ERC y que este asociado o identificado con ciertas variables y otro que posea personas con IRC, el cual estuviera caracterizado o explicado por otras variables. En muchas situaciones los grupos son desconocidos a priori y queremos disponer de un procedimiento objetivo para obtener los grupos existentes y clasificar las observaciones.

Un tercer objetivo relacionado con el anterior aparece cuando los grupos están bien definidos a priori y queremos clasificar nuevas observaciones. Por ejemplo, queremos clasificar a clientes que solicitan créditos como fiables o no; y para el abordaje de la IRC, se trata de clasificar personas como enfermas de IRC o que no tienen ERC, a través de la determinación de factores de riesgo.

Para alcanzar estos tres objetivos una herramienta importante es entender la estructura de dependencia entre las variables, ya que las relaciones entre las variables son las que permiten resumirlas en variables indicadoras, encontrar grupos no aparentes por las variables individuales. Un problema distinto es relacionar dos conjuntos de variables. Por ejemplo, podemos disponer de un conjunto de variables de capacidad intelectual y otros de resultados profesionales y queremos relacionar ambos conjuntos de variables. En particular, los dos grupos de variables pueden corresponder a las mismas variables medidas en dos momentos distintos en el tiempo o en el espacio y queremos ver la relación entre ambos conjuntos. El presente estudio de la IRC se basa prácticamente en los tres primeros objetivos [31].

Las técnicas de análisis multivariante tienen aplicaciones en todos los campos científicos y comenzaron desarrollándose para resolver problemas de clasificación en Biología, se extendieron para encontrar variables indicadoras y factores en Psicometría, Marketing y las Ciencias Sociales y han alcanzado una gran aplicación en Ingeniería y Ciencias de la Computación como herramientas para resumir la

información y diseñar sistemas de clasificación automática y de reconocimiento de patrones. Algunos pocos ejemplos indicativos de sus aplicaciones en distintas disciplinas, son: [31]

- ✓ Medicina: Identificar tumores mediante imágenes digitales, determinar factores de riesgo relacionados con la IRC.
- ✓ Administración de Empresas: Construir tipologías de clientes.
- ✓ Agricultura: Clasificar terrenos de cultivo por fotos aéreas.
- ✓ Arqueología: Clasificar restos arqueológicos.
- ✓ Biometría: Identificar los factores que determinan la forma de un organismo vivo.
- ✓ Ciencias de la Computación: Diseñar algoritmos de clasificación automática.
- ✓ Ciencias de la Educación: Investigar la efectividad del aprendizaje a distancia.
- ✓ Ciencias del medio ambiente: Investigar las dimensiones de la contaminación ambiental.
- ✓ Economía: Identificar las dimensiones del desarrollo económico.
- ✓ Geología: Clasificar sedimentos.
- ✓ Historia: Determinar la importancia relativa de los factores que caracterizan los periodos pre revolucionarios.
- ✓ Ingeniería: Transmitir óptimamente señales por canales digitales.
- ✓ Psicología: Determinar los factores que componen la inteligencia humana
- ✓ Sociología y Ciencia Política: Construir tipologías de los votantes de un partido.

5.4.2. DATOS MULTIVARIANTES

Se comienza observando un conjunto de variables en un conjunto de elementos (por ejemplo, personas) de una población. La información de partida dependiendo del tipo de método estudiado puede ser de varios tipos. La más habitual es una tabla donde aparecen los valores de p variables observadas sobre n elementos. Las variables pueden ser cuantitativas, cuando su valor se exprese numéricamente, como la edad de una persona, su estatura, su peso, su nivel de glucosa, o cualitativas, cuando su valor sea un atributo o categoría, como el género, ocupación laboral, tipo de enfermedad, el municipio de nacimiento, etc. Las variables cualitativas pueden clasificarse en binarias, cuando toman únicamente dos valores posibles, como, por ejemplo: el género (mujer, hombre), Tipo de diagnóstico (persona sin ERC, persona con IRC), o generales cuando toman muchos valores posibles, como el estado civil de la persona (soltero, acompañado, casado, divorciado, viudo) [31].

Las variables binarias se pueden codificar como numéricas (por ejemplo, la variable género se convierte en numérica asignando 1 al varón y el 2 a la mujer). Las variables cualitativas con más de dos categorías pueden también codificarse numéricamente, una manera de hacerlo es convertir las categorías en variables binarias. Por ejemplo, supongamos la variable obesidad, "obesi" y para simplificar supongamos que las categorías posibles son delgadez (D), normal (N), obeso (O), sobrepeso (S). Tenemos p = 4 categorías que podemos representar con p – 1 = 3 variables binarias definidas como: [31]

a) $x_1 = 1$, si obesi = D, $x_1 = 0$ en otro caso.
b) $x_2 = 1$, si obesi = N, $x_2 = 0$ en otro caso.
c) $x_3 = 1$, si obesi = O, $x_3 = 0$ en otro caso.

La Tabla 5.12, presenta la codificación de la variable atributo "obesi" en las tres variables binarias, x_1, x_2, x_3.

Tabla 5.12. Ejemplo de codificación de variable: obesidad.

obesi	x_1	x_2	x_3
D	1	0	0
N	0	1	0
O	0	0	1
S	0	0	0

Naturalmente la variable "obesi" podría también haberse codificado dando valores numéricos arbitrarios a las categorías, por ejemplo, D = 1, N = 2, O = 3, S = 4. Cuando los atributos pueden interpretarse en función de los valores de una variable continua tiene más sentido codificarla con números que indiquen el orden de las categorías. Por ejemplo, los niveles escolares como parvularia, primer ciclo, segundo ciclo, tercer ciclo, bachillerato y universidad, tienen sentido codificarlos con los números 1, 2, 3, 4, 5, pero que solo tienen un sentido de orden [31].

La matriz de datos

Suponiendo que se observan p variables en un conjunto de n elementos. Cada una de estas p variables se denomina una variable **escalar o univariante** y el conjunto de las p variables forman una variable vectorial o multivariante. Los valores de las p variables escalares en cada uno de los n elementos pueden representarse en una matriz $\boldsymbol{X}$, de dimensiones $n \times p$, que se llama matriz de datos. Se denota por x_{ij} al elemento genérico de esta matriz, que representa el valor de la variable escalar j sobre el

individuo i. Es decir: [31] datos x_{ij} donde $i = 1, \dots, n$ representa el individuo; $j = 1, \dots, p$ representa la variable, la matriz de datos $\boldsymbol{X}$ se puede representar como: [31]

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} \cdots & x_{1p} \\ x_{21} & x_{22} \ddots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}$$

donde cada variable $\boldsymbol{x}_i'$ es un vector fila, $1 \times p$, que representa los valores de las p variables sobre el individuo i. Alternativamente, podemos representar la matriz $\boldsymbol{X}$ por columnas: [31]

$$\boldsymbol{X} = \left[\boldsymbol{x}_1 \dots \boldsymbol{x}_p\right]$$

donde ahora cada variable es un vector columna, $n \times 1$, que representa la variable escalar x_j medida en los n elementos de la población. Llamaremos $\boldsymbol{x} = (x_1, \dots, x_p)'$ a la variable multivariante formada por las p variables escalares que toma los valores particulares $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, en los n elementos observados [31].

El vector de medias

Puede calcularse, como el caso escalar, promediando las medidas de cada elemento, que ahora son vectores: [31]

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

La matriz de varianzas y covarianzas

Como se comentaba en el análisis bivariante, para variables escalares la variabilidad respecto a la media se mide habitualmente por la varianza y la relación lineal entre dos variables se mide por la covarianza.

Esta información para una variable multivariante puede representarse de forma compacta en la matriz de varianzas y covarianzas, como: [31]

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X},$$

es decir: [31]

$$\boldsymbol{S} = \begin{pmatrix} s_1^2 & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_p^2 \end{pmatrix}$$

que es una matriz cuadrada y simétrica que contiene en la diagonal las varianzas y fuera de la diagonal las covarianzas entre las variables.

Medidas de dependencia lineal

Un objetivo fundamental de la descripción de los datos multivariantes es comprender la estructura de dependencias entre las variables. Estas dependencias pueden estudiarse:

1. por pares de variables;
2. entre una variable y todas las demás;
3. entre pares de variables, pero eliminando el efecto de las demás variables;
4. entre el conjunto de todas las variables [31].

Por lo que en el análisis global de la IRC se desarrollan los 3 primeros puntos. En este apartado se expone una breve introducción de la teoría referente a (1) mientras que un resumen de (2) y (3) se considera en el marco teórico de Regresión Logística.

Dependencia por pares: La matriz de correlación. Como se decía en el análisis bivariante la dependencia lineal entre dos variables se puede estudiar también mediante el coeficiente de correlación lineal. Si se tienen muchas variables, la dependencia por pares se mide por la matriz de correlación. Se llama matriz de correlación $\boldsymbol{R}$, a la matriz cuadrada y simétrica que tiene unos en la diagonal principal y fuera de ella los coeficientes de correlación lineal por pares de variables, por lo tanto, se escribe: [31]

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ \vdots & & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

5.4.3. ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

Dado que un problema central en el análisis datos multivariantes es la reducción de dimensionalidad [31], se expone la idea básica de componentes principales que es el principio de otras técnicas multivariantes como el análisis de correspondencia y correspondencia múltiple utilizadas en la presente caracterización de la IRC.

El análisis de componentes principales tiene por objetivo: dadas n observaciones de p variables, se analiza si es posible representar o describir adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. Por lo que se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información. Por ejemplo, con variables con alta dependencia es frecuente que un pequeño número de nuevas variables (menos del 20% de las originales) expliquen la mayor parte (más del 80%) de la variabilidad original. La utilidad de esta técnica es doble: [31]

1. Permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio general p -dimensional. En este sentido componentes principales es el primer paso para identificar posibles variables latentes. o no observadas, que están generando la variabilidad de los datos.
2. Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.

Supongamos que se dispone de los valores de p-variables en n elementos de una población dispuestos en una matriz $\boldsymbol{X}$ de dimensiones $n \times p$. El problema que se desea resolver es como encontrar un espacio de dimensión más reducida que represente adecuadamente o con precisión los datos [31].

Enfoque estadístico

Representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las p variables originales por una nueva variable z_1, en la dirección de un vector $\boldsymbol{a}_1 = (a_{11}, \dots, a_{1p})'$ de norma unidad[9] que resuma óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión. La condición para prever con la mínima pérdida de información los datos observados, es utilizar la variable de máxima variabilidad. Este enfoque puede extenderse para obtener el mejor subespacio resumen de los datos de dimensión 2. Para ello se calcula el plano que mejor aproxima a los puntos. El problema se reduce a encontrar una nueva dirección definida por un vector unitario $\boldsymbol{a}_2$, que, sin pérdida de generalidad, puede tomarse

[9] Este vector representa la línea recta que mejor aproxima a los puntos y la raíz cuadrada de la suma de sus componentes es igual a 1

ortogonal a $\boldsymbol{a}_1$, y que verifique la condición de que la proyección de un punto sobre este eje 2 maximice las distancias entre los puntos proyectados. Estadísticamente esto equivale a encontrar una segunda variable z_2, incorrelada con la anterior, y que tenga varianza máxima. En general, la componente z_r $(r < p)$ tendrá varianza máxima entre todas las combinaciones lineales de las p variables X originales, con la condición de estar incorrelada con las $z_1, \dots, z_{r-1}$, previamente obtenidas [31].

Calculo del primer componente

El primer componente principal será la combinación lineal de las variables originales que tenga varianza máxima. Los valores de este primer componente en los n individuos se representarán por un vector $\boldsymbol{z}_1$, dado por: [31]

$$\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{a}_1$$

Suponiendo para simplificar, que las variables originales tienen media cero, entonces, $\boldsymbol{z}_1$ tendrá media nula. Su varianza será: [31]

$$Var(z_1) = \frac{1}{n}\boldsymbol{z}_1'\boldsymbol{z}_1 = \frac{1}{n}\boldsymbol{a}_1'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}_1 = \boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1$$

donde $\boldsymbol{S}$ es la matriz de varianzas y covarianzas de las observaciones. Se puede maximizar la varianza sin límite aumentando el módulo del vector $\boldsymbol{a}_1$. Para que esta maximización tenga solución se impone una restricción al módulo del vector $\boldsymbol{a}_1$, y, sin pérdida de generalidad, se impone que $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$. Se introduce esta restricción mediante el multiplicador de Lagrange: [31]

$$M = \boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1 - \lambda(\boldsymbol{a}_1'\boldsymbol{a}_1 - 1)$$

y se maximiza esta expresión de la forma habitual derivando respecto a los componentes de $\boldsymbol{a}_1$ e igualando a cero. Entonces: [31]

$$\frac{\partial M}{\partial \boldsymbol{a}_1} = 2\boldsymbol{S}\boldsymbol{a}_1 - 2\lambda\boldsymbol{a}_1 = 0$$

cuya solución es: [31]

$$\boldsymbol{S}\boldsymbol{a}_1 = \lambda\boldsymbol{a}_1$$

que implica que $\boldsymbol{a}_1$ es un vector propio (o autovector) de la matriz $\boldsymbol{S}$, y λ su correspondiente valor propio (o autovalor). Para determinar qué valor propio de $\boldsymbol{S}$ es la solución de esta ecuación, se multiplica por la izquierda por $\boldsymbol{a}_1'$ y se tiene que,

$$\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1 = \lambda\boldsymbol{a}_1'\boldsymbol{a}_1 = \lambda$$

y se concluye que λ es la varianza de z_1. Como esta es la cantidad que queremos maximizar, λ será el mayor valor propio de la matriz $\boldsymbol{S}$. Su vector asociado $\boldsymbol{a}_1$, define los coeficientes de cada variable en el primer componente principal. [31]

Cálculo del segundo componente

Se trata de obtener el mejor plano de proyección de las variables $\boldsymbol{X}$. Se calcula estableciendo como función objetivo que la suma de las varianzas de z_1 y z_2[10] sea máxima, donde $\boldsymbol{a}_1$ y $\boldsymbol{a}_2$ son los vectores que definen el plano. La función objetivo será: [31]

$$\emptyset = \boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1 + \boldsymbol{a}_2'\boldsymbol{S}\boldsymbol{a}_2 - \lambda_1(\boldsymbol{a}_1'\boldsymbol{a}_1 - 1) - \lambda_2(\boldsymbol{a}_2'\boldsymbol{a}_2 - 1)$$

que incorpora las restricciones de que las direcciones deben de tener módulo unitario $(\boldsymbol{a}_i'\boldsymbol{a}_i) = 1,\ i = 1, 2$. Derivando e igualando a cero: [31]

$$\frac{\partial\emptyset}{\partial\boldsymbol{a}_1} = 2\boldsymbol{S}\boldsymbol{a}_1 - 2\lambda_1\boldsymbol{a}_1 = 0$$

$$\frac{\partial\emptyset}{\partial\boldsymbol{a}_2} = 2\boldsymbol{S}\boldsymbol{a}_2 - 2\lambda_2\boldsymbol{a}_2 = 0$$

La solución de este sistema es: [31]

$$\boldsymbol{S}\boldsymbol{a}_1 = \lambda_1\boldsymbol{a}_1,$$

$$\boldsymbol{S}\boldsymbol{a}_2 = \lambda_2\boldsymbol{a}_2$$

que indica que $\boldsymbol{a}_1$ y $\boldsymbol{a}_2$ deben ser vectores propios de $\boldsymbol{S}$. Si se toman los vectores propios de norma uno y se sustituyen en la función objetivo, se obtiene que, en el máximo, esta función es

[10] En donde los vectores de valores en los n individuos son $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{a}_1$y $\boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{a}_2$.

$$\phi = \lambda_1 + \lambda_2$$

Por lo tanto, es claro que λ_1 y λ_2 deben ser los dos autovalores mayores de la matriz $\boldsymbol{S}$ y $\boldsymbol{a}_1$ y $\boldsymbol{a}_2$ sus correspondientes autovectores. Se observa que la covarianza entre z_1 y z_2, dada por $\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_2$ es cero ya que $\boldsymbol{a}_1'\boldsymbol{a}_2 = 0$, y las variables z_1 y z_2 estarán incorreladas [31].

Propiedades de los componentes

Los componentes principales como nuevas variables tienen las propiedades siguientes: [31]

(1) Conservan la variabilidad inicial: la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales. Como $Varianza(z_h) = \lambda_h$ y la suma de las raíces características es la traza de la matriz de varianzas y covarianzas se puede plantear como:

$$traza(\boldsymbol{S}) = Varianza(x_1) + \ldots + Varianza(x_p) = \lambda_1 + \ldots + \lambda_p,$$

Por tanto $\sum_{j=1}^{p} Varianza(x_i) = \sum_{j=1}^{p} \lambda_i = \sum_{j=1}^{p} Varianza(z_i)$. Las nuevas variables z_i tienen conjuntamente la misma variabilidad que las variables originales, pero su distribución es muy distinta en los dos conjuntos.

(2) La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz.

En efecto, como la varianza del componente h es λ_h, el valor propio que define el componente, y la suma de todas las varianzas de las variables originales es $\sum_{j=1}^{p} \lambda_i$, igual, como acabamos de ver, a la suma de las varianzas de los componentes, la proporción de variabilidad total explicada por el componente h es $\lambda_h / \sum_{j=1}^{p} \lambda_i$.

(3) Las covarianzas entre cada componente principal y las variables x vienen dadas por el producto de las coordenadas del vector propio que define el componente por el valor propio:

$$Covarianza(z_h; x_1, \ldots, x_p) = \lambda_h \boldsymbol{a}_h = (\lambda_h a_{h1}, \ldots, \lambda_h a_{hp},$$

donde $\boldsymbol{a}_h$ es el vector de coeficientes de la componente z_h.

(4) La correlación entre un componente principal y una variable x es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable. Para comprobarlo:

$$Correlación(z_h;\ x_j) = \frac{Covarianza(z_h; x_j)}{\sqrt{Varianza(z_h)Varianza(x_j)}} = \frac{\lambda_h a_{hj}}{\sqrt{\lambda_h s_j^2}} = a_{hj}\frac{\sqrt{\lambda_h}}{s_j}.$$

5.4.4. ANÁLISIS DE CONGLOMERADOS (AC)

El análisis de conglomerados (clusters) tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes o semejanzas entre ellos. Normalmente se agrupan las observaciones, pero el análisis de conglomerados puede también aplicarse para agrupar variables. Estos métodos se conocen también con el nombre de métodos de clasificación automática o no supervisada, o de reconocimiento de patrones sin supervisión. El nombre de no supervisados se aplica para distinguirlos del análisis discriminante. El análisis de conglomerados estudia tres tipos de problemas: [31]

a) Partición de los datos. Se dispone de datos que sospechamos son heterogéneos y se desea dividirlos en un número de grupos prefijado, de manera que:
   1. cada elemento pertenezca a uno y solo uno de los grupos;
   2. todo elemento quede clasificado;
   3. cada grupo sea internamente homogéneo.

b) Construcción de jerarquías. Se trata de estructurar los elementos de un conjunto de forma jerárquica por su similitud. Por ejemplo, se tiene una encuesta de atributos de distintas profesiones y se quiere ordenarlas por similitud. Una clasificación jerárquica implica que los datos se ordenan en niveles, de manera que los niveles superiores contienen a los inferiores. La jerarquía construida permite obtener una partición de los datos en grupos.

c) Clasificación de variables. En problemas con muchas variables es interesante hacer un estudio exploratorio inicial para dividir las variables en grupos. Este estudio puede orientar para plantear los modelos formales para reducir la dimensión. Las variables pueden clasificarse en grupos o estructurarse en una jerarquía.

Los métodos de partición utilizan la matriz de datos, pero los algoritmos jerárquicos utilizan la matriz de distancias o similitudes entre elementos. Para agrupar variables se parte de la matriz de relación entre variables: para variables continuas suele ser la matriz de correlación, y para variables discretas, se construye, a partir de la distancia ji-cuadrado [31].

En el presente análisis se utiliza la construcción de jerarquías para hacer grupos de variables para luego aplicar el análisis de correspondencia múltiple en cada grupo.

Algoritmos Jerárquicos

Dada una matriz de distancias o de similitudes se desea clasificar los elementos en una jerarquía. Los algoritmos existentes funcionan de manera que los elementos son sucesivamente asignados a los grupos, pero la asignación es irrevocable, es decir, una vez hecha, no se cuestiona nunca más. Los algoritmos son de dos tipos: [31]

1. De aglomeración. Parten de los elementos individuales y los van agregando en grupos.
2. De división. Parten del conjunto de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales.

Los algoritmos de aglomeración requieren menos tiempo de cálculo y son los más utilizados.

Métodos Aglomerativos

Los algoritmos aglomerativos que se utilizan tienen siempre la misma estructura y sólo se diferencian en la forma de calcular las distancias entre grupos. Su estructura es: [31]

1. Comenzar con tantas clases (grupos) como elementos n. Las distancias entre clases son las distancias entre elementos originales.
2. Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos una clase.
3. Sustituir los dos elementos utilizados en (2) para definir la clase por un nuevo elemento que represente la clase construida. Las distancias entre este nuevo elemento y los anteriores se calculan con algún criterio para definir distancias entre elementos o grupos de elementos.
4. Volver a (2) y repetir (2) y (3) hasta que se tengan todos los elementos agrupados en una clase única.

Criterios para definir distancias entre grupos

Se supone que se tiene un grupo A con $n_a$elementos, y un grupo B con n_b elementos, y que ambos se fusionan para crear un grupo (AB) con $n_a + n_b$ elementos. La

distancia del nuevo grupo (AB), a otro grupo C con n_c elementos, se calcula por ejemplo con alguna de siguientes reglas: [31]

1. **Encadenamiento simple o vecino más próximo.** La distancia entre los dos nuevos grupos es la menor de las distancias entre grupos antes de la fusión. Es decir:

$$d(C;AB) = \min(d_{CA}, d_{CB})$$

Una forma simple de calcular con un ordenador el mínimo entre las dos distancias es utilizar que

$$\min(d_{CA}, d_{CB}) = \frac{1}{2}(d_{CA} + d_{CB} - |d_{CA} - d_{CB}|)$$

En efecto, si $d_{CB} > d_{CA}$ el término en valor absoluto es $d_{CB} - d_{CA}$ y el resultado de la operación es d_{CA}, la menor de las distancias. Si $d_{CA} > d_{CB}$ el segundo término es $d_{CA} - d_{CB}$ y se obtiene d_{CB}.

Como este criterio sólo depende del orden de las distancias será invariante ante transformaciones monótonas: se obtiene la misma jerarquía, aunque las distancias sean numéricamente distintas. Se ha comprobado que este criterio tiende a producir grupos alargados, que pueden incluir elementos muy distintos en los extremos.

2. **El método de Ward.** Un proceso algo diferente de construir el agrupamiento jerárquico ha sido propuesto por Ward y Wishart. La diferencia con los métodos anteriores es que ahora se parte de los elementos directamente, en lugar de utilizar la matriz de distancias, y se define una medida global de la heterogeneidad de una agrupación de observaciones en grupos. Esta medida es $\boldsymbol{W}$, que es la suma de las distancias euclidianas al cuadrado entre cada elemento y la media de su grupo, que viene dada por:

$$\boldsymbol{W} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\boldsymbol{x}_{ig} - \bar{\boldsymbol{x}}_g)' (\boldsymbol{x}_{ig} - \bar{\boldsymbol{x}}_g)$$

donde $\bar{\boldsymbol{x}}_g$ es la media del grupo g. El criterio comienza suponiendo que cada dato forma un grupo, $g = n$ y por tanto $\boldsymbol{W}$ es cero. A continuación, se unen los elementos que produzcan el incremento mínimo de $\boldsymbol{W}$. Esto implica tomar los más próximos con la distancia euclidiana. En la siguiente etapa se tiene $n-1$ grupos, $n-2$ de un elemento y uno de dos elementos. Se Decide de nuevo unir dos grupos para que $\boldsymbol{W}$ crezca lo menos posible, con lo que se pasa a $n - 2$ grupos y así sucesivamente hasta tener un único grupo. Los valores de $\boldsymbol{W}$ van indicando el crecimiento del criterio al formar grupos y pueden utilizarse para decidir cuántos

grupos naturales contienen los datos. En cada etapa, los grupos que deben unirse para minimizar $\boldsymbol{W}$ son aquellos tales que:

$$min \frac{n_a n_b}{n_a + n_b} (\overline{\boldsymbol{x}}_a - \overline{\boldsymbol{x}}_b)'(\overline{\boldsymbol{x}}_a - \overline{\boldsymbol{x}}_b)$$

Se recomienda analizar qué tipo de criterio de distancia es más razonable utilizar para los datos que se quieren agrupar y, en caso de duda, probar con varios y comparar los resultados.

El dendrograma

El dendrograma, o árbol jerárquico, es una representación gráfica del resultado del proceso de agrupamiento en forma de árbol. Los criterios para definir distancias, que hemos presentado, tienen la propiedad de que, si consideramos tres grupos, A, B, C, se verifica que: [31]

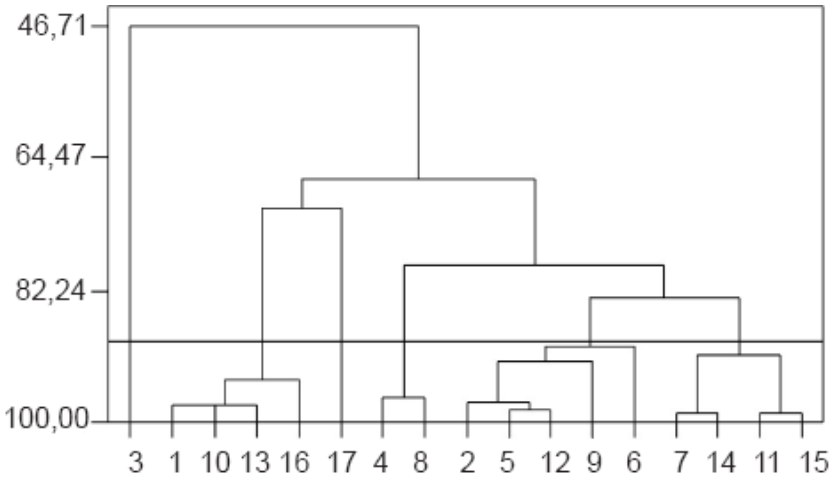
$$d(A, C) \leq max\{d(A, B), D(B, C)\}$$

y una medida de distancia que tiene esta propiedad se denomina ultramétrica. Esta propiedad es más fuerte que la propiedad triangular, ya que una ultramétrica es siempre una distancia. En efecto si $d^2(A, B)$ es menor o igual que el máximo de $d^2(A, C)$, $d^2(B, C)$, forzosamente será menor o igual que la suma $d^2(A, B) + d^2(B, C)$. El dendrograma es la representación de una ultramétrica, y se construye (Véase Figura 5.2): [31]

1. En la parte inferior del gráfico se disponen los n elementos iniciales.
2. Las uniones entre elementos se representan por tres líneas rectas. Dos dirigidas a los elementos que se unen y que son perpendiculares al eje de los elementos y una paralela a este eje que se sitúa al nivel en que se unen.
3. El proceso se repite hasta que todos los elementos están conectados por líneas rectas.

Si se corta el dendrograma a un nivel de distancia dado, se obtiene una clasificación del número de grupos existentes a ese nivel y los elementos que los forman [31].

Figura 5.2. Ejemplo de representación de un dendrograma.



5.4.4.1. CONGLOMERADOS POR VARIABLES

El análisis de conglomerados de variables es un procedimiento exploratorio que puede sugerir procedimientos de reducción de la dimensión, como por ejemplo el análisis factorial. La idea es construir una matriz de distancias o similitudes entre variables y aplicar a esta matriz un algoritmo jerárquico de clasificación [31].

Medidas de distancia y similitud entre variables

Las medidas habituales de asociación entre variables continuas son la covarianza y la correlación. Estas medidas tienen en cuenta únicamente las relaciones lineales. Alternativamente, podríamos construir una medida de distancia entre dos variables x_j y x_h representando cada variable como un punto en $\Re^n$ y calculando la distancia euclidiana entre los dos puntos. Esta medida es: [31]

$$d_{jh}^2 = \sum_{i=1}^{n} (x_{ij} - x_{ih})^2$$
$$= \sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ih}^2 - 2\sum_{i=1}^{n} x_{ij}x_{ih}$$

Para que la distancia no dependa de las unidades, las variables deben estar estandarizadas. En otro caso, la distancia entre dos variables podría alterarse arbitrariamente mediante transformaciones lineales de éstas. (Por ejemplo, midiendo las estaturas en metros, en lugar de en centímetros y en desviaciones respecto a la media poblacional en lugar de con carácter absoluto). Suponiendo, por tanto, que se

trabaja con variables estandarizadas de media cero y varianza uno, se obtiene que d_{jh}^2 se reduce a: [31]

$$d_{jh}^2 = 2n(1 - r_{jh})$$

Se observa que: [31]

(a) Si $r_{jh} = 1$, la distancia es cero, indicando que las dos variables son idénticas.

(b) Si $r_{jh} = 0$, las dos variables están incorreladas y la distancia es $d_{jh} = \sqrt{2n}$.

(c) Si $r_{jh} < 0$, las dos variables tienen correlación negativa, y la distancia tomará su valor máximo $\sqrt{4n}$, cuando las dos variables tengan correlación −1.

Esta medida de distancia puede estandarizarse para que sus valores estén entre cero y uno prescindiendo de la constante n y tomando $d_{jh} = \sqrt{(1 - r_{jh})/2}$ [31].

Para variables cualitativas binarias se puede construir una medida de similitud construyendo una *tabla de asociación entre variables*. Para ello se cuenta: a) el número de elementos donde están presentes ambas características, es decir 1 y 0, b) el número de elementos donde está solo el 1, c) el número de elementos donde esta solo el 0, y d) el número de elementos donde no están ninguna de las dos características, o sea, ni 1 y 0. En estas tablas se verifica que, si n es el número de individuos, $n = a + b + c + d$, podemos construir coeficientes de similitud como se hizo con los elementos. Alternativamente, esta tabla de asociación entre variables es una tabla de contingencia y una medida de distancia es el valor de la ji-cuadrado: [31]

$$\chi^2 = \frac{(ad - bc)^2}{(a + b)(a + c)(c + d)(b + d)}.$$

Es más habitual definir la distancia por el coeficiente de contingencia

$$d_{ij} = 1 - \sqrt{\frac{\chi^2}{n}}.$$

5.4.5. ANÁLISIS DE CORRESPONDENCIAS SIMPLE (ACS)

El análisis de correspondencias simple o simplemente análisis de correspondencia es una técnica descriptiva para representar tablas de contingencia. Constituye el equivalente de componentes principales para variables cualitativas. Desde este

enfoque, la información de partida es una matriz de dimensiones $I \times J$, que representa las frecuencias absolutas observadas de dos variables cualitativas en n elementos (personas). La primera variable se representa por filas, y suponemos que toma I valores posibles, y la segunda se representa por columnas, y toma J valores posibles. En general, una tabla de contingencia es un conjunto de números positivos dispuestos en una matriz, donde el número en cada casilla representa la frecuencia absoluta observada para esa combinación de las dos variables (Véase Tabla 5.13) [31].

Tabla 5.13. Ejemplo de tabla de contingencia para la combinación de variables: Estado de salud de las personas concerniente a la IRC y grupos de edades.

	Estado	**Grupos de edades**						Total
		18 a 29	**30 a 39**	**40 a 49**	**50 a 59**	**60 a 69**	**>=70**	
	Personas con IRC	1	5	13	22	14	21	76
	Personas sin ERC	266	129	109	71	31	18	624
	Total	267	134	122	93	45	39	700

El análisis de correspondencias es un procedimiento para resumir la información contenida en una tabla de contingencia. Puede interpretarse de dos formas equivalentes. La primera, como una manera de representar las variables en un espacio de dimensión menor, de forma análoga a componentes principales, pero definiendo la distancia entre los puntos de manera coherente con la interpretación de los datos y en lugar de utilizar la distancia euclidiana se utiliza la distancia ji-cuadrado. Desde este enfoque, el análisis de correspondencias es el equivalente de componentes principales para datos cualitativos, de ahí que se trate como una técnica multivariante [31].

Búsqueda de la mejor proyección

Se llama $\boldsymbol{F}$ a la matriz de frecuencias relativas obtenida dividiendo cada casilla por n, el total de elementos observados. Se llama f_{ij} a las frecuencias relativas que verifican

$$\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = 1$$

La matriz $\boldsymbol{F}$ puede considerarse por filas o por columnas. Cualquier análisis lógico de esta matriz debe de ser equivalente al aplicado a su transpuesta, ya que la elección de la variable que se coloca en filas, en lugar de en columnas, es arbitraria, y no debe

influir en el análisis. Por lo que solo se presenta el análisis por filas de esta matriz, que es simétrico al análisis por columnas [31].

Proyección por filas

Se trata de analizar la matriz $\boldsymbol{F}$ por filas. Entonces las I filas se pueden tomar como I puntos en el espacio $\Re^J$.Se busca una representación de estos *I* puntos en un espacio de dimensión menor que permita apreciar sus distancias relativas. El objetivo es el mismo que con componentes principales, pero ahora se tiene en cuenta las peculiaridades de este tipo de datos. Estas peculiaridades provienen de que la frecuencia relativa de cada fila es distinta, lo que implica que: [31]

(1) Todas las filas (puntos en $\Re^J$) no tienen el mismo peso, ya que algunas contienen más datos que otras. Al representar el conjunto de las filas (puntos) debemos dar más peso a aquellas filas que contienen más datos.
(2) La distancia euclidiana entre puntos no es una buena medida de su proximidad y debemos modificar esta distancia.

Comenzando con el primer punto, cada fila de la matriz $\boldsymbol{F}$ tiene una frecuencia relativa $f_{i.} = \sum_{j=1}^{J} f_{ij}$, y el conjunto de estas frecuencias relativas se calcula con: [31]

$$\boldsymbol{f} = \boldsymbol{F}'\mathbf{1}$$

se debe dar a cada fila un peso proporcional a su frecuencia relativa y los términos del vector $\boldsymbol{f}$ pueden directamente considerarse como pesos, ya que son números positivos que suman uno [31].

Con relación a la medida de distancia a utilizar entre las filas, se observa que la distancia euclidiana no es una buena medida de las diferencias reales entre las estructuras de las filas. Se puede dar la situación de que las frecuencias relativas de las filas son muy distintas, y sin embargo tienen exactamente la misma estructura relativa: puede suceder simplemente que una fila tenga más del doble de frecuencias que otra fila, pero la distribución de frecuencias es idéntica en ambas filas [31].

Si se calcula la distancia euclidiana entre estas dos filas se obtiene un valor alto, que no refleja una estructura distinta de las filas sino sólo que tienen distinta frecuencia relativa. Suponiendo que se divide cada casilla por la frecuencia relativa de la fila $f_{i.}$, se obtienen los números que aparecen en las filas, los cuales representan la frecuencia relativa de la variable columna condicionada a la variable fila. Ahora las dos

filas son idénticas, y esto es coherente con una distancia euclidiana cero entre ambas [31].

Para analizar qué medida de distancia se debe utilizar, se le llama $\boldsymbol{R}$ a la matriz de frecuencias relativas condicionadas al total de la fila, que se obtiene con: [31]

$$\boldsymbol{R} = \boldsymbol{D}_f^{-1}\boldsymbol{F}$$

donde $\boldsymbol{D}_f$ es una matriz diagonal $I \times I$ con los términos del vector $\boldsymbol{f}$, $f_{i.}$, frecuencias relativas de las filas, en la diagonal principal. Esta operación transforma la matriz original de frecuencias relativas $\boldsymbol{F}$, en otra matriz cuyas casillas por filas suman uno. Cada fila de esta matriz representa la distribución de la variable en columnas condicionada al atributo que representa la fila [31].

Se denota $\boldsymbol{r}_i'$ a la fila i de la matriz $\boldsymbol{R}$ de frecuencias relativas condicionadas por filas, que puede considerarse un punto (o un vector) en el espacio $\Re^J$. Como la suma de los componentes de $\boldsymbol{r}_i'$ es uno, todos los puntos están en un espacio de dimensión $J-1$. Se desea proyectar estos puntos en un espacio de dimensión menor de manera que las filas que tengan la misma estructura estén próximas, y las que tengan una estructura muy diferente, alejadas. Para ello, se debe definir una medida de distancia entre dos filas $\boldsymbol{r}_a$ y $\boldsymbol{r}_b$. Una posibilidad es utilizar la distancia euclidiana, pero esta distancia tiene el inconveniente de tratar igual a todos los componentes de estos vectores [31].

Para obtener comparaciones razonables entre estas frecuencias relativas se tiene que tener en cuenta la frecuencia relativa de aparición del atributo que se estudia. En atributos raros, pequeñas diferencias absolutas pueden ser grandes diferencias relativas, mientras que, en atributos con gran frecuencia, la misma diferencia será poco importante. Una manera intuitiva de construir las comparaciones es ponderar las diferencias en frecuencia relativa entre dos atributos inversamente proporcional a la frecuencia de este atributo. Es decir, en lugar de sumar los términos $(r_{aj} - r_{bj})^2 = (f_{aj}/f_{a.} - f_{bj}/f_{b.})^2$ que miden la diferencia que las filas a y b tienen en la columna j se suman los términos $(r_{aj} - r_{bj})^2/f_{.j}$ donde $f_{.j} = \sum_{i=1}^{I} f_{ij}$ es la frecuencia relativa de la columna j. La expresión de la distancia entre dos filas, $\boldsymbol{r}_a$ y $\boldsymbol{r}_b$ de $\boldsymbol{R}$ vendrá dada en esta métrica por: [31]

$$D^2(\boldsymbol{r}_a, \boldsymbol{r}_b) = \sum_{j=1}^{J} \left(\frac{f_{aj}}{f_{a.}} - \frac{f_{bj}}{f_{b.}}\right)^2 \frac{1}{f_{.j}} = \sum_{j=1}^{J} \frac{(r_{aj} - r_{bj})^2}{f_{.j}}$$

que puede escribirse matricialmente como: [31]

$$D^2(\boldsymbol{r}_a, \boldsymbol{r}_b) = (\boldsymbol{r}_a - \boldsymbol{r}_b)' \boldsymbol{D}_c^{-1} (\boldsymbol{r}_a - \boldsymbol{r}_b)$$

A esta distancia se le conoce como distancia χ^2. $\boldsymbol{D}_c$ es una matriz diagonal con términos $f_{.j}$. Se simplifica el problema definiendo una matriz de datos transformada, sobre la que tiene sentido considerar la distancia euclidiana entre filas. Llamando: [31]

$$\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{D}_c^{-1/2} = \boldsymbol{D}_f^{-1}\boldsymbol{F}\boldsymbol{D}_c^{-1/2}$$

Se obtiene una matriz $\boldsymbol{Y}$ que contiene términos del tipo

$$y_{ij} = \left\{ \frac{f_{ij}}{f_{i.} f_{.j}^{1/2}} \right\}$$

que ya no suman uno ni por filas ni por columnas. Las casillas de esta matriz representan las frecuencias relativas condicionadas por filas $\frac{f_{ij}}{f_{i.}}$, pero estandarizadas por su variabilidad, que depende de la raíz cuadrada de la frecuencia relativa de la columna. De esta manera las casillas son directamente comparables entre sí [31].

De esta manera se tendría una matriz de datos estándar, con observaciones en filas y variables en columnas, y así averiguar cómo proyectarla de manera que se preserven las distancias relativas entre las filas, es decir, las filas con estructura similar aparezcan próximas en la proyección. Esto implica encontrar una dirección $\boldsymbol{a}$ de norma unidad,

$$\boldsymbol{a}'\boldsymbol{a} = 1$$

tal que el vector de puntos proyectados sobre esta dirección,

$$\boldsymbol{y}_p(\boldsymbol{a}) = \boldsymbol{Y}\boldsymbol{a}$$

tenga variabilidad máxima. El vector $\boldsymbol{a}$ se encontrará maximizando $\boldsymbol{y}_p(\boldsymbol{a})'\boldsymbol{y}_p(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{Y}'\boldsymbol{Y}\boldsymbol{a}$ con la condición $\boldsymbol{a}'\boldsymbol{a} = 1$, tal como en componentes principales: el vector $\boldsymbol{a}$ es un vector propio de la matriz $\boldsymbol{Y}'\boldsymbol{Y}$. Sin embargo, este tratamiento de la matriz $\boldsymbol{Y}$ como una matriz de variables continuas no es del todo correcto porque las filas tienen una distinta frecuencia relativa $f_{i.}$, y por tanto deben tener distinto peso. Aquellas filas con mayor frecuencia relativa deben de tener más peso en la representación que aquellas otras con frecuencia relativa muy baja, de manera que las filas con gran número de

individuos estén bien representadas, aunque esto sea a costa de representar peor las filas con pocos elementos. En consecuencia, se proporciona a cada fila un peso proporcional al número de datos que contiene. Esto puede hacerse maximizando la suma de cuadrados ponderada.

$$m = \boldsymbol{a}'\boldsymbol{Y}'\boldsymbol{D}_f\boldsymbol{Y}\boldsymbol{a}$$

que equivale a

$$m = \boldsymbol{a}'\boldsymbol{D}_c^{-1/2}\boldsymbol{F}'\boldsymbol{D}_f^{-1}\boldsymbol{F}\boldsymbol{D}_c^{-1/2}\boldsymbol{a}$$

Alternativamente se puede construir una matriz de datos $\boldsymbol{Z}$ definida por

$$\boldsymbol{Z} = \boldsymbol{D}_f^{-1/2}\boldsymbol{F}\boldsymbol{D}_c^{-1/2}$$

cuyos componentes son

$$z_{ij} = \left\{\frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}}\right\}$$

y que estandariza las frecuencias relativas en cada casilla por el producto de las raíces cuadradas de las frecuencias relativas totales de la fila y la columna, y, además, escribir el problema de encontrar el vector $\boldsymbol{a}$ como el problema de maximizar $m = \boldsymbol{a}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{a}$ sujeto a la restricción $\boldsymbol{a}'\boldsymbol{a} = 1$. Este es el problema resuelto en componentes principales, cuya solución es: [31]

$$\boldsymbol{D}_c^{-1/2}\boldsymbol{F}'\boldsymbol{D}_f^{-1}\boldsymbol{F}\boldsymbol{D}_c^{-1/2}\boldsymbol{a} = \lambda\boldsymbol{a}$$

y $\boldsymbol{a}$ debe ser un vector propio de la matriz $\boldsymbol{Z}'\boldsymbol{Z}$ y λ su valor propio. Se puede comprobar que la matriz $\boldsymbol{Z}'\boldsymbol{Z}$ tiene como mayor valor propio siempre el 1 y como vector propio $\boldsymbol{D}_c^{1/2}$. Para observar esta situación, si se multiplica la expresión anterior por $\boldsymbol{D}_c^{-1/2}$ se obtiene: [31]

$$\boldsymbol{D}_c^{-1}\boldsymbol{F}'\boldsymbol{D}_f^{-1}\boldsymbol{F}\left(\boldsymbol{D}_c^{-\frac{1}{2}}\boldsymbol{a}\right) = \boldsymbol{\lambda}(\boldsymbol{D}_c^{-1/2}\boldsymbol{a})$$

Las matrices de $\boldsymbol{D}_f^{-1}\boldsymbol{F}$ y $\boldsymbol{F}\boldsymbol{D}_c^{-1}$ representan matrices de frecuencias relativas por filas y por columnas y su suma por filas y columnas, respectivamente, es 1. Por tanto $\boldsymbol{D}_f^{-1}\boldsymbol{F}\mathbf{1} = \mathbf{1}$ y $\boldsymbol{D}_c^{-1}\boldsymbol{F}'\mathbf{1} = \mathbf{1}$, que implica que la matriz $\boldsymbol{D}_c^{-1}\boldsymbol{F}'\boldsymbol{D}_f^{-1}\boldsymbol{F}$ tiene un valor propio 1 unido a un vector propio **1**. En consecuencia, haciendo $\boldsymbol{D}_c^{-1/2}\boldsymbol{a} = \mathbf{1}$ se concluye que la

matriz $\boldsymbol{Z'Z}$ tiene un valor propio igual a uno con vector propio $\boldsymbol{D_c^{1/2}}$. Se puede demostrar que esta es una solución trivial, que no da información sobre la estructura de las filas. Por tal razón se toma el valor propio mayor, menor que la unidad y su vector propio asociado $\boldsymbol{a}$. Entonces proyectando la matriz $\boldsymbol{Y}$ sobre la dirección $\boldsymbol{a}$ encontrada: [31]

$$\boldsymbol{y}_f(\boldsymbol{a}) = \boldsymbol{Ya} = \boldsymbol{D_f^{-1}FD_c^{-1/2}a}$$

y el vector $\boldsymbol{y}_f(\boldsymbol{a})$ es la mejor representación de las filas de la tabla de contingencia en una dimensión. Análogamente, si se extra el vector propio ligado al siguiente mayor valor propio se obtiene una segunda coordenada y así poder representar las filas en un espacio de dimensión dos. Las coordenadas de la representación de cada fila vendrán dadas por las filas de la matriz

$$\boldsymbol{C}_f = \boldsymbol{YA}_2 = \boldsymbol{D_f^{-1}FD_c^{-1/2}A}_2$$

donde $\boldsymbol{A}_2 = [\boldsymbol{a}_1\ \boldsymbol{a}_2]$ contiene en columnas los dos vectores propios $\boldsymbol{Z'Z}$. La matriz $\boldsymbol{C}_f$ es $I \times 2$ y las dos coordenadas de cada fila proporcionan la mejor representación de las fila de la matriz $\boldsymbol{F}$ en un espacio de dos dimensiones. El procedimiento se extiende sin dificultad para representaciones en más dimensiones, calculando vectores propios adicionales de la matriz $\boldsymbol{Z'Z}$ [31].

Proyección de las columnas

Se aplica a las columnas de la matriz $\boldsymbol{F}$ un análisis equivalente al de las filas. Las columnas serán ahora puntos en $\Re^I$. Llamando

$$\boldsymbol{c} = \boldsymbol{F'1}$$

al vector de frecuencias relativas de las columnas y $\boldsymbol{D}_c$ a la matriz diagonal que contiene estas frecuencias relativas en la diagonal principal, de acuerdo con el apartado anterior, la mejor representación de los J puntos (columnas) en un espacio de dimensión menor, con la métrica χ^2 conducirá, por simetría, a estudiar la matriz $\boldsymbol{D_c^{-1}F'D_f^{-1/2}}$. Se Observa que, si ahora se considera la matriz $\boldsymbol{F'}$ y se vuelve al problema de representarla por filas (que es equivalente a representar $\boldsymbol{F}$ por columnas), el problema es idéntico al que se ha resuelto en el apartado anterior. Ahora la matriz que contiene las frecuencias relativas de las filas $\boldsymbol{F'}$ es $\boldsymbol{D}_c$ y la que contiene la de las

columnas es $\boldsymbol{D}_f$. Intercambiando el papel de estas matrices, las direcciones de proyección son los vectores propios de la matriz

$$\boldsymbol{ZZ'} = \boldsymbol{D}_f^{-1/2}\boldsymbol{F}\boldsymbol{D}_c^{-1}\boldsymbol{F'}\boldsymbol{D}_f^{-1/2}$$

donde $\boldsymbol{Z}$ es la matriz $I \times J$ definida en el apartado anterior. Como $\boldsymbol{Z'Z}$ y $\boldsymbol{ZZ'}$ tienen los mismos valores propios no nulos, esa matriz tendrá un valor propio unidad ligado al vector propio $\mathbf{1}$. Esta solución trivial no se considera. Llamando $\boldsymbol{b}$ al vector propio ligado al mayor valor propio distinto de la unidad de $\boldsymbol{ZZ'}$, la mejor representación de las columnas de la matriz en un espacio de dimensión uno vendrá dada por

$$\boldsymbol{y}_c(\boldsymbol{b}) = \boldsymbol{Y'b} = \boldsymbol{D}_C^{-1}\boldsymbol{F'}\boldsymbol{D}_f^{-1/2}\boldsymbol{b}$$

y, análogamente, la mejor representación en dimensión dos de las columnas de la matriz vendrá dada por las coordenadas definidas por las filas de la matriz

$$\boldsymbol{C}_c = \boldsymbol{Y'B}_2 = \boldsymbol{D}_C^{-1}\boldsymbol{F'}\boldsymbol{D}_f^{-1/2}\boldsymbol{B}_2$$

donde $\boldsymbol{B}_2 = [\boldsymbol{b}_1\ \boldsymbol{b}_2]$ contiene en columnas los dos vectores propios ligados a los valores propios mayores de $\boldsymbol{ZZ'}$ y menores que la unidad. La matriz $\boldsymbol{C}_c$ es $J \times 2$ y cada fila es la mejor representación de las columnas de la matriz $\boldsymbol{F}$ en un espacio de dos dimensiones [31].

Análisis conjunto

Dada la simetría del problema conviene representar conjuntamente las filas y las columnas de la matriz. Se observa que las matrices $\boldsymbol{Z'Z}$ y $\boldsymbol{ZZ'}$ tienen los mismos valores propios no nulos y que los vectores propios de ambas matrices que corresponden al mismo valor propio están relacionados. En efecto, si $\boldsymbol{a}_i$ es un vector propio de $\boldsymbol{Z'Z}$ ligado al valor propio λ_i: [31]

$$\boldsymbol{Z'Z}\boldsymbol{a}_i = \lambda_i\boldsymbol{a}_i$$

entonces, multiplicando por $\boldsymbol{Z}$

$$\boldsymbol{ZZ'}(\boldsymbol{Za}_i) = \lambda_i(\boldsymbol{Za}_i)$$

y se obtiene que $\boldsymbol{b}_i = \boldsymbol{Za}_i$ es un vector propio de $\boldsymbol{ZZ'}$ ligado al valor propio λ_i. Una manera rápida de obtener estos vectores propios es calcular directamente los vectores propios de la matriz de dimensión más pequeña, $\boldsymbol{Z'Z}$ o $\boldsymbol{ZZ'}$, y obtener los otros

vectores propios como $\boldsymbol{Z}\boldsymbol{a}_i$ o $\boldsymbol{Z}'\boldsymbol{b}_i$. Alternativamente se puede utilizar la descomposición en valores singulares de la matriz $\boldsymbol{Z}$ o $\boldsymbol{Z}'$. Esta descomposición aplicada a $\boldsymbol{Z}$ es

$$\boldsymbol{Z} = \boldsymbol{B}_r \boldsymbol{D}_r \boldsymbol{A}_r' = \sum_{i=1}^{r} \lambda_i^{1/2} \boldsymbol{b}_i \boldsymbol{a}_i'$$

donde $\boldsymbol{B}_r$ contiene en columnas los vectores propios de $\boldsymbol{Z}\boldsymbol{Z}'$, $\boldsymbol{A}_r$ los de $\boldsymbol{Z}'\boldsymbol{Z}$ y D_r es diagonal y contiene los valores singulares, $\lambda_i^{1/2}$, o raíces de los valores propios no nulos y $r = \min(I, J)$. Entonces la representación de las filas se obtiene con $\boldsymbol{y}_f(\boldsymbol{b})$ y la de las columnas con $\boldsymbol{y}_c(\boldsymbol{b})$. La representación de la matriz $\boldsymbol{Z}$ con h dimensiones (habitualmente $h = 2$) implica aproximar esta matriz mediante $\widehat{\boldsymbol{Z}}_h = \boldsymbol{B}_h \boldsymbol{D}_h \boldsymbol{A}_h'$. Esto es equivalente, por $\boldsymbol{Z} = \boldsymbol{D}_f^{-1/2} \boldsymbol{F} \boldsymbol{D}_c^{-1/2}$, a una aproximación a la tabla de contingencia observada mediante

$$\widehat{\boldsymbol{F}}_h = \boldsymbol{D}_f^{1/2} \widehat{\boldsymbol{Z}}_h \boldsymbol{D}_c^{1/2},$$

y una forma de juzgar la aproximación que se está utilizando es reconstruir la tabla de contingencia con la expresión anterior [31].

Si se desea eliminar el valor propio unidad desde el principio, dado que no aporta información de interés, se puede reemplazar la matriz $\boldsymbol{F}$ por $\boldsymbol{F} - \widehat{\boldsymbol{F}}_e$, donde $\widehat{\boldsymbol{F}}_e$ es la matriz de frecuencias esperadas que viene dada por

$$\widehat{\boldsymbol{F}}_e = \frac{1}{n} \boldsymbol{r}\boldsymbol{c}'$$

Por lo que $\boldsymbol{F} - \widehat{\boldsymbol{F}}_e$ ya no tiene el valor propio igual a la unidad [31].

La proporción de variabilidad explicada por cada dimensión se calcula como en componentes principales descartando el valor propio igual a uno y tomando la proporción que representa cada valor propio con relación al resto [31].

La distancia ji-cuadrado (o chi-cuadrado)

El contraste de independencia entre las variables fila y columna en una tabla de contingencia $I \times J$ se realiza con el estadístico

$$\chi^2 = \sum_{i=1}^{n} \frac{(frecuencias\ observadas - frecuencias\ esperadas)^2}{frecuencias\ esperadas}$$

que, en la hipótesis de independencia, sigue una distribución χ^2 con $(I-1)\times(J-1)$ grados de libertad. De acuerdo con la notación anterior, la frecuencia esperada en cada celda de la fila i, suponiendo independencia de filas y columnas, se obtendrá repartiendo el total de la fila $nf_{i.}$, proporcionalmente a la frecuencia relativa de cada columna $f_{.j}$. Por tanto, el estadístico χ^2 para contrastar la independencia puede escribirse: [31]

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}}$$

donde $f_{i.} = \sum_{j=1}^{J} f_{ij}$ es la frecuencia relativa de la fila i y $f_{.j} = \sum_{i=1}^{I} f_{ij}$ la de columna j. Como

$$\frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}} = \frac{nf_{i.}}{f_{.j}} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}^2}$$

En la prueba de independencia, si x e y representan a la variable fila y columna respectivamente, se dice que la hipótesis nula es H_0: x e y son independientes y la alternativa es H_1: x e y son dependientes [31].

Dado un nivel de significancia, como por ejemplo 0.05, y con un $\chi^2_{observado}$, la hipótesis nula se rechaza si $P\left[\chi^2_{(I-1)(j-1)} \geq \chi^2_{observado}\right] \leq 0.05$, lo que significa que la variable fila y columna son dependientes [31].

5.4.6. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES (ACM)

En la sección anterior, se expone la teoría del análisis de correspondencias que trata acerca de la relación entre dos variables cualitativas. En cambio, en esta sección se presenta un resumen teórico de la relación existente entre más de dos variables cualitativas, mediante *el análisis de correspondencias múltiples,* de forma abreviada ACM. En ACM se investiga el tipo de asociación existente entre variables y su intensidad. Por lo que esta técnica se aplica a tablas de contingencias en las que por filas se tienen n individuos y por columnas Q variables cualitativas o categóricas. Se puede llevar a cabo el ACM sobre una matriz que contenga los datos codificados de forma binaria, la *matriz binaria* o bien sobre una matriz formada por todos los cruzamientos posibles entre las variables, la *matriz de Burt* [34].

Por lo general se analiza la relación existente entre más de dos variables cualitativas en el contexto de un solo fenómeno de interés. Por ejemplo, podrían ser las respuestas a preguntas relacionadas con la utilización de ciertos agroquímicos en la agricultura, o datos que describan los antecedentes personales de enfermedades relacionadas con la IRC. Lo importante es que las variables sean "homogéneas", es decir, que sean sustantivamente similares [34].

Matriz binaria

Si se consideran 4 preguntas o variables acerca de la utilización de agroquímicos, la estructura de datos en la base original toma el formato dado en la Tabla 5.14 [34].

Tabla 5.14. Formato de datos originales en una matriz binaria para el ACM.

P1: ¿Utiliza Folidol?	P2: ¿Utiliza Karate?	P3: ¿Utiliza Lannate?	¿Utiliza Volaton?
1=si	2=no	1=si	1=si
2=no	1=si	1=si	2=no
2=no	2=no	2=no	1=si
1=si	2=no	1=si	1=si
. . .	. . .	. . .	. . .
… y así sucesivamente para la muestra de N=700 personas			

La representación de la matriz binaria correspondiente para desarrollar el ACM se muestra en la Tabla 5.15, en donde se codifican las respuestas de forma binaria, es decir, el "1" significa que fue lo que respondió la persona en una pregunta y el "0" indica que fue lo que no respondió [34].

Tabla 5.15. Ejemplo de codificación binaria para la aplicación del ACM.

P1: Utiliza Folidol?		P2: Utiliza Karate?		P3: Utiliza Lannate?		Utiliza volaton?	
Si	No	Si	No	Si	No	Si	No
1	0	0	1	1	0	1	0
0	1	1	0	1	0	0	1
0	1	0	1	0	1	1	0
1	0	0	1	1	0	1	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
… y así sucesivamente para la muestra de N=700 personas							

Se puede definir el ACM como el ACS[11] de la matriz binaria. El cálculo de la variabilidad o inercia total de la matriz binaria es muy simple. Depende solo del número de variables y del número de categorías. Suponiendo que se tiene Q variables y que cada variable q, tiene J_q categorías, J indica el número total de categorías: $J = \sum_{q=1}^{Q} J_q$ (En el ejemplo anteriormente expuesto, $Q = 4$, $J_q = 2$, $q = 1, \dots, Q$ y $J = 8$). La matriz binaria, simbolizada por $\boldsymbol{Z}$, con J columnas, es una matriz compuesta formada por tablas $\boldsymbol{Z}_q$ agrupadas lateralmente, una para cada variable. En cada tabla, los valores marginales de las filas, o sea, las frecuencias relativas de las filas, son iguales a una columna de unos. Por tanto, la inercia total de la matriz binaria es igual a la media de la inercia de las tablas que la componen. Cada tabla $\boldsymbol{Z}_q$ tiene un solo 1 en cada fila, los restantes valores son ceros. En consecuencia, en todas las tablas, las inercias de todos los ejes principales serán iguales a 1. Y, por tanto, la inercia total de la tabla $\boldsymbol{Z}_q$ será igual a su dimensionalidad, es decir, igual a $J_q - 1$. La inercia de $\boldsymbol{Z}$ será la media de las inercias de las tablas que la componen: [34]

$$inercia(\boldsymbol{Z}) = \frac{1}{Q}\sum_{q=1}^{Q} inercia(\boldsymbol{Z}_q) = \frac{1}{Q}\sum_{q=1}^{Q}(J_q - 1) = \frac{J-Q}{Q}$$

Dado que $J - Q$ es la dimensionalidad de $\boldsymbol{Z}$, la inercia media por dimensión será $\frac{1}{Q}$, el cual se utiliza como umbral para decidir para qué ejes o dimensiones es interesante interpretar el ACM (similar al valor umbral de 1 de los valores propios en el ACP) [34].

Una estructura alternativa de datos para el ACM es la matriz compuesta por todas las tablas resultantes de cruzar todas las variables de interés dos a dos, la *matriz de Burt*, para los datos del ejemplo que se está considerando, es una matriz compuesta de 4×4, formada por 16 tablas. Con excepción de las tablas de la diagonal las restantes 12 se obtienen cruzando los valores de dos variables de las 700 personas. La matriz de Burt es simétrica, por tanto, fuera de la diagonal, solo hay seis cruzamientos distintos que se transponen a ambos lados de la diagonal de la matriz compuesta. Las tablas de la diagonal corresponden a los cruces de las variables por ellas mismas, son matrices diagonales con las frecuencias marginales de la variable en su diagonal. La

[11] Análisis de Correspondencias Simple

matriz de Burt, $\boldsymbol{B}$, se relaciona de forma sencilla, con la matriz binaria $\boldsymbol{Z}$ de la manera siguiente: [34]

$$\boldsymbol{B} = \boldsymbol{Z}'\boldsymbol{Z}$$

La otra forma "*clásica*" de definir el ACM es el ACS de la matriz de $\boldsymbol{B}$. Dado que $\boldsymbol{B}$ es una matriz simétrica, las soluciones de filas y de columnas son idénticas.

5.4.7. METODO DE DEPENDENCIA: REGRESIÓN LOGÍSTICA

5.4.7.1. INTRODUCCIÓN

Un método de dependencia supone que las variables analizadas están divididas en dos grupos: las **variables dependientes** y las **variables independientes**. El objetivo de los métodos de dependencia consiste en determinar si el conjunto de variables independientes afecta al conjunto de variables dependientes y de qué forma [35]. En el presente estudio secundario de la IRC, se ha considerado una variable dependiente cualitativa binaria o dicotómica y, con el nombre de "Estado", que toma los valores de $1 = persona\ con\ IRC\ (y = si)$ y $0 = persona\ \text{sin}\ ERC\ (y = no)$, y al menos una posible variable independiente x_j , donde $j = 1, \dots, p\ variables$, ya sean cuantitativas o cualitativas [36].

En tales circunstancias, uno de los métodos de dependencia idóneo para estudiar la relación entre una o más variables independientes x_j y una variable dependiente de tipo dicotómica ($y = si$ o $y = no$), es decir, que solo admite dos categorías que definen opciones o características mutuamente excluyentes u opuestas, es el **modelo de regresión logística**, el cual se utiliza muy a menudo en el análisis de datos procedentes de investigaciones propias del ámbito de las ciencias de la salud [36]. De manera que la idea es intentar construir un modelo que explique los valores de la variable dependiente o de clasificación de dos tipos de población. El problema de discriminación se convierte en prever el valor de la variable ficticia o dicotómica y, en un nuevo elemento cuando se conocen los valores que toman las variables x_j en la persona. Por ejemplo, si el valor previsto está más próximo a cero que a uno, se clasificará al individuo en la población de no enfermos con ERC, en caso contrario se clasificará en la población con IRC [31].

En otras palabras un modelo de regresión logística permite predecir o estimar la probabilidad de que una persona sufra IRC ($y = si$) o no padezca de ERC ($y = no$) en función de determinadas características individuales (x_j) como: edad (grupos de

edades), sexo (masculino/femenino), presión arterial sistólica mayor o igual a 140 mmHg (sí o no), presión arterial diastólica mayor o igual a 90 mmHg (sí o no), fumador (si o no) antecedente personal o familiar de hipertensión arterial (si o no) [36]**.**

Los resultados de ese hipotético modelo de regresión logística podrían indicar que un hombre fumador y con un nivel de colesterol alto tiene una probabilidad elevada de padecer IRC, en comparación con un individuo que presenta otras características, tales como ser mujer y/o no fumar. También podría indicar que la probabilidad de padecer IRC aumenta con la edad y con la existencia de antecedentes personales o familiares de hipertensión arterial [36].

Algunas características de la regresión logística son que las variables independientes x_j pueden ser cualitativas binarias (sexo masculino o femenino) o categóricas (nivel de escolaridad: sin estudios, estudios primarios, bachiller, estudios universitarios) y cuantitativas (edad en años). Las variables independientes también se denominan predictivas o predictores, determinantes o explicativas. Las variables x_j , pueden determinar, en mayor o menor magnitud, la variable dependiente y, la cual también se denomina variable explicada, determinada, respuesta, predecida, predicha o criterio. Así, la variable independiente "**antecedente personal o familiar de hipertensión arterial**" podría ser la más importante al predecir la probabilidad de sufrir IRC [36].

El modelo de regresión logística también es aplicable a variables dependientes y que son politómicas (tres o más categorías). En el presente análisis secundario de IRC se aborda la regresión logística en el que la variable dependiente es dicotómica ($y = si\ IRC,$ o $y = no\ ERC$). Los resultados de un supuesto modelo de regresión logística que permiten estimar la probabilidad de que un individuo sufra IRC en función de unas determinadas características individuales permiten identificar y estimar la importancia y contribución relativa de determinadas características individuales x_j , denominadas factores de riesgo, como podrían ser fumar o tener hipertensión arterial, en la probabilidad de padecer IRC. El modelo también permite estimar o predecir la magnitud del riesgo global de padecer dicha enfermedad, cuando coinciden dos o más factores de riesgo en un mismo individuo. Así por ejemplo el hecho de ser fumador incrementaría la probabilidad de padecer IRC en un hombre con hipertensión arterial [36].

5.4.7.2. EL MODELO DE REGRESIÓN LOGÍSTICA

El modelo de regresión logística se utiliza para predecir la probabilidad estimada $P(y)$ de que la variable dependiente y presente uno de los dos valores posibles ($1 = si\ o\ 2 = no$) en función de los diferentes valores que adoptan el conjunto de variables independientes x_j . Normalmente se intenta predecir la probabilidad de que se produzca el acontecimiento o suceso definido como $y = 1$. En otras palabras, el modelo de regresión logística permite relacionar una variable dependiente dicotómica con una o más variables independientes cuantitativas y/o cualitativas. Las variables categóricas dicotómicas son aquellas que definen, mediante los indicadores $(1, 0)$, dos características mutuamente excluyentes y opuestas, como son la presencia "1" o ausencia "0" de un acontecimiento o suceso, como por ejemplo el acontecimiento que se podría predecir es la presencia de IRC [36].

Los objetivos del modelo de regresión logística, al estudiar la relación entre una variable dependiente dicotómica y y una o más variables independientes x_j, son: [36]

1. Determinar la existencia o ausencia de relación entre una o más variables independientes x_j y la variable dependiente y.
2. Medir la magnitud de dicha relación.
3. Estimar o predecir la probabilidad de que se produzca un suceso o acontecimiento definido como $\mathrm{y} = 1$ en función de los valores que adoptan las variables independientes $\mathrm{x_j}$.

El tipo de diseño del estudio y la sistemática de recogida de datos condiciona la interpretación de los resultados del modelo de regresión logística. En aquellos estudios en que los valores definidos por las variables independientes preceden en el tiempo al suceso o acontecimiento señalado por la variable dependiente, la relación entre ambos tipos de variables se explica en términos de predicción o determinación de la variable dependiente por una o más variables independientes. En cambio, en los estudios en que las características definidas por ambos tipos de variables –independientes y dependiente- se miden en el mismo momento en el tiempo, su relación se interpreta en términos de correlación o asociación [36].

La regresión logística se incluye dentro del conjunto de las denominadas técnicas estadísticas de análisis de datos. De manera particular en la regresión logística el concepto de proporción es importante porque permite establecer las primeras diferencias respecto al modelo de regresión lineal. En este último modelo mencionado

se predice el valor medio de la variable dependiente (y podría ser continua o categórica) a partir de una o más variables independientes (x_j). En cambio, los modelos de regresión logística permiten predecir la proporción de una de las dos categorías de la variable dependiente (y = dicotómica) en función de una o más variables independientes x_j [36].

5.4.7.3. COMPONENTES DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA MULTIPLE

Se refiere al problema de discriminación entre dos poblaciones, en el que se define una variable de clasificación o dependiente, y que tome el valor uno cuando el elemento pertenece a la primera población, P_1, y cero cuando pertenece a una segunda, P_2. Entonces la muestra consistirá en n elementos del tipo $(y_i, \boldsymbol{x}_i)$, donde y_i es el valor que toma la variable binaria de clasificación en el individuo i y $\boldsymbol{x}_i$ un vector de variables explicativas, por lo que, la idea es construir un modelo para prever el valor de la variable ficticia binaria en un nuevo elemento cuando se conocen las variables $\boldsymbol{x_i}$. El primer enfoque simple es formular el siguiente modelo de regresión: [31]

$$y = \alpha + \boldsymbol{\beta}'\boldsymbol{x} + u$$

Donde α representa el término independiente o constante, $\boldsymbol{\beta}$ representa el vector de coeficientes de regresión, los cuales cada uno de ellos está asociado a una variable independiente x_j contenida en el vector $\boldsymbol{x}$ y u es una variable aleatoria que se le conoce como perturbación. Sin embargo, este modelo presenta problemas de interpretación, Por lo que si se toma esperanzas en la ecuación anterior para $\boldsymbol{x} = \boldsymbol{x}_i$, donde i representa a un individuo, se tiene que: [31]

$$E[y/\boldsymbol{x}_i] = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_i$$

Ya que los valores de y dependen o vienen dados según los valores que toma $\boldsymbol{x}_i$ y se supone que el valor esperado de la perturbación es igual a cero. Llamando a P a la probabilidad estimada de que y tome el valor 1 cuando $\boldsymbol{x} = \boldsymbol{x}_i$: [31]

$$P = P(y = 1/\boldsymbol{x}_i)$$

Como y es binomial y toma los valores posibles uno y cero con probabilidades P y $1 - P$, su esperanza será: [31]

$$E[y/\boldsymbol{x}_i] = P \times 1 + (1 - P) \times 0 = P$$

Por lo tanto, se concluye que: [31]

$$P = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_i \text{ ó } P = a + \widehat{\boldsymbol{\beta}}'\boldsymbol{x}_i$$

Para diferenciar que α y $\boldsymbol{\beta}'$ representan valores correspondientes a una población, mientras que a y $\widehat{\boldsymbol{\beta}}'$ representan los valores estimados en la muestra, el termino α también se puede representar como β_0, de la misma forma en lo que respecta a la muestra, el termino a también se puede representar como $\hat{\beta}_0$.

Para que el modelo construido proporcione directamente la probabilidad de pertenecer a cada población, se debe transformar la variable respuesta para garantizar que la respuesta prevista este entre cero y uno. Escribiendo: [31]

$$P = F(\alpha + \boldsymbol{\beta}'\boldsymbol{x}_i)$$

P estará entre cero y uno si escogemos F para que tenga esa propiedad. La clase de funciones no decrecientes acotadas entre cero y uno es la clase de las funciones de distribución, por lo que el problema se resuelve tomando como F cualquier función de distribución. Una solución considerada es tomar como F la función de distribución logística, dada por: [31]

$$P = \frac{1}{1 + e^{-\alpha - \boldsymbol{\beta}'\boldsymbol{x}_i}}$$

Esta función tiene la ventaja de la continuidad. Además, como: [31]

$$1 - P = \frac{e^{-\alpha - \boldsymbol{\beta}'\boldsymbol{x}_i}}{1 + e^{-\alpha - \boldsymbol{\beta}'\boldsymbol{x}_i}} = \frac{1}{1 + e^{\alpha + \boldsymbol{\beta}'\boldsymbol{x}_i}}$$

resulta que: [31]

$$L_i = \ln\left(\frac{P}{1 - P}\right) = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_i$$

es un modelo lineal en esta transformación que se denomina $logit$. La variable $Logit$, L, representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a ambas poblaciones, y al ser una función lineal de las variables explicativas, facilita la estimación y la interpretación del modelo.

Método de estimación de los parámetros del modelo de regresión logística: [36]

Los parámetros de la ecuación de regresión logística se estiman por el **método de máxima verosimilitud.** El método se fundamenta en la estimación de los valores de α y de las β poblacionales que maximizan la función logística para el conjunto de valores muestrales. Los valores estimados a y $\hat{\beta}$ para los parámetros poblacionales α y β, respectivamente, deben reflejar lo mejor posible los datos observados en la muestra seleccionada. La ecuación logística es una expresión de la probabilidad de obtener los valores observados en la muestra en función de los parámetros incluidos en el modelo. Los parámetros estimados tienden a seguir un patrón de distribución asintóticamente normal, por lo que los coeficientes estimados $\hat{\beta}_i$ divididos por sus respectivos $s(\hat{\beta}_i)$ (error estándar en la estimación de $\hat{\beta}_i$) producen un valor de "W", llamado estadístico de Wald, que permite contrastar la significación estadística de los coeficientes estimados, mediante su comparación con una distribución normal estandarizada de valores z o chi-cuadrado.

Las principales asunciones a tener en cuenta en la elaboración de modelos de regresión logística son: [36]

1. El modelo deber estar especificado de manera correcta, por lo que las probabilidades estimadas $P(y = 1)$ son el resultado de una función logística que incluye las variables independientes x_j. En otras palabras, los $logit$ (L_i) son funciones lineales de las variables independientes x_j.
2. En el modelo no se omiten variables independientes x_j que son importantes en la predicción de la variable dependiente y.
3. Las variables independientes x_j incluidas en el modelo están en medidas sin error.
4. Los elementos de la muestra u observaciones son independientes entre sí.
5. Ninguna de las variables independientes x_j incluidas en el modelo es una función lineal de otra(s) variables $\mathrm{x_j}$. El fenómeno de relación lineal entre variables independientes se denomina multicolinealidad.

5.4.7.4. INTERPRETACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA

Los parámetros del modelo son β_0, la ordenada en el origen, y $\boldsymbol{\beta}_1 = (\beta_1, \dots, \beta_p)$, las pendientes. A veces se utilizan también como parámetros $\exp(\beta_0)$ y $\exp(\beta_i)$ [12], e indican cuánto se modifican las probabilidades por unidad de cambio en las variables x_j. En efecto, se tiene que

$$O_i = \frac{P_i}{1 - P_i} = \exp(\alpha) \prod_{j=1}^{p} \exp(\beta_j)^{x_j}$$

Supongamos dos elementos, i y k, con todos los valores de las variables iguales excepto la variable h y $x_{ih} = x_{kh} + 1$. El cociente de los ratios de probabilidades para estas dos observaciones es: [31]

$$\frac{O_i}{O_k} = e^{\beta_h}$$

e indica cuánto se modifica el ratio de probabilidades cuando la variable x_h aumenta una unidad. Suponiendo que se sustituye $\hat{P} = 0.5$ en el modelo $Logit$, entonces,

$$ln \frac{P_i}{1 - P_i} = \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip} = 0,$$

es decir,

$$x_{i1} = -\frac{\beta_0}{\beta_1} - \sum_{j=2}^{p} \frac{\beta_j \beta_{ij}}{\beta_1}$$

y x_{i1} representa el valor de x_1 que hace igualmente probable que un elemento, cuyas restantes variables son $x_{i2}, \dots, x_{ip}$, pertenezca a la primera o la segunda población.

[12] Que se denominan los *odds ratios* o ratios de probabilidades

5.4.7.5. CONTRASTES DEL MODELO DE REGRESION LOGISTICA

En primer lugar, se exponen aspectos básicos referentes a la función de verosimilitud del modelo, para poder comprender de manera lógica los contrastes.

Función de verosimilitud [31]

Supondremos una muestra aleatoria de datos (x_i, y_i), $i = 1, \dots, n$. La función de probabilidades para una respuesta y_i cualquiera es

$$P(y_i) = P_i^{y_i}(1 - P_i)^{1-y_i}, \quad y_i = 0, 1$$

y para la muestra

$$P(y_1, \dots, y_n) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}$$

Tomando logaritmos tenemos que

$$logP(\boldsymbol{y}) = \sum_{i=1}^{n} y_i \log\left(\frac{P_i}{1 - P_i}\right) + \sum_{i=1}^{n} \log(1 - P_i)$$

La función soporte (de verosimilitud en logaritmos) puede escribirse como

$$logP(\beta) = \sum_{i=1}^{n} (y_i logP_i + (1 - y_i)\log(1 - P_i))$$

donde $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ es un vector de $P + 1$ componentes, incluyendo la constante β_0 que determina las probabilidades P_i. Maximizar la verosimilitud puede expresarse como minimizar una función que mide la desviación entre los datos y el modelo. Se puede definir la desviación de un modelo mediante $D(\theta) = -2L(\theta)$ donde L es la función soporte o el logaritmo de la verosimilitud, y por tanto la desviación del modelo será:

$$D(\beta) = -2\sum_{i=1}^{n} (y_i logP_i + (1 - y_i)\log(1 - P_i)).$$

La función de verosimilitud, también se puede expresar en función de los parámetros de interés de la siguiente manera:

$$L(\beta) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i' \boldsymbol{\beta} - \sum_{i=1}^{n} log(1 + e^{x_i' \boldsymbol{\beta}})$$

Contrastes [31]

Si queremos contrastar si una variable o grupo de variables incluidas dentro de la ecuación es significativo, podemos construir un contraste de la razón de verosimilitudes comparando el máximo de la función de verosimilitud para el modelo con y sin estas variables. Supongamos que $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, donde $\boldsymbol{\beta}_1$ tiene dimensión $p - s$, y $\boldsymbol{\beta}_2$ tiene dimensión s. Se desea contrastar si el vector de parámetros:

$$H_0: \boldsymbol{\beta}_2 = 0,$$

frente a la alternativa

$$H_1: \boldsymbol{\beta}_2 \neq 0$$

El contrate de razón de verosimilitudes utiliza que $\lambda = 2L(H_1) - 2L(H_0)$, donde $L(H_1)$ es el máximo del soporte cuando estimamos los parámetros bajo H_1, y $L(H_0)$ es el máximo cuando estimamos los parámetros bajo H_0, es decir, si H_0 es cierta, una χ_s^2. Una manera equivalente de definir el contraste es llamar $D(H_0) = -2L(\widehat{\boldsymbol{\beta}}_1)$ que mide la desviación entre los datos y el modelo, cuando el modelo se estima bajo H_0, es decir, suponiendo que $\boldsymbol{\beta}_2 = 0$, y $D(H_1) = -2L(\widehat{\boldsymbol{\beta}}_1 \widehat{\boldsymbol{\beta}}_2)$ a la desviación bajo H_1. La desviación será menor con el modelo con más parámetros (la verosimilitud será siempre mayor bajo H_1 y, si H_0 es cierta, la diferencia de desviaciones, que es el contraste de verosimilitudes

$$\chi_s^2 = D(H_0) - D(H_1) = 2L(\widehat{\boldsymbol{\beta}}_1 \widehat{\boldsymbol{\beta}}_2) - 2L(\widehat{\boldsymbol{\beta}}_1)$$

se distribuye como una χ_s^2 con s grados de libertad. En particular esta prueba puede aplicarse para comprobar si un parámetro es significativo y debe dejarse en el modelo. Sin embargo, es más habitual en estos casos comparar el parámetro estimado con su desviación típica. Los cocientes

$$w_j = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}$$

se denominan estadísticos de Wald y en muestras grandes se distribuyen, si el verdadero valor del parámetro es cero, como una normal estándar. Una medida global del ajuste es:

$$R^2 = 1 - \frac{D(\widehat{\boldsymbol{\beta}})}{D_0} = 1 - \frac{\widehat{\boldsymbol{\beta}}}{L(\beta_0)}$$

donde el numerador es la desviación (verosimilitud en el máximo) para el modelo con parámetros estimados $\widehat{\boldsymbol{\beta}}$ y el denominador la desviación (verosimilitud) para el modelo que sólo incluye la constante β_0. Observemos que, en este último caso, la estimación de la probabilidad P_i es constante para todos los datos e igual a m/n siendo m el número de elementos en la muestra con la variable $y = 1$. Entonces, sustituyendo en la ecuación de desviación de un modelo, de tal manera que se obtenga la desviación máxima que corresponde al modelo más simple posible con sólo β_0, lo cual asigna la misma probabilidad a todos los datos, es:

$$D_0 = -2L(\beta_0) = -2mlog(m) - 2(n - m)\, log(n - m) + 2n \log n.$$

Por otro lado, si el ajuste es perfecto, es decir todas las observaciones con $y = 1$ tienen $P_i = 1$ y las de $y = 0$ tienen $P_i = 0$, entonces, la desviación es cero y $L(\widehat{\boldsymbol{\beta}}) = 0$ y $R^2 = 1$. Por el contrario, si las variables explicativas no influyen nada la desviación con las variables explicativas será igual que sin ellas, $L(\widehat{\boldsymbol{\beta}}) = L(\beta_0)$ y $R^2 = 0$.

6. DISCUSIÓN Y ANALISIS DE RESULTADOS

6.1. ANÁLISIS EXPLORATORIO UNIVARIANTE

En esta sección se expone los razonamientos del procedimiento para la formación de información primaria, es decir, la base de datos muestral a utilizar en los análisis estadísticos. En segundo lugar, se muestra la aplicación del análisis univariado a las variables seleccionadas, lo cual permite seleccionar, excluir y describir variables.

6.1.1. PROCEDIMIENTO PARA LA FORMACIÓN DE INFORMACIÓN PRIMARIA

Entre las 192 variables de la base de datos de Nefrolempa se encuentran dos que se distribuyen completamente a lo largo de la muestra seleccionada de 700 personas, con las cuales se crea la variable respuesta o dependiente y, que clasifica a 76 personas con IRC y 624 sin ERC, la cual es nombrada y puede ser definida como:

$$Estado = y = \begin{cases} 1 = persona\ con\ IRC \\ 0 = persona\ \text{sin}\ ERC \end{cases}$$

Luego, en el conjunto de 190 variables restantes, se encuentran características sociales, epidemiológicas y clínicas, entre las cuales se seleccionan las posibles variables explicativas (independientes), es decir, los posibles factores de riesgo o variables que se suponen podrían estar asociadas con la IRC. Dicha selección se efectúa considerando los siguientes criterios:

a) Se seleccionan variables que se suponen podrían tener algún tipo de asociación razonable y coherente con la IRC, o en otras palabras variables que podrían estar dentro del contexto o marco de estudio de la IRC.
b) Se seleccionan variables tomando como referencia las variables que fueron consideradas como posibles factores de riesgo asociados con la IRC, en el estudio Nefrolempa: [28]
    1) Sexo
    2) Edad categorizada
    3) Antecedentes familiares de ERC
    4) Diabetes mellitus
    5) Hipertensión arterial
    6) Síndrome metabólico
    7) Dislipidemia
    8) Hábito de fumar

9) Contacto con plantas medicinales
10) Enfermedades infecciosas
11) Consumo de antiinflamatorios no asteroideos
12) Contacto con agroquímicos
13) Consumo de alcohol
14) Ocupación
15) Obesidad
16) Antecedentes familiares de HTA
17) Antecedentes familiares de DM

c) Se excluyen del estudio, variables con información redundante respecto a otras seleccionadas para los análisis de la IRC.
d) Se excluyen de los análisis de la IRC, variables con muchos valores ausentes.

Por tanto, tomando en cuenta los cuatro criterios anteriores se seleccionan de la base de datos del estudio Nefrolempa, los posibles factores de riesgo o variables que se suponen podrían estar asociadas con la IRC, las cuales son:

1) Sexo
2) Grupo de edades
3) Ocupación laboral
4) Uso de plantas medicinales
5) Contacto con agroquímicos
6) Uso de medicamentos antiinflamatorios no esteroides (AINES)
7) Ocurrencia de dislipidemia
8) Ocurrencia de diabetes mellitus
9) Clasificación de hipertensión arterial (JNC7 - 2003)
10) Clasificación de IMC

6.1.2. ANÁLISIS UNIVARIADO DE POSIBLES VARIABLES EXPLICATIVAS Y LA VARIABLE DEPENDIENTE

En esta sección se expone el análisis univariado de las variables que se seleccionaron en la sección 6.1.1, con el objeto de conocer el tipo de información contenida, identificar datos ausentes, justificar la posible exclusión de variables basado en la ausencia de muchos valores, describir recuentos y porcentajes por medio de tablas y gráfico de barras, de manera que revele las variables más destacadas, lo cual ayuda en las decisiones sobre cuáles variables pueden ser incluidas en un Análisis de Correspondencias Múltiples (ACM).

A continuación, se presentan recuentos y porcentajes de variables seleccionadas en la sección anterior, por medio de tablas y gráficos de barras.

En primer lugar, la prevalencia de IRC es: 10.9% (76/700 personas), según la muestra seleccionada de la región del Bajo Lempa, mientras que el 89.1% (624/700 personas) no padecen de ERC. En los siguientes gráficos (Véase Figura 6.1 y Figura 6.2), se muestran las distribuciones de estos recuentos y porcentajes o prevalencias:

Figura 6.1. Recuento de personas con IRC y sin ERC, en la muestra seleccionada de 700 personas de la región del Bajo Lempa.

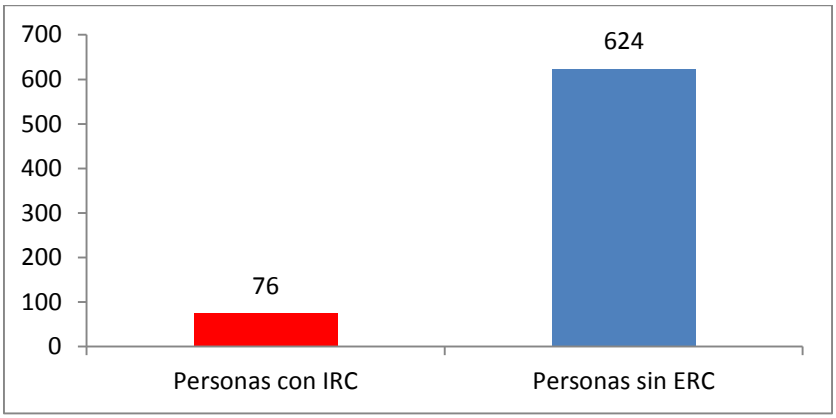
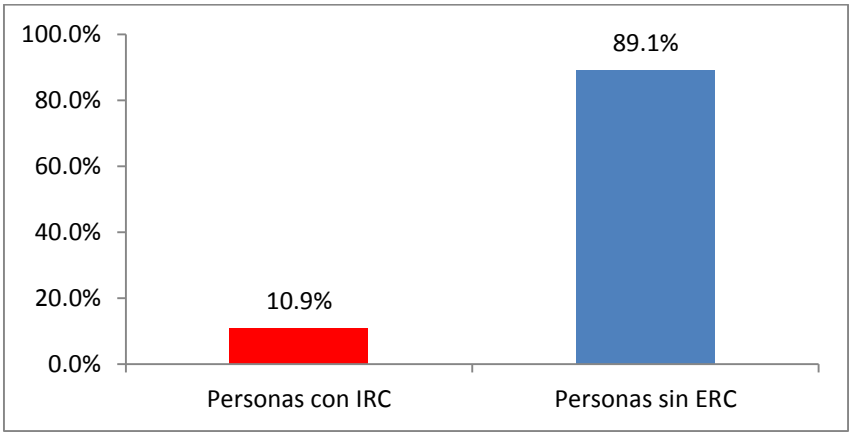


Figura 6.2. Prevalencia de personas con IRC y sin ERC, en la muestra seleccionada de 700 personas de la región del Bajo Lempa.



En la Tabla 6.1, se muestran los recuentos y porcentajes de las posibles variables explicativas, para conocer las categorías más destacadas y a su vez identificar la cantidad de valores ausentes en cada variable.

Tabla 6.1. Recuentos y porcentajes de la muestra (700 personas) según posibles variables asociadas con la IRC.

Nombre de Variable	Categorías	Recuento	Porcentaje (%)
Sexo	Femenino	396	56.6
	Masculino	304	43.4
	Total	700	100.0
Grupo de edades	18 a 29	267	38.1
	30 a 39	134	19.1
	40 a 49	122	17.4
	50 a 59	93	13.3
	60 a 69	45	6.4
	mayor o igual a 70	39	5.6
	Total	700	100.0
	Ausentes	0	0.0
Ocupación laboral	Agricultor	230	32.9
	Agricultor y ama de casa	5	0.7
	Agricultor y fumigador	31	4.4
	Ama de casa	278	39.7
	Desempleado	15	2.1
	Estudiante	2	0.3
	Enfermera	39	5.6
	Laboratorio clínico	1	0.1
	Médico	1	0.1
	Odontólogo	2	0.3
	Psicólogo	1	0.1
	Promotor de salud	4	0.6
	Otros	91	13.0
	Total	700	100.0
	Ausentes	0	0.0
Uso de plantas medicinales	Si usa	446	63.9
	No usa	252	36.1
	Total	698	100.0
	Ausentes	2	0.3
Contacto con agroquímicos	Con contacto	410	58.7
	Sin contacto	289	41.3
	Total	699	100.0
	Ausentes	1	0.1
Uso de medicamentos antiinflamatorios no esteroides (AINES)	Si usa	521	74.5
	No usa	178	25.5
	Total	699	100.0
	Ausente	1	0.1
Ocurrencia de dislipidemia	Con dislipidemia	439	62.7
	Sin dislipidemia	261	37.3
	Total	700	100.0
	Ausentes	0	0.0
Ocurrencia de diabetes mellitus	Con diabetes	67	9.6
	Sin diabetes	633	90.4
	Total	700	100.0
	Ausentes	0	0.0
Clasificación de hipertensión arterial (JNC7 - 2003)	Normal	407	58.2
	Pre-hipertensión	255	36.5
	Hipertensión arterial estadio 1	22	3.1
	Hipertensión arterial estadio 2	15	2.1
	Total	699	100.0
	Ausentes	1	0.1
Clasificación de IMC	Peso bajo	17	2.5
	Peso normal	276	40.6
	Sobrepeso	240	35.3

Nombre de Variable	Categorías	Recuento	Porcentaje (%)
	Obesidad	147	21.6
	Total	680	100.0
	Ausentes	20	2.9

En la Tabla 6.1, se observa que las variables “Grupo de edades”, “Ocupación laboral”, “Clasificación de hipertensión arterial (JNC7 - 2003)”, “Clasificación de IMC”, son politómicas, las cuales a excepción de “Grupo de edades”, son transformadas a dicotómicas con la finalidad de establecer en análisis posteriores asociaciones más convenientes con la IRC. En tales circunstancias la operacionalización de estas nuevas variables dicotómicas se puede ver en la Tabla 6.2.

Tabla 6.2. Operacionalización de variables dicotómicas posiblemente asociadas con la IRC, anteriormente definidas como politómicas.

Nombre anterior de variable politómica	Nombre actual de variable transformada en dicotómica	Código de variable dicotómica	Descripción de variable dicotómica	Valores de variable dicotómica
Ocupación laboral	Ocupación Agrícola	Agricultor	Si la persona es o no agricultor(a)	1=agricultor(a) (si) 0=no agricultor(a) (no)
Clasificación de hipertensión arterial (JNC7 - 2003)	Ocurrencia de hipertensión arterial	Ocurre_hipertensión	Si la persona tiene o no hipertensión	1=con hipertensión (si) 2=sin hipertensión (no)
Clasificación de IMC	Ocurrencia de obesidad	Ocurre_obesidad	Si la persona es o no obesa	1=con obesidad (si) 0=sin obesidad (no)

Luego en la Tabla 6.3, se presenta los recuentos y porcentajes de las variables dicotómicas definidas en la Tabla 6.2.

Tabla 6.3. Recuentos y porcentajes de variables dicotómicas posiblemente asociadas con la IRC, anteriormente definidas como politómicas.

Nombre de variable	Categorías	Recuento	Porcentaje
Ocupación agrícola	Agricultor(a)	266	38.0%
	No agricultor(a)	434	62.0%
	Total	700	100.0%
	Ausentes	0	0.0%
Ocurrencia de hipertensión arterial	Con hipertensión	37	5.3%
	Sin hipertensión	662	94.7%
	Total	699	100.0%
	Ausentes	1	0.1%
Ocurrencia de obesidad	Con obesidad	387	56.9%
	Sin obesidad	293	43.1%
	Total	680	100.0%
	Ausentes	20	2.9%

6.2. ANÁLISIS EXPLORATORIO BIVARIADO ENTRE POSIBLES VARIABLES EXPLICATIVAS Y LA VARIABLE DEPENDIENTE

En esta sección, se presenta un análisis bivariado que describe las prevalencias de las posibles variables explicativas, distribuidas por “Sexo” y la variable dependiente y (Estado). También, se detalla la prevalencia de la IRC según posibles variables explicativas, junto con la cantidad de personas con IRC y sin ERC por cada categoría. Así mismo, se presentan prevalencias a través de gráficos de barra con una estructura de correspondencia de dos o tres variables, con el objeto de visualizar de mejor manera los resultados. Luego se muestran los análisis de asociaciones entre la variable respuesta y el resto de variables, en base a la hipótesis nula de independencia entre dos variables, los cuales son contrastados por medio del estadístico chi-cuadrado, a un nivel de significancia de 5%. Por lo cual, las variables que están asociadas o correlacionadas significativamente con la variable respuesta, podrían ser tomadas en cuenta en las posibilidades de incluirlas en la construcción de un modelo de regresión logística binaria.

6.2.1. ANÁLISIS BIVARIADO DESCRIPTIVO ENTRE LA VARIABLE DEPENDIENTE Y LOS POSIBLES FACTORES DE RIESGO ASOCIADOS CON LA IRC

En primer lugar, se presenta en la Tabla 6.4, los porcentajes de variables que se suponen podrían estar asociadas con la variable respuesta y (Estado), distribuidos primeramente por consideraciones informativas por “Sexo”.

Tabla 6.4. Prevalencia de posibles variables explicativas por Sexo (sexo femenino: 396; sexo masculino: 304).

Nombre de Variable	Categorías	Sexo (%)	
		Masculino	Femenino
Grupo de edades	18 a 29	37.2	38.9
	30 a 39	18.1	19.9
	40 a 49	14.8	19.4
	50 a 59	13.8	12.9
	60 a 69	7.9	5.3
	Igual o mayor a 70	8.2	3.5
	Total	100.0	100.0
Ocupación agrícola	Agricultor(a)	80.3	5.6
	No agricultor(a)	19.7	94.4
	Total	100.0	100.0
Uso de plantas medicinales	Si usa	58.2	68.3
	No usa	41.8	31.7
	Total	100.0	100.0
Contacto con agroquímicos	Con contacto	89.8	34.7
	Sin contacto	10.2	65.3
	Total	100.0	100.0
Uso de medicamentos antiinflamatorios no	Si usa	71.7	76.7

Nombre de Variable	Categorías	Sexo (%)	
		Masculino	Femenino
esteroides (AINES)	No usa	28.3	23.3
	Total	100.0	100.0
Ocurrencia de dislipidemia	Con dislipidemia	64.1	61.6
	Sin dislipidemia	35.9	38.4
	Total	100.	100.0
Ocurrencia de diabetes mellitus	Con diabetes	9.2	9.8
	Sin diabetes	90.8	90.2
	Total	100.0	100.0
Ocurrencia de hipertensión arterial	Con hipertensión	5.3%	5.3%
	Sin hipertensión	94.7%	94.7%
	Total	100.0	100.0
Ocurrencia de obesidad	Con obesidad	45.2%	65.8%
	Sin obesidad	54.8%	34.2%
	Total	100.0	100.0

De la misma manera, se muestra en la Tabla 6.5 las prevalencias de los posibles factores de riesgo asociados con la IRC, según la variable dependiente y (Estado).

Tabla 6.5. Prevalencia de posibles variables explicativas por Estado (personas con IRC: 76; personas sin ERC: 624).

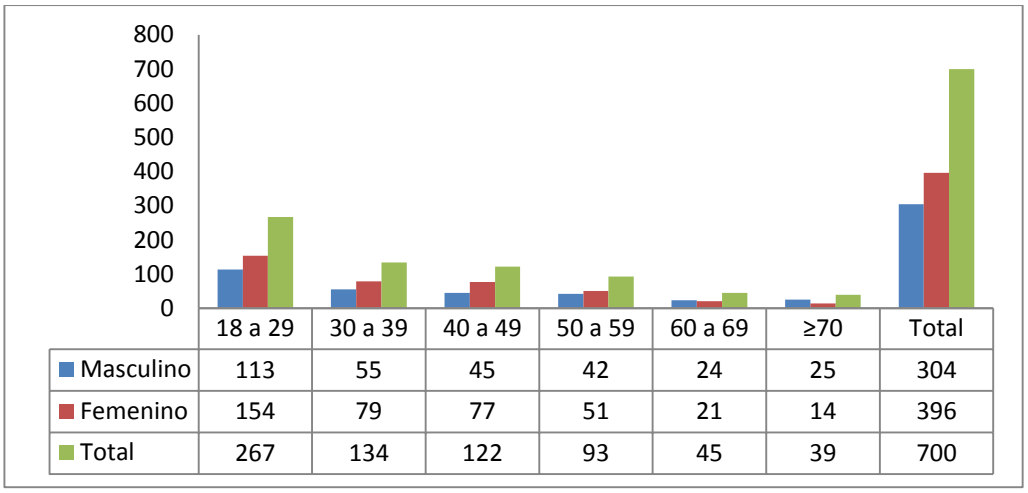
Nombre de variable	Categorías	Estado (%)	
		IRC	No ERC
Sexo	Femenino	23.7	60.6
	Masculino	76.3	39.4
	Total	100.0	100.0
Grupo de edades	18 a 29	1.3	42.6
	30 a 39	6.6	20.7
	40 a 49	17.1	17.5
	50 a 59	28.9	11.4
	60 a 69	18.4	5.0
	Igual o mayor a 70	27.6	2.9
	Total	100.0	100.0
Ocupación agrícola	Agricultor(a)	69.7%	34.1%
	No agricultor(a)	30.3%	65.9%
	Total	100	100
Uso de plantas medicinales	Si usa	68.4	63.3
	No usa	31.6	36.7
	Total	100.0	100.0
Contacto con agroquímicos	Con contacto	80.3	56.0
	Sin contacto	19.7	44.0
	Total	100.0	100.0
Uso de medicamentos antiinflamatorios no esteroides (AINES)	Si usa	72.4	74.8
	No usa	27.6	25.2
	Total	100.0	100.0
Ocurrencia de Dislipidemia	Con dislipidemia	73.7	61.4
	Sin dislipidemia	26.3	38.6
	Total	100.	100.0
Ocurrencia de diabetes mellitus	Con diabetes	21.1	8.2
	Sin diabetes	78.9	91.8
	Total	100.0	100.0
Ocurrencia de hipertensión arterial	Con hipertensión	19.7%	3.5%
	Sin hipertensión	80.3%	96.5%
	Total	100.0	100.0
Ocurrencia de obesidad	Con obesidad	48.7%	57.9%

Nombre de variable	Categorías	Estado (%)	
		IRC	No ERC
	Sin obesidad	51.3%	42.1%
	Total	100.0	100.0

A continuación, desde la Figura 6.3 hasta la

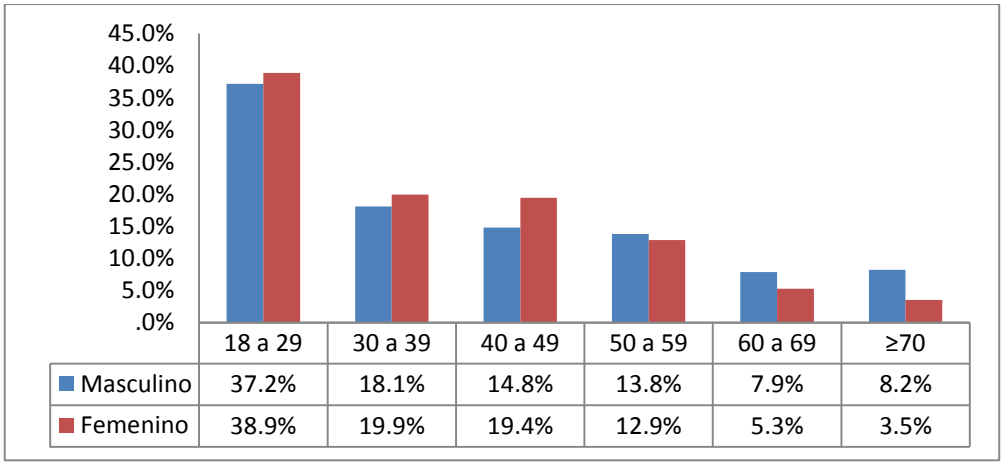
Figura 6.6 se muestran gráficos de barra para ilustrar de mejor manera recuentos y prevalencias de la variable politómica “Grupo de edades”, según “Sexo” y “Estado”.

Figura 6.3. Distribución de la muestra por Sexo y Grupo de edades.



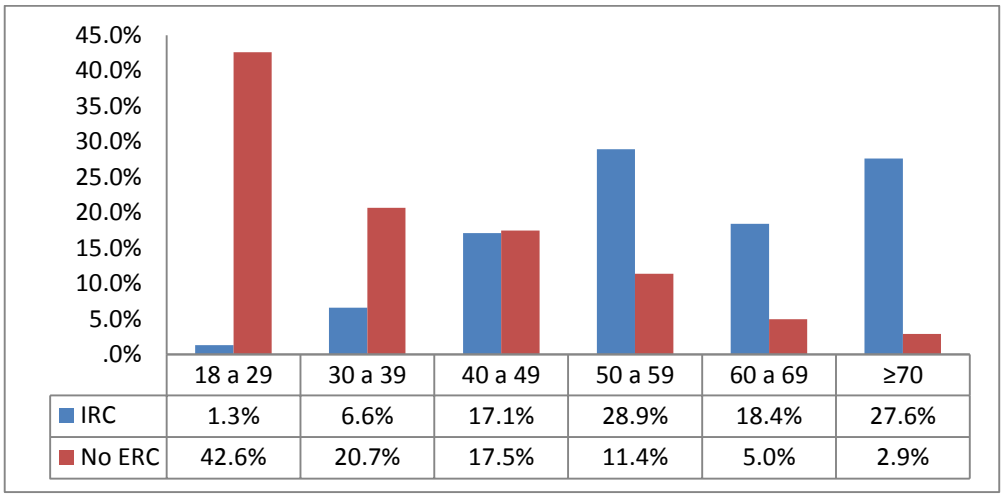
	18 a 29	30 a 39	40 a 49	50 a 59	60 a 69	≥70	Total
Masculino	113	55	45	42	24	25	304
Femenino	154	79	77	51	21	14	396
Total	267	134	122	93	45	39	700

Figura 6.4. Prevalencia de Grupo de edades por Sexo.



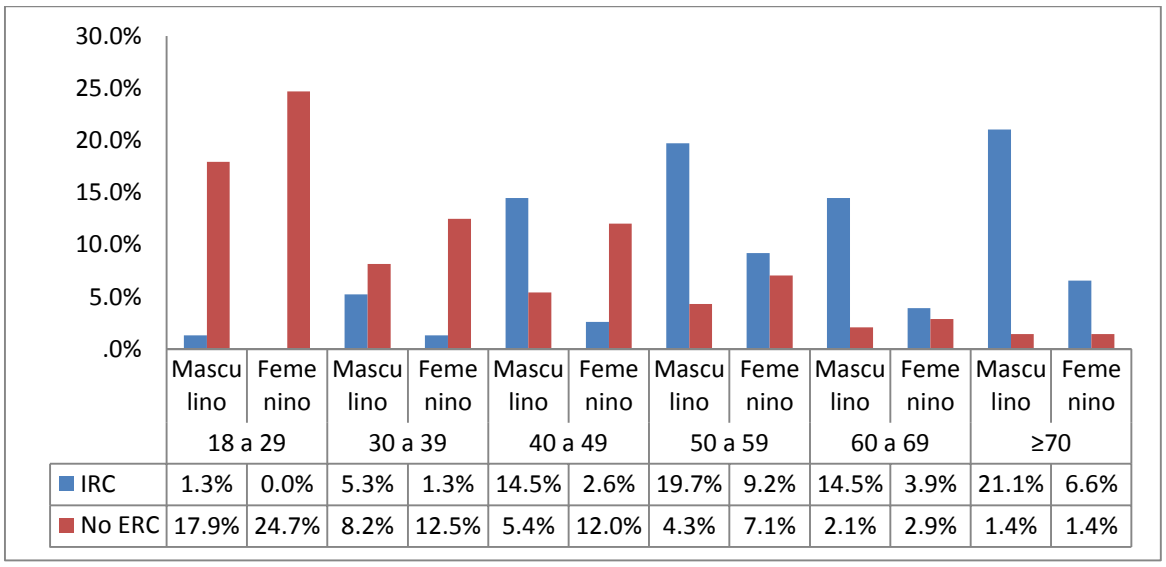
	18 a 29	30 a 39	40 a 49	50 a 59	60 a 69	≥70
Masculino	37.2%	18.1%	14.8%	13.8%	7.9%	8.2%
Femenino	38.9%	19.9%	19.4%	12.9%	5.3%	3.5%

Figura 6.5. Prevalencia de Grupo de edades por Estado.



	18 a 29	30 a 39	40 a 49	50 a 59	60 a 69	≥70
IRC	1.3%	6.6%	17.1%	28.9%	18.4%	27.6%
No ERC	42.6%	20.7%	17.5%	11.4%	5.0%	2.9%

Figura 6.6. Prevalencia de Grupo de edades y Sexo según Estado.



	Masculino	Femenino	Masculino	Femenino	Masculino	Femenino	Masculino	Femenino	Masculino	Femenino	Masculino	Femenino
	18 a 29		30 a 39		40 a 49		50 a 59		60 a 69		≥70	
IRC	1.3%	0.0%	5.3%	1.3%	14.5%	2.6%	19.7%	9.2%	14.5%	3.9%	21.1%	6.6%
No ERC	17.9%	24.7%	8.2%	12.5%	5.4%	12.0%	4.3%	7.1%	2.1%	2.9%	1.4%	1.4%

Seguidamente se presentan recuentos y porcentajes de la variable dependiente y (Estado), según los posibles factores de riesgo asociados con la IRC, resumidos en la Tabla 6.6.

Tabla 6.6. Prevalencia y recuentos de la variable dependiente y (IRC: persona con IRC; No ERC: persona sin ERC) según posibles variables explicativas.

Nombre de variable	Categorías	Estado					
		IRC		No ERC		Total	
		Recuento	Porcentaje	Recuento	Porcentaje	Recuento	Porcentaje
Sexo	Masculino	58	19.1%	246	80.9%	304	100.0%
	Femenino	18	4.5%	378	95.5%	396	100.0%
Grupo de edades	18 a 29	1	0.4%	266	99.6%	267	100.0%
	30 a 39	5	3.7%	129	96.3%	134	100.0%
	40 a 49	13	10.7%	109	89.3%	122	100.0%
	50 a 59	22	23.7%	71	76.3%	93	100.0%
	60 a 69	14	31.1%	31	68.9%	45	100.0%
	≥70	21	53.8%	18	46.2%	39	100.0%
Ocupación agrícola	Agricultor(a)	53	19.9%	213	80.1%	266	100.0%
	No agricultor(a)	23	5.3%	411	94.7%	434	100.0%
Uso de plantas medicinales	Si usa	52	11.7%	394	88.3%	446	100.0%
	No usa	24	9.5%	228	90.5%	252	100.0%
Contacto con agroquímicos	Con contacto	61	14.9%	349	85.1%	410	100.0%
	Sin contacto	15	5.2%	274	94.8%	289	100.0%
Uso de medicamentos antiinflamatorios no esteroides (AINES)	Si usa	55	10.6%	466	89.4%	521	100.0%
	No usa	21	11.8%	157	88.2%	178	100.0%
Ocurrencia de dislipidemia	Con dislipidemia	56	12.8%	383	87.2%	439	100.0%
	Sin dislipidemia	20	7.7%	241	92.3%	261	100.0%
Ocurrencia de diabetes mellitus	Con diabetes	16	23.9%	51	76.1%	67	100.0%
	Sin diabetes	60	9.5%	573	90.5%	633	100.0%
Ocurrencia de hipertensión arterial	Con hipertensión	15	40.5%	22	59.5%	37	100.0%
	Sin hipertensión	61	9.2%	601	90.8%	662	100.0%
Ocurrencia de obesidad	Con obesidad	37	9.6%	350	90.4%	387	100.0%
	Sin obesidad	39	13.3%	254	86.7%	293	100.0%

A continuación, se muestra un análisis bivariado en el que se determinan asociaciones entre la variable respuesta y (Estado) y las posibles variables explicativas.

6.2.2. ANÁLISIS BIVARIADO DE CORRESPONDENCIAS SIMPLES: PRUEBAS DE INDEPENDENCIA ENTRE LA VARIABLE DEPENDIENTE Y LOS POSIBLES FACTORES RIESGO ASOCIADOS CON LA IRC (CONTRASTE DEL ESTADÍSTICO CHI-CUADRADO)

En esta sección, se muestra el análisis bivariado de asociaciones entre variables explicativas y la variable dependiente y (Estado), por medio de la prueba del estadístico chi-cuadrado para la independencia entre dos variables, a un nivel probabilístico de significancia de 5%. En primer lugar, en la Tabla 6.7 se expone los resultados de las pruebas de asociación concernientes a variables que resultaron correlacionadas de manera significativa con la variable respuesta.

Tabla 6.7. Resultados del análisis bivariado de asociaciones significativas entre posibles variables explicativas y la variable dependiente y (Estado).

Pruebas de chi-cuadrado de Pearson			
Nombre de posible variable explicativa	**Chi-cuadrado**	**Grados de libertad**	**Probabilidad de significancia**
Sexo	37.533	1	0.000
Grupo de edades	146.633	5	0.000
Ocupación agrícola	36.448	1	0.000
Contacto con agroquímicos	16.417	1	0.000
Ocurrencia de Dislipidemia	4.388	1	0.036
Ocurrencia de diabetes mellitus	12.984	1	0.000
Ocurrencia de hipertensión arterial	35.485	1	0.000

De igual manera, en la Tabla 6.8 se muestran los resultados de las pruebas de asociaciones bivariadas, realizadas mediante el estadístico chi-cuadrado, para mostrar cuales son las variables que no están correlacionadas de manera significativa con la variable dependiente “Estado”.

Tabla 6.8. Resultados del análisis bivariado de asociaciones no significativas entre variables y la variable dependiente y (Estado).

Pruebas de chi-cuadrado de Pearson			
Nombre de variable	**Chi-cuadrado**	**Grados de libertad**	**Probabilidad de significancia**
Uso de plantas medicinales	0.757	1	0.384
Uso de medicamentos antiinflamatorios no esteroides (AINES)	0.211	1	0.646
Ocurrencia de obesidad	2.362	1	0.124

Por lo tanto, en análisis multivariantes posteriores se les da prioridad a las variables que de manera bivariante resultan estar asociadas significativamente con la variable respuesta, es decir, las variables de la Tabla 6.7. Por tanto dichas variables son consideradas como más confiadamente posibles variables explicativas o factores de riesgo asociados con la IRC.

6.3. DETERMINACIÓN DE VARIABLES ASOCIACIADAS CON LA INSUFICIENCIA RENAL CRÓNICA A TRAVÉS DE UN ABORDAJE DESCRIPTIVO MULTIVARIANTE

Esta sección trata sobre la determinación de correlaciones o asociaciones entre la variable dependiente y (Estado: 1 = persona con IRC, 0 = persona sin ERC) y las posibles variables explicativas identificadas en la sección 6.2.2, a través de un enfoque descriptivo multivariante, utilizando la técnica de **Análisis de Correspondencias Múltiples (ACM)**, ya que los factores de riesgo que podrían estar asociados con la IRC se manifiestan en un momento determinado de manera conjunta en el individuo y no de forma aislada.

Por lo tanto, el procedimiento de análisis consiste en seleccionar grupos de posibles variables explicativas con la característica de que sean aproximadamente homogéneas dentro de cada grupo. Luego se determinan asociaciones entre cada grupo y la variable dependiente y (Estado) a través del diagrama conjunto de puntos de categorías proporcionado por el ACM. Teniendo en cuenta que cuando no se logra obtener una clara interpretación de las asociaciones en el gráfico de categorías, debido a la cantidad de variables incluidas o por que el porcentaje de variabilidad de datos resultante es inadecuado, es decir, por debajo de 50%, se utiliza en tales circunstancias la técnica de **Análisis de Conglomerados Jerárquico (ACJ)**, para poder agrupar variables en grupos aproximadamente homogéneos en función de las

similitudes entre ellas generadas en base a la distancia chi-cuadrada, y de esta manera tener una orientación o guía para saber con qué grupo de variables es posible obtener asociaciones creíbles con la variable respuesta que se puedan explicar a través del ACM.

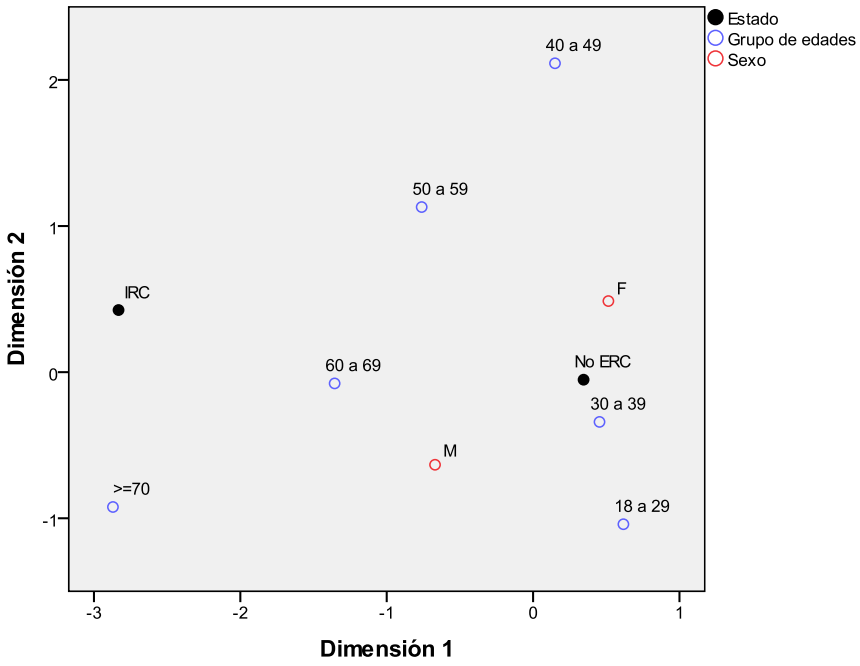
Determinación de asociaciones entre Estado y Sexo, Grupo de edades, a través del ACM

En el análisis bivariado se presentaba a través de tablas y gráficos de barra la correspondencia entre “Estado” y “Sexo”, y luego “Sexo” y “Grupo de edades". Sin embargo, es importante determinar de manera conjunta, el comportamiento de la existencia de relaciones lineales entre las categorías de estas tres variables, a través de una herramienta alternativa como es el caso del ACM, de la siguiente manera:

Tabla 6.9. Porcentajes de variabilidad 1.

Dimensión	Varianza (Autovalor)	% de la varianza
1	1.561	52.038
2	1.035	34.493
Total	2.596	86.531
Media	1.298	43.265

Figura 6.7. Diagrama conjunto de puntos de categorías 1.



La Tabla 6.9, muestra que el porcentaje de variabilidad de las variables “Grupo de edades”, “Sexo” y “Estado”, explicada por las dos dimensiones que construyen el diagrama conjunto de puntos de categorías que se muestra en la Figura 6.7 es igual a 86.5%, el cual indica un porcentaje significativo de confianza para determinar asociaciones entre categorías de variables, a través del gráfico de la Figura 6.7.

Por lo tanto, se observa que la categoría que representa el grupo de personas con edades mayores o iguales a 70 años es la que se asocia mayormente con la categoría que representa a las personas con IRC, o viceversa. También se observa que el grupo de edades de 60 a 69 años se asocia un poco más con la IRC que con el hecho de no padecer ERC. En conclusión, según los resultados obtenidos por el ACM, se plantea lo siguiente.

Se destaca que la IRC se correlaciona o se identifica más con las categorías:

- Grupo de edades mayores o iguales a 70 años.
- Grupo de edades de 60 a 69 años.

Luego, se estima que las categorías que están más asociadas con el hecho de no padecer ERC, que con la IRC son:

- Grupo de edades de 30 a 39 años.
- Grupo de edades de 18 a 29 años.
- Sexo femenino

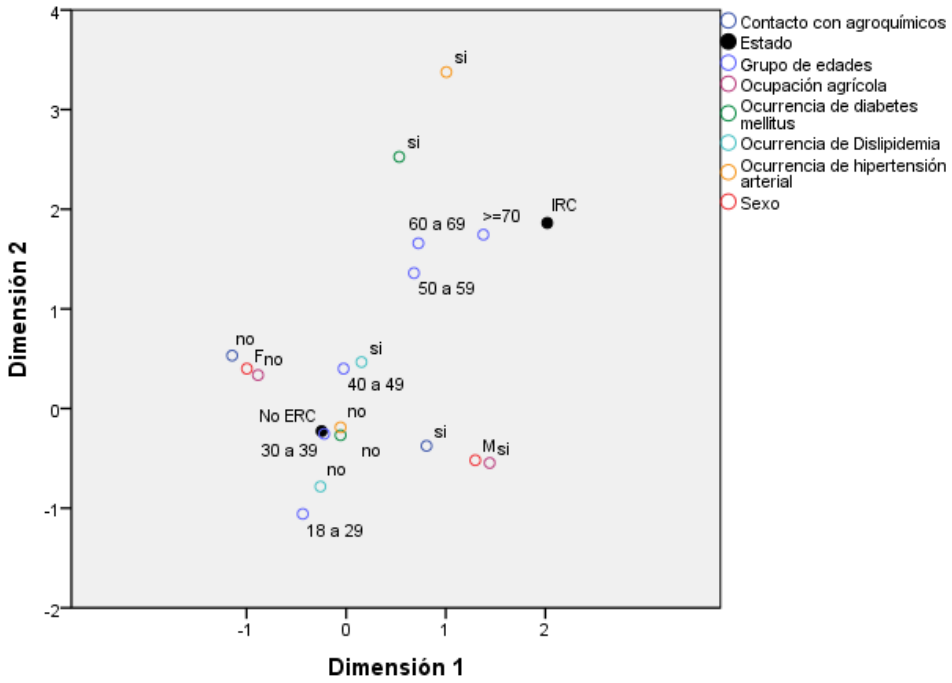
Otro aspecto importante por mencionar es que la categoría “masculino” se asocia más con la IRC que la categoría “femenino”, la cual a su vez se asocia más a la ausencia de ERC que con la IRC. O en otras palabras, se podría decir que en base a esta muestra de datos, las mujeres tienden a identificarse más con el hecho de no padecer ERC que con la IRC.

Determinación de asociaciones entre Estado y posibles variables explicativas, a través del ACM

Tabla 6.10. Porcentajes de variabilidad 2

Dimensión	Varianza (Autovalor)	% de la varianza
1	2.419	30.232
2	1.758	21.971
Total	4.176	52.203
Media	2.088	26.101

Figura 6.8. Diagrama conjunto de puntos de categorías 2.



La Tabla 6.10, muestra que el porcentaje de variabilidad de categorías explicada por las dos dimensiones que construyen el diagrama de categorías de la Figura 6.8, es igual a 52.1%, el cual es un porcentaje moderado de confianza que puede permitir identificar asociaciones entre categorías, mediante el gráfico de la Figura 6.8.

De esta manera se puede destacar de manera general, que las categorías que están más correlacionadas o asociadas con la IRC que con el no padecimiento de la ERC son:

- Grupo de edades mayores o iguales a 70 años.
- Grupo de edades de 60 a 69 años.
- Grupo de edades de 50 a 59 años.
- Con hipertensión
- Con diabetes mellitus

También, se aprecia que el hecho de no padecer ERC se correlaciona o se identifica de mejor manera con las categorías:

- Grupo de edades de 30 a 39 años.
- Sin hipertensión arterial
- Sin diabetes mellitus

Luego se logra estimar que la categoría que representa el grupo de edades de 18 a 29 años se identifica más con la ausencia de ERC, que con la IRC.

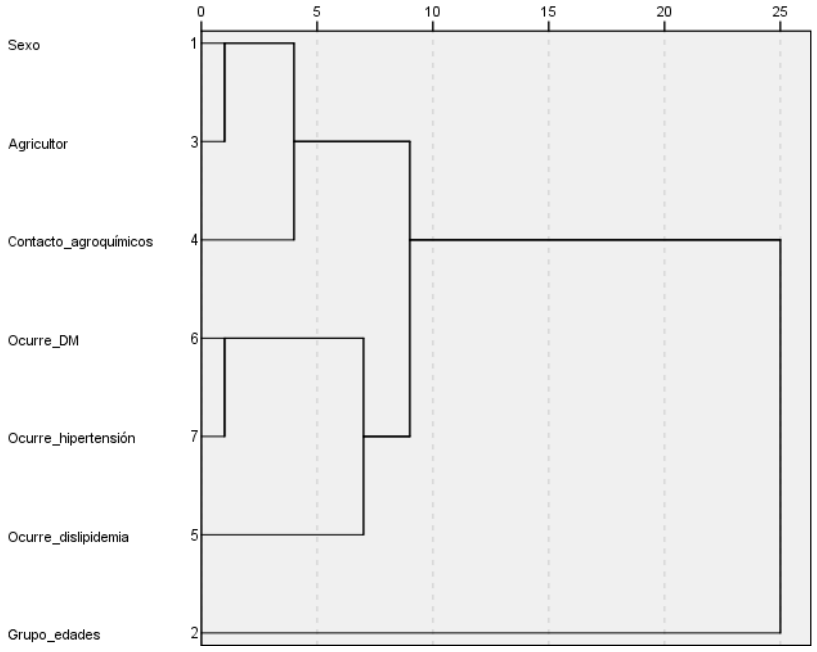
Del mismo modo, se alcanza a evaluar que la categoría que representa a las personas con dislipidemia está más asociada con la IRC, que la categoría que representa a las personas sin dislipidemia, la cual a su vez, se asocia más al hecho de no padecer ERC que con la IRC.

Determinación de asociaciones entre Estado y posibles variables explicativas, a través de ACJ Y ACM

En primer lugar se aplica un ACJ a las posibles variables explicativas, para conocer sus similitudes y así poder agruparlas en grupos aproximadamente homogéneos, para luego se aplica el ACM por cada grupo formado unido con la variable respuesta y (Estado), de manera que se puedan determinar asociaciones más admisibles.

ACJ aplicado a posibles variables explicativas

Figura 6.9. Dendrograma para la formación de grupos de posibles variables explicativas.



En el dendrograma de la Figura 6.9, se observa que se pueden formar los siguientes tres grupos de variables:

Grupo 1 de posibles variables explicativas:

- Sexo
- Ocupación agrícola
- Contacto con agroquímicos

Grupo 2 de posibles variables explicativas:

- Ocurrencia de diabetes mellitus
- Ocurrencia de hipertensión arterial
- Ocurrencia de dislipidemia

Grupo 3 de posibles variables explicativas:

- Grupo de edades

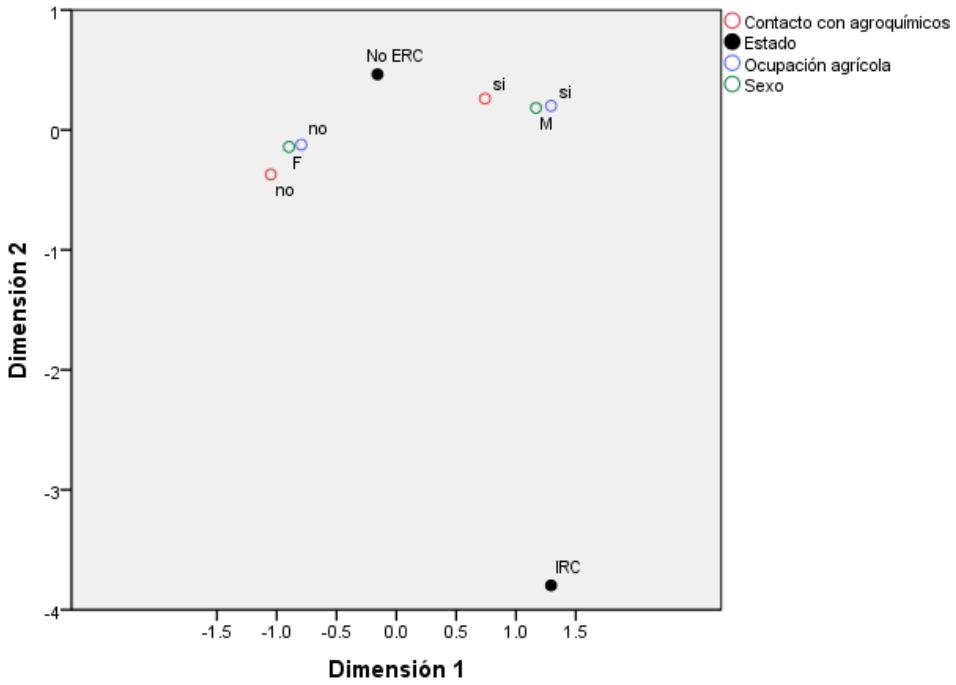
De tal manera que a continuación, se presenta la determinación de asociaciones entre la variable dependiente y los grupos 1 y 2 de posibles variables explicativas anteriormente formados, para determinar asociaciones por medio de la aplicación del ACM.

Determinación de asociaciones entre Estado y grupo 1 de posibles variables explicativas, a través del ACM

Tabla 6.11. Porcentajes de variabilidad 3

Dimensión	Varianza (Autovalor)	% de la varianza
1	2.335	58.373
2	0.909	22.718
Total	3.244	81.091
Media	1.622	40.546

Figura 6.10. Diagrama conjunto de puntos de categorías 3



En la Tabla 6.11 se observa que el porcentaje de variabilidad de categorías explicada por las dos dimensiones que construyen el gráfico de categorías de la Figura 6.10, es igual a 81.1%, el cual es un porcentaje significativo que permite determinar asociaciones plausibles entre categorías, mediante el diagrama de la Figura 6.10.

En consecuencia se puede enfatizar de manera conjunta, que el tipo de asociación entre las categorías de las variables “Contacto con agroquímicos”, “Ocupación agrícola” y “Sexo”, se identifica más con el hecho de no padecer ERC, que con el fenómeno de la IRC.

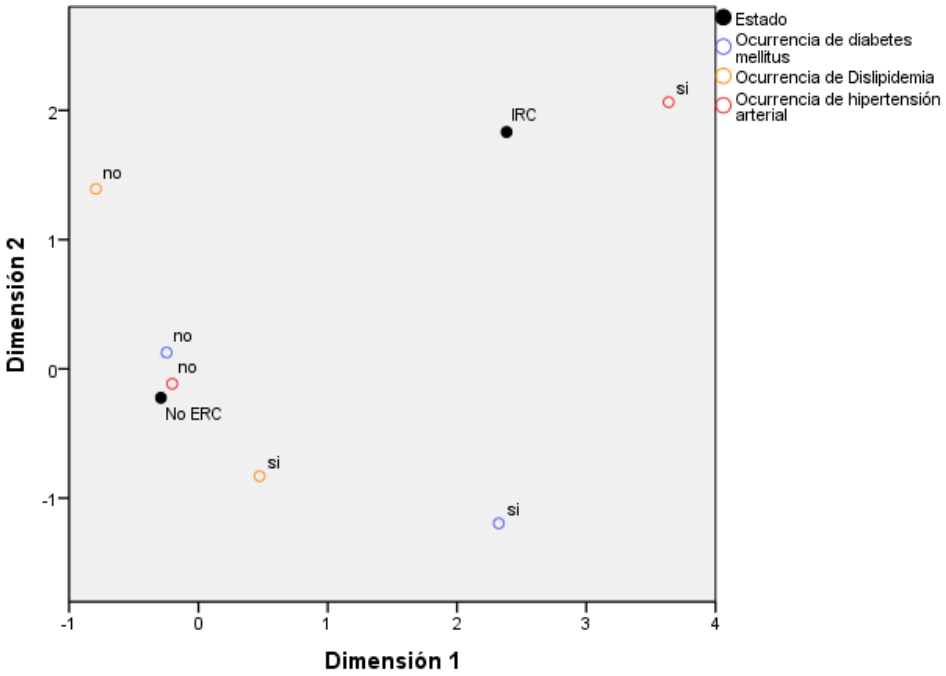
También se resalta la situación de que las mujeres agricultoras se identifican un poco más con el hecho de no padecer ERC, que los hombres agricultores.

Determinación de asociaciones entre Estado y grupo de posibles variables explicativas, a través del ACM

Tabla 6.12. Porcentajes de variabilidad 4

Dimensión	Varianza (Autovalor)	% de la varianza
1	1.415	35.371
2	0.955	23.866
Total	2.369	59.237
Media	1.185	29.619

Figura 6.11. Diagrama conjunto de puntos de categorías 4



La Tabla 6.12 muestra que el porcentaje de variabilidad de categorías explicada por las dos dimensiones que construyen el gráfico de la Figura 6.11 es aproximadamente igual a 60%, el cual indica un porcentaje moderado de confianza que puede permitir determinar asociaciones entre categorías, por medio del diagrama de la Figura 6.11.

Por consiguiente se observa que la categoría que esta mayormente asociada con la IRC es:

- Con hipertensión arterial

Por otro lado se tiene que las categorías más asociadas con el hecho de no padecer ERC:

- Sin diabetes mellitus
- Sin hipertensión arterial

Además se recalca que la categoría que representa a las personas sin diabetes mellitus está más próxima o más asociada con la ausencia de ERC, que la categoría que representa a las personas con diabetes.

6.4. REGRESIÓN LOGÍSTICA PARA LA DETERMINACIÓN DE FACTORES DE RIESGO ASOCIADOS CON LA INSUFICIENCIA RENAL CRÓNICA

Esta sección consiste sobre la determinación de factores de riesgo asociados con la IRC, a través de la aplicación de un modelo de regresión logística binaria multivariante, lo cual para realizarlo se utilizaran las posibles variables explicativas (independientes) o factores de riesgo que se suponen podrían estar asociados con la IRC, las cuales, dichas variables fueron identificadas en la sección 6.2.2.

Se muestran el $odds\ ratio$ (OR) de cada asociación significativa entre la variable respuesta y una posible variable explicativa, que luego será comparado con el OR obtenido por el modelo. Luego, se verifica el nivel de multicolinealidad entre las posibles variables explicativas, por medio de una matriz de correlación, para prevenir inconvenientes en las estimaciones del modelo.

Por consiguiente, se construye el modelo de regresión logística incluyendo las posibles variables explicativas que resultaron dentro del análisis bivariado asociadas significativamente con la variable respuesta. Luego se evalúa el modelo de regresión logística obtenido, por medio de pruebas de bondad de ajuste a los datos, siempre y cuando proporcione asociaciones significativas con la variable dependiente y que clasifique o pronostique aceptablemente a la mayor parte de los individuos. Finalmente se realizan los análisis e interpretaciones de los resultados.

En primer lugar en la Tabla 6.13 se muestran los OR que miden las magnitudes de las asociaciones significativas bivariadas entre la variable respuesta y las posibles variables explicativas.

Tabla 6.13. Análisis bivariado para el cálculo de $odds\ ratio$ (OR) de asociaciones entre las posibles variables explicativas y la variable dependiente y (Estado).

		Estado		OR
		IRC	No ERC	
Sexo	**Masculino**	58	246	$(58 \times 378)/(18 \times 246) = 4.95$
	Femenino	18	378	
Grupo de edades	**30 a 39**	5	129	$(5 \times 266)/(1 \times 129) = 10.31$
	18 a 29 (categoría de referencia o de menor riesgo de sufrir IRC)	1	266	
	40 a 49	13	109	$(13 \times 266)/(1 \times 109) = 31.72$

	18 a 29 (categoría de referencia o de menor riesgo de sufrir IRC)	1	266	
	50 a 59	22	71	$(22 \times 266)/(1 \times 71) = 82.42$
	18 a 29 (categoría de referencia o de menor riesgo de sufrir IRC)	1	266	
	60 a 69	14	31	$(14 \times 266)/(1 \times 31) = 120.13$
	18 a 29 (categoría de referencia o de menor riesgo de sufrir IRC)	1	266	
	Mayor o igual a 70	21	18	$(21 \times 266)/(1 \times 18) = 310.33$
	18 a 29 (categoría de referencia o de menor riesgo de sufrir IRC)	1	266	
Ocupación agrícola	**Agricultor(a)**	53	213	$(53 \times 411)/(23 \times 213) = 4.45$
	No Agricultor(a)	23	411	
Contacto con agroquímicos	**Con contacto**	61	349	$(61 \times 274)/(15 \times 349) = 3.19$
	Sin contacto	15	274	
Ocurrencia de Dislipidemia	**Con dislipidemia**	56	383	$(56 \times 241)/(20 \times 383) = 1.76$
	Sin dislipidemia	20	241	
Ocurrencia de diabetes mellitus	**Con diabetes**	16	51	$(16 \times 573)/(60 \times 51) = 3.00$
	Sin diabetes	60	573	
Ocurrencia de hipertensión arterial	**Con hipertensión**	15	22	$(15 \times 601)/(61 \times 22) = 6.72$
	Sin hipertensión	61	601	

Luego en la Tabla 6.14 se muestra una matriz que verifica en alguna medida, el grado de multicolinealidad que pudiera existir entre las posibles variables explicativas, con la finalidad de observar si existe algún nivel de correlación multivariante que podría causar inconvenientes respecto a las estimaciones que conciernen en la elaboración del modelo.

Tabla 6.14. Matriz de correlaciones entre las posibles variables explicativas, para la evaluación de multicolinealidad.

	Sexo	**Grupo de edades**	**Ocupación agrícola**	**Contacto con agroquímicos**	**Ocurrencia de Dislipidemia**	**Ocurrencia de diabetes mellitus**	**Ocurrencia de hipertensión arterial**
Sexo	1.0	-0.077	0.762	0.553	0.022	-0.012	-0.001
Grupo de edades	-0.077	1.0	-0.104	-0.037	-0.208	-0.292	-0.269
Ocupación agrícola	0.762	-0.104	1.0	0.528	0.020	-0.008	-0.001

	Sexo	Grupo de edades	Ocupación agrícola	Contacto con agroquímicos	Ocurrencia de Dislipidemia	Ocurrencia de diabetes mellitus	Ocurrencia de hipertensión arterial
Contacto con agroquímicos	0.553	-0.037	0.528	1.0	0.018	-0.033	-0.010
Ocurrencia de Dislipidemia	0.022	-0.208	0.020	0.018	1.0	0.127	0.105
Ocurrencia de diabetes mellitus	-0.012	-0.292	-0.008	-0.033	0.127	1.0	0.143
Ocurrencia de hipertensión arterial	-0.001	-0.269	-0.001	-0.010	0.105	0.143	1.0

Por tanto en la Tabla 6.14, se observa evidentemente, que solamente existen correlaciones moderadas entre las variables “Sexo”, “Contacto con agroquímicos” y “Ocupación agrícola”. Sin embargo, se incluyen en la elaboración del modelo dado el sentido clínico y epidemiológico que tienen con la IRC. Por consiguiente, se muestra a continuación efectivamente el listado de las posibles variables explicativas que intervienen en la construcción del modelo de regresión logística binaria múltiple:

1) Sexo.
2) Grupo de edades.
3) Ocupación agrícola.
4) Contacto con agroquímicos
5) Ocurrencia de dislipidemia.
6) Ocurrencia de diabetes mellitus.
7) Ocurrencia de hipertensión arterial.

6.4.1. RESULTADOS DEL PROCEDIMIENTO DE CONSTRUCCIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA MÚLTIPLE

Para la construcción del modelo de regresión logística se utilizó el **método por pasos hacia atrás (Razón de verosimilitud)**; implementado en el programa estadístico SPSS [13] [37]. En este caso, el método parte de un modelo con las 7 posibles variables explicativas verificadas en la sección anterior que luego va eliminando del modelo, una a una, cada variable que no sea significativa, hasta que todas las variables incluidas

[13] SPSS es un software comercial que se utiliza con licencia.

sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste. La evaluación de estimación del modelo se contrasta a un nivel de significancia de 5%. En el procedimiento de construcción del modelo se introduce un **punto de corte para la clasificación de 0.12**, el cual permite determinar de mejor manera la tabla de clasificación de los casos pronosticados y observados en la variable dependiente y (Estado) [37]. De manera que los resultados del bloque 1 del método por pasos hacia atrás que construye el modelo de regresión logística, se exponen en la siguiente sección.

6.4.1.1. RESULTADOS DEL BLOQUE 1: MÉTODO POR PASOS HACIA ATRÁS (RAZÓN DE VEROSIMILITUD):

Tabla 6.15. Resumen del procesamiento de los casos que intervienen en la construcción del modelo de regresión logística binaria múltiple, según el método por pasos hacia atrás.

	Número de datos	Porcentaje
Casos Incluidos en el análisis	699	99.9
Casos excluidos del análisis (casos ausentes, por tener algún valor faltante)	1	0.1
Total	700	100.0
Casos no seleccionados	0	0.0
Total general	700	100.0

Tabla 6.16. Resultados de variables en el modelo de regresión logística binaria múltiple, según el método por pasos hacia atrás.

Número de paso de la iteración	Código de variable	Coeficiente β	Error estándar	Estadístico Wald	Grados de libertad	Valor de significancia	OR	Intervalo de confianza de 95% para OR	
								Inferior	Superior
Paso 4[a]	**Sexo**	1.188	0.488	5.919	1	0.015	3.282	1.260	8.549
	Grupo_edades			53.591	5	0.000			
	Grupo_edades1=30 a 39 años	2.317	1.105	4.396	1	0.036	10.142	1.163	88.446
	Grupo_edades2=40 a 49 años	3.491	1.050	11.054	1	0.001	32.819	4.191	256.983
	Grupo_edades3=50 a 59 años	4.272	1.041	16.850	1	0.000	71.669	9.321	551.051
	Grupo_edades4=60 a 69 años	4.613	1.065	18.778	1	0.000	100.810	12.512	812.218
	Grupo_edades5=mayor o igual a 70 años	5.533	1.067	26.883	1	0.000	252.779	31.223	2046.478
	Agricultor	0.637	0.470	1.834	1	0.176	1.891	0.752	4.754
	Ocurre_hipertensión	1.163	0.438	7.065	1	0.008	3.200	1.357	7.544
	Constante	-6.651	1.035	41.304	1	0.000	0.001		
Paso 5[a]	**Sexo**	1.693	0.318	28.349	1	0.000	5.438	2.916	10.143
	Grupo_edades			53.735	5	0.000			
	Grupo_edades1=30 a 39 años	2.317	1.105	4.399	1	0.036	10.149	1.164	88.483
	Grupo_edades2=40 a 49 años	3.534	1.050	11.341	1	0.001	34.274	4.382	268.109
	Grupo_edades3=50 a 59 años	4.342	1.040	17.442	1	0.000	76.875	10.018	589.903
	Grupo_edades4=60 a 69 años	4.635	1.064	18.991	1	0.000	103.077	12.816	829.043
	Grupo_edades5=mayor o igual a 70 años	5.521	1.066	26.828	1	0.000	249.781	30.925	2017.458
	Ocurre_hipertensión	1.135	0.432	6.882	1	0.009	3.110	1.332	7.259
	Constante	-6.638	1.035	41.174	1	0.000	0.001		

a. Variables introducidas en el paso 1: Sexo, Grupo_edades, Agricultor, Contacto_agroquimicos, Ocurre_dislipidemia, Ocurre_DM, Ocurre_hipertensión.

La Tabla 6.16 muestra los dos últimos pasos o iteraciones, es decir, el cuarto y quinto paso, que corresponden al procedimiento de construcción de un modelo de regresión logística realizado mediante el método por pasos hacia atrás (criterio: razón de verosimilitud) en el que se incluyen las 7 posibles variables independientes verificadas en la sección anterior. Se observa que en el cuarto paso está incluida la variable "Ocupación agrícola", la cual resulta estar no asociada significativamente con la variable respuesta, estadísticamente hablando, sin embargo llama la atención que su correspondiente OR es igual a 1.891, casi próximo a 2, por lo que alternativamente bajo ciertas circunstancias podría ser tentativo plantear un modelo a partir de los resultados que proporciona el cuarto paso. No obstante se tiene que dicho procedimiento termina en el quinto paso, proporcionando un modelo más adecuado en términos estadísticos con 3 variables explicativas, entre las que se encuentra la variable politómica "Grupo de edades", la cual el programa transformó en 5 variables dicotómicas, luego de haberle indicado la instrucción de que tomara como variable de referencia o de menor riesgo de padecer IRC el grupo de menor edad, o sea, el de 18 a 29 años y como variable más expuesta o de mayor riesgo de sufrir IRC, el grupo de mayor edad, es decir, mayor o igual a 70 años. Consecuentemente se realiza a continuación las evaluaciones y pruebas respectivas del modelo proporcionado por el quinto paso.

En este sentido se observa en el quinto paso que los errores estándar de los coeficientes son pequeños, lo cual es adecuado para el buen ajuste de un modelo. En base a un nivel de significancia de 5% se contrastan los estadísticos de Wald, los cuales resultan significativos, permitiendo así a los OR que miden la magnitud de las asociaciones bivariadas entre las variables explicativas y la variable respuesta y (Estado), aportar información que sea válida. Por consiguiente, en la Tabla 6.17 se evalúa la bondad de ajuste del modelo de regresión logística binaria, propuesto por el resultado del quinto paso de la Tabla 6.16.

Evaluación de bondad de ajuste del modelo obtenido por el método por pasos hacia atrás (Razón de verosimilitud):

Tabla 6.17. Prueba omnibus sobre los coeficientes del modelo obtenido, según el método por pasos hacia atrás.

		Chi-cuadrado	Grados de libertad	Valor de significancia
Paso 1	**Paso**	172.465	11	0.000
	Bloque	172.465	11	0.000
	Modelo	172.465	11	0.000
Paso 2[a]	**Paso**	-0.476	1	0.490
	Bloque	171.989	10	0.000
	Modelo	171.989	10	0.000
Paso 3[a]	**Paso**	-0.774	1	0.379
	Bloque	171.215	9	0.000
	Modelo	171.215	9	0.000
Paso 4[a]	**Paso**	-0.817	1	0.366
	Bloque	170.398	8	0.000
	Modelo	170.398	8	0.000
Paso 5[a]	**Paso**	-1.876	1	0.171
	Bloque	168.522	7	0.000
	Modelo	168.522	7	0.000

a. Un valor de chi cuadrado negativo indica que ha disminuido el valor de chi cuadrado con respecto al paso anterior.

Los resultados de la Tabla 6.17, expresan que el proceso iterativo del método por pasos hacia atrás termina en el quinto paso con un modelo que es significativamente mejor que el modelo con todas las variables [38]. Esto se puede notar observando cómo han ido disminuyendo los grados de libertad del contraste chi-cuadrado de la prueba omnibus sobre el modelo, desde 11 en el primer paso (se tienen 12 coeficientes: el de la constante, los de 7 variables iniciales y los de 5 categorías como resultado de la transformación de la variable politómica “Grupo de edades”, en 5 variables dicotómicas) hasta 7 en el quinto paso (es decir el método se quedó con 7 variables dicotómicas más la constante en la ecuación final de regresión).

De manera que la Tabla 6.17 aporta información sobre el ajuste del modelo, por lo que el interés se centra en la fila **Modelo** del quinto paso, el cual se refiere al contraste de razón de verosimilitudes en base al estadístico chi-cuadrado a un nivel de significancia de 5%.

En el contraste se evalúa la hipótesis nula de que los coeficientes β de todas las variables (excepto la constante) tomadas en cuenta por el modelo, son cero. En consecuencia, la significación estadística de 0.000 indica que el modelo obtenido en el quinto paso del método hacia atrás, mejora el ajuste de forma significativa con respecto a lo que se tenía en los pasos anteriores.

Tabla 6.18. Resumen del modelo obtenido, según el método por pasos hacia atrás.

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	308.231[a]	0.219	0.440
2	308.706[a]	0.218	0.439
3	309.480[a]	0.217	0.437
4	310.297[a]	0.216	0.435
5	312.173[a]	0.214	0.431

a. La estimación ha finalizado en el número de iteración 8 porque las estimaciones de los parámetros han cambiado en menos de 0.001.

La Tabla 6.18 muestra información sobre los ajustes de los diferentes modelos que ha ido generando en cada paso el método por pasos hacia atrás, a través de tres medidas de resumen del modelo complementarias a la tabla de la prueba omnibus, para evaluar de forma global su validez, las cuales son: el -2log de la verosimilitud o desviación del modelo a los datos y los dos coeficientes de determinación (R^2), parecidos al que se obtiene en regresión lineal, que expresan la proporción de la variación explicada por el modelo [37, 38].

Por lo tanto, el resultado de interés es el proporcionado en el quinto paso. De manera que el R cuadrado de Cox y Snell es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables independientes [37]. En este caso el R cuadrado de Cox y Snell es un valor no tanto discreto (0.214) el cual indica que el 21.4% de la variación de la variable dependiente es explicada por las variables que quedaron en el modelo.

La R cuadrado de Nagelkerke es una versión corregida de la R cuadrado de Cox y Snell. La R cuadrado de Cox y Snell tiene un valor máximo inferior a 1, incluso para un modelo "perfecto". La R cuadrado de Nagelkerke corrige la escala del estadístico para cubrir el rango completo de 0 a 1 [37]. Por consiguiente, en este caso, se tiene un R cuadrado de Nagelkerke de 43.1%, lo cual es un porcentaje moderado.

Tabla 6.19. Prueba de Hosmer y Lemeshow del modelo obtenido, según el método por pasos hacia atrás.

Paso	Chi-cuadrado	Grados de libertad	Valor de significancia
1	3.629	8	0.889
2	3.815	7	0.801
3	5.122	7	0.645
4	1.609	7	0.978
5	1.947	6	0.925

La Tabla 6.19 muestra información de otra prueba conocida como Hosmer y Lemeshow, en base al estadístico chi-cuadrado, para evaluar de manera global la bondad de ajuste de un modelo de regresión logística obtenido en cada paso del método por pasos hacia atrás [37]. Por lo que el interés se centra en observar el estadístico chi-cuadrado brindado por el quinto paso, el cual para contrastarlo es en base a un nivel de significancia de 5%.

Una característica de esta prueba, es que se desea que no haya significación (¡lo contrario a lo que suele ser habitual!) [37]. En consecuencia, en este caso se observa en el quinto paso, que la prueba no es significativa, lo cual indica que no hay motivos para pensar que los resultados predichos sean diferentes de los observados (o que si hay diferencias pueden explicarse razonablemente por el azar o error del muestreo) y que el modelo puede considerarse aceptable.

Tabla 6.20. Clasificación de valores observados y pronosticados en la variable dependiente por el modelo obtenido, según método por pasos hacia atrás.

	Observado		Pronosticado		
			Estado		Porcentaje correcto
			No ERC	IRC	
Paso 1	**Estado**	**No ERC**	514	109	82.5
		IRC	12	64	84.2
	Porcentaje global				82.7
Paso 2	**Estado**	**No ERC**	510	113	81.9
		IRC	12	64	84.2
	Porcentaje global				82.1
Paso 3	**Estado**	**No ERC**	518	105	83.1
		IRC	13	63	82.9
	Porcentaje global				83.1
Paso 4	**Estado**	**No ERC**	517	106	83.0

		IRC	13	63	82.9
	Porcentaje global				83.0
	Estado	**No ERC**	515	108	82.7
Paso 5		**IRC**	14	62	81.6
	Porcentaje global				82.5

El valor de corte es: 0.12

La Tabla 6.20 muestra un resumen de las tablas de clasificación de $2x2$, que se han ido obteniendo en cada paso del método por pasos hacia atrás, para clasificar a los individuos de la muestra según la concordancia de los valores predichos o estimados por el modelo en cada paso frente a los valores realmente observados, en base a un punto de corte de clasificación de 0.12, lo cual es una forma de evaluar la ecuación de regresión y el modelo obtenido [38]. En cada paso la mejoría es escasa en el porcentaje de predicción.

De manera que el resultado de interés es el brindado por el quinto paso, en donde se puede apreciar como el modelo obtenido clasifica correctamente 62 de las 76 personas con IRC ($y = 1$), es decir, el 81.6% (62/76); por el contrario clasifica correctamente 515 de las 624 personas sin ERC ($y = 0$), o sea, el 82.7%. Y de forma global se puede decir que el modelo tiene una capacidad de clasificar correctamente al 82.5% ([62+515]/699) de los casos o individuos analizados, aunque hay que decir que clasifica un poco mejor a las personas sin ERC, que los que tienen IRC.

En consecuencia se puede decir que las variables que quedaron en el quinto paso del procedimiento de construcción de un modelo de regresión logística realizado mediante el método por pasos hacia atrás (Razón de verosimilitud), se pueden considerar como explicativas de la variable dependiente y (Estado).

En la siguiente sección se muestran los resultados de regresión logística, obtenidos a través del método introducir implementado en el SPSS, luego de incluir las variables explicativas determinadas por el modelo encontrado mediante el método por pasos hacia atrás (Razón de verosimilitud) realizado en esta sección, con el objeto de refinar los resultados.

6.4.1.2. RESULTADOS DEL BLOQUE 1: MÉTODO INTRODUCIR

Se observa que los resultados del modelo de regresión logística obtenidos por el método introducir (Véase desde la Tabla 6.21 hasta la Tabla 6.26) son iguales a los resultados brindados en el quinto paso del método por pasos hacia atrás (Véase desde la Tabla 6.15 hasta la Tabla 6.20).

Adicionalmente se muestra en esta sección un gráfico correspondiente a la curva ROC (Véase Figura 6.12), en donde el eje vertical es la susceptibilidad o sensibilidad del modelo, el cual se refiere al porcentaje de observaciones que posee el evento $y = 1$, y que han sido correctamente clasificadas. El eje horizontal es (1 - especificidad) 100%, en donde la especificidad es el porcentaje de observaciones del otro evento $y = 0$, que han sido correctamente clasificadas [38].

Luego, la Tabla 6.27 muestra el área bajo la curva ROC, que es el poder de discriminación o clasificación del modelo construido, el cual es contrastado bajo la hipótesis de no discriminación o clasificación del modelo, que en la gráfica corresponde a los puntos que caen sobre la diagonal [38]. Una ecuación sin poder de clasificación alguno tendría una especificidad, sensibilidad y un total de clasificaciones correctas igual a 50% (por el simple azar), por lo que un modelo puede considerarse aceptable si tanto la especificidad y la sensibilidad tienen un nivel alto, de al menos el 75% [37].

Como se ve en la tabla de clasificación, se dice, que la sensibilidad es igual a 81.6%, y la especificidad es 82.7%, y vemos en la Tabla 6.27 que el poder de clasificación del modelo construido es el área bajo la curva ROC, que posee un valor de 89.9% del máximo posible, el cual es estadísticamente significativo.

Tabla 6.21. Resumen del procesamiento de los casos que intervienen en la construcción del modelo de regresión logística binaria múltiple, según el método introducir.

	Número de datos	Porcentaje
Casos Incluidos en el análisis	699	99.9
Casos excluidos del análisis (casos ausentes, por tener algún valor faltante)	1	0.1
Total	700	100.0
Casos no seleccionados	0	0.0
Total general	700	100.0

Tabla 6.22. Resultados de variables explicativas en el modelo de regresión logística binaria múltiple, obtenidos por el método introducir.

Número de paso	Código de variable	Coeficiente β	Error estándar	Estadístico Wald	Grados de libertad	Valor de significancia	OR	Intervalo de confianza de 95% para OR	
								Inferior	Superior
Paso 1[a]	**Sexo**	1.693	0.318	28.349	1	0.000	5.438	2.916	10.143
	Grupo_edades			53.735	5	0.000			
	Grupo_edades1=30 a 39 años	2.317	1.105	4.399	1	0.036	10.149	1.164	88.483
	Grupo_edades2=40 a 49 años	3.534	1.050	11.341	1	0.001	34.274	4.382	268.109
	Grupo_edades3=50 a 59 años	4.342	1.040	17.442	1	0.000	76.875	10.018	589.903
	Grupo_edades4=60 a 69 años	4.635	1.064	18.991	1	0.000	103.077	12.816	829.043
	Grupo_edades5=mayor o igual a 70 años	5.521	1.066	26.828	1	0.000	249.781	30.925	2017.458
	Ocurre_hipertensión	1.135	0.432	6.882	1	0.009	3.110	1.332	7.259
	Constante	-6.638	1.035	41.174	1	0.000	0.001		

a. Variables introducidas: Sexo, Grupo_edades, Agricultor, Ocurre_hipertensión.

Tabla 6.23. Pruebas omnibus sobre los coeficientes del modelo, obtenidos por el método introducir.

		Chi-cuadrado	Grados de libertad	Valor de significancia
Paso 1	Paso	168.522	7	0.000
	Bloque	168.522	7	0.000
	Modelo	168.522	7	0.000

Tabla 6.24. Resumen del modelo obtenido por el método introducir.

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	312.173[a]	0.214	0.431

a.La estimación ha finalizado en el número de iteración 8 porque las estimaciones de los parámetros han cambiado en menos de 0.001

Tabla 6.25. Prueba de Hosmer y Lemeshow del modelo obtenido por el método introducir.

Paso	Chi-cuadrado	Grados de libertad	Valor de significancia
1	1.947	6	0.925

Tabla 6.26. Clasificación de valores observados y pronosticados en la variable dependiente por el modelo obtenido en el método introducir.

	Observado		Pronosticado		
			Estado		Porcentaje correcto
			No ERC	IRC	
Paso 1	Estado	No ERC	515	108	82.7
		IRC	14	62	81.6
	Porcentaje global				82.5

El valor de corte es: 0.12

Figura 6.12. Curva ROC resultante del modelo obtenido.

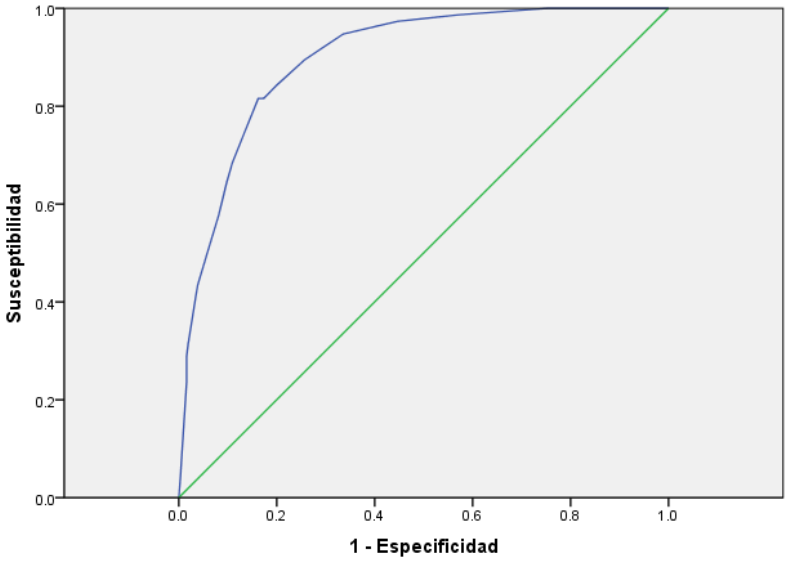


Tabla 6.27. Área bajo la curva ROC del modelo obtenido.

Área	Error típico	Significación asintótica [a]	Intervalo de confianza asintótico al 95%	
			Límite inferior	Límite superior
0.899	0.016	0.000	0.867	0.930

a. Hipótesis nula: área = 0.5

Consecuentemente, en base a los resultados anteriores proporcionados por los métodos que construyeron el modelo de regresión logística obtenido, se muestra en la Tabla 6.28 la lista de variables explicativas que representan aceptablemente factores de riesgo asociados con la IRC.

Tabla 6.28. Declaración de variables que representan factores de riesgo asociados con la IRC, según el modelo de regresión logística obtenido.

Nombre de variable	Nombre de variable	Código de variable
Sexo	Sexo	Sexo
Grupo de edades	Grupo de edades	Grupo_edades
Ocurrencia de hipertensión	Ocurrencia de hipertensión	Ocurre_hipertensión

6.4.2. PLANTEAMIENTO E INTERPRETACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA MÚLTIPLE

El planteamiento del modelo de regresión logística binaria múltiple, según resultados de la sección 6.4.1, es el siguiente:

$$Logit(P) = -6.638 + 1.693Sexo + 2.317Grupo\_edades1 + 3.534Grupo\_edades2 + 4.342Grupo\_edades3 + 4.635Grupo\_edades4 + 5.521Grupo\_edades5 + 1.135curre\_hipertensión + \varepsilon$$

Dónde P es la probabilidad de padecer IRC o no tener ERC. Luego, las variables explicativas incluidas en el modelo anterior, se pueden reconocer como variables que representan factores de riesgo asociados con la IRC, y ε representa a los residuos o errores del modelo. En la Tabla 6.29 se muestran medidas de influencia entre las variables explicativas y la variable dependiente y (Estado).

Tabla 6.29. Parámetros de influencia que declaran la existencia de variables que representan factores de riesgo asociados con la IRC, según el modelo de regresión logística obtenido.

Nombre de variable explicativa	Categorías de la variable explicativa	Coeficiente β de variable explicativa	Odds ratio (OR) obtenido por el modelo	Odds ratio (OR) obtenido en el análisis bivariado
Sexo	Masculino	1.693	5.438	4.95
	Femenino			
Grupo de edades	Grupo_edades1=30 a 39 años	2.317	10.149	10.31
	18 a 29 años (categoría de referencia o de menor riesgo de sufrir IRC)			
	Grupo_edades2=40 a 49 años	3.534	34.274	31.72
	18 a 29 años (categoría de referencia o de menor riesgo de sufrir IRC)			
	Grupo_edades3=50 a 59 años	4.342	76.875	82.42
	18 a 29 años (categoría de referencia o de menor riesgo de sufrir IRC)			
	Grupo_edades4=60 a 69 años	4.635	103.077	120.13
	18 a 29 años (categoría de referencia o de menor riesgo de sufrir IRC)			
	Grupo_edades5=mayores o iguales a 70 años	5.521	249.781	310.33
	18 a 29 años (categoría de referencia o de menor riesgo de sufrir IRC)			
Ocurrencia de hipertensión Arterial	Con hipertensión	1.135	3.110	6.72
	Sin hipertensión			

En la Tabla 6.29, al observar los OR obtenidos por medio del modelo de regresión logística binaria múltiple, se tiene lo siguiente:

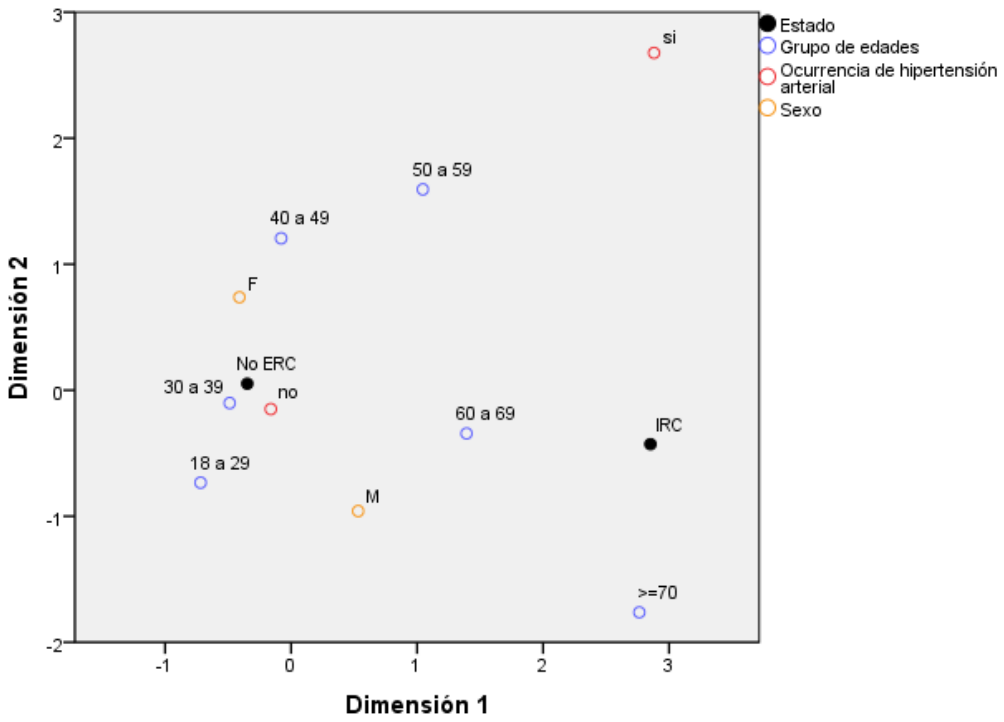
- El riesgo de que un hombre padezca IRC es aproximadamente 5 veces más de que la padezca una mujer.
- El riesgo de que la IRC la padezca una persona que esté entre las edades de 30 a 39 años, es 10 veces mayor de que la padezca alguien que se encuentre entre las edades de 18 a 29 años.
- El riesgo de que la IRC aparezca en una persona que esté entre las edades de 40 a 49 años, es 34 veces más de que suceda en alguien que se encuentre entre las edades de 18 a 29 años.
- El riesgo de aparición de la IRC en una persona que se halle entre las edades de 50 a 59 años, es aproximadamente 77 veces mayor de que aparezca en un individuo que esté entre las edades de 18 a 29 años.
- El riesgo de que la IRC se manifieste en una persona que esté entre las edades de 60 a 69 años, es 103 veces mayor de que aparezca en alguien que se localice entre las edades de 18 a 29 años.
- El riesgo de que la IRC la padezca una persona que tuviere una edad mayor o igual a 70 años, es aproximadamente 250 veces mayor de que la sufra un individuo que esté entre las edades de 18 a 29 años.
- El riesgo de que una persona con hipertensión arterial sufra IRC es 3 veces más de que la posea alguien que no tenga hipertensión arterial.

Finalmente a continuación se presenta un Análisis Correspondencias Múltiples (ACM), aplicado sobre la variable dependiente y (Estado) y las variables representan factores de riesgo asociadas con la IRC. Se provee este ACM con el objeto de observar de manera conjunta la forma en que se comportan las asociaciones entre las categorías de estas variables:

Tabla 6.30. Porcentajes de variabilidad de variables explicativas y la variable dependiente y (Estado: 1=IRC, 0=No ERC).

Dimensión	Varianza (Autovalor)	% de la varianza
1	1.707	42.678
2	1.110	27.744
Total	2.817	70.422
Media	1.408	35.211

Figura 6.13. Diagrama conjunto de puntos de categorías de variables explicativas y de la variable dependiente y (Estado: 1=IRC, 0=No ERC).



La Tabla 6.30 muestra que el porcentaje de variabilidad de categorías explicada por las dos dimensiones que construyen el diagrama de dispersión de categorías de la Figura 6.13 es igual a 70.4%, por lo cual es posible determinar asociaciones que sean aceptables, a través de la Figura 6.13.

En efecto, se destaca que de entre todas las categorías de variables explicativas las que están más asociadas con la IRC son:

- Grupo de edades mayores o iguales a 70 años (>= 70)
- Grupo de edades de 60 a 69 años
- Con hipertensión arterial

También, se aprecia que las categorías de variables explicativas que se asocian en alguna medida con la IRC, pero que se asocian un poco más al hecho de no padecer ERC son:

- Grupo de edades de 50 y 59 años
- Masculino

Asimismo se valora que las categorías de variables explicativas que están más asociadas o correlacionadas con la ausencia de ERC que con la IRC son:

- Grupo de edad de 18 a 29 años
- Grupo de edad de 30 a 39 años
- Sin hipertensión arterial
- Femenino

Además se logra estimar de manera visual que existe una discriminación o separación clara entre los sexos, es decir:

- El sexo femenino se asocia más con la ausencia de ERC que con la IRC.
- Las mujeres tienen a asociarse o identificarse más con la ausencia de ERC que los hombres.

7. CONCLUSIONES

1. A través de un abordaje descriptivo multivariante se encontró que las categorías de variables que más se asocian con la IRC son:
    - Grupo de edad de 60 a 69 años.
    - Grupo de edad de mayor o igual a 70 años.
    - Con hipertensión arterial.
2. En base a un enfoque descriptivo multivariante se halló que las personas con diabetes mellitus tienden a asociarse más con la IRC, que las personas que no tienen diabetes mellitus. Así mismo, las personas que no padecen diabetes, tienden a identificarse más con el hecho de no padecer ERC que con la IRC.
3. En el análisis descriptivo multivariante se puede ver que el contacto con productos agroquímicos y la ocupación agrícola, se asocian o se identifican más con la ausencia de ERC que con la IRC.
4. Las variables que representan factores de riesgo asociados con la IRC, según el modelo de regresión logística binaria múltiple encontrado son:
    - Sexo.
    - Grupo de edades.
    - Ocurrencia de hipertensión arterial.
5. En base a los resultados de análisis multivariantes se tiene que los hombres tienden a asociarse o identificarse más con la IRC que las mujeres, la cuales a su vez tienden a identificarse más con la ausencia de ERC que con la IRC.
6. Apoyado en los resultados multivariantes se puede decir que las personas con edades mayores o iguales a 50 años, tienden a correr mucho más riesgo de padecer IRC, que las personas con edades menores de 30 años.
7. Según los resultados de los análisis multivariantes se puede mencionar que las personas con hipertensión arterial tienden a asociarse o identificarse más con el hecho de padecer IRC. O en otras palabras el padecer hipertensión arterial puede representar un factor de riesgo asociado con la IRC.

8. REFERENCIAS BIBLIOGRÁFICAS.

1. Ministerio de Salud de El Salvador. OPS. Asociación de Nefrología e Hipertensión Arterial de El Salvador. Recomendaciones del Primer Taller de Salud Renal al Ministerio de Salud y Asistencia Social de El Salvador. San Salvador. 2010. (http://nefrologíaelsalvador.com/wp-content/uploads/2013/05/DECLARACION-TALLER-SALUDRENAL-160310.doc (acceso 21/10/2013).

2. Orantes, C. M., Herrera, R., Almaguer, M., Brizuela, E. G., Hernández, C. E., Bayarre, H., et al. Chronic kidney disease and associated risk factors in the Bajo Lempa region of El Salvador: Nefrolempa study, 2009. MEDICC review. 2011; 13(4):2.

3. National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Am J Kidney Dis. 2002 Feb;39 (2 Suppl 1):S1-266.

4. Meguid El Nahas A, Bello AK. Chronic kidney disease: the global change. Lancet. 2005 Jan 22-28; 365(9456):331–40.

5. United States Renal Data System (USRDS). 2004 annual report. Am J Kidney Dis. 2005; 45 (Suppl. 1).

6. Ansell D, Feest T, editors. UK renal registry report 2004. Bristol: UK Renal Registry; 2004.

7. Barsoum RS. Chronic Kidney Disease in the Developing World. New Engl J Med. 2006 Mar 9; 354(10):997–9.

8. Trabanino RG, Aguilar R, Silva CR, Mercado MO, Merino RL. Nefropatía terminal en pacientes de un hospital de referencia en El Salvador. Rev Panam Salud Pública. 2002 Sep;12(3):202–6.Spanish.

9. Flores Reyna R, Jenkins Molieri JJ, Vega Manzano R, Chicas Labor A, Leiva Merino R, Calderón GR, et al. Enfermedad renal terminal: Hallazgos preliminares de un reciente estudio en el Salvador. San Salvador: Pan American Health Organization; 2003.Spanish.

10. García-Trabanino R, Domínguez J, Jansá JM, Oliver A. Proteinuria e insuficiencia renal crónica en la costa de El Salvador: detección con métodos de bajo costo y factores asociados. Nefrología. 2005;25 (1):31–8.Spanish.

11. Cuadra SN, Jakobsson K, Hogstedt C, Wesseling C. Enfermedad Renal Crónica: Evaluación del conocimiento actual y la viabilidad para la colaboración de su investigación a nivel regional en América Central. Heredia (CR): SALTRA: IRE-UNA; 2006. 76 p. Spanish.

12. Hernández W. Nacimiento y Desarrollo del Río Lempa [Internet]. San Salvador: Servicio Nacional de Estudios Territoriales (SV); 2005 May [cited2011 Apr 21]. 14 p. Available from: http://www.snet.gob.sv/Geologia/Nacimiento EvolucionRLempa.pdf. Spanish.

13. Arnaiz Quintana A. Tierras pagadas a precio de sangre. Testimonios y retratos del Bajo Lempa Usuluteco. 2nd ed. San Salvador: Editorial Catalunya; 2008.Spanish.

14. Subsecretaria de innovación y calidad, Dirección General de Evaluación del Desempeño, Facultad de Medicina de la Universidad Nacional Autónoma de México. Estudio de Insuficiencia Renal Crónica y Atención Mediante Tratamiento de Sustitución. Ciudad de México: 2008. Disponible en: http://dged-salud.blogspot.com/2009/04/resultados-de-la-evaluacion-del-estudio.html.

15. Instituto Nacional de Salud (INS). [Página principal en internet]. El Salvador: Dr. Carlos Manuel Orantes Navarro; c2015. P.e.: [aprox. 2 pantallas]. Disponible en:http://www.ins.salud.gob.sv/index.php/temas/investigacion/investigacionensaludrenal.

16. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL Jr, et al. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 Report. JAMA. 2003 May14; 289(19):2560–71.

17. ADA. American Diabetes Association. Clinical Practice Recommendations 2004.Diabetes Care.2004; 27:S5–10.

18. Hohenberger EF, Kimling H. Compendio de Urianálisis con tiras reactivas. Mannheim (DE): Roche; 2008.Spanish.

19. Alfonzo JP. Definiciones de sobrepeso y obesidad. In: Alfonzo JP, editor. Obesidad. Epidemia del siglo XXI. Havana: Editorial Científico-Técnica; 2008. p. 175–92.Spanish.

20. Expert Panel on the Detection, Evaluation and Treatment of High Blood Cholesterol in Adults. Executive summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). JAMA. 2001 May16; 285(19):2486–97.

21. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. Ann Intern Med. 1999 Mar16; 130(6):461–70.

22. KDOQI. Clinical practice guidelines and clinical practice recommendations for diabetes and chronic kidney disease. Am J Kidney Dis. 2007 Feb; 49(2 Suppl 2):S12–154.

23. Cuadra, Steven N. et al. Enfermedad Renal Crónica: Evaluación del conocimiento actual y la viabilidad para la colaboración de su investigación a nivel regional en América Central. Heredia, Costa Rica: SALTRA, IRET-UNA, 2006. 76 p.

24. Sergio Arce Bustabad y cols. “Trasplante renal y enfermedad renal crónica. Sistema de leyes integradoras”. La Habana: Editorial Ciencias Médicas, 2009. 323 p.

25. Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. Nephron 1976; 16:31-41.

26. Hogan, Michelle KDIGO conference proposes changes to CKD Classification, but not to the definition. Nephrology Times. December 2009: 2, 12, pp 9-12.

27. Levey, S. Andrew et al. Definición y clasificación de la enfermedad renal crónica: Propuesta KDIGO (Kidney Disease: Improving Global Outcomes) Kidney International (Edición en español (2005) 1, 135–146.

28. ORANTES, C. M. Enfermedad renal crónica y factores de riesgo en el Bajo Lempa, El Salvador: Estudio Nefrolempa. *San Salvador: Ministry of Health (SV)*; 2010.

29. Criterios diagnósticos del Síndrome Metabólico. National Cholesterol Education Program – Adult Treatment Panel III (JAMA 2001; 285:2486-97)

30. Hernández, Bernardo y Velasco-Mondragón, Héctor. Encuestas Transversales. Salud pública de México/volumen 42, No. 5, septiembre-octubre de 2000

31. Peña, D. *Análisis de datos multivariantes*. Vol. 24. Madrid: McGraw-Hill; 2002.

32. Rojo JL, Gómez BF, Abascal HF, De la Mora JF, Sans JA. [Página principal en internet]. Valladolid: Universidad de Valladolid; c2008 [citado 25 octubre 2015]. P.e.: [aprox. 2 pantallas]. Disponible en:http://www5.uva.es/estadmed/datos/univariante/univar.htm.

33. Rojo JL, Gómez BF, Abascal HF, De la Mora JF, Sans JA. [Página principal en internet]. Valladolid: Universidad de Valladolid; c2008 [citado 25 octubre 2015]. P.e.: [aprox. 3 pantallas]. Disponible en: http://www5.uva.es/estadmed/datos/bivariante/bivar.htm.

34. Greenacre, M. J. *La práctica del análisis de correspondencias*. Fundación BBVA; 2008.

35. Salvador Figueras, M. Introducción al Análisis Multivariante, [en línea] 5campus.com, Estadística. España: Universidad de Zaragoza; 2000. Disponible en: http://www.5campus.com/leccion/anamul.

36. Albert J. Jovel. Análisis de Regresión Logística, Cuadernos Metodológicos. Centro de Investigaciones sociológicas. Disponible en: http://libreria.cis.es/libros/analisis-de-regresion-logistica/9788474762167/

37. Aguayo, M. (2007). Como hacer una regresión logística con SPSS, paso a paso. *Documentos Fabis-org.*

38. Aguayo, M., & Lore, E. (2007). Cómo hacer una Regresión Logística binaria paso a paso II análisis multivariante. *Fundación Andalucia Beturia para la Investigación en Salud, Dot*, (0702013), 1-35.