UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Ramón Álvarez-Vaz

Estadística multivariante y análisis espaciotemporal

Uso en vigilancia epidemiológica en salud bucal

bibliotecaplural

bibliotecaplural

Ramón Álvarez Vaz

Estadística multivariante y análisis espaciotemporal

Uso en vigilancia epidemiológica en salud bucal

COMISIÓN SECTORIAL DE INVESTIGACIÓN CIENTÍFICA
CSIC
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

La publicación de este libro fue realizada con el apoyo de la Comisión Sectorial de Investigación Científica (CSIC) de la Universidad de la República.
Los libros publicados en la presente colección han sido evaluados por académicos de reconocida trayectoria en las temáticas respectivas.
La Subcomisión de Apoyo a Publicaciones de la CSIC, integrada por Luis Bértola, Magdalena Coll, Mónica Lladó, Alejandra López Gómez, Vania Markarian, Sergio Martínez y Aníbal Parodi ha sido la encargada de recomendar los evaluadores para la convocatoria 2020.

Apoyo en la producción editorial del equipo de Ediciones Universitarias:

Ediciones Universitarias,
Unidad de Comunicación de la Universidad de la República (UCUR)
J. E. Rodó 1866
Montevideo, cp 11200, Uruguay
Tels.: (+598) 2409 7720 - (+598) 2408 5714
Correo electrónico: ucur@udelar.edu.uy
https://udelar.edu.uy/portal/institucional/comunicacion/ediciones-universitarias

isbn: 978-9974-0-2234-8
e-isbn: 978-9974-0-2239-3

Dedico este trabajo a mi viejo y a mi hijo Julián, a quienes extraño y hoy, me hubiese gustado poder compartirlo. A mi madre, por haberme inculcado desde niño el gusto por el estudio.

También a mis compañeras y compañeros de estudio de la carrera de Técnico en Estadística por permitirnos mostrar una vez más que se podía llegar, si se nos daba la oportunidad.

Por último y lo más importante: el tiempo dedicado a este trabajo es tiempo que dejé de compartir con mi esposa y compañera de la vida, Luisa, y mi hijo Bruno y mi hija Lucía, que son mi orgullo.

Tabla de contenidos

Índice de figuras

Índice de tablas

Lista de abreviaturas

Fuentes de datos

PRNSB2011 Primer Relevamiento Nacional de Salud Bucal en población joven y adulta

RACA2012 Relevamiento y análisis de caries dental en adolescentes de 12 años de la ciudad de Montevideo,Uruguay

RPAFO2015 Relevamiento en población que se asiste Facultad de Odontología durante 2015-2016

DPBIO2009 Determinación del perfil biológico de pacientes asistidos en la clínica de ortodoncia del Instituto Universitario Centro de Estudio y Diagnóstico de las Disgnacias del Uruguay IUCEDDU

Organismos

INE Instituto Nacional de Estadística

ASSE Administración de Servicios de Salud del Estado (ASSE)

MSP Ministerio de Salud Pública

OMS Organización Mundial de la Salud

Técnicas estadísticas

SNA Análisis de redes sociales (Social Network Analysis)

ENT Enfermedades no transmisibles

RLog Regresión logística

RLin Regresión lineal

RB Regresión Beta

RPoi Regresión Poisson

MLG Modelos lineales generalizados

TRI Teoría de respuesta al ítem

MCont Modelos de conteo

PIG Poisson inversa gaussiana

IG Inversa gaussiana

PG Poisson generalizada

DP Doble Poisson

MPoi Modelo de Poisson

MHPoi Modelo hurdle Poisson

MhBN Modelo hurdle binomial negativo

MBN Modelo binomial negativo

NB-I Binomial negativa tipo I

NB-II Binomial negativa tipo II

MBN.p Modelo binomial Negativo-P

POI Distribución de Poisson

MH Modelos Hurdle o con obstáculos

MEC Modelos con exceso de Ceros

BN Distribución binomial Negativa

AF Análisis factorial

AC Análisis de cluster o conglomerados

ACP Análisis de componentes principales

ACM Análisis factorial de correspondencias múltiples

CART Classification and Regression Trees

AD Análisis discriminante probabilístico

MRL Modelos de Regresión lineal

MDC Modelos de datos categóricos

ADC Análisis de Datos composicionales

GIS Sistema de información geográfica

SM Simulación Monte Carlo

REM Razón estandarizada de mortalidad

AIC Akaike Information Criteria

BIC Bayesian information criterion

Multiniv Análisis multinivel

CCI Coeficiente de correlación intraclase

Vario Variogramas

SI Muestreo aleatorio Simple sin reposición de tamaño n

SY Muestreo sistemático

STSI Muestreo estratificado

SIC Muestreo por conglomerados

MMult Muestreo en varias etapas

UPM Unidades primarias de muestreo

USM Unidades secundarias de muestreo

TK Tabla de Kish

DEff Efecto 'diseño'

PEC PostEstratificación completa o calibración sobre totales de celdas conocidos

Rake PostEstratificación incompleta o calibración sobre las marginales conocidas

TNR Tasa de no respuesta

ECT Encuestas de corte transversal (cross-section)

EL Encuestas longitudinales (panel data)

IM Índice de Moran

IGea Índice de Geary

Métricas

Se Sensibilidad

Esp Especificidad

ROC Receiving operating Curve

VP+ Valor predictivo positivo

VP- Valor predictivo negativo

AUC Área bajo la curva

Kappa Índice de concordancia Kappa

EE Error estándar

BC Centralidad de intermediación

CC Centralidad de cercanía

EBC Centralidad de intermediación de enlaces

Índices e indicadores odontológicos

ceo Índice CPO en niñas y niños

CPO Índice CPO

IG Índice gingival

CPITN Índice periodontal Comunitario de necesidad de tratamiento

CPI Índice periodontal comunitario

Odont Odontograma

ICDAS International Caries Detection and Assessment System

CIE-OE Clasificación Internacional de Enfermedades Aplicada a la Odontología y Estomatología

DAI Indice de estética dental

TTM Trastornos temporomandibulares

Acrónimos epidemiológicos

CIE10 Décima versión de la Clasificación Internacional de Enfermedades

VE Vigilancia epidemiológica

ET Enfermedades transmisibles

NT Necesidades de tratamiento

APCA Age-period-cohort Analysis

Acrónimos Software

FSF Free Software Foundation,

GNU GNU No es Unix

Encuestas Estándares

ECH Encuesta Continua de Hogares

STPES Encuesta STEPS de la OMS

Agradecimientos

Cuando corresponde agradecer, el espacio nunca es suficiente y uno tiene miedo de que la memoria traicione y queden personas por considerar. Este es un trabajo colectivo en el que participaron muchas personas —colegas, docentes, compañeros de trabajo—que fui conociendo a lo largo de varios años.

Del IESTA, que es mi casa, un especial agradecimiento a los jóvenes integrantes Eugenia Riaño, Elena Vernazza y Fernando Massa, hijos postizos académicos. En especial a Fernando, con quien desarrollé una parte importante de los hallazgos que presento en este libro y de quien me siento orgulloso; me quedo tranquilo en ver en él un relevo para cuando ya me dedique a otras cosas. También a Jorge Blanco, que fue mi maestro y hoy no está para que pueda agradecerle en persona.

A Silvia Rodríguez, más próxima generacionalmente a mí, un reconocimiento a su especial pedido de que yo no dejara y tratara de divulgar los resultados cuando las fuerzas flaqueaban por estar haciendo todo esto ya de veterano.

También agradecer a Cecilia Acuña, bibliotecóloga del IESTA y amiga, por su eterno apoyo en la búsqueda de artículos y libros, así como su agudo sentido para ayudarnos a los investigadores a visibilizar mejor nuestro trabajo con herramientas que ayudan en la labor cotidiana.

De los integrantes del Servicio de Epidemiología y Estadística de la Facultad de Odontología, donde tengo la suerte de participar, un especial reconocimiento a Susana Lorenzo y Anunzziatta Fabruccini por su paciencia en acostumbrarse a convivir y discutir con alguien como yo, que provengo del mundo de los números, y aprender a dialogar. También un agradecimiento por su apoyo cuando tuve que consultarlos a Graciela Buño, Alicia Picapedra y demás docentes de la Facultad de Odontología.

También a mis compañeros del Área de Epidemiología y Estadística de la Comisión Honoraria de Salud Cardiovascular, Virginia Estragó y Matías Muñoz —otro hijo postizo—, por las largas charlas semanales en donde reflexionamos y tratamos de pensar los problemas de la salud en su globalidad, mezclando clínica, epidemiología y estadística, sin perder el rigor y, sobre todo, el buen humor.

A otros investigadores de muy diferentes áreas de la salud, de quienes me enriquecí mucho como investigador: Paula Moliterno, Estela Skapino, Juan Dapueto, Fiorella Cavalleri, José Boggia, Alejandra López, entre otros.

Agradezco a mis tutores, los profesores Dr. Enrique Barrios y PhD. Gabriel Camaño, por su invalorable apoyo

No puedo cerrar esta sección sin agradecer al Dr. Hugo Dibarboure Rossini, porque principio tienen las cosas y fue gracias a él, compañero de aventuras en este asunto, que primero decidí terminar la maestría y luego imitarlo e ir siempre por más, tratando de hacer el doctorado, que me permite hoy generar esta obra. Gracias a él es que puedo

sentirme contento de lo que produje y pongo a consideración de lectoras y lectores, esperando que sea de utilidad.

Montevideo, Mayo de 2024

Presentación de la colección Biblioteca Plural

Vivimos en una sociedad atravesada por tensiones y conflictos, en un mundo que se encuentra en constante cambio. Pronunciadas desigualdades ponen en duda la noción de progreso, mientras la riqueza se concentra cada vez más en menos manos y la catástrofe climática se desenvuelve cada día frente a nuestros ojos. Pero también nuevas generaciones cuestionan las formas instituidas, se abren nuevos campos de conocimiento y la ciencia y la cultura se enfrentan a sus propios dilemas.

La pluralidad de abordajes, visiones y respuestas constituye una virtud para potenciar la creación y uso socialmente valioso del conocimiento. Es por ello que hace más de una década surge la colección Biblioteca Plural.

Año tras año investigadores e investigadoras de nuestra casa de estudios trabajan en cada área de conocimiento. Para hacerlo utilizan su creatividad, disciplina y capacidad de innovación, algunos de los elementos sustantivos para las transformaciones más profundas. La difusión de los resultados de esas actividades es también parte del mandato de una institución como la nuestra: democratizar el conocimiento.

Las universidades públicas latinoamericanas tenemos una gran responsabilidad en este sentido, en tanto de nuestras instituciones emana la mayor parte del conocimiento que se produce en la región. El caso de la Universidad de la República es emblemático: aquí se genera el ochenta por ciento de la producción nacional de conocimiento científico. Esta tarea, realizada con un profundo compromiso con la sociedad de la que se es parte, es uno de los valores fundamentales de la universidad latinoamericana.

Esta colección busca condensar el trabajo riguroso de nuestros investigadores e investigadoras. Un trabajo sostenido por el esfuerzo continuo de la sociedad uruguaya, enmarcado en las funciones que ella encarga a la Universidad de la República a través de su Ley Orgánica.

De eso se trata Biblioteca Plural: investigación de calidad, generada en la universidad pública, encomendada por la ciudadanía y puesta a su disposición.

Rodrigo Arim
Rector de la Universidad de la República

Presentación

En el ámbito de la salud pública es necesario conocer en profundidad las características de las poblaciones y de los problemas de salud. Para eso se recurre a las fuentes de datos existentes. Entre ellas se destacan las estadísticas vitales; los registros de problemas específicos de salud —por ejemplo, los registros de cáncer, que son registros de base poblacional—, que permiten, entre otras cosas, establecer la incidencia de la enfermedad, y los registros de enfermedades de etiología infecciosa con notificación obligatoria, en los que se basan los sistemas de vigilancia epidemiológica. Cuando la información que el investigador en biomedicina necesita no está disponible, se debe recurrir a mecanismos de generación a través de los diferentes diseños de estudios sanitarios, como los mecanismos de muestreo y las encuestas.

Sin embargo, pueden existir limitaciones en los indicadores que suelen utilizarse en epidemiología y salud pública, ya que muchas veces no toman en cuenta la estructura multivariada de la información o, si la toman, lo hacen a través de algoritmos de cálculo que generan indicadores univariados para ganar en simplicidad y no miden correctamente, por lo tanto, los fenómenos en estudio.

Teniendo en cuenta los antecedentes planteados, se reformulará cómo se considera la información que ya se viene recogiendo, para los cuales existen varios índices recomendados de la Organización Mundial de la Salud Bucal y otros índices epidemiológicos sobre estado de la salud bucal. Para ello, se usarán diferentes técnicas estadísticas, algunas de uso frecuente y que sirven para resolver el problema de preservar la estructura multivariada de la información y, a su vez, técnicas más nuevas que provienen de otras disciplinas. Se sistematizará un conjunto de indicadores epidemiológicos alternativos a los clásicos usando la misma información habitualmente, pero considerándola de forma diferente, transformando los algoritmos de cálculos. Esto permitirá hacer predicciones en función de las características epidemiológicas de las personas con el uso de métodos estadísticos adecuados y presentar indicadores que den cuenta de la distribución espaciotemporal de las patologías bucales en estudio.

Este libro está basado en la tesis de doctorado del Programa para la Investigación Biológica Sistematización y creación de indicadores e índices para la vigilancia epidemiológica en salud bucal: uso de técnicas estadísticas multivariantes y de análisis espaciotemporal, (Álvarez-Vaz, 2020). Se compone de tres partes. En la parte I, el capítulo 1 presenta los problemas y necesidades de esta sistematización de nuevos indicadores para medir adecuadamente las patologías en salud bucal, que se presentan en el capítulo 2; el capítulo 3 presenta los aspectos metodológicos y estadísticos que aparecen como solución a los problemas. La parte II desarrolla, en 5 capítulos, varias aplicaciones en las que se emplean varios de los indicadores presentados en la parte I. Por último la parte III consta de un capítulo de conclusiones y recomendaciones a futuro.

Parte I

Aspectos metodológicos

Capítulo 1

Introducción

En el ámbito de la salud pública, es necesario conocer en profundidad los problemas de salud y las características de las poblaciones en las que se pretende intervenir para mejorar sus indicadores, para lo cual se requieren diagnósticos de situación como punto de partida antes de todo plan estratégico y conjunto de acciones. Tal como plantea, por ejemplo, Ramis (Oriol, 1997), existen diferentes fuentes de datos para generar indicadores. Algunas de ellas son las estadísticas vitales; los registros de problemas específicos de salud, tales como los registros de cáncer (que son registros de base poblacional), que permiten, entre otras cosas, establecer la incidencia de la enfermedad, y los registros de enfermedades de etiología infecciosa con notificación obligatoria, en los que se basan los sistemas de vigilancia epidemiológica. Cuando la información que el especialista en biomedicina necesita no está disponible a través de alguna de las fuentes antes mencionadas, se debe recurrir a mecanismos de generación en los que se toman en cuenta la forma de selección de los individuos y el manejo del tiempo en la evaluación de los resultados. Por ello, se recurre a valiosas herramientas epidemiológicas que utilizan el método de investigación, como son los estudios clínicos con diseño de casos y controles, los de cohorte, los estudios experimentales —ensayos clínicos—, y las encuestas de base poblacional mediante muestreo probabilístico complejo —encuestas de corte transversal (cross-section)—y encuestas longitudinales (panel data) (Lilienfeld y Lilenfield, 1980), (Särndal et al., 1992), (Clayton y Hills, 1993), (Martínez et al., 1997), (Rothman y Greenland, 1998), (Silva, 2000), (Vittinghoff et al., 2005). Toda la información recolectada se puede sistematizar y clasificar en forma protocolizada a través de la Clasificación Internacional de Enfermedades versión 10 (CIE10).[1]

1.1. Sistematización de la información y elaboración de indicadores

Esta sistematización de las diferentes fuentes de información permitiría, en rigor, construir un sistema de información (Oriol, 1997). Este es uno de los pilares necesarios para la verdadera Vigilancia Epidemiológica (VE), la que permitirá luego poder hacer intervenciones en salud, para poder modificar la situación.

1. https://icd.who.int/browse10/2019/en

Por otra parte, la VE, pensada como un monitoreo permanente de la situación sanitaria, debe estar acompañada de una serie de indicadores epidemiológicos construídos con la información sistematizada que permiten evaluar si son necesarias diferentes intervenciones y, a su vez, priorizar entre las diferentes alternativas posibles. Para eso, la epidemiología se nutre de las herramientas de la demografía y la administración apoyándose, a su vez, en la estadística para construir una gran diversidad de indicadores que deben cumplir ciertas características, tales como la validez de aspecto, la validez de contenido, la validez de criterio o concurrencia, la capacidad predictiva y fiabilidad o reproducibilidad (Silva, 1998).

Todos estos conceptos son válidos para la vigilancia en salud pública en general y en particular en la salud oral, donde existen indicadores recomendados por la Organización Mundial de la Salud (OMS) (Organización Mundial de la Salud, 1997). Se usan en encuestas de base poblacional, por ejemplo, en Brasil (Ministério da Saúde, 2003).

Hasta la fecha, los antecedentes de estudios en Uruguay consideran la aplicación de algunos de esos índices (Beca et al., 1996), (Bianco et al., 1997). Sin embargo, estos indicadores son inadecuados para ser usados en la toma de decisiones que generen acciones concretas con el propósito de mejorar la salud oral de la población. Esto se debe a que no considera toda la información necesaria que debe usarse en la epidemiología más moderna, que toma en cuenta la distribución de los fenómenos en el tiempo y en el espacio, lo que implica la necesidad de georreferenciar la información y de construir indicadores que deben ser integrados al proceso de vigilancia. Por otra parte, los indicadores epidemiológicos clásicamente usados en salud oral no toman en cuenta la estructura multivariada de la información epidemiológica relevada. Usar técnicas estadísticas multivariantes recientes puede ayudar a tener perfiles epidemiológicos más completos, fundamentales para mejorar la planificación en salud. Esta característica de uso de indicadores limitados (al no tomar en cuenta la estructura, multivariada de la información o algoritmos de cálculos que generan indicadores univariados para ganar en simplicidad) se da en otros dominios de la salud pública, no solo en la salud oral, por lo menos en nuestro país.

Otro aspecto que se debe considerar en la creación de los sistemas de vigilancias es la forma en que los indicadores clásicos o los nuevos que se propongan se pueden combinar, ya que provienen de diferentes fuentes de información. En Uruguay se pueden destacar dos tipos de fuentes en salud oral, que tienen niveles de existencia en el tiempo y desarrollo o profundidad muy heterogéneos. Son los registros de atención en el primer nivel de atención en el ámbito de atención del sector público y privado y las encuestas de base poblacional, que son muy pocas las que se han desarrollado con alcance nacional. Estas dos grandes fuentes de información hacen que, en realidad, la población sobre la que se pretende hacer vigilancia en salud oral pueda ser dividida en tres grupos de personas: la población de las personas que consultan, la población de las personas que tienen cobertura de algún tipo y no consultan y la población en general.

Teniendo en cuenta las diferentes fuentes de información y las poblaciones que la VE considera, surge un nuevo problema: cómo construir y combinar indicadores que trabajen con información que se genera mediante registros, muestras probabilísticas con diseño muestral complejo, que puede ser de tipo longitudinal o de corte transversal, y que deban trabajar sobre marcos muestrales con multiplicidad, con diferentes probabilidades de inclusión y con sesgos de selección (Särndal et al., 1992),(Nithila et al., 1998),(Thomas y Weber, 2001), (Oakes y Kaufman, 2006), (Kim y Dailey, 2008), (Ministerio de Salud Pública, 2009), (Álvarez-Vaz, 2010), (Chattopadhyay, 2011).

Todos los aspectos manejados hasta el momento muestran un vacío muy importante en el manejo de la información para la vigilancia epidemiológica, que no solamente se da para ENT de la que forman parte las patologías orales. Las Enfermedades Transmisibles (ET), como los virus respiratorios, la hepatitis A, el VIH, el dengue. Estas deben ser monitoreadas con sistemas de vigilancia compuestos por las mismas herramientas metodológicas y estadísticas, adaptadas necesariamente al caso de las ET, donde el concepto espcaiotemporal de la información epidemiológica es clave, en virtud de la dinámica propia de esas enfermedades.

En resumen, este diagnóstico en cuanto a la necesidad de un manejo adecuado de la información en salud justifica esta línea de trabajo en la que se plantean los siguientes objetivos.

1.2. Objetivo general

Teniendo en cuenta los antecedentes antes planteados con respecto a las fuentes de información en salud en general, y en salud bucal en particular, se propone sistematizar y construir un conjunto de indicadores alternativos y complementarios a los que ya existen en las estudios sanitarios en salud oral, reformulando la forma de considerar la información que ya se viene recogiendo, usando las técnicas estadísticas pertinentes.

1.3. Objetivos específicos

1. Se pretende elaborar un conjunto de indicadores epidemiológicos que combine los clásicos al considerar la estructurada multivariada de la información independientemente de la fuente de datos, usando técnicas estadísticas de uso poco habitual en investigación epidemiológica en nuestro país (ver sección 3.3).

2. Otro objetivo es desarrollar un conjunto de indicadores epidemiológicos alternativos a los clásicos usando la misma información que suele utilizarse, pero considerándola de forma diferente, transformando los algoritmos de cálculos, permitiendo hacer predicciones en función de características epidemiológicas de las personas con el uso de métodos estadísticos adecuados (ver sección 3.4).

3. También se presentarán indicadores que den cuenta de la distribución espaciotemporal de las patologías orales en estudio (ver sección 4).

4. Por último, se identificarán los problemas que surgen del análisis epidemiológico al no considerar el proceso de generación de la información (ver sección 3.6 y 3.7).

1.4. Estructura de la obra

En este libro se presenta y se referencia una serie de publicaciones que, a juicio del autor, son importantes para la obra y para formar parte de ella, ya que muestran varias aplicaciones de los desarrollos presentados en las diferentes secciones del capítulo 3.

Algunas de estas producciones se han elaborado en primer término como trabajos presentados en jornadas académicas o congresos que luego derivaron como documentos de trabajo o en artículos en revistas y otras que directamente se hicieron bajo el formato de artículos en revistas desde su inicio. La elección del formato con el cual se presentan los diferentes trabajos responde a la necesidad de poder mostrarlos todos y en la secuencia temporal en que fueron creados o publicados, si corresponde, en coautoría, no como autor principal, en el marco del trabajo de investigación. Por otra parte, algunos temas presentados como problemas a estudiar en el capítulo 3, se desarrollan en capítulos específicos desde el 5 al 8, siendo ésta producción en primer autoría, donde el énfasis está puesto, sobre todo, en aplicaciones donde lo que más importa y resulta novedoso es el uso de técnicas estadísticas muy poco usadas en el ámbito de la biomedicina. Por este motivo, los capítulos se caracterizan por tener un desarrollo exhaustivo de las técnicas usadas.

Como consecuencia, los temas que se desarrollan en los capítulos de aplicaciones que se consignan y que son una parte de lo desarrollado en la tesis Sistematización y creación de indicadores e índices para la vigilancia epidemiológica en salud bucal: uso de técnicas estadísticas multivariantes y de análisis espaciotemporal, (Álvarez-Vaz, 2020) son, para el autor de la obra, donde se logra el mayor aporte al conocimiento de otros investigadores biomédicos. Se espera que puedan tomarlos como nuevas herramientas de solución a sus problemas. Se entiende que esta es la forma adecuada de presentar la línea de investigación desarrollada en la tesis, a través de de la producción escrita publicada en artículos y complementada con cinco de los ocho capítulos desarrollados en las aplicaciones de la tesis, que muestran la respuesta a los objetivos planteados en las secciones 1.2 y 1.3 del capítulo 1. A su vez, las aplicaciones y sus resultados se retoman en el capítulo 9, donde se sintetiza el trabajo.

Capítulo 2

Antecedentes

En lo que respecta a la salud bucal, se deben evaluar muchas dimensiones de manera individual. Algunos ejemplos son el estado de las piezas dentales (odontograma), el estado de las mucosas o tejidos (examen local) y los síntomas asociados con los aspectos funcionales (articulares, oclusión) y hábitos (higiene, dieta). Teniendo en cuenta las dimensiones, es necesario ver las patologías que aparecen, y la forma de medirlas da lugar a varios indicadores que también se presentan a lo largo del capítulo.

2.1. Patologías en salud oral

Los problemas de salud en la mayor parte del mundo han cambiado. Las ENT son las de mayor prevalencia, en las que el nuevo patrón global está estrechamente relacionado con los estilos de vida de las sociedades modernas (Breilh, 2010). Este cambio también ha tenido impacto en la salud bucal, ya que las enfermedades bucodentales son uno de principales problemas en la salud pública por su alta prevalencia e incidencia en todas regiones del mundo (Peterson, 2004).

Antes de presentar las patologías bucales más importantes y los índices para evaluarlas, es necesario consignar que, en odontología, existe una forma de referirse a las piezas dentales, que reciben una numeración coherente también con la medición de patologías de la mucosa. Las piezas se suelen agrupar en sextantes (inferiores y superiores) o cuadrantes (inferiores y superiores y, a su vez, izquierdos y derechos). La figura 2.1 es un odontograma en el que puede verse la disposición de las piezas en la boca. Las piezas numeradas del 55 al 75 corresponden a la temporarias.

2.1.1. Caries

El concepto actual de caries dental define a la enfermedad como un proceso dinámico localizado en la superficie dentaria cubierta por una biopelícula, caracterizado por desequilibrios en los procesos de desmineralización y remineralización que ocurren constantemente en la cavidad bucal. A lo largo de un determinado período de tiempo, la predominancia de momentos de pérdida mineral de los tejidos duros del diente (sobre todo iones calcio y fosfato) para la biopelícula y saliva resulta en la lesión de caries (Holst et al., 2001). Algunos componentes del proceso de las caries actúan en la superficie del diente (saliva, biopelícula, dieta, acceso al flúor), mientras que existe otro conjunto de factores que determinan el comportamiento de la persona (conocimiento, actitud, ingre-

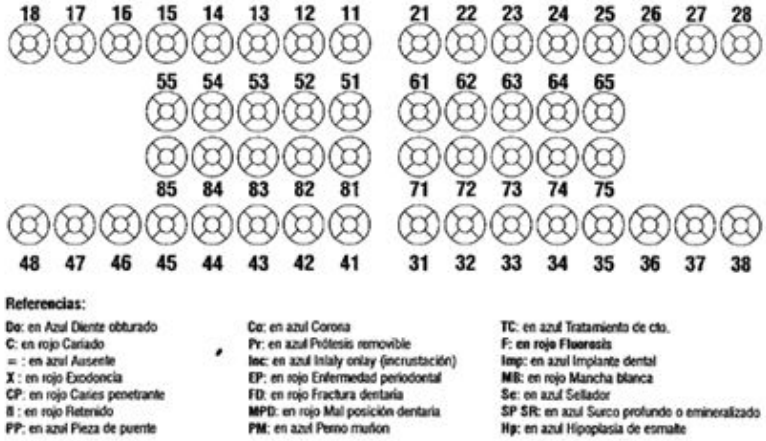

Figura 2.1: Odontograma

sos, nivel educativo y socioeconómico) (Fejerscov, 2004). El proceso de la enfermedad debe estar sujeto a un control permanente a lo largo de la vida para evitar consecuencias irreversibles en etapas posteriores (Maltz y Jardim, 2010).

2.1.2. Enfermedad periodontal

Las enfermedades periodontales son enfermedades infecciosas inflamatorias que, según el grado de afectación de los tejidos que rodean y soportan al diente, se dividen en gingivitis y periodontitis (Wolf et al., 2005). La gingivitis es un proceso inflamatorio de la encía sin pérdida de inserción. El epitelio de unión se mantiene unido al diente en el nivel original, sin migración apical ni pérdida de soporte periodontal. Esto se produce por la acumulación inespecífica de especies bacterianas (biopelícula) y se elimina mediante un buen control de esta (Cortés, 1999). En caso de depresión del estado inmunitario, presencia de factores de riesgo y mediadores proinflamatorios, así como el aumento de bacterias periodontopatógenas, es posible que a partir de una gingivitis se desarrolle una periodontitis. Esta ocurre cuando la inflamación de la encía pierde inserción, hay migración apical del epitelio de unión y pérdida del soporte periodontal (desintegración del colágeno y destrucción ósea) (Wolf et al., 2005). Mediante un consenso entre expertos, se han propuesto criterios para determinar casos de periodontitis para usar en diversos estudios de investigación.

El 5.º Workshop Europeo del 2005 propone dos niveles de criterios para la definición de un caso de periodontitis:

1. la presencia de pérdida de inserción proximal $\geq 3mm$ en dos o más piezas no adyacentes permite la definición de casos incipientes;

2. la presencia de pérdida de inserción proximal $> 5mm$ en por lo menos el 30 % de

las piezas posteriores permite la definición de casos más específicos (identifica casos de extensión y severidad).

Este criterio no está diseñado para evaluar la prevalencia de la periodontitis, está enfocado en identificar los factores de riesgo (Tonetti y Claffey, 2005). En el mismo sentido, la Academia Americana de Periodoncia en 2007 presentó un criterio de diagnóstico que planteaba los siguientes criterios:

1. ausencia de periodontitis moderada o severa;
2. periodontitis moderadas LAC (límite amelocementario hasta el borde de la encía) $\geq 4mm$ (no en la misma pieza) o mayor de dos sitios con profundidad al sondaje $\geq 5mm$;
3. periodontitis severa en dos o más sitios interproximales con LAC $\geq 6mm$ y más de un sitio interproximal con profundidad al sondaje $\geq 5mm$ (Page y Eke, 2007).

Contrario a lo que se creía en el pasado, no toda gingivitis evoluciona a periodontitis. El mecanismo por el cual la simple inflamación gingival progresa a la destrucción de los tejidos periodontales aún no está explicado en su totalidad, aunque el modelo etiológico multifactorial de las enfermedades crónicas humanas, puede explicar esta evolución. Así pues, la periodontitis es una enfermedad infecciosa en la que el agente microbiano es necesario pero no suficiente para causarla, lo que implica la necesidad de un huésped susceptible para su inicio y desarrollo, de tal manera que el estilo de vida, factores sistémicos y factores psicosociales desempeñan un papel importante en la etiopatogenia de la periodontitis (Manau y Echeverría, 1999).

2.1.3. Erosión

Existen varias formas de presentar esta patología, pero una que parece muy adecuada es la que surge como resultado del proceso de pérdida de tejido erosivo que se da en el desgaste fisiológico de los dientes. En ese proceso las características clínicas que aparecen son una pérdida inicial de brillo de los dientes, seguida de un aplanamiento de las estructuras convexas y, con la exposición continua al ácido, se forman concavidades en superficies lisas o surcos y ahuecamientos en las superficies incisales u oclusales (Ganss, 2006). A su vez, Ganss advierte que la erosión dental debe distinguirse de otras formas de desgaste, pero también puede contribuir a la pérdida general de tejido al suavizar la superficie, mejorando así los procesos de desgaste físico. La determinación de la erosión dental como condición o patología es relativamente fácil en caso de que haya dolor o complicaciones endodónticas, pero es ambigua en términos de función o estética. Este autor habla de que el impacto de la erosión dental en la salud bucal en la mayoría de los casos se describe mejor como una condición, ya que el ácido es de origen no patológico.

Para entender mejor la pérdida del tejido erosivo es importante tener en cuenta la variedad de procesos que se dan a lo largo de la duración de las piezas dentales, donde

aparece la fricción de material exógeno (por ejemplo, durante la masticación, el cepillado de dientes o las herramientas de sujeción), forzadas sobre sustancias dentales (abrasión), el efecto de los dientes antagónicos (desgaste), el impacto de la tracción y fuerzas compresivas durante la flexión del diente (abfracción) y la disolución química del mineral de las piezas (erosión).

Estos cuatro procesos pueden ampliarse de acuerdo a una terminología muy precisa, para la cual hay una definición y etiología de tipos físico y químico:

Abrasión: Desgaste físico como resultado de procesos mecánicos que involucran sustancias u objetos (tres desgaste del cuerpo). Los factores etiológicos son procedimientos de higiene bucal (por ejemplo, cepillado o uso de hilo dental excesivos, efecto de abrasivos en pastas dentales), hábitos (por ejemplo, objetos de sujeción) o exposición ocupacional a partículas abrasivas. La morfología resultante de los defectos puede ser difusa o localizada, dependiendo del impacto predominante. Los defectos en forma de cuña también se atribuyen a la abrasión; una forma especial de abrasión es la demasticación, producto de masticar comida. La pérdida de tejido se localiza en superficies incisales u oclusales y depende de la abrasividad de la dieta individual.

Desgaste: Desgaste físico como consecuencia de la acción de dientes antagónicos sin sustancias extrañas que intervengan (dos desgastes del cuerpo). Los rasgos característicos son facetas planas antagónicas con márgenes definidos.

Abfracción: Desgaste físico como resultado de la tensión en la región cemento-esmalte, lo que provoca microfracturas en esmalte y dentina (desgaste por fatiga). Los defectos en forma de cuña también se atribuyen a la abfracción.

Erosión: Desgaste químico como resultado de ácidos extrínsecos o intrínsecos o quelantes que actúa sobre superficies dentales sin placa, donde las características clínicas del desgaste (tribo) químico presentan pérdida de la estructura de la superficie, apariencia derretida; Ranurado en superficies oclusales-incisales y concavidades someras coronal de la unión cemento-esmalte.

Teniendo en cuenta los diferentes aspectos ya analizados, es importante ver la diferencia que puede darse para la erosión al ser considerada como condición o como patología, dependiendo de los conceptos de salud y enfermedad que se manejen. La erosión puede entenderse como patología cuando ocurre en combinación con dolor o complicaciones endodónticas agudas. Tambien es razonable considerar que la pérdida de tejido es una característica normal del envejecimiento. Sin embargo cuando el desgaste erosivo avanzado es asintomático, resulta difícil poder diferenciar entre un estado fisiológico y uno patológico, por lo cual el debate aún continúa. En términos generales, la erosión dental puede describirse como una condición provocada por ácidos de origen no patológico.

2.1.4. Maloclusión

Para hacer una muy breve introducción a esta patología se puede hablar de ella en función del criterio que parte de la relación anteroposterior entre los primeros molares superiores e inferiores. En ortodoncia se ha propuesto un gran número de clasificaciones, pero no se ha reemplazado al sistema de Angle, ya que este método es aceptado y conocido universalmente (Álvarez-Vaz et al., 2011). Se puede presentar el gradiente de la maloclusión con la siguiente clasificación:

1. Maloclusión de clase I o neutroclusión: la cúspide mesiovestibular del primer molar permanente superior ocluye en el surco mesiovestibular del primer molar permanente inferior (posición de máxima intercuspidación).

2. Maloclusión de clase II o distoclusión: la cúspide mesiovestibular del primer molar permanente superior ocluye por delante del surco mesiovestibular del primer molar permanente inferior.

3. Maloclusión de clase III o mesioclusión: la cúspide mesiovestibular del primer molar permanente superior ocluye por detrás del surco mesiovestibular del primer molar permanente inferior cuando los maxilares están en máxima intercuspidación.

Al hablar de las maloclusiones, es muy difícil establecer claramente su etiología, ya que estas son de origen multifactorial. Sin embargo, es posible definir dos componentes principales en la etiología de las maloclusiones, que son la predisposición genética (que se refiere a los genes que dictan la herencia de una maloclusión) y los los factores exógenos o ambientales (incluye todos los elementos capaces de condicionar una maloclusión durante el desarrollo craneofacial). De la interacción recíproca de estos factores dependerá la manifestación de una determinada maloclusión.

La oclusión no solo comprende la relación y la interdigitación de los dientes, sino también las relaciones de estos con los tejidos blandos y duros que los rodean. En el presente, la maloclusión se define como una disposición de los dientes que crea un problema para el individuo, bien sea estético, en relación con el mal alineamiento, protrusión con autoestima perjudicada por la maloclusión funcional debido a dificultades en el movimiento mandibular o cualquier combinación de estos (Graber y Swaim, 1991) (Proffit, 1990), (Vig, 1990). Los autores señalan la necesidad de que, para prevenir, antes se ha de conocer e identificar mejor la etiología de las maloclusiones. La concepción actual de la etiología de las maloclusiones es integralmente diferente al vigente a principios de siglo, cuando se creía que cada individuo nacía con pleno potencial para llegar a alcanzar una dentición completa y una oclusión correcta. Para el pensamiento de entonces, la maloclusión resultaba de la acción de fuerzas ambientales que desviaban el desarrollo, pero el potencial genético siempre apuntaba hacia el logro de una normoclusión, tal como fue descrita por Angle.

Hoy en día, y luego de casi cincuenta años de investigación en este área, se considera que, en la mayoría de los casos, las maloclusiones resultan de una de estas dos situaciones: una discrepancia relativa del tamaño de los dientes y de los huesos y una desarmonía en el desarrollo de las bases óseas maxilares. Hay igual predisposición a tener dientes grandes que a desarrollar una mandíbula progénica. La carga genética influye de una forma decisiva en la mayoría de las maloclusiones junto con una constelación de factores ambientales que matizan su expresión final en la oclusión. El reconocimiento de la etiología de las maloclusiones es la clave del plan de tratamiento ortodóncico, puesto que este debe ser etiológico y no sintomático. El diagnóstico ortodóncico debe tratar de identificar el agente causal, el protagonismo de la herencia y la multiplicidad de causas que intervienen en el mismo cuadro de maloclusión en distintos momentos del desarrollo y con diferente intensidad.

La observación clínica de los pacientes, de sus hermanos y progenitores conduce a la idea de que la herencia juega un papel importante en la estructura cráneofacial y dental de las maloclusiones. Durante muchos años, lo que explicaba el aumento de prevalencia de maloclusiones era la heredabilidad independiente de variables como, por ejemplo, heredar el tamaño de dientes de un progenitor y el tamaño de maxilares de otro progenitor. Esta idea, aunque todavía se maneja en ocasiones, no es compatible con el conocimiento actual de la herencia poligénica. De acuerdo con los hallazgos actuales en el campo sobre la etiología de maloclusiones, se determina que no son monogénicas, sino poligénicas. El gen que interviene en la expresión de la característica genética apenas contribuye a las malformaciones fenotípicas. Cuando se manifiesta el efecto de otros genes puede presentar poligenia aditiva. Esa es la razón de por qué las características o anomalías de herencia poligénica muestran un cuadro clínico menos nítido que la monogénica, que se traducen por un fenotipo relativamente uniforme.

La mayoría de las maloclusiones de clase II en las que suele existir un patrón heredado de déficit mandibular y en las de clase III se constata una clara tendencia familiar y racial. A su vez, los problemas de excesos verticales también tienen un importante componente hereditario. Sin embargo, estas maloclusiones esqueléticas heredadas pueden agravarse por la presencia de factores ambientales. La herencia también influye en el tamaño y forma dentaria, en el número de piezas e incluso en la cronología y patrón eruptivo. Sobre el factor de la herencia, sólo se puede actuar con la detección precoz y el consejo genético, aunque en un futuro próximo, y según los descubrimientos recientes del genoma humano, será posible influir directamente en la genética para prevenir las maloclusiones.

Si bien las maloclusiones tienen un importante componente genético, existen factores externos que pueden afectar la situación de equilibrio en la que se encuentran las estructuras dentales y esqueléticas. El efecto de una fuerza ambiental que rompa esta situación de equilibrio depende sobre todo de su duración, frecuencia e intensidad. Con

el paso del tiempo, parece evidenciarse este fenómeno cuando se compara la prevalencia de maloclusiones en la actualidad con la de poblaciones primitivas o contemporáneas sin un estilo de vida de sociedad urbana industrializada.

En estudios llevados a cabo por antropólogos, se observa una frecuencia baja de Maloclusión en grupos humanos primitivos alejados de la civilización. Los individuos poseen normooclusiones aceptables que se deterioran tan pronto como se cambian los hábitos dietéticos y se usan alimentos blandos y refinados; en una o dos generaciones se alcanza el grado de prevalencia de maloclusiones propio de las sociedades industrializadas. Este cambio es tan rápido que difícilmente puede ser atribuido al papel de la herencia, por lo que se sugiere que la reducción de la consistencia y dureza de los alimentos disminuiría el estímulo funcional de crecimiento y que la dieta blanda sería el factor más importante en la alta incidencia actual de la maloclusión. Tanto los estudios realizados sobre grupos humanos como en animales de experimentación soportan la evidente contribución del estímulo funcional de la masticación al normal desarrollo de los maxilares.

La falta de uso del aparato masticatorio en el hombre civilizado condiciona una atrofia que se manifiesta en maloclusiones de distinto signo, alta incidencia y variable intensidad. De ese modo, se aceleraría la tendencia evolutiva normal hacia la reducción del tamaño de los maxilares y se favorecerían situaciones como el incremento en la prevalencia del apiñamiento de las generaciones actuales y otros factores ambientales. Una de las causas ambientales de maloclusión más importante la constituyen los hábitos de larga duración que pueden alterar la función y equilibrio normal de dientes y maxilares. En el contexto de maloclusión en niños, se puede ver que los hábitos de presión interfieren en el crecimiento normal y en la función de la musculatura orofacial. Entre estos están la interposición lingual (deglución atípica); la succión digital, de las cuales la más común es la succión del pulgar cuando se lo sostene en posición vertical; el uso prolongado del chupete; la respiración oral, que puede aparecer como consecuencia de la reducción en el pasaje aéreo de la nariz o de la nasofaringe por circunstancias de tipo mecánico o alérgico.

El problema aparece cuando se prolonga en el tiempo. La aparición de una maloclusión debida a un hábito depende del número de horas (duración y frecuencia) en el que este actúe más que de la intensidad. Otros factores ambientales que influyen en la etiología de la maloclusión son la pérdida prematura de dientes, la caries dental, los traumatismos y las patologías tumorales y quísticas. Las maloclusiones son un problema real de salud pública al que se enfrenta la práctica odontológica diaria, por lo que debe es necesario evaluar la oclusión de forma adecuada y hacer predicciones acertadas respecto a su futuro.

2.1.5. Trastornos temporomandibulares

Los trastornos temporomandibulares (TTM) se pueden definir como un conjunto de condiciones dolorosas o disfuncionales en los músculos masticatorios o en la articulación temporomandibular (ATM). Entre estos se encuentran la parafunción oral de apretamiento dental durante el día y el bruxismo del sueño, (Willeman Bastos Tesch et al.,

2014). En general, afectan a las mujeres durante los años reproductivos y su prevalencia disminuye bruscamente con la edad. Además de la predisposición relacionada con el sexo desde el nacimiento y la edad, la exposición a determinados factores predispone a las personas a desarrollar TTM.

Esta patología merece estudio, ya que estos mismos autores consignan que los TTM ocupan el tercer lugar en prevalencia entre los dolores crónicos. En esta escala, están los dolores de cabeza primarios en primer lugar y el dolor de espalda en segundo lugar.

El dolor crónico debido a los TTM provoca un impacto significativo en la calidad de vida de quienes los padecen. Se verifican desde leves cambios en la alimentación hasta la manifestación de conducta depresiva profundamente discapacitante, todos aspectos se dan en un contexto de una interferencia importante en las actividades diarias. Por otra parte, la demanda de tratamiento para los TTM parece estar relacionada, sobre todo, presencia de dolor intenso y los distintos grados en los que este se puede manifestar. El alivio del dolor es el indicador más confiable para que los pacientes y los médicos juzguen el éxito del tratamiento.

Dado que los TTM son una patología que tiene un origen multifactorial, se presentan a continuación algunos de estos (Willeman Bastos Tesch et al., 2014):

Endocrinológicos. Los receptores de estrógeno se localizan en los tejidos de la ATM y la expresión de los TTM ha demostrado sufrir modulación por medio del proceso inflamatorio y de la concentración de estrógeno, lo que suministra un sustrato molecular para los efectos directos de estas hormonas en la articulación.

Otorrinoralingología. Las estructuras craneofaciales y periorales, directamente involucradas en los TTM, forman el sistema estomatognático, responsable de muchos procesos fisiológicos vitales como comer, respirar, tragar y la comunicación verbal y no verbal. Indirectamente, estas estructuras también pueden participar en la fisiopatología de los procesos que se le solían atribuir solo a la actividad neuronal de sus regiones anatómicas específicas.

Neurología. El dolor orofacial crónico es una respuesta del cuerpo, no solo un fenómeno local, ya que siempre implica el procesamiento simultáneo de diferentes tipos de informaciones para diversos niveles de integración. El dolor no es una consecuencia pasiva de la simple transferencia de estímulos periféricos para un centro de dolor en la corteza. El dolor es un proceso activo generado en parte en la periferia y en parte en el sistema nervioso central, donde varios cambios plásticos que involucran el aprendizaje y la memoria contribuyen a esta experiencia. La definición de dolor de la Asociación Internacional para el Estudio del Dolor (IASP)[1] sirve para reforzar esta conclusión: «El dolor es una experiencia sensorial y emocional desagradable, asociada con un daño tisular real o potencial, o descrita en términos de tal

1. https://www.iasp-pain.org/

daño». De este modo, la experiencia dolorosa es siempre subjetiva y es el resultado de la ocurrencia simultánea de diferentes factores, entre ellos la fisiopatología, las experiencias del pasado y el contexto social.

2.1.6. Lesiones de mucosa

Entre las lesiones de mucosa se encuentran, en primer término, el cáncer bucal, dada su malignidad, y, la candidosis, las úlceras crónicas, la gingivitis necrotizante aguda, los abscesos crónicos, la hiperplasia paraprotética y otras entidades como estomatitis, subplaca, fibromas, nevus, hemangiomas (Casnati et al., 2013).

La leucoplasia, utilizada por primera vez por Ernest Schwimmer a finales del siglo XIX. El término procede del griego leuco-, que significa 'blanco' y plakos-, que significa 'placa'. En 1978, la Organización Mundial de la Salud (OMS) pretendió consensuar la terminología utilizada hasta el momento y precisó su definición como una mancha blanca que no puede caracterizarse como otra entidad clínica ni patológica (Escribano-Bermejo y Bascones-Martínez, 2009). El liquen plano (LP) es una enfermedad inflamatoria crónica, de etiología desconocida (se reconoce una base autoinmune), mucocutánea con manifestaciones orales muy frecuentes, una clínica e histología características y de curso evolutivo benigno, aunque en ocasiones puede llegar a sufrir una degeneración maligna (Blanco Carrión et al., 2008).

Un aspecto que esta íntimamente ligado a las lesiones es su localización, de acuerdo a la siguiente topografía, en paladar, rebordes, labios, mucosa yugal, lengua, surcos, comisura.

2.2. Indicadores epidemiológicos usados en salud oral

Luego de presentadas las patologías mas importantes reseñadas hasta el momento, en las siguientes secciones se plantean los índices usados para medirlas.

2.2.1. Índice CPO

El Índice Cariados, Perdidos, Obturados (CPO) es el más utilizado para medir la enfermedad de caries dental. Fue descripto por Klein y Palmer en 1937 y adoptado por la OMS (Cortés, 1999). Este índice mide el presente y el pasado de la caries dental en un individuo o una población, puede aplicarse en la dentición permanente CPO o en la dentición decidua Índice CPO en niñas y niños medido por el (ceo), luego de las modificaciones introducidas por Gruebbell en 1944. En la sigla, C hace referencia al número de dientes afectados con lesiones de caries dentales no tratadas, P al número de dientes perdidos por causa de caries dental y O al número de dientes obturados o restaurados como consecuencia de caries dentales. El CPO es un índice agregado que se obtiene como resultado de la suma de estos valores. Cuando la unidad de observación es el diente, el índice se expresa como CPO-D o ceo-d, mientras que si la unidad observada es la superfi-

cie dentaria, el índice será CPO-S o ceo-s (Arana et al., 2005). En el caso de un individuo adulto, el índice CPO puede adoptar valores de 0 a 28 (excluyendo terceros molares), cuando la unidad es el diente, o valores de 0 a 128 cuando la unidad es la superficie dentaria (siendo 12 dientes anteriores con 4 y 16 dientes posteriores con 5 superficies). En cambio, en una población es el promedio del grupo (Arana et al., 2005), (Cortés, 1999), (Burt et al., 2008). Además, es un índice irreversible, ya que una vez producida la caries dental, está no revertirá, sino que se mantendrá en ese estadío o podrá ser obturada o extraída. En consecuencia, el índice solo puede incrementarse o permanecer estable. Por ello, solo puede variar la constitución de cada componente (cariado, perdido u obturado) en el valor total del CPO (Arana et al., 2005). El CPO cumple con una serie de características: es simple, versátil, estadísticamente manejable y fiable cuando los examinadores han sido entrenados. Sin embargo, también presenta limitaciones:

1. Sus valores no están relacionados con el número de dientes en riesgo, puesto que un valor de CPO es un recuento de aquellos dientes que el examinador juzgó como afectados por caries, no tiene denominador. El valor de CPO, por lo tanto, no muestra directamente la intensidad del daño en un solo individuo.

2. El CPO da el mismo valor a las piezas dentarias con lesiones de caries no tratadas, a las perdidas y a las obturadas. Esta filosofía es deficiente para muchos propósitos;

3. Los datos del CPO son de poca utilidad para la estimación de las necesidades de tratamiento. Cuando se cuentan los dientes perdidos en el componente P solo es válido contar aquellas piezas perdidas a causa de caries dental. Los dientes pueden perderse por razones periodontales en adultos mayores y por razones de ortodoncia en adolescentes. Se requieren, entonces, reglas de decisión para determinar cómo hacer frente a estos casos.

4. El índice CPO puede sobrestimar la experiencia de caries en dientes a través de restauraciones preventivas y sellantes de fosas y fisuras. La tendencia en un estudio epidemiológico es a contar estas restauraciones dentro del componente 'O' del CPO, a pesar de que estos dientes son sanos. El valor del CPO se verá inflado. Las restauraciones de resina compuesta que se han colocado por razones estéticas y los sellantes no deberían ser incluidas en el CPO, deberían ser contabilizadas por separado.

5. Los valores de CPO no pueden compararse de un grupo a otro sin tener en cuenta el criterio diagnóstico de caries. No existe un criterio universal de lo que es un diente con caries dental. La comparación entre grupos en los que la caries dental fue registrada como lesión cavitada con otro donde se registró desde lesiones no cavitadas es claramente inválida. El CPO como medida de incidencia y severidad de Caries dental es muy anticuado y, en realidad, puede ser más válido como medida de tratamiento recibido. Es cuestionable usar un índice para una enfermedad que

es tan dependiente de los juicios de tratamiento de muchos profesionales y de la combinación de un tratamiento previo (es decir, componente P y O) con necesidad de tratamiento actual (componente C). No se usa en otros lugares de la vigilancia en salud pública (Burt et al., 2008).

A pesar de estas limitaciones, el CPO, con algunas especificaciones, sigue siendo el índice más utilizado para expresar el estado de caries dental en una población.

2.2.2. Índice ICDAS

El International Caries Detection and Assessment System (ICDAS), traducido como Sistema Internacional para la Detección y Evaluación de Caries, es un sistema internacional de detección y diagnóstico de caries consensuado en Baltimore, Maryland, Estados Unidos, en el año 2005, para la práctica clínica, la investigación y el desarrollo de programas de salud pública. Su objetivo es desarrollar un método visual para la detección de caries en fase temprana y que sirva, además, para evaluar el gradiente de enfermedad y el nivel de actividad de la patología[2] (Gugnani et al., 2011).

Un estudio llevado a cabo por el Departamento de Cariología, Ciencias de la Restauración y Endodoncia de la Facultad de Odontología de la Universidad de Michigan en 2007, en el que se evaluaron las cualidades del índice, demostró que, adenás de ser sencillo, es práctico y tiene validez de contenido, discriminatoria y externa con una correlación con el examen histológico de las fosas y fisuras en dientes extraídos. Por eso los autores consideran que es un método especialmente útil para la detección temprana de caries de esmalte y que permite, a su vez, la planificación de terapia de remineralización individual. Este índice también permite hacer el seguimiento del patrón de caries de una determinada población, aspecto clave para la VE (Ismail et al., 2007). El sistema tiene entre el 70 % y el 85 % de sensibilidad y una especificidad del 80 % al 90 % para detectar caries en dentición temporaria y permanente. Esta variabilidad en el grado de acierto depende del nivel de entrenamiento y calibración del personal que hace el examen clínico, con un índice Kappa $\geq 0{,}65$ (Jablonski-Momeni et al., 2008), (Diniz et al., 2009).

Teniendo en cuenta las codificaciones de la (CIE-OE), la Orgamización Mundial de la Salud (OMS), basada en el criterio de diente cariado, perdido y obturado (CPO-D) y el sistema (ICDAS) completo (ICDAS EPI e ICDAS combinado), se puede comparar entre sí para ver su relación con el umbral visual.

El ICDAS completo presenta siete categorías. La primera es para dientes sanos (código 0, en color verde) y las dos siguientes son para caries limitadas al esmalte, que pueden ser mancha blanca o marrón (códigos 1 y 2, marcadas en color amarillo). Las categorías tercera y cuarta (código 3 y 4, en color rojo) se interpretan como caries que se extienden al esmalte y dentina, pero sin dentina expuesta. Por último, para mostrar más gravedad, las dos categorías restantes (códigos 5 y 6), aplican a caries con dentina expuesta. A su

2. https://www.sdpt.net/ICDAS.htm

vez, para las categorías 1 a 6, en cada estadío este puede ser estar inactiva la lesión o, al contrario, estar activo. Estos aspectos se reflejan por consignar qué es positivo, qué corresponde a activo, mientras que el valor negativo corresponde a inactivo.

2.2.3. Índice gingival (IG)

El índice gingival (IG), elaborado por Löe y Silness a comienzos de 1960, considera la inflamación de la encía en tres grados:

1. 0: encía sana,
2. 1: inflamación leve sin hemorragia al sondaje,
3. 2: inflamación moderada con hemorragia al sondaje,
4. 3: fuerte inflamación, tendencia a la hemorragia espontánea, eventual ulceración.

Se mide la salud gingival en cuatro zonas: vestibular, lingual, mesial y distal en seis dientes índices de Ramfjörd (16, 21, 24, 36, 41, 44). El IG se obtiene sumando todas las puntuaciones y dividiendo por el número de superficies observadas. El síntoma de hemorragia solo está presente a partir del grado 2 (Cortés, 1999), (Wolf et al., 2005).

2.2.4. Índice periodontal comunitario de necesidad de tratamiento

El Community Periodontal Index of Treatment Needs (CPITN) fue desarrollado por la OMS (Organization, 1987) y se usa, sobre todo, en estudios epidemiológicos. Además de determinar el grado de gravedad de la gingivitis (hemorragia) y la periodontitis (profundidad de bolsa), a partir de sus datos indica también el tipo y alcance del tratamiento necesario. No tiene en cuenta la pérdida de inserción de un diente, sino los parámetros que hay que tratar: inflamación gingival, hemorragia, sarro dental y profundidad de sondaje. Se determina dividiendo la boca por sextantes, limitados por los caninos, y utilizando una sonda CP11 (OMS) y se anota la afectación más grave del sextante. El algoritmo de aplicación es el siguiente:

1. Antes de examinar se identifican las marcas de la sonda CP11 (presenta una bolita de 0,5 mm en su extremo y una banda oscura situada entre 3,5 mm y 5,5 mm). Se debe usar con una presión ligera (20 gr) y esperar veinte segundos para observar el sangrado.
2. Se divide la boca en seis sextantes, uno anterior y dos posteriores en cada arco. Los sextantes se limitan en individuos mayores de veinte años utilizando todos los dientes presentes 17-14, 13-23, 24-27, 37-34, 33-43 y 44-47 o utilizando dientes índices 17-16, 11, 26-27, 47-46, 31 y 36-37. En menores de 20 años se eliminan los segundos molares. Para que un sextante pueda ser medido debe presentar dos piezas sin indicación de avulsión;

3. Se deben observar parámetros como sangrado gingival, cálculos gingivales y bolsas periodontales y registrar los siguientes códigos:

   - código 0: sano, ausencia de signos patológicos;
   - código 1: hemorragia al sondaje suave;
   - código 2: cálculo supra-subgingivales;
   - código 3: bolsa periodontal de $4 - 5$mm (banda negra parcialmente oculta);
   - código 4: bolsa ≥ 6 mm (banda negra oculta).

   El resultado de estas mediciones se convierte en Necesidades de tratamiento (NT), que se categorizan, en relación con los anteriores códigos, de la siguiente manera:

   - $NT = 0$: no necesita tratamiento (código 0);
   - $NT = 1$: necesita instrucción de higiene oral (código 1);
   - $NT = 2$: necesita la eliminación de cálculo u obturaciones desbordantes (códigos 2 y 3);
   - $NT = 3$: necesita tratamiento complejo (código 4).

   Cada categoría de necesidad de tratamiento incluye a la anterior, es decir, un individuo que necesite tratamiento complejo también necesitará que antes se elimine el sarro y se le instruya sobre higiene oral.

2.2.5. Índice periodontal comunitario

El Community Periodontal Index (CPI) es el índice actualmente recomendado por la OMS (Organization, 1987) y es una variante mejorada de CPITN. Se mantienen la división por sextantes y las instrucciones sobre los dientes que deben ser examinados y los códigos de 0 a 4 son iguales. En el caso de que falten el diente o los dientes índices en un sextante, todos los demás deben ser sondeados, pero se excluye el sondaje de caras distales de terceros molares. La sonda es similar a la del CPITN, con dos marcas añadidas en los 8,5 mm y 11,5 mm. Se agrega el registro la falta de inserción de los dientes afectados por la periodontitis, la cual es la crítica más importante al CPITN. La profundidad de bolsa proporciona cierta información sobre la cantidad de pérdida de inserción, pero esta medida no es fiable si existe recesión gingival o inflamación severa. La pérdida de inserción nos da información sobre la destrucción acumulada a lo largo de la historia de vida del individuo. Se registra inmediatamente después de realizar el CPI para ese sextante en particular, a partir de los siguientes códigos y criterios:

- código 0: pérdida de inserción es de 0 a 3 mm (CAL no visible y código CPI de 0-3);

  Si el grado de de CPI es de 4 o si la CAL es visible:

- código 1: Pérdida de inserción entre 4 a 5 mm (CAL dentro de la banda negra);
- código 2: Pérdida de inserción de 6 a 8 mm (CAL entre el límite superior de la banda negra y anillo de 8,5 mm);
- código 3: Pérdida de inserción de 9 a 11 mm (CAL entre anillos de 8.5 a 11.5 mm);
- código 4: Pérdida de inserción de 12 mm o más (CAL más allá del anillo de 11,5 mm).

La pérdida de inserción no se registra en los niños y adolescentes menores de quince años.

2.2.6. Índice de erosión

Desde un punto de vista clínico, el desgaste erosivo de los dientes es un fenómeno de la superficie que se produce en áreas accesibles al diagnóstico visual. El procedimiento de diagnóstico es, por lo tanto, un enfoque visual más que instrumental.

Se presenta una clasificación de los índices que (Ganss, 2006) a partir de lo que sugieren (Smith y Knight, 1984) para el desgaste dental en general, que incluyen los criterios de diagnóstico para el desgaste dental erosivo (Eccles, 1979).

Para el desgaste general se puede manejar el siguiente índice Basic Erosive Wear Examination (BEWE), donde se deben considerar las siguientes localizaciones:

- B: bucal o lingual;
- C: cervical;
- I: incisal;
- L: lingual o palatino;
- O: oclusal.
- 0 - B/L/O/I. Sin pérdida de características superficiales del esmalte, C sin pérdida de contorno.
- 1 - B/L/O/I. Pérdida de características superficiales del esmalte, C con mínima pérdida de contorno
- 2 - B/L/O. Pérdida de esmalte que expone la dentina a menos de un tercio de la superficie, I Pérdida de esmalte apenas exponiendo dentina, C con defecto de menos de 1 mm de profundidad.
- 3 - B/L/O. Pérdida de esmalte que expone la dentina a menos de un tercio de la superficie, I Pérdida de esmalte y pérdida sustancial de dentina, C con Defecto de menos de 1 a 2 mm de profundidad.

- 4 - B/L/O. Pérdida completa de esmalte (o exposición pulpar o de la dentina secundaria), C con defecto de más de 2 mm.

Del trabajo de Eccles surge una clasificación de erosión dental en base a la apariencia clínica que da cuenta del gradiente y localización de la misma:

- Clase I: Etapas tempranas de erosión, ausencia de crestas de desarrollo, superficie lisa y vidriada que se produce principalmente en las superficies labiales de incisivos maxilares y caninos;
- Clase II: la dentina facial de esta clase está involucrada por menos de un tercio de la superficie. Tipo 1: lesión cóncava ovoide o creciente en la región cervical de la superficie, que debe diferenciarse de las lesiones en forma de cuña. Tipo 2: lesión irregular completamente en la corona que tiene un aspecto perforado donde el esmalte está ausente del piso.
- Clase IIIa. Destrucción más extensa de la dentina, particularmente de los dientes anteriores. La mayoría de las lesiones afectan una gran parte de la superficie, pero algunas están localizadas y ahuecadas.
- Clase IIIb. Lesiones linguales o palatales de las superficies en más de un tercio de su área. Los bordes incisales se vuelven translúcidos debido a la pérdida de dentina, que parece suave y, en algunos casos, es plana o ahuecada. Los márgenes gingival y proximal tienen un aspecto blanco, grabado.
- Clase IIIc Incisal u oclusal. Los bordes incisales o las superficies oclusales están implicados en la dentina, el aplanamiento o el ahuecamiento, las restauraciones se ven elevadas por encima de la superficie del diente circundante, los bordes incisales parecen translúcidos debido al esmalte socavado
- Clase IIId. Todos los dientes están severamente afectados y las superficies labiales y linguales están muy involucradas.

2.2.7. Índice éstético dental

El Dental Aesthetic Index (DAI) como consigna (Ourens et al., 2013) es el índice usado para la evaluación de la maloclusión y se calcula con una ecuación en la que intervienen los diferentes componentes de este, con la siguiente forma:

$$6 * V_1 + 3 * (V_2 + V_3 + V_4) + V_5 + V_6 + V_7 * 2 + V_8 * 4 + V_9 * 4 + V_{10} * 4 + 13 \quad (2.1)$$

En la ecuación 2.1 puede verse que se tienen en cuenta diferentes dimensiones de las anomalías en cuanto a los dientes separados o apiñados, la presencia de dientes visibles perdidos y también las dimensiones con respecto a la mordida. Así se desarrolla a continuación:

- V_1: dientes visibles perdidos (DVP),
- V_2: apiñamiento (Api),
- V_3: separación (Sep),
- V_4: diastema (Dia),
- V_5: máxima Irregularidad maxilar anterior (MiMaxia),
- V_6: máxima Irregularidad mandibular anterior (MiMandia),
- V_7: superposición anterior del maxilar superior (Sums),
- V_8: superposición anterior de la mandíbula (Sam),
- V_9: mordida abierta anterior (Maa),
- V_10: relación molar anterior (Rma).

Cada uno de los indicadores que componen el DAI pueden ser evaluados por separado o en subdimensiones, como las anomalías de dentición, espacio y oclusión.

El siguiente índice se puede categorizar para evaluar el gradiente de la enfermedad a través de un nuevo índice que tiene cuatro niveles con los siguientes umbrales de corte:

1. oclusión normal o maloclusión leve (valores entre 13 y 25),
2. maloclusión definida (valores entre 26 y 30),
3. maloclusión severa (valores entre 31 y 35),
4. maloclusión muy severa (valores mayores de 35).

Los indicadores que componen el índice se pueden categorizar con la siguiente pauta:

- apiñamiento de incisivos (sin apiñamiento, con apiñamiento en 1 o 2 segmentos),
- espaciamiento en la región de incisivos (sin espaciamiento, con espaciamiento en 1 o 2 segmentos),
- diastema (sin diastema o diastema > 0 mm),
- irregularidad mandíbular (irregularidad de 0 a 1 mm, irregularidad ≥ 2 mm),
- irregularidad maxilar (irregularidad de 0 a 1 mm, irregularidad ≥ 2 mm),
- overjet (overjet ≤ 0 mm, overjet < 3 mm, overjet ≥ 3 mm),
- mordida abierta anterior (sin mordida abierta, mordida abierta > 1 mm).

2.2.8. Indicadores para TTM

Todos los aspectos mencionados en la sección 2.1.5 pueden medirse de forma muy variada, para lo cual se presenta, a modo de ejemplo, un instrumento utilizado en Uruguay en 2008 en un relevamiento poblacional con alcance nacional llevado adelante por docentes de la Facultad de Odontología de la Udelar (Riva et al., 2011).

Este instrumento contiene varios bloques de preguntas de tipo binarias (presencia o ausencia) que pueden considerarse como síntomas y signos:

- dificultad o dolor al abrir grande la boca;
- bloqueo de la mandíbula al abrir la boca, dificultades funcionales;
- ruidos articulares;
- dolor de oídos o alrededor de ellos;
- traumatismos;
- dolor de cabeza;
- relato de síntomas de bruxismo nocturno como: sensación de haber dormido apretando los dientesal levantarse, dolor de cabeza al levantarse, autovaloración del paciente de su grado de estrés, necesidad o consulta hechas por problemas articulares o síntomas vinculados al bruxismo.

A partir de la presencia de los síntomas y signos se puede construir indicadores de presencia de cada uno:

$$ISint_i = \sum_{j=1}^{j=4} Sint_{ij} \tag{2.2}$$

$$ISig_i = \sum_{j=1}^{j=5} Sig_{ij} \tag{2.3}$$

Con los dos indicadores definidos en (2.2) y en (2.3) se puede elaborar la prevalencia de síntomas y de signos.

2.2.9. Indicadores para lesiones de mucosa

Del mismo modo que para las TTM, en estos indicadores se usan medidas de prevalencia para cada una de las topografías y localizaciones de las lesiones de mucosas presentadas.

Capítulo 3

Metodología estadística

En este capítulo se presentan los dos pilares que deben tenerse en cuenta al sistematizar la información para el desarrollo de sistemas de vigilancia epidemiológica. Estos son, por una parte, los datos, sus fuentes y los mecanismos para generarlos y, por otra parte, la técnicas estadísticas empleadas para el análisis, lo que permite la creación de indicadores plausibles de ser resumidos y estudiados en el tiempo y espacio. Para eso se presentan fuentes de datos, algunas metodologías para el análisis de indicadores ya establecidas y usadas de rutina y también nuevas formas de uso de la información que involucran el uso metodologías estadísticas no convencionales.

Al investigador en biomedicina lector de este capítulo deben quedarle claros los métodos estadísticos que hoy aparecen en la mayoría de los artículos de revistas arbitradas. Debería conocer, al menos, su funcionamiento y el contexto en el que estos se usan.

A través de las diferentes secciones de este capítulo se presentan las técnicas estadísticas que los especialistas de la salud oral suelen usar. Estas se plantean con un nivel de desarrollo relativamente sencillo, desde el punto de vista estadístico, para que el investigador del área de la biomedicina y potencial lector de la obra las encuentre familiares, pero sí se exponen los supuestos en los que estas se basan y cómo se debe proceder para que su uso sea adecuado. Cuando sea necesario profundizar en las técnicas estadísticas, el lector podrá encontrar más detalle en los capítulos correspondientes a las aplicaciones, donde se tratan en forma mas rigurosa o se presentan por primera vez en cada caso. En las secciones del anexo de metodología estadística A, también se puede encontrar mayor nivel de detalle de algunas de las técnicas que se consideran más relevantes para que el lector pueda profundizar.

Esta forma de ir presentando las técnicas estadísticas vale tanto para los tres grupos de indicadores que se desarrollan en la sección 1.3 como para otras situaciones donde el investigador biomédico debe hacer uso de la estadística y también las que se vinculan con el mecanismo de generación de los datos (generalmente por muestreo) 3.6 y en el manejo de información auxiliar que aparecen en la sección 3.7.

3.1. Fuentes de datos

Tal como se presentó en la sección 3.1, existen diferentes fuentes de datos para medir la salud oral y bucal, como los sistemas de registros que pueden crearse al juntar datos provenientes de consultorios en el plano individual; los sistemas de registros donde se

agrega la información que se genera en el primer nivel de atención (policlínicas odontológicas) del ámbito privado y público, y, por último, las encuestas, que pueden ser de base poblacional o relativas a poblaciones mas pequeñas, pero debidamente identificadas.

3.1.1. Sistemas de registros

Existen diferentes sistemas de registros que se articulan en el primer nivel de atención del sector de salud del ámbito privado y público en los diferentes efectores (entre ellos, la atención que se hace a nivel de la salud pública a través de la Administración de Servicios de Salud del Estado (ASSE) o en ámbitos municipales) y que están pensados para la gestión de atención, pero que tienen mucha información con un gran potencial de uso epidemiológico. Además, está la información que se sistematiza a través del sistema Rediente, sistema desarrollado a medida para la Facultad de Odontología de la Universidad de la República.

Según consta en su sitio web,[1] «REDIENTE es una aplicación de software del sector salud que tiene como objetivo la salud bucal y la mejora en el ejercicio de la profesión y la docencia mediante el uso apropiado de la información clínica disponible. REDIENTE ofrece la Historia Clínica de los pacientes on-line (Intranet o Internet), es una herramienta de supervisión de la eficiencia profesional, una herramienta de investigación epidemiológica, así como una aplicación para la administración de servicios de la salud.». Este sistema permite obtener en todo momento indicadores de salud útiles para la gestión de calidad de la asistencia, para la supervisión docente y para el cumplimiento de normas. Rediente, además de proponer un formato unificado de registro, la historia clínica odontológica (hco), pone en manos del paciente los datos básicos de su tratamiento en el carnet rediente, apoya al docente y facilita el ejercicio profesional y del estudiante. Las instituciones como la Facultad de Odontología, ASSE, las mutualistas y seguros, las intendencias y ongs o los consultorios colectivos y particulares encuentran en rediente la información normalizada para conectarla a sus sistemas de reserva de horas, de facturación o de gestión. Es muy importante tomar esta definición como ejemplo de sistema de información estadístico-epidemiológico en el que es posible aplicar los indicadores propuestos.

Sobre esta sistematización de los sistemas de registros disponibles según sea el objetivo de estudio, conociendo su estructura, se podrían implementar mediante diferentes diseños muestrales adecuados el monitoreo o vigilancia usando la información que surge de los sistemas de registros. Esta opción es la que permite que la vigilancia sea realizable mientras no exista un sistema similar al de Rediente, que sea obligatorio y universal y de registro electrónico. Esta situación es análoga a la que se da en el monitoreo de la morbilidad por causas externas que está impulsando el equipo técnico de vigilancia de enfermedades no transmisibles del Ministerio de Salud Pública (MSP), que ha establecido vigilancia por muestreo en las emergencias hospitalarias.

1. https://www.bullseye.com.uy/portfolio/rediente/

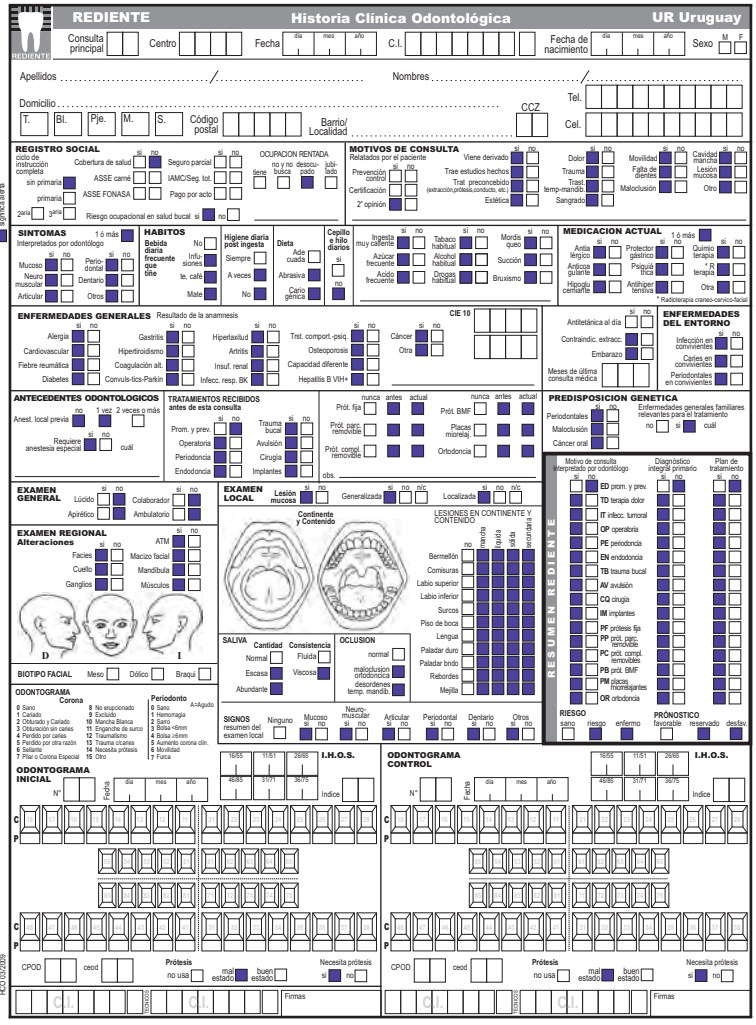

Figura 3.1: Historia clínica odontológica de Rediente

3.1.2. Encuestas de base poblacional

Se va a considerar la encuesta como un tipo de estudio en el que el mecanismo para generar información es la aplicación de un cuestionario sobre los individuos, ya que la información no puede, en principio, ser extraída de otra fuente de datos, como podría ser la historia clínica en el caso de otros tipos de estudio. Ramis, en (Oriol, 1997), señala que las encuestas son una fuente de datos no individualizable sobre la salud de las personas para ser usada en su medición, en la utilización de los servicios sanitarios, etcétera. Desde el punto de vista epidemiológico y de la salud pública, este mecanismo es el que va a permitir medir diferentes fenómenos en la población, de una forma homogénea, al aplicar un cuestionario donde se releve a través de preguntas estándares. Además, esta herramienta permite la evaluación a través de un corte transversal al aplicarse en un momento dado del tiempo como en 3.1.2, pero puede ser pensada para que se la aplique a través del tiempo como se detalla en 3.1.2, donde aparecen algunas ventajas extra y aspectos que también deben ser tenidos en cuenta.

Otro aspecto clave es el mecanismo estadístico de selección de los individuos en las encuestas, ya que no siempre se recurre al muestreo probabilístico, con las consiguientes limitaciones en cuanto al alcance de los resultados y el error cometido. En este trabajo, para que los desarrollos presentados sean válidos que para las encuestas consideradas, se está empleando algún mecanismo de muestreo aleatorio, de los cuales se hace una reseña en la sección 3.6.

Encuesta de corte tranversal

En este caso, la encuesta de corte tranversal (ECT) refiere al estudio en un momento dado, lo que coincide con los estudios de prevalencia, en los que se podrá evaluar una serie de variables dependientes e independientes.

Encuestas longitudinales

Las encuestas longitudinales EL, se refieren a estudios basados en observaciones repetidas efectuadas sobre las mismas unidades de muestreo a lo largo del tiempo, que pueden ser personas, hogares, empresas, entre otras.

La medición periódica de elementos permite hacer el seguimiento de la población objetivo. Logra captar su dinámica en el tiempo, donde el objetivo de medir cambios en la población a nivel macro puede lograrse mediante la comparación de resultados de encuestas transversales (cross-section) convencionales llevadas a cabo en distintos momentos del tiempo. La justificación de la utilización de encuestas de panel radica en el interés de medir cambios individuales (o micro) en poblaciones específicas. Los resultados particulares en cada instancia de medición (estimaciones transversales) pueden ser obtenidos sin perjuicio de lo anterior y, aunque no sea el objetivo principal de las encuestas de panel, son muy importantes.

Los distintos momentos del tiempo en los que las encuestas son llevadas a cabo se denominan olas. La duración del panel y el período entre las olas se definen en la etapa del diseño de la encuesta.

Algunos autores, como (Fuller y Braid, 1999), distinguen tres variaciones de panel: puto, rotativo y suplementado. El panel puro es aquel en que las mismas unidades son observadas en distintos momentos del tiempo. La muestra es extraída por única vez al inicio del estudio y todas las unidades seleccionadas serán observadas a lo largo de la duración del panel. Una unidad que no fue seleccionada al principio nunca pertenecerá al panel. El uso de este procedimiento en estudios mensuales (u otros estudios de carácter muy repetitivo) se basa en las ventajas que ofrece frente a métodos no rotativos: se evita sobrecargar a los respondentes logrando obtener mayor tasa de respuesta (las encuestas repetitivas tienen la característica de aburrir a los entrevistados causando abandono del panel por parte de ellos), Esto introduce un sesgo que se conoce como sesgo de abandono o attrition bias, (Álvarez-Vaz, 2017).

Una vez hecho el diagnóstico de las fuentes de información y decidido cuál es el mecanismo adecuado de generación de la información (sea trabajando con sistemas de registros o mediante encuestas por muestreo) se consideran una serie de indicadores, que se detallan a continuación, para evaluar luego cómo aplicarlos en cada fuente de información y, por último, analizar la sustentabilidad del sistema de vigilancia.

Como manera de presentar los diferentes indicadores que se usarán en los capítulos que forman parte de la obra y los artículos en los que se participa en coautoría y que se referencian, se ordenan los indicadores y se resumen del siguiente modo:

1. Indicadores clásicos que responden a la forma de medir las diferentes patologías descritas en la sección 2.1 del capítulo 2, tal cual fueron originalmente pensados.

2. Indicadores que combinan muchas variables odontológicas que no se suelen tratar en esa forma tal cual surge de la experiencia nacional e internacional en la bibliografía consultada y que se denominarán Indicadores combinados, los que se presentan en la sección 3.3.

3. Nuevos indicadores que surgen como reformulaciones de indicadores ya validados y aceptados, pero que incorporan modificaciones en sus algoritmos de cálculos, considerando métodos estadísticos más nuevos o que mejor se adecúan, dada la natulareza de los fenómenos bajo estudio, y que se presentan como Indicadores alternativos en la sección 3.4.

4. Indicadores que, mas allá de ser indicadores clásicos, combinados o alternativos, incorporan las dimensiones de tiempo y espacio asociados al fenómeno morboso en la salud bucal. Se presentan como Indicadores espacio-temporales en la sección .

3.2. Indicadores clásicos usados en salud bucal

Sin intentar ser exhaustivos, existen varios indicadores de las diferentes dimensiones que se determinan a nivel individual en salud bucal que se presentaron en el capítulo 2.2 y que pueden ser considerados a nivel colectivo desde una perspectiva epidemiológica.

Dentro de los que corresponden a la patología caries se consideran entonces los indicadores como el Índice CPO en niñas y niños (ceo), CPO, ICDAS. De los índices que dan cuenta del estado de las piezas dentales, es necesario hacer definiciones y establecer una nomenclatura de los diferentes unidades de observación:

- i, individuo; j, diente; k, cuadrante; s, superficie; g, grupo o subpoblación (se podría tener, por ejemplo, dos subopoblaciones: hombres y mujeres).
- Piezas dentales, $d^g_{i,j,k}$;
- Cuadrantes, $q^g_{i,.,k}$ (formados por piezas).
- Sextantes, $se^g_{i,.,k}$ (formados por piezas).
- Superficies de cada pieza $s^g_{i,j,l}$.

Con la información relativa al estado de las piezas dentales, se toma como primer ejemplo el CPO para razonar cómo es el proceso de construcción del indicador, sabiendo que si se considerasen el ceo o el ICDAS habría que hacer las modificaciones que correspondan.

P	T	U	V	W	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC
Motivo de consulta	c18	c17	c16	c15	c33	c34	c35	c36	c37	c38	C	P	O	CPO
TRATAMIENTO	~~8~~	~~6~~	~~3~~	~~3~~	~~6~~	~~2~~	~~1~~	~~3~~	~~3~~	~~6~~	4	8	8	20
EXTRACCION	~~8~~	~~3~~	~~3~~	~~3~~	~~2~~	~~8~~	~~3~~	~~3~~	~~3~~	~~8~~	~~3~~	~~17~~	0	20
EXTRACCION	~~2~~	~~3~~	~~1~~	~~2~~	~~8~~	~~3~~	~~3~~	~~3~~	~~3~~	~~8~~	~~10~~	~~5~~	~~2~~	17
EXTRACCION	8	3	3	3	0	2	3	3	3	8	~~8~~	~~14~~	~~8~~	16
PROTESIS	0	3	3	3	0	0	1	3	3	3	1	15	4	20
PROTESIS	3	3	3	3	0	0	1	3	3	3	1	23	1	25
PROTESIS	3	3	3	2	0	0	2	3	3	1	2	15	2	19
PROTESIS	0	2	3	3	0	0	0	3	3	8	2	12	0	14
PROTESIS	3	0	0	0	3	3	3	3	3	3	2	11	0	13
EXTRACCION	2	2	3	0	0	0	0	3	2	0	6	4	0	10
EXTRACCIONES REST	3	3	3	3	0	2	3	3	3	3	3	18	2	23
INFECCION	3	3	3	3	3	3	3	3	3	3	3	26	0	29

Figura 3.2: Datos individuales: cálculo del CPO

El CPO es un índice unidimensional que cuenta el número de dientes cariados C, perdidos P y obturados O. Ha sido utilizado durante mucho tiempo como una forma de determinar la historia de salud, medido a través de las caries de un conjunto de individuos. Los valores bajos de CPO indican un buen estatus de salud oral, mostrando que las piezas dentales tienen poca historia de enfermedad. Por lo general, las personas tienen,

salvo excepciones, un total de 28 a 32 piezas repartidas en 4 cuadrantes, 2 inferiores izquierdos y derechos y a su vez 2 superiores izquierdos y derechos , con un total de 7 piezas por cuadrante. Cuando las personas tienen incluso los terceros molares (lo que se habitualmente se llaman 'muelas del juicio') se puede tener hasta 32 piezas, con un total de 8 por cuadrante. De esta manera, para una persona en particular se puede evaluar el estado de las piezas a través del índice que se detalla en la siguiente ecuación:

$$CPO_i^g = \sum_j^n C_{i,j,k}^g + \sum_j^n P_{i,j,k}^g + \sum_j^n O_{i,j,k}^g \tag{3.1}$$

Sin embargo, el primer problema que presenta este indicador es que enmascara toda la variabilidad de las diferentes dimensiones que mide (2 de enfermedad presente (C,P) y 1 de enfermedad curada O). Por ejemplo, un mismo valor de CPO de 12 puede estar indicando situaciones muy diversas, como de una persona con 8 piezas obturadas y 4 con caries y, de otra, 5 con caries y 7 perdidas. En ambos casos, los niveles de enfermedad son importantes (tienen $\frac{12}{28} = 42{,}8\,\%$ de su piezas afectadas, es decir no sanas), pero no se sabe si la carga de enfermedad es la misma, ya que las piezas obturadas ponen de manifiesto una enfermedad pasada.

En virtud del ejemplo presentado antes, es necesario manejar alternativas, como utilizar los 3 componentes del CPO por separado, transformado en tasas, proporciones o índices basados en medidas de entropía. Esto implica considerar la misma información, pero analizándola de otra manera, teniendo en cuenta que se suele recoger información en el ámbito individual sobre características sociodemográficas que pueden estar asociadas a los niveles de enfermedad bucal. Medida con el CPO, se propone integrar esas características personales para evaluar diferencias para los componentes del CPO a través de modelos estadísticos, que son modelos predictivos, pero que, a su vez, son herramientas descriptivas importantes.

3.2.1. Modelos de regresión en epidemiología

Desde el punto de vista epidemiológico, es deseable recurrir a modelos parsimoniosos pero adecuados, en el sentido de que sean fáciles de estimar, sencillos de usar, que la información esté disponible y que los investigadores en ciencias médicas los entiendan y los adopten.

Las alternativas pueden ser:

1. $Y = f(X_{ij})$ (una única variable de respuesta) [tipo1];

2. $Y = f(X_{ij})$ (donde Y son varias variables de respuesta a la vez que dependen de un conjunto de variables explicativas. Estos modelos se denominan modelos generalizados aditivos vectoriales ,que se conocen como (VGAM),Yee y Hastie (2003) [tipo2].

Dentro de los muchos modelos que el investigador en biomedicina debería conocer del tipo 1, están los Modelos de regresión lineal (MRL) y los Modelos de datos categóricos (MDC) (Agresti, 2005), que pueden ser entendidos como casos particulares de los Modelos lineales generalizados (MLG). En este trabajo se consideran algunos modelos del tipo 1 que cada vez más pasan a estar difundidos en el campo de la epidemiología. Es por esto que se le debe prestar, entonces, especial atención a los modelos de conteo, mientras que otros modelos de tipo 2, aunque muy importantes, no se consideran en detalle en este trabajo.

3.2.2. Modelo de regresión logística

La variable de respuesta Y_i es una variable aleatoria Bernoulli. Sus resultados posibles son éxito y fracaso, codificados como $\{0, 1\}$, distribución de probabilidad: $P(Y_i = 1) = \pi_i$, $P(Y_i = 0) = 1 - \pi_i$, y[2]: $E(Y_i|X) = \pi_i$.

$$P(Y = 1|X) = \pi = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{3.2}$$

$$P(Y_i = 1|X_i) = \pi_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_p X_{pi}}}{1 + e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_p X_{pi}}} \tag{3.3}$$

El modelo planteado en (3.3) puede ser linealizado realizando la siguiente transformación:

$$\log_e(\frac{\pi}{1-\pi}) \tag{3.4}$$

Esa transformación se llama transformación Logit. El cociente $\frac{\pi}{1-\pi} = e^{X\beta}$ es lo que se denomina odds, que es un cociente de probabilidades: probabilidad de éxito sobre probabilidad de fracaso. Se lo considera como medida de riesgo. Por ejemplo, si se dice que los odds son 4 a 1 ($odds = 4$) de que una persona x padezca una enfermedad, es equivalente a decir que la probabilidad de enfermedad es 0,8.

Predecir el valor de la variable de respuesta en función de ciertos valores de las variables explicativas implica determinar cual es el valor crítico a partir del cual las estimaciones del valor esperado implican un valor de 1 para la variable de respuesta. Valores grandes de π_i implican $Y_i = 1$ mientras que valores chicos de π_i implican $Y_i = 0$. El problema del investigador es determinar cuándo una estimación se considera como un valor muy grande o muy chico. Existen distintas aproximaciones a tener en cuenta para determinar el punto de corte:

1. El punto de corte es 0,5. La regla de decisión es: si $\hat{\pi}_i > 0{,}5$ entonces $Y_i = 1$ La aproximación es válida si:

   a es igualmente probable que ocurra 0 o 1,

   b los costos de predecir en forma incorrecta 0 o 1 son prácticamente los mismos.

2. A los efectos de simplificar la notación en el texto se usará $\pi_i = \pi(X_i) = E(Y_i|X_i)$

2. Encontrar el mejor punto de corte para los datos. Esto implica calcular distintos puntos de corte y evaluar, en cada caso, cómo se pronostican las n observaciones. Se emplea el punto de corte con menor porcentaje de error. Este método es apropiado cuando:

    a los datos son una muestra aleatoria de la población, por lo que representan las proporciones adecuadas de 0 y 1,

    b los costos de predecir en forma incorrecta 0 o 1 son prácticamente los mismos.

3. Usar probabilidades a priori. Pueden usarse cuando los datos no son una muestra aleatoria de la población y cuando el costo de predecir incorrectamente 1 no es el mismo que el de predecir incorrectamente 0. De esta forma, se minimizará el valor esperado de predecir incorrectamente.

Como regla general, en la elección del punto de corte se busca optimizar el cociente entre los éxitos observados clasificados como éxitos y el total de éxitos observados —Sensibilidad (Se)—y el cociente entre los fracasos observados clasificados como fracasos y el total de fracasos observados —Especificidad (Esp)—del modelo. Estos dos conceptos se basan en un punto de corte específico a partir del cual se clasifica una observación como éxito o fracaso.

Una forma de evaluar la performance es ver el área bajo la curva: Sea n_1 la cantidad de observaciones con $y = 1$, n_0 la cantidad de observaciones con $y = 0$. Se crean $n_1 \times n_0$ pares donde cada individuo con $y = 1$ es apareado con individuo con $y = 0$. Del conjunto $n_1 \times n_0$ pares se cuenta la cantidad de veces que el individuo con $y = 1$ tiene la probabilidad más alta de las dos probabilidades. La proporción de esos individuos es el área bajo la Receiver Operating Curve (ROC).

3.2.3. Modelos de conteo

Los modelos de conteo (MCont) son una serie de modelos en los que la variable de respuesta refiere al número de veces que un evento ocurre. El evento de conteo es la realización de una variable aleatoria no negativa (Cameron, 1998), (Hirji, 2006), (Winkelmann, 2008).

Cuando este tipo de respuesta se desea explicar a través de modelos de regresión, se puede trabajar con el marco conceptual de la teoría de los MLG que ya autores como (Nelder y Wedderburn, 1972) presentaron a principios de los años setenta.

A partir de los trabajos de (Mullahy, 1986), (Zeileis y Kleiber, 2008), (Zeileis, 2006) y (Hilbe, 2011), los modelos se pueden clasificar como:

La densidad para la variable aleatoria Poisson es la que se consigna en (3.5)

$$f(y;\mu) \quad = \quad \frac{\exp(-\mu)\cdot\mu^y}{y!} \tag{3.5}$$

En la tabla 3.1 aparece como alternativa al modelo de Poisson cuando, por ejemplo, existe sobredispersión y da lugar a un Modelo binomial negativo (MBN) como surge de (3.6).

Tipo	Distribución	Descripción
MLG	Poi	Regresión Poisson estimado por máxima verosimilitud (MV)
	BN	regresión Binomial Negativa
	BNf	Regresión Binomial Negativa estimada por MV y que incluye parámetro de forma

Tabla 3.1: Diferentes alternativas a los modelos de conteo

$$f(y|\mu) = \int_0^\infty \frac{\exp(-\mu) \cdot \mu^y}{y!} f_\Gamma(\mu) d\mu \tag{3.6}$$

$$\mu \sim \text{Gamma}(\alpha, \beta) \tag{3.7}$$

donde el parámetro α controla la forma de la distribución y β su escala.

En principio, el investigador en biomedicina puede encontrar que la expresión de la (BN) es compleja, pero de la integral que allí aparece conceptualmente debe considerarse que, en realidad, en el caso de la BN, lo que se tiene es un modelo en el que la tasa de fallas μ no es fija y, por lo tanto, deben considerarse un promedio de ella, lo que se logra con la expresión analítica de la ecuación (3.6). En las secciones del capítulo 5 se retoman con más profundidad, explicitando cuándo es necesario usar alguno de los tipos presentados en la tabla 3.1 u otras variantes cuando estos dos modelos básicos en epidemiología no son suficientes para enfrentar los problemas de conteo.

Cualquiera de los modelos de conteo detallados en la sección 3.2.3 pueden presentar problemas adicionales como el exceso de ceros, situación en la que la cantidad de conteos iguales a 0 es excesivamente mayor a lo esperado. Para eso es necesario que estosse profundicen mediante reformulación a través de otros modelos (Mullahy, 1986) que se presentan en el capítulo 6 con una aplicación para el caso de los componentes C, P y O, en los que, además, se consideran otras distribuciones de probabilidad válidas para estos modelos.

3.3. Indicadores combinados

Los indicadores que se pueden agrupar en esta categoría son los que surgen de considerar técnicas descriptivas multivariantes como lo son el Análisis factorial (AF) y el Análisis de cluster o conglomerados (AC), técnicas que no son muy usadas actualmente en el ámbito de la epidemiología en Uruguay. El análisis factorial permitiría descartar las variables que son menos «importantes» (usando Análisis de componentes principales (ACP) o Análisis factorial de correspondencias múltiples (ACM)), (Blanco, 2006), (Cuadras, 2007).

3.3.1. Análisis factorial

La forma de presentar las técnicas de AF es considerarlas como una forma ordenada y sistemática de reducir el número de variables al crear índices (combinaciones lineales de las variables originales,a las que se le da una importancia diferente a cada una). La ventaja de estos métodos es que se adecúan al tipo de variable considerada, sean estas de tipo cuantitativa o cualitativa ordinal o cualitativa nominal. Un primer ejemplo es trabajar para el caso del CPO y su descomposición, considerando el total de C, P, y O por sextante. Esto hace que se deban considerar 6 variables, las que pueden mediante la aplicación de ACP, simplificarse en 2 o 3 factores al combinarse linealmente. Si, por otra, parte se analiza cómo es la asociación intracuadrante con respecto al CPO, con la ayuda del ACM se puede simplificar el problema al transformar las 6 variables multinominales de 3 categorías en 2 o 3 factores.

La otra ventaja que tiene la creación de estos indicadores (los factores) es que pasan a ser variables cuantitativas y que tienen, por lo tanto, las propiedades de ser resumibles y graficables. Así, es posible combinar variables de distinta naturaleza (cuantitativas y cualitativas) que de otra manera deberían ser trabajadas en forma separada. A su vez, como última ventaja de estos nuevos indicadores, aparece la posibilidad de crear tipologías o grupos de poblaciones con perfiles epidemiológicos bien diferenciados de acuerdo a las patologías y los factores de riesgos asociados con la ayuda del AC, que se presenta en la sección 3.3.2.

3.3.2. Métodos de clustering

Una forma de presentar los métodos de clustering es como una forma adecuada de separar los individuos en clases o categorías disjuntas de modo que los individuos que pertenecen a los grupos presentan más homogeneidad entre sí que si se los considera por separado. Esta metodología, por lo tanto, es una forma de crear subpoblaciones no observables o difíciles de definir a priori, planteando el problema en forma inversa a cuando se desea describir una población (toda o una muestra), ya que logra poner de manifiesto cuáles son las características que más asemejan o diferencian a los individuos.

Existen muy variados métodos de clustering, entre los cuales se encuentran los de tipo jerárquico y no jerárquico, sobre los cuales se pueden aplicar diferentes tipos de distancia, que toman en cuenta métricas diferentes que se adaptan al tipo de variable considerada. En los métodos que se reseñan en la obra se sigue la lógica de que la información se observa a través de un corte transversal.

Cuando los datos son de tipo longitudinal, los métodos de clustering son otros y, por motivos de extensión del trabajo, no se estudian aquí.

En el capítulo 8 se presenta una aplicación para la construcción de perfiles epidemiológicos mediante diferentes técnicas de clustering.

3.4. Indicadores alternativos

Se proponen nuevos indicadores construídos a partir del CPO que lo descomponen usando los componentes por separado al transformarlos en proporciones (Cribari-Neto y Zeileis, 2010).

Otra forma de trabajar es descomponer el CPO usando una nueva variable $Y = (CPO)$ formada por tres proporciones diferentes y usar modelos generalizados para respuesta multivariada. Modelar el CPO como un vector permite analizar relaciones entre las proporciones que componen dicho índice que se pierden cuando se colapsan en un único indicador. Esto se realiza con modelos de regresión multivariada (Yee y Hastie, 2003), (Yee, 2010).

3.4.1. Métodos CART

Este nuevo enfoque (poco usado en la epidemiología de salud bucal, aunque cada vez más) está basado en técnicas estadísticas de aprendizaje supervisado, como los métodos de Classification and Regression Trees (CART), que es español se conocen como árboles de clasificación y regresión. Esta metodología permite la construcción de modelos basados en técnicas no paramétricas, lo que supone muchas menos restricciones de distribución de probabilidad en las variables consideradas. Esto, a su vez, permite encontrar las variables que mejor discriminan el comportamiento de una variable de respuesta o dependiente de tipo categórica. La aplicación immediata se hace sobre variables que clasifican en ausencia o presencia de una patología en diferentes niveles de patología (maloclusión, enfermedad periodontal) como, por ejemplo, la maloclusión en escolares (Álvarez-Vaz et al., 2011). La gran ventaja de estas técnicas es que prescinden de un modelo analítico explícito (como puede ser los modelos de regresión lineal múltiple, de regresión logística o análisis discriminante probabilístico), lo que las hace más fácilmente usables e interpretables por los no especialistas en estadística (Breiman et al., 1984), (Abernathy et al., 1987).

3.4.2. Modelos probabilísticos para ajustar tasas

Una posibilidad es pensar en transformaciones de la variable de respuesta, que al ser variables de conteo se pueden reformular como tasas o proporciones, para el caso presentado los índices CPO o algunos de sus componentes relativizados contra diferentes totales: 32 (máximo número de piezas, número de piezas presentes, etc.). La ventaja de estas transformaciones es que existen modelos probabilísticos conocidos para trabajar con tasas, proporciones o índices de concentración, de los cuales se conocen muchas ca-

racterísticas necesarias a la hora de usarlos para hacer inferencia estadística. Es necesario tener presente que las proporciones a estimar que están en el rango $(0, 1)$ no cumplen el supuesto de normalidad, por lo que pueden existir asimetrías muy importantes, y que la varianza puede cambiar entre los individuos.

A modo de ejemplo, se pueden relativizar los componentes del CPO convirtiéndolos en proporciones usando los 3 componentes del CPO del siguiente modo:

- $\frac{\sum O_i}{\sum O_i + \sum C_i}$ nivel de cobertura de la enfermedad previa a la entrada al programa;
- $\frac{\sum C_i}{\sum O_i + \sum C_i}$ indicador de estadío de la enfermedad en el momento actual;
- $\frac{\sum S_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$ indicador de salud en caries en el momento actual;
- $\frac{\sum P_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$ indicador de necesidad de prótesis en el momento actual.

En el capítulo 7 se presenta la aplicación de la regresión BETA y se detalla la formulación del modelo predictivo, que puede hacerse a partir de los trabajos de (Kieschnick y McCullough, 2003), (Salinas-Rodríguez et al., 2009).

3.4.3. Índices basados en la teoría de la información

En este caso, para poder discriminar el comportamiento en la población de los diferentes indicadores a evaluar en la salud bucal, se pueden adaptar índices basados en la teoría de la información de Shanon, ampliamente usados en economía y demografía económica.

Dentro de este grupo se pueden encontrar diferentes casos de lo que son los índices de entropía (Shannon y Weaver, 1949). En el capítulo 11 de la tesis de doctorado (Álvarez-Vaz, 2020) se proponen diferentes medidas de desigualdad con una aplicación a una encuesta de base poblacional en niños, que se presenta en la sección 4.2.

A su vez, cualquier indicador que se elabore usando técnicas no detalladas en este capítulo se considera como un indicador alternativo.

3.5. Indicadores espaciotemporales

Los indicadores que se pueden aplicar son de tres tipos y buscan tener un manejo en el tiempo y espacio de los fenómenos morbosos de la salud bucal, tal como se trabaja en la epidemiología de las ET. Para cada tipo de indicador se plantean conceptos clásicos de epidemiología descriptiva, que se acompañan de técnicas estadísticas muy sencillas que ya tienen varios años y que, a su vez, presentan algunas limitaciones. No obstante, se presentan alternativas para luego mostrar algunas metodologías que suponen el uso de técnicas más modernas y que hoy en día pueden usarse gracias al desarrollo computacional.

3.5.1. Agregaciones espaciales

Existen muchos métodos para detectar agregaciones espaciales que se basan es establecer distancias entre objetos, que en este caso son los puntos de muestreo o recolección de información; estos puntos ubicados espacialmente pueden representar la agregación de casos o o eventos en una determinada localización geográfica mostrando un patrón espacial que pueda considerarse como no aleatorio (las localizaciones pueden corresponder a unidades geográficas como estados, o departamentos, ciudades, barrios o conjuntos de manzanas). Para la evaluación de estas situaciones pueden usarse:

Método Grimson

Esta técnica fue diseñada para la detección de agregaciones espaciales de alguna enfermedad, dentro de una región que se divide en áreas más pequeñas. Se usa cuando hay sospecha de que las áreas de mayor riesgo tienden a agruparse espacialmente. El test estadístico consiste en comparar el número de pares de áreas adyacentes de alto riesgo contra un umbral o número, bajo el supuesto de que dichas áreas de alto riesgo se distribuyen aleatoriamente en la región de análisis (Grimson et al., 1981). Para poder evaluar el estadístico de prueba es necesario conocer cada una de las áreas y los índices que identifican cada una para saber si son áreas adyacentes a ella o que comparten fronteras. A su vez saber el valor de las tasas de incidencia de la patología bajo estudio en cada área y finalmente tener un umbral de corte para las tasas de incidencia, que permita definir área de riesgo. A partir de N (número total de áreas), - n (número de áreas de alto riesgo), P (número de pares de áreas adyacentes de alto riesgo), n_i (número de áreas adyacentes al área $i (i = 1, ..., N)$) y la media (μ) y varianza (σ^2) de estos N valores.

$$P_i = Pr[Poisson(E(P)) \geq P] \tag{3.8}$$

$$P_i = Pr[N(0,1) \geq \frac{P - E(P)}{\sqrt{V(P)}}] \tag{3.9}$$

Método Ohno

Es un método desarrollado para identificar patrones geográficos de mortalidad o morbilidad observados en el mapa de una región dividida en áreas más pequeñas (municipios, comarcas o áreas sanitarias, por ejemplo). La lógica que se sigue en este tipo de test es suponer que áreas adyacentes tienden a presentar niveles de la enfermedad más similares de lo que se esperaría por azar, con el supuesto de la agregación espacial (Ohno et al., 1979). Por lo tanto, para poder calcular el estadístico de prueba se necesita la identificación de cada una de las N áreas en que se divide la región de estudio, los índices que identifican para cada área para poder identificar áreas adyacentes y tasas de incidencia de cada una. A su vez, el número de categorías (k) en las que se quieren agrupar las áreas en función de sus tasas de incidencia y los $k - 1$ puntos de corte de las tasas de incidencia que definen esas categorías. El primer paso del método consiste en clasificar las áreas según el nivel de riesgo de enfermedad que presenta cada una, es decir, atender

a los puntos de corte definidos para las tasas de incidencia. Así, a partir de los $k-1$ puntos de corte $V_i(i=1,...,k-1)$ y las tasas de incidencia de la enfermedad para cada área $T_i(i=1,...,N)$, las k categorías de riesgo se determinan del siguiente modo:

1. se calcula el número de pares de áreas, dentro de cada categoría, que son adyacentes, es decir, el número de pares de áreas que cumplen a la vez en el mismo nivel de riesgo (son concordantes) y tienen también contigüidad espacial (son adyacentes);

2. con estos insumos se calcula un estadístico que tiene distribución χ^2, donde se compara el número observado de pares de áreas adyacentes y concordantes en la categoría i-ésima $(AC_i, i=1,...,k)$ y un número esperado $(EAC_i, i=1,...,k)$ bajo la hipótesis de que la distribución es uniforme en toda la región de estudio.

$$\chi^2 = \left(\frac{AC_i - EAC_i}{\sqrt{EAC_i}}\right)^2, i+1,....K \tag{3.10}$$

donde $EAC_i = \frac{AN_iN_i(N_i-1)*0,5}{N(N-1)*0,5}$, con A número total de pares adyacentes.
Este análisis se puede complementar con un estadístico global.

$$\chi^2 = \left(\frac{AC - EAC}{\sqrt{EAC}}\right)^2, i+1,....K \tag{3.11}$$

donde $AC = \sum_{i=1}^{i=k} AC_i$ y $EAC = \sum_{i=1}^{i=k} EAC_i$, que permite detectar agrupaciones espaciales en todas las regiones.

3.5.2. Agregaciones temporales

Para la detección de agregaciones temporales se pueden usar indicadores que buscan dar cuenta de si el número o proporción de casos de la patología en estudio que aparece en intervalos de tiempo consecutivos sucede con una frecuencia diferente a la esperada si se tratara de una distribución aleatoria. Algunos de los métodos para casos individuales se consigna a continuación.

Método Chen

En este caso, para evaluar si existe una agregación temporal se deben considerar T, la tasa esperada de incidencia anual de la enfermedad (puede ser una tasa histórica o del pasado reciente que haya prevalecido hasta antes de la sospecha de epidemia); P, que tiene en cuenta el tamaño de la población expuesta riesgo; FP, la tasa de falsos positivos, es decir, la probabilidad de rechazar la hipótesis nula siendo cierta a causa de la aparición de casos falsos de la enfermedad en el período estudiado, siendo que esta probabilidad puede verse como el nivel de significación de la prueba. Debe usarse también la fecha inicial como fecha anterior al primer caso considerado (Chen et al., 1982). Con la información presentada se calcula la longitud temporal máxima (o intervalo de referencia), que esta dada por:

$$LC = -log(1 - FP^{1/N})\ LE;\ LE = \frac{12}{PT} \tag{3.12}$$

donde N es el número total de casos y LE es la longitud esperada del intervalo entre casos consecutivos.

Método Sets

Esta técnica fue diseñada para estudiar y controlar la incidencia de malformaciones y de otras enfermedades raras, como ciertos tipos de cáncer. Se basa en el Set, que puede definirse como el intervalo de tiempo entre dos casos consecutivos de una enfermedad o, en general, entre dos eventos consecutivos. La lógica que hay detrás de este test es que si cada Set de una secuencia de casos es menor que cierto valor de referencia, se toma como umbral para indicar la existencia de una agregación de casos que no es aleatoria y es significativa en el área de estudio (Chen, 1978), (Chen, 1979), (Chen et al., 1983), (Chen, 1986). Para aplicar el método Sets es necesario tener la serie de casos con sus respectivas fechas de ocurrencia, T (tasa esperada de incidencia anual de la enfermedad), P (el tamaño de la población expuesta a riesgo), A (número de años en falsos positivos). Se supone que, al aumentar A, se reduce la probabilidad de rechazar la hipótesis nula que sea cierta, a causa de la aparición de casos falsos de la enfermedad en el período estudiado. También se necesita la fecha inicial. Se toma en cuenta cada secuencia de n casos, en las que se comparan los Sets de la secuencia con el intervalo o valor de referencia, que se determina como un múltiplo del tiempo esperado entre casos consecutivos (inverso del número esperado de casos anuales).

$$R = \frac{k}{\text{número esperados de casos}} = \frac{k}{P * T} \tag{3.13}$$

Método Poisson

El método de Poisson, como su nombre indica, está basado en la distribución de Poisson y y es el más sencillo de los usados para detectar agregaciones de casos en el tiempo (Weatherall y Haskey, 1976). Es necesario conocer la serie temporal de casos observados O_i y la serie temporal de casos esperados E_i

$$P_i = Pr[Poisson(E_i) \geq O_i] \tag{3.14}$$

La decisión es si $P_i \leq \alpha$, con nivel α de significación.

Método Scan

El estadístico Scan, en el que se basa este test, es el número máximo de casos observados en una ventana temporal móvil de tamaño fijo, que se mueve de forma continua a lo largo del tiempo. Si hay indicio de que hay agrupación de casos en el tiempo, entonces el número máximo de casos en la ventana temporal será grande. El tamaño de la ventana se basa en la duración esperada de una epidemia, (Naus, 1965), (Naus, 1982). Se necesita contar con la serie temporal de casos observados $X_i, i = 1, ..., T$, y w: el tamaño de la

ventana, expresada en las mismas unidades de tiempo. Si $X_{t,t+w}$ representa el número de casos en el período $(t, t+w]$, el estadístico Scan puede expresarse como

$$S_w = \underbrace{max}_{0 \leq t \leq Tw} Y_{t,t+w} \tag{3.15}$$

En este test se rechaza la hipótesis nula, con un nivel de α, si la probabilidad de que el estadístico sea mayor que el valor observado en la serie es mayor o igual que α:

$$Pr(S_w \leq n) \geq \alpha$$

donde el valor p depende del número total de casos observados (N), la amplitud relativa de la ventana respecto al período total de estudio $(r = w/T)$ y el número máximo de casos observados en todas las ventanas temporales de amplitud $w(n)$. Se consigna como $Pr(n, N, r)$, que debe entenderse como la probabilidad de que si N puntos están distribuidos aleatoriamente a lo largo de una línea de longitud 1 exista un intervalo de longitud r que contenga al menos n de ellos. Como no se conoce la distribución del estadístico Scan bajo la hipótesis nula, el valor p debe ser aproximado Simulación Monte Carlo (SM) o mediante aproximaciones basadas en la distribución binomial.

Método Texas

El método Texas es un procedimiento estadístico que permite detectar agregaciones temporales cuando se considera la incidencia de una enfermedad de baja frecuencia, desarrollado para ser usado en el contexto de una comunidad potencialmente expuesta a una fuente contaminante (Hardy et al., 1983), (Hardy et al., 1990). Esta técnica se basa en una regla de decisión que funciona en dos etapas y que utiliza la Razón estandarizada de mortalidad (REM) para identificar un aumento significativo en la mortalidad o morbilidad de una región con respecto a lo esperado. El método establece dos umbrales $(U_1 < U2)$, con los que se definen una fase de alerta, al ser un posible indicio de problemas. Cuando se supera el umbral U_2 ya deja de ser una alerta para decidir una acción o intervención. Se necesita la serie temporal de casos observados, O_i, la serie temporal de casos esperados E_i, P_1, que es la probabilidad de alerta y P_2, la probabilidad de acción $(P_2 < P_1)$. Para llevar adelante la aplicación del cálculo de valores umbrales y la regla de decisión es necesario decidir si en función de la cantidad de casos se puede usar una aproximación por la distribución normal o la distribución exacta que podría ser distribución de Poisson (DPoi), para evaluar el número de casos esperado. Para $n < 30$ se determinan los valores umbrales, P_1,P_2. Para cada período se calcula la probabilidad de observar al menos los O_i casos observados bajo supuesto DPoi con media E_i:

$$P_i = Pr[Poisson(E_i) \geq O_i] \tag{3.16}$$

Se aplica la siguiente regla:

- ALERTA: $U_1 < P_i < U_2$

- ACTUACIÓN: $P_i \geq U_2$ o -$U_1 < P_i < U_2$ y $U_1 < P_{i-1} < U_2$ (2 alertas consecutivas)

Cuando $n \geq 30$ los umbrales se determinan:

$$U_i = 1 + \frac{\phi^{-1}(1 - P_i)}{\sqrt{(E_i)}}; \tag{3.17}$$

Se calcula la Razón Estandarizada de Mortalidad (REM) de cada período como el cociente entre el número de casos observados y esperados, para luego aplicar la siguiente regla:

- - ALERTA: $U_1 < SMR_i < U_2$
- - ACTUACIÓN: $SMR_i \geq U_2$ o
- $U_1 < SMR_i < U_2$ y $U_1 < SMR_{i-1} < U_2$ (2 alertas consecutivas)

Método Cusum

A partir de una técnica creada para el control de calidad en la industria, la epidemiología usa este método para detectar incrementos en la incidencia de las patologías, para lo cual es necesario tener:

- la serie temporal de casos observados (en días, semanas, cuatrisemanas o meses), O_i,
- el número esperado de casos por unidad de tiempo, E,
- el $Cusum$ inicial.

Usando el número esperado de casos (E) se obtienen dos valores K, el de refencia y h, el de alarma, tratando de maximizar la capacidad de detección del método, pero intentando minimizar la probabilidad de producir falsas alarmas. Este método funciona decidiendo al comparar el valor $Cusum$, calculado para cada intervalo de tiempo, con el valor h. Si $Cusum > h$, se detecta un cambio significativo en la frecuencia de la enfermedad y se declara una alerta. Para calcular $Cusum$ se procede del modo siguiente:

- Para cada período k se calcula: $S_k = Max\{0; O_{i-K}\}$ si $E < 9$
- $S_k = Max\left\{0; \frac{O_i - K}{\sqrt{E}}\right\}$ si $E \geq 9$

El $Cusum$ del período i es la suma acumulada de valores S_k desde el período inicial, mientras no se haya declarado ninguna alerta o, en caso contrario, desde el período siguiente al de la última alerta y hasta el período i. Para calcular el primer valor acumulado $Cusum_1$ se le suman a S_1 el valor de $Cusum_0$. Gráficamente puede evaluarse viendo cómo, a través del tiempo, la trayectoria que sigue el estadístico $Cusum$, evaluando si no cruza un umbral máximo, donde se dispara la alerta (Scrucca, 2004).

Método Ederer-Myers-Mantel

Se lo usa para detectar agrupamientos en el tiempo en varias series de tiempo. El método que trabaja con conteos de casos busca encontrar la frecuencia máxima entre subintervalos disjuntos, para lo cual un valor significativamente grande indicaría evidencia de cluster temporal. El estadístico de prueba es $M = max(n_1, ..., n_m), n = n_1 + ... + n_m$, donde n_i representa el conteo para cada intervalo, considerañdose despreciables los cambios en la población expuesta a riesgo. El estadístico de prueba es el siguiente

$$\chi_1^2 = \frac{(|\sum M - \sum E(M)| - 0{,}5)^2}{\sum(Var(M))} \tag{3.18}$$

siendo $E(M)$ y $V(M)$ el valor esperado de M y su varianza, respectivamente, con fórmulas de cálculo de $E(M)$ y $V(M)$, por lo cual se puede evaluar el test mediante SM (Stark y Mantel, 1967), (Tango, 2010), (Rowlingson y Diggle, 2017).

Índice de Tango

Este índice busca poner en evidencia el agrupamiento de casos en el tiempo a través del siguiente estadístico:

$$C = r^t Ar = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{n_i n_j}{n^2} a_{ij} \tag{3.19}$$

donde $r^t = \frac{(n_1,...,n_m)}{n}$ y a_{ij} es una medida de la cercanía entre el i-ésimo y el j-ésimo intervalo, subintervalo $a_{ij} = exp(-|i-j|)$, y donde el estadístico de prueba se puede calcular mediante SM (Gómez-Rubio et al., 2005),(French, 2015b), (Tango, 2010).

3.5.3. Agregaciones espaciotemporales

Se busca encontrar si existen clusters en el espacio, por un lado y en el tiempo, en forma simultánea, pensando entonces que exista interacción. Aquí la interacción equivale a suponer que los casos cercanos en el espacio son, además, cercanos en el tiempo, lo que obliga considerar que la localización de un caso depende de la localización del caso que lo precede. Una metodología adecuada para tratar ambos componentes es a través de un método estadístico temporalmente dinámico y espacialmente descriptivo, para lo cual pueden usarse algunos tests, como:

Test de Knox

Este método fue diseñado por (Knox, 1964) para detectar agregaciones espaciotemporales que ocurren cuando los casos observados de una enfermedad en determinada región guardan cercanía, tanto en espacio como en tiempo. El método de Knox tiene la ventaja de que no es necesario conocer el tamaño ni las características de la población en estudio. Para aplicar el método de Knox es necesario disponer de los siguientes datos: las coordenadas espaciales X e Y, correspondientes a cada uno de los casos de enfermedad ocurridos

durante el período de estudio, que se determinan en un sistema cartesiano, tomando como origen un punto arbitrario dentro de la región estudiada y están expresadas en una unidad de longitud; las fechas de ocurrencia de los casos (dd/mm/aa). Además, debe definirse de antemano qué se entenderá por proximidad espacial y temporal. Para ello, el investigador tiene que fijar dos valores críticos, e y t, que constituyen, respectivamente, la distancia máxima aceptada entre dos casos para que puedan ser considerados cercanos espacialmente (distancia espacial crítica) y el tiempo máximo entre la ocurrencia de dos casos para considerarse próximos en tiempo (distancia temporal crítica). Los criterios para definir la cercanía en el espacio y en el tiempo dependerán de las características epidemiológicas de la enfermedad. El procedimiento consiste en calcular las distancias espaciales y temporales entre todos los pares de casos y, a partir de los valores críticos establecidos, definir dos variables dicotómicas que expresen si un par de casos están o no próximos en el espacio y en el tiempo. Para calcular la distancia espacial entre dos casos (x_i, y_i) y (x_j, y_j) se utiliza la distancia euclidiana (la recta más corta que une dos puntos) (Knox, 1964).

Test de Kulldorf

Este test es una variante al de Knox, que se modifica para situaciones en las que hay una tasa de crecimiento poblacional variable en diferentes regiones y en las que se construye un estadístico que se basa en replicar varias muestras de los casos observados, permutando su ubicación espacial y temporal, con probabilidad proporcional al tamaño de la población de esa región y período (French, 2015a). Se evalúa mediante Simulación Monte Carlo (SM) y permutaciones de varianza.

Test de Diggle

Es otra variante al método de Knox. Permite un procedimiento capaz de evaluar varios diagnósticos útiles para interacciones y, a su vez, hacer tests espaciotemporales para evitar múltiples pruebas (Tango, 2010).

Test k-NN de Jacquez

Este test busca contar el número observado de pares de casos cercanos en espacio y tiempo en los que la medida de cercanía está definida por los k vecinos más próximos y un valor significativamente grande indicaría evidencia de agrupamiento de espacio-tiempo de la enfermedad bajo estudio. Se define un estadístico

$$T_k = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i^S a_j^T \tag{3.20}$$

donde $a_i^S a_j^T$ expresan la cercanía en espacio y tiempo:

$$a_{ij}^S = \begin{cases} 1, \text{si el caso j es un KNN del caso i}(\neq j) \text{ en el espacio} \\ 0, \text{en otro caso} \end{cases} \tag{3.21}$$

$$a_{ij}^{T} = \begin{cases} 1, \text{si el caso j es un KNN del caso i}(\neq j) \text{ en el tiempo} \\ 0, \text{en otro caso} \end{cases} \tag{3.22}$$

Dados los valores de k, se evalúa si el test muestra evidencia a favor de agrupaciones mediante SM, permutando los tiempos (ubicaciones espaciales) entre las ubicaciones espaciales fijas (tiempos) (Tango, 2010).

Estos indicadores deben considerarse como herramientas epidemiológicas que están pensadas para la descripción. El cálculo en todos ellos es relativamente sencillo y dejan de lado el modelado.

Además de esta batería de indicadores que permite guiar una acción de intervención para modificar la situación epidemiológica fuera de la común, se puede pensar en la construcción de mapas de riesgo que permiten visualizar la situación del fenómeno en términos de la distribución territorial que se haya considerado. Esos mapas de riesgo que pueden ser fácilmente interpretados tienen un proceso de construcción más complejo. Estos nuevos indicadores suponen manejar un herramental estadístico más avanzado, en tanto tienen que manejar procesos estocásticos, con los cuales hay que estimar Variogramas (Vario), para lo cual es necesario estimar funciones de autocorrelación espacial.

La autocorrelación espacial puede definirse como el fenómeno por el cual la similitud geográfica (observaciones próximas espacialmente) se une con la similitud de valores. Así, valores altos o bajos de una variable aleatoria tienden a agruparse en el espacio (autocorrelación espacial positiva) o bien se sitúan en localizaciones rodeadas de unidades vecinas con valores disímiles (autocorrelación espacial negativa). Los estadísticos que suelen utilizarse para medir autocorrelación espacial son, entre otros, el Índice de Moran (IM) y el Índice de Geary (IGea) (Bivand et al., 2008). El análisis de la autocorrelación espacial permite descubrir si se cumple la hipótesis de que una variable tiene una distribución aleatoria o si, por el contrario, existe una asociación significativa de valores similares entre zonas vecinas.

En una segunda instancia se puede hacer interpolación por método de krigging sobre los puntos de muestreo georreferenciados. A partir de esta interpolación direccional se pueden obtener curvas de nivel que se pueden representar gráficamente en mapas. Para evaluar su aplicabilidad existen diferentes herramientas de Sistemas de Información Geográfica (GIS) que permiten generar esos indicadores, conociendo los algoritmos que manejan éstos y los supuestos que están detrás de éstos (Waller y Gotway, 2004), (Lawson y Kleinman, 2005), (Pfeiffer et al., 2008), (Lawson, 2009), (Tango, 2010).

3.6. Aspectos a considerar en el trabajo con muestras probabilísticas

En esta sección se plantean algunos aspectos relativos al muestreo de poblaciones finitas que suele ser necesario tener en cuenta en el ámbito de la epidemiología. Se manejan

una serie de elementos también básicos para poder trabajar adecuadamente en función del mecanismo de generación de los datos, lo que garantiza la validez de los métodos estadísticos que se usan habitualmente. La forma de presentarlos sigue la lógica que se maneja en libros como (Särndal et al., 1992) y, en particular, en el libro (Álvarez-Vaz, 2017).

Por lo tanto, al considerar el concepto de diseño muestral se debe tener en cuenta que se hará un manejo diferente al que se hace con la teoría estadística tradicional. Mientras que en este último caso el carácter de la aleatoriedad surge de un modelo hipotético, en el de esta nueva lógica de trabajo es el mecanismo de extracción de la muestra el que modula el carácter randómico de lo observado (Cochran, 1977), (Martínez et al., 1997). Los diseños muestrales no dependen de los valores de las variables de interés en la muestra, es decir, no dependen de los y_k observados, aunque sí pueden depender de los valores de variables auxiliares x_k.

Para eso se van a considerar diferentes aspectos que se toman en cuenta en la estadística en esta nueva forma de trabajar en los diseños muestrales, como son los conceptos de población, muestra y diseño muestral y, desde la epidemiología, apropiarse del lenguaje para entender todo el proceso de construcción del mecanismo de muestreo. Esto permitirá que se conozca por qué se hace de cierta manera y las consecuencias que tiene en términos de costos, precisión, capacidad de desagregación (el estudio en dominios), que se trata también en la sección 3.6.6. Todo el desarrollo que se hace más adelante es diferente del que se propone, por ejemplo, en textos clásicos como Hansen et al. (1953), Kish (1965), Raj (1968), Cochran (1977), utilizados en la formación en epidemiología. El enfoque que se adopta está basado, en cambio, en el trabajo Model Assisted Survey Sampling (Särndal et al., 1992) y en Estimation in Surveys with Nonresponse (Särndal y Lundström, 2005).

3.6.1. Diseño SI

En el diseño Simple (SI) se define una población U de la que de N elementos se extraen n elementos de manera independiente y sin reponer.

Hay C_n^N muestras posibles de tamaño n. El diseño muestral viene dado por

$$p(s) = \begin{cases} \left(C_n^N\right)^{-1} & \forall s \text{ de tamaño } n \\ 0 \text{ en otro caso} \end{cases}$$

Para poder llevar adelante el proceso de estimación de totales (las medias y proporciones surgen de estos totales) es necesario poder calcular las probabilidades de inclusión de primer y segundo orden, que son

$$\pi_k = \frac{n}{N}, \quad \forall k \in U \tag{3.23}$$

$$\begin{cases} P(k\ y\ l \in S) = \pi_{kl} = \frac{n(n-1)}{N(N-1)}\ \forall k \neq l \in U \\ \Delta_{kl} = -\frac{n(N-n)}{N^2(N-1)}\ \forall k \neq l \in U \end{cases} \tag{3.24}$$

En particular, el estimador π del total poblacional de la variable y, t_y, viene dado por $\hat{y}$, que es insesgado.

$$\hat{t}_y = \sum_S \check{y}_k = \sum_S \frac{y_k}{n/N} = N \sum_S y_k = N\bar{y}_S \tag{3.25}$$

Cuando se tiene un diseño aleatorio simple (SI), de tamaño n sobre una población de tamaño N, se tiene que la varianza del estimador π de un total poblacional y un estimador insesgado de esta vienen dados, respectivamente, por

$$\begin{aligned} V_{\mathrm{SI}}\left(\hat{t}_\pi\right) &= N^2\left(1-f\right)\frac{S^2_{y_U}}{n} \\ \hat{V}_{\mathrm{SI}}\left(\hat{t}_\pi\right) &= N^2\left(1-f\right)\frac{S^2_{y_S}}{n} \end{aligned} \tag{3.26}$$

donde $f = \frac{n}{N}$ se denomina fracción de muestreo; $(1-f)$ factor de corrección por población finita y $S^2_{y_U} = \frac{1}{N-1}\sum_U \left(y_k - \bar{y}_U\right)^2$ donde $\bar{y}_U = \frac{1}{N}\sum_U y_k$. Por último, $S^2_{y_S} = \frac{1}{n-1}\sum_S \left(y_k - \bar{y}_S\right)^2$ es la varianza muestral.

3.6.2. Efecto diseño (DEFF)

El diseño SI es el que se toma como base de comparación para otros diseños que se verán más adelante, ya que hay que sacrificar la pérdida de precisión (tener más varianza) para ganar la posibilidad de tener un diseño más económico. Estos diseños son:

- Muestreo sistemático (SY),
- Muestreo estratificado (STSI),
- Muestreo por conglomerados (SIC),
- Muestreo en varias etapas (MMult).

Tal como se plantea en (Álvarez-Vaz, 2010) y (Álvarez-Vaz, 2017), no es frecuente que se use un diseño puro, sino que es necesario combinar los diseños antes mencionados, por lo cual se termina trabajando con MMult. Por este motivo, la pregunta es cómo comparar la eficiencia entre distintos diseños y saber por cuál optar. La respuesta es que esto depende de distintos factores:

- distribución de los valores de la variable de interés en la población,
- parámetro a estimar,
- estimador utilizado,

- disponibilidad de información auxiliar.

Una forma de medir la pérdida o ganancia de eficiencia por cambiar de diseño de muestreo es comparar las varianzas entre el diseño que se quiera usar y el diseño SI.

El efecto del diseño p(s) para estimar t con el estimador $\hat{t}_\pi$ es

$$\text{Deff}\left(\ p(s)\ ,\ \hat{t}_\pi\right) = \frac{V_{p(s)}(\hat{t}_\pi)}{V_{SI}(\hat{t}_\pi)} \tag{3.27}$$

Lo importante es recordar que se está evaluando la eficiencia de $p(s)$ (el diseño en uso) relativizándola al diseño que se utiliza como base de comparación al muestreo (SI).

3.6.3. Diseños en varias etapas

Con la misma lógica de considerar dos, tres o más etapas interesa ver cómo pueden establecerse diferentes diseños en forma flexible combinando los tipos de diseños puros vistos antes, según la necesidad de cada caso. Para eso se toma como ejemplo una de las situaciones planteadas en el manual WHO Steps Surveilllance Manual (World Health Organization, 2006) de donde se extraen la figura que sigue, que corresponde a uno de los diferentes diseños. La lógica del diseño empleado siempre es disponer de un marco muestral adecuado que permita seleccionar las unidades primarias de muestro (UPM) y, en segunda instancia, dependiendo de la calidad de la información, trabajar con dos, tres o más etapas de muestreo.

Un diseño recomendado por el equipo de STEPS es muy similar al de la figura 3.3, donde se listan todos los residentes y, en la última etapa de muestreo, se seleccionan los participantes mediante Tabla de Kish (TK).

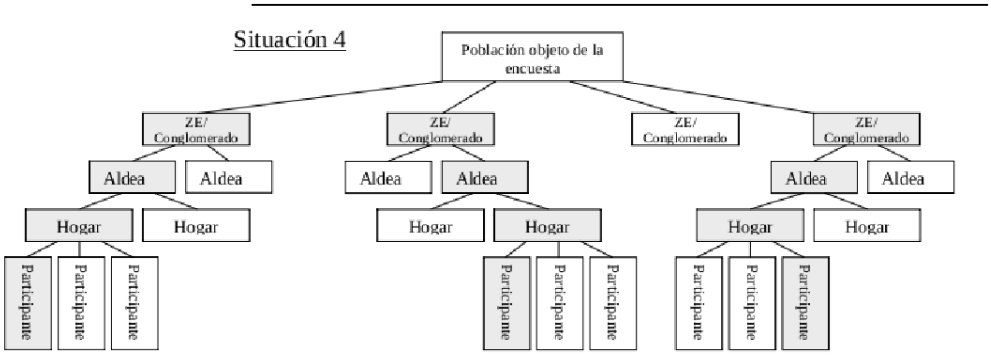

Figura 3.3: Diseño en 3 etapas, figura extraída del manual de encuestas (World Health Organization, 2006)

Por último, se puede seguir esta serie de recomendaciones, que tiene en cuenta, a su vez, los fundamentos téoricos vistos en las secciones precedentes para cada tipo de diseño de muestreo:

- El número de unidades (conglomerados o individuos) de cada etapa depende del número que se tenga de cada una en el marco, en cada etapa de muestreo.
- Es necesario evaluar el tamaño de muestra deseado global y para cada estrato (por ejemplo: edad*sexo).
- Es más recomendable muestrear mayor cantidad de conglomerados de menor tamaño que muestrear un menor número de conglomerados más grandes.
- Hay que considerar el costo asociado de muestrear muchas localidades.

3.6.4. Cálculo de la varianza por aproximación

Hasta ahora se ha presentado cómo calcular los totales de una población de acuerdo al diseño de muestreo seguido, pero en los estudios sanitarios se trabaja con muestreos complejos (Álvarez-Vaz, 2010),(Álvarez-Vaz, 2017). En general, se habla de diseños complejos cuando el diseño de muestreo no es mediante SI, sino que mediante varias etapas que suponen considerar estratos o conglomerados, tal como presentan en su trabajo Guillén et al. (Oct) y Cañizares Pérez et al. (Mar), y que no permiten ignorar el diseño para no tener estimaciones puntuales sesgadas. Es fundamental calcular correctamente el error de estimación a través de la varianza.

En general, es necesario en los diseños complejos estimar y medir errores de otros parámetros más complicados que los totales, como los promedios, ratios de totales, medianas, proporciones, cuantiles y coeficientes de regresión. Todas esas cantidades se pueden expresar como funciones de los totales poblacionales:

$$\theta = f(t_1, t_2, ..., t_q)$$

con $t_j = \sum_U y_{jk}$. Si se supone que θ es una función lineal de los totales $\theta = a_0 + \sum_{j=1}^{q} a_j t_j$, entonces va a ser muy sencillo poder estimar y calcular errores porque se va a tener $\hat{\theta} = f(\hat{t}_1, \hat{t}_2, ..., \hat{t}_q) = a_0 + \sum_{j=1}^{q} a_j \hat{t}_{j\pi}$ para lo que se va a poder medir el error a través de

$$V(\hat{\theta}) = V(\sum_{j=1}^{q} a_j \hat{t}_j) = \sum_{j=1}^{q} \sum_{j'=1}^{q} a_j a_{j'} C(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) \tag{3.28}$$

La clave, entonces, de poder plantear esto es que el cálculo de $V(\hat{\theta})$ se facilita, ya que se hace por aproximación mediante métodos numéricos. Este es uno de los motivos por los cuales no se puede efectuar estos cálculos con paquetes estadísticos convencionales (Álvarez-Vaz, 2017).

Un intervalo aproximado para un total t, al nivel de confianza (1-α) viene dado por

$$\hat{t}_0 \pm \phi_{1-\alpha/2} \left[\hat{V}\left(\hat{t}_\pi\right) \right]^{1/2} \tag{3.29}$$

con $\Phi\left(z_{1-\alpha/2}\right) = 1 - \alpha/2$

donde $\Phi(z)$ es la función de distribución de una variable $N(0,\ 1)$.

3.6.5. Cálculos de los tamaños de muestras

El error muestral que se comete en la estimación de un parámetro depende del tipo de diseño establecido para seleccionar los elementos que integran la muestra. Los diseños complejos suelen ser más ineficientes que si se trabaja con SI. A su vez, puede pasar que, si se logra estratificar el STSI, este puede ser más eficiente que el SI, pero no es una situación frecuente en los diseños en varias etapas. Por ese motivo, en general, a tamaños de muestra iguales, el error es mayor si se utiliza un diseño complejo en lugar de muestreo aleatorio simple SI, tal como se vio en la sección 3.6.2.

Con un diseño SI la determinación del tamaño depende del nivel de confianza aceptado, α; de la precisión deseada, ni del parámetro θ que se desea evaluar.

Para ese parámetro se desea poder dar un intervalo de confianza que verifique la siguiente relación:

$$\hat{\theta}(m) \pm \phi_{1-\alpha/2}\sqrt{\hat{V}[\hat{\theta}]} \tag{3.30}$$

donde $\theta(m)$ puede ser media, proporción, odds ratio, riesgo relativo, coeficiente de concordancia, valores de sensibilidad y especificidad en pruebas diagnósticas.

$$P(|\hat{\theta} - \theta| > \epsilon) < \alpha$$

es el IC que, con determinadas condiciones, puede suponerse con distribución asintótica normal y que puede reformularse como

$$P\left(|\hat{\theta} - \theta| > \Phi_{1-\alpha/2}\sqrt{var(\hat{\theta})}\right) = \alpha$$

Para eso, a partir del cálculo que se hace en el supuesto de SI se puede derivar cuál es el factor de inflación de la varianza a través del DEFF o trabajar con las fórmulas que permiten obtener el tamaño de muestra en función de la precisión bajo supuesto SI y corregirla por el efecto de diseño. Esto implica decidir un rango para el DEFF, que puede estar entre 1 y 3, según recomendaciones de la obra Adecuacy of sample size in health studies (Lwanga y Lemeshow, 1991).

3.6.6. Estudio en dominios

Sin importar cuál sea la estrategia de muestreo adoptada en función de la información disponible, del alcance del estudio y, por lo tanto, del diseño de muestreo resultante, el investigador en biomedicina necesita, con frecuencia, poder estimar, no solo para toda la población U, sino para alguna subpoblación, por ejemplo, en función de características de las personas (si fuese una encuesta a personas), como el sexo, tramo de edad, lugar de nacimiento.

Todas estas categorías para cada atributo medido forman lo que se denomina dominios, que parten a la población $U = 1, 2, ..., N$ en subpoblaciones o subconjuntos $U_1, U_2, ..., U_D$ y donde N_d es el tamaño de la población U_d.

Lamentablemente, esa necesidad de poder tener estimaciones en esos dominios no se plantea desde el inicio, lo que acarrea problemas luego, al analizar la información y generalizar resultados, ya que los tamaños muestrales obtenidos son pequeños y aleatorios. Este aspecto agrega una dificultad, ya que las precisiones, que son función del tamaño de muestra, cambian, lo que hace que no se pueda dar resultados desagregados por dominio en muchos casos, por no haber sido contemplado en el diseño ni en el cálculo del tamaño muestral (Álvarez-Vaz, 2017).

Por último, otro aspecto a tener en cuenta al trabajar con diseños complejos es que, en general, sin importar cuál haya sido el que se desarrolló, los pesos muestrales o expansores que se calculan a partir de las probabilidades de inclusión y son necesarios para la estimación puntual de parámetros y de los errores de estimación, opera un aspecto no deseado, pero que no se puede soslayar: existe, por lo general, un proceso de pérdida de información que se denomina no respuesta, que no es aleatorio. Las personas que no contestan no son una muestra aleatoria de la muestra originalmente propuesta y sorteada, por lo cual es necesario estudiarla e incorporarla al proceso de estimación a través de lo que se denomina proceso de calibración de pesos muestrales.

Este proceso supone modificar los pesos originalmente calculados, para que algunas variables que el investigador considera fundamentales, como el sexo, la edad, la región geográfica, el nivel socioeconómico, sean semejantes en la muestra finalmente recolectada a la de totales poblacionales de referencia. Para eso es necesario que el investigador decida cuál de toda esa información desea controlar para evitar sesgos. Luego deberá ajustarla, según disponga de ella en forma parcial o total.

Los procesos de calibración pueden ser esencialmente dos: Postestratificación completa o calibración sobre totales de celdas conocidos (PEC) y Postestratificación incompleta o calibración sobre las marginales conocidas (RAKE).

Para el caso de PEC la calibración puede ser usada en particular si se dispone de una tabla de frecuencia con tantas celdas como combinación de categorías de cada variable de control se tenga, donde se conozcan los totales poblacionales en cada una de ellas. Por ejemplo, si se desea controlar por edad y sexo, es necesario saber si la edad se estableció en 4 tramos, conocer los totales por sexo y tramo etario.

En cambio, la calibración por RAKE se hace, una vez determinada la variable por la cual se desea controlar, cuando:

- La marginal de la población es conocida, pero los N_{ij} de cada celda no se conocen. Los dos conjuntos de marginales pueden provenir de diferentes fuentes de datos, pero la clasificación cruzada falta. Por necesidad, hay que calibrar sobre las marginales conocidas.

- Hay algunas celdas vacías o tienen muy pocas observaciones, lo que haría que el proceso de ajuste sea inestable.

En libros como Särndal et al. (1992), Silva (2000) y Álvarez-Vaz (2017) se puede encontrar mucho más detalle de los dos procesos de ajuste que, a su vez, requieren programas informáticos especializados.

3.7. Ajustes de los indicadores mediante información auxiliar

Es importante tener en cuenta que, en la práctica, hay muchas situaciones en las que cualquiera de los indices presentados en las secciones anteriores debe ser ajustado porque se violan los supuestos en los que están basados, por ejemplo, al trabajar con muestras probabilísticas, no se verifica la independencia entre observaciones. Para eso, una forma de corregir esos aspectos es trabajar con Análisis Multinivel (Multiniv), tal como se presenta en la sección 3.7.1.

3.7.1. Modelos multinivel

El análisis multinivel (Multiniv) se desarrolló inicialmente para la investigación educativa (Goldstein, 1991). Cuando se analiza la performance de estudiantes, hay que tener en cuenta que las observaciones que corresponden a individuos de la misma clase o grupo no pueden ser consideradas como independientes. Los estudiantes que pertenecen a una misma clase pueden verse como partes de una jerarquía, de manera tal que los individuos en esta situación siguen una especie de clusterización. Si los individuos forman grupos o clusters, se podría esperar que dos de ellos, seleccionados de un mismo grupo, tiendan a ser más parecidos que dos individuos seleccionados de otros grupos. Por ejemplo, los niños aprenden en las clases las condiciones de su grupo, como características de los maestros y la capacidad de otros niños en la clase, lo que puede influir en el logro educativo de cada uno.

Esta situación se puede ver como una estructura de dos niveles de datos, donde el primer nivel son los estudiantes y el segundo nivel son las clases. Las clases que en este caso están en el nivel 2 de la jerarquía también forman parte de una estructura jerárquica más amplia donde las escuelas pueden estar clusterizadas entre sí (Gelman y Hill, 2006). Por lo tanto, para evaluar esas dependencias se recurre a los modelos multinivel (también conocidos como modelos jerárquicos lineales, modelos mixtos, modelos de efectos aleatorios y modelos de componentes de la varianza) para analizar los datos con una estructura jerárquica.

Al igual que en el caso de la escuela, esto se da en otras disciplinas como en la salud, cuando se lleva a cabo un estudio para investigar la relación entre el total de colesterol (CT) y la edad.

$$\text{CT} = \beta_0 + \beta_1 * \text{edad} + \beta_2 * \text{género} + \varepsilon \tag{3.31}$$

Los pacientes en estudio pueden provenir de diferentes médicos consultantes (es como habitualmente se hacen los estudios), por lo cual se puede pensar que existe una estructura jerárquica que responde al médico de quien provienen los pacientes del estudio. Eso se puede resolver incorporando variables dummies que tengan en cuenta el médico tratante. Si hubiese 12 médicos involucrados, serían necesarias 11 variables dummies en el modelo, lo que implicaría una pérdida de potencia muy grande. Más aún si se desease comparar el modelo para un valor dado de edad en las mujeres, por ejemplo, habría 11 modelos lineales univariados con diferentes interceptos, pero una misma pendiente. La idea que está detrás del Multiniv es que en lugar de considerar los interceptos por separado, se considere la varianza del intercepto (Hox, 1995).

Si, además, se puede suponer que las observaciones están clusterizadas entre los médicos, el modelo convencional que incorpora interacción entre edad y tipo de médico es el de la ecuación (3.32).

$$\begin{aligned} \text{CT} = \beta_0 + \beta_1 * \text{edad} + \beta_2 * \text{D1} \ + ...\beta_m * \text{Dm-1} + \\ ... + \beta_{m+1} * \text{D1*edad} + ...\beta_{2m-1} * \text{D11*edad} + \epsilon \end{aligned} \tag{3.32}$$

Si lo que interesa para el ejemplo planteado no es tener una pendiente diferente para cada médico, sino un efecto global de interacción, con el Multiniv se puede estimar la varianza de la pendiente en lugar de considerar una pendiente diferente para cada combinación de edad y médico (Twisk, 2006).

Antes de que el análisis multinivel fuera desarrollado, el problema de observaciones correlacionadas se abordaba de dos maneras: ignorando el hecho de que estaban correlacionadas o combinando las observaciones en un único valor. Por lo tanto, se debía calcular una especie de valor medio de las observaciones para cada grupo en primer término para luego utilizar estos promedios en un análisis de regresión estándar. Este método se conoce como método de agregación.

Los dos modelos de análisis multinivel con intercepto aleatorio o con intercepto y pendiente aleatoria se estiman mediante máxima verosimilitud y coinciden con los modelos de regresión que en la literatura de investigación se presentan con variedad de nombres tales como modelo de coeficientes aleatorios (de Leeuw y Kreft,1986; Longford, 1993), modelo de componentes de varianza (Longford, 1986) o, más recientemente como un caso particular de los modelos mixtos que se presentan en (McCulloch, 2001). En general, se puede ver el Multiniv como un caso de modelo mixto.

$$\mathcal{Y} = \underbrace{\mathcal{X}\beta}_{\text{Coeficientes fijos}} + \underbrace{\mathcal{U}\delta}_{\text{Coeficientes aleatorios}} + \epsilon \tag{3.33}$$

que en forma escalar se puede representar como lo plantean de Leeuw y Meijer (2008)

$$y_i = \sum_{q=1}^{q=r} x_{qi}\beta_q + \sum_{s=1}^{s=p} u_{si}\delta_s + \epsilon_i \tag{3.34}$$

Una de las ventajas fundamentales del análisis multinivel es que puede ser utilizado para el análisis de otros tipos de variables de respuesta también, como el análisis multinivel logístico (AML), el análisis multinivel de Poisson (AMP) para variables de conteo e incluso el análisis de supervivencia de varios niveles.

El Multiniv también puede ser usado para el modelado de datos longitudinales, donde los datos están correlacionados, ya que se trata, por ejemplo, de pacientes que se repiten y para los cuales se mide una variable tiempodependiente. Este tipo de datos tiene una forma jerárquica que responde a individuos o unidades que se miden en más de una ocasión, como sucede en los estudios de crecimiento humano. Aquí las ocasiones se agrupan dentro de los individuos que representan las unidades de nivel 2 con las medidas en diferentes momentos que representan el nivel 1. Por lo general, hay mucha más variación entre los individuos que en las ocasiones dentro de los individuos (Hox, 1995).

3.8. Software a utilizar

Se propone trabajar con el sistema R (R Core Team, 2017), que cumple con la ventaja de ser multiplataforma (es decir, que el mismo código puede ser usado con diferentes sistemas operativos), es software libre y está excelentemente bien documentado. Para presentarle el software al lector que pudiese no conocerlo, se opta por traducir al español la definición que está publicada en su sitio web:[3], traduciéndolo al español.

«R es un lenguaje y entorno para computación estadística y gráficos. Es un proyecto GNU que es similar al lenguaje y al entorno S que fue desarrollado en los Laboratorios Bell (anteriormente AT&T, ahora Lucent Technologies) por John Chambers y sus colegas. R puede considerarse como una implementación diferente de S. Hay algunas diferencias importantes, pero gran parte del código escrito para S se ejecuta sin alterar en R.

»R proporciona una amplia variedad de técnicas estadísticas (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupación, entre otras) y técnicas gráficas. Además, es altamente extensible. El lenguaje S suele ser el vehículo elegido para la investigación en metodología estadística y R proporciona una ruta de código abierto para participar en esa actividad.

»Una de las fortalezas de R es la facilidad con la que se puede producir gráficos de calidad de publicación bien diseñados e incluir símbolos matemáticos y fórmulas donde sea necesario. Se ha prestado gran atención a los valores predeterminados para las opciones de diseño menores en gráficos, pero el usuario mantiene el control total.

»R está disponible como software libre según los términos de la GNU No es Unix (GNU)[4] de Free Software Foundation (FSF)[5] en forma de código fuente. Compila y se

3. https://www.r-project.org/about.html
4. https://www.gnu.org/home.es.html
5. https://www.fsf.org/

ejecuta en una amplia variedad de plataformas UNIX y sistemas similares, entre ellos FreeBSD, Linux, Windows y MacOS.

»3.8.1. El entorno R

»R es un conjunto integrado de instalaciones de software para la manipulación de datos, el cálculo y la visualización gráfica. Incluye una instalación de manejo y almacenamiento de datos, un conjunto de operadores para cálculos sobre matrices, una colección grande, coherente e integrada de herramientas intermedias para el análisis de datos, facilidades gráficas para análisis de datos y visualización en pantalla o en papel y un lenguaje de programación bien desarrollado, simple y efectivo que incluye estructuras de control, funciones recursivas definidas por el usuario y facilidades de entrada y salida. El término entorno pretende caracterizarlo como un sistema totalmente planificado y coherente, en lugar de una acumulación incremental de herramientas muy específicas e inflexibles, como suele ocurrir con otros programas de análisis de datos.

»R, al igual que S, está diseñado en torno a un verdadero lenguaje informático y permite que los usuarios agreguen funcionalidad adicional mediante la definición de nuevas funciones. Gran parte del sistema está escrito en el dialecto R de S, lo que facilita a los usuarios seguir las elecciones algorítmicas realizadas. Para tareas intensivas en computación, el código C, C ++ y Fortran se pueden vincular y llamar en tiempo de ejecución. Los usuarios avanzados pueden escribir código C para manipular objetos R directamente.

»Muchos usuarios piensan en R como un sistema de estadísticas. Preferimos pensar en un entorno en el que se implementan técnicas estadísticas. R se puede extender (fácilmente) a través de paquetes. Hay aproximadamente ocho mil paquetes suministrados con la distribución R y muchos más están disponibles a través de la red de sitios de Internet CRAN que cubren una amplia gama de estadísticas modernas.

»R tiene su propio formato de documentación similar a LaTeX, que se utiliza para proporcionar documentación completa, tanto en línea como en papel.»

Actualmente, es la herramienta mas usada en el campo de la estadística aplicada a la epidemiología y un paradigma en la enseñanza e investigación de estadística en las universidades más importantes.

En cada una de las aplicaciones presentadas en la parte II se detallan las librerías (conjuntos de subrutinas) usadas en cada caso para la resolución de los problemas, tanto que se use el herramental estadístico presentado en las secciones 3.2, 3.3, y 3.4, como otras técnicas usadas en las aplicaciones y que, dada su complejidad, no se entra en detalle, salvo una presentación mínima para entender su aplicación en el contexto.

Parte II

Aplicaciones

Capítulo 4

Datos usados en las aplicaciones

En el desarrollo de los capítulos del libro se usan fuentes de datos primarias, dos de ellas encuestas de base poblacional. Los relevamientos utilizados se dieron en el marco de los estudios de posgrado de otras investigadoras y otros investigadores, donde participó el autor, de los cuales fueron surgiendo varios artículos en coautoría —y de ellos se hace referencia en este libro—. Al igual que en la tesis de doctorado (Álvarez-Vaz, 2020), se opta por adjudicarles una sigla y nombre de fantasía, aunque cada uno de los estudios tiene las referencias a las publicaciones que se hicieron, donde aparecen tal como se los identifica:

- Primer relevamiento nacional de salud bucal en población joven y adulta. Fue llevado adelante por la Facultad de Odontología y supervisado por el servicio de Epidemiología y Estadística, de la cual el autor de la obra forma parte desde 2007. Ver sección 4.1.

- Relevamiento y análisis de caries dental en adolescentes de 12 años de la ciudad de Montevideo, Uruguay. Esta encuesta se desarrolló entre agosto de 2011 y julio de 2012 para evaluar el estado de salud bucal de los escolares de 12 años de escuelas públicas y privadas (Relevamiento de Facultad de Odontología con financiación de la ANII). Ver sección 4.2.

- Relevamiento en población que se asistió en Facultad de Odontología entre 2015 y 2016. Forma parte del proyecto CSIC I+D llevado adelante entre junio de 2015 y junio de 2016. Ver sección 4.3.

4.1. Primer relevamiento nacional de salud bucal en población joven y adulta

Es la encuesta hecha en el marco del Primer relevamiento nacional de salud bucal (PRNSB2011), donde se relevaron 1485 personas, cuya metodología se publicó en (Lorenzo et al., 2013a), (Olmos et al., 2013), (Lorenzo et al., 2013b), (Ourens et al., 2013), (Casnati et al., 2013). Se basó en los criterios sugeridos por la Organización Mundial de la Salud (OMS) para estudios poblacionales (1997), que se adaptaron utilizando un diseño muestral complejo en 2 fases.

En la primera fase se trabajó con el conjunto de personas de los 3 tramos de edad que se consignan en la tabla 4.1, pertenecientes a localidades de 20.000 habitantes o más visitadas en la Encuesta Continua de Hogares (ECH) para 4 olas del 2010. La ECH es una encuesta nacional que considera nueve zonas de carácter geográfico y socioeconómico. Está basada en un diseño muestral estratificado por conglomerados polietápico. En la primera etapa, las unidades primarias de muestreo (UPM) son las secciones censales y, en la segunda etapa, las unidades secundarias de muestreo (USM) son los hogares. Esta encuesta se lleva a cabo cada dos meses. Las 4 olas de la encuesta hacen referencia a la cantidad mínima necesaria para lograr el tamaño de muestra calculado para cada grupo de edad. Tenemos un total de 4000 personas, de las cuales hay 1500 pertenecientes al grupo de edad de 15 a 24 y 2500 de otros grupos de edad; En la segunda fase, todas las personas pertenecientes al grupo de edad de 15 a 24 son seleccionadas entre las 4000 de la primera fase (alrededor de 1500). Los 2500 restantes que pertenecen a otros grupos de edad son seleccionados por muestreo aleatorio simple.

El tamaño de muestra se determinó a partir del mínimo necesario para estimar prevalencias con un error del 5 % y un margen de confianza confianza. Para eso, se utilizó como referencia la prevalencia de caries del relevamiento nacional de Brasil del año 2003, considerando las patologías más prevalentes en adultos de ambos países. Se establecieron 6 dominios de estimación que surgieron de cruzar los grupos de edad definidos y caracterizados por la OMS con 2 regiones (Montevideo y el interior del país).

El Instituto Nacional de Estadística (INE) estuvo a cargo del sorteo de la muestra y proporcionó los expansores asociados al diseño. Se relevaron las personas sorteadas de las siguientes 14 ciudades, Artigas, Canelones, Ciudad de la Costa, La Paz, Las Piedras, Colonia, Florida, Maldonado, San Carlos, Paysandú, Salto, San José, Rivera y Tacuarembó, correspondientes a 10 departamentos.

4.1.1. Calibración de la muestra para PRNSB2011

El cálculo para el tamaño de muestra tiene en cuenta los 6 dominios de estimación y considera para el tramo de 15 a 19 una prevalencia del 54 % para paradenciopatías y del 52 % en caries para los los restantes tramos etarios, sin diferenciar Montevideo y el interior. Tiene precisión del $\delta = 5\,\%$, confianza del 95 %, tasa de no respuesta del 20 % y DEFF de 1,5. Al usar la ecuación 4.1, resultan los tamaños que se consignan en la tabla 4.1

$$n = \left[\frac{(\phi_{1-\alpha/2})^2 * \pi * (1-\pi)}{(\delta)^2}\right] * \left[\frac{Deff}{(1-TNR)}\right] \tag{4.1}$$

Teniendo en cuenta la tasa de no respuesta se hizo un ajuste de calibración posterior mediante posestratificación. Las variables usadas para la calibración fueron el sexo y la edad. Esto se llevó a cabo con la librería survey (Lumley, 2009).

6 Dominios de estimación= (3) Edad y (2) Región

	Tramo etario	Montevideo	Interior	Total
	15-19	715	715	1430
	35-44	394	394	788
	65-74	394	394	788
	Total	1503	1503	3006

Tasa de respuesta por dominio

	Tramo etario	Montevideo	Interior
	15-19	78	58,5
	35-44	67	58,1
	65-74	77	69,8
	Total	74,8	61,3

Tabla 4.1: Tamaño de muestra alcanzado por dominio para encuesta PRNSB2011

.

4.2. Relevamiento y análisis de caries dental en adolescentes escolarizados de 12 años de la ciudad de Montevideo, Uruguay

El Relevamiento y análisis de caries dental en adolescentes (RACA2012) se desarrolló entre agosto de 2011 y julio de 2012. Su objetivo fue evaluar el estado de la salud bucal de los escolares de 12 años de edad de escuelas públicas y privadas.

Para calcular el tamaño de la muestra se utilizaron los siguientes parámetros: prevalencia de caries del 60 % (22), intervalo de confianza del 95 % (CI), nivel de precisión del 4 % y efecto diseño DEFF del 1.3, al que se añadió una tasa de no respuesta del 30 %. Por lo tanto, el tamaño de muestra necesario para este estudio fue de 1235 individuos. Se adoptó una muestra bietápica estratificada por conglomerados. Las unidades primarias de muestreo (UPM) son las escuelas públicas y privadas de Montevideo. Cuarenta y cuatro escuelas fueron seleccionadas al azar, 32 públicas y 12 privadas. Todos los niños de 12 años que asistían a estas escuelas fueron invitados a participar en el estudio, independientemente del año escolar en el que estuviesen.

Para la primera etapa se usaron las escuelas de 5.º y 6.º año del sector privado y público como marco muestral y se estratificaron en 3 estratos. Se sortearon 2 muestras independientes, una por estrato. Para el diseño usado las UPM, en este caso las escuelas, se seleccionaron con muestreo $\pi - ps$ (probabilidad proporcional al tamaño), es decir, la variable total de niños matriculados que figuraban en el marco muestral. Se usó la librería sampling. De esta manera, se tiene para las muestra sorteadas las UPM seleccionadas.

A partir de las π_{ik} probabilidades de inclusión se pueden calcular los expansores o pesos muestrales que son, en este caso:

$$\frac{1}{\pi_{ik}} = w_{i1} \tag{4.2}$$

es decir que cada escuela pesa por w_{ik}.
Los cálculos están hechos con base en la información disponible, que es el total de niños por escuela, de los cuales una parte debe ser descartada, ya que solo pueden ser incluidos los niños de 12 años. Esto hace que se tenga total de niños y total de niños elegibles. Luego, dependiendo de la cantidad de niños relevados en la segunda etapa, se tiene un tamaño de muestra aleatorio, por lo cual para el cálculo de los ponderadores es necesario considerar que cada niño tiene un ponderador que depende de la cantidad de niños elegibles:

$$w_{i2} = \frac{1}{N_{j\ ik}} \tag{4.3}$$

donde N_j es el total de niños elegibles en la escuela j-ésima.

Para tener los pesos muestrales de ambas etapas es necesario combinarlos de manera multiplicativa:

$$w_{i12} = w_{i1} * w_{i2} \tag{4.4}$$

Los w_{i1} fueron calculados con base en el total y no el total de elegibles, por lo que para cada escuela se debe considerar un factor extra, que es:

$$f_{j1} = \frac{\text{Total}_j}{\text{Total elegible}_j} \tag{4.5}$$

Con la información disponible se puede usar una f_{j1} variable o usar un factor de corrección fijo que se puede estimar como estimador de razón, lo que se hizo y dio un valor de $1,45$. Esto debe interpretarse como que hay en promedio un $45\,\%$ más de niñas que de niños elegibles.

4.2.1. Calibración de la muestra

Para la calibración de la muestra del estudio RACA2012 se deben de tener en cuenta dos aspectos: La no respuesta y el desbalanceo de alguna variable que se desee controlar. En este caso, para la muestra finalmente relevada con un $35\,\%$ de no respuesta promedio y desbalanceo con sobrerrepresentación de mujeres, fue necesaria una calibración, que se hizo usando la librería survey (Lumley, 2009).

4.3. Relevamiento en población que se asiste en Facultad de Odontología

Los datos del Relevamiento en población que se asiste en Facultad de Odontología (RPAFO2015) provienen de un estudio sobre personas que demandan atención en la Facultad de Odontología y que son evaluadas por los odontólogos del Servicio de Registros de la Facultad, desarrollado en el marco de un proyecto I+D de la CSIC en 2014. Se aplica una muestra de 602 personas que consultaron en el período entre mayo de 2015 y junio de 2016, que se seleccionaron mediante muestreo sistemático. Se les aplicó un cuestionario sociodemográfico y un examen completo de la boca para evaluar el estado de las piezas dentales y de la mucosa, además de medidas antropométricas, de PA y de glicemia. El tamaño muestral se determinó para poder medir prevalencias de hasta 25 % con un margen de error $\delta = 0{,}05$ y un nivel de confianza $1 - \alpha = 0{,}95$ y cubrir hasta una tasa de no respuesta del 90 %. Por último, de los 640 calculados en principio se obtuvieron 602, que representa una fracción de muestreo de alrededor del 15 % del total de personas que consultan por año.

A partir de estas tres fuentes de datos se presentan en esta obra solo algunas aplicaciones que consideran algunas de las patologías bucales presentadas en el capítulo 2.

Capítulo 5

Aplicación de modelos de regresión binaria y de modelos de conteo básicos al estudio de la enfermedad caries en una encuesta poblacional

5.1. Introducción

En este capítulo se presenta una situación muy frecuente en el ámbito de la epidemiología, como lo es el trabajo con variables de respuesta en escala cuantitativa que se dicotomizan para un determinado umbral y el uso de modelos predictivos, pero con la pérdida del gradiente de la enfermedad.

Se muestra, entonces, cómo funcionan los modelos en la escala original, que para el caso de algunos de los componentes del CPO, por ser variables discretas pensadas para evaluar conteos, deben modelarse con distribuciones de probabilidad adecuadas, como la Poi o la BN, que son las más usadas en la aplicaciones biomédicas frente a este tipo de problema. Sin embargo, esas dos distribuciones no siempre son las adecuadas y existen varias alternativas que se presentan en el capítulo 6 y se desarrolla su utilización, sobre todo cuando los modelos de conteo son patológicos al presentar sobredispersión o excesos de 0.

Se presentan y comparan los resultados usando la estrategia de trabajar dicomotizando el conteo de caries (C) a través de la RLog vs una regresión que conserve la escala original de medida preservando el gradiente de la enfermedad, para lo cual se deben usar las diferentes distribuciones de probabilidad para modelar conteos.

5.2. Medición del componente C del CPO

Cuando se quiere hacer modelos predictivos para cualquiera de los componentes del CPO es necesario decidir cuál es la herramienta más adecuada para la naturaleza de los datos que se consideran en el CPO. Hay que tener en cuenta que para cualquiera de sus componentes se puede trabajar con variables binarias que dan cuenta de la presencia o ausencia de cualquiera de los componentes o trabajar con variables aleatorias de conteo que ponen de manifiesto el gradiente del componente. Para eso, es necesario evaluar diferentes modelos predictivos que en forma muy general se presentaron en la sección 3.2.1 y que se desarrollan a continuación.

5.3. Modelos de regresión

Ya en el capítulo 3 se propone el uso de modelos parsimoniosos pero adecuados, en el sentido de que sean fáciles de estimar, sencillos de usar, que la información esté disponible y que los especialistas en ciencias médicas los entiendan y adopten.

Cuando la variable de respuesta es categórica (Silva, 1995), (Agresti, 2005), se puede recurrir a la RLog, en la que el investigador biomédico logra trabajar con un modelo predictivo paramétrico. En este proyecto se consideran algunos modelos que están poco difundidos en el campo de la biomedicina a pesar de ser muy necesarios. Por ello, se le presta especial atención a los modelos de conteo.

5.3.1. Método de regresión logística

Cuando la variable de respuesta Y_i es una variable aleatoria Bernoulli, los resultados posibles son éxito y fracaso. Se codifican como $\{0, 1\}$, distribución de probabilidad: $P(Y_i = 1) = \pi_i$, $P(Y_i = 0) = 1 - \pi_i$ y valor esperado, se maneja el modelo ya presentado en la sección 3.2.2 del capítulo 3.

$$P(Y = 1|X) = \pi = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{5.1}$$

$$P(Y_i = 1|X_i) = \pi_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_p X_{pi}}}{1 + e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_p X_{pi}}} \tag{5.2}$$

5.3.2. Modelos de conteo básicos

Los modelos de conteo que ya se presentaron en la sección 3.2.3 pueden adoptar diferentes formas, que se basan en la teoría de los MLG detallada en la sección A.3 del apéndice A.

Con esas restricciones de ser variables con recorrido discreto, existen varias alternativas. Las usadas con más frecuencia son la Poi y la BN, pero que en muchas circunstancias no son las más adecuadas, para lo cual existen varias alternativas, algunas muy poco usadas, que permiten mejorar el ajuste cuando algunos supuestos básicos que están implícitos en la POI no se cumplen, como que la media sea igual a la varianza, lo que se conoce como equidispersión.

Modelo de Poisson

La distribución más sencilla para modelar datos de conteo es la Poi, que tiene función de masa de probabilidad:

$$f(y; \mu) \quad = \quad \frac{\exp(-\mu) \cdot \mu^y}{y!} \tag{5.3}$$

Es del tipo (A.4) donde la regresión de Poisson puede ser vista como un caso particular de la teoría de MLG. En este caso la función de "enlace" es $g(\mu) = \log(\mu)$ lo que produce una relación log-lineal entre la media y el predictor lineal. La varianza en el modelo de Poisson se supone idéntica a la media, con lo que se fija el parámetro de dispersión como

$\phi = 1$ y la función de varianza como $V(\mu) = \mu$. A menudo, en la práctica no se verifica, por lo que se vuelve necesario trabajar con los modelos que siguen.

Modelo cuasi-Poisson

Otra forma de tratar con el exceso de dispersión es el uso de la función de regresión media y la función de la varianza en un modelo Poisson pero estimando el parámetro de dispersión ϕ sin restricciones. De esta manera no se fija a ϕ, sino que se estima a partir de los datos. Esta estrategia produce las mismas estimaciones para los coeficientes del modelo estándar de Poisson, pero la inferencia se adecúa al problema de sobredispersión en el tratamiento de la varianza (Zeileis, 2004, 2006).

Modelo binomial negativo

Otra forma de trabajar con la sobredispersión considerando un modelo de conteo que asume distribución binomial negativa, $y_i|x_i$ que surge como un proceso Γ de mezcla de distribuciones de Poisson. Para eso, se parametriza su función de masa de probabilidad como

$$f(y;\mu,\theta) \quad = \int_0^\infty \frac{\exp(-\mu)\cdot\mu^y}{y!} f_\Gamma(\mu)d\mu = \quad \frac{\Gamma(y+\theta)}{\Gamma(\theta)\cdot y!}\cdot\frac{\mu^y\cdot\theta^\theta}{(\mu+\theta)^{y+\theta}} \tag{5.4}$$

donde están los parámetros de media μ y forma θ; $\Gamma(\cdot)$ es la función Gamma. Para cada parámetro fijo θ, este es del tipo de familia exponencial que aparece en la ecuación (A.4). Tiene parámetro de disersión $\phi = 1$, pero con función de varianza $V(\mu) = \mu + \frac{\mu^2}{\theta}$ (Cameron y Trivedi, 1990), (Hilbe, 2011).

Luego de haber presentado los detalles básicos de los MCont y las diferentes alternativas de distribuciones, el objetivo es comparar lo que sucede al considerar para el componente C de caries un modelo de respuesta binaria, presencia o ausencia de caries, con respecto a un modelo de conteo.

5.4. Aplicación: primer relevamiento nacional de salud bucal en población joven y adulta uruguaya (2011)

Tal como se aclaró en el capítulo 4, el PRNSB2011 se caracteriza por tener un diseño muestral en dos fases, donde a partir de la importante Tasa de no respuesta (TNR), se debió hacer un proceso de calibrado mediante PEC, usando como variables de control el sexo y la edad. Este procedimiento asegura que no haya desbalance para esas dos variables, pero genera un juego de pesos muestrales calibrados que difieren de los originales, por lo cual se estudia la relación entre ambos juegos de expansores, tal como se ve en las figuras 5.1 y 5.2.

De esas figuras se desprende que la deformación que sufren estos pesos dependen del sexo. Los hombres son los que más aumentan sus valores

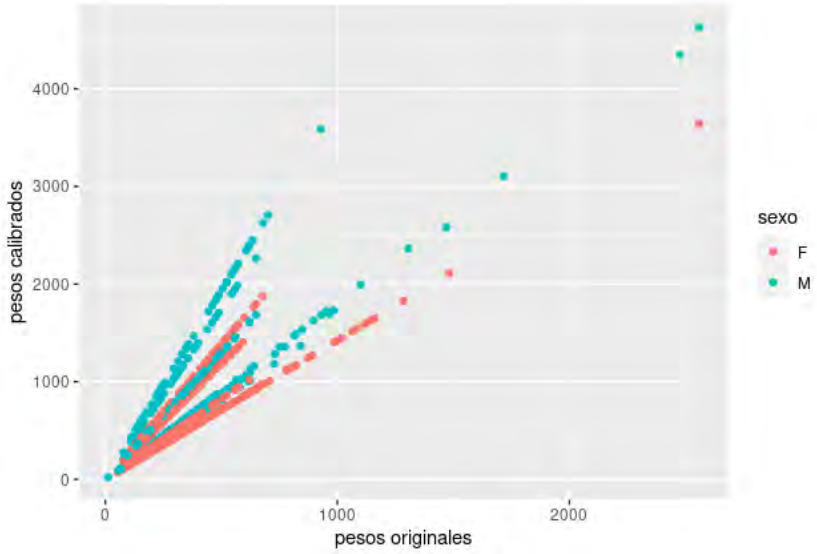

Figura 5.1: Comparación por sexo de los pesos muestrales originales vs. pesos calibrados

Para la edad se puede ver que el incremento de los expansores se da en los tres tramos de edad. A su vez, en ambas figuras se ve que hay observaciones con pesos, luego del proceso de calibrado, que pueden considerarse extremas. Como tendrían un impacto muy severo, se decide truncar, fijando como límite inferior el percentil 10 del peso calibrado y como límite superior el percentil 90 del peso calibrado.

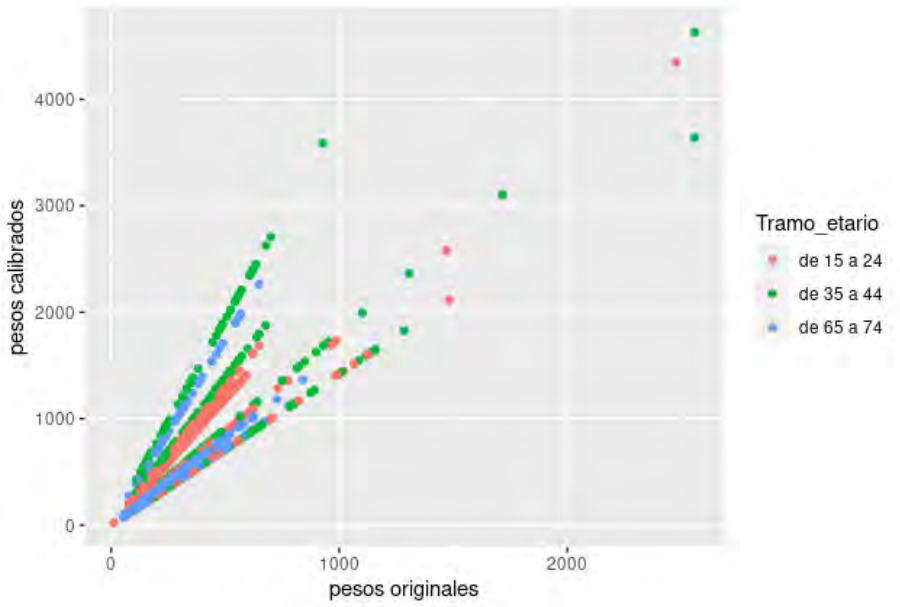

Figura 5.2: Comparación por edad de los pesos muestrales originales vs. pesos calibrados

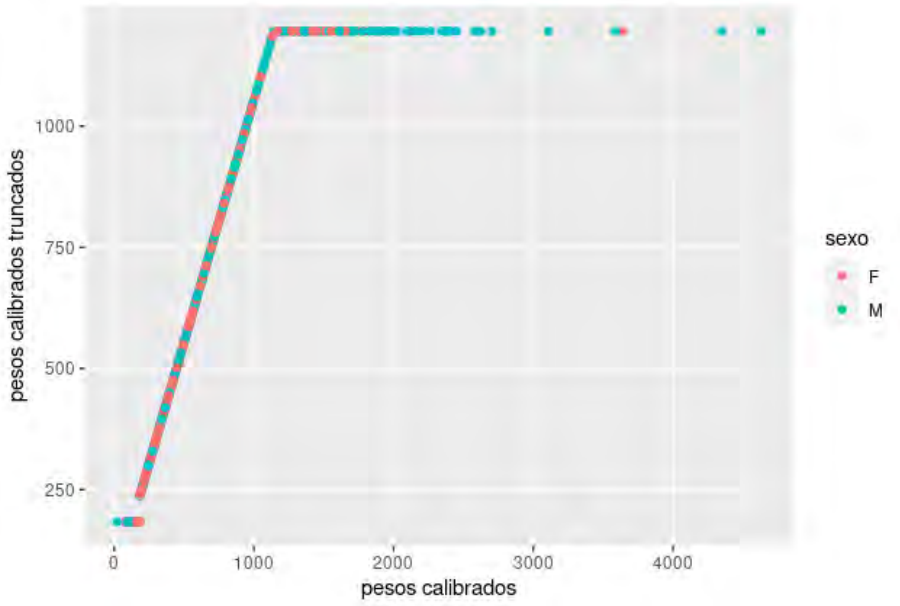

Figura 5.3: Comparación de los pesos muestrales calibrados vs. pesos calibrados truncados

A continuación, se presentan las estimaciones puntuales del componente C, que es el que se trabajará a lo largo del capítulo

Medias del componente C				
Dominio	$\bar{x}$ para C	se	DEff para C	Intervalo de confianza
Sexo				
F	1,47	0,09	1,20	(1,30;1,64)
M	1,52	0,11	1,14	(1,60;1,74)
Tramo etario				
de 15 a 24	1,62	0,09	1,11	(1,42;1,81)
de 35 a 44	1,91	0,16	1,12	(1,59;2,23)
de 65 a 74	0,71	0,08	1,34	(0,54;0,86)
Global				
Global	1,5	0,07	1,19	(1,36;1,63)

Tabla 5.1: Medias de componente C

Descripción de variables explicativas		
Variable	Nombre	Descripción
V(1)	CPO	(nivel de CPO)
V(2)	edad	edad en tramos (3 niveles)
V(3)	sexo	sexo (2 niveles)
V(4)	estrato	estrato sociodegráfico (3 niveles de Montevideo y 1 del interior)
V(5)	alcohol	nivel de consumo de alcohol (2 niveles)
V(6)	tabaco	tabaco (2 niveles)
V(7)	INSE	Índice de Nivel Socioeconómico (3 niveles)

Tabla 5.2: Conjunto de variables regresoras usadas para los modelos en PRNSB2011

5.5. Estimación del componente C como variable binaria

Para poder trabajar el componente C como variable binaria que está expresando la prevalencia de caries (dejando de lado la extensión), se tienen los siguientes resultados.

En toda la sección se trabaja con el mismo conjunto de variables explicativas que aparecen en la tabla 5.4

Para evaluar la bondad de ajuste del modelo de la tabla 5.4, es necesario ver la predicción a través de la curva ROC. Sin embargo, debe tenerse en cuenta que los datos

Prevalencias del Componente C				
Dominio	$\hat{p}$ para C	se	DEff para C	Intervalo de confianza
Sexo				
F	48,8	0,09	1,20	(45,0;52,6)
M	49,1	0,11	1,14	(44,9;53,4)
Tramo etario				
de 15 a 24	52,2	0,01	1,23	(48,4;56,6)
de 35 a 44	57,7	0,03	1,21	(52,0;63,4)
de 65 a 74	29,9	0,02	1,33	(24,7;35,0)
Global	51	0,01	1,22	(46,1;56,8)

Tabla 5.3: Prevalencia de caries

Variables	Coeficientes	E.E.	t	Pr($>\|t\|$)
(Intercepto)	0,233	0,182	1,28	0,201
tramo etario (35 a 44)	-0,119	0,177	-0,67	0,500
tramo etario (65 a 74)	-1,569	0,280	-5,59	0,000
estrato (2 (M))	-0,204	0,216	-0,94	0,345
estrato (3 (M))	-0,629	0,259	-2,43	0,015
estrato (4 (Int))	0,246	0,172	1,43	0,152
tabaco (Si)	0,521	0,145	3,58	0,000
CPO	0,032	0,010	3,11	0,002
INSE (MEDIO)	-0,665	0,132	-5,01	0,000
INSE (ALTO)	-1,096	0,363	-3,01	0,002

Tabla 5.4: Modelo de regresión logística para la prevalencia de Caries

provienen de una encuesta con diseño muestral complejo, por lo que, para elaborar la Se y Esp, que surge de ir variando los puntos de corte de la predicción del modelo, deberían surgir de tablas de doble entrada entre los valores observados y los pronosticados, creadas considerando la expansión de los datos.

Para eso se presenta una curva creada como si no existiese este problema, ya que la mayoría de los paquetes estadísticos no lo consideran, el sistema R que tiene varias librerías para elaborar una curva ROC, tampoco lo resuelve.

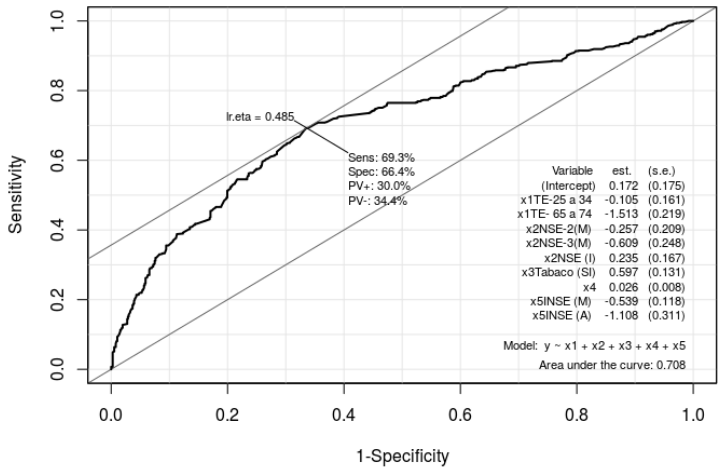

Figura 5.4: Curva roc para modelo estimado en tabla 5.4, sin considerar pesos muestrales

Para superar el problema planteado antes, se estima una curva que si toma en cuenta los pesos muestrales, usando el siguiente procedimiento:

Se considera el vector de valores pronosticados (en términos de probabilidad), el cual se deciliza y establece puntos de corte, creando una variable categórica sobre la que se totalizan los pesos muestrales.

Construcción de curva ROC usando diseño muestral		
Deciles de probabilidad	NO	SI
(0,066,0,257]	66211	25945
(0,257,0,327]	57462	22526
(0,327,0,394]	66639	25720
(0,394,0,454]	60874	39308
(0,454,0,486]	37456	28710
(0,486,0,540]	42301	32631
(0,540,0,599]	36957	64809
(0,599,0,648]	35964	50631
(0,648,0,724]	23362	67281
(0,724,0,872]	13457	68054

Tabla 5.5: Tabla de puntos de corte para la probabilidad, considerando los pesos muestrales

Sobre la tabla 5.5, se aplica luego una ROC que permite ver la performance global a través del área bajo la curva (AUC), que es un indicador de bondad de la capacidad predictiva.

Se puede observar que las 2 curvas difieren, mostrando más cantidad de puntos para la figura 5.4, ya que se hace en cada uno de los 1468 encuestados del modelo estimado sin expandir, mientras que la curva de la figura 5.5 muestra apenas 10 puntos, que son los que surgen de la tabla 5.5. Sin embargo, los valores de AUC son similares, pero, tal como se plantea en la sección 5.7, es necesario ver los sesgos que surgen al obviar los pesos muestrales.

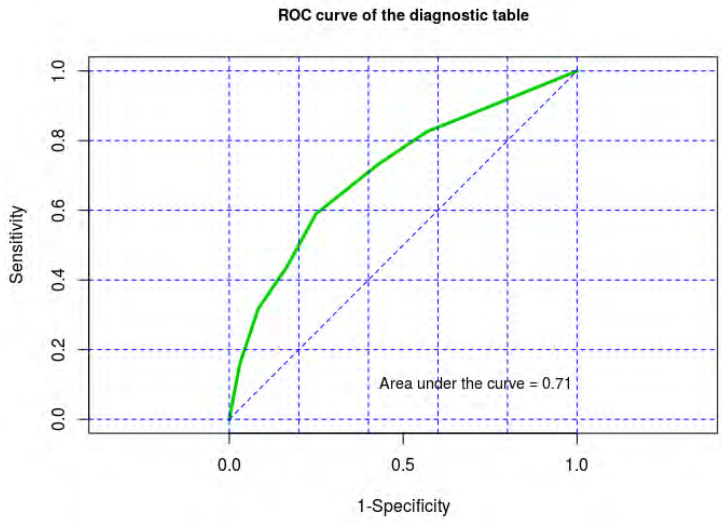

Figura 5.5: Curva Roc para modelo estimado en tabla 5.4, considerando pesos muestrales

5.6. Estimación del componente C como variable de conteo

Antes de pasar a modelar el conteo de C, para el juego de variables explicativas presentadas en la sección anterior, parece prudente ver cuál es el comportamiento de C, para lo cual en la tabla 5.6 se consigna la diferencia entre la distribución usando la expansión, que surge de considerar los datos crudos, que también se ve en las siguientes figuras.

Observando la figura 5.6 de la página 99 y la tabla 5.6 de la página 99, puede apreciarse que la diferencia es mínima y, en realidad, el impacto de no considerar el diseño muestral se verá en la etapa de modelización.

En las figuras 5.7 y 5.8 pueden verse algunas diferencias entre el comportamiento por sexo, edad y por estrato económico.

Conteos	% con los datos expandidos	% con los datos sin expandir
0	50,97	52,00
1	18,09	17,69
2	9,73	8,75
3	6,67	6,51
4	5,99	5,97
5	2,25	2,37
6	2,18	2,17
7	0,96	1,02
8	0,58	0,75
9	0,45	0,47
10	0,99	0,81
11	0,14	0,27
12	0,20	0,27
13	0,33	0,41
15	0,07	0,07
16	0,08	0,14
17	0,09	0,14
18	0,24	0,20

Tabla 5.6: Frecuencias relativas de datos expandidos y sin expandir para componente C de caries

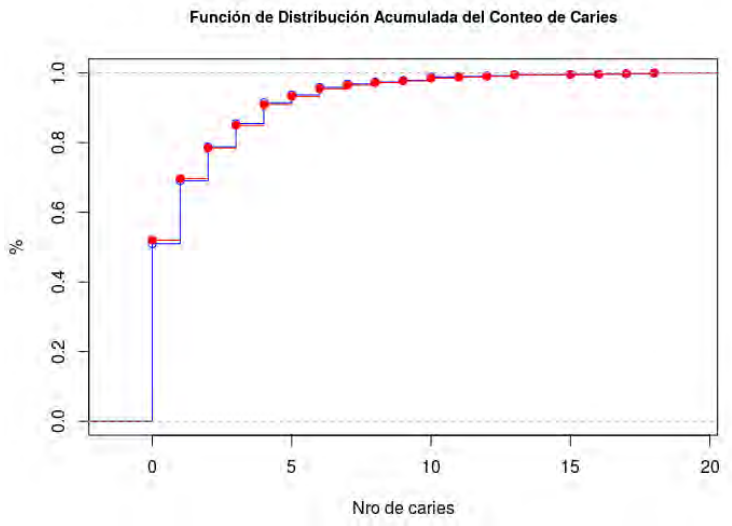

Figura 5.6: Curva de Distribución acumulada del conteo de caries, usando pesos expandidos y pesos crudos

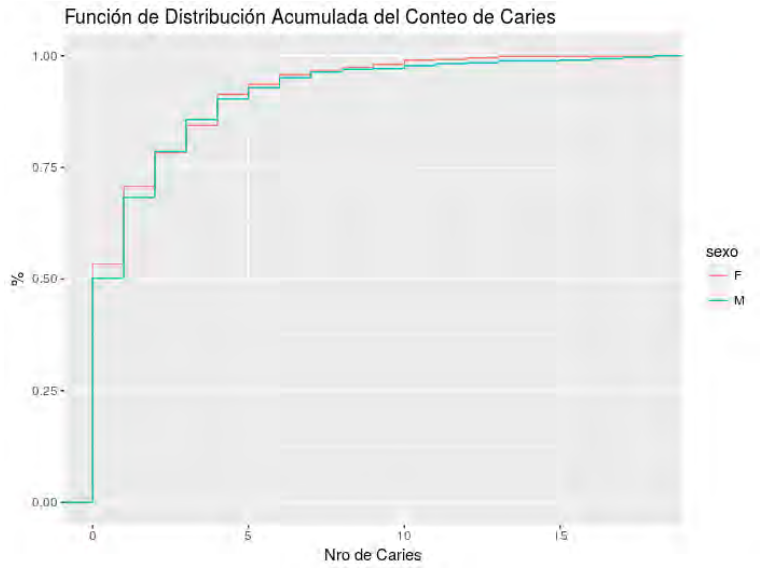

Figura 5.7: Curva de distribución acumulada del conteo de caries por sexo usando pesos sin expandir

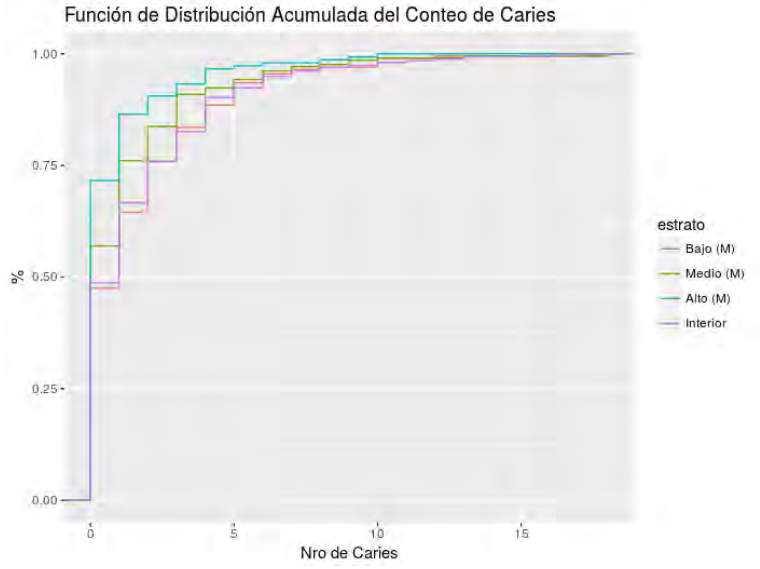

Figura 5.8: Curva de distribución acumulada del conteo de caries por tramo etario usando pesos sin expandir

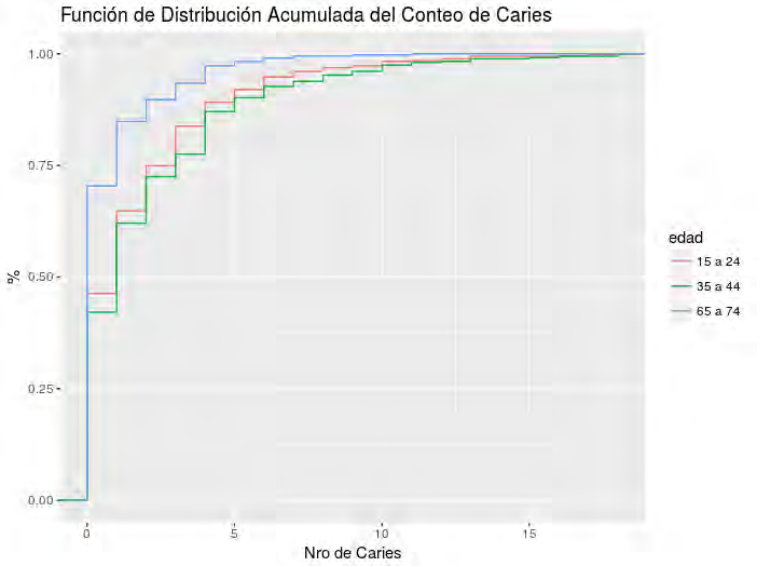

Figura 5.9: Curva de distribución acumulada del conteo de caries por estrato socioeconómico usando pesos sin expandir

El motivo de usar funciones de distribución acumuladas es que permiten ver de mejor modo la variación entre los valores de la variable de conteo C, que no interesa agruparla y visualizarla en un histograma, y también mejora el constraste visual al estratificarla por algún atributo. Debe recordarse que las figuras están hechas sin expandir los datos.

Si ahora se estima el conteo de caries a través de un modelo de regresión de Poisson, son significativas todas excepto el sexo.

Componente C				
Variables	Coeficientes	E.E.	t	Pr(> |t|)
(Intercepto)	0,391	0,122	3,20	0,001
tramo etario (35 a 44)	-0,513	0,147	-3,48	0,0005
tramo etario (65 a 74)	-1,850	0,225	-8,19	0,0000
estrato (2 (M))	-0,051	0,166	-0,31	0,755
estrato (3 (M))	-0,355	0,204	-1,74	0,081
estrato (4 (Int))	0,265	0,104	2,54	0,011
tabaco (Si)	0,429	0,088	4,88	0,0000
alcohol (Si)	-0,189	0,096	-1,97	0,049
CPO	0,052	0,007	7,09	0,0000
INSE (MEDIO)	-0,464	0,088	-5,26	0,0000
INSE (ALTO)	-0,974	0,295	-3,30	0,0010

Tabla 5.7: Modelo de regresión Poisson para el conteo de caries

Resta ver la calidad del ajuste de este modelo, teniendo en cuenta dos aspectos que ya a priori podrían hacer pensar que la distribución no es adecuada. Se tiene en cuenta el exceso de ceros y la sobredispersión.

	Media	Varianza	N.º de O
Componente C	1,50	6,14	50 %

Tabla 5.8: Evaluación de la sobredispersión y del exceso de 0

De la tabla 5.8 surge que la varianza es mucho mayor que la esperada para variable con distribución de Poisson con $\lambda = 1,5$. Por otra parte, la masa de probabilidad en 0 es del 50 %, mientras que ese valor no puede superar el 22 %.

Con esos dos motivos, es claro que el modelo planteado en la tabla 5.7 no es adecuado. A pesar de eso, si se comparan los coeficientes para ambos modelos se puede ver el impacto de trabajar con el modelo de tipo binario y el de tipo de conteo.

A continuación se presenta el resultado de un modelo de Poisson mal aplicado, cuya justificación se presenta en la sección 5.7. El error consiste en estimar un modelo Poisson sobre la variable que es de tipo binaria (es decir el conteo de C desaparece y se transforma en 0 y 1).

5.7. Discusión

En la sección anterior se modeló el componente C del CPO, es decir, la patología caries, usando dos tipos de modelos: uno que da cuenta de la prevalencia de C a través de un modelo de RLog, que implica una transformación de la variable original, siendo un modelo

	Conteo	Observado	Pronosticado	Residuo
1	0	51,91	31,25	20,66
2	1	17,64	29,02	-11,37
3	2	8,79	18,50	-9,72
4	3	6,54	10,14	-3,61
5	4	5,99	5,25	0,75
6	5	2,38	2,70	-0,31
7	6	2,18	1,42	0,76
8	7	1,02	0,77	0,26
9	8	0,75	0,43	0,32
10	9	0,48	0,24	0,24
11	10	0,82	0,13	0,68
12	11	0,27	0,08	0,20
13	12	0,27	0,04	0,23
14	13	0,41	0,02	0,39
15	14	0,00	0,01	-0,01
16	15	0,07	0,01	0,06
17	16	0,14	0,00	0,13
18	17	0,14	0,00	0,14
19	18	0,20	0,00	0,20

Variables	Modelo binario		Modelo de conteo	
	Coeficientes	Pr(> |t|)	Coeficientes	Pr(> |t|)
(Intercepto)	0,233	0,201	0,391	0,001
tramo etario (35 a 44)	-0,119	0,500	-0,513	0,000
tramo etario (65 a 74)	-1,569	0,0000	-1,85	0,0000
estrato (2 (M))	-0,204	0,345	-0,051	0,755
estrato (3 (M))	-0,629	0,015	-0,355	0,081
estrato (4 (Int))	0,246	0,152	0,265	0,011
tabaco (Si)	0,521	0,000	0,429	0,0000
alcohol (Si)	-	-	0,429	0,0000
CPO	0,0323	0,010	0,052	0,000
INSE (MEDIO)	-0,665	0,000	-0,464	0,0000
INSE (ALTO)	-1,096	0,002	-0,974	0,001

Tabla 5.9: Comparación de modelo de RLog y modelos RPoi

Variables	Modelo de conteo		Modelo de conteo incorrecto	
	Coeficientes	Pr(> |t|)	Coeficientes	Pr(> |t|)
(Intercepto)	0,391	0,001	-0,591	0,0001
tramo etario (35 a 44)	-0,513	0,0005	-0,072	0,363
tramo etario (65 a 74)	-1,850	0,0000	-0,793	0,0001
estrato (2 (M))	-0,051	0,755	-0,107	0,0000
estrato (3 (M))	-0,355	0,081	-0,392	0,011
estrato (4 (Int))	0,265	0,011	0,102	0,148
tabaco (Si)	0,429	0,0000	0,213	0,0000
alcohol (Si)	-0,189	0,049	-0,0006	0,001
CPO	0,052	0,0000	0,013	0,001
INSE (MEDIO)	-0,464	0,0000	-0,300	0,000
INSE (ALTO)	-0,974	0,0010	-0,536	0,016

Tabla 5.10: Comparación de modelo de RPoi y modelo de RPoi mal estimado

más sencillo y difundido pero que tiene la desventaja de no poder ver como el gradiente de enfermedad se asocia con las variables explicativas. Por otra lado el considerar la variable original que es de conteo usando la regresión de Poisson presenta un modelo un poco más complejo pero más adecuado. Interesa, por lo tanto, saber cuál es la desventaja al usar un modelo más sencillo. Para eso, en la tabla 5.9 se comparan los coeficientes de ambos modelos, en los que se puede advertir que, por un lado, tienen signos iguales para los coeficientes de las variables explicativas, pero en valor absoluto los coeficientes difieren y para entender lo que se pierde de información se propone aplicar ambos modelos a una persona con un perfil dado y consignar la probabilidad: P(de 35 a 44, Estrato 4, Fuma, C=8, CPO=32, BAJO)=0,843, y si se le aplica el modelo de conteo da un valor medio de caries de 9,35, cercano al verdadero valor observado, que es de 8.

Si se cambia de perfil tenemos que P(de 65 a 74, Estrato 2, NO fuma, C=0, CPO=21, MEDIO)=0,066 y aplicándole el modelo de conteo se tendría 1 como media.

Con ambos ejemplos y a pesar de que el modelo de conteo tiene problemas de ajuste, la gran ventaja sobre el binario es que cuando la probabilidad es alta (que se considera que va a tener caries), no se sabe si va a tener 1, 2 o más caries.

Mención aparte merece el mal uso de un modelo que trabaja como si fuera de conteo, pero un recorrido truncado, ya que tiene solo dos valores 0 (que significa que no tiene caries) y 1 (que tiene cierta cantidad de caries). Ese aspecto del método de estimación no lo toma en cuenta, si no se le advierte, lo cual lleva a estimar un modelo que no es correcto. El motivo de este proceder es porque algunos investigadores en biomedicina, haciendo abuso del método, prefieren reportar razones de prevalencias (conocidas como IRR), que es lo surge de los modelos de conteo al hacer la exponenciación de los coeficientes para su

interpretación, justificando este proceder, porque la razón de prevalencias puede parecer más sencilla de interpretar que los odds ratios que surgen de la RLog.

Eso se puede ver si a un mismo perfil de encuestado se le aplica el modelo de conteo bien estimado y se compara con los resultados del modelo de Poisson mal estimado, como está consignado en la tabla 5.10. Si se comparan en términos de IRR 2 categorías consecutivas para el modelo de conteo correcto, por ejemplo el tramo etario, se tiene para 2 categorías la siguiente situación:

Modelo	35 a 44	65 a 74	IRR
Correcto	0,601	0,145	3,9
Incorrecto	0,929	0,452	2,05

Tabla 5.11: Comparación del IRR sobre modelo correcto e incorrecto

que estaría mostrando que el impacto en el riesgo de 2 categorías adyacentes es la mitad al usar el modelo incorrecto que si se usa el correctamente tratado.

Antes de finalizar y pasar a las conclusiones, es importante retomar un aspecto mencionado antes para la evaluación del ajuste del modelo de RLog, presentado en la sección 5.5, donde se advertía de la inconveniencia de usar la curva ROC sin considerar los pesos muestrales asociados al diseño muestral. Si bien se presentan las curvas construidas de ambos modos, la que debe ser tenida en cuenta es la que aparece en la figura 5.5, en color verde. Algunos autores como (Yao et al., 2015) en trabajos recientes manifiestan que «La curva (ROC) se puede utilizar para evaluar el rendimiento de las pruebas de diagnóstico. El área bajo la curva ROC (AUC) es un índice de resumen ampliamente utilizado para comparar múltiples curvas ROC. Se han desarrollado métodos paramétricos y no paramétricos para estimar y comparar las AUC. Sin embargo, estos métodos generalmente solo son aplicables a los datos recopilados a partir de muestras aleatorias simples y no a estudios y estudios epidemiológicos que utilizan diseños de muestra complejos, como el muestreo estratificado y / o de diseños con múltiples etapas de muestreo y pesos muestrales para la ponderación de la muestra. Estas muestras complejas pueden inflar las variaciones de la correlación intra-clúster y alterar las propiedades de los estadísticos de prueba». (Yao et al., 2015) en su artículo Estimation of ROC Curve with Complex Survey Data muestra mediante simulación Monte Carlo que la varianza de la AUC se calcula de forma diferente de la habitual. También expone que comparar 3 modelos de RLog a través de la AUC puede llevar a resultados contradictorios si se opta por un modelo con respecto a otro, cuando realmente no existen diferencias.

5.8. Conclusiones

A modo de conclusión, luego de haber presentado 2 formas de modelar la patología caries en una encuesta poblacional, si bien la hipótesis de trabajo era que el modelo de conteo podía ser más informativo al poder evaluar el impacto de las variables explicativas en el gradiente de enfermedad, se puede decir que la distribución de la variable de conteo es patológica, desde el punto de vista estadístico, al presentar sobredispersión y exceso de 0. Para ese tipo de anomalía se estudia en detalle y se presentan alternativas para poder trabajar en el capítulo 6. Lamentablemente, dado que los datos provienen de un diseño muestral complejo, en este capítulo no se pudo ensayar como alternativa la regresión vía MBN, que podría, en parte, mitigar la sobredispersión, ya que la librería survey (Lumley, 2009) no permite trabajar con esa distribución.

Parece, por lo tanto, más honesto trabajar con la RLog, que no sería tan informativa y que muestra tener una performance global del 71 % cuando se elabora usando los pesos muestrales. También tiene valores similares a los que se usan sin tener en cuenta este aspecto, que aparece reseñado en la figura 5.4, donde algunos indicadores, como el punto de corte óptimo $lr.\eta$, que estaría indicando la Se, Esp y el Valor predictivo positivo (VP+) y Valor predictivo negativo (VP-), deberían ser tomados con precaución.

Las variables relevantes que parecen modular la propensión a tener caries son el tramo etario (los adultos mayores), con un signo negativo, producto de que a esa edad los adultos mayores tienen mucha pérdida dentaria (y por ende pocas piezas con caries), el estrato de mayor nivel sociodemográfico y geográfico de Montevideo (estrato 3 (M)), el hábito de fumar con signo negativo, una mayor carga de CPO y el INSE con valor negativo, reflejando que a mayor INSE , menos probabilidad de tener caries, que podría ser interpretado en términos de poder adquisitivo para paliar la enfermedad o como un aspecto también, educativo, que está muy relacionado con el INSE. Por último, y finalizando el manejo inadecuado del modelo de conteo sobre una variable con recorrido truncado, es un aspecto que el investigador en biomedicina debe conocer cuando, al trabajar, para facilitar la interpretación, se abusa de la técnica, tanto si es él quien lo hace u otro investigador con más formación en estadística.

Capítulo 6

Estudio de los componentes del CPO en el estudio RPAFO2015 mediante modelos de conteo alternativos

6.1. Introducción

La idea en este capítulo es comparar los resultados del estudio RPAFO2015 con otros modelos de conteo que superan los inconvenientes que tienen algunos casos de sobredispersión, que no solamente se logran vencer con distribuciones como la BN (que tiene varias parametrizaciones), sino con otras distribuciones de probabilidad. Entre estas distribuciones, algunas alternativas son las Poisson inversa generalizada (PIG), todas distribuciones que permiten evaluar el gradiente de enfermedad que se perdía al usar modelos de respuesta binaria como se mostró en el capítulo 5. En la revisión de la literatura en varios trabajos publicados en revistas especializadas de biomedicina y epidemiología bien posicionadas no se le presta mucha atención a estos aspectos. En estos trabajos no queda claro el motivo por el cual optan por alternativas, muchas veces, porque se opta por alternativas al MPoi, sino que tampoco se trabaja la capacidad de ajuste (ver capacidad predictiva). Los autores suelen dedicarse a ver las variables y ajustar modelos, lo que los lleva a usar aquellos con variables significativas, pero que podrían ser muy pobres prediciendo. Este último aspecto es relevante, ya que con base en esos modelos los investigadores terminan elaborando teoría para explicar patologías en función de variables que no son buenas predictoras en la discusión.

6.2. Diferentes distribuciones de probabilidad para modelos de conteo

Casi tan importante como poder modelar adecuadamente los modelos de conteo es identificar las posibles distribuciones de probabilidad. Para eso, en la tabla 6.1 se presentan diferentes alternativas de modelos de probabilidad, en las que se modula la varianza en función de un factor de inflación γ. En estas queda determinada, por lo tanto, una forma de variar la varianza que puede ser lineal, como en el caso de la (BN- tipo II) y en forma cuadrática para la (BN-tipo I). Para los casos de las PIG y Poisson Generalizada (PG) las funciones de varianza son polinomios de grado 3 en μ (Hilbe, 2014).

Se puede comenzar por el caso más sencillo, el modelo de Poisson, que tiene la siguiente función de cuantía:

Modelo	Media	Varianza
Poisson	μ	μ
NB-I	μ	$\mu(1+\gamma) = \mu + \gamma\mu$
NB-II	μ	$\mu(1+\gamma\mu) = \mu + \gamma\mu^2$
Binomial Negativa ρ	μ	$\mu(1+\gamma\mu^\rho) = \mu + \gamma\mu^\rho$
PIG	μ	$\mu(1+\gamma\mu^2) = \mu + \gamma\mu^3$
PG	μ	$\mu(1+\gamma\mu)^2 = \mu + 2\gamma\mu^3 + \gamma^2\mu^3$

Tabla 6.1: Relación entre media y varianza para diferentes modelos de conteo

$$f(y;\mu) \quad = \quad \frac{\exp(-\mu)\cdot\mu^y}{y!} \tag{6.1}$$

Cuando existe sobredispersión se puede trabajar con un modelo MBN, que es una mezcla de distribuciones de Poisson con diferentes μ_i y el proceso de mezcla está dado por una distribución Γ para μ, donde $\mu \sim \text{Gamma}(\alpha, \beta)$. En realidad es lo que se llama mezcla Poisson-Gamma,

$$f(y|\mu) = \int_0^\infty \frac{\exp(-\mu)\cdot\mu^y}{y!} f_\Gamma(\mu) d\mu \tag{6.2}$$

$$\mu \sim \text{Gamma}(\alpha, \beta) \tag{6.3}$$

donde el parámetro α controla la forma de la distribución y β su escala.

$$f(\mu) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu} & \text{si } \mu \geq 0 \quad (\alpha > 0,\ \beta > 0) \\ \\ 0 & \text{en otro caso} \end{cases} \tag{6.4}$$

En la figura 6.1 puede verse una posible forma de variar el parámetro μ, de acuerdo a una distribución $Gamma(\alpha, \beta)$

Manejando la notación de (Hilbe, 2014), la BN de tipo II se puede expresar como:

$$f(y;\mu,\gamma) = \binom{y + \frac{1}{\gamma} - 1}{\frac{1}{\gamma} - 1} \left(\frac{1}{1+\gamma\mu}\right)^{\frac{1}{\gamma}} \left(\frac{\gamma\mu}{1+\gamma\mu}\right)^y \tag{6.5}$$

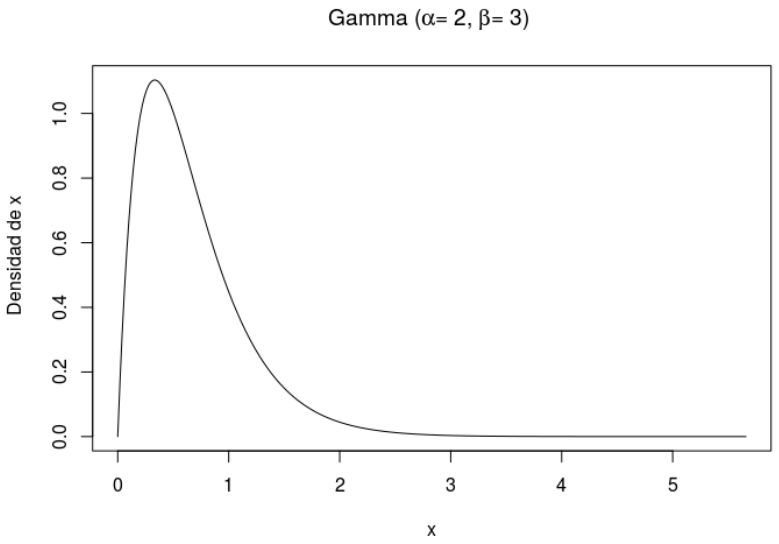

Figura 6.1: Densidad Gamma(α, β) para parámetro μ en distribución BN

Otra distribución de probabilidad para la forma de variar μ es del tipo Inversa Gaussiana (IGau) (μ, ϕ). Es una distribución muy usada en análisis de sobrevida (médico y actuarial) y se caracteriza por ser asimétrica con parámetros de media μ y precisión ϕ. Esto da origen a la PIG, que se debe interpretar de forma similar al proceso de mezcla de Poisson-Gamma para la BN, con la diferencia que en este caso es una mezcla de Poisson y de Inversa Gaussiana, con $\alpha = \frac{1}{\phi}$ (Giner y Smyth, 2016), (Wheeler, 2016).

$$f(y;\mu,\alpha) = \begin{cases} \sqrt{\frac{\phi}{2\pi y^3}} exp \frac{-\phi(y-\mu)^2}{2\mu^2 y}, & \text{si } 0 < y < \infty \\ 0 & \text{en otro caso} \end{cases} \tag{6.6}$$

Cuando se presentan datos que muestran que el parámetro de dispersión γ puede no ser fijo a lo largo de todas las observaciones, como es el caso de la BN y la PIG, es necesario incorporar un parámetro extra ρ que interviene en lo que se conoce como MBN.p. Si se tiene en cuenta el planteo de la función de varianza que surgía de la tabla 6.1 para la BN tipo I y la BN tipo II la diferencia es una potencia del término en $\gamma\mu$, donde ρ debe estimarse junto con el resto de los parámetros (Hilbe, 2011):

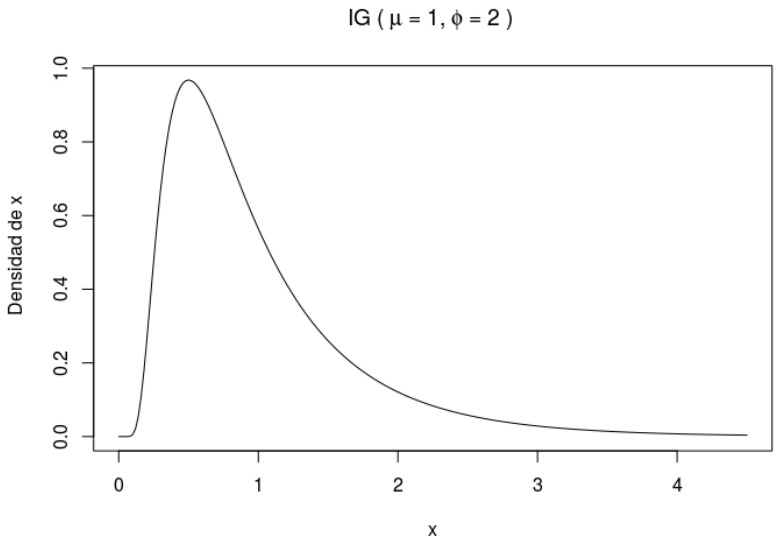

Figura 6.2: Densidad IG(μ, ϕ), para parámetro μ en distribución PIG

$$\begin{cases} \text{Binomial negativa (BN - tipo I)} & \mu & \mu(1+\gamma\mu) = \mu + \gamma\mu \\ \text{Binomial negativa (BN - tipo II)} & \mu & \mu(1+\gamma) = \mu + \gamma\mu^2 \\ \text{Binomial negativa} - \rho & \mu & \mu(1+\gamma\mu^\rho) = \mu + \gamma\mu^\rho \end{cases}$$

Por último, si bien es poco frecuente encontrar datos con subdispersión, existe otra alternativa de distribución de probabilidad, que es la PG, que tiene la siguiente expresión:

$$f(y;\theta,\delta) = \frac{\theta_1(\theta_1+\delta y)^{y-1}e^{-\theta_1-\delta y}}{y!}, \quad y = 0,1,2,.... \tag{6.7}$$

En su libro (Hilbe, 2014), el autor presenta un estudio sobre el tiempo en días de internación (TDI) de pacientes con patología coronaria que reciben dos tipos de procedimientos quirúgicos y que muestran tener una media para TDI de $8,8$ y un desvío estándar de $6,9$ con una sobredispersión de $3,2$. Sin embargo, de las 3589 observaciones originales se intenta modelar el TDI de los que tienen menos de 8 días, y son 1982 los pacientes que verifican esa restricción, con una media de TDI de $4,4$, un desvío estándar de $2,30$ y una dispersión de $0,79$, lo que indica que no es conveniente en el contexto de regresión usar un MPoi o un MBN. Una alternativa es, precisamente, usar la distribución PG.

Es importante, entonces, más allá de ver cada modelo antes planteado, compararlos visualmente, por lo cual para un valor dado de $\mu = 4$ y $\gamma = 0,5$ se presenta como se diferencian entre estos.

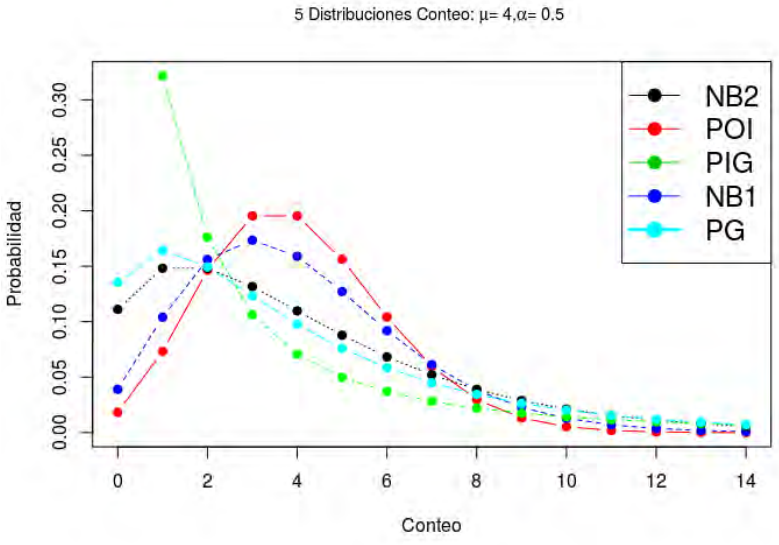

Figura 6.3: Comparación de diferentes distribuciones de probabilidad para modelos de conteo

Sin embargo, en la práctica hay datos que presentan otras patologías, como el exceso de 0 o, dado el problema, la no presencia de 0, por ejemplo los clásicos estudios de días de internación (donde se puede definir que el mínimo es 1), en los que, por definición, el recorrido está truncado. Para eso, se puede recurrir a los modelos que siguen.

6.2.1. Modelos Hurdle

Los modelos Hurdle (MH), que podrían considerarse como modelos con obstáculos, son aquellos que combinan dos procesos de conteo, uno para los 0, con $f_{\text{cero}}(y; z, \gamma)$ (censurado por la derecha en $y = 1$), y otro para conteos > 0 $f_{\text{cont}}(y; x, \beta)$ (truncado por la izquierda en $y = 1$), que puede ser de tipo Poisson o binomial negativo (Mullahy, 1986),(Cameron, 1998).

$$f_{\text{hurdle}}(y; x, z, \beta, \gamma) = \begin{cases} f_{\text{cero}}(0; z, \gamma) & \text{si } y = 0, \\ \frac{(1-f_{\text{cero}}(0;z,\gamma))\cdot f_{\text{cont}}(y;x,\beta)}{(1-f_{\text{cont}}(0;x,\beta))} / & \text{si } y > 0 \end{cases} \tag{6.8}$$

Los parámetros del modelo β, γ y los potenciales parámetros de dispersión θ (si f_{cont} o f_{cero} o ambos tienen densidad negativa binomial) se estiman por máxima verosimilitud, con la ventaja de que los componentes del conteo y de Hurdle pueden maximizarse en forma separada.

La regresión sobre la media tiene la siguiente ecuación:

$$\log(\mu_i) \quad = \quad x_i^\top \beta + \log(1 - f_{\text{cero}}(0; z_i, \gamma)) - \log(1 - f_{\text{cont}}(0; x_i, \beta)), \tag{6.9}$$

6.2.2. Modelos con exceso de ceros (MEC)

Los modelos con exceso de ceros (MEC) toman como casos particulares Poisson con excesos de 0 (PEC) y de binomial negativa con exceso de 0(BNEC) respectivamente son modelos de mezcla que combinan un componente de conteo y una masa de probabilidad en cero con el modelo restante (Poisson o BN) para los conteos > 0 (Cameron, 1998) (Hilbe, 2014).

En este caso, hay dos fuentes de 0 para el modelo, que provienen de la masa puntual en 0 $I_{\{0\}}(y)$ y del modelo de conteo con distribución $f_{\text{cont}}(y; x, \beta)$. La probabilidad de observar un conteo de 0 se incrementa con probabilidad $\pi = f_{\text{cero}}(0; z, \gamma)$

$$f_{\text{ceroinfl}}(y; x, z, \beta, \gamma) \quad = \quad f_{\text{cero}}(0; z, \gamma) \cdot I_{\{0\}}(y) + (1 - f_{\text{cero}}(0; z, \gamma)) \cdot f_{\text{cont}}(y; x, \beta), \quad (6.10)$$

donde $I(\cdot)$ es la función indicadora y la probabilidad no observada π de pertenecer al componente de masa puntual se modela con un MLG de tipo binomial $\pi = g^{-1}(z^\top \gamma)$.

La ecuación de regresión para la media, usando la función de enlace canónico, es

$$\mu_i \quad = \quad \pi_i \cdot 0 \; + \; (1 - \pi_i) \cdot \exp(x_i^\top \beta), \quad (6.11)$$

A partir de las alternativas de modelos planteados hasta aquí, se busca identificar los modelos de probabilidad que mejor se ajustan al componente C del CPO de los planteados en la tabla 6.1. Luego en contextos de modelos de regresión, el objetivo es ver cuál es la mejor alternativa, que puede ser un modelo sencillo al trabajar con una sola distribución de probabilidad o un modelo más complejo de tipo combinado (ver secciones 6.2.1,6.2.2), necesarios para el caso de que exista sobredispersión y exceso de 0.

6.3. Aplicación de modelos de conteo alternativos en RPAFO2015 para los componentes C, P y O

Para evaluar cómo funcionan las distribuciones y los modelos presentados, se trabaja con los datos provenientes del estudio RPAFO2015 en el marco del proyecto de Investigación y Desarrollo (I+D) de la CSIC. En ese estudio se relevan 602 pacientes que consultan en el servicio de registro de Facultad de Odontología, se analiza el componente C para identificar su distribución y se estiman modelos de regresión usando las siguientes variables explicativas.

Como resumen de los parámetros de dispersión implícitos en los MCont, se presentan las alternativas de librerías en (R Core Team, 2017) para su estimación.

- $y \sim Poi(\mu) \longrightarrow$ equidispersión y se usa la librería `stats` (R Core Team, 2017).
- $y \sim BN(\mu, \sigma) \longrightarrow$ para el tratamiento de la variabilidad de parámetros dispersión, $\lambda \sim Gamma(\mu, \lambda)$, se usa librería MASS (Venables y Ripley, 2002a), `COUNT`, (Hilbe, 2016).

- $y \sim BN(\mu, \sigma, \rho) \longrightarrow$ variabilidad de parámetros dispersión a través de observaciones, $\lambda \sim IG(\mu, \phi)$, se trabaja con librería `gamlss` (Rigby y Stasinopoulos, 2005).
- $y \sim PIG \longrightarrow$ variabilidad de parámetros dispersión, $\lambda \sim IG(\mu, \lambda)$, se trabaja con librería `gamlss` (Rigby y Stasinopoulos, 2005).

Se detallan las variables explicativas que se usarán para modelar los componentes del CPO en la tabla 6.2:

Descripción de variables explicativas		
Variable	Nombre	Descripción
V(1)	CPO	(Nivel de CPO) (C)
V(2)	edad	edad en años(C)
V(3)	sexo	Sexo (2 niveles)
V(4)	niveledu	Nivel educativo (4 niveles)
V(5)	ingresos	Ingresos percibidos (3 niveles)
V(6)	alcohol	Nivel de consumo de alcohol (3 niveles)
V(7)	bebiazuc	Nro de dias que consume bebidas azucaradas (C)
V(8)	fumaactual	Fuma actualmente (2 niveles)

Tabla 6.2: Conjunto de variables regresoras usadas para los modelos de conteo en RPAFO2015

En principio, la distribución para los tres componentes y el CPO es la siguiente:

Componentes	n	$\bar{x}$	Desvío Estándar	mediana	mín	máx
C	602	2,50	3,05	2	0	25
P	602	10,17	8,48	8	0	32
O	602	3,66	4,10	2	0	22
CPO	602	16,33	8,11	17	0	32

Tabla 6.3: Medidas de resumen de los componentes de CPO

Si bien se estiman modelos para los tres componentes y el CPO, se muestra con particular detalle lo referente a C, dada la forma que presenta en este caso para los datos de la RPAFO2015 y la importancia que tiene desde el punto de vista epidemiológico.

Usando la librería gamlss (Rigby y Stasinopoulos, 2005), se puede estimar la mejor distribución paramétrica para ajustar la variable C. Si bien al aplicar la función fitDist() los resultados muestran modelos que consideran inflación de 0, solo se presentan los cinco modelos parámetricos vistos en la sección 6.2. En la figura 6.5 se presentan los ajustes que se hicieron para el componente C, considerando que el parámetro μ estimado por la media muestral $\bar{x} = 2,5$, lo que lleva tener que evaluar una alternativa mejor. El PG es el

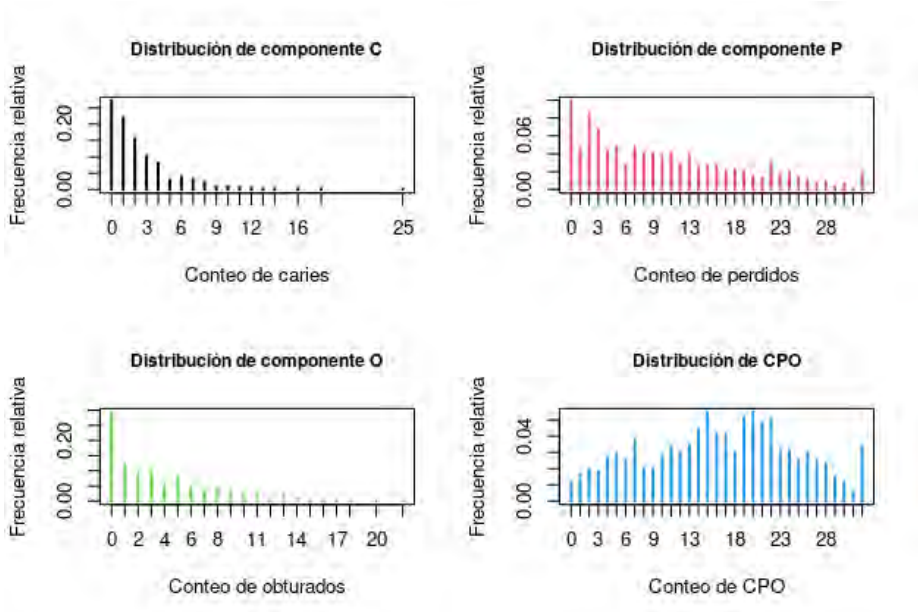

Figura 6.4: Distribución para el CPO y sus tres componentes

mejor, tal come se ve en la tabla 6.4, donde aparece el criterio de ajuste por el estadístico AIC y la figura 6.6, que muestra que es la mejor alternativa.

Modelo ajustado para componente C			
Tipo de MCont	AIC	Parámetros	
PG	2522,3	$\mu = 2,5$	$\gamma = 0,370$
NBII	2525,0	$\mu = 2,5$	$\gamma = 2,44$
NBI	2525,1	$\mu = 2,5$	$\gamma = 0,979$
PIG	2525,8	$\mu = 2,5$	$\gamma = 1,23$
PO	3150,2	$\mu = 2,5$	$\gamma = 0$

Tabla 6.4: Ajuste de la distribución del componente C

La figura 6.7 muestra que se cumplen las características para el ajuste sea adecuado, tal como que los residuos tengan media 0, distribución aparentemente gaussiana a partir de la densidad y cuantiles empíricos que coinciden con los teóricos, donde no aparece un patrón o sesgo en las observaciones.

La tabla 6.4 muestra un resumen de los mejores modelos MCont ajustados para el CPO y sus 3 componentes. En cada modelo ajustado se lleva a cabo el mismo proceso iterativo para encontrar el que tenga mejores indicadores de bondad, como el Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC) y la devianza global y, a su vez, se hace el diagnóstico de ajuste a través de los residuos. Para entender mejor

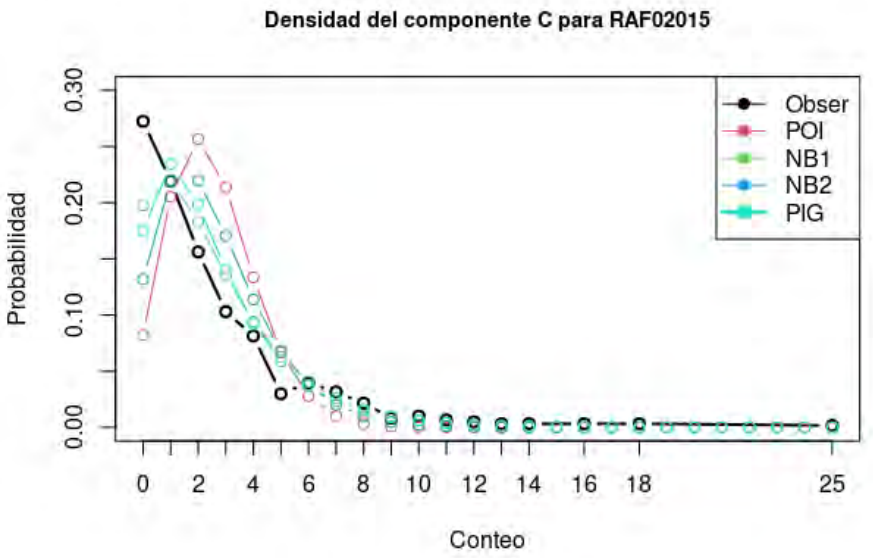

Figura 6.5: Distribución de diferentes MCont para C, dado el valor de $\mu = 2,5$

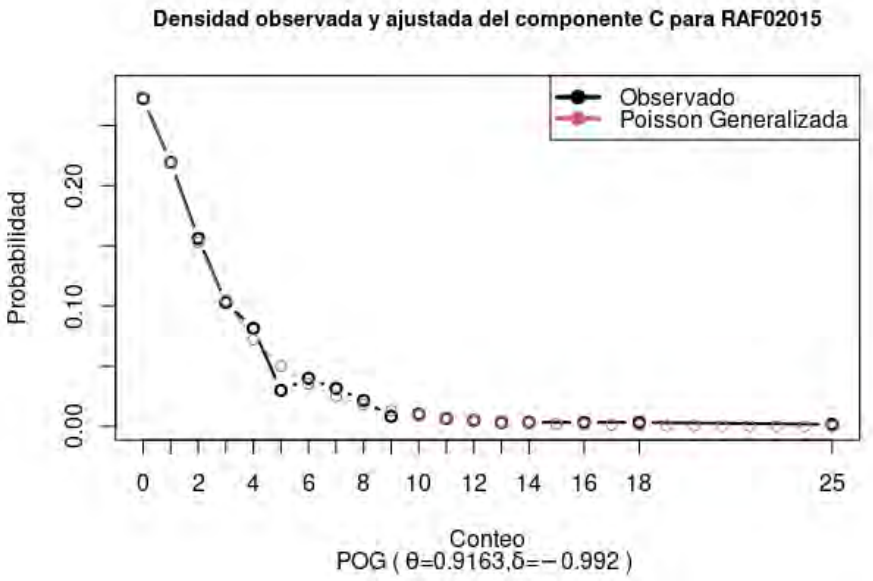

Figura 6.6: Ajuste del C, dado el valor de $\mu = 2,5$ para un modelo PG

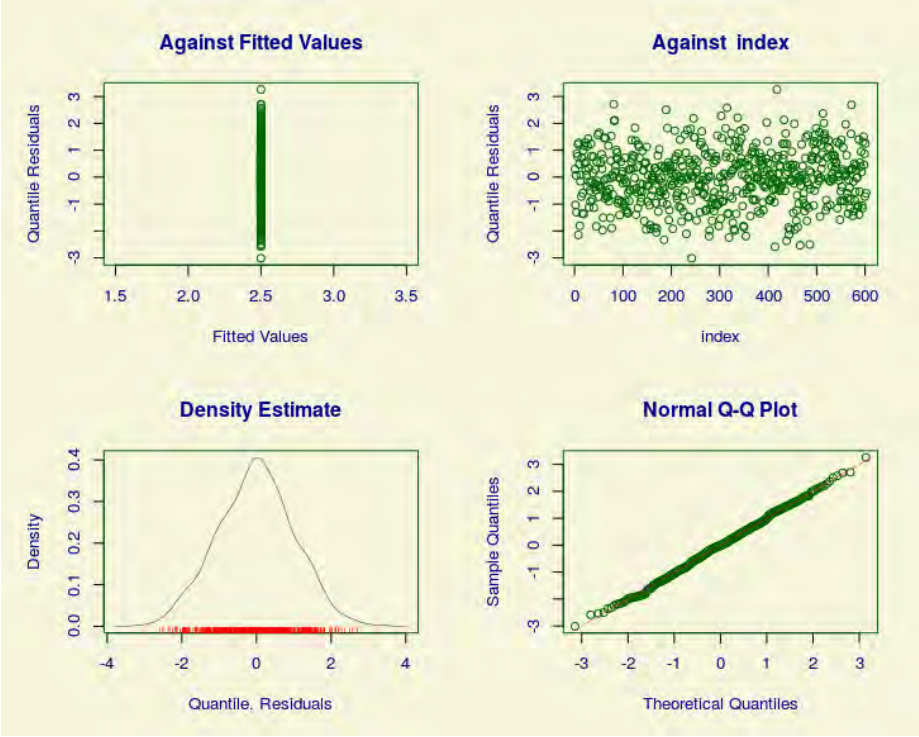

Figura 6.7: Gráficos de bondad de ajuste para el modelo de conteo PG para C

los resultados encontrados se presenta en la siguiente tabla la bondad de ajuste para cada componente del CPO y la jerarquía que queda para los diferentes MCont presentados en 6.2. En la tabla 6.5 solo se consignan los MCont presentados en detalle con anterioridad.

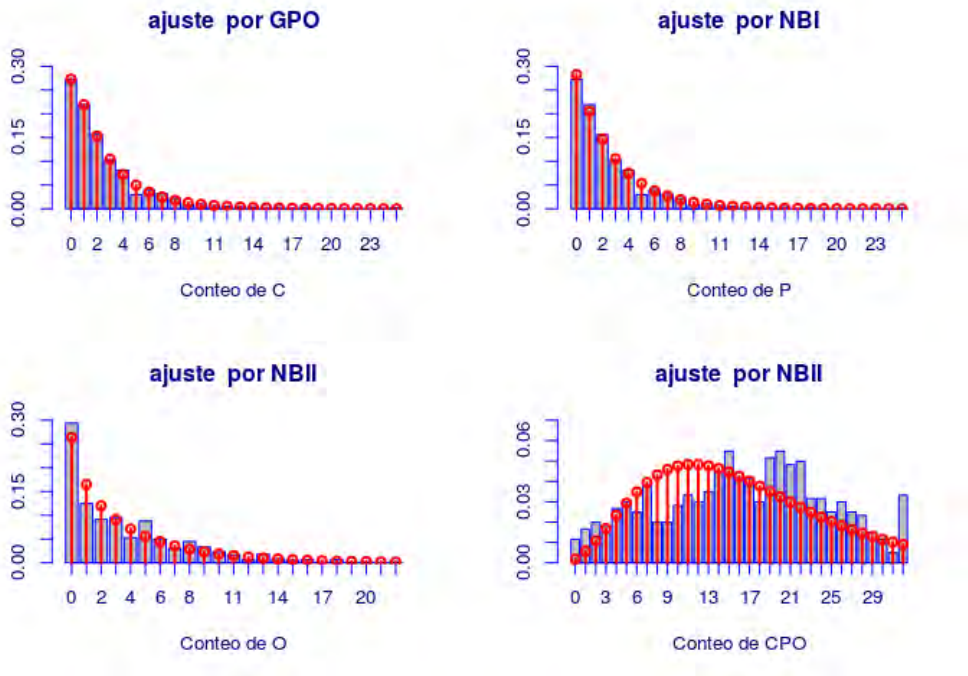

Figura 6.8: Representación gráfica de los modelos de probablidad ajustados para CPO y sus tres componentes

Un aspecto a tener en cuenta es que los coeficientes en la tabla 6.5 están expresados a través de la función de enlace, que es de tipo logarítmico, por lo cual debe reexpresarse usando la base de logaritmos naturales.

Se presentan dos escenarios para poder manejar las distribuciones tratadas en detalle en la sección 6.2 en el escenario A y una nueva distribución, que se detalló en (6.12) para el escenario B.

Ranking de modelos de conteo por componente	
modelos para componente C	
Modelo	Ranking
Poisson	29
Binomial negativa (BN - tipo I)	13
Binomial negativa (BN - tipo II)	12
PIG	14
PG	1
modelos para componente P	
Modelo	Ranking
Poisson	29
Binomial negativa (BN - tipo I)	9
Binomial negativa (BN - tipo II)	10
PIG	22
PG	21
modelos para componente O	
Modelo	Ranking
Poisson	29
Binomial negativa (BN - tipo I)	21
Binomial negativa (BN - tipo II)	11
PIG	22
PG	21
modelos para componente CPO	
Modelo	Ranking
Poisson	27
Binomial negativa (BN - tipo I)	10
Binomial negativa (BN - tipo II)	9
PIG	19
PG	16

Tabla 6.5: Ranking de ajuste de los MCont para CPO y sus tres componentes

Modelo ajustado para componente C				
Tipo	Parámetros			
PG	Coeficientes	EE	t	$\Pr(>\|t\|)$
	μ= 0,916	0,049	18,44	$< 0,001$
	$\gamma = -0,992$	0,0789	-12,56	$< 0,001$
	Devianza Global =2518,39	AIC=2522,39	SBC=2531,2	
Modelo ajustado para componente P				
Tipo	parámetros			
NB-I	Coeficientes	EE	t	$\Pr(>\|t\|)$
	$\mu = 2,32$	0,039	58,63	$< 0,001$
	$\gamma = -0,168$	0,066	-2,35	0,011
	Devianza Global =4048	AIC=4052	SBC=4060	
Modelo ajustado para componente O				
Tipo	parámetros			
NB-II	Coeficientes	EE	t	$\Pr(>\|t\|)$
	$\mu = 1,29$	0,05	25,16 $< 0,001$	
	$\gamma = 1,57$	0,095	16,60	$< 0,001$
	Devianza Global =2904,1	AIC=2908,1	SBC=2917,1	
Modelo ajustado para componente O				
Tipo	parámetros			
NB-II	Coeficientes	EE	t	$\Pr(>\|t\|)$
	$\mu = 2,59$	0,023	119,80	$< 0,001$
	$\gamma = 1,46$	0,078	18,8	$< 0,001$
	Devianza Global =4304,1	AIC=4308,1	SBC=4316,1	

Tabla 6.6: Ajuste de la distribución de CPO y sus tres componentes (Escenario A)

Por último, pueden considerarse otros modelos alternativos a los de conteo, presentados en la sección 6.2, como es el Doble Poisson (DP), que aparece para P, O y CPO y que consiste en un modelo de mezcla, caracterizado por dos parámetros:

$$f(y|\mu,\gamma) = (\frac{1}{\gamma})^{1/2}[\frac{(e^{-y}y^{y})}{y!}][(e\mu)/y]^{y/\gamma}.K \tag{6.12}$$

donde K es una constante de normalización para que $f(y|\mu,\gamma)$ sume 1 en todo el recorrido de sus valores.

Antes de pasar a la etapa de elaboración de modelos de pronóstico con las variables regresoras ya presentadas, en la figura 6.8 se presentan los mejores modelos ajustados para el CPO y sus tres componentes. A continuación se presentan las medidas de resumen para las variables regresoras del componente C, ya que por ser de los el más relevante de los tres desde el punto de vista epidemiológico y el más frecuentemente estudiado, será el único que se analice mediante modelos de regresión en este capítulo.

Medidas de resumen						
Variables	n	$\bar{x}$	Desvío Estándar	mediana	min	max
V(1) CPO	602,00	16,33	8,11	17,00	0,00	32,00
V(2) edad	602,00	45,02	16,85	44,00	18,00	85,00
V(7) bebiazuc	583,00	3,11	2,95	2,00	0,00	7,00
Variables	Tablas de frecuencia					
V(3) sexo		Femenino: 352		Masculino: 250		
V(4) niveledu	1: 170	2: 175	3: 162	4: 93	NA: 20	
V(5) ingresos	1: 325	2: 189	3: 58	NA: 30		
V(6) alcohol	NO consume: 214	mensual: 364	semanal/diario: 21	NA 3		
V(8) fumactual	NO: 399	SI: 199	NA: 4			

Tabla 6.7: Medidas de resumen de las variables regresoras para componente C

	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	-0,349	0,153	-2,27	0,0234
CPO	0,0578	0,007	8,00	0,000
cedad	-0,034	0,004	-8,62	0,0000
sexo(Masculino)	0,285	0,085	3,33	0,0009
bebiazuc	0,049	0,015	3,23	0,001
ingresos(2)	-0,127	0,094	-1,36	0,175
ingresos(3)	-0,378	0,170	-2,22	0,026
Parámetro de dispersión =2,463				

Tabla 6.8: Modelo de regresión cuasi-Poisson para componente C

El modelo estimado presentado en la tabla 6.8 toma en cuenta la sobredispersión, en este caso es casi $2,5$. Observando el modelo se podría pensar que existe una relación entre el número de C y el sexo, la edad, la ingesta de bebidas azucaradas y las personas con mayor ingreso.

Trabajando con un modelo de probabilidad de tipo NBII, se ve un mejor ajuste los datos como aparece en la tabla 6.5, los resultados

	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	-0,4259	0,1526	-2,79	0,0052
CPO	0,0621	0,0076	8,17	0,000
edad	-0.0322	0.0039	-8,28	0.000
sexo(Masculino)	0,3126	0,0899	3,48	0,0005
bebiazuc	0,0461	0,0156	2,95	0,0032
ingresos(2)	-0,0928	0,0966	-0,96	0,3367
ingresos(3)	-0,3956	0,1616	-2,45	0,0143
Parámetro de dispersión para cuasipoisson =2,463				

Tabla 6.9: Modelo de regresión NBI para componente C

Coeficientes para parámetro θ				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	-0,455	0,162	-2,804	0,0052
CPO	0,0637	0,008	7,487	$< 0,0001$
edad	-0,0321	0,0038	-8,237	$< 0,001$
sexo(Masculino)	0,319	0,091	3,496	$< 0,001$
bebiazuc	0,0455	0,016	2,833	$< 0,005$
ingresos(2)	-0,088	0,097	-0,909	$0,363$
ingresos(3)	-0,396	0,158	-2,505	$0,012$
Coeficientes para parámetro δ				
Variables	Coeficientes	EE	Valor z	Pr(> |z|)
(Intercepto)	-1,472	0,105	-14,02	$< 0,001$

Tabla 6.10: Modelo de regresión PG para componente C

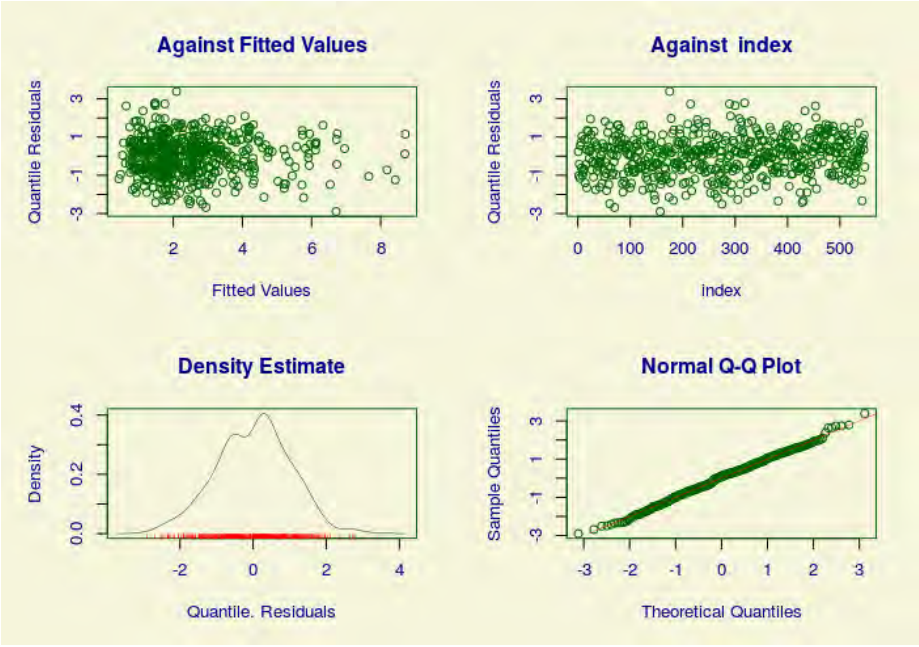

Figura 6.9: Gráficos de bondad de ajuste para el modelo de regresión con distribución PG para C

Si en lugar de estimar un modelo MH con distribución Poisson para el componente truncado se usa la distribución BN, en la que aparece el el parámetro de dispersión, los resultados cambian, tal como se muestra en la tabla 6.12 y donde ambos se pueden comparar para observar la mejoría en usar el modelo más complejo a través del test de Wald.

```
Wald test
Model 1: C ~ CPO + edad + sexo + bebiazuc + ingresos.rec2
| CPO + edad +sexo + ingresos.rec2
Model 2: C ~ CPO + edad + sexo + bebiazuc |
CPO + edad + sexo + ingresos.rec2
Res.Df Df  Chisq Pr(>Chisq)
1    541
2    542 -2 5,2924    0,07092 .
```

El test indicaría que los cambios al usar el modelo con BN para el componente truncado no son relevantes.

La tabla 6.13 muestra un resumen de la performance de los diferentes modelos y, además, indicadores de ajuste, para los que la mejor opción es el modelo con obstáculos y distribución de conteo truncado de tipo BN, aunque igual subestima la cantidad de personas sanas de caries. Esto marca que el modelo hallado parece adecuarse a una

Modelo de conteo MH con distribución Poisson Coeficientes para modelo con función de enlace logit				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	1,715	0,316	5,42	$< 0,0001$
CPO	0,057	0,017	3,33	$< 0,0001$
edad	-0,035	0,007	-4,50	$< 0,0001$
sexo Masculino	0,438	0,209	2,09	0,0363
ingresos(2)	-0,139	0,220	-0,63	0,527
ingresos(3)	-0,835	0,310	-2,68	0,007
Modelo de conteo con distribución Poisson truncado Coeficientes del modelo				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	1,27	0,112	11,25	$< 0,0001$
CPO	0,057	0,005	11,13	$< 0,0001$
edad	-0,030	0,002	-11,18	$< 0,0001$
sexo(Masculino)	0,244	0,059	3,82	$< 0,0001$
bebiazuc	0,045	0,010	4,27	$< 0,0001$

Tabla 6.11: Modelo de regresión MH para componente C, con distribución Poisson

Modelo de Conteo MH con distribución BN Coeficientes para modelo con función de enlace logit				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	1,715	0,316	5,42	$< 0,001$
CPO	0,057	0,017	3,33	$< 0,001$
edad	-0,035	0,007	-4,50	$< 0,001$
sexoMasculino	0,438	0,209	2,09	0,036
ingresos.rec22	-0,139	0,220	-0,63	0,527
ingresos.rec23	-0,835	0,310	-2,68	0,007
Coeficiente $\sigma = 2,034$ para modelo BN				
Modelo de conteo con distribución BN truncada Coeficientes del modelo				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	0,968	0,188	5,12	$< 0,001$
CPO	0,071	0,009	7,31	$< 0,001$
edad	-0,034	0,004	-7,61	$< 0,001$
sexoMasculino	0,270	0,100	2,69	$< 0,001$
bebiazuc	0,046	0,018	2,55	0,011
Log(σ)	0,712	0,208	3,41	$< 0,001$

Tabla 6.12: Modelo de regresión MH para componente C, con distribución binomial negativa

	Modelos vía (MLG)		Modelos con obstáculos	
Tipo	Poisson	BN	Poisson MH-Hurdle	BN-MH
# de parámetros	7	8	13	12
(AIC)	2471	2182	2340	2212
(BIC)	2500	2216		
$Log(\mathcal{L})$	-1228	-1083	-1156	-1094
$\sum_1^{602} x_i = 0$	82	143	146	146

Tabla 6.13: Performance de los diferentes modelos de regresión para componente C mediante modelos parámetricos (MLG) y modelos con obstáculos

población más enferma, por lo que se plantea la necesidad de seguir indagando para identificar un mejor modelo.

6.4. Discusión sobre la distribución y el modelado de los componentes de CPO

Luego de desarrollados los modelos de probabilidad en la sección 6.2, de los cuales se presenta un ejemplo en la figura 6.5, puede verse la forma de la distribución de los cuatro componentes del estudio RPAFO2015. En esta sección se discute, en primer lugar, sobre modelado del componente C, dada su importancia para la epidemiología y la salud pública. Luego se muestra el comportamiento observado para los restantes componentes del CPO.

Puede verse, por lo tanto, que si se trabaja solo con la distribución de C con independencia del resto de las variables (es decir, la distribución incondicional o en un contexto sin variables regresoras), puede decirse que el mejor MCont asociado con las distribuciones más sencillas es el que corresponde a la BN-II, como se ve en la figura 6.5, pero sin perder de vista que el ajuste es pobre para los cinco modelos paramétricos presentados en 6.2. Por ese motivo, tal como se adelantó en la sección 6.3, la mejor alternativa es la PG. Así lo muestra la tabla 6.6, donde, de un total de 29 distribuciones de conteo que se pueden estimar a través de la librería gamlss, (Rigby y Stasinopoulos, 2005), la PG está en el primer lugar. Entre todas esas distribuciones de conteo están las básicas, que se tratan en la sección 6.2 y que corresponden a 6.2.1 y 6.2.2, y varias de sus variantes que, dada su complejidad matemática, se dejan de lado. La figura 6.8 da cuenta de la calidad del ajuste que muestra que los residuos tienen media 0, distribución aparentemente gaussiana, a partir de la densidad y cuantiles empíricos que coinciden con los teóricos, donde no aparece un patrón o sesgo en las observaciones.

Si, ahora, se considera el resto de los componentes del CPO, su ajuste a través de MCont básicos es insuficiente, salvo para el componente P. Para el caso de O, con un ajuste por una BN-II los datos muestran un exceso de 0 y una diferencia importante.

Para terminar esta parte del análisis, el CPO es la variable de conteo que, dado su comportamiento de no monotonía y al ser multimodal, parece difícil de responder a un modelo con distribución conocido. Por ende, parece adecuado considerar un proceso de mezcla.

En cambio, en un contexto de regresión, en las tabla s 6.9 y 6.10 se presentan los resultados de ajustar por un MPoi y una BN, dada la sobredispersión, donde de las ocho variables regresoras ya consideradas en la tabla 6.2 solo resultan significativas el CPO, la edad, el sexo, la cantidad de días en la semana en los que se ingieren bebidas azucaradas y el ingreso de quienes participan del estudio. Para ambos modelos, cuyos coeficientes muestran valores similares, se consideran las variables que son significativas y la asociación con el logaritmo del número medio de caries muestra resultados esperables. La cantidad de caries se asocia con un nivel mayor de CPO, un aumento promedio de una caries por cada punto extra en el CPO y un decremento de casi una carie por cada año por encima de la media (lo que podría explicarse porque, al aumentar la edad, las personas tienen menos piezas presentes). A su vez, el consumo diario de bebidas azucaradas tiene un aumento promedio de una caries por cada día en que la persona consume bebidas. Un aspecto importante para ambos modelos es la importancia del coeficiente asociado al sexo masculino, ya que el número medio de caries es de casi 1,36 con respecto a las mujeres. Por último, los ingresos muestran una asociación negativa, en tanto las personas que están en el último tramo de ingresos tienen una reducción importante en el número de caries con respecto a las del primer tramo, que es la referencia.

Todos estos resultados son muy similares para ambas versiones de los modelos y podrían resultar en la elaboración de teoría epidemiológica. A su vez, podrían llevar al investigador en biomedicina no advertido a cometer errores si no toma en cuenta un aspecto que en general no se considera: la capacidad predictiva del modelo estimado.

Para este caso, en el que el conteo muestra comportamiento patológico al tener sobredispersión y exceso importante de 0, es fundamental verificar el grado de ajuste. No alcanza, por lo tanto, con que las variables sean significativas, ya que no tener en cuenta este aspecto puede llevar al investigador en biomedicina que trabaje con modelos similares a cometer errores muy importantes pautando asociaciones entre variables que, en la práctica, no se dan.

En particular, interesa ver qué sucede con el conteo de 0. El componente C muestra que hay 164 personas sanas de caries, es decir, con un conteo igual a 0, mientras que los dos modelos básicos estimados para el caso de Poisson pronostican 82 personas sin caries y en el modelo de la BN el conteo es de 143. Por eso motivo, los modelos con obstáculos o MH con distribución de conteo truncada que se presentan en las tablas 6.11 y 6.12 tienen una mejor capacidad de detectar personas con C=0. A su vez, las variables regresoras de cada parte del modelo pautan un aspecto muy importante que se detalla a continuación. La parte del modelo MH que se estima mediante un logit, que permite saltar el obstáculo, es la que indica qué perfil tiene la persona para tener o no caries. En el caso del MH

Poisson (MHPoi) son el CPO, la edad, el sexo, el nivel de ingesta de bebidas azucaradas y el ingreso, mientras que para la parte del modelo truncado la ingesta no se debe tener en cuenta y el ingreso pasa a tener un efecto contrario al cambiar de signo. Para el caso del modelo MH binomial negativo (MHBN), las variables que permiten saltar el obstáculo a través del modelo logit son CPO, la edad, el sexo y el ingreso. Para la parte del modelo truncado desaparece el ingreso, mientras que la ingesta de bebidas azucaradas aparece como una variable moduladora en el aumento del conteo de caries.

6.5. Conclusiones para los modelos de conteo para el estudio RPAFO2015

De los resultados encontrados resulta fundamental recordar los pasos que debería seguir el investigador biomédico al trabajar con este tipo de datos que presentan varias patologías desde el punto de vista estadístico, por lo cual no se pueden usar los modelos básicos de conteo:

- Examinar antes cuáles pueden ser las distribuciones que reproducen los conteos (en este caso C,P,O), independientemente de los modelos que se deseen elaborar para encontrar asociaciones.
- No alcanza con encontrar variables significativas con las cuales desarrollar teoría epidemiológica que tenga sentido para el investigador, ya que estaría basada en modelos que no son comparables con los datos en estudio; en el caso presentado las asociaciones son válidas para para poblaciones menos enfermas (hay menos gente con caries).
- Es necesario usar modelos combinados más complejos, como los que se componen de dos submodelos.
    - Uno modelo que trabaja con personas que no tienen Caries (si ésta fuera la variable de conteo, por ejemplo), aspecto que se modela con el componente 1 del modelo (MH);
    - Cuando las personas tienen caries, la cantidad de éstas se modelan con el componente 2 del modelo (MH).

Podría suceder que las variables regresoras que se usan para salir del obstáculo no fueran las mismas que las que contribuyen a modelar el conteo truncado (> 0). Incluso si fueran las mismas, podrían cambiar el sentido de la asociación o su intensidad con coeficientes con distintos valores.

Si bien no se trabajó con el resto de los componentes del CPO o el CPO mismo en un contexto de regresión, ya se vio que las distribuciones no son fácilmente identificables a través de un modelo explícito de los presentados, sino que hay que pensar en identificar mezclas de distribuciones.

Como último comentario, siempre es deseable manejar modelos parsimoniosos, pero es deber del investigador conocer sus limitaciones y lograr un equilibrio entre lo sencillo y lo adecuado.

Capítulo 7

Uso de la regresión Beta para la creación de indicadores alternativos para la vigilancia en salud oral

7.1. Introducción

Tal como se manifestó en capítulos anteriores, los indicadores generalmente utilizados en la epidemiología y salud pública pueden tener limitaciones, ya que muchas veces no toman en cuenta la estructurada multivariada de la información o, si la toman, lo hacen a través de algoritmos de cálculos que generan indicadores univariados para ganar en simplicidad y no miden correctamente, por lo tanto, los fenómenos en estudio. Esta característica de uso de indicadores limitados (al no tener en cuenta la estructurada multivariada de la información o algoritmos de cálculos que generan indicadores univariados para ganar en simplicidad) se da en otros dominios de la salud pública y no solo en salud bucal, al menos en nuestro país.

7.1.1. Antecedentes

Para entender los aspectos mencionados, en este capítulo se podrían considerar aquellos indicadores que tienen que ver con el estado de las piezas dentales y que se identifican con *ceo* para los niños, CPO en adultos e ICDASII como una variante al CPO que evalúa de forma más detallada y gradualista estadíos o niveles de enfermedad. Para el desarrollo de las indicadores alternativos se toma como ejemplo el CPO que tiene la siguiente expresión.

$$CPO_i^g = \sum_j^n C_{i,j,k}^g + \sum_j^n P_{i,j,k}^g + \sum_j^n O_{i,j,k}^g \tag{7.1}$$

En virtud de las limitaciones del índice CPO, un índice univariado que enmascara muchas configuraciones de los componentes aislados, es necesario manejar alternativas, como utilizar los tres componentes del CPO por separado y transformarlos en tasas, proporciones o índices basados en medidas de entropía, es decir, considerar la misma información, pero analizarla de otra manera en el contexto de los modelos de regresión.

7.1.2. Modelos probabilísticos para ajustar tasas

Una posibilidad para considerar modelos de regresión es pensar en transformaciones de la variable de respuesta, que al ser variables de conteo se pueden reformular como tasas o proporciones para el caso, presentando los índices del CPO o algunos de sus compo-

nentes relativizados contra diferentes totales: 32 (máximo número de piezas, número de piezas presentes, etc). La ventaja de estas transformaciones es que existen modelos probabilísticos conocidos para trabajar con proporciones, de los cuales se conocen muchas características necesarias a la hora de usarlos para hacer inferencias, considerando que:

- las proporciones a estimar están en el rango $(0, 1)$;
- otras distribuciones pueden estar acotadas en el intervalo (a, b) y pueden ser reparametrizadas en el rango $(0, 1)$;
- no cumplen el supuesto de normalidad;
- pueden existir asimetrías muy importantes;
- la varianza puede cambiar, lo que obliga a manejar otros modelos.

A modo de ejemplo, para este capítulo, se relativizan los componentes del CPO para convertirlos en proporciones usando indistintamente los tres componentes del CPO o el componente S (diente sano) del siguiente modo:

Descripción de proporciones	
Proporción	Cálculo
Nivel I: Enfermedad pasada y presente	
(prop1)	$\frac{\sum O_i}{\sum O_i + \sum C_i}$
(prop2)	$\frac{\sum C_i}{\sum O_i + \sum C_i}$
Nivel II: Enfermedad pasada y presente (CPO)	
(prop3)	$\frac{\sum C_i}{\sum O_i + \sum C_i + \sum P_i}$
(prop4)	$\frac{\sum P_i}{\sum O_i + \sum C_i + \sum P_i}$
(prop5)	$\frac{\sum O_i}{\sum O_i + \sum C_i + \sum P_i}$
Nivel III: Enfermedad pasada y presente + sanos	
(prop6)	$\frac{\sum S_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$
(prop7)	$\frac{\sum P_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$
(prop8)	$\frac{\sum C_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$
(prop9)	$\frac{\sum O_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$

Tabla 7.1: Transformación de los componentes de CPO en proporciones

En la tabla 7.1 se detalla lo que representa conceptualmente cada derivado de los componentes del CPO

1. $\frac{\sum O_i}{\sum O_i+\sum C_i}$ nivel de cobertura de la enfermedad (prop1)
2. $\frac{\sum C_i}{\sum O_i+\sum C_i}$ indicador de estadio de la enfermedad en el momento actual (prop2)
3. $\frac{\sum C_i}{\sum O_i+\sum C_i+\sum P_i}$ índice de C en el momento actual (prop3)
4. $\frac{\sum P_i}{\sum O_i+\sum C_i+\sum P_i}$ índice de P en el momento actual (prop4)
5. $\frac{\sum O_i}{\sum O_i+\sum C_i+\sum P_i}$ nivel de cobertura de la enfermedad (considerando oezas perdidas)(prop5)
6. $\frac{\sum S_i}{\sum O_i+\sum C_i+\sum P_i+\sum S_i}$ indicador de salud en caries en el momento actual (prop6)
7. $\frac{\sum P_i}{\sum O_i+\sum C_i+\sum P_i+\sum S_i}$ indicador de necesidad de prótesis en el momento actual (prop7)
8. $\frac{\sum C_i}{\sum O_i+\sum C_i+\sum P_i+\sum S_i}$ indicador de necesidad de obturación en el momento actual (prop8)
9. $\frac{\sum O_i}{\sum O_i+\sum C_i+\sum P_i+\sum S_i}$ indicador de enfermedad pasada en el momento actual (prop9)

En la sección 7.3 se amplía y detalla el motivo para considerar tres niveles de indicadores medidos con diferentes proporciones.

7.1.3. Formulación del modelo de probabilidad Beta

De los modelos que se podrían utilizar para modelar proporciones, se considera el de probabilidad Beta que se presenta a continuación. Luego será incorporado en los modelos de regresión como un caso particular de MLG.

$$f(y;p,q)=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}y^{p-1}(1-y)^{q-1},\quad 0<y<1, \tag{7.2}$$

- con $0<\mu<1$, $\phi>0$. Cribari-Neto y Zeileis (Cribari-Neto y Zeileis, 2010) hacen una reparametrización $\mu=\frac{p}{p+q}$ y $\phi=p+q$.

  Se puede escribir $y\sim\mathcal{B}(\mu,\phi)$. Por lo tanto, $E(y)=\mu$, $VAR(y)=\mu(1-\mu)/(1+\phi)$.

- El parámetro ϕ es un parámetro de precisión, ya que para μ fijo, cuanto más grande es ϕ, más pequeña es la varianza de y; ϕ^{-1} es un parámetro de dispersión.

$$f(y;\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\quad 0<y<1, \tag{7.3}$$

En la Figura ?? se pueden ver ejemplos de distribución BETA al cambiar los parámetros de dispersión o precisión.

La formulación del modelo predictivo se puede hacer a partir de los trabajos de (Kieschnick y McCullough, 2003) y (Salinas-Rodríguez et al., 2009), en los cuales

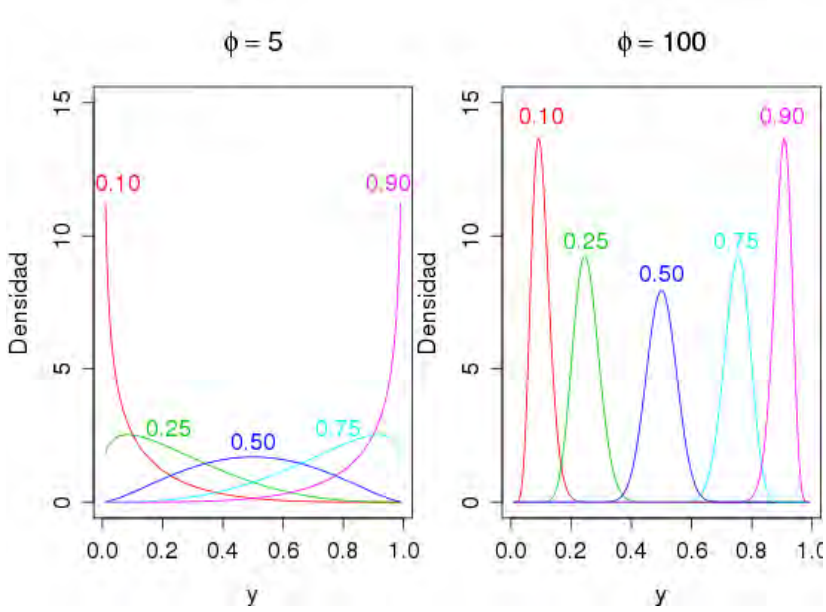

Figura 7.1: Densidades Beta en el intervalo $(0, 1)$ para diferentes valores de μ y ϕ

- si se tiene $y_1, \ldots, y_n$ una muestra aleatoria tal que $y_i \sim \mathcal{B}(\mu_i, \phi)$, $i = 1, \ldots, n$;
- el modelo de regresión Beta (MRB) es $\mu_i = g^{-1}(x_i^\top)$, donde , el vector de parámetros de regresión, se estima por máxima verosimilitud (ML).

$$g(\mu_i) = x_i^\top = \eta_i, \tag{7.4}$$

El modelado de la dispersión se puede hacer a través de:

$$VAR(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi} = \frac{g^{-1}(x_i^\top)[1-g^{-1}(x_i^\top)]}{1+\phi}. \tag{7.5}$$

$$\begin{aligned} \ell_i(\mu_i, \phi) &= \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) + (\mu_i\phi - 1)\log y_i \\ &\quad + \{(1-\mu_i)\phi - 1\}\log(1-y_i). \end{aligned} \tag{7.6}$$

La funciones de enlaces (link) para modelos de regresión Beta (RB) pueden ser:

- logit $g(\mu) = \log(\mu/(1-\mu))$;
- probit $g(\mu) = \Phi^{-1}(\mu)$, con $\Phi(\cdot)$ función de distribución normal estandarizada;
- log-log complementaria $g(\mu) = \log\{-\log(1-\mu)\}$;
- log-log $g(\mu) = -\log\{-\log(\mu)\}$.

Cuando para el modelo Beta existe heterocedasticidad, el parámetro de precisión no es constante en todas las observaciones. Es por esto que es necesario modelarlo como se hizo con la media.

En particular, $y_i \sim \mathcal{B}(\mu_i, \phi_i)$, $i = 1, \ldots, n$, y

$$g_1(\mu_i) = \eta_{1i} = x_i^\top, \tag{7.7}$$

$$g_2(\phi_i) = \eta_{2i} = z_i^\top \gamma, \tag{7.8}$$

donde $= (_1, \ldots, _k)^\top$, $\gamma = (\gamma_1, \ldots, \gamma_h)^\top$, $k + h < n$ son los coeficientes de regresión de ambas ecuaciones, η_{1i} y η_{2i} son predictores lineales, x_i y z_i son los vectores de regresión, los que se estiman por ML remplazando ϕ por ϕ_i en la ecuación 7.6.

Un aspecto que debe ser considerado es que la variable Beta está definida en el intervalo $(0, 1)$. Si, previo a la transformación en proporción de la variable de conteo, tenemos 0 o el total contra el que se normaliza coincide con el máximo de la variable de conteo y la proporción vale 1, es necesaria una transformación extra que trunque los extremos. Esta debe tener la siguiente característica: si se tiene una variable y_i que modela una proporción, se puede aplicar la transformación de Smithson-Verkuilen (Verkuilen y Smithson, 2012)

$$y_i^* = \frac{y_i(n-1) + 0,5}{n}$$

que funciona truncando menos cuanto mayor sea n, lo que permite estimar los parámetros.

7.2. Aplicación de modelos de RB

Estos resultados surgen del análisis de los componentes del CPO, considerando varias transformaciones en proporciones, con diferentes criterios de elaboración. Desde el punto de vista epidemiológico representan diferentes dimensiones del problema en el contexto del estudio RPAFO2015. Tanto para los gráficos como para los modelos se trabaja con el lenguaje *R* (R Core Team, 2016), en particular para los modelos de conteo la librería `MASS` (Venables y Ripley, 2002b) y la librería `Betareg` (Cribari-Neto y Zeileis, 2010). La distribución de los componentes del CPO es la que se muestra en la figura 7.2, donde se ve una gran asimetría para el componente de caries y para el de obturación, lo que habla de una población con una carga de enfermedad actual o pasada media, pero con una mayor presencia del componente *P*. Este es un resultado esperable, teniendo en cuenta que se trata de las personas que pertenecen al estudio antes mencionado y que tienen, en general, una mayor carga de enfermedad.

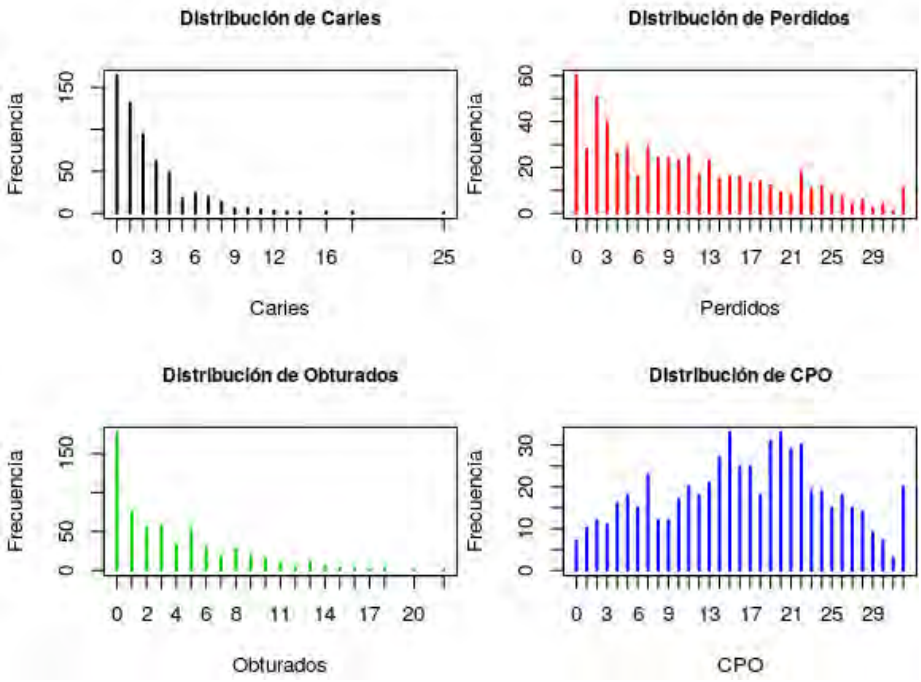

Figura 7.2: Densidades empíricas de los componentes del CPO

También es importante ver cómo cambian los resultados al trabajar con los conteos cuando se los reexpresa en proporciones, eligiendo diferentes formas de normalizar, tal como se presentaron en 7.1.2.

variables	n	$\bar{x}$	SE	mediana
prop1	560	0,54	0,38	0,60
prop2	560	0,46	0,38	0,40
prop3	602	0,48	0,25	0,47
prop4	602	0,32	0,27	0,28
prop5	595	0,19	0,23	0,11
prop6	595	0,54	0,30	0,57
prop7	595	0,27	0,28	0,19
prop8	602	0,08	0,10	0,06
prop9	602	0,12	0,13	0,07

Tabla 7.2: Distribuciones de las proporciones de los componentes del CPO

Observando el comportamiento de los componentes C, O, P y S se ve que es necesario hacer transformaciones para tener los valores en el rango $(0, 1)$ transformando con $y_i^* = \frac{y_i(n-1)+0,5}{n}$.

Para evaluar si se pueden obtener mejoras, se presentan varios modelos de regresión Beta a los resultados que se manejaron en el capítulo 6. Esto se hace transformando los diferentes componentes del CPO en proporciones. Si bien en los componentes no existen datos faltantes, como se ve en la tabla 7.2, al hacer la transformación en proporciones aparecen datos faltantes según cual sea la transformación, ya que tienen denominadores con conteos nulos. Teniendo en cuenta este detalle, se analiza si ese pérdida es aleatoria o si, por el contrario, tiene un patrón definido.

Como forma general se opta por modelar todas las proporciones con un mismo grupo de variables:

Descrición de variables explicativas		
Variable	Nombre	Descripción
V(1)	CPO	(Nivel de CPO)
V(2)	edad	edad en años
V(3)	sexo	Sexo (2 niveles)
V(4)	niveledu	Nivel educativo (4 niveles)
V(5)	ingresos	Ingresos percibidos (3 niveles)
V(6)	alcohol	Nivel de consumo de alcohol (3 niveles)
V(7)	bebiazuc	Número de días que consume bebidas azucaradas
V(8)	fumaactual	Fuma actualmente (2 niveles)

Tabla 7.3: Conjunto de variables regresoras usadas para los modelos de regresión Beta en RPAFO2015

Para cada modelo evaluado para las diferentes proporciones definidas en la sección 7.1.2 se considera este bloque de ocho variables explicativas y se reporta el modelo final luego de que se eliminaron las variables no significativas. A su vez, en cada caso se verifica la no existencia de heterocedasticidad, que daría lugar a un único parámetro ϕ de dispersión. Se procede a modelar la dispersión asociada a cada variable, tal como se detalla en la ecuación (7.8). Por su parte, el criterio para considerar si es pertinente la modelización de la heterocedasticidad surge de la evaluación previa a través del test de Breusch-Pagan, que se reporta en cada caso.

Se comienza por evaluar, para la tasa *prop*1, cuáles son las variables que mejor explican el comportamiento en el modelo 1:

Coeficientes (Modelo 1 con función de enlace logit)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	-1,172	0,219	-5,34	$< 0,0001$
CPO	-0,034	0,009	-3,46	$< 0,0001$
edad	0,029	0,004	6,39	$< 0,0001$
sexo(Masculino)	-0,312	0,117	-2,65	0,008
N.EDU(2)	0,419	0,155	2,70	0,006
N.EDU(3)	0,652	0,161	4,02	$< 0,0001$
N.EDU(4)	0,798	0,187	4,27	$< 0,0001$
ingresos.(2)	-0,0152	0,153	-0,09	0,920
ingresos.(3)	0,408	0,171	2,37	0,017
ingresos.(4)	0,569	0,195	2,90	0,003
(Parámetro ϕ de precisión)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(ϕ)	0,64862	0,03141	20,65	$< 0,0001$
(Test de Breusch-Pagan)				
BP = 16,122, df = 9, p-value = 0,064				

Tabla 7.4: Modelo 1 de regresión Beta para *prop*1 en estudio RPAFO2015

El modelo 2 sería el que considera la relación entre *prop*2 y el grupo de variables explicativas, pero, por cómo está construida, esta proporción es el complemento de *prop*1, por lo cual el modelo sería el mismo que para *prop*1.

Se cambia de grupo de proporciones del nivel I y se trabaja con las dos proporciones ya presentadas, considerando solamente las piezas presentes, es decir, dejando de lado el componente O, pasando al grupo nivel II.

Se incorporan variantes a *prop*1 y *prop*2 al usar el componente O, en el que el primer modelo es para *prop*3, consignado en la tabla 7.5, que muestra que las variables regresoras no son las mismas, ya que el nivel educativo desaparece y la ingesta de bebidas azucaradas aparece como un aspecto a ser tenido en cuenta. Por otra parte, también es necesario modelar la heterocedasticidad, donde la edad aparece como dato relevante.

El modelo que se encuentra para *prop*3 es:

Coeficientes (Modelo 3 con función de enlace logit)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	-0,327	0,188	1,73	$0,082$
edad	-0,045	0,003	-12,95	$0,0001$
sexo(Masculino)	0,267	0,092	2,87	$< 0,0001$
ingresos(2)	-0,083	0,119	-0,69	0,484
ingresos(3)	-0,127	0,138	-0,91	0,359
ingresos(4)	-0,391	0,157	-2,47	0,013
bebiazuc	0,030	0,016	1,89	0,058
(Test de Breusch-Pagan)				
86,994, df = 6, p-value ¡2,2e-16				
(Parámetro ϕ de precisión)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercepto)	-0,707	0,16	-4,35	$< 0,0001$
edad	0,038	0,01	10,36	$< 0,0001$

Tabla 7.5: Modelo 3 de Regresión Beta para *prop*3 en estudio RPAFO2015

También para la modelización de prop4, que trabaja con el componente P, al considerar al grupo de variables explicativas deben descartarse alcohol e ingesta de bebidas azucaradas y al detectarse que existe heterocedasticidad en el parámetro de precisión, se modela lo que finalmente da como resultado lo que aparece en 7.6.

Coeficientes (Modelo 4 on función de enlace logit)				
Variables	Coeficiente	EE	Valor z	r(> |z|)
(Intercepto)	-3,816	0,12	-29,46	< ,0001
CPO	0,156	0,00	27,70	< ,0001
edad	0,011	0,00	4,89	< ,0001
N.EDU(2)	-0,297	0,08	-3,70	< ,0001
N.EDU(3)	-0,510	0,08	-6,09	< ,0001
N.EDU(4)	-0,575	0,09	-5,97	$< 0,0001$
(Test de Breusch-Pagan) BP = 61,69, df = 5, p-value = 5,436e-12				
(Parámetro ϕ de precisión)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
(Intercept)	2,544	0,17	14,55	$< 0,0001$
CPO	-0,062	0,01	-7,06	$< 0,0001$
edad	0,013	0,01	3,16	$< 0,0001$
fuma actualmente-Si	0,244	0,11	2,17	$< 0,0001$

Tabla 7.6: Modelo 4 de regresión Beta para *prop*4 en estudio RPAFO2015

La última proporción del bloque de nivel II que considera el componente O muestra las siguientes relaciones:

Pasando a las proporciones del nivel III, donde aparece por primera vez la dimensión de la salud a través del componente S de piezas sanas, el indicador de salud en el momento actual (*prop*6) tiene un modelo ajustado que se presenta en la tabla 7.8.

En la figura 7.3 se muestra cómo queda la curva de la proporción de piezas sanas en función del nivel de CPO, controlando para los 2 últimos niveles de ingreso (3 y 4) y de ingesta de alcohol (mensual y semananal/diario).

Coeficientes (Modelo 5 con función de enlace logit)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	-0,706	0,236	-2,98	0,0027
CPO	-0,077	0,009	-7,87	$< 0,0001$
edad	0,007	0,004	1,602	0,109
sexo Masculino	-0,232	0,099	-2,34	0,019
niveledu.rec2	0,453	0,127	3,55	0,000
niveledu.rec3	0,729	0,134	5,41	$< 0,0001$
niveledu.rec4	0,903	0,156	5,76	$< 0,0001$
ingresos.rec12	0,010	0,126	0,08	0,933
ingresos.rec13	0,526	0,149	3,52	0,000
ingresos.rec14	0,661	0,167	3,94	$< 0,0001$
alcohol.rec2-mensual	0,168	0,103	1,63	0,102
alcohol.rec3-semanal/diario	-0,0715	0,261	-0,27	0,784517
bebiazuc	-0,049	0,019	-2,50	0,012
(Test de Breusch-Pagan)				
BP = 78,1, df = 12, p-value = $< 0,0001$				
(Parámetro ϕ de precisión)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	-0,924	0,178	-5,19	$< 0,0001$
CPO	0,046	0,009	4,68	$< 0,0001$
edad	0,013	0,004	2,87	0,004
bebiazuc	0,050	0,019	2,52	0,011

Tabla 7.7: Modelo 5 de Regresión Beta para prop5 en estudio RPAFO2015

Coeficientes (Modelo6 con función de enlace logit)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	2,659	0,047	56,17	$< 0,0001$
CPO	-0,167	0,002	-72,80	$< 0,0001$
ingresos(2)	-0,056	0,0394	-1,443	0,148
ingresos(3)	-0,036	0,0437	-0,84	0,398
ingresos(4)	0,143	0,053	2,66	0,007
alcohol(mensual)	-0,072	0,032	-2,25	0,024
alcohol(semanal/diario)	-0,0476	0,094	-0,502	0,615
(Test de Breusch-Pagan)				
19,282, df = 6, p-value = 0,003				
(Parámetro ϕ de precisión)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	4,508	0,167	26,91	$< 0,0001$
edad	-0,017	0,003	-5,18	$< 0,0001$

Tabla 7.8: Modelo 6 de regresión Beta para *prop*6 en estudio RPAFO2015

Para el indicador de necesidad de prótesis en el momento actual (prop7) el modelo ajustado es el que sigue:

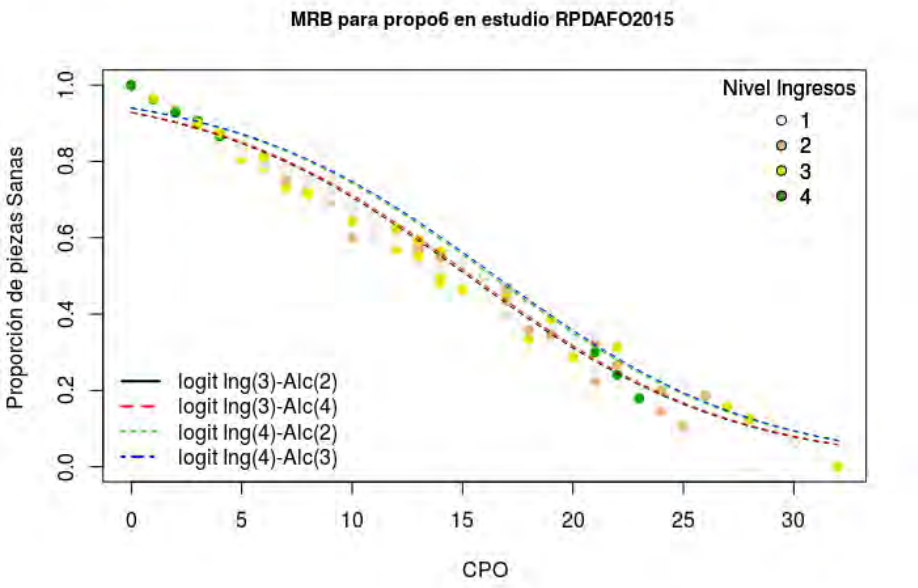

Figura 7.3: Relación entre *prop*6 y CPO para (modelo6)

Coeficientes (Modelo 7 con función de enlace logit)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	-2,242	0,176	-12,71	$< 0,0001$
CPO	0,073	0,009	8,11	$< 0,0001$
edad	-0,036	0,004	-7,95	$< 0,0001$
sexo(Masculino)	0,240	0,081	2,95	0,003
ingresos(2)	-0,097	0,105	-0,92	0,356
ingresos(3)	-0,161	0,119	-1,35	0,176
ingresos(4).rec14	-0,410	0,138	-2,96	0,003
bebiazuc	0,031	0,014	2,19	0,0283
(Test de Breusch-Pagan)				
BP = 74,838, df = 9, p-value = 1,701e-12				
(Parámetro ϕ de precisión)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	2,054	0,197	10,42	$< 0,0001$
CPO	-0,064	0,011	-5,81	$< 0,0001$
edad	0,024	0,005	4,43	$< 0,0001$

Tabla 7.9: Modelo 7 de Regresión Beta para *prop*7 en estudio RPAFO2015

Coeficientes (Modelo 8 con función de enlace logit)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	-3,862	0,140	-27,50	$< 0,0001$
CPO	0,155	0,0057	26,90	$< 0,0001$
edad	0,0122	0,002	4,99	$< 0,0001$
N.EDU(2)	-0,262	0,082	-3,19	0,00138
N.EDU(3)	-0,457	0,085	-5,33	$< 0,0001$
N.EDU(4)	-0,513	0,099	-5,13	$< 0,0001$
ingresos(2)	-0,033	0,081	-0,41	0,678
ingresos(3)	-0,130	0,096	-1,35	0,176
ingresos(4)	-0,214	0,109	-1,96	0,049
fuma actualmente -SI	0,130	0,067	1,91	0,054
(Test de Breusch-Pagan)				
BP = 74,838, df = 9, p-value = 1,701e-12				
(Parámetro ϕ de precisión)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	2,721	0,171	15,89	$< 0,0001$
CPO -0,054	0,009	-5,95	$< 0,0001$	
edad 0,009	0,004	2,14	0,0324	

Tabla 7.10: Modelo 8 de regresión Beta para prop8 en estudio RPAFO2015

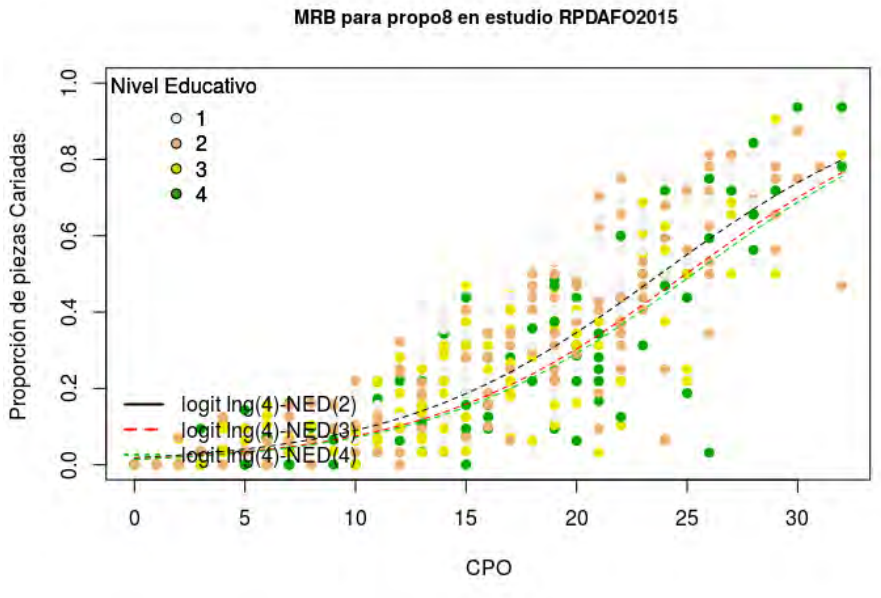

Figura 7.4: Relación entre *prop*8 y CPO para modelo8

Coeficientes (Modelo 9 con función de enlace logit)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	-3,626	0,210	-17,268	$< 0,0001$
edad	0,020	0,003	6,104	$< 0,0001$
sexo(Masculino)	-0,221	0,08	-2,571	0,010
N.EDU(2)	0,439	0,114	3,83	0,0001
N.EDU(2)	0,621	0,117	5,29	$< 0,0001$
N.EDU(2)	0,825	0,134	6,12	$< 0,0001$
ingresos(2)	-0,012	0,110	-0,10	0,913
ingresos(3)	0,451	0,123	3,65	0,000
ingresos(4)	0,299	0,141	2,12	0,033
alcohol (mensual)	0,211	0,090	2,34	0,019
alcohol(semanal/diario)	0,028	0,247	0,11	0,906
(Test de Breusch-Pagan)				
BP = 74,838, df = 9, p-value = 1,701e-12				
(Parámetro ϕ de precisión)				
Variables	Coeficiente	EE	Valor z	Pr(> |z|)
Intercepto	2,036	0,194	10,49	$< 0,0001$
edad	-0,011	0,004	-2,91	0,003

Tabla 7.11: Modelo 9 de regresión Beta para *prop*9 en estudio RPAFO2015

Por último, para ordenar los resultados y facilitar su discusión en la sección 7.3, la tabla 7.12 permite discernir la performance de cada modelo.

Tipo y calidad de ajuste para cada modelo estimado				
Proporción	Modelo	Pseudo-Rcuadrado	Heterocedasticidad	Grupo
*prop*1	modelo 1	0,171	No	Insuficiente
*prop*2	modelo 2	0,171	No	Insuficiente
*prop*3	modelo 3	0,158	Sí	Insuficiente
*prop*4	modelo 4	0,684	Sí	Suficiente
*prop*5	modelo 5	0,202	Sí	Insuficiente
*prop*6	modelo 6	0,778	Sí	Suficiente
*prop*7	modelo 7	0,115	Sí	Insuficiente
*prop*8	modelo 8	0,687	Sí	Suficiente
*prop*9	modelo 9	0,151	Sí	Insuficiente

Tabla 7.12: Ranking de los modelos ajustados para cada tipo de proporción en estudio RPAFO2015

7.3. Discusión sobre los modelos de RB estimados

En la sección 7.2, que contiene los resultados, se fueron viendo diferentes modelos de RB, siguiendo una lógica común de usar las mismas variables explicativas para todas las proporciones que, con transformaciones, podían dar cuenta de las diferentes dimensiones del CPO. Estas proporciones se agruparon en tres niveles siguiendo una jerarquía en términos de lo que pueden llegar a representar desde lo conceptual. Las del nivel I son indicadores que muestran la enfermedad actual (C) o pasada (O), pero que dejan de lado el componente O que de por sí representa una patología que debe ser revertida mediante la colocación de prótesis. Los otros tres indicadores del grupo del nivel II ya consideran la totalidad de los componentes de CPO, clásicamente utilizados en el análisis de epidemiología. Estos muestran ventajas sobre los que se pueden alcanzar mediante el uso de MCont, ya tratados en el capítulo 6 con mucho detalle. Los cuatro últimos indicadores presentan la ventaja de que incorporan una dimensión que no se usa siempre, las piezas sanas (es decir, que aún no sufrieron patología cariogénica). Tienen, además, como característica, que el componente S (sanos) pasa a formar parte del algoritmo de cálculo al estar en el denominador. A su vez, también puede introducir una proporción con la *prop*6, que es la única que muestra la salud y no su ausencia, como si lo hacen los indicadores manejados en el grupo nivel I y nivel II.

Siguiendo este análisis jerárquico, en la sección aparecen modelos que no siempre contienen las mismas variables explicativas y en los que, a su vez, aparece una fuente

de variabilidad extra, evaluada a través del componente ϕ de dispersión. Las variables que deben ser tenidas en cuenta no son las mismas, lo que supone una necesidad de reflexión para el investigador en biomedicina que deba trabajar con esta metodología. Para explicar dimensiones que están íntimamente relacionadas, como los componentes del CPO, el mecanismo indirecto (ya que no se puede hablar de causalidad), por el cual existe más o menos patología es diferente según cuál sea la que se considere. A su vez, la heterogeneidad, que no debe ser vista como algo inadecuado, sino como una forma de entender mejor la casuística observada; también parece estar regida por diferentes mecanismos, ya que no siempre son las mismas variables, lo que hace que ajustar sea necesario.

Teniendo en cuenta estos aspectos metodológicos, los modelos estimados se pueden clasificar en términos de la dificultad que presentan, ya que los que muestran heterocedasticidad deben ser estimados de nuevo, incorporando el juego de variables adecuado para controlar esa anomalía, la cantidad de variables necesarias para lograr el mejor ajuste y su capacidad predictiva.

Tal como se explica en el capítulo 3, lo ideal es que el especialista en ciencias médicas disponga de modelos estadísticos parsimoniosos y logre un adecuado equilibrio entre sencillez y buenos resultados. Sin embargo, no debe olvidarse que en un contexto de modelización el investigador no puede quedarse en la etapa de identificación de las variables significativas, sino que debe evaluar su utilidad en términos de la capacidad predictiva.

Es por eso que se puede decir que hay dos grupos de modelos identificados en la tabla 7.12, los suficientes y los insuficientes. El modelo suficiente implica que logra todos los cometidos planteados y, como podrá advertirse, se considera que tiene una buena capacidad predictiva, que en este caso se mide por el Pseudo-Rcuadrado, entre otros indicadores de ajuste. Por lo tanto, de los nueve modelos, solo tres pueden considerarse suficientes, lo que los vuelve recomendables. A continuación, se los compara.

Si bien dos de los tres son los del grupo nivel III y, por lo tanto, intentan modelar indicadores que tienen el componente de salud la comparación puede hacerse desde el lado de que variables explicativas son las que intervienen para llegar al mejor modelo para el modelo en sí y para la modelización de la heterocedasticidad.

A modo de resumen, se presenta la tabla 7.13 donde aparece que el nivel de CPO está presente en los tres modelos y lo mismo sucede con la edad. En el modelo 4, se ve que el componente P está asociado con el nivel educativo y muestra que, a mayor nivel, menor proporción de caries. A su vez, el hábito de fumar es relevante para marcar la heterogeneidad que aumenta la proporción de P. El modelo 6, que trabaja con la proporción de dientes sanos, muestra una asociación positiva con mayor ingreso y una relación inversa con la ingesta de alcohol y el hábito de fumar. La edad es la variable que modula la heterogeneidad.

Descripción de variables explicativas de los modelos						
Variable	Modelo 4		Modelo 6		Modelo 8	
	Modelo	Het	Modelo	Het	Modelo	Het
Nivel de CPO	Sí	Sí	Sí		Sí	Sí
edad en años	Sí	Sí		Sí	Sí	Sí
Sexo						
Nivel educativo	Sí				Sí	
Ingresos percibidos			Sí			
Nivel de consumo de alcohol			Sí			
Dias que consume bebidas azucaradas						
Fuma actualmente		Sí			Sí	

Tabla 7.13: Síntesis de los modelos considerados suficientes para el estudio RPA-FO2015

Por último, el modelo 8 muestra que la proporción del componente C se asocia de manera positiva con la edad y el CPO, algo esperable, mientras que los ingresos y el nivel educativo se asocian de manera negativa. La edad y el CPO son también los que aparecen en la heterogeneidad.

Antes de concluir, es importante destacar que no es que las otras proporciones para las que se estimaron modelos sirvan, sino que los modelos hallados son pobres, lo que obliga a pensar en cambiar, posiblemente, la función de enlace que, por motivos de extensión, en este capítulo no se trabajaron aunque se presentaron en la sección 7.1.2.

7.4. Conclusiones y futuros pasos

¿Cómo seguir? Hasta el momento, los resultados para los modelos de conteo trabajados mediante la RB, tienen resultados satisfactorios y complementarios a los hallados en el capítulo 6. Resta mejorar los resultados para los modelos que se clasificaron como insuficientes. Hay que tener en cuenta para el caso de la reparametrización que se usa para el MRB donde la variable de respuesta se transforma en una proporción, que aparece una dificultad extra, que podría llamarse efecto de Granularidad, ya que la variable de respuesta no puede tomar todos los valores de $(0, 1)$, pues cambian en incrementos de $\frac{1}{32}$, lo que puede ser un inconveniente a la hora de evaluar la predicción. En consecuencia, este aspecto, sumado con el de la posibilidad de usar otras funciones de enlaces, en lugar de la logit usada en este capítulo parecen ser los caminos a seguir. Se debe pensar si también es posible obtener mejorías estratificando o segmentando los datos mediante el uso de técnicas estadísticas como los CART.

Capítulo 8

Elaboración de perfiles epidemiológicos en estudios sanitarios mediante técnicas de clustering binario y análisis de redes

8.1. Introducción

Las enfermedades no transmisibles ENT, categoría que agrupa enfermedades como las cardiovasculares, diabetes, cáncer y enfermedades respiratorias crónicas, son hoy la causa de mortalidad más importante en el mundo. Cuentan el 63 % de las muertes globales, y casi el 40 % de ellas se producen entre los 30 y 70 años, lo que implica una carga social, dado que este es el tramo de edad de las personas económicamente activas. A su vez, la carga de este tipo de enfermedades en los países con bajos y medianos ingresos es del 86 % de las muertes prematuras (Skapino y Álvarez-Vaz, 2016). Este conjunto de enfermedades es, a su vez, responsable de un gran aumento de la discapacidad en varios países del mundo, donde en particular para los de menores ingresos, se producen a edades más tempranas, y esto se traduce en discapacidades por períodos más prolongados previo a la muerte.

Los estilos de vida que se relacionan con, por ejemplo, la alimentación inadecuada (World Cancer Research Fund International, 2007), (Cook et al., 2007), (Food and Agriculture Organization of the United Nations, 2010), (Bhupathiraju y Tucker, 2011), el sedentarismo, el consumo nocivo de alcohol y de tabaco juegan un papel importante en estas enfermedades. A su vez, actúan de forma directa o indirecta, creando otros factores de riesgo como la obesidad, los trastornos del metabolismo de los hidratos de carbono, la hipertensión arterial (HTA) y las dislipemias, (Skapino y Álvarez-Vaz, 2016). Este tipo de problema de la salud pública tiene un impacto muy grande en el desarrollo macroeconómico de los países, tal como consignan la OMS y el Foro Económico Mundial. Ambos sostienen que, de mantenerse estáticos los niveles de intervención y de continuar las cifras de ENT su ritmo de crecimiento, la pérdida económica acumulativa a causa de estas patologías en los países con ingresos medios y bajos superarán los U§S 7 trillones en el período 2011-2025. Modificar los estilos de vida es prioritario para la salud pública mundial y en ese sentido, se llevan a cabo programas preventivos que promueven una disminución importante en el número de muertes prematuras.

En los estudios epidemiológicos que investigan las ENT se trabaja con variables binarias que reflejan la presencia de determinadas enfermedades que, a su vez, están asociadas

con otro conjunto de enfermedades, denominadas comorbilidades. Estas también se miden a través de variables binarias y, en general, se asumen como factores de riesgo de las primeras. En el ámbito de los estudios epidemiológicos existen situaciones donde se manejan ENT, en particular en salud bucal, y ambos tipos de variables pueden ser intercambiables en cuanto a quién cumple el papel de factor de riesgo.

En este trabajo el objetivo es obtener perfiles epidemiológicos bien diferenciados en base a los atributos binarios, partiendo de un conjunto de variables, sin discriminar cuáles son variables de respuesta. Para esto, se propone la creación de grupos mediante dos estrategias:

1. usar un método de clustering basado en el algoritmo k-modes;
2. hacer análisis de redes sociales (SNA) a partir de las mismas variables construyendo la matriz de adyacencias sobre la cual aplicar una batería de métricas (closeness, betweenness, modularity, clustering) sobre los nodos y enlaces, que permite detectar comunidades.

El trabajo está organizado de la siguiente forma: en la sección 8.2 y 8.3 se presentan las técnicas a aplicar y en la sección 8.4 se presentan brevemente el problema en estudio y los datos usados. Luego, se muestran los resultados en la sección 8.5, se discuten los hallazgos en 8.6 y, para terminar, en 8.7 se presentan las conclusiones y futuros pasos.

8.2. Metodología A: clustering a través de algoritmo k-modes

Se usa parte de la metodología propuesta por (Tsekouras et al., 2005) para clasificar atributos categóricos a través del algoritmo mixtoFuzzy C-modes y usada por (Álvarez-Vaz y Massa, 2012) para encontrar perfiles de infección parasitaria en escolares de Montevideo. En ambos trabajos, cada individuo se clasifica con anterioridad con algún método de clustering y luego pasa por una etapa de difusión, donde pasa a pertenecer a más de un cluster con diferentes grados de participación o membresía. En el método original antes mencionado se utilizaba el algoritmo k-modes, que es de tipo modal y que no es más que un caso particular de un k-prototipo descrito por (Huang, 1997). En este caso el algoritmo tiene una lógica de funcionamiento similar a la del algoritmo k-means. Dada la naturaleza de las variables que son binarias, es necesario usar otras medidas de disimilaridad y método basado en frecuencias para actualizar los modos (Weihs et al., 2005).

Por lo tanto, del método mixto original de (Tsekouras et al., 2005) se trabaja solo con el algorimo k-modes, que aplica la siguiente disimilaridad: siendo x_i, y_i dos individuos de los que se miden los atributos,

$$d(x_i, y_i) = \sum_{1}^{m} \delta(x_j, y_j); \quad \delta(x_j, y_j) = \begin{cases} 0 & \text{si} \quad x_j = y_j \\ 1 & \text{si} \quad x_j \neq y_j \end{cases} \tag{8.1}$$

El algoritmo trabaja de la siguiente manera, a través de 4 pasos:

1. Selecciona k modos iniciales, uno para cada cluster con x, y variables categóricas binarias en este caso, actualizando el modo;

2. Calcula la cantidad de disimilitudes de cada individuo con dichos modos y lo agrupará con el centroide que presente menor cantidad de diferencias (mayor similaridad);

3. Luego de que todos los individuos han sido asignados, reestima la disimilaridad de los objetos contra el modo actual y si encuentra que un individuo que está más próximo del modo de otro grupo lo reasigna, actualizando los modos de ambos grupos que se modificaron;

4. Repite el paso 3 hasta que ningún individuo haya cambiado de cluster hasta haber visitado todo el conjunto de datos.

El resultado de este algoritmo es entonces una partición de los individuos en grupos cuyo represente es el perfil modal, es decir la combinación de respuestas que es más frecuente en cada cluster.

8.3. Metodología B: análisis de redes

En esta sección se presentan con brevedad las métricas que se usan para la caracterización de las redes sociales, que suele usarse en SNA. Para su presentación se seguirá la notación de del libro Statistical Analysis of Network Data with R (Kolaczyk y Csárdi, 2014), (Luke, 2015) aunque textos seminales como (Wasserman y Faust, 1994), (Borgatti et al., 2013) son una guía también.

Antes de presentar algunas de las métricas más relevantes para describir una red, es necesario definir conceptos básicos. Una red o grafo es una estructura matemática que está formada por dos tipos de elementos: nodos y enlaces. Los nodos pueden ser personas, variables o alguna otra entidad, mientras que los enlaces son las relaciones que existen entre los nodos. Se escribe como $G(V, E)$, donde V son los nodos y E los enlaces.

Si se observa la Figura 8.1, se ve que el nodo 8 está conectado con el nodo $3, 10, 5, 2, 9, 1, 6, 7$, mientras que el nodo 4 está con todos los nodos salvo el 8, como aparece más abajo.

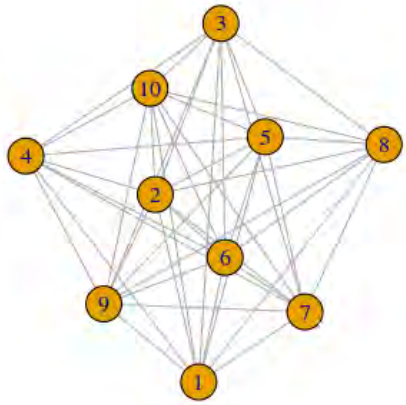

Figura 8.1: Ejemplo de red para diez personas

```
1--2  1--3  1--4  1--5  1--6  1--7  1--8  1--9
1--10 2--3  2--4  2--5  2--6  2--7  2--8  2--9
2--10 3--4  3--5  3--6  3--7  3--8  3--9  3--10
4--5  4--6  4--7  4--9  4--10 5--6  5--7  5--8
5--9  5--10 6--7  6--8  6--9  6--10 7--8  7--9
7--10 8--9  8--10 9--10
```

En el contexto de este trabajo, que se presenta en la sección 8.4, los nodos son personas y los enlaces surgen de saber si esas personas comparten ciertas variables, en este caso patologías y hábitos de vida.

Para entender la descripción que se hace del problema desde la perspectiva del SNA, es primordial presentar un conjunto de métricas que sirva para resumir la información y caracterizar la estructura de la red a través de lo que se conoce como la topología del grafo o de la red.

8.3.1. Grados de los vértices

Los grados d_v de un vértice v de un grafo $G(V, E)$ es el número de aristas en E incidentes sobre V. A partir de esta medida, se puede definir f_d como la fracción de vértices de $v \in V$ con grado $d_v = d$. El conjunto $\{f_d\} d \geq 0$ es lo que se llama distribución de grados de G.

Para las redes ponderadas, una generalización útil del grado es la noción de fuerza de vértice, que se obtiene sumando los pesos de los bordes de un vértice dado.

8.3.2. Centralidad de los vértices

Las medidas de centralidad de intermediación tienen por objeto resumir en qué medida un vértice se encuentra entre otros pares de vértices (Freeman, 1979), lo que se conoce como betweenness centrality o Centralidad de intermediación (BC)

$$c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{8.2}$$

$\sigma(s,t|v)$ es el número total de caminos más cortos entre s y t que pasan a través de v, y $\sigma(s,t)$ es el número total de caminos más cortos entre s y t (independientemente de si pasan o no por v). Esta medida de centralidad puede rescalarse al intervalo $[0,1]$ mediante un factor de $(N_v - 1)(N_v - 2)$ / 2, siendo N_v el número de vértices del grafo $G(V,E)$.

Las medidas de centralidad de proximidad intentan capturar la noción de que un vértice es central si está cerca de muchos otros vértices(Freeman, 1979), (Brandes, 2001). El enfoque estándar, introducido por (Sabidussi, 1966), implica dejar que la centralidad varíe inversamente con una medida de la distancia total de un vértice de todos los demás, lo que se conoce como closeness centrality:

$$c_{CL}(v) = \frac{1}{\sum_{u \in V} dist(v,u)} \tag{8.3}$$

$dist(v,u)$ es la distancia geodésica entre los vértices $u, v \in V$. Para comparar entre otras medidas de centralidad, esta medida se puede reescalar al intervalo $[0,1]$ a través de la multiplicación por un factor $N_v - 1$.

Por último, otras medidas de centralidad se basan en nociones de prestigio o rango. Es decir, buscan capturar la idea de que cuanto más centrales sean los vecinos de un vértice, más central es el vértice en sí mismo. Estas medidas pueden expresarse en términos de vectores propios de soluciones de sistemas lineales de ecuaciones y hay muchas medidas de centralidad de vectores propios.

De acuerdo a (Bonacich, 1987),(Bonacich y Lloyd, 2001):

$$C_{E_i}(v) = \alpha \sum_{\{u,v\} \in E} C_{E_i}(u) \tag{8.4}$$

El vector $C_{E_i} = (C_{E_i}(1), ..., C_{E_i}(N_v))^T$ es la solución al autovalor para $AC_{E_i} = \lambda^{-1} C_{E_i}$, donde A es la matriz de adyacencia para el grafo $G(V, E)$. Bonacich sostiene que una elección óptima de α^{-1} es el autovalor más grande de A y, por lo tanto, C_{E_i} es el autovector correspondiente. Cuando G es no dirigido, el valor propio más alto de A será simple y su autovector tendrá valores distintos de cero y del mismo signo.

8.3.3. Descripción de los enlaces

Se puede extender la idea de intermediación para los enlaces, aspecto que se denomina edge betweenness centrality o Centralidad de intermediación de enlaces (EBC) y que es una extensión de la intermediación de nodos que asigna a cada enlace un valor que refleja el número de caminos más cortos, shortest paths, que atraviesan ese enlace. Para otra medidas de centralidad que caractericen los enlaces se puede consultar (Brandes y Erlebach, 2005).

8.3.4. Cohesión de la red

Existen varias maneras de evaluar la cohesión de una red, dependiendo del problema, donde pueden usarse triádas o componentes gigantes o lo que se denomina cliques, que no son más que subconjuntos de nodos totalmente cohesivos, en el sentido de que todos los vértices dentro del subconjunto están conectados por enlaces. Se pueden definir cliques de tamaño 1, que en este caso son los nodos v, mientras que cliques de tamaño 2 representan los enlaces (e). Los cliques o subgrafos de tamaño 3 son lo que (Kolaczyk y Csárdi, 2014) también denomina triangles, de manera que, al ir aumentado el tamaño de los cliques, se puede ver cuál es la estructura del grafo bajo análisis. Por el proceso de construcción antes descrito, al aumentar el tamaño del clique, los últimos contienen los niveles más bajos, por lo cual (Kolaczyk y Csárdi, 2014) definen un concepto de clique máximo que surge de considerar un subgrafo que no es subconjunto de un clique mayor y que da una métrica extra que los autores denominan clique number que corresponde al tamaño del clique máximo.

8.3.5. Conectividad

Una noción de conectividad es la que tiene que ver con el hecho de que si, dado un subconjunto de k vértices (o enlaces) se quitan del grafo, el subgrafo restante aún permanece conectado. En particular, un grafo $G(V, E)$ se llama k-vértice-conectado si el número de vértices $N_v > k$ y el resultado de eliminar cualquier subconjunto de vértices $X \in V$ de cardinalidad $|X| < k$,X deja un subgrafo conectado. A su vez, si $G(V, E)$ se denomina k-borde-conectado si $N_v \leq 2$ y si, al eliminar cualquier subconjunto de aristas $Y \in E$ de

cardinalidad $|Y| < K$, deja un subgrafo que está conectado.

De esa manera se define como conectividad de vértice (enlace) de $G(V, E)$ al entero más grande, tal que G es k-vértice- (k-borde-) conectado. (Kolaczyk y Csárdi, 2014) manifiestan que se puede demostrar que la conectividad del vértice está acotada por la conectividad de enlace que, a su vez, está acotada por el grado mínimo d entre los vértices en G.

8.3.6. Clustering de la red

Cuando se habla de partición de la red $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ de un conjunto $\mathcal{S}$ se hace referencia a su división en clases naturales, tales que estas son disjuntas entre sí y, a su vez, la unión de ellas reproducen el conjunto de partida ($\bigcup_{k=1}^{K} C_k = S$). A su vez, es importante también evaluar si un subconjunto de nodos (algunas de esas clases) es cohesivo, para lo cual se entiende que es así si los nodos están bien conectados entre sí. Al mismo tiempo, están relativamente bien separados de los nodos restantes. Así, los algoritmos de particionado buscan una partición $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ de un grafo $G = (V, E)$, de manera que los conjuntos $E(C_k, C_{k'})$ de enlaces conectando nodos de C_k en $C_{k'}$ sea relativamente pequeña en comparación con el conjunto $E(C_k) = E(C_k$ o $E(C_{k'})$ de enlaces que conectan nodos al interior de C_k.

Una primera forma de evaluar el particionado de la red es a través de clustering jerárquico de tipo aglomerativo, donde se incorpora una función de costo, que refleja la cohesión, con lo cual surge el concepto de modularidad de $\mathcal{C}$, que define $f_{ij}(\mathcal{C})$ como la fracción de enlaces de la red original que conectan nodos de C_i con nodos de C_j:

$$mod(C) = \sum_{k=1}^{K} [f_{kk}(C) - f_{kk}^{*}]^2, \tag{8.5}$$

donde f_{kk}^{*} es el valor esperado de f bajo el supuesto de un modelo aleatorio de asignación de enlaces. Valores grandes de la modularidad sugieren que $\mathcal{C}$ captura una estructura no trivial de grupos (es decir, que existen grupos), a la inversa que si los enlaces se asignasen al azar.

8.3.7. Enlace selectivo (asortatividad)

Otro aspecto importante para evaluar la topología de una red es la evaluación de lo que se denomina enlace selectivo entre nodos, de acuerdo a algunas características. Se miden con lo que se conoce como coeficiente de asortatividad, que tiene una lógica muy similar a la de los coeficientes de correlación. Este concepto también se conoce, a veces, como homofilia, y expresa la tendencia de las personas a relacionarse con otras que se les parecen.

Cuando la característica que se estudia es de tipo categórico (nominal u ordinal) la medida homofilia es

$$r_a = \frac{\sum_i f_{ii} - \sum_i f_{i+} f_{+i}}{1 - \sum_i f_{i+} f_{+i}} \tag{8.6}$$

donde f_{ij} es la fracción de enlaces en $G(V, E)$ que unen un nodo en la i-ésima categoría con un nodo en la j-ésima categoría y f_{i+},f_{+i} expresan la suma de la i-ésima fila y columna respectivamente, de la matriz resultante $\mathbf{f}$ de frecuencias (Newman, 2002), (Newman, 2003).

El coeficiente descrito en la ecuación (8.6) está acotado en el intervalo $[-1, 1]$, expresando que si es cercano a 0, la mezcla de nodos en el grafo no difiere de la que se obtendría al asignar los enlaces al azar, preservando la distribución de grados marginal; cuando el coeficiente se acerca a 1 o -1 existe una mezcla selectiva perfecta.

Cuando los nodos tienen una característica de interés que es continua para evaluar la homofilia, se consideran como (x_e, y_e) los valores que toman los nodos enlazados por el enlace e, para lo cual se usa el coeficiente de correlación de Pearson de los pares (x_e, y_e):

$$r = \frac{\sum_{x,y} xy - (f_{xy} - f_{x+} f_{+y})}{\sigma_x \sigma_y} \tag{8.7}$$

8.4. Descripción del problema en estudio

Se trabaja con los datos del estudio RPAFO2015 con las personas que demandan atención en la Facultad de Odontología de la Universidad de la República durante el período 2015-2016. En particular, se consideran los siguientes atributos que conforman tres bloques de variables:

Variable	Descripción	Bloque	Tipo
V1	Fuma a diario	1	Comportamental
V2	Consumo nocivo de alcohol	1	Comportamental
V3	Actividad física insuficiente	1	Comportamental
V4	IMC sobrepeso/obesidad,	2	ENT
V5	Razón de Cintura-Cadera	2	ENT
V6	Hipertensión	2	ENT
V7	Diabetes	2	ENT
V8	Prev. bolsa	3	Odontológicas
V9	Pérdida dentaria	3	Odontológicas
V10	Prevalencia de caries	3	Odontológicas
V11	Prevalencia de PIP	3	Odontológicas

Tabla 8.1: Bloques de variables ENT utilizadas

Las primeras 3 variables constituyen factores de riesgo (bloque 1) que muchas veces

se asocian con las variables del bloque 2 que son las ENT, que son a su vez patologías (enfermedades) y que también son factores de riesgo a su vez para las variables del bloque de patologías odontológicas (bloque 3).

Variable	Descripción	Prevalencia
V1	Fuma a diario	33,1
V2	Consumo nocivo de alcohol	9,8
V3	Actividad física insuficiente	44,7
V4	IMC sobrepeso/obesidad	57,3
V5	Razón de cintura/cadera	56,3
V6	Hipertensión	43,2
V7	Diabetes	21,3
V8	Presencia bolsa	58,6
V9	Pérdida dentaria	59,6
V10	Prevalencia de caries	72,8
V11	Prevalencia de PIP	63,6

Tabla 8.2: Prevalencias de variables estudiadas

Para el análisis global se trabaja con el (R Core Team, 2016), para la determinación de los clusters con el algoritmo presentado en la metodología A 8.2 se usa la librería `k-modes` (Weihs et al., 2005). A través de la función k-modes de determinan los clusters. Para el análisis de los datos desde la perspectiva de redes se trabaja con la librería igraph (Csardi y Nepusz, 2006).

La definición de presencia o ausencia de los factores de riesgo del bloque 1 y del bloque 2 siguen la lógica de la encuesta nacional de factores de riesgo de Uruguay. En particular, se usa un nuevo indicador llamado razón de cintura (V5) que compara la circunferencia de cintura para hombres y mujeres contra valores de referencia máximos, por lo cual razones de cintura mayor a 1 indican patología. En cuanto a las patologías bucales, se trabaja con la presencia de caries (V10), la pérdida de inserción de las piezas, también considerada como prevalencia de PIP (V11), la pérdida dentaria (V9), que en este caso representa falta de dentición funcional. Por último, la presencia de bolsa es un indicador relativo a la patología de la mucosa en cada sextante en los que se divide la boca.

Se presentan los umbrales de corte para cada variable:

- Alcohol: se considera consumo nocivo a las categorías alto y muy alto.
- Actividad física insuficiente: la realización de menos de 75 minutos por semana de actividad física vigorosa o menos de 150 minutos por semana de actividad física moderada.
- Hipertensión arterial una medición de presión arterial sistólica (PAS) mayor o igual

a 140mmHg o presión arterial diastólica (PAD) mayor o igual a 90 mmHg o el autorreporte de ser hipertenso.

- Diabetes/glicemia alterada en ayunas (GAA) a quienes tenían glicemia capilar mayor o igual a 200 mg/dl con o sin ayuno o antecedentes de diabetes diagnosticada por su médico y proporcionado a través del autorreporte del paciente.
- Razón de cintura ($M > 94$, $F > 80$).
- Bolsa $\Longrightarrow > 3mm$.
- PIP $\Longrightarrow > 4mm$ en algunos de los sextantes.
- Pérdida Dentaria $\Longrightarrow <$ 20 dientes.

8.5. Resultados

8.5.1. Análisis con k-modes

La tabla de datos que se considera de RPAFO2015 y que se resume en la tabla 8.2 deja de lado una observación que se caracteriza por ser una persona que tiene 0 en las 11 variables, lo que estaría representando a una persona que está en principio sana y no presenta factores de riesgo. Recordando que el estudio era sobre personas que demandan atención, puede resultar raro considerarla, ya que el resto es gente que está enferma o tiene alguno de los factores de riesgo. En este caso, se trata de una persona que tiene otras patologías bucales, que no son las que se consideran en el bloque 3 de la tabla. Por otra parte, para hacerlo comparable con el análisis de redes que se presenta más adelante, es necesario dejarlo de lado, ya que sería un nodo de la red que estaría desconectado de los 601 encuestados restantes. Por último, el no considerar esta observación no altera las prevalencias reportadas en la tabla 8.2.

En la tabla 8.3 se puede ver el comportamiento cambiando de cantidad de clusters, para lo cual se definen cuatro escenarios en los que se presentan varias métricas.

Métricas para Descrición de los grupos					
Escenario 1 (2 clusters)	1	2			
tamaño	439	162			
withindiff	1548	462			
densidad	3,52	2,85			
Escenario 2 (3 clusters)	1	2	3		
tamaño	265	113	223		
withindiff	707	265	525		
densidad	2,66	2,34	2,35		
Escenario 3 (4 clusters)	1	2	3	4	
tamaño	245	106	204	46	
withindiff	634	238	459	101	
densidad	2,58	2,24	2,25	2,19	
Escenario 4 (5 clusters)	1	2	3	4	5
tamaño	219	76	198	44	64
withindiff	534	166	438	94	132
densidad	2,43	2,18	2,21	2,13	2,06

Tabla 8.3: Caracterización de los clusters mediante algoritmo k-modes

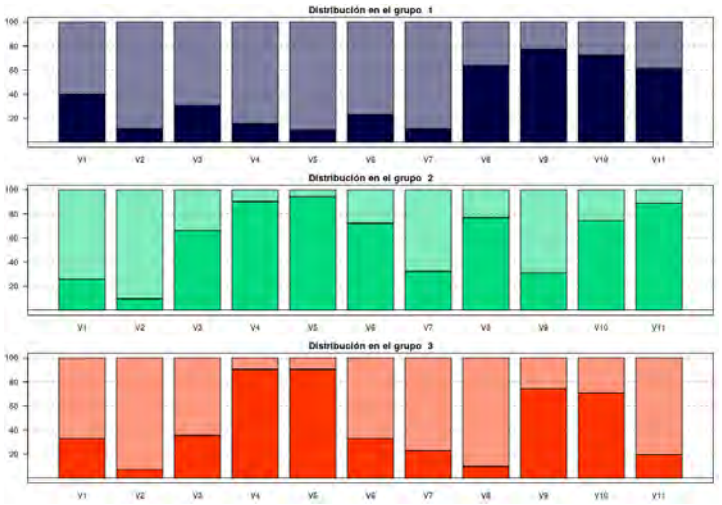

Figura 8.2: Prevalencias de las variables en cada grupo (3 grupos)

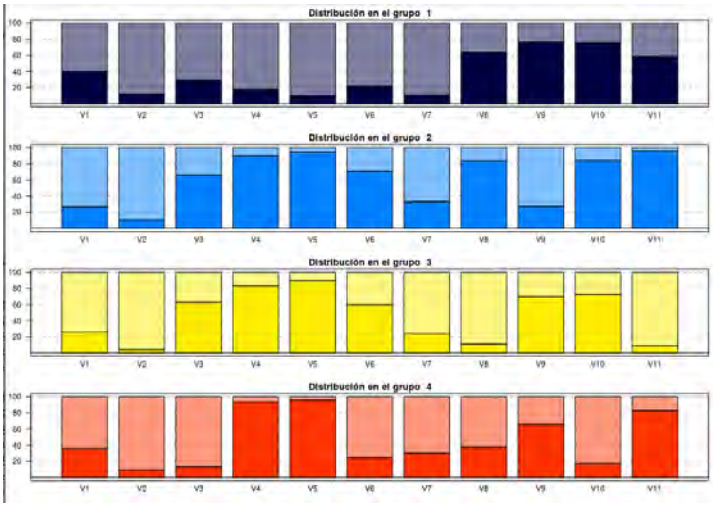

Figura 8.3: Prevalencias de las variables en cada grupo (4 grupos)

Se considera el tamaño en cada grupo y la métrica withindiff que expresa la distancia de simple matching intracluster, que expresa la cantidad de desacuerdos para todas las variables dentro de cada cluster. También se construye una métrica que combina las dos primeras y que se expresa como densidad, relativizando el total de desacuerdos con respecto al total de observaciones dentro de cada cluster. Esta última es un indicador de homogeneidad interna, y se puede ver cómo varía y decrece en algún cluster al cambiar la cantidad de grupos que se establece como restricción. Analizando la densidad se puede ver que esta decrece a medida que hay más cantidad de grupos, con lo cual los escenarios 2 y 3, que consideran 3 y 4 grupos, parecen ser una buena solución.

En la figura 8.2 y figura 8.3 se ve cómo es la prevalencia de las variables a la interna de cada bloque para el escenario 2 y 3 respectivamente, puesto que la opción de cuatro grupos es la mejor y los grupos muestran en promedio la cantidad de desacuerdos intracluster es más baja (densidad más baja) que para el caso del escenario 2 con 3 grupos.

Se puede apreciar el perfil modal de cada grupo para las soluciones planteadas en la tabla 8.4.

En la tabla 8.5 se presenta la prevalencia en cada grupo para los tres bloques de variables y que se puede comparar con la prevalencia global de cada variable para la opción de tres grupos, donde cada grupo es G(1), G(2), G(3).

En la tabla 8.6 se hace una caracterización de los grupos en función de atributos personales y, por otra parte, la carga de variables intragrupo (cantidad de respuestas afirmativas en las variables de los bloques 1, 2 y 3 de la tabla 8.2).

Perfil modal											
Escenario k=3 clusters											
Grupo	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	no	no	no	normal	normal	no	no	si	si	si	si
2	no	no	no	sobrepeso	alterada	no	no	no	si	si	no
3	no	no	si	sobrepeso	alterada	si	no	si	no	si	si
Escenario k=4 clusters											
Grupo	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	no	no	no	normal	normal	no	no	si	si	si	si
2	no	no	no	sobrepeso	alterada	no	no	no	si	si	no
3	no	no	si	sobrepeso	alterada	si	no	si	no	si	si
4	si	no	si	sobrepeso	alterada	no	no	si	no	si	si

Tabla 8.4: Perfil modal de las variables en cada grupo

Prevalencia de cada variable				
	G(1)	G(2)	G(3)	Total
Fuma a diario	39,5	25,6	32,7	33,1
Consumo nocivo de alcohol	11,3	9,4	7,1	9,8
Actividad física insuficiente	30,8	65,9	35,4	44,7
IMC sobrepeso/obesidad	15,8	90,1	90,3	57,3
Razón de cintura/cadera	10,2	94,2	90,3	56,3
Hipertensión	23,3	72,2	32,7	43,2
Diabetes	11,3	32,3	23,0	21,3
Presencia bolsa	63,9	77,1	9,7	58,6
Pérdida dentaria	77,4	30,9	74,3	59,6
Prevalencia de caries	72,2	74,4	70,8	72,8
Prevalencia de pip	61,3	88,8	19,5	63,6
Tamaño	265	113	223	601

Tabla 8.5: Perfiles de los grupos creados mediante k-modes

Perfiles de los grupos				
Total de respuestas (SI)	G(1)	G(2)	G(3)	Total
1	13	0	0	13
2	24	2	0	26
3	51	12	1	64
4	69	34	14	117
5	51	30	39	120
6	38	23	52	113
7	16	11	54	81
8	3	1	40	44
9	0	0	20	20
10	0	0	3	3
Sexo, edad				
Femenino	124	78	149	351
Masculino	141	35	74	250
Edad promedio	37,8	41,5	55,5	45
Edad mediana	35	39	56	44
Ingreso				
0 a 17000	141	62	122	325
17000 a 24000	40	21	48	109
24000 a 32000	46	12	21	79
32000 a 40000	8	9	15	32
màs de 40000	12	4	10	26
No sabe	9	1	5	15
Rehúsa	9	4	2	15
Total	265	113	223	601

Tabla 8.6: Asociación de los clusters con algoritmo k-modes con características sociodemográficas

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	Observados
0	0	1	1	1	1	0	1	0	1	1	12
0	0	1	0	0	0	0	0	1	1	0	10
0	0	0	0	0	0	0	0	1	0	0	8
0	0	0	0	0	0	0	0	1	1	0	8
0	0	0	0	0	0	0	1	1	1	0	8
0	0	0	1	1	1	1	1	0	1	1	8

Tabla 8.7: Patrones de respuestas más frecuentes

Por último, se presentan los patrones de respuesta más frecuentes de un total de 313 observados de 2048 posibles (2^{11})

8.5.2. Análisis con SNA

A continuación se presenta el análisis de los datos mediante SNA. Para la detección de comunidades se usan varias librerías (Butts, 2016), (Csardi y Nepusz, 2006), (Kolaczyk y Csárdi, 2017).

Al trabajar con los datos desde la perpectiva del SNA se construye un grafo, al cual se le asocia la matriz de adyacencias, sobre la que se crean las comunidades mediante los algoritmos Random Walk y Fast Greedy. La matriz de adyacencias surge de considerar los nodos como las 601 personas del RPAFO2015 y se establece que 2 nodos están conectados si comparten alguna de las 11 variables que se usaron al crear los grupos en la sección 8.5.2. En las diferentes figuras la representación del grafo y la posición de los nodos es siempre la misma. Para eso se usa un layout aleatorio, pero se usa la misma semilla para iniciar la visualización para tener siempre en el mismo lugar los mismos nodos y poder hacer que las figuras sean comparables.

Los 601 individuos considerados en la sección 8.5.1, al quitar el único individuo que quedaría aislado, se genera un grafo conectado, que se describe con base en métricas, como los grados y la frecuencia con las que se da cada patrón, al igual que medidas de centralidad.

En la tabla 8.8 se puede ver que la configuración que aparece con más frecuencia es la que corresponde al nodo 206, con 12 nodos iguales que tiene un alto número de enlaces y que pertenece al cluster 3 para la solución de 3 grupos y para el de 4 grupos. Se trata de un nodo tipo que tiene presente las 3 patologías del tipo (ENT) del bloque 2 y 2 de las 4 bucales. A este perfil del nodo se le antepone el nodo 97 (que es uno de los dos con esta configuración) y que está mucho menos conectado y se caracteriza por ser una persona que solo presenta el hábito de fumar. El resto de las filas de la tabla 8.8 muestra otras configuraciones que corresponden a nodos con mayor número de enlaces (mayor valor de la variable grados), pero que aparecen menos veces. Esto verifica que están asignados a diferentes clusters. Una característica común a los seis nodos es que todos los individuos presentan las cuatro patologías bucales.

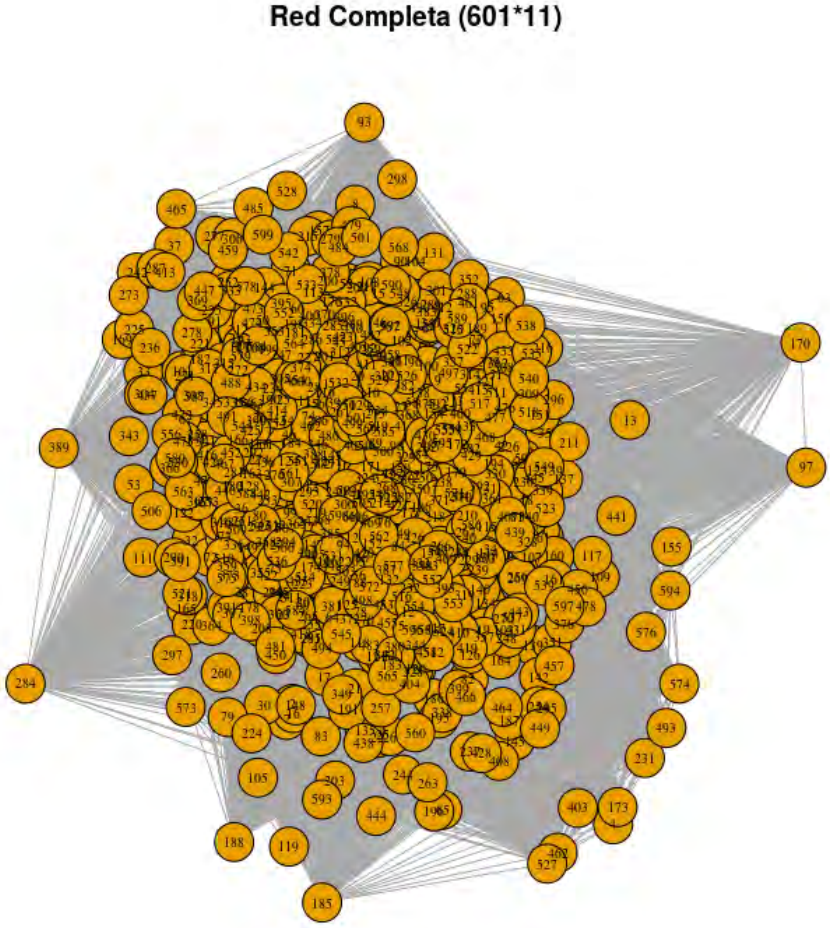

Figura 8.4: Red generada con 601 individuos analizados

Perfiles por frecuencia nodo	patrón	k3	k4	grados	frecuencia
206	0-0-1-1-1-1-0-1-0-1-1	3	3	586	12
97	1-0-0-0-0-0-0-0-0-0-0	1	1	198	2
Perfiles con más grados					
205	1-0-1-0-1-1-0-1-1-1-1	3	3	600	13
175	1-0-1-1-1-0-0-1-1-1-1	3	4	600	13
281	1-0-1-1-1-1-0-1-1-1-1	3	3	600	13
506	1-0-1-1-1-1-1-1-1-1-1	3	3	600	13
324	1-1-1-1-1-0-0-1-1-1-1	3	4	600	13
171	1-1-1-1-1-1-0-1-1-1-1	3	3	600	13

Tabla 8.8: Descripción de algunos nodos

Sobre el grafo se aplican dos algoritmos de detección de comunidades. La modularidad es mayor para la solución de dos grupos que surge del algoritmo de random walk.

Luego, sobre la red creada se proyectan los grupos creados mediante el algoritmo de k-modes, que eran 3 contra los 4 clusters asimilados a las comunidades detectadas.

	Algoritmo	(Cluster	Fast Greedy)
Algoritmo Cluster Random Walk	1	2	Total
1	0	177	177
2	76	0	76
3	146	3	149
4	54	145	199
Total	276	325	601

Tabla 8.9: Comparación de los dos métodos de detección de comunidades

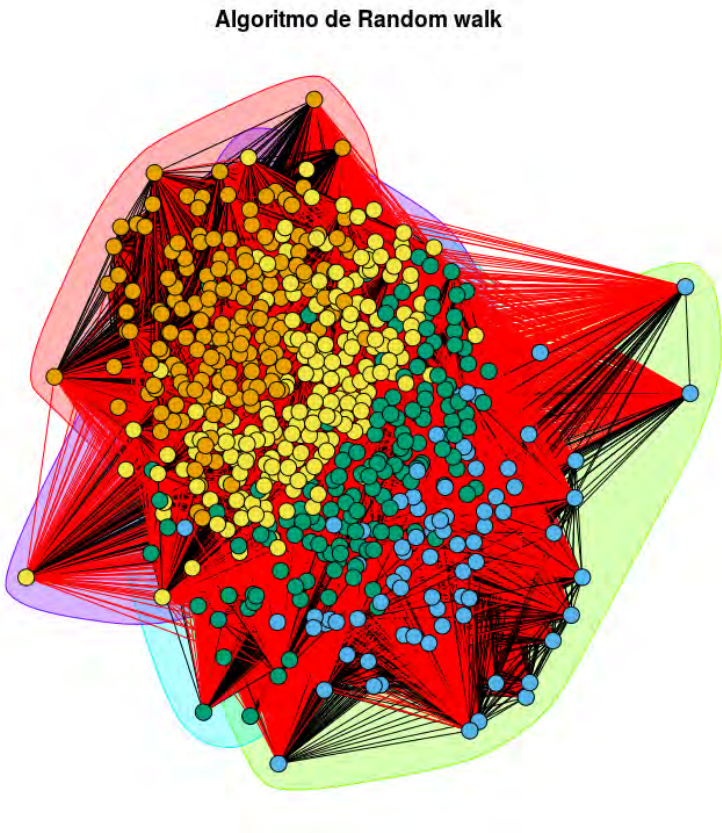

Figura 8.5: Comunidades identificadas con algoritmo Random Walk

	(Cluster		k-modes)	
Cluster Algoritmo Random Walk	1	2	3	Total
1	132	1	44	177
2	0	76	0	76
3	72	76	1	149
4	83	13	103	199
Total	287	166	148	601

Tabla 8.10: Relación entre las clusters creados mediante los algoritmos Random Walk y k-modes

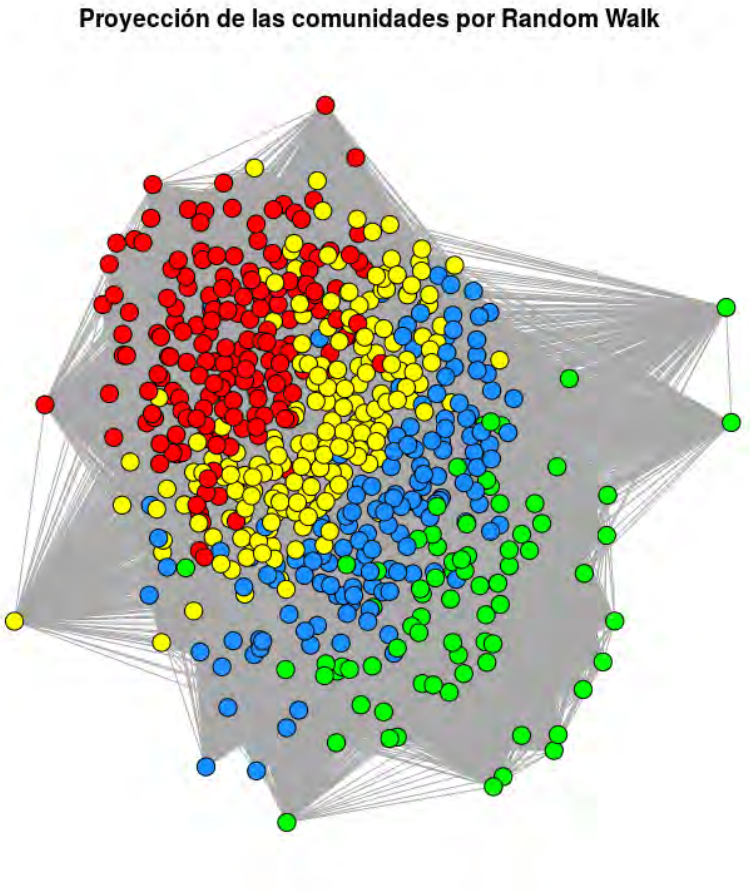

Figura 8.6: Proyección en la red de los grupos encontrados con algoritmo Random Walk

			(Método	Random walk)		(Método Fast Greedy)	
Variable	1	2	3	4	1	2	Total
V1	14,7	44,7	44,5	36,1	50,7	18,1	33,1
V2	7,9	9,2	12,1	10,0	10,8	8,9	9,8
V3	55,3	27,6	38,9	46,2	38,0	50,4	44,7
V4	91,5	0,00	10,0	84,4	18,8	90,1	57,3
V5	96,0	0,00	4,0	81,9	15,6	91,8	56,3
V6	67,2	5,3	29,5	47,7	25,4	58,4	43,2
V7	35,6	0,00	9,4	25,6	9,0	31,6	21,3
V8	58,7	21,0	78,5	58,3	60,1	57,5	58,6
V9	7,8	94,7	71,8	83,4	81,1	41,5	59,6
V10	59,9	71,0	78,5	80,9	80,4	66,6	72,8
V11	80,1	3,9	79,8	59,8	55,8	70,5	63,6
Tamaño	177	76	149	199	276	325	

Tabla 8.11: Perfiles de los grupos creados mediante SNA

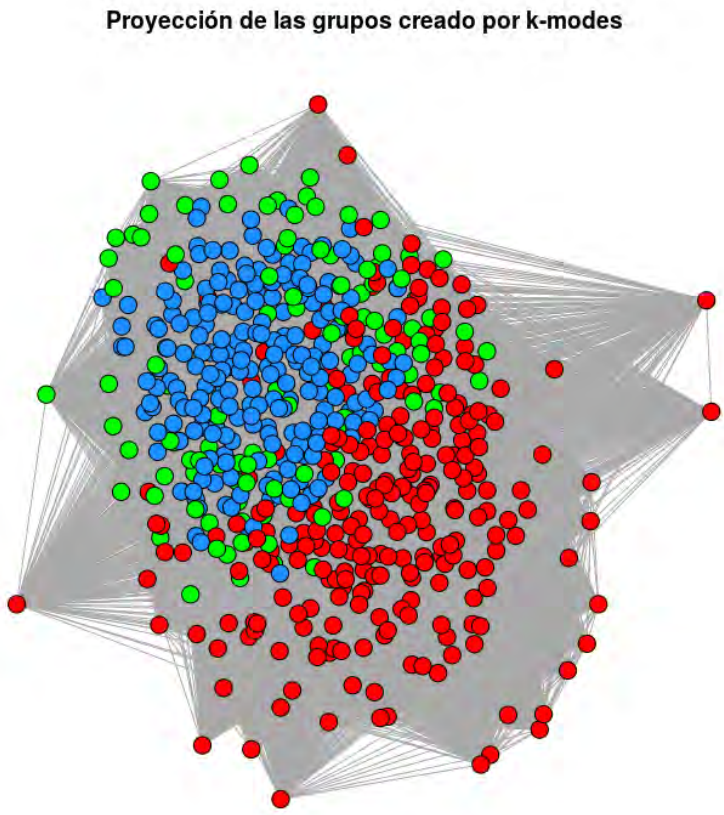

Figura 8.7: Proyección en la red de los grupos encontrados con algoritmo k-modes

8.6. Discusión

Teniendo en cuenta los clusters creados con la metodología de cluster por el método k-modes, observando los resultados de la tablas 8.6 y 8.5 se puede decir que se establece una agrupación con las siguientes características:

- G(1) con un total de $n = 265$ individuos que presentan poca actividad física y una prevalencia de consumo de tabaco y alcohol superior a la media, alta prevalencia de enfermedad periodontal (bolsa y PIP), mientras que tiene un perfil por debajo del promedio en términos de la prevalencia de las variables del bloque 2 (ENT). A su vez, es el grupo en el que la proporción de hombres es mayor a la media, más joven y con un perfil de ingresos similar a la distribución global;
- G(2) con menos individuos $n = 113$, que tiene baja proporción de fumadores y baja prevalencia de pérdida dentaria, con prevalencia menor de caries y enfermedades periodontales y, a su vez, con prevalencias de las ENT muy aumentadas con respecto al promedio. Es un grupo en el que la proporción de mujeres duplica la de hombres, que tienen una edad promedio por debajo de la media global, pero por encima de la del grupo 1. El ingreso muestra un perfil similar al promedio general.
- G(3) con $n = 223$ individuos, en el que muestran una alta proporción de conductas comportamentales negativas (que estaría dejándolo con un grupo con más perfil de riesgo). Sin embargo, las ENT están fuertemente presentes, salvo la HTA y donde solo se ve incrementada la pérdida dentaria, mientras que el resto de la patología odontológica es menor.

Antes de pasar a la descripción de los individuos con la metodología de redes, es importante tener en cuenta la participación o carga de variables con presencia de respuestas afirmativas (que debe interpretarse como más factores de riesgo, más ENT o más patología bucal). Un análisis de ese indicador muestra un trasiego de grupo con un aumento monótono, en el que se puede decir que el G(1) es el grupo que se caracteriza por tener menos carga de patologías (variables del grupo 2 y 3) o de hábitos inadecuados, mientras que el G(3) se caracteriza por tener mucha concentración de respuestas SI, lo que se puede entender como un grupo con más carga de enfermedad y de factores de riesgo.

Con respecto a la caracterización de los grupos que surgen de las comunidades detectadas con el SNA, se comentan los siguientes hallazgos con respecto al método Fast Greeedy:

- El grupo 1 tiene un total de $n = 276$ individuos que muestran mayor prevalencia de las variables comportamentales (lo que supone un aumento de los factores de riesgo, salvo para la actividad física), mientras que muestra tener una disminución sostenida de la prevalencia de las variable ENT. En cuanto a las variables del bloque 3, los individuos del grupo 1 muestran un perfil con más carga de patología odontológica.

- El grupo 2 tiene más individuos, $n = 325$, y muestra una disminución del consumo de tabaco y de alcohol, mientras que la prevalencia de actividad física inadecuada está por encima del promedio. A su vez, el grupo se caracteriza por tener una alteración de las ENT muy por encima de los valores medios y un comportamiento en cuanto a las variables odontológicas de menor prevalencia (están menos enfermos).

A su vez, las figuras 8.5 y 8.6 muestran las comunidades detectadas, claramente diferenciadas. El cluster 1 es el rojo, el 2 el verde, el 3 el azul y el 4 el amarillo. Cuando se analiza la proyección en la red de los clusters creados con el método k-modes el cluster 1 aparece bastante diferenciado del cluster 2, más solapado con el cluster 3.

Como todo problema de clustering, este no escapa a la situación en la que no se sabe el número de clusters con exactitud, sino más bien aproximaciones a un número óptimo. Teniendo en cuenta, a su vez, que para el caso de la metodología A se usa el algoritmo k-modes (equivalente del k-means pero para variables categóricas), no es de extrañar que los individuos una vez clasificados en los clusters muestren perfiles que se alejan de los centroides de los grupos (que son los perfiles modales), a pesar de que están más cerca de estos que si estuviesen en otros clusters. Esta situación puede mejorarse extendiendo el método a uno mixto mediante un proceso de difusión, donde cada individuo puede extenderse a otros clusters con diferentes grados de membresía y cada individuo pertenece a más de un cluster con diferentes grados de participación, partiendo previamente de una clasificación previa, que en este caso es el algoritmo k-modes. Es importante advertir que el fenómeno de difusión que podría darse en los clusters creados con el algoritmo k-means puede verse acrecentado por la propia naturaleza de las variables binarias en las que la variabilidad intracluster que surge al aplicar el algoritmo k-means está mucho más discretizada.

8.7. Conclusiones

Con los resultados encontrados hasta el momento aplicando estas metodologías, se puede decir que ambas detectan distintas cantidades de grupos bien diferenciados. Tienen una explicación en el contexto del problema para el caso de la partición encontrada mediante k-modes, que es bastante difusa, mientras que la partición creada mediante el SNA se muestra más estable.

A su vez, tal como se dice en la introducción, el objetivo fue plantear alternativas de elaboración de perfiles mediante diferentes metodologías, pero que parte de una estrategia metodológica particular y que consiste en no tener en cuenta para nada la jerarquía en las variables al ponerlas a todas en los tres bloques en igualdad de condiciones para segmentar la población bajo estudio. La literatura muestra que, en general, la aproximación es de tipo modelizante, donde hay claramente un bloque de variables explicadas, que sería el bloque 3 de las variables odontológicas, las que pueden ser explicadas por las variables comportamentales (bloque 1), que configuran factores de riesgos y las variables ENT

(bloque 2) y que son de por sí ya patologías y a su vez factores de riesgo. Por eso se proponen varios caminos que puedan ayudar a entender mejor las tipologías resultantes mediante la metodología de k-modes y la de detección de comunidades mediante SNA, que se plantean en el capítulo 9.

Parte III

Discusión y conclusiones generales

Capítulo 9

Conclusiones

9.1. Consideraciones sobre la parte I del libro

En los capítulos de la parte I se introduce al especialista en biomedicina y se estudian la justificación y la necesidad de elaborar una sistematización en el ámbito de la salud bucal como subdisciplina de la biomedicina. Por último, se expone la metodología necesaria para lograrlo.

El capítulo 1 presenta una introducción acerca de los motivos en salud pública por los cuales es necesario conocer información sistematizable a través de fuentes de datos y estrategias metodológicas para intervenir y modificar situaciones. A partir de esa información sistematizada se puede plantear hacer VE, donde juegan un rol fundamental los indicadores que la resumen, pero que deben cumplir determinados requisitos para hacerlos comparables y reproducibles. Sin embargo, esos mismos indicadores a menudo adolescen de ser limitados en su potencial de análisis, en virtud de su operacionalización en el intento de facilitar sus cálculos. Este es el disparador del objetivo general y de los objetivos específicos que se plantean al final del capítulo, buscando elaborar una tipología de tres tipos de indicadores: los Indicadores alternativos, los Indicadores combinados y los Indicadores espacio-temporales. Esa tipología se le va presentando al especialista en biomedicina a lo largo de este libro preservando un aspecto relevante: aprovechar al máximo toda la información que cada uno de estos contenga, es decir nunca perder la estructura multivariada aunque eso implique tener que trabajar con indicadores e índices más elaborados, pero con otras características. Esto no consiste en pedir más información, sino en tratarla de forma diferente, usando técnicas estadísticas tal vez un poco más sofisticadas y, en muchas casos, desconocidas en este ámbito.

El capítulo 2 presenta una síntesis de las patologías en salud bucal más relevantes para ser tratadas desde la vigilancia epidemiológica, la caries, la enfermedad periodontal, la erosión, la maloclusión, los trastornos temporomandibulares y las lesiones de mucosa. En ese capítulo se resume el alcance y la importancia de hacer vigilancia sobre estas patologías para el investigador en biomedicina no especialista en salud bucal. A su vez, los indicadores tradicionales y sus algoritmos de cálculo tal como se presentan son los que justifican la necesidad de darles otra mirada con la ayuda de técnicas estadísticas más modernas y abarcativas y que potencian el análisis de las patologías.

El capítulo 3, el más extenso de esta primera parte, pone en conocimiento del investigador biomédico el conjunto de técnicas estadísticas mínimas que no debería dominar, sino conocer, para poder resolver una gran cantidad de problemas de la biomedicina, que se trabajan en la mayoría de los artículos de las revistas relevantes.

El conjunto de técnicas se va presentando como solución a los tipos de indicadores consignados en la secciones 1.2 y 1.3 y abarcan los modelos de regresión (discreto y continuo), muy usados en biomedicina. Aparecen variantes, como los modelos de conteo (dada la naturaleza de las variables), que obligan a recurrir a modelos más elaborados, que precisan supuestos más restrictivos como los modelos de conteo MCont. Estos se introducen muy brevemente para que el lector se familiarice con ellos, dado que luego, en las aplicaciones de la parte II, se ven más en profundidad su fundamento y, sobre todo, su aplicación. En ese intento gradualista de ir mostrando los demás tipos de indicadores como los Indicadores combinados es que se presentan técnicas estadísticas que funcionan combinando muchos indicadores básicos (de eso su nombre) y donde surge el análisis factorial y el de clustering. Se presentan mínimamente y se profundizan en los apéndices A.

Frente a la necesidad de medir mejor el mismo fenómeno, surge la necesidad de generar Indicadores alternativos. Se lleva a cabo una simple transformación de los datos trabajando con tasas o proporciones, lo que lleva a considerar una clase de modelos no tan extendidos en el ámbito de la biomedicina —como los modelos de RB que, como extensión de los MLG, son modelos probabílisticos para ajustas tasas que permiten relajar las restricciones que estos consideran, al tener recorrido truncado, y en donde no se puede trabajar con los clásicos modelos de RLin. Como última situación de indicadores alternativos se le presentan al lector los índices derivados de indicadores convencionales que forman parte de los índices basados en teoría de la información, provenientes de la economía y la matemática, de gran utilidad para el análisis de la desigualdad. Se presentan con mucho de detalle en En el capítulo 11 de la tesis de doctorado (Álvarez-Vaz, 2020), con una importante aplicación en una encuesta poblacional.

Para terminar la presentación de la tipología de indicadores se completa con la que aparece en la sección 4, donde esta clase de indicadores Indicadores espaciotemporales se divide en tres de acuerdo a las dimensiones que intentan considerar:

Agregaciones espaciales: interesa ver si en una determinada localización geográfica aparece o un patrón espacial que pueda considerarse no aleatorio.

Agregaciones temporales: se pueden utilizar indicadores que buscan decidir si el número o proporción de casos de la patología en estudio que aparece en intervalos de tiempo consecutivos suceden con una frecuencia diferente a la esperada si se tratara de una distribución aleatoria.

Agregaciones espaciotemporales: se busca encontrar si existen clusters en el espacio y en el tiempo en simultáneo, pensando entonces que exista interacción. En este caso,

la interacción equivale a suponer que los casos cercanos en el espacio son, además, cercanos en el tiempo, lo que obliga a considerar que la localización de un caso depende de la localización del caso que lo precede.

Luego de presentada la tipología de indicadores y sus respectivas técnicas, el autor de este libro considera como aspecto complementario el herramental básico a ser conocido y tomado en cuenta por el investigador en biomedicina. Se presentan en las dos secciones que siguen y que plantean algunos aspectos relativos al muestreo de poblaciones finitas que en el ámbito de la epidemiología muchas veces se debe tener en cuenta.

Por un lado, se presentan los recaudos a tener en cuenta para poder trabajar de manera adecuada en función del mecanismo de generación de los datos, lo que suele hacerse mediante muestras probabilísticas que garantizan la validez de los métodos estadísticos que puedan usarse. Es así que surge el impacto al trabajar con diseños muestrales complejos (como alternativa a los procedimientos convencionales de muestreo aleatorio simple), con modificaciones en la estimación de la varianza, en el tamaño de muestra necesario y demás consideraciones, como el tratamiento de la no respuesta, que se amplian apenas en parte en el apéndice A, pero que se detallan con gran cantidad de referencias bibliográficas.

Un último aspecto a tener en cuenta por el lector es el que tiene que ver con la necesidad de ajustar los indicadores usando información auxiliar, cuando en la práctica cualquiera de los índices presentados en las secciones anteriores debe ser modificados porque se violan los supuestos en los que están basados, por ejemplo, al trabajar con muestras probabilísticas, en las que no se verifica la independencia entre observaciones. Para eso una forma de corregir esos aspectos es trabajar con Multiniv.

9.2. Consideraciones sobre la parte II del libro

La segunda parte del libro es donde se ve el potencial de la tipología de indicadores presentados y las técnicas estadísticas usadas, que en la producción biblográfica en la que participa el autor en coautoría en varios artículos de revistas que se referencian no se ve, dada la especificidad de estas publicaciones (revistas clínicas o epidemiológicas del ámbito de la odontología). Es por eso que se hace énfasis en los detalles más técnicos y estadísticos.

Esta segunda parte se divide en cinco capítulos, uno de los cuales presenta con mucho detalle las tres fuentes de datos a ser tratadas en los capítulos siguientes. El capítulo 4 presenta tres estudios que fueron llevados adelante por docentes de la Facultad de Odontología y del IESTA de la Facultad de Ciencias Económicas y de Administración de la Universidad de la República. Hoy son la información epidemiológica más importante a escala poblacional de las patologías bucales presentadas en el capítulo 1. El autor participó en todas ellas con un papel fundamental en el diseño metodológico y en la producción

bibliográfica que de ellas surge hasta la fecha. Dos de ellas son encuestas poblacionales en población joven y adulta PRNSB2011 y en escolares de 12 años de Montevideo RACA2012, respectivamente, con diseños muestrales complicados que justifican su uso en solo dos de las cinco aplicaciones que se detallan a continuación.

Antes de resumir los hallazgos más relevantes en cada capítulo de esta parte II, las lectoras y los lectores deben recordar que solo se trataron algunas de las patologías presentadas en el capítulo 1, como el CPO y sus componentes y la enfermedad periodontal. Por otra parte, la elección de la patología y sobre qué estudio trabajar no es caprichosa, ya que, como se advirtió en la sección 3.6, el diseño muestral limita el uso con los debidos recaudos de muchas técnicas estadísticas.

El capítulo 5 trata sobre el impacto de trabajar con una variable de conteo, en este caso la caries, que se dicotomiza para un determinado umbral. Se trabaja con modelos predictivos, pero con la pérdida del gradiente de la enfermedad. Como se trabaja con los datos del PRNSB2011, hay problemas de diseño muestral que dificultan el análisis, con una característica: el conteo de caries tiene problemas severos al presentar una muy importante sobredispersión y, a su vez, un exceso de 0 extremadamente grande, lo que se traduce en que el modelo a usar no sea el de Poisson, sino cualquiera de los otros modelos de conteo. Lamentablemente, dado el diseño muestral complejo, con el software usado solo es posible usar la RPoi. Además, el autor muestra lo que sucede con una aplicación incorrecta de un modelo de Poisson sobre la variable con recurrido truncado para facilitar la interpretación en términos de razón de prevalencias, pero donde se llegan a coeficientes muy diferentes si se usara el modelo de conteo adecuadamente como contraejemplo.

En el capítulo 6, se trabaja con el estudio RPAFO2015, que tiene la ventaja de tener un diseño muestral autoponderado, lo que le permite al especialista en biomedicina manejar modelos de conteo más complicados con total tranquilidad. A lo largo de este capítulo se van presentado diferentes distribuciones de probabilidad válidas para los modelos de conteo, como la MBN, la PIG, la PG y la DP. Luego de su presentación teórica se aplican C, P y O del CPO para ajustar los componentes. Dado que persiste un comportamiento patológico (en términos estadísticos) de las variables utilizadas, se opta por presentar otra serie de modelos que sí son conocidos y usados en varias publicaciones internacionales como el MH o el MEC, que son modelos paramétricos más complicados por ser modelos combinados. Lo más relevante de este capítulo es que el modelo que mejor capta el conteo de caries es de tipo combinado, en este caso MH, y que, desde el punto de vista epidemiológico, las variables regresoras que explican que se salte el obstáculo —cuando la persona tiene caries— no son las mismas si la parte del modelo de conteo del método combinado MH fuese Poisson o binomial negativa.

En el capítulo 7 se continúa trabajando con el mismo estudio del capítulo 6 cambiando la perspectiva de análisis para mostrar una aplicación de indicadores de tipo alternativo. A partir de los diferentes componentes del CPO se incluye por primera vez el componente S (de dientes sanos) mediante transformación en proporciones, con el uso de RB, con una

presentación detallada. Se logra encontrar para el mismo set de variables explicativas que las usadas en el capítulo 6, relaciones relevantes entre los componentes del CPO y el S. La proporción de dientes sanos $\frac{\sum S_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$ muestra una asociación positiva con mayor ingreso y relación inversa con la ingesta de alcohol y el hábito de fumar. La edad es la variable que modula la heterogeneidad y la de dientes cariados $\frac{\sum C_i}{\sum O_i + \sum C_i + \sum P_i + \sum S_i}$ que se asocia positivamente con la edad y el CPO, algo esperable, mientras que los ingresos y el nivel educativo se asocian negativamente. En estos casos, la edad y el CPO aparecen en la heterogeneidad.

En estos tres primeros capítulos de la parte II del libro se trató de forma muy estricta de considerar que un modelo es bueno no solo si tiene variables estadísticamente significativas, sino si muestra una buena capacidad de ajuste. Los demás modelos estimados si bien muestran variables significativas, no están recomendados por el autor, dado el pobre ajuste de los datos.

El capítulo 8 tiene un enfoque diferente. Se usa una lógica de indicador combinado, pero con una diferencia metodológica muy importante: se cambia la filosofía de trabajo dejando de lado los modelos predictivos hasta ahora manejados para lograr resolver de otro modo como se asocian los grupos de patologías bucales. En el contexto de los estudios epidemiológicos donde se indaga por las ENT se trabaja con variables binarias que reflejan la ausencia o presencia de determinadas enfermedades que, a su vez, se asocian con otro conjunto de enfermedades, las comorbilidades, medidas también con variables binarias que, en general, se asumen como factores de riesgo de las primeras. En el ámbito de los estudios epidemiológicos en salud bucal, ambos tipos de variables pueden ser intercambiables en cuanto a quien hace el rol de factor de riesgo.

Con esta premisa, en el capítulo 8 se busca obtener perfiles epidemiológicos bien diferenciados con base en los atributos binarios partiendo de un conjunto de variables, sin discriminar cuáles son variables de respuesta. Se propone la creación de grupos mediante dos estrategias:

1. usar un método de clustering basado en el algoritmo k-modes,

2. usar el análisis de redes SNA y construir la matriz de adyacencias sobre la que se aplican una batería de métricas (closeness, betweenness, modularity, clustering) sobre los nodos y enlaces para detectar comunidades.

Los resultados encontrados permiten decir que ambas metodologías detectan distintas cantidades de grupos que, si bien están diferenciados y tienen una explicación en el contexto del problema para el caso de la partición encontrada mediante k-modes, esta es bastante difusa, mientras que la partición creada mediante el SNA se muestra como más estable.

Por último, antes de pasar a la sección 9.3, donde se dejarán planteadas algunas futuras líneas de acción, es importante mostrar a las lectoras y los lectores del libro el

debe con respecto a uno de los tres tipos de indicadores desarrollados y presentados en la sección 3.5, para los cuales no hay ninguna aplicación en la segunda parte del libro.

El motivo por el cual no se pueden presentar aplicaciones con este tipo de indicador es la falta de datos longitudinales en el área de la salud bucal. Si el lector presta atención, en el capítulo 1 lo esperable era trabajar con el sistema Rediente. Sin embargo, este sistema, que tenía mucha información relativa a las personas que a lo largo de muchos años se han atendido en la Facultad de Odontología, es un sistema cerrado que solo genera indicadores agregados y no permite trabajar con los microdatos, por lo cual no es posible aplicar cualquiera de las metodologías planteadas en la sección 3.5. Por otra parte el sistema Rediente se descontinuó y sustituyó por la HIFO (Historia Clínica de la Facultad de Odontología), que es un buen sistema de seguimientos de pacientes para su gestión, pero no está pensado como un sistema de información estadístico y tampoco permite extraer información individual para aplicar los índices e indicadores presentados.

Lamentablemente, el resto de las fuentes de datos trabajadas, que son de corte transversal, tampoco fueron concebidas en su diseño muestral para usar cualquier método estadístico. Para contemplar el uso de ese herramental y tal vez solamente en el caso del estudio RACA2012 se podría ensayar el uso de algunos indicadores para mostrar agregaciones espaciales, pero con cuidado de no abusar de ellas. Tal como se consignó, el manejo de matrices de contigüidad, aspecto clave en estas técnicas, es lo más difícil de establecer.

Resta advertir a lectoras y lectores que, si bien no hay aplicaciones de este tipo, es fundamental que el investigador en biomedicina sepa que el manejo del tiempo y espacio también es un aspecto muy relevante en el mecanismo de generación de los datos, tanto como el saber si tienen asociado un diseño muestral que pueda condicionar el análisis y de ahí el motivo de la inclusión como una sección relevante del capítulo 3.

En la tabla 9.1 se presenta un resumen de los tipos de indicadores sistematizados y en qué capítulo se presentan, con la particularidad que en varios aplicaciones se combinan estos. Es importante recordar que este libro es un subproducto de la tesis (Álvarez-Vaz, 2020), que contiene ocho capítulos de aplicaciones, de los cuales solo se integran cuatro en este libro y que aparecen en la tabla 9.1:

	Tipo de indicadores usados		
Capítulo	I. Clásicos	I. Combinados	I. Alternativos
Capítulo 5	Si		
Capítulo 6	Si		
Capítulo 7	Si		Si
Capítulo 8	Si	Si	Si

Tabla 9.1: Resumen de los diferentes tipos de Indicadores según capítulo donde aparece una aplicación en parte II del libro

Algunos de los capítulos de la tesis que no forman parte del libro han sido reelaborados y publicados bajo formato de artículos (Álvarez-Vaz y Sassi, 2020), (Álvarez-Vaz y Massa, 2020a) o de capítulos de libros (Álvarez-Vaz et al., 2020). Otros, como el capítulo 7, aparecen en (Álvarez-Vaz y Massa, 2020b).

9.3. Consideraciones generales y planes a futuro

En las dos secciones anteriores se resume un trabajo muy extenso que involucra aspectos teóricos y metodológicos que luego se plasman en un número de aplicaciones, casi en su totalidad inéditas. Cada una de estas, cuando se termina, deja planteadas las limitaciones encontradas y también eventuales desarrollos complementarios.

Un primer desafío para el autor como estadístico y especialista en salud a la vez es de tratar de socializar los metodologías y resultados encontrados entre el grupo de investigadores del área de la biomedicina con quienes el autor comparte investigaciones en coautoría, pertenezcan al área de la salud bucal, en el que se basa este trabajo o con otros de áreas muy diversas como la nefrología, la nutrición, la calidad de vida y los especialistas de las ENT. El propósito de esta obra es que las investigadoras y los investigadores en biomedicina en Uruguay, cualquiera sea su área de especialidad, vea este trabajo como una aproximación a poder usar el herramental estadístico que se presenta y potenciar su trabajo con el uso de estos para sentirse competitivos a nivel internacional.

Otra cuestión que interesa dejar planteada por el autor, es que en cada aplicación que se presenta se trabajó con extremo rigor, cuidando no abusar de la técnica estadística y preservando dos aspectos primordiales: no perder la estructura multivariada pero solo usar modelos con buen nivel de ajuste, ya que de otro modo las investigadoras y los investigadores en biomedicina estarían desarrollando teoría clínica o epidemiológica, independientemente del problema en estudio), que estaría describiendo otra realidad. Por otra parte, no se trabajó en todos los casos particionando los datos con muestras de aprendizajes para los modelos de regresión, en los cuales testear los modelos estimados, aspecto no menor y que en general desde una perpectiva estadística, suele hacerse como forma de evaluar la robustez de los modelos y mitigar la eventual sobrestimación de los mismos. El motivo de esta decisión es no agregar un elemento de complejidad extra y entender más aún, partiendo del supuesto de que, en general, en muchas revistas de biomedicina ese aspecto metodológico no se sabe si se lleva a cabo, pues no se consigna.

Para terminar y dejar algunas líneas de trabajo se propone para:

Modelos de conteo: trabajar para solucionar las limitaciones del uso de todos estos en el contexto de estudios con diseños muestrales complejos, que en el trabajo desarrollado en el libro, dadas las limitaciones del software usado, fue posible trabajar con un solo tipo, con lo cual la solución puede ser desarrollar subrutinas que contemplen los aspectos requeridos de manera adecuada o cambiar de paquete estadístico.

Por ejemplo, el Stata trabaja con modelos de conteo con diseños complejos, pero, lamentablemente, además de ser un software privativo, no existe documentación técnica que exlique en detalle cómo lleva a cabo los procesos de estimación.

Extensión sobre modelos probabilísticos para ajustar tasas: profundizar en su uso para otras patologías bucales o de otro tipo, trabajando con resultados que tiene que ver con una mejora en la estimación mediante árboles de regresión (combinación de los métodos CART y los modelos de RB y modelos de variable de clase latente que explican mejor la heterogeneidad, así como el exceso de O y 1, que se dan en los RB (Ospina y Ferrari, 2012), (Grün et al., 2011), (Zeileis et al., 2008).

Análisis de redes: Incorporar cada vez más este tipo de análisis. En realidad, el problema en estudio de las patologías de las ENT da lugar a una red bipartita, de la cual solo se considera el modo 1, donde los nodos son las personas y los enlaces la cantidad de ENT que comparten. Además, dado que en las ENT hay otros atributos de las personas que pueden justificar su inclusión en el análisis, se puede pensar en elaborar una red en la que los nodos sean las ENT y otros indicadores epidemiológicos de las personas y el peso de los enlaces la cantidad de personas que comparten cada atributo. Lo más relevante de evaluar en la red a estudiar es el enlace selectivo entre nodos de acuerdo a algunas características, usando el coeficiente de asortatividad. Este concepto, conocido como homofilia, estaría expresando la tendencia de las personas que son parte de la red de atención de la Facultad de Odontología a relacionarse con personas que se les parecen, (Krackhardt y Stern, 1988). Una fortaleza de este método de SNA es que frente al enfoque tradicional, dedicado a estimar modelos por separado para cada ENT y que resulta insuficiente porque tal vez exista una propensión global a padecer ENT, aparece como una solución alternativa, relajando la restricción de independencia entre observaciones, condición necesaria para que su uso sea válido. Se propone un análisis de redes alternativo, validando modelos estadísticos (hay que recordar que lo utilizado con la lógica del SNA es solo con un enfoque de descripción), donde algunos de los atributos evaluados en la caracterización se pueden usar como variables explicativas, usando la teoría de los modelos exponenciales aleatorios en grafos (ERGM),(Kolaczyk y Csárdi, 2014).

Agregaciones espaciotemporales: Dado que el Uruguay no tiene datos longitudinales disponibles para investigadores en salud bucal, se propone trabajar con otras fuentes de datos de otros países. Un ejemplo son las lesiones por accidente de tránsito e internaciones que surgen de diferentes fuentes para el caso de Brasil, como el Sistema de Informacões de Mortalidade (SIM), del Sistema de Internacões Hospitalares (SIH-SUS) y del Instituto Brasileiro de Geografia e Estatística (IBGE) a través del sistema (DATASUS). Estas tres fuentes de datos, que tienen veinte años, con frecuencia mensual, permitirían aplicar muchas de las herramientas presentadas en la sección 3.5. Al momento de presentar este libro, el autor se encuentra trabajando

con esos datos para la elaboración de clusters de trayectorias de tasas de mortalidad y letalidad de cada unos de los 27 estados en los que se divide Brasil, desde una perspectiva descriptiva mediante metodología de clusters longitudinales, (Genolini et al., 2015) para luego complementarla con la metodología de Age-period-cohort Analysis (APCA), (Yang y Land, 2013)).

Bibliografía

Abernathy, J. R., Graves, R. C., Bohannan, H. M., Stamm, J. W., Greenberg, B. G., y Disney, J. A. (1987). Development and application of a prediction model for dental caries. Community Dentistry and Oral Epidemiology, 15:24–28.

Agresti, A. (2005). Categorical Data Analysis. John Wiley & Sons.

Aida, J., Kondo, K., Kondo, N., Watt, R. G., Sheiham, A., y Tsakos, G. (2011). Income inequality, social capital and self-rated health and dental status in older japanese. Soc Sci Med, 73(10):1561–8.

Álvarez-Vaz, R. (2010). Métodos de muestreo para estudios sanitarios con uso de información auxiliar. Tesis de maestría, Maestría en Epidemiología-Facultad de Medicina-Udelar.

Álvarez-Vaz, R. (2017). Métodos de muestreo para estudios sanitarios: una aplicación para la Encuesta de Factores de Riesgo 2006 en Uruguay. Ediciones Universitarias, Unidad de Comunicación de la Universidad de la Repb́lica (UCUR).

Álvarez-Vaz, R. (2020). Sistematización y creación de Indicadores e Índices para la vigilancia epidemiológica en Salud Bucal. Tesis de Doctorado, PROINBIO, Escuela de Graduados, Facultad de Medicina, Udelar.

Álvarez-Vaz, R. y Massa, F. (2012). Determinación de tipologías de infecciones parasitarias intestinales, en escolares mediante, técnicas de clustering sobre datos binarios. Documento de Trabajo Serie DT (12 / 05) - ISSN : 1688-6453, IESTA.

Álvarez-Vaz, R. y Massa, F. (2020a). Aplicación de técnicas de clustering para la elaboración de perfiles epidemiológicos en estudios sanitarios. Saberes, 12(2).

Álvarez-Vaz, R. y Massa, F. (2020b). Aplicación de análisis de redes para la elaboración de perfiles epidemiolgícos en estudios sanitarios. En Zacarías Flores, José Dionisio and Vázquez Guevara,Víctor Hugo, editor, Actualidad en Probabilidad y Estadística, pp. 87–102. Benemérita Universidad Autónoma de Puebla.

Álvarez-Vaz, R., Massa, F., y Vernazza, E. (2020). Creación de indicadores alternativos para la vigilancia en Salud oral mediante Regresión Beta. En Zacarías Flores, José Dionisio and Vázquez Guevara,Víctor Hugo, editor, Actualidad en Probabilidad y Estadística, pp. 195–212. Benemérita Universidad Autónoma de Puebla.

Álvarez-Vaz, R. y Sassi, C. (2020). Determinación del sexo mediante técnicas de clasificación supervisada. Revista de la Facultad de Ciencias de la Universidad Nacional de Colombia, 9(1):6–24.

Álvarez-Vaz, R., Riaño, M., Mesa, M., Buño, G., y Nalbarte, L. (2011). Maloclusión en niños en edad escolar: Análisis de los factores de riesgo. Colección Biblioteca Plural de la CSIC. Departamento de Publicaciones, Unidad de Comunicación de la Universidad de la República (UCUR).

Arana, A., E, B., y Salazar, F. (2005). Diagnóstico de caries dental, pp. 113–120.

Bacallao, J. (2013). Ensayo crítico acerca de la medición de las desigualdades sociales en salud. Tesis de doctorado, Universidad de Ciencias Médicas de La Habana.

Bacallao, J., Castillo-Salgado, C., Schneider, M. C., Mujica, O. J., Loyola, E., y Manuel, V. (2002). Índices para medir las desigualdades de salud de carácter social basados en la noción de entropía. Revista Panamericana de Salud Pública, 12:429 – 435.

Beca, J., Ferrara, A., y Lorenzo, S. (1996). Prevalencia de caries dental a los 12 años en la ciudad de Montevideo. Rev Tecnol Odonto, 6:29–34.

Bhupathiraju, S. y Tucker, K. (2011). Coronary heart disease prevention: Nutrients, foods, and dietary patterns. Clin Chim Acta, 412(17-18):1493–514.

Bianco, P., Dominguez, M., y C.Beca (1997). Aproximación a los determinantes sociales de la enfermedad caries en niños de 12 años. An Fac Odont, 27:5–38.

Bivand, R., Pebesma, E., y Gómez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Use R! Springer.

Blanco, J. (2006). Introducción al Análisis Multivariado:Teoría y aplicaciones a la realidad latinoamericana. IESTA, Universidad de la República.

Blanco Carrión, A., Otero Rey, E., Peña María Mallón, M., y Diniz Freitas, M. (2008). Diagnóstico del liquen plano oral. Avances en Odontoestomatología, 24:11 – 31.

Bonacich, P. (1987). Power and centrality: A family of measures. American Journal of Sociology, 5:1170.

Bonacich, P. y Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. Social Networks, 23:191 – 201.

Borgatti, S. P., Everett, M. G., y Johnson, J. (2013). Analyzing Social Networks. SAGE Publications Ltd.

Brandes, U. (2001). A faster algorithm for betweenness centrality. The Journal of Mathematical Sociology, 25(2):163–177.

Brandes, U. y Erlebach, T. (2005). Network analysis: methodological foundations. Nmero 3418. Springer, Berlin ; New York.

Breilh, J. (2010). La epidemiología crítica: una nueva forma de mirar la salud en el espacio urbano. Salud Colectiva, 31:152–157.

Breiman, L., J.Friedman, Stone, C. J., y Olshen, R. (1984). Classification and Regression Trees. Chapman and Hall/CRC.

Breslow, N. y Beckwith, J. (1982). Epidemiological features of wilms' tumor: results of the national wilms' tumor study. J Natl Cancer Inst., 68(3):429–36.

Burt, B., Baelum, V., y Fejerskov, O. (2008). Dental Caries: The disease and its clinical management, caplo Dental Caries: The disease and its clinical management.The epidemiology of dental caries, pp. 124–145.

Butts, C. T. (2016). sna: Tools for Social Network Analysis. R package version 2.4.

Cañizares Pérez, M., Barroso Utra, I., Alfonso León, A., García Roche, R., Alfonso Sagué, K., Chang de la Rosa, M., Bonet Gorbea, M., y León, E. (2004 Mar). Estimate methods used with complex sampling designs: their application in the cuban 2001 health survey. Rev Panam Salud Publica, 15(3):176–84.

Cameron, A. (1998). Regression analysis of count data. Cambridge University Press, Cambridge, UK New York, NY, USA.

Cameron, A. C. y Trivedi, P. K. (1990). Regression-based tests for overdispersion in the poisson model. Journal of Econometrics, 46:347–364.

Casnati, B., Álvarez-Vaz, R., Massa, F., Lorenzo, S., Angulo, M., y Carzoglio, J. (2013). Prevalencia y factores de riesgo de las lesiones de la mucosa oral en la población urbana del Uruguay. Odontoestomatología, 15:58 – 67.

Chattopadhyay, A. (2011). Oral Health Epidemiology: Principles and Practice. Jones and Bartlett.

Chen, R. (1978). A surveillance system for congenital malformations. Journal American Statistical Association, (73):323–7.

Chen, R. (1979). Statistical techniques in birth defects surveillance systems. Contr. Epidemiol. Biostat., (1):184–9.

Chen, R. (1986). Revised values for the parameters of the sets technique for monitoring the incidence rate of a rare disease. Meth Inform Med., (25):47–9.

Chen, R., Mantel, N., Connelly, R., e Isacson, P. (1982). A monitoring system for chronic diseases. Meth Inform Med., (21):86–90.

Chen, R., McDowell, M., Terzian, E., y Weatherall, J. (1983). Eurocat guide to monitoring methods for malformation registers.

Clayton, D. y Hills, M. (1993). Statistical Models in Epidemiology. Oxford Science Publications.

Cochran, W. (1977). Sampling Techniques. John Wiley & Sons, New York.

Cook, N., Cutler, J., Obarzanek, E., Buring, J., Rexrode, K., y Kumanyika, S. (2007). Long term effects of dietary sodium reduction on cardiovascular disease outcomes: observational follow-up of the trials of hypertension prevention (TOHP). British Medical Journal, 334(7599):885–8.

Cortés, F. (1999). Medición de la enfermedad en odontología comunitaria., pp. 303–325. Masson.

Cribari-Neto, F. y Zeileis, A. (2010). Beta Regression in R. Journal of Statistical Software, 34(2):1–24.

Csardi, G. y Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems:1695.

CSDH, C. o. S. D. o. H. (2008). A conceptual framework for action on the social determinants of health. discussion paper for the commission on social determinants of health draft. Technical report, WHO.

Cuadras, C. (2007). Nuevos Métodos de Análisis Multivariante. CMC Editions.

de Leeuw, J. y Meijer, E. (2008). Handbook of Multilevel Analysis. Springer.

Diniz, M. B., Rodrigues, J. A., Hug, I., De Cássia Loiola Cordeiro, R., y Lussi, A. (2009). Reproducibility and accuracy of the ICDAS-II for occlusal caries detection. Community Dentistry and Oral Epidemiology, 37(5):399–404.

Do, L. G., Spencer, A. J., Slade, G. D., Ha, D. H., Roberts-Thomson, K. F., y Liu, P. (2010). Trend of income-related inequality of child oral health in australia. J Dent Res, 89(9):959–64.

Eccles, J. (1979). Dental erosion of nonindustrial origin. a clinical survey and classification. The Journal of Prosthetic Dentistry, 42(6):649 – 653.

Elani, H. W., Harper, S., Allison, P. J., Bedos, C., y Kaufman, J. S. (2012). Socioeconomic inequalities and oral health in canada and the united states. J Dent Res, 91(9):865–70. .

Escofier, B. y Pagés, J. (2008). Analyse Factorielle Mutiple. Dunod.

Escribano-Bermejo, M. y Bascones-Martínez, A. (2009). Leucoplasia oral: Conceptos actuales. Avances en Odontoestomatología, 25:83 – 97.

Fejerscov, O. (2004). Changing paradigms in concept son dental caries: Consequences for oral health care. Caries Research, 38:182–191.

Finch, W. (2014). Multilevel modeling using R. CRC Press, Boca Ratón, Florida.

Food and Agriculture Organization of the United Nations (2010). Fats and fatty acids in human nutrition. Report of an experte consultation 10-14 november 2008, FAO.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. Social Networks, 1(3):215.

French, J. (2015a). smacpod: Statistical Methods for the Analysis of Case-Control Point Data. R package version 1.4.1.

French, J. (2015b). smerc: Statistical Methods for Regional Counts. R package version 0.2.2.

Fuller, W. A. y Braid, J. (1999). Estimation for supplemented pannels. Sankya: The Indian J of Statistics, 61(Serie b):58–70.

Ganss, C. (2006). Dental erosion : from diagnosis to therapy. Karger, Basel New York.

Gelman, A. y Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Genolini, C., Alacoque, X., Sentenac, M., y Arnaud, C. (2015). kml and kml3d: R packages to cluster longitudinal data. Journal of Statistical Software, 65(4):1–34.

Giner, G. y Smyth, G. K. (2016). Statmod: probability calculations for the inverse gaussian distribution. R Journal, 8(1):339–351.

Goldstein, L. P. H. (1991). New statistical methods for analysing social structures: An introduction to multilevel models. Educational Research Journal, 17(4):387–393.

Gómez-Rubio, V., Ferrándiz-Ferragud, J., y Lopez-Quílez, V. (2005). Detecting clusters of disease with r. Journal of Geographical Systems, 7(2):189–206.

Graber, T. y Swaim, B. (1991). Ortodoncia. Principios Generales y Técnicas. Editorial Médica Panamericana S.A.

Graham, H. (2004a). Social determinants and their unequal distribution: clarifying policy understandings. The Milbank Quarterly, 82(1):101–24. Graham, Hilary Milbank Q. 2004;82(1):101-24.

Graham, H. (2004b). Tackling inequalities in health in england: remedying health disadvantages, narrowing health gaps or reducing health gradients? Journal of Social Policy, 33(01):16.

Grimson, R., Wang, K., y Johnson, P. (1981). Searching for hierarchical clusters of disease: spatial patterns of sudden infant death syndrome. Social Science Medicine, (15):287–93.

Grün, B., Kosmidis, I., y Zeileis, A. (2011). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. Working Paper 2011-22, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck.

Gugnani, N., Pandit, I., Srivastava, N., Gupta, M., y Sharma, M. (2011). International caries detection and assessment system (icdas): A new concept. International Journal of Clinical Pediatric Dentistry, 4(2),:93–100.

Guillén, M., Juncá, S., Rué, M., y Aragay, J. (2000 Sep-Oct). Effect of sample design in the analysis of complex surveys. application to the health survey of catalonia. Gaceta Sanitaria, 14(5):399–402.

Hansen, M. H., Hurwitz, W. N., y Madow, W. G. (1953). Sample Survey Methods and Theory, Vol. I (Concepts and Discussion) and Vol. II,. John Wiley & Sons, New York.

Hardy, R., Buffler, P., Prichard, H., Schroeder, G., Cooper, S., Crane, M., y Doak, C. (1983). Monitoring for health effects of low-level radioactive waste disposal: a feasibility study. Technical report, Texas Dept. of Health. submitted to Texas Low-level Radioactive Waste Disposal Authority.

Hardy, R., Schroeder, G., Cooper, S., Buffler, P., Prichard, H., y Crane, M. (1990). A surveillance system for assessing health effects from hazardous exposures. 1(132):532–42.

Hilbe, J. (2011). Negative binomial regression. Cambridge University Press, Cambridge, UK New York.

Hilbe, J. (2014). Modeling count data. Cambridge University Press, New York, NY, USA.

Hilbe, J. (2016). COUNT: Functions, Data and Code for Count Data. R package version 1.3.4.

Hirji, K. (2006). Exact analysis of discrete data. Chapman & Hall/CRC, Boca Raton.

Holst, D., Schuller, A., Aleksejunienė, J., y EriKsen, H. (2001). Caries in populations – teorical, casual approach. European Journal Of Oral Sciences, 109(3):143–148.

Hox, J. (1995). Multilevel Analysis Techniques and Applications. Lawrence Erlbaum Associates, Mahwah, N.J.

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. in kdd: Techniques and applications. Technical report, World Scientific.

Ismail, A. I., Sohn, W., Tellez, M., Amaya, A., Sen, A., Hasson, H., y Pitts, N. B. (2007). The international caries detection and assessment system (icdas): an integrated system for measuring dental caries. Community Dentistry and Oral Epidemiology, 35(3):170–178.

Jablonski-Momeni, A., Stachniss, V., Ricketts, D., Heinzel-Gutenbrunner, M., y Pieper, K. (2008). Reproducibility and Accuracy of the ICDAS-II for Detection of Occlusal Caries in vitro. Caries Research, 42(2):79–87.

Keppel, K., Pamuk, E., Lynch, J., Carter-Pokras, O., Kim, I., Mays, V., Pearcy, J., Schoenbach, V., y Weissman, J. (2005). Methodological issues in measuring health disparities. Vital Health Stat, 141:1–16.

Keppel, K., Pearcy, J., Jeffrey, N., y Klein, R. (2004). Measuring progress in healthy people 2010. Healthy People Statistical Notes, (25).

Kieschnick, R. y McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Statistical Modelling: An International Journal, 3(3):193.

Kim, J. S. y Dailey, R. J. (2008). Biostatistics for Oral Healthcare. Blackwell, Oxford.

Kish, L. (1965). Survey Sampling. John Wiley & Sons, New York.

Knox, G. (1964). The detection of space-time interactions. Applied Stat., (13):25–9.

Kolaczyk, E. y Csárdi, G. (2014). Statistical analysis of network data with R. Springer, New York.

Kolaczyk, E. y Csárdi, G. (2017). sand: Statistical Analysis of Network Data with R. R package version 1.0.3.

Krackhardt, D. y Stern, R. N. (1988). Informal networks and organizational crises: An experimental simulation. Social psychology quarterly, pp. 123–140.

Lawson, A. B. (2009). Bayesian Disease Mapping Hierarchical Modeling In Spatial Epidemiology. Chapman & Hall/Crc Interdisciplinar y Statistics Series.

Lawson, A. B. y Kleinman, K. (2005). Spatial and Syndromic Surveillance for Public Health. John Wiley & Sons Ltd.

Lilienfeld, A. y Lilenfield, D. (1980). Foundations of Epidemiology. Oxford University Press.

Lorenzo, S., Álvarez-Vaz, R., Blanco, S., y Peres, M. (2013a). Primer Relevamiento Nacional de Salud Bucal en población joven y adulta uruguaya: Aspectos metodológicos. Odontoestomatología, 15:8 – 25.

Lorenzo, S., Piccardo, V., Alvarez, F., Massa, F., y Alvarez Vaz, R. (2013b). Enfermedad Periodontal en la población joven y adulta uruguaya del Interior del país: Relevamiento Nacional 2010-2011. Odontoestomatología, 15:35 – 46.

Luke, D. (2015). A user's guide to network analysis in R. Springer.

Lumley, T. (2009). Survey: analysis of complex survey samples. R package version 3.16. R package version 3.16.

Lwanga, S. y Lemeshow, S. (1991.). Determinación del tamaño de las muestras en los estudios sanitarios:.

Mackenbach, J. P. y Kunst, A. E. (1997). Measuring the magnitude of socio-economic inequalities in health: an overview of available measures illustrated with two examples from europe. Soc Sci Med, 44(6):757–71.

Maltz, M. y Jardim, J. Alves, L. (2010). Health promotion and caries dental caries. Braz Oral Res, 24(Spec Iss 1):18–25.

Manau, C. y Echeverría, J. (1999). Enfermedades periodontales, pp. 137–152.

Martínez, N. J., Antó, B. J., Castellanos, P., Gili, M. M., Marset, C. P., y Navarro, L. V., editores (1997). Salud Pública. Mc Graw Hill-Interamericana.

McCulloch, C. (2001). Generalized, linear, and mixed models. John Wiley & Sons, New York.

Ministério da Saúde (2002-2003). Condições de saúde bucal da população brasileira. Technical report, Ministério da Saúde.Série C. Projetos, Programas e Relatórios.

Ministerio de Salud Pública (2009). 1a Encuesta Nacional de Factores de Riesgo de Enfermedades Crónicas No Transmisibles. Ministerio de Salud Pública, División Epidemiología.

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics, 33:341–365.

Narvai, P., Frazo, P., Roncalli, A., y Antunes, J. (2006). Caries dentaria no brasil: declínio, polarizacão, iniqüidade e exclusão social. Rev Panam Salud Publica, 19(6):8.

Naus, J. (1965). The distribution of the size of the maximum cluster of points on a line. Journal American Statistical Association, (60):532–38.

Naus, J. (1982). Approximations for distributions of scan statistics. Journal American Statistical Association, (77):177–83.

Nelder, J. A. y Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society A, 135:370–384.

Newman, M. E. J. (2002). Assortative mixing in networks. Phys. Rev. Lett., 89:208701.

Newman, M. E. J. (2003). Mixing patterns in networks. Phys. Rev. E, 67.

Nithila, A., Bourgeois, D., Barmes, D. E., y Murtomaa, H. (1998). Banco Mundial de Datos sobre Salud Bucodental de la OMS, 1986–1996: panorámica de las encuestas de salud bucodental a los 12 años de edad. Pan Am J Public Health, 4(6):411–419.

Oakes, J. M. y Kaufman, J. S. (2006). Methods in Social Epidemiology. Jossey Bass- A Wiley imprint.

Ohno, Y., Aoki, K., y Aoki, N. (1979). A test of significance for geographic clusters of disease. International Journal for Parasitology, (8):273–81.

Olmos, P., Piovesan, S., Musto, M., Lorenzo, S., Álvarez-Vaz, R., y Massa, F. (2013). Caries dental. La enfermedad oral más prevalente: Primer Estudio poblacional en jóvenes y adultos uruguayos del interior del país. Odontoestomatología, 15:26 – 34.

Organización Mundial de la Salud (1997). Encuestas de Salud Bucodental: Métodos Básicos. Organización Mundial de la Salud, 4ta edici

Organization, W. H. (1987). Oral Health Survey Basic Methods. World Health Organization, 3 edici

Oriol, R. J. (1997). Salud Pública, caplo 8, pp. 139–147. Mc Graw Hill-Interamericana.

Ospina, R. y Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis, Volume 56(Issue 6):1609–1623.

Ourens, M., Celeste, R., Hilgert, J. B., Lorenzo, S., Hugo Neves, F., Álvarez-Vaz, R., y Abegg, C. (2013). Prevalencia de maloclusiones en adolescentes y adultos jóvenes del interior del Uruguay. Relevamiento nacional de salud bucal 2010-2011. Odontoestomatología, 15:47 – 57.

Page, R. y Eke, P. (2007). Case definitions for use in population-based surveillance of periodontititis. Journal Of Periodontology, 78(7):1387–1399.

Petersen, P. E. (2009). Global policy for improvement of oral health in the 21st century–implications to oral health research of world health assembly 2007, world health organization. Community Dent Oral Epidemiol, 37(1):1–8.

Peterson, P. (2004). Challenges to improvement of oral health in the 21st century - the approach of the who global oral health programme. International Dental Journal, 54:329–343.

Pfeiffer, D. U., Robinson, T. P., Stevenson, M., y Stevens, K. B. (2008). Spatial Analysis in Epidemiology. Oxford University Press.

Proffit, W. (1990). Reactor paper: risk assessment for developmental problems-where are we now?, pp. 162–163. Bader, J D. Ed. Risk assessment in den- tistry. Chapell Hill, NC: University of North Carolina,.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raj, D. (1968). Sampling Theory. McGraw-Hill, Inc., New York.

Rigby, R. A. y Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). Applied Statistics, 54:507–554.

Riva, R., Sanguinetti, M., Rodríguez, A., Guzzetti, L., Lorenzo, S., Álvarez-Vaz, R., y Massa, F. (2011). Prevalencia de trastornos témporo mandibulares y bruxismo en Uruguay: PARTE I. Odontoestomatología, 13:54 – 71.

Rothman, K. J. y Greenland, S. (1998). Modern Epidemiology. Lippincott Williams & Wilkins, 2 edici

Rowlingson, B. y Diggle, P. (2017). splancs: Spatial and Space-Time Point Pattern Analysis. R package version 2.01-40.

Sabidussi, G. (1966). The centrality index of a graph. Psychometrika, 31(4):581–603.

Salinas-Rodríguez, A., Manrique-Espinoza, B., y Sosa-Rubí, S. G. (2009). Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud. Salud Pública de México, 51:397–406.

Sanders, A. (2007). Social determinants of oral health: conditions linked to socioeconomic of inequalities in oral health in the australian population. Technical report, University of Adelaide.

Santiago Pérez, M. I., Hervada Vidal, X., Naveira Barbeito, G., Silva, L. C., Fariñas, H., Vázquez, E., Bacallao, J., y Mujica, O. J. (2010). El programa epidat: usos y perspectivas. Revista Panamericana de Salud Pública, 27:80 – 82.

Särndal, C.-E. y Lundström, S. (2005). Estimation in Surveys with Nonresponse. John Wiley & Sons.

Särndal, C.-E., Swensson, B., y Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag.

Schneider, M., Castillo-Salgado, C., Bacallao, J., Loyola, E., Mujica, O., y Vidaurre, M. (2002). Métodos de medición de las desigualdades de salud. Revista Panamericana de Salud Pública, 12(6):398–415.

Scrucca, L. (2004). qcc: An R package for quality control charting and statistical process control. R News, 4(1):11–17.

Shannon, C. E. y Weaver, W. (1949). The Mathematical Theory of Communication. Univ of Illinois Press.

Silva, L. C. (1995). Excursión a la Regresión Logística en Ciencias de la Salud. Díaz de Santos, Madrid.

Silva, L. C. (1998). Cultura Estadística e investigación científica en el campo de la salud: una mirada crítica. Díaz de Santos.

Silva, L. C. (2000). Diseño razonado de muestras y captación de datos para la investigación sanitaria. Díaz de Santos, Madrid.

Skapino, E. y Álvarez-Vaz, R. (2016). Prevalencia de factores de riesgo de enfermedades crónicas no transmisibles en funcionarios de una institución bancaria del Uruguay. Revista Uruguaya de Cardiología, 31:246 – 255.

Smith, B. y Knight, J. (1984). An index for measuring the wear of teeth. British Dental Journal, 156::435–438.

Stark, C. R. y Mantel, N. (1967). Lack of seasonal or temporal spatial clustering of down's syndrome births in michigan. American Journal of Epidemiology, (86):199–213.

Tango, T. . (2010). Statistical Methods for Disease Clustering. Statistics for Biology and Health. Springer.

Thomas, J. C. y Weber, D. J. (2001). Epidemiologic Methods for the Study of Infectious Diseases. Oxford University Press, USA, 1 ediciISBN: 0195121120.

Tonetti, M. y Claffey, N. (2005). European workshop in periodontology group c. advances in the progression of periodontitis and proposal of definitions of a periodontitis case and disease progression for use in risk factors research. group c. Clinic Periodontol, 32(6):210–213.

Tsakos, G., Demakakos, P., Breeze, E., y Watt, R. G. (2011). Social gradients in oral health in older adults: findings from the english longitudinal survey of aging. Am J Public Health, 101(10):1892–9.

Tsekouras, G., Papageorgiou, D., Kotsiantis, S., Kalloniatis, C., y Pintelas, P. (2005). Fuzzy clustering of categorical attributes and its use in analyzing cultural data. World Academy of Science, Engineering and Technology, 1:87–91.

Twisk, J. W. (2006). Applied Multilevel Analysis. Cambridge University Press.

Venables, W. y Ripley, B. (2002a). Modern Applied Statistics with S. Springer-Verlag, New York, 4th edici

Venables, W. N. y Ripley, B. D. (2002b). Modern Applied Statistics with S. Springer, New York, 4 ediciISBN 0-387-95457-0.

Verkuilen, J. y Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. Journal of Educational and Behavioral Statistics, 37(1):82–113.

Vig, P. (1990). Risk assessment applied to dentofacial deformity: a consideration of postnatal environmental factors, pp. 156–161. Bader, J D. Ed. Risk assessment in dentistry. Chapell Hill, NC: University of North Carolina,.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., y McCulloch, C. E. (2005). Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. Springer, 1 edici344.

Wagstaff, A. (2002). Inequality aversion, health inequality and health achievement. Research Working Paper 2765, The World Bank Development Research Group Public Services and Human Development Network Health, Nutrition and Population Team.

Wagstaff, A., Paci, P., y van Doorslaer, E. (1991). On the measurement of inequalities in health. Soc Sci Med., 33:545–57.

Waller, L. A. y Gotway, C. A. (2004). Applied Spatial Statistics for Public Health Data. John Wiley & Sons Inc.

Wasserman, S. y Faust, K. (1994). Social network analysis: methods and applications. Nmero 8 en Structural analysis in the social sciences. Cambridge University Press, Cambridge ; New York.

Weatherall, J. y Haskey, J. (1976). Surveillance of malformations. Br Med Bull., (32):39–44.

Weihs, C., Ligges, U., Luebke, K., y Raabe, N. (2005). klar analyzing german business cycles. En Baier, D., Decker, R., y Schmidt-Thieme, L., editores, Data Analysis and Decision Support, pp. 335–343, Berlin. Springer-Verlag.

Wheeler, B. (2016). SuppDists: Supplementary Distributions. R package version 1.1-9.4.

Whitehead, M. (1992). The concepts and principles of equity and health. Int J Health Serv, 22(3):429–45.

Willeman Bastos Tesch, L. V., Souza Tesch, R. d., y Pereira Jr., F. J. (2014). Trastornos temporomandibulares y dolor orofacial crónico: al final, á qué área pertenecen? Revista de la Sociedad Española del Dolor, 21:70 – 74.

Winkelmann, R. (2008). Econometric analysis of count data. Springer, Berlin.

Wolf, H., Rateitschak, E., y Rateitschak, K. (2005). Atlas en color de odontología. Elsevier-Masson., 3 edici

World Cancer Research Fund International (2007). Food, nutrition, physical activity and the prevention of cancer: a global perspective. Technical report, World Cancer Research Fund, American Institute for Cancer Research.

World Health Organization (2000). Global strategy for the prevention and control of noncommunicable diseases. Technical report, WHO.

World Health Organization (2005). Preventing chronic diseases: a vital investment:who global report. Technical report, WHO.

World Health Organization (2006). Who STEPS Surveillance Manual. World Health Organization.

Yang, Y. y Land, K. (2013). Age-Period-Cohort Analysis New Models, Methods, and Empirical Applications. CRC Press.

Yao, W., Li, Z., y Graubard, B. I. (2015). Estimation of roc curve with complex survey data. Statistics in medicine, 34(8):1293–1303.

Yee, T. y Hastie, T. J. (2003). Reduced-rank vector generalized linear models. Statistical Modelling, 3:15–41.

Yee, T. W. (2010). The vgam package for categorical data analysis. Journal of Statistical Software, 32(10):1–34.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. Journal of Statistical Software, 11(10):1–17.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. Journal of Statistical Software, 16(9):1–16.

Zeileis, A. y Kleiber, C. (2008). AER: Applied Econometrics with R. R package version 0.9-0.

Zeileis, A., Kleiber, C., y Jackman, S. (2008). Regression models for count data in r. Journal of Statistical Software, 27(8):1–25.

Parte IV

Apéndice

Apéndice A

Aspectos metodológicos estadísticos

A continuación se presenta una sucinta descripción teórica de las herramientas estadísticas utilizadas en la tesis de (Álvarez-Vaz, 2020) que fue adaptada en esta obra.

A.1. Análisis factorial

El análisis factorial es una técnica multivariada que pretende describir un conjunto de variables a través de un número reducido de variables sintéticas no observables llamadas factores. Uno de los objetivos es reducir las dimensiones de la tabla de datos con pérdida de información mínima.

Existen diversos métodos de análisis factorial: el análisis de componentes principales (ACP) trata las matrices cruzadas de individuos por variables cuantitativas; el análisis de correspondencias simple (ACS) manipula tablas de contingencia; el análisis de correspondencias múltiples (ACM) trabaja con las llamadas tablas disjuntas, en las que se cruzan individuos por variables cualitativas, entre otros. Este último permite analizar una población de I individuos descritos por J variables categóricas. Una variable categórica representa una propiedad que hace referencia a cualidades del objeto de estudio, que no pueden ser cuantificadas, como el género y la raza.(Escofier y Pagés, 2008).

La tabla de datos puede representarse de dos maneras. La primera es mediante la tabla disyuntiva completa (TDC), en la que las filas representan a los individuos y las columnas a las modalidades de las variables. Se definen las siguientes variables indicadoras:

$$x_{ik} \begin{cases} 1 & \text{si el individuo i posee la modalidad k} \\ 0 & \text{en otro caso} \end{cases} \tag{A.1}$$

por lo tanto, los valores de la TDC serán solo 0 y 1.

Las modalidades son excluyentes, por lo que dentro de una misma variable deberá encontrarse un único 1. Por lo tanto, la suma de las filas es la cantidad de variables J. La suma de las columnas es la cantidad de individuos que poseen la modalidad k, I_k.

La segunda forma de representar la tabla de datos es con la tabla de Burt, que contiene el conjunto de tablas de contingencia entre variables 2 a 2. La tabla de Burt no es exactamente una tabla de contingencia, sino una yuxtaposición de ellas. Esta tabla es análoga a una matriz de correlaciones, en tanto resume el conjunto de las relaciones entre las variables tomadas 2 a 2.

En el ACM intervienen tres objetos: los individuos, las variables y las modalidades.

1. Estudio de los individuos: Uno de los objetivos del ACM es construir tipologías de individuos. Esta tipología debe estar basada en una noción de semejanza tal que dos individuos están tanto más próximos cuanto mayor es el número de modalidades que tienen en común.

2. Estudio de las variables: se pueden adoptar dos puntos de vista en el estudio de las variables. El primero es el estudio de la relación entre dos variables, que necesita considerar la tabla de contingencia que cruza sus modalidades. El segundo es el de reducir las dimensiones, es decir, obtener un número pequeño de variables sintéticas numéricas.

3. Estudio de las modalidades: una modalidad puede ser clasificada como:

   - variable indicadora definida sobre el conjunto de los individuos. Esto es una columna de la TDC.
   - clase de individuos de la que se conoce la distribución sobre el conjunto de las modalidades. Puede ser una fila o una columna de la tabla de Burt.

   La noción de semejanza entre modalidades depende del punto de vista adoptado. En el primer caso, dos modalidades se parecen tanto más cuanto mayor sea su presencia o ausencia simultánea en un gran número de individuos. Las otras modalidades no intervienen. En el segundo caso, dos modalidades se parecen tanto más cuanto mayor o menor sea su asociación con las mismas modalidades.

A.2. Análisis factorial múltiple

El AFM es una herramienta desarrollada por (Escofier y Pagés, 2008) que permite el análisis simultáneo de variables estructuradas en grupos, equilibrando la influencia de cada uno de ellos. Los grupos de variables pueden ser de diferente naturaleza y número. Las únicas condiciones son que las variables de un mismo grupo sean de la misma naturaleza —cuantitativa o cualitativa— y que las tablas tengan la misma cantidad de individuos.

La influencia de un grupo de variables en un análisis conjunto se puede dar a través de dos elementos:

- el número de variables: cuanto mayor sea el número, mayor la influencia;
- la estructura del grupo: cuanto más relacionadas estén, más determinan los principales factores.

Una de las formas de ponderar cada variable dentro de un grupo puede determinarse por el inverso del valor propio asociado al primer eje factorial del ACM o del ACP realizado en cada uno de ellos, denotado como λ_j^1. El AFM permite estudiar:

- relaciones entre grupos, además de medir su grado de semejanza,
- relaciones entre las variables de un grupo y las del resto de los grupos,
- semejanzas entre los individuos vistos a través de los diferentes grupos de variables.

Es un método que pone en evidencia factores comunes a todos los grupos, factores comunes a algunos grupos y factores específicos de algunos grupos. Los objetos de estudio de esta técnica —individuos, variables y grupos de variables— están en los espacios R^K, R^I y R^{K_j} respectivamente, donde K es el número de variables, I el número de individuos y K_j el total de variables en el grupo j.

A.2.1. Influencia de la ponderación de los grupos

A cada grupo de variables j le corresponde una nube que representa los individuos, N_I^j situada en el espacio R^{K_j}. La ponderación que trata de equilibrar el papel de los grupos de variables consiste en dividir por λ_j^1 el peso inicial de cada variable del grupo j. Este coeficiente, al ser idéntico para todas las variables dentro del grupo j, no modifica la forma de la nube N_I^j. En cambio, normaliza estas nubes en el sentido de que, con estos pesos, la inercia máxima de toda nube N_I^j en cualquier dirección vale 1.

Por otra parte, el conjunto de variables se encuentra representado en la nube N_I situada en el espacio R^I. En esta nube el cuadrado de la distancia entre dos puntos i y l es la suma de los cuadrados de su distancia en las N_I^j. Sea i^j el punto i en la nube j, entonces

$$d^2(i,l) = \sum_{j \in J} d^2(i^j, l^j) \tag{A.2}$$

En la distancia entre dos elementos de la nube N_I, la influencia de los distintos grupos no se equilibra si las distancias en las distintas nubes N_I^j son del distinto orden de magnitud. Multiplicar los pesos iniciales de las variables del grupo j por un coefciente α_j es un medio de equilibrar la influencia de los grupos, puesto que la distancia se calcula entonces como

$$d^2(i,l) = \sum_{j \in J} \alpha_j d^2(i^j, l^j) \tag{A.3}$$

Con la ponderación $\alpha_j = 1/\lambda_j^1$, ningún grupo puede ser preponderante en la primera dirección de inercia de la nube media. No obstante, el número de direcciones de N_I sobre las que influye el grupo j crece con la dimensión de N_I^j.

En el espacio R^I est'an situados varios tipos de elementos que se desean estudiar:

1. las variables iniciales,
2. los factores obtenidos del ACP o ACM de cada uno de los grupos por separado,

3. los factores comunes a varios conjuntos de variables.

En AFM la ponderación de las variables de un grupo tiene en cuenta a la vez el número de variables y su relación. Los ejes factoriales maximizan la inercia de las proyecciones de todas las variables. La inercia proyectada de cada nube N_I^j puede ser interpretada como la contribución de un grupo. La ponderación de los grupos por $1/\lambda_j^1$ equilibra su influencia, ya que la contribución de un grupo o la construcción de un eje está limitada por el valor 1. Surge de nuevo la idea según la cual ningún grupo puede por sí solo determinar el primer eje y un grupo influye sobre más ejes cuanto mayor sea su dimensión.

A.2.2. Implementación

La implementación comprende dos etapas. En la primera se analiza cada grupo de variables por separado. Esta primera etapa es necesaria para calcular el inverso de valor propio del ACP o ACM de cada grupo que pondera o sobrepondera las variables en la segunda etapa y los primeros factores de cada grupo que se tratan como variables suplementarias en la segunda etapa. La segunda etapa es un ACP conjunto de las variables de todos los grupos ponderadas. Para los grupos cualitativos este ACP es equivalente a un ACM. Un ACM es equivalente a un AFM en el que cada grupo está constituido por las indicadoras asociadas a una misma variable.

A.3. Aspectos de la teoría de los GLMC para modelos de conteo

Se describe en este apartado la dependencia de una variable escalar y_i $(i = 1, \ldots, n)$ con un vector de regresores x_i. La distribución condicional $y_i|x_i$ es una familia exponencial lineal en la que la función de masa de probabilidad es

$$f(y; \lambda, \phi) \quad = \quad \exp\left(\frac{y \cdot \lambda - b(\lambda)}{\phi} + c(y, \phi)\right), \tag{A.4}$$

teniendo en cuenta que λ es el parámetro canónico dependiente de los regresores a través de un predictor lineal y ϕ es el parámetro de dispersión. A partir de las funciones $b(\cdot)$ y $c(\cdot)$, que son conocidas, se determina cuál miembro de la famila se usará, por ejemplo la distribución normal, binomial o de Poisson. En este caso, la media condicional y la varianza de y_i vienen dadas por $E[y_i \,|\, x_i] = \mu_i = b'(\lambda_i)$ y $VAR[y_i \,|\, x_i] = \phi \cdot b''(\lambda_i)$.

Teniendo en cuenta el parámetro de dispersión ϕ, la distribución y_i queda determinada por su media. La varianza es proporcional a $V(\mu) = b''(\lambda(\mu))$, que se conoce como función de varianza.

La dependencia de la media condicional $E[y_i \,|\, x_i] = \mu_i$ en los regresores x_i queda especificada como

$$g(\mu_i) \quad = \quad x_i^\top \beta, \tag{A.5}$$

donde $g(\cdot)$ es lo que se conoce como función de enlace, mientras que β es el vector de coeficientes de regresión, estimados en general por máxima verosimilitud (MVS) a través de un procedimiento iterativo de mínimos cuadrados ponderados.

En lugar de considerar los modelos GLM para la verosimilitud completa (tal como se determinaba con la ecuación (A.4)), en este caso los GLM se ven como modelos de regresión para la media (tal como se especificaba en la ecuación (A.5)), mientras que las funciones de estimación usadas para ajustar el modelo se derivan de una familia particular.

A.4. Árboles de clasificación

La construcción del árbol se basa en tomar distintas decisiones de forma sucesiva. Se comienza con un nodo raíz que contiene todas las observaciones, que se divide en subgrupos determinados por la partición de la variable elegida, generando nodos descendentes. Estos subgrupos se dividen usando la dicotomización de una segunda variable, y así sucesivamente hasta alcanzar los nodos terminales, que es cuando se logra un grupo lo más homogéneo posible y se obtiene la mayor representación de una clase. Se consideran todas las particiones posibles y cada partición está jerarquizada según un criterio de calidad. La regla de clasificación es sencilla: en cada nodo de decisión se verifica si el valor de cierta variable es mayor que cierto valor específico. Si es mayor se sigue el camino de la derecha y si es menor el de la izquierda.

Para el proceso de partición existe un conjunto de preguntas binarias $\mathcal{Q}$. Si la variable es cuantitativa, la pregunta será del tipo: $X_m \leq v$; si la variable es cualitativa, la pregunta será del tipo: $x_m \in \mathscr{S}$ —$\mathscr{S}$ es un subconjunto de $(b_1, b_2, ..., b_L)$ en que existen L clases posibles para la variable X—. Si la respuesta es positiva va a la izquierda, si la respuesta es negativa va a la derecha. Así se originan nodos hijos que surgen de la partición del nodo anterior. El proceso de partición de los nodos culmina con la obtención del árbol máximo. Luego se puede proceder a la poda para reducir los nodos terminales hasta llegar a un árbol óptimo.

Las particiones tienen como objetivo incrementar la homogeneidad de los subconjuntos que resulten de ella. A cada partición está asociada una medida de impureza. Cuando la impureza es mínima, las observaciones de un nodo pertenecen a una misma clase, mientras que cuando es máxima, las observaciones de un nodo pertenecen en igual proporción a las distintas clases. La medida de impureza de un nodo es el resultado de evaluar la función de impureza en ese nodo, tomando como información las proporciones $p(j/t)$, probabilidad de que una observación que está en el nodo t pertenezca a la categoría j. Existe una impureza global (para todo el árbol) y la impureza de cada nodo. Llamamos Φ a la función de impureza

$$i(t) = \Phi(p(1/t), p(2/t), ..., p(J/t)) \tag{A.6}$$

con $p(j/t) = \frac{N_j(t)}{N(t)}$.

Como medidas de impureza se tiene:

1. índice de Gini: $\sum p(i/t)p(j/t) = i(t)$
2. deviance: $D(t) = -2\sum n_{jt}p(j|t)log(p(j|t))$

La evaluación de cada subárbol se hace mediante estimaciones de las tasas de clasificación incorrecta (por sustitución, por muestra de validación o por validación cruzada).

A.5. Diseños en fases

Hay situaciones en las que existe poca información en el marco muestral relativa a los elementos de una población o esta información es de poca utilidad, con alternativas para trabajar usando diseños de tipo SI o SIC, donde se combinan con los π estimadores, pero la precisión se logra minimizar al manejar tamaños muestrales extremadamente grandes, lo que hace que no sea una estrategia de muestreo a seguir. Otra alternativa es reunir información para construir un nuevo y muy informativo marco muestral y usar luego un diseño apropiado para combinarlo con métodos de regresión. En este caso se logra disminuir el tamaño muestral, pero con un gran costo ocasionado por la construcción del nuevo marco muestral. d La última alternativa es usar lo que se denomina muestreo en dos fases, que consiste en usar en la primera etapa un diseño sencillo $p_a(.)$ con una muestra muy grande de s_a elementos. De ellos s_a se reúne información auxiliar extra. En la segunda etapa, con la ayuda de la información auxiliar, se logra seleccionar una segunda muestra s a partir de s_a con un diseño $p(.|s_a)$, donde en la submuestra s se observa la variable y en estudio. Esta nueva forma de proceder es muy importante y cada vez más usada en epidemiología, en estudios de casos y controles, cohortes, de los cuales se puede citar el estudio de tumores de Wilms (Breslow y Beckwith, 1982). Este procedimiento presenta, además, una ventaja adicional, ya que se puede usar para el tratamiento de la no respuesta. Como se expresa en (Särndal et al., 1992), en una encuesta con no respuesta la selección de la muestra probabilística pensada en un principio puede ser vista como la primera fase de selección y el conjunto r/s como la submuestra de la fase 2.

A.6. Evaluación de la necesidad del análisis multinivel

Un aspecto importante a considerar que puede resultar una guía para saber si es necesario usar el enfoque multinivel es estudiar el coeficiente de correlación intraclase (CCI) entre individuos que están clusterizados o anidados en niveles jerárquicos superiores (para el caso de los estudiantes: clase, escuela, distrito escolar; para el estudio sobre colesterol: los médicos, hospitales, ciudades) (Finch, 2014).

$$ICC = \rho = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{A.7}$$

donde τ^2 representa la varianza entre clusters y σ^2 es la varianza dentro de los clusters. Un valor elevado de ρ indicaría que la variable de respuesta está asociada con la pertenencia de la observación a un cluster o, lo que es lo mismo, los individuos dentro del mismo grupo (por ejemplo, la escuela) son más parecidos en la variable medida de lo que son estudiantes de otro cluster.

Para evaluar el CCI (Finch, 2014) propone hacer un análisis de la varianza (ANOVA) debiendo estimar $\hat{\rho}$ usando para eso $\hat{\tau^2}$ y $\hat{\sigma^2}$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(n_j - 1)S_j^2}{N - C} \tag{A.8}$$

donde S_j^2 es la varianza intraclase

$$S_j^2 = \frac{\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)^2}{n_j - 1} \tag{A.9}$$

n_j es el tamaño del cluster j, N es el tamaño total de la muestra y C es el número total de clusters. En otras palabras, σ^2 es el promedio ponderado de la varianza intraclusters.

Para estimar $\hat{\tau^2}$ se debe calcular la varianza ponderada interclase

$$\hat{S}_B^2 = \frac{\sum_{i=1}^{c}(\overline{y}_j - \overline{y})^2}{\tilde{n}(C-1)} \tag{A.10}$$

donde $\overline{y}_j$ es la media de la variable de respuesta en el cluster j y $\overline{y}$ es la media global

$$\tilde{n} = \frac{1}{C-1}\left[N - \frac{\sum_{i=1}^{c} n_j^2}{N}\right] \tag{A.11}$$

Como no se puede usar directamente $\hat{S}_B^2$ como la estimación de τ^2, ya que está impactada por la variación aleatoria de los individuos dentro de los clusters, se corrige mediante la siguiente relación

$$\hat{\tau}^2 = \hat{S}_B^2 - \frac{\hat{\sigma}^2}{\tilde{n}} \tag{A.12}$$

de donde puede calcularse

$$\rho_I = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{S}_B^2} \tag{A.13}$$

Datos del autor

Ramón Álvarez Vaz es técnico en Estadística, magíster en Epidemiología y doctor en Ciencias de la Salud. Actualmente desarrolla su actividad laboral en la Facultad de Ciencias Económicas y de Administración y en la Facultad de Odontología. Se dedica a la investigación en estadística aplicada en torno al uso e incorporación de herramental estadístico (https://orcid.org/0000-0002-2505-4238)

Tiene publicados en la colección Biblioteca Plural sus libros Maloclusión en niños en edad escolar: Análisis de los factores de riesgo (2011), Métodos de muestreo para estudios sanitarios: una aplicación para la Encuesta de Factores de Riesgo 2006 en Uruguay (2017) y La investigación en estadística desde la mirada de cuatro jóvenes investigadores (2018), este último en coautoría.

En salud pública, es fundamental conocer las características de las poblaciones y sus problemas de salud sobre la base de fuentes de datos como estadísticas vitales, registros específicos de enfermedades y sistemas de vigilancia epidemiológica.

Cuando estos datos no están disponibles, se recurre a estudios sanitarios, como encuestas y muestreos. Sin embargo, los indicadores epidemiológicos tradicionales tienen limitaciones, ya que a menudo no consideran la estructura multivariada de la información y la simplifican a través de algoritmos.

Para abordar esto, se proponen nuevos enfoques en los que se utilizan técnicas estadísticas avanzadas, que preservan la complejidad de los datos y permiten generar indicadores más precisos.

Estos métodos mejorados permitirán predecir la distribución espaciotemporal de las patologías bucales, utilizando información habitual de manera diferente y más precisa.

COMISIÓN SECTORIAL DE INVESTIGACIÓN CIENTÍFICA
CSIC

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

ISBN: 978-9974-0-2234-8