proinbio
Programa para la Investigación Biomédica

ACTA

TESIS DE DOCTORADO EN CIENCIAS MÉDICAS

Título: *"Sistematización y creación de Indicadores e Indices para la vigilancia epidemiologíca en Salud Bucal: Uso de Técnicas estadísticas multivariantes y de Análisis espacio-temporal"*

De: Msc. Rámon Álvarez Vaz

12 de febrero de 2020

INTEGRACIÓN DEL TRIBUNAL:

Presidente Dr. Marco Brito Correa
Dra. Cecilia Severi
Dr. Gonzalo Perera.

Dir. Académico Dr. Enrique Barrios
Co-director: PhD. Gabriel Camaño.

Fallo del Tribunal Nota[1] EXCELENTE Escala numérica[2]: 12 -

1 Excelente. Muy satisfactorio. Satisfactorio. Aceptable y No aprobado.
2 Utilizar escala de 1 al 12

proinbio
Programa para la Investigación Biomédica

JUICIO COMPLEMENTARIO ESCRITO (HASTA UN MÁXIMO DE 300 PALABRAS)

Excelente trabajo, haciendo uso exhaustivo y riguroso de diversas técnicas estadísticas aplicadas pertinentemente a la epidemiología particularmente odontológica.
El texto y la presentación fueron muy claros y completo

Dr. Marco Brito Correa Dra. Cecilia Severi Dr. Gonzalo Perera.

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Sistematización y creación de Indicadores e Índices para la vigilancia epidemiológica en Salud Bucal:

Uso de técnicas Estadísticas Multivariantes y de Análisis Espacio-Temporal

Ramón Álvarez-Vaz

Programa de Investigación en Biomedicina
Facultad de Medicina
Universidad de la República

Montevideo – Uruguay
Febrero de 2020

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Sistematización y creación de Indicadores e Índices para la vigilancia epidemiológica en Salud Bucal:

Uso de técnicas Estadísticas Multivariantes y de Análisis Espacio-Temporal

Ramón Álvarez-Vaz

Tesis de Doctorado para el Programa de Investigación en Biomedicina, Facultad de Medicina de la Universidad de la República, como requisitos necesarios para la obtención del título de Doctor en Ciencias Médicas

Director:
Dr. Enrique Barrios

Codirector:
PhD. Gabriel Camaño

Director académico:
Dr. Enrique Barrios

Montevideo – Uruguay
Febrero de 2020

Álvarez-Vaz, Ramón

Sistematización y creación de Indicadores e Índices para la vigilancia epidemiológica en Salud Bucal: / Ramón Álvarez-Vaz. - Montevideo: Universidad de la República, Facultad de Medicina, 2020.

xxxvi, 302 p.: il.; 29, 7cm.

Director:

Enrique Barrios

Codirector:

Gabriel Camaño

Director académico:

Enrique Barrios

Tesis de Doctorado – Universidad de la República, Programa de Investigación en Biomedicina, 2020.

Referencias bibliográficas: p. 246 – 267.

1. Salud bucal, 2. Vigilancia epidemiológica, 3. Indicadores, 4. Análisis Multivariante, 5. Análisis Espacio-Temporal. I. Barrios, Enrique, *et al.* II. Universidad de la República, Programa de Investigación en Biomedicina. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

PhD. Marcos Brito Correa

PhD. Cecilia Severi

PhD. Gonzalo Perera

Montevideo – Uruguay
Febrero de 2020

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 mis hijos Bruno y Lucía, que son mi orgullo.

Agradecimientos

Cuando corresponde agradecer, el espacio nunca es suficiente y aparece el miedo de que la memoria nos traicione y queden personas por considerar. Este es un trabajo colectivo en el que participaron muchas personas (colegas, docentes, compañeras y 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, por su infinita paciencia y por apoyarme en este desafío que me toma ya en la etapa madura de mi vida y de mi actividad como investigador. En especial a Fernando, con quien desarrollé una parte importante de mis hallazgos de la tesis y de quién me siento orgulloso y me quedo tranquilo en ver en él un relevo, para cuando ya me dedique a otras cosas.

Para Silvia Rodríguez, más próxima generacionalmente a mí un reconocimiento a su especial pedido de que yo no dejara, cuando las fuerzas flaqueaban por estar haciendo todo esto ya de veterano. También decirle, ahora que ella está en la misma etapa, que va a llegar, a pesar de los obstáculos que hoy en día aún implica ser mujer, madre en el ambiente académico y querer cumplir en sus muchas tareas y avanzaro, pero estoy seguro que lo va a lograr.

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 mejor visibilizar 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. Para Susana (encargada del Servicio) compañera en esta locura, de hacer el doctorado del Pro.In.Bio ya maduros mis felicitaciones porque acaba de lograrlo; a ella y a la Tana (Anunzziatta) por la paciencia de ambas en acostumbrarse a convivir y discutir con

alguien como yo, proviniendo del mundo de los números y aprender a dialogar. También un agradecimiento por su apoyo cuando tuve que consultarlos, a Graciela Buño, Alicia Picapedra, Ernesto Andrade, Marcelo Kreiner 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 reflexionar y tratar 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.

En este período del doctorado también tuve la suerte de profundizar con otros investigadores de muy diferentes áreas de la salud de quien me enriquecí mucho como investigador: Paula Moliterno, Estela Skapino, Juan Dapueto, Fiorella Cavalleri, Alejandra López, entre otras y otros. Muy especialmente quiero agradecer al Dr. José Boggia con quien compartí muchas charlas,trabajos, tuve la suerte de poder publicar con él, y sobre todo de contar con su apoyo, planteando firmememente antes sus colegas del Pro.in.Bio, que mi trabajo era muy diferente pero a su vez importante para la biomedicina, lo cual me permitió llegar al día de hoy, donde puede presentar mi trabajo de tesis y defenderlo. Pepe es (y no lo descubro yo) un magnífico investigador, joven, generoso y ciertamente visionario, a quien agradezco sus consejos y con quien me encanta poder compartir la pasión por investigar.

Agradezco a mis tutores su invalorable apoyo, ya que Enrique (Prof. Dr. Enrique Barrios) siempre me hizo sentir que iba por el lado correcto, dándome la libertad de hacer una tesis muy poco ortodoxa pero haciéndose tiempo para ver mis avances, y hacerme sugerencias a pesar de que mucho de los aspectos que fui incorporando a la tesis no eran estrictamente epidemiológicos. A Gabriel (PhD. Prof. Gabriel Camaño), mi otro tutor, y director del IESTA, por ser incondicional a mi intento de traer la estadística a las ciencias médicas, conociendo bien él el esfuerzo que eso supone y que intuyo es lo que nos cautivó y llevó a estudiar hace casi 35 años estadística, cuando nos conocimos, tratando de comprender y ayudar a entender mejor los problemas de la salud, la economía y de otros disciplinas científicas (primero como colega y luego como codirector de tesis).

Un especial agradecimiento a los autores de UdelarTex, Mihdí Caballero, Pablo Castrillo, Virginia Bertolotti y demás integrantes de la Comisión

Académica de Posgrado (CAP) de la Udelar. En (http://heisenberg.csic.edu.uy/TallerTesis/UdelaRTeX.git) se puede encontrar el modelo prediseñado (clase o plantilla) en LaTeXconcebido para los que escriben sus tesis de posgrado en programas de la Udelar, con la que pude armar este trabajo de tesis.

Y no puedo cerrar esta sección sin agradecer a Hugo (Dr. Hugo Dibarboure), porque principio tienen las cosas y es gracias a él, compañero de aventuras en este asunto de ayudarme primero a terminar la maestría y luego escucharme y apoyarme cuando decidí imitarlo e ir siempre por más, tratando de hacer el doctorado. 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, Febrero de 2020

"Decir lo que se piensa, hacer lo que se dice"..*Extraído del último discurso del General Seregni en el Paraninfo de la Universidad de la República, en ocasión de recibir el doctorado honoris causa por parte de la Udelar, 19 de Marzo de 2004*

Resumen

En el ámbito de la salud pública, existe la necesidad de conocer en profundidad las características de las poblaciones y los problemas de salud y de ese modo poder intervenir para mejorarlos. Esto significa que es necesario por lo menos tener una idea de la situación de partida y para eso se recurre a las fuentes de datos existentes. Entre ellas se destacan las estadísticas vitales; los registros de problemas específicos de salud (los registros de cáncer por ejemplo que son registros de base poblacional), que permiten entre otras cosas establecer la incidencia de la enfermedad; 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 diferentes mecanismos de generación a través del método científico de los diferentes diseños de estudios sanitarios, que incluyen los mecanismos de muestreo y las encuestas.

Sin embargo pueden existir limitaciones en los indicadores generalmente utilizados en la epidemiología y salud pública, 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 cálculo que generan indicadores univariados para ganar en simplicidad, y no miden por lo tanto correctamente los fenómenos bajo estudio.

Teniendo en cuenta los antecedentes antes planteados con respecto a las fuentes de información en salud y en salud bucal en particular, se propone presentar un conjunto de indicadores alternativos y complementarios a los que ya existen. Se reformulará la forma de considerar la información que ya se viene recogiendo y para los cuales existen ya varias índices recomendados de la Organización Mundial de la Salud (CPO, CPI, ICDAS, IHOS) y otros índices epidemiológicos sobre estado de la salud bucal, usando para estos 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 muevas que provienen de otras disciplinas.

Para eso la propuesta a desarrollar consiste en tratar de elaborar y sistematizar este conjunto de indicadores epidemiológicos a través de técnicas estadísticas multivariantes de aprendizaje supervisado y no supervisado, combinados con otras como los modelos de conteo, la teoría de respuesta al ítem, el análisis de redes sociales y la teoría de la información, técnicas que no son muy

usadas en el ámbito de la epidemiología en Uruguay. Con ellas se espera poder construir tipologías o grupos de poblaciones con perfiles epidemiológicos bien diferenciados de acuerdo a las patologías y los factores de riesgos asociados.
Por último incorporar la dimensión espacio-temporal, indispensable en la vigilancia epidemiológica actual e intentar sistematizar las fuentes de información disponibles para poder proponer los nuevos indicadores y evaluar finalmente la aplicabilidad de los mismos y la sustentabilidad de los sistemas de vigilancia integrados por estos indicadores, en el tiempo y distribución territorial del país.

Abstract

In the field of Public Health, there is a need to know in depth the characteristics of populations and health problems. This means that it is necessary to at least have an idea of the starting situation and for that, the existing data sources are used. Among them, vital statistics stand out; records of specific health problems (cancer registries that are population-based registries), which allow, among other things, to set up the incidence of the disease; records of diseases of infectious etiology, with mandatory notification on which the epidemiological surveillance systems are based.

When the information that the researcher in biomedicine needs is not available, different generation mechanisms must be resorted through scientific method with different designs of health studies, including the sampling mechanisms and the surveys.

However, there may be limitations in the indicators generally used in epidemiology and public health, since many times they do not take into account the multivariate structured information or if they take it, they do it through simple calculation algorithms to generate univariate indicators and therefore do not correctly measure the phenomena under study.

From this background previously raised regarding the sources of health information and oral health in particular, we propose to present a set of alternative and complementary indicators to those that already exist. The way of considering the information that has already been collected and for which there are already several recommended indexes of the World Health Organization (DMF, CPI, ICDAS, IHOS) and other epidemiological indexes on the state of oral health, will be reformulated, using different statistical techniques, some of them frequently used to solve the problem of preserving the multivariate structure of the information and in turn apply techniques that come from other disciplines.

For that, the proposal to be developed consists of trying to elaborate and systematize this set of epidemiological indicators through multivariate statistical techniques of supervised and unsupervised learning, combined with others such as Counting Models, the Item Response Theory, Social Networks Analysis and Theory of Information, techniques that are not widely used in the field of epidemiology in Uruguay. With these strategies it is expected to be able to build typologies or groups of people with well differentiated epidemiological profiles according to several pathologies and the risk factors.

Finally, try to incorporate the space-time dimension, indispensable in the current epidemiological surveillance and systematize the available information sources to be able to propose the new indicators and after evaluate the applicability of them and the sustainability of the surveillance systems integrated by these indicators, in time and space in the country.

Producción escrita que surge del trabajo de tesis

Se detallan a continuación los diferentes productos generados, algunos como publicaciones en revistas o libros y otros presentados en el formato de "conference paper" en jornadas académicas y congresos nacionales e internacionales. En ambas secciones se sigue el orden cronológico de cada producto y en caso de que algunos de ellos hayan sido elaborados como parte de tesis en el marco de algún programa de posgrado (maestría o doctorado) se menciona explícitamente. Para más detalles dejo enlace a mi **ORCID :** https://orcid.org/0000-0002-2505-4238

Producción escrita en coautoría

En esta sección se consigna la producción escrita, en donde participo como coautor, es decir no soy primer autor del artículo, capítulo de libro, memoria de diferentes congresos o lo que corresponda.

Artículo -"Primer Relevamiento Nacional de Salud Bucal en población joven y adulta uruguaya: Aspectos metodológicos", (Lorenzo *et al.*, 2013a), https://hdl.handle.net/20.500.12008/2626.

Artículo -"Enfermedad Periodontal en la población joven y adulta uruguaya del Interior del país: Relevamiento Nacional 2010-2011", (Lorenzo *et al.*, 2013b), https://hdl.handle.net/20.500.12008/2627.

Artículo -'Caries dental. La enfermedad oral más prevalente: Primer Estudio poblacional en jóvenes y adultos uruguayos del interior del país, (Olmos *et al.*, 2013), https://hdl.handle.net/20.500.12008/2634.

Artículo -"Prevalencia y factores de riesgo de las lesiones de la mucosa oral en la población urbana del Uruguay", (Casnati *et al.*, 2013), https:

//hdl.handle.net/20.500.12008/2573.

Artículo -"Prevalencia de maloclusiones en adolescentes y adultos jóvenes del interior del Uruguay. Relevamiento nacional de salud bucal 2010-2011", (Ourens *et al.*, 2013), https://hdl.handle.net/20.500.12008/2635.

Artículo -"Craniofacial Pain of Cardiac Origin Is Associated with Inferior Wall Ischemia", (Kreiner *et al.*, 2014), https://doi.org/10.11607/ofph.1257

Artículo -"Erosive Tooth Wear among 12-Year-Old Schoolchildren: A Population-Based Cross-Sectional Study in Montevideo, Uruguay", (Alvarez Loureiro *et al.*, 2015), https://doi.org/10.1159/000368421.

Artículo -"Periodontal conditions and associated factors among adults and the elderly: findings from the first National Oral Health Survey in Uruguay" (Tesis de Doctorado-Pro.in.Bio), (Lorenzo *et al.*, 2015), https://hdl.handle.net/20.500.12008/11099 .

Artículo -"Craniofacial Pain Can Be the Sole Prodromal Symptom of an Acute Myocardial Infarction. an Interdisciplinary Study", (Kreiner *et al.*, 2010), https://hdl.handle.net/20.500.12008/11091.

Artículo -"Comparative effectiveness of water and salt community-based fluoridation methods in preventing dental caries among schoolchildren",(Tesis de Maestría), (Fabruccini *et al.*, 2016), https://doi.org/10.1111/cdoe.12251.

Artículo -"The role of contextual and individual factors on periodontal disease in Uruguayan adults", (Tesis de Doctorado-Pro.in.Bio), (Lorenzo-Erro *et al.*, 2018), https://doi.org/10.1590/1807-3107bor-2018.vol32.0062.

Artículo -"Nível socioeconômico na primeira infância e oclusopatia em adolescentes e adultos jovens no Uruguai",(Tesis de Maestría), (Goettems *et al.*, 2018), http://dx.doi.org/10.1590/0102-311X00051017.

Trabajos Completos arbitrados en Eventos -"Item Response Theory modelling assessment of oral health in a Uruguayan population study" 33rd International Workshop on Statistical Modelling Volume II, july 2018, (Álvarez-Vaz y Massa, 2018c).

Trabajos Completos arbitrados en Eventos - "Characterization of morbidity and mortality from automobile accidents using temporal spatial epidemiological tools in Brazil between 2000 and 2015", (Álvarez-Vaz *et*

al., 2018c).

Artículo -"Sex Determination in a Brazilian Sample from Cranial Morphometric Parameters". Carlos Sassi, Alicia Picapedra, Ramón Álvarez Vaz, Eduardo Da Luz Jr, Luiz Francesquini Jr. En fase de referato en la revista 'Journal of Forensic Odontoestomatology' enviada en Octubre de 2018.

Working paper "Modeling in an oral health study through two statistical methods in Uruguay", abril 2019, *bioRxiv*, doi: http://dx.doi.org/10.1101/611921.

Producción escrita como primer autor

En esta sección se consigna la producción escrita, en donde participo como autor principal, solo o en coutoría para artículo, capítulo de libro, memoria de diferentes congresos, documentos de trabajo, según corresponda.

Trabajos Completos arbitrados en Eventos -"Modelos predictivos para los componentes del CPO mediante Regresión Beta en una encuesta de base poblacional", (Álvarez-Vaz y Vernazza, 2012).

Documento de Trabajo -"Distribución Bernoulli Multivariada. Una aplicación a la salud oral", (Álvarez-Vaz y Massa, 2014), https://hdl.handle.net/20.500.12008/10540.

Trabajos Completos arbitrados en Eventos -"Elaboración de perfiles epidemiológicos en estudios sanitarios mediante técnicas de clustering difuso", (Álvarez-Vaz y Massa, 2017).

Artículo - "Aplicación de técnicas de clustering para la elaboración de perfiles epidemiológicos en estudios sanitarios" en arbitraje, enviado a revista Saberes de Estadística de la Universidad de Rosario, Argentina, (Álvarez-Vaz y Massa, 2018a).

Capítulo de libro -"Uso de la Distribución Bernoulli Multivariada en salud bucal" capítulo de libro, Puebla, México,en prensa (Álvarez-Vaz y Massa, 2018d), https://iesta.fcea.udelar.edu.uy/publicaciones/libros/.

Trabajos Completos arbitrados en Eventos -"Uso de la Distribución Bernoulli Multivariada en salud" bucal, (Álvarez-Vaz y Massa, 2018e), https://iesta.fcea.udelar.edu.uy/publicaciones/libros/.

Trabajos Completos arbitrados en Eventos -"Evaluación de la Salud Bucal a través de la Teoría de la respuesta al Ítem en un estudio poblacional en Uruguay", Puebla, México, (Álvarez-Vaz y Massa, 2018b).

Trabajos Completos arbitrados en Eventos -"Comparing 2 methods of statistical modeling of oral health in a population study in Uruguay", (Álvarez-Vaz *et al.*, 2018a).

Trabajos Completos arbitrados en Eventos -"Visualization for the multivariate structure of the components of the DMFT using analysis of compositional data", (Álvarez-Vaz *et al.*, 2018b).

Documento de Trabajo - "Medición Y caracterización de las desigualdades de salud bucal entre escolares de 12 años de Montevideo, Uruguay"(Álvarez-Vaz, 2019), https://iesta.fcea.udelar.edu.uy/wp-content/uploads/2020/12/ddt_02_19.pdf.

Trabajos Completos arbitrados en Eventos -"Creación de indicadores alternativos para la vigilancia en salud oral mediante Regresión Beta", Puebla, México, (Álvarez-Vaz *et al.*, 2019c).

Trabajos Completos arbitrados en Eventos -"Aplicación de análisis de redes para la elaboración de perfiles epidemiológicos en estudios sanitarios", Puebla, México, (Álvarez-Vaz *et al.*, 2019a).

Artículo -"Determinación del sexo mediante técnicas de clasificación supervisada", enviado junio 2019 a arbitraje y aceptado Noviembre 2019 en Revista de Facultad de Ciencias de la Universidad Nacional de Colombia, (Álvarez-Vaz y Sassi, 2020).

Artículo -"Modelos de conteo alternativos para los componentes del Cpo en un estudio poblacional", enviado junio 2019 a arbitraje de Revista de Facultad de Ciencias de la Universidad Nacional de Colombia, (Álvarez-Vaz *et al.*, 2019b)

Lista de figuras

Lista de tablas

Lista de abreviaturas y siglas

A. Multinivel Análisis multinivel 62, 63, 64, 235

AC Análisis de Cluster o Conglomerados 40, 41

ACM Análisis Factorial de Correspondencias Múltiples 40, 41

ACP Análisis de Componentes Principales 40, 41

AD Análisis Discriminante Probabilístico 213

ADC Análisis de Datos Composicionales xxxiv, 170, 174, 175, 182, 183, 239

AF Análisis Factorial 40, 41

AIC Akaike Information Criteria 105, 112

ASSE Administración de servicios de salud (ASSE) 32

AUC Área bajo la curva 84, 92

Age-period-cohort Analysis Age-period-cohort Analysis 245

Árboles de Clasificación y Regresión Árboles de Clasificación y Regresión 42, 43, 136, 214, 215, 224

BIC Bayesian information criterion 105, 112

BN Distribución Binomial Negativa xviii, 39, 40, 75, 77, 78, 94, 95, 96, 97, 111, 112, 113, 114, 115

BN - tipo I Binomial Negativa 95, 107

BN - tipo II Binomial Negativa 95, 107

CC Curva de concentración 190, 199

CCC Curva de Carga 202, 206, 207

CCI Correlación Intraclase 282

CIE-OE Clasificación Internacional de Enfermedades Aplicada a la Odontología y Estomatología 21

CIE10 Décima versión de la Clasificación Internacional de las Enfermedades 2

CL Curva de Lorenz 189, 190, 199

CPI Índice Periodontal Comunitario 24, 185, 186

CPITN Índice Periodontal Comunitario De Necesidad De Tratamiento 22

CPO Índice CPO 19, 35, 185, 186

CPO-grama CPO-grama xxxiv, 177, 178, 179, 181, 182

CTE Cociente de Tasas Extremas 188

Classification and Regression Trees Classification and Regression Trees xxxv, 214

D. Poisson Distribución de Poisson 39, 49, 75, 77

DAI Indice de Estética Dental xxxi, 26, 27

DIC Distancia Intercanina 216

DKL Discrepancia de Kullback-Liebler 193, 195

DMD Diámetro Mesiodistal 216

DP Doble Poisson 108, 237, 285

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 xxxv, 74, 215, 229

DT Diferencia de Tasas Extremas 188

Deff Efecto "diseño" 60

ECSC Escuela de Contexto Socioeconómico Crítico 196, 197

ECT Encuestas de Corte Transversal (cross-section) 33

ECU Escuela Pública Común Urbana 196, 197

EL Encuestas Longitudinales (panel data) 33

ENT Enfermedades No Transmisibles xxii, 5, 8, 32, 137, 146, 153, 159, 160, 165, 238, 242, 243

EP Escuela Privada 196, 197

ET Enfermedades Transmisibles 5, 44

ETC Escuela Pública de Tiempo Completo 196, 197

Esp Especificidad 38, 83, 93, 221, 228

GT Gráficos Triangulares xxxiv, 171, 173, 177, 179, 181, 182

I de Geary Índice de Geary 54

I de Gini Índice de Gini 189, 207

I de Hoover Índice de Hoover 193, 195, 206

I de Moran Índice de Moran 54

I de Rao Índice de Rao 212

I de Redundancia Índice de redundancia 194

I de Theil Índice de Theil 193, 195, 207

I. Alternativos Indicadores Alternativos 34, 233, 234

I. Clásicos Indicadores Clásicos 34

I. Combinados Indicadores Combinados 34, 233, 234

I. Espacio-Temporales Indicadores Espacio-Temporales 35, 233, 234

I.M. Canino Índice Mandibular Canino 212

ICDAS International Caries Detection and Assessment System 21, 35

ICM Índices Caninos Maxilares Estándares 217, 218

IG Índice Gingival 22

IG Inversa Gaussiana 97

INE Instituto Nacional de Estadística 71

INSE Índice de Nivel Socioeconómico 197

IPK Índice de Pearcy-Keppel 191, 192, 205

IPKp Índice de Pearcy-Keppel Ponderado 192, 205

Kappa Índice de Concordancia Kappa 21

M. Binomial Negativo(p) Modelo Binomial Negativo-P 97

M. de Poisson Modelo de Poisson 39, 94, 95, 98, 114

MBN Modelo Binomial Negativo 40, 93, 95, 98, 237

MC Modelos de Conteo xix, xxii, xxxiii, 39, 78, 94, 101, 103, 105, 113, 116, 133, 233

MDC Modelos de Datos Categóricos 37

MEC Modelos con Exceso de Ceros 100, 237

MH Modelos Hurdle o con obstáculos xxii, 99, 111, 112, 115, 237

MLG Modelos Lineales Generalizados 37, 39, 77, 112, 120, 161, 234, 238

MMult Muestreo en varias etapas 56

MRL Modelos de Regresión Lineal 37

MSP Ministerio de Salud 32

NT Necesidades de Tratamiento 23

OMS Organización Mundial de la Salud 3, 21, 138

PEC Post-Estratificación Completa o Calibración sobre totales de celdas conocidos 61, 78

PG Poisson Generalizada xix, 95, 98, 103, 104, 107, 110, 113, 237, 285

PIG Poisson Inversa Gaussiana xix, 94, 95, 97, 237

PRNSB2011 Primer Relevamiento Nacional de Salud Bucal en población joven y adulta xxi, xxxii, 70, 71, 78, 81, 236

R. Lin Regresión Lineal 234

R. Logística Regresión Logística xxi, 43, 76, 90, 91, 92, 93, 240

R. Poisson Regresión Poisson xxi, 90, 236

RACA2012 Relevamiento y Análisis de Caries Dental en Adolescentes de 12 años de la ciudad de Montevideo,Uruguay xxiv, xxxii, xxxiv, 72, 73, 196, 204, 236, 239, 241

RAKE Post-Estratificación Incompleta o Calibración sobre las marginales conocidas 61, 62

RAP Riesgo Atribuible Poblacional 188, 189, 204, 205

RAPr Riesgo Atribuible Poblacional Porcentual 188, 189, 205

REM Razón Estandarizada de Mortalidad 49, 50

ROC Curva ROC 39, 83, 92, 240

RPAFO2015 Relevamiento en Población que se asiste Facultad de Odontología durante 2015-2016 xxi, xxii, xxxii, xxxiii, xxxiv, 73, 94, 101, 102, 113, 122, 125, 126, 127, 128, 129, 130, 131, 132, 135, 146, 148, 153, 164, 175, 181, 182, 236, 239

Regresión Beta Regresión Beta xxxiii, 121, 122, 133, 136, 234, 237, 243

SI Muestreo Aleatorio Simple sin reposición de tamaño n xxxi, 55, 56, 58, 59

SIC Muestreo por Conglomerados 56

SIG Sistema de Información Geográfica 54

SM Simulación Monte Carlo 49, 51, 53

SNA Análisis de Redes Sociales (Social Network Analysis) xxxiii, 139, 140, 153, 159, 238, 244

STSI Muestreo Estratificado 56, 59

SY Muestreo Sistemático 56

Sen Sensibilidad 38, 83, 93, 221, 228

TK Tabla de Kish 57

TNR Tasa de No Respuesta 78

TRI Teoría de Respuesta al Ítem 161, 238, 243

TTM Trastornos temporomandibulares xxxi, 17, 27, 28

UG Unidades Geodemográficas 186, 188, 192, 193, 196, 197, 206, 207, 239

UPM Unidades Primarias de Muestreo 57

VE Vigilancia Epidemiológica 3, 21, 232

VEG Índice de Varianza entre grupos 191, 192, 205

VEGr Índice de Varianza relativa entre grupos 192

VP+ Valor Predictivo Positivo 93

VP- Valor Predictivo Negativo 93

Vario Variogramas 54

Tabla de contenidos

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 en (Oriol, 1997), existen diferentes fuentes de datos para generar indicadores. Entre ellas se encuentran las estadísticas vitales; 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; 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 algunas de las fuentes antes mencionadas, se debe recurrir a diferentes 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. Así entonces 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 actualmente en forma protocolizada a través de la Décima versión de la Clasificación Interna-

cional de las Enfermedades (CIE10) (www.who.int/classifications/en/).

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), que es 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.

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 y 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, la administración, apoyándose a su vez en la Estadística para construir una gran diversidad de indicadores que deben cumplir algunas características tal como presenta por ejemplo, (Silva, 1998)

- validez de aspecto
- validez de contenido
- validez de criterio o concurrencia
- capacidad predictiva
- fiabilidad o reproducibilidad

Todos estos conceptos son válidos para la vigilancia en salud pública 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), y que se usan en encuestas de base poblacional como por ejemplo en Brasil, (Ministério da Saúde, 2003).

Hasta la fecha los antecedentes de estudios a nivel nacional consideran la aplicación de algunos de esos índices, (Beca *et al.*, 1996), (Bianco *et al.*, 1997).

Sin embargo estos indicadores son inadecuados a la hora de ser usados en la toma de decisiones para generar acciones concretas con el propósito de me-

jorar la salud oral de la población. Esto se debe a que no considera toda la información necesaria a ser usada 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 georeferenciar la información y 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; el usar técnicas estadísticas multivariantes recientes pueden 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 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 solamente en salud oral, por lo menos en nuestro país.

Otro aspecto a ser considerado 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 al provenir de diferentes fuentes de información. En Uruguay se pueden destacar 2 tipos de fuentes en salud oral y que tienen niveles de existencia en el tiempo y desarrollo o profundidad muy heterogéneos.

- registros de atención en el primer nivel de atención en el ámbito de atención del sector público y privado;
- encuestas de base poblacional (Son muy pocas las que se han desarrollado con alcance nacional).

Estas 2 grandes fuentes de información hacen que en realidad la población sobre la que se pretende hacer vigilancia en salud oral puede ser dividida en 3 grupos de personas

- población de las personas que consultan;
- población de las personas que tienen cobertura de algún tipo y no consultan;
- población a nivel 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 necesariamente 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 las ENT de la que forman parte las patologías orales. Las Enfermedades Transmisibles (ET) como por ejemplo los virus respiratorios, la Hepatitis A, el VIH, el Dengue, 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 "espacio-temporal" 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 el anteproyecto de tesis presentado para acceder al programa de doctorado (que se adjunta en apéndice C) y de 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 bucal, reformulando la forma de considerar la información que ya se viene recogiendo, usando las técnicas estadísticas pertinentes.

1.3. Objetivos Específicos

1. Elaborar un conjunto de indicadores epidemiológicos combinando 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. Desarrollar un conjunto de indicadores epidemiológicos alternativos a los clásicos, usando la misma información habitualmente utilizada 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. Presentar indicadores que den cuenta de la distribución espacio-temporales de las patologías orales en estudio (ver sección 4).
4. Identificar los problemas que surgen en el análisis epidemiológico al no considerar el proceso de generación de la información (ver sección 3.6 y 3.7).

1.4. Estructura de la tesis

En este trabajo de tesis se presentan y se referencian una serie de publicaciones que a juicio del autor son importantes integrar a la tesis, ya que en estos se 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 de bajo que formato se presentan los diferentes trabajos responde a la necesidad de poder ser mostrados todos y en la secuencia temporal en que fueron creados o publicados si corresponde en coautoría, no siendo 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 12, siendo ésta producción en primer autoría, donde el énfasis es esencialmente en aplicaciones donde lo que más importa y resulta novedoso es el uso de técnicas estadísticas muy poco usadas para el ámbito de la biomedicina, siendo capítulos que se caracterizan por tener un desarrollo exhaustivo de las técnicas usadas.

Por lo tanto los temas que se desarrollan en los capítulos de aplicaciones son para el autor de la tesis, donde se logra el mayor aporte, para el conoci-

miento de otros investigadores biomédicos, que puedan tomarlos como nuevas herramientas de solución a sus problemas. Es entonces ésta la forma adecuada de presentar la línea de investigación con producción escrita publicada en artículos, que se complementa con los capítulos desarrollados en las aplicaciones, logrando responder los objetivos planteados en las secciones 1.2 y 1.3 del capítulo 1. De cualquier modo todas las aplicaciones y sus resultados se retoman en el capítulo 13, donde se sintetiza el trabajo de tesis.

Capítulo 2

Antecedentes

A nivel de la salud bucal existen muchas dimensiones que deben ser evaluadas a nivel individual. Algunos ejemplos son:

- estados de las piezas dentales (Odontograma);
- estado de las mucosas o tejidos (Examen Local);
- síntomas asociados con los aspectos funcionales (articulares, oclusión) y hábitos (higiene,dieta).

Por lo tanto teniendo en cuenta las dimensiones antes planteadas que se consideren, es necesario ver las patologías que aparecen y la forma de medirla lo que da lugar a varios indicadores que también se presentan a lo largo del capítulo.

2.1. Diferentes patologías en Salud Bucal

Los problemas de salud en la mayor parte del mundo han cambiado, siendo las enfermedades no transmisibles (ENT) las de mayor prevalencia, donde este nuevo patrón global está estrechamente relacionado a los estilos de vida de las sociedades modernas (Breilh, 2010). Este cambio también impactó en la salud bucal, siendo las enfermedades bucodentales unos de principales problemas en la salud pública debido a su alta prevalencia e incidencia en todas regiones del mundo, (Peterson, 2004).

Antes de pasar a presentar las mas importantes patologías bucales y los índices para evaluarlas es necesario consignar previamente que en odontología existe una forma de referirse las piezas dentales, la que reciben cierta numeración que es coherente también con la medición de patologías de la mucosa.

Las piezas se suelen también agrupar en sextantes (inferiores y superiores) o cuadrantes (inferiores y superiores y a su vez izquierdos y derechos). En la Figura 2.1 denominado Odontograma puede verse la disposición de las piezas en la boca

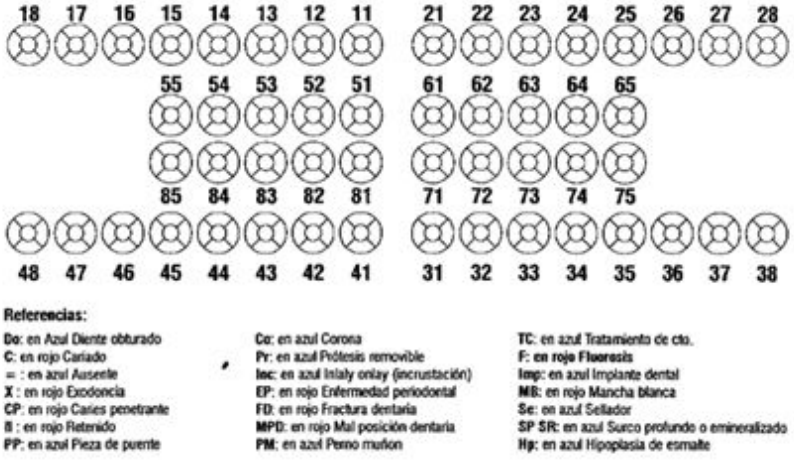

Figura 2.1: Odontograma.

donde 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 bio-película, caracterizado por desequilibrios en los procesos desmineralización-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 (principalmente iones calcio y fosfato) para la bio-película y saliva resulta en el establecimiento de la lesión de Caries (Holst *et al.*, 2001). Algunos componentes del proceso de Caries actúan en la superficie del diente (saliva, bio-película, dieta, acceso al flúor), mientras que existe otro conjunto de factores que determinan el comportamiento de la persona (conocimiento, actitud, ingresos, nivel educativo y socioeconómico) (Fejerscov, 2004). El proceso de la enfermedad debe ser sujeto a un control permanente a lo largo de la vida con el fin de evitar consecuencias irreversibles en etapas posteriores (Maltz y Jardim, 2010).

2.1.2. Enfermedad Periodontal

Las enfermedades periodontales son enfermedades infecciosa-inflamatorias, que de acuerdo al 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. Se produce por la acumulación inespecífica de especies bacterianas (bio-película) y se elimina mediante un buen control de la misma, (Cortés, 1999). En el caso de una depresión del estado inmunitario, presencia de factores de riesgo y mediadores pro-inflamatorios, así como el aumento de bacterias periodontopatógenas, es posible que a partir de una Gingivitis se desarrolle una Periodontitis: esta se produce cuando la inflamación de la encía pierde inserción, existe 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 existen criterios para determinar "caso de Periodontitis" para usar en diversos estudios de investigación. Por un lado el quinto Workshop Europeo del 2005, propone 2 niveles de criterios para la definición de un caso de Periodontitis:

1. presencia de pérdida de inserción proximal $\geq 3mm$ en 2 o más piezas no adyacentes, permite definición de caso incipientes;
2. presencia de pérdida de inserción proximal $> 5mm$ en por lo menos 30 % de las piezas posteriores, permite definición de caso más específico (identifica casos de extensión y severidad). Este criterio no está diseñado para evaluar la prevalencia de 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. 0 ausencia de Periodontitis moderada o severa
2. 1 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. 2 Periodontitis severa 2 o más sitios inter-proximales con LAC $\geq 6mm$ y más de un sitio inter-proximal con profundidad al sondaje $\geq 5mm$ (Page y Eke, 2007).

Contrariamente 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 completamente explicado. Aunque el modelo etiológico multifactorial de las enfermedades crónicas humanas, puede explicar esta evolución. Así pues, la Periodontitis es una enfermedad infecciosa, donde el agente microbiano es necesario pero no suficiente para causarla, requiriéndose 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 del resumen que en el capítulo 2 de (Ganss, 2006) muestra 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 / oclusales. 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 el caso de 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, siendo el ácido 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, cepillado de dientes, 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.

Terminología definición y etiología

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, excesivos cepillado / uso de hilo dental, 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 y / 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 intervienen (dos desgaste del cuerpo);
Los rasgos característicos son facetas planas antagónicas con márgenes definidos.

Abfracción Desgaste físico como resultado de la tensión o tensión de corte en la región cemento-esmalte provocando 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;
Actuando sobre superficies dentales sin placa;
Características clínicas características del desgaste (tribo) químico con 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 tratados anteriormente, es importante ver la diferencia que puede darse para la erosión al ser considerada

bien como condición o patología dependiendo de los conceptos de salud y enfermedad que se manejen. La erosión podría considerarse como patología cuando ocurre en combinación con dolor o complicaciones endodónticas agudas. Tambien sería 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 patológico, por lo cual el debate continua actualmente. En términos generales, la erosión dental podría 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 la misma en función del criterio que prevalece que existe y es a partir de la relación anteroposterior entre los primeros molares superiores e inferiores. En Ortodoncia, se han propuesto un gran número de clasificaciones, pero no se ha reemplazado al sistema de Angle, ya que éste método es considerado y conocido universalmente, (Álvarez-Vaz *et al.*, 2011) Se puede entonces presentar el gradiente de la maloclusión con la siguiente clasificación:

1. maloclusión 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: cuando 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 éstas son de origen multifactorial. Sin embargo, se pueden 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; los factores exógenos o ambientales, que incluye todos los elementos capaces de condicionar una maloclusión durante el desarrollo cráneofacial.

De la interacción recíproca de estos factores, dependerá la manifestación de una determinada maloclusión.

La oclusión comprende no sólo la relación y la interdigitación de los dientes, sino también las relaciones de éstos con los tejidos blandos y duros que los rodean. Actualmente la maloclusión se define como una disposición de los dientes que crea un problema para el individuo, bien sea estético referido por el mal alineamiento y/o protrusión; de autoestima perjudicada por la maloclusión funcional debido a dificultades en el movimiento mandibular, o cualquier combinación de éstos, (Graber y Swaim, 1991).

Autores como (Proffit, 1990), (Vig, 1990) coinciden en señalar, la necesidad de que, para prevenir, antes se han de conocer e identificar mejor la etiología de las maloclusiones. El concepto 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 correcta oclusión. 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 normooclusión, tal como fue descrita por Angle.

En el momento actual y luego de casi cincuenta años de investigación en esta 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 unos dientes grandes que a desarrollar una mandíbula progénica, y 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 el tratamiento 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, de sus 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, desde luego 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 que las características o anomalías de herencia poligénica muestran un cuadro clínico menos nítido que la monogénica, que se traducen por un fenotipo relativamente uniforme.

En la mayoría de las maloclusiones Clase II suele existir un patrón heredado de déficit mandibular, y en las de Clase III existe una clara tendencia familiar y racial, y en los problemas de excesos verticales que 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 "herencia", sólo se puede actuar con la detección precoz y el consejo genético, aunque en un futuro próximo y según los recientes descubrimientos del genoma humano, será posible influir directamente a nivel genético para prevenir las maloclusiones.

Si bien las maloclusiones tienen un importante componente genético, existen sin embargo 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 fundamentalmente 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 realizados por antropólogos, se observa una frecuencia baja de maloclusión en grupos humanos primitivos alejados de la civilización. Los individuos poseen aceptables normooclusiones 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 favorecería, junto a otros factores ambientales, situaciones como el incremento en la prevalencia del apiñamiento de las generaciones actuales. 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 la 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 puede mencionarse:

- interposición lingual (deglución atípica);
- succión digital, entre los que se encuentra como la más común la succión del pulgar, sosteniéndolo en posición vertical;
- uso prolongado del chupete;
- respiración oral, la cual 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 actúe el hábito, más que de la intensidad. Otros factores ambientales, que influyen en la etiología de la maloclusión, lo constituyen la pérdida prematura de dientes, Caries dental, traumatismos y patologías tumorales y quísticas. Las maloclusiones son un problema real de salud pública con el que se enfrenta la práctica odontológica diaria, por lo que debe saberse evaluar la oclusión de forma adecuada y hacer predicciones acertadas respecto a su futuro.

2.1.5. Trastornos temporomandibulares

Los TTM se pueden definir como un conjunto de condiciones dolorosas y/o disfuncionales en los músculos masticatorios y/o en la articulación temporomandibular (ATM). 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 al desarrollo de TTM. Entre estos se encuentran la parafunción oral de apretamiento dental durante el día y el bruxismo del sueño, (Willeman Bastos Tesch *et al.*, 2014). Es una patología a ser considerada ya que estos mismos autores consignan que el TTM ocupa el tercer lugar en prevalencia entre los dolores crónicos, donde en primer lugar están los dolores de cabeza primarios, seguidos en segundo lugar por el dolor de espalda.

El dolor crónico debido al TTM provoca un impacto muy importante en la calidad de vida de sus víctimas. Se verifican desde leves cambios en la alimentación hasta la manifestaciones 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 predominantemente relacionada con la presencia del dolor y su intensidad, siendo el alivio de este el indicador más confiable para que los pacientes y los médicos juzguen el éxito del tratamiento.

Siendo los TTM una patología que tiene un origen multifactorial, se presentan a continuación algunos de éstos, (Willeman Bastos Tesch *et al.*, 2014)

Endocrinológicos Los receptores de estrógeno se localizan en los tejidos de la ATM y su expresión 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 previamente atribuidos, exclusivamente, a la actividad neuronal de sus regiones anatómicas específicas.

Neurología El dolor orofacial crónico es una respuesta del cuerpo, no sólo 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 [sigla en inglés]) 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 daño'*. De este modo, la experiencia dolorosa es siempre subjetiva y es el resultado de la ocurrencia simultánea de diferentes factores, incluyendo la fisiopatología, las experiencias del pasado y el contexto social.

2.1.6. Lesiones de Mucosa

Dentro de las lesiones de mucosa se encuentran en primer término, dada su malignidad el cáncer bucal y por otra la Candidosis, las úlceras crónicas, la Gingivitis necrotizante aguda, los abscesos crónicos, Hiperplasia Paraprotética y otras entidades como Estomatitis, Subplaca, Fibromas, Nevus, Hemangiomas (Casnati *et al.*, 2013).

La Leucoplasia utilizada por primera vez por E. Schwimmer a finales del siglo XIX, procede de las palabras griegas 'leuco' que significa blanco y 'plakos' que significa placa. En 1978, la Organización Mundial de la Salud (OMS) pretendió a 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 la localización de las mismas, de acuerdo a la siguiente topografía, como Paladar, Rebordes,

Labios, Mucosa yugal, Lengua, Surcos, Comisura.

2.2. Indicadores Epidemiológicos usados en Salud Bucal

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 CPO (CPO) es el más utilizado para medir la enfermedad Caries dental. Fue descrito por Klein y Palmer en 1937 y adoptado por la Organización Mundial de la Salud (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 y/o en la dentición decidua ceo, luego de las modificaciones introducidas por Gruebbell en 1944. La sigla "C" describe el número dientes afectados con lesiones de Caries dental no tratadas, "P" el número de dientes perdidos por causa de Caries dental y "O" el número de dientes obturados o restaurados como consecuencia de Caries dental. El CPO es un índice agregado, que se obtiene como el 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 superficie dentaria, el índice será CPO-S ó ceo-s, (Arana *et al.*, 2005). En el caso de un individuo adulto, el índice CPO puede adoptar valores de 0-28 (excluyendo terceros molares) cuando la unidad es el diente, ó valores de 0-128 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, una vez producida la lesión de Caries dental está no revertirá, manteniéndose en ese estadío, o bien podrá ser obturada o extraída. En consecuencia el índice sólo puede incrementarse o permanecer estable, pudiendo solo 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 presenta limitaciones:

1. Sus valores no están relacionados con el número de dientes en riesgo, un valor de CPO es un recuento de aquellos dientes que el examinador juzgo 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 así también 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 sólo 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 contar estas restauraciones dentro del componente "O" del CPO, a pesar de que estos dientes son sanos. El valor del CPO será así inflado. Las restauraciones de resina compuesta que se han colocado por razones estéticas y los sellantes no deberían ser incluidas en el CPO, y deben 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 donde 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 anticuada, y en realidad puede ser más válida como medida de tratamiento recibido. Es cuestionable usar un índice para la 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, con algunas especificaciones el CPO sigue siendo el principal índice 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, USA en el año 2005, para la práctica clínica, la investigación y el desarrollo de programas de salud pública, con el objetivo de desarrollar un método visual para la detección de la Caries, en fase tan temprana y que además sirva para evaluar el gradiente de enfermedad, así como el nivel de actividad de la patología (https://www.sdpt.net/ICDAS.htm), (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, donde se evaluaron los cualidades del índice, demostró además de ser sencillo, práctico, tener validez de contenido, validez discriminatoria y validez externa con una correlación con el examen histológico de las fosas y fisuras en dientes extraídos. Por tal motivo los autores lo consideran 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; también este índice permitiría hacer el seguimiento del patrón de Caries de una determinada población, aspecto clave para la VE, (Ismail *et al.*, 2007).

El sistema tiene 70 al 85 % de sensibilidad y una especificidad de 80 al 90 % para detectar Caries, en dentición temporaria y permanente. Esta variabilidad en el grado de acierto dependiendo del nivel de entrenamiento y calibración del personal que hace el examen clínico, con un Índice de Concordancia Kappa (Kappa) ≥ 0.65, (Jablonski-Momeni *et al.*, 2008), (Diniz *et al.*, 2009).

Teniendo en cuenta las codificaciones de la Clasificación Internacional - de Enfermedades Aplicada a la Odontología y Estomatología (CIE-OE) , la 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 7 categorías, la primera para dientes sano (código 0, en color verde) y las dos siguientes para Caries limitadas al esmalte, mancha blanca / marrón (códigos 1 y 2, marcadas en color amarillo). Dos siguientes categorías (código 3 y 4, en color rojo) que se interpretan como Caries que se extienden al esmalte y dentina, pero sin dentina expuesta. Y por último para mostrar más gravedad dos categorías restantes (códigos 5 y 6), que aplican a Caries con dentina expuesta. A su vez para las categorías de 1 a 6, en cada estadío este puede ser estar inactiva la lesión o al contrario estar activo, aspectos que se reflejan por consignar que es positivo, que corresponde a activo, mientras que el valor negativo corresponde a inactivo.

2.2.3. Índice Gingival (IG)

El Índice Gingival (IG) que fue elaborado por Löe y Silness a comienzos de 1960, considera la inflamación de la encía en tres grados, donde

1. 0: es 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 lugares zona 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 sólo está presente a partir del grado 2, (Cortés, 1999), (Wolf *et al.*, 2005).

2.2.4. Índice Periodontal Comunitario De Necesidad De Tratamiento (CPITN)

El Índice Periodontal Comunitario De Necesidad De Tratamiento (CPITN) es desarrollado por la OMS (Organization, 1987), utilizándose sobre todo en estudios epidemiológicos. Su característica es que es un índice que 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. El CPITN no tiene en cuenta la pérdida de

inserción de un diente, sino sólo 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, utilizando una sonda de la OMS CP11 y se anota la afectación más grave del sextante. El algoritmo de aplicación es el siguiente:

1. Antes de examinar se identifica las marcas de la sonda CP11 (presenta una bolita de 0,5 mm en su extremo, una banda oscura situada entre 3,5-5,5mm), que debe ser usada con una presión ligera (20 gr), y esperar 20 segundos para observar el sangrado;
2. Se divide la boca en 6 sextantes; uno anterior y dos posteriores en cada arco. Los sextantes se limitan en individuos > 20 años así: utilizando todos los dientes presentes 17-14, 13-23, 24-27, 37-34, 33-43 y 44-47 ó pueden utilizarse dientes índices 17-16, 11, 26-27, 47-46, 31 y 36-37, en < 20 años se eliminan los segundos molares. Para que un sextante pueda ser medido debe presentan 2 piezas sin indicación de avulsión;
3. Se debe 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 y/o subgingivales;
   - Código 3: bolsa periodontal de $4-5$mm (banda negra parcialmente oculta);
   - Código 4: bolsa $\geq 6mm$ (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);
   - $NT = 2$: eliminación de cálculo y/u obturaciones desbordantes (código 2 y 3);
   - $NT = 3$: necesita tratamiento complejo (código 4).

   Cada categoría de necesidad de tratamiento incluye a la anterior, un individuo que necesite tratamiento complejo también necesitará previamente eliminación del sarro e instrucción de higiene oral.

2.2.5. Índice Periodontal Comunitario (CPI)

Es el índice actualmente recomendado por la OMS y conocido como Índice Periodontal Comunitario (CPI),(Organization, 1987) y es una variante mejorada del CPITN. Se mantienen la división por sextantes y las instrucciones sobre los dientes que deben ser examinados, los códigos de 0-4 son iguales al CPTIN. En el caso que faltan el diente o los dientes índices en un sextante, todos los demás deben ser sondeados, excluyéndose el sondaje de caras distales de terceros molares. La sonda utilizada es similar a la de CPITN, con dos marcas añadidas en los $8.5mm$ y $11.5mm$. Se agrega el registro la falta de inserción de los dientes afectados por la Periodontitis, la cuál es la crítica mayor importante al CPTN. 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, siguiendo los siguientes códigos y criterios:

- Código 0: Pérdida de inserción es de $0-3mm$ (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 $4-5mm$ (CAL dentro de la banda negra);
- Código 2: Pérdida de inserción de $6-8mm$ (CAL entre el límite superior de la banda negra y anillo de $8.5mm$);
- Código 3: Pérdida de inserción de 9-11 mm (CAL entre anillos de $8.5-11.5mm$);
- Código 4: Pérdida de inserción de $12mm$ o más (CAL más allá del anillo de $11.5mm$).

La pérdida de inserción no se registra en los niños y adolescentes menores de 15 años.

2.2.6. Índice de Erosión

El desgaste erosivo de los dientes desde un punto de vista clínico 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) presentan en el capítulo 4 a partir de lo que sugieren (Smith y Knight, 1984) para el desgaste dental en general, y los que incluyen los criterios de diagnóstico para el desgaste dental erosivo, (Eccles, 1979).

Para el desgaste general se puede manejar el siguiente índice, donde se debe considerar las siguientes localizaciones

- B bucal or lingual;
- C cervical;
- I incisal;
- L lingual or 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 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 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 Defecto de menos de 1–2 mm de profundidad;
- 4 - B/L/O Pérdida completa de esmalte, o exposición pulpar o exposición de la dentina secundaria, I exposición pulpar o exposición de la dentina secundaria, Defecto de más de 2 mm.

Para el índice de Eccles,(Eccles, 1979) las etapas son

- 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;
- La dentina facial de Clase II 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;

- Facial de clase IIIa Destrucción más extensa de la dentina, particularmente de los dientes anteriores, la mayoría de las lesiones afectan a una gran parte de la superficie, pero algunas están localizadas y ahuecadas;
- Clase IIIb Lesiones linguales o palatales de las superficies durante más de un tercio de su área, los bordes incisales se vuelven translúcidos debido a la pérdida de dentina, la dentina 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 Dientes severamente afectados, donde las superficies labiales y linguales están muy involucradas.

2.2.7. DAI

El DAI (que podría traducirse como Índice Estético Dental) es el índice usado para la evaluación de la maloclusión se calcula a través de una ecuación donde intervienen los diferentes componentes del mismo, (Ourens *et al.*, 2013), con la siguiente forma:

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

donde se consideran diferentes dimensiones de las anomalías en cuanto a los diente separados o apiñados, presencia de dientes visibles perdidos y por otro lado las dimensiones con respecto a la mordida

- 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 an (Rma).

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

El índice que es agregado, se puede categorizar para evaluar el gradiente de la enfermedad a través de un nuevo índice que tiene 4 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).

Las 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,diastema $> 0mm$);
- Irregularidad mandíbular (irregularidad de 0 a 1 mm, irregularidad $\geq 2mm$);
- Irregularidad maxilar (irregularidad de 0 a 1 mm, irregularidad $\geq 2mm$);
- Overjet, (overjet $\leq 0mm$, overjet $< 3mm$, $overjet \geq 3mm$);
- Mordida abierta anterior (sin mordida abierta, mordida abierta $> 1mm$).

2.2.8. Indicadores para TTM

Todos los aspectos mencionados en sección 2.1.5 pueden medirse de muy variada manera, 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, (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 nocturnos como: sensación al levantarse de haber dormido apretando los dientes, dolor de cabeza al levantarse, autovaloración del paciente de su grado de stress, necesidad y/o consulta realizadas por problemas articulares o síntomas vinculados al bruxismo.

A partir de la presencia de los síntomas y signos se pueden 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 2 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 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 2 pilares que deben considerarse al sistematizar la información para el desarrollo de sistemas de vigilancia epidemiológica y que son, por un lado los datos, las fuentes de éstos y los mecanismos de generación de los mismos; 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 diferentes fuentes de datos, algunas metodologías para el análisis de indicadores ya establecidos y usados de rutina y también de nuevas formas de uso de la información, que involucran el uso metodologías estadísticas no convencionales.

Por lo tanto para el investigador en biomedicina, lector de este capítulo debe quedarle claro cuales son los métodos estadísticos que actualmente aparecen en la mayoría de los artículos de revistas arbitradas y que mínimamente debería conocer su funcionamiento, y el contexto en el que éstos se usan.

En este capítulo, a través de las diferentes secciones, se presentan las técnicas estadísticas habitualmente usadas por los especialistas en la salud oral, 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 tesis, les resulte familiar pero planteando los supuestos en los que éstos se basan y como se debe proceder para que su uso sea adecuado. Cuando es 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 diferentes secciones del apéndice de metodología estadística A, también aparecen con mayor nivel de detalle algunas de las técnicas, que se considera mas 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 proponían desarrollar en la sección 1.3, como para otras situaciones donde el investigador biomédico debe hacer uso de la estadística, como 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 donde medir la salud oral y bucal, como los sistemas de registros que pueden crearse al juntar datos a nivel individual provenientes de consultorios, 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 (incluyendo la atención que se hace a nivel de salud pública, 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 (http://www.rediente.org), sistema desarrollado a medida para la Facultad de Odontología, Udelar.

Rediente según consta en su página web "*es un sistema que facilita el registro y la evaluación de la salud bucal. El uso de rediente inicia con el llenado de la historia clínica al lado del paciente, que se lleva su propio carnet Rediente y continua con el ingreso de datos a la base de datos. Esta base 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*

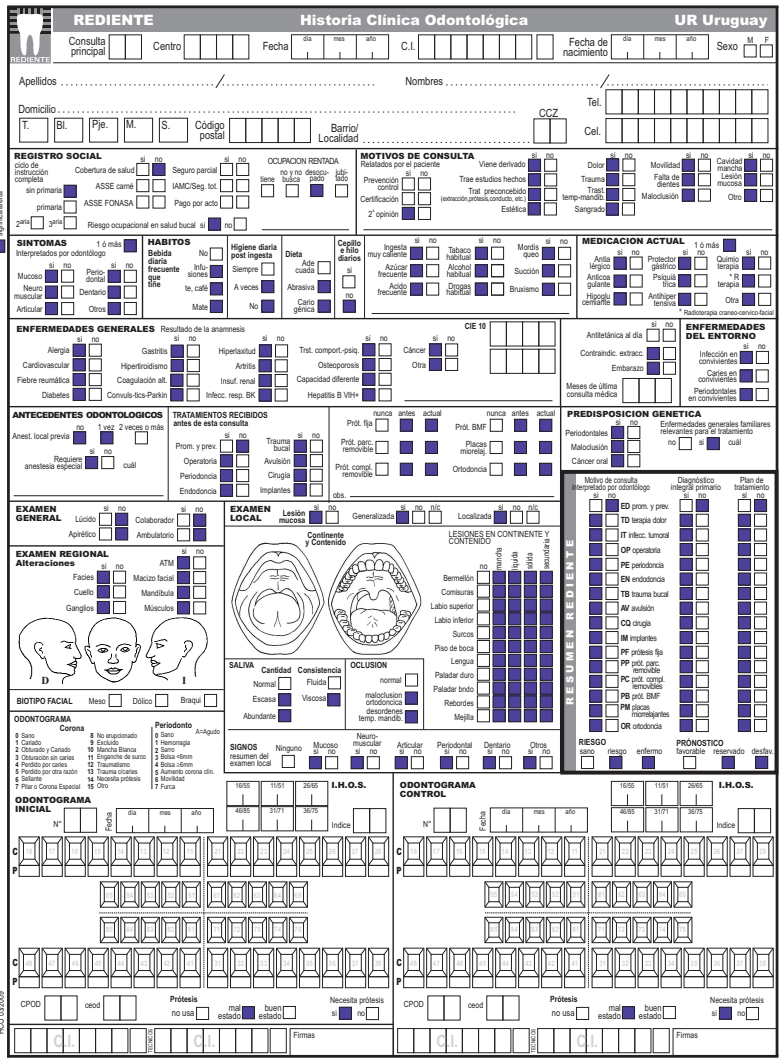

Figura 3.1: Historia Clínica Odontológica de Rediente.

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, Administración de servicios de salud (ASSE) (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". En función de la descripción que antecede es muy importante tomarla como ejemplo sistema de información estadístico-epidemiológico donde aplicar eventualmente 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. Esta opción es la que puede hacer realizable la vigilancia 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 actualmente impulsando el equipo técnico de vigilancia de enfermedades no transmisibles ENT, del Ministerio de Salud (MSP), estableciendo vigilancia por muestreo en las emergencias hospitalarias.

3.1.2. Encuestas de base poblacional

Se va a considerar a la encuesta como un tipo de estudio donde el mecanismo para la generación de 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 puede ser la historia clínica en el caso de los otros tipos de estudio). Nuevamente Ramis en (Oriol, 1997) refiere a las encuestas como una fuente de datos no individualizable sobre la salud de las personas para ser usada en la medición de ésta, en la utilización de los servicios sanitarios, etc. Desde el punto de vista epidemiológico y de la salud pública este mecanismo es el que va a permitir a nivel poblacional medir diferentes fenómenos, de una forma homogénea, al aplicar un cuestionario donde se releva exactamente igual 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 ser aplicada a través del tiempo como se detalla en 3.1.2, donde aparecen algunas ventajas extras 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. De ahora en más en este trabajo de tesis se supone que en las encuestas se está empleando algún mecanismo de muestreo aleatorio, de los cuales se hace una reseña en la sección 3.6.

Encuestas de Corte Transversal (cross-section) (ECT)

En este caso la encuesta de corte tranversal refiere al estudio en un momento dado, lo que coincide con los estudios de prevalencia, donde se podrá evaluar una serie de variables, algunas de las cuales podrán ser consideradas como dependientes y el resto como independientes.

Encuestas Longitudinales (panel data) (EL)

Las encuestas de panel se refieren a estudios basados en observaciones repetidas efectuadas sobre los mismas unidades de muestreo a lo largo del tiempo, que pueden ser personas, hogares, empresas, etc.
La medición periódica de elementos permite realizar un *seguimiento* de la población objetivo, logrando 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 realizadas 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 mismas se definen en la etapa del diseño de la encuesta.
Algunos autores como (Fuller y Braid, 1999) distinguen tres variaciones de panel: panel puro, panel rotativo, panel 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), lo que 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 cual 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 como aplicarlos en cada fuente de información, para finalmente analizar la sustentabilidad del sistema de vigilancia.

Como forma de presentar los diferentes indicadores a ser usados en los capítulos que forman parte de la tesis y los artículos en los que se participa en coautoría y que se referencian, se ordenan los indicadores y se resumen de acuerdo al criterio de respetar la propuesta original en el anteproyecto de tesis, que se adjunta en el apéndice C.

1. Indicadores Clásicos (I. Clásicos) que responden a la forma de medir las diferentes patologías descritas en el capítulo 2, en la sección 2.1, tal cual fueron originalmente pensados;
2. Indicadores que combinan muchas variables odontológicas que habitualmente no se tratan en esa forma tal cual surge de la experiencia nacional e internacional en la bibliografía consultada y que se denominarán Indicadores Combinados (I. Combinados), los que se presentan en la sección 3.3;
3. Los 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 (I. Alternativos) en la sección 3.4;
4. Un cuarto grupo de indicadores que mas allá de ser indicadores clásicos, combinados o alternativos incorporan las dimensiones de tiempo y espa-

cio asociados al fenómeno morboso en la salud bucal se presentan como Indicadores Espacio-Temporales (I. 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 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);
- Las piezas dentales $d^g_{i,j,k}$;
- Los cuadrantes $q^g_{i,.,k}$ (formados por piezas);
- Los sextantes $se^g_{i,.,k}$ (formados por piezas);
- Las superficies de cada pieza $s^g_{i,j,l}$

Con la información relativa a el estado de las piezas dentales, se toma como primer ejemplo el CPO para razonar como es el proceso de construcción del indicador, sabiendo que si se considerasen el ceo o el ICDASII 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	0	3	3	0	2	1	3	3	8	4	8	8	20
EXTRACCION	8	3	3	3	2	0	3	3	3	8	3	17	0	20
EXTRACCION	2	3	1	2	0	3	3	3	3	0	10	5	2	17
EXTRACCION	8	3	3	3	0	2	3	3	3	8	2	14	0	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 "status" de salud bucal, mostrando que las piezas dentales tienen poca historia de enfermedad. Generalmente las personas tienen, salvo excepciones, un total de $28 - 32$ piezas, repartidas en 4 cuadrantes, 2 inferiores y a su vez 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 con 5 cariadas y 7 perdidas. En ambos casos, los niveles de enfermedad son importantes (tienen 12/28 % 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 enfermedad pasada.

En virtud del ejemplo antes presentado es necesario manejar alternativas, como utilizar los 3 componentes del CPO por separado, transformado en tasas, proporciones o índices basados en medidas de entropía; es decir considerar la misma información pero analizándola de otra manera.

Más aún, teniendo en cuenta que habitualmente se recoge información a nivel individual sobre características sociodemográficas que pueden estar a asociadas a los niveles de enfermedad bucal, medida a través del CPO, se propone integrar esas características personales para evaluar diferencias para los componentes del CPO a través de modelos estadísticos, que son modelos de tipo *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, de uso sencillo, que la información esté disponible y que los investigadores en ciencias médicas lo entiendan y lo 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 , [**tipo2**],que se conocen como **(VGAM)**,Yee y Hastie (2003)

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 vistos todos como casos particulares de los Modelos Lineales Generalizados (MLG). En este trabajo de tesis se consideran algunos modelos del *tipo 1* que cada vez más pasan a estar difundidos en el campo de la epidemiología, prestando entonces especial atención a los modelos de conteo, mientras que otros modelos de *tipo 2*, siendo 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*, con resultados posibles: *éxito, fracaso* codificados como $\{0, 1\}$, distribución de probabilidad: $P(Y_i = 1) = \pi_i$, $P(Y_i = 0) = 1 - \pi_i$ y valor esperado [1]: $E(Y_i|X) = \pi_i$.

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

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

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

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

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

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 y se lo considera como "medida de riesgo". Por ejemplo cuando 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 cuando una estimación es considerada 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 ó 1;
   b los costos de predecir en forma incorrecta 0 ó 1 son prácticamente los mismos;
2. Encontrar el mejor punto de corte para los datos. Implica calcular distintos puntos de corte y evaluar en cada caso como son pronosticadas las n observaciones. Se emplea aquel 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 ó 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 buscan optimizar la Sensibilidad (Sen) y la Especificidad (Esp) del modelo. La Sen se define como el cociente entre los éxitos observados clasificados como éxitos y el total de éxitos observados, mientras que la Esp se define como el cociente entre los

fracasos observados clasificados como fracasos y el total de fracasos observados. Los 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 Curva ROC (ROC).

3.2.3. Modelos de Conteo

Los Modelos de conteo (MC) son una serie de modelos donde la variable de respuesta refiere al número de veces que un evento ocurre, donde 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 70.

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

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

Tabla 3.1: Diferentes alternativas a los modelos de Conteo.

Modelo de Poisson (M. de Poisson)

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

En la Tabla 3.1 aparece como alternativa al modelo de Poisson, cuando por

ejemplo existe sobredispersión y da lugar a un modelo Modelo Binomial Negativo (MBN)

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

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

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 la integral que allí aparece conceptualmente debe considerarse como que en realidad en el caso de la BN lo que se tiene es un modelo donde la tasa de fallas μ no es fija y por lo tanto lo deben considerarse un promedio de la misma, 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 cuando es necesario el uso de algunos de los 2 tipos presentados en la Tabla 3.1 u otras variantes cuando estos 2 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 una profundización de los mismos, reformulándolos 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, donde 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 el Análisis Factorial (AF) y el Análisis de Cluster o Conglomerados (AC), técnicas que no son muy usadas 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 Muĺtiples (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, donde se le da un peso o importancia diferente a cada una). La ventaja de estos métodos es que se adecuan al tipo de variable considerada, sean éstas de tipo cuantitativa o cualitativa ordinal o cualitativa nominal. Un primer ejemplo es trabajar para el caso del CPO, con una descomposición del mismo, que considere 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 está analizando como es la asociación intra cuadrante con respecto al CPO, con la ayuda del ACM, se puede simplificar el problema al transformar las 6 variables multinomiales 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. Se puede así 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 con la ayuda del AC, que se presenta en la sección 3.3.2 crear *tipologías* o grupos de poblaciones con perfiles epidemiológicos bien diferenciados de acuerdo a las patologías y los factores de riesgos asociados.

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 les considera por separado. Es por lo tanto esta metodología una forma de crear subpoblaciones no observables o por lo pronto 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 cuales son las características que mas asemejan o diferencian a los individuos.
Existen muy variados métodos de clustering, dentro de los cuales se encuentran por ejemplo 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 estos métodos que se reseñan en la tesis se sigue la lógica de que la información vista es 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 de tesis no se consideran en la misma.
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, en el que se descompone el mismo usando por separado los componentes 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

Se considera un nuevo enfoque (poco usado en la epidemiología de salud bucal aunque cada vez más) basado en técnicas estadísticas de *aprendizaje supervisado*, como son los métodos Árboles de Clasificación y Regresión (Árboles de Clasificación y Regresión). Esta metodología permite la construcción de modelos basados en técnicas no paramétricas, lo que supone muchas menos restricciones de distribuciones de probabilidad en las variables consideradas, permitiendo encontrar las variables que mejor discriminan el comportamiento de una variable de respuesta o dependiente de tipo categórica; la aplicación immediata es 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 facilmente usables e interpretables por los no especialistas en estadística, (Breiman *et al.*, 1984),(Abernathy *et al.*, 1987).

En el capítulo 12 se presenta una aplicación donde se combinan los métodos Árboles de Clasificación y Regresión junto con la R. Logística para el reconocimiento del sexo a partir de diferentes medidas de los dientes caninos en humanos. En la sección A.5 del apéndice A se presenta en forma un poco más detallada la metodología en las que se basan los CART.

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 características necesarias a la hora de usarlos para hacer *Inferencia estadística*. Hay que considerar que las proporciones a estimar que están en el rango $(0, 1)$, no cumplen el supuesto de Normalidad, pudiendo existir asimetrías muy importantes y la varianza que puede cambiar a lo largo de 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**, donde se detalla la formulación del modelo predictivo que puede ser hecha a partir de los trabajos de, (Kieschnick y McCullough, 2003), (Salinas-Rodríguez *et al.*, 2009).

3.4.3. Índices basados en Teoría de la Información

En este caso se pueden adaptar, para poder discriminar el comportamiento en la población de los diferentes indicadores a evaluar en la salud bucal, 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 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 Indicador alternativo.

3.5. Indicadores espacio-temporales

Los indicadores que se pueden aplicar son de 3 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 pero que ya tienen varios años y que a su vez presentan algunas limitaciones. No obstante se presentan para luego mostrar algunas metodologías alternativas que suponen el uso de técnicas mas modernas y que hoy en día pueden usarse gracias al desarrollo computacional existente.

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 Ohno Es un método desarrollado para identificar patrones geográficos de mortalidad o morbilidad observados visualmente 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, bajo el supuesto de agregación espacial, (Ohno *et al.*, 1979). Por lo tanto para poder calcular el estadístico de pueba 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, tasas de incidencia de cada área. 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, atendiendo 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: Se calcula el número de pares de

Categorías	Punto de corte	Definición	Número de áreas
1	V_1	$T_i < V_1$	N_1
2	V_1, V_2	$V_1 < T_i < V_2$	N_2
...	...	...	...
$k-1$	V_{k-2}, V_{k-1}	$V_{k-2} < T_i < V_{k-1}$	N_{k-1}
k	V_{k-1}, V_k	$V_{k-1} < T_i < V_k$	N_k

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). 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

$$\chi^2 = \left(\frac{AC_i - EAC_i}{\sqrt{EAC_i}} \right)^2, i+1, ...K \tag{3.8}$$

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

$$\chi^2 = \left(\frac{AC - EAC}{\sqrt{EAC}} \right)^2, i+1, ...K \tag{3.9}$$

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.

método Grimson Esta técnica 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, los índices que identifican a cada una, de modo de saber si son áreas que adyacentes a ella, o que comparten fronteras. A su vez las tasas de incidencia de la patología bajo estudio en cada área y finalmente 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.

$$P_i = Pr[Poisson(E(P)) \geq P] \tag{3.10}$$

$$P_i = Pr[N(0,1) \geq \frac{P - E(P)}{\sqrt{V(P)}}] \tag{3.11}$$

3.5.2. Agregaciones Temporales

Para la detección de agregaciones temporales se puede 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. Algunos de los métodos si los casos son individuales son:

Método Chen En este caso para evaluar si existe una agregación temporal se debe considerar T, la tasa esperada de incidencia anual de la enfermedad (puede ser una tasa histórica o del pasado reciente que ha prevalecido hasta antes de la sospecha de epidemia), P que considera el tamaño de la población expuesta riesgo, FP, la tasa de falsos positivos, es decir, que en este caso debe verse como la probabilidad de rechazar la hipótesis nula siendo cierta, a causa de la aparición de casos falsos de la enfermedad en el período estudiado. Esto 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 antes presentada se calcula la longitud temporal máxima o intervalo de referencia, que esta dada por

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

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

Método Sets La técnica Sets 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, el 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 se necesitan 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 siendo 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, donde se compara 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)

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

Si los casos son agrupados pueden emplearse por ejemplo:

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). Se necesita 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 $P_i \leq \alpha$, 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

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

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), 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 la detección de 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, 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, 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 D. Poisson, para evaluar el número de casos esperado. Para $n < 30$ se determinan los valores umbrales, P_1,P_2, se calcula para cada período, la probabilidad de observar al menos los O_i casos observados bajo supuesto de distribución Poisson , (D. Poisson) 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 $n \geq 30$ se determinan los umbrales

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

Se calcula la REM de cada período como el cociente entre el número de casos observados y el esperados, para aplicar luego 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 los cual es necesario tener la

- La serie temporal de casos observados (en días, semanas, cuatrisemanas o meses): O_i
- E: el número esperado de casos por unidad de tiempo.
- 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 cuidando de 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 como 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 Utilizado 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, considerańdose despreciable 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

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

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 espacio-temporales

En este caso se busca encontrar si existen clusters en el espacio, por un lado y en el tiempo, en forma simultánea, 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 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 que sea 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 espacio-temporales, 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. Estas coordenadas 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/aaaa). 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, los cuales 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 expresan 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 euclídeana (la recta más corta que une dos puntos), (Knox, 1964).

Test de Kulldorf Este test es una variante al de Knox antes presentado que se modifica para situaciones donde hay una tasa de crecimiento poblacional variable en diferentes regiones y donde 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 ta-

maño de la población de esa región y período, (French, 2015a). También el test en este caso se evalúa mediante SM y permutaciones de varianza;

Test de Diggle Es otra variante al método de Knox, que permite un procedimiento capaz de evaluar varios diagnósticos útiles para Interacciones y a su vez hacer pruebas espacio-temporales para evitar múltiples pruebas, (Tango, 2010)

Test k-NN de Jacquez Este test busca contar el número observado de pares de casos se cercanos en espacio y tiempo, donde la medida de cercanía está definida por los k vecinos más próximos, y donde 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 donde el cálculo es relativamente sencillo y donde se deja 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 su-

ponen manejar un herramental estadístico más avanzado al tener 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 ser definida 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 (I de Moran) y el Índice de Geary (I de Geary) (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 por el contrario, existe una asociación significativa de valores similares o no, entre zonas vecinas.

En una segunda instancia se puede hacer interpolación por *método de krigging* sobre los puntos de muestreo georeferenciados. 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 Sistema de Información Geográfica (SIG) que permiten generar esos indicadores y cuales son los algoritmos que manejan 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 al trabajar con muestras probabilísticas

En esta sección se plantean algunos aspectos relativos al muestreo de poblaciones finitas que en el ámbito de la epidemiología muchas veces se debe tener en cuenta. 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 habitualmente se usan. La forma de presentarlos sigue la lógica que se manejan 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, donde el carácter de la aleatoriedad, surge de un modelo hipotético; en el caso 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, o sea, 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 desde la estadística se toman en cuenta en esta nueva forma de trabajar en los diseños muestrales como son los conceptos de *población, muestra, diseño muestral* y desde la epidemiología apropiarse del lenguaje para entender todo el proceso de construcción del mecanismo de muestreo, para conocer por que se hace de cierta manera, y las consecuencias que tiene en términos de los costos, de la precisión, de la 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) habitualmente utilizados en la formación en epidemiología, y que son los que más se usan. 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

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

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

$$\begin{cases} P(k\ y\ l \in S) = \pi_{kl} = \frac{n(n-1)}{N(N-1)}\ \forall k \neq l \in U \\ \Delta_{kl} = -\frac{n(N-n)}{N^2(N-1)}\ \forall k \neq l \in U \end{cases} \tag{3.23}$$

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

$$\hat{t}_y = \sum_S \check{y}_k = \sum_S \frac{y_k}{n/N} = N \sum_S y_k = N\bar{y}_S \tag{3.24}$$

Cuando se tiene un diseño aleatorio simple que se abrevia como *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 ésta, vienen dados, respectivamente, por

$$\begin{aligned} V_{\text{SI}}\left(\hat{t}_\pi\right) &= N^2\left(1-f\right)\frac{S^2_{y_U}}{n} \\ \hat{V}_{\text{SI}}\left(\hat{t}_\pi\right) &= N^2\left(1-f\right)\frac{S^2_{y_S}}{n} \end{aligned} \tag{3.25}$$

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 más adelante se verán, ya que hay que sacrificar la pérdida de precisión o lo que es lo mismo tener más *varianza* para ganar la posibilidad de tener un diseño más económico

- Muestreo Sistemático (SY);
- Muestreo Estratificado (STSI);
- Muestreo por Conglomerados (SIC);
- Muestreo en varias etapas (MMult)

Tal como se plantea en (Álvarez-Vaz, 2010),(Álvarez-Vaz, 2017) en realidad rara vez se usa un diseño puro, sino que es necesario combinar los diseños antes mencionados, por lo cual se termina trabajando con diseños muestrales en varias etapas, MMult.

Entonces la pregunta es cómo comparar la eficiencia entre distintos diseños y saber por cual optar y la respuesta es que esto depende de distintos factores:

- La distribución de los valores de la variable de interés en la población;
- El parámetro a estimar;
- El estimador utilizado;
- La 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

$$\text{Deff}\left(\ p(s)\ ,\ \hat{t}_\pi\right) = \frac{V_{p(s)}(\hat{t}_\pi)}{V_{SI}(\hat{t}_\pi)} \tag{3.26}$$

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

Siguiendo la misma lógica de considerar dos, tres o más etapas interesa ver como pueden establecerse diferentes diseños en forma flexible combinando según la necesidad de cada caso los diferentes tipos de diseños "puros" vistos antes. Para eso se toman como ejemplo una de las diferentes situaciones planteadas en el manual *WHO Steps Surveilllance Manual* (World Health Organization, 2006) de donde se extraen la figuras que sigue correspondiente a uno de los diferentes diseños. Siempre la lógica del diseño empleado es disponer de un marco muestral adecuado, donde poder seleccionar las Unidades Primarias de Muestreo (UPM) y en una segunda instancia dependiendo de la calidad de la información trabajar con dos, tres o más etapas de muestreo subsiguientes.

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 usando la Tabla de Kish (TK).

Se puede seguir por último esta serie de recomendaciones (que tienen en cuenta a su vez los fundamentos téoricos vistos en las secciones precedentes para cada tipo de diseño de muestreo)

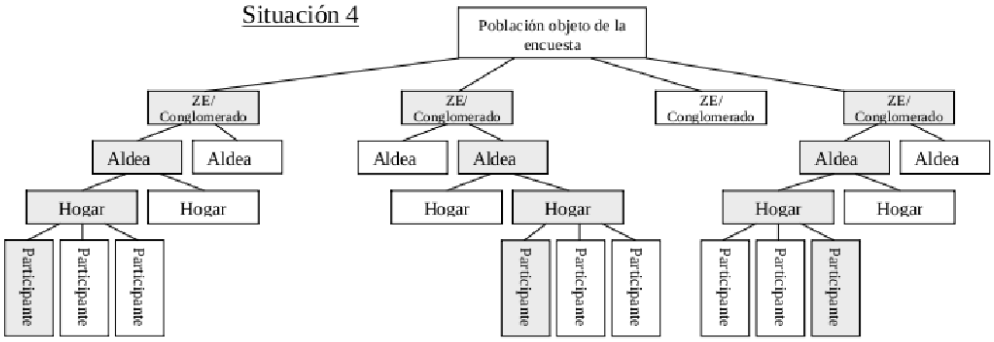

Figura 3.3: Diseño en 3 etapas, figura extraída del manual de encuestas (World Health Organization, 2006).

- 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
- 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 fueron presentando como calcular los totales de una población de acuerdo al diseño de muestreo seguido, pero sin embargo 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 se usan varias etapas, que suponen considerar estratos y/o conglomerados, tal como presentan en su trabajo (Guillén *et al.*, Oct), (Cañizares Pérez *et al.*, Mar), y donde no se puede 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 son los *promedios, ratios de totales, medianas, proporciones, cuantiles, coeficientes de regresión.* Todas esas cantidades se pueden expresar como funciones de los totales pobla-

cionales.

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

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 el que se va a poder medir el error a través de

$$V(\hat{\theta}) = V(\sum_{j=1}^{q} a_j \hat{t}_j) = \sum_{j=1}^{q} \sum_{j'=1}^{q} a_j a_{j'} C(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) \tag{3.27}$$

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 pueden 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

$$\hat{t}_0 \pm \phi_{1-\alpha/2} \left[\hat{V}\left(\hat{t}_\pi\right) \right]^{1/2} \tag{3.28}$$

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. En general los diseños complejos son mas ineficientes que si se trabaja con SI. A su vez puede pasar que si se logra estratificar el STSI puede ser mas 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 que bajo muestreo aleatorio simple SI, tal como se vió en la sección 3.6.2.

Bajo un diseño SI la determinación del tamaño depende

- del nivel de confianza aceptado, α;
- la precisión deseada;
- el parámetro θ que se desea evaluar

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

$$\hat{\theta}(m) \pm \phi_{1-\alpha/2}\sqrt{\hat{V}[\hat{\theta}]} \tag{3.29}$$

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 bajo 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 bajo el supuesto de SI se puede exactamente derivar cual es el *factor de inflación de la varianza* a través del Efecto "diseño" (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 en general puede estar entre $[1; 3]$ siguiendo recomendaciones de *Adecuacy of sample size in health studies*, (Lwanga y Lemeshow, 1991).

3.6.6. Estudio en dominios

Sin importar cual sea la estrategia de muestreo adoptada, en función de la información disponible, los alcances del estudio y por lo tanto el diseño de muestreo resultante, muy frecuentemente el investigador en biomedicina tiene necesidad de poder estimar, no solamente para toda la población U, sino para alguna subpoblación de ella, como puede ser 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 forma lo que se denomina *dominios*, donde se parte 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 extra, ya que las precisiones, que son función del tamaño de muestra cambian, lo que hace que no se pueda muchas veces dar resultados desagregados por dominio, por no haber sido contemplado en el diseño, ni en el cálculo del tamaño muestral, (Álvarez-Vaz, 2017).

Finalmente otro aspecto a tener en cuenta al trabajar con diseños complejos, es que en general sin importar cual 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, en general opera un aspecto no deseado pero que no se puede soslayar y es que existe en general un proceso de pérdida de información, que se denomina *No respuesta* y que no es aleatorio. Es decir que 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 para algunas variables que el investigador considera fundamentales, como puede ser 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 por cual de toda esa información desea controlar para evitar sesgos y que luego deberá ajustar, según disponga de la misma en forma parcial o total.

Los procesos de calibración pueden ser esencialmente 2

- Post-Estratificación Completa o Calibración sobre totales de celdas conocidos (PEC);
- Post-Estratificación 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 sepa los totales poblacionales en cada una de ellas. Por ejemplo, si se desea controlar por edad y sexo, es necesario si la edad se estableció en 4 tramos, conocer los totales por sexo y tramo etario.

En cambio, la calibración por RAKE se hace cuando, una vez determinado por el investigador, por cual variable 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, que harían que el proceso de ajuste sea inestable.

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

3.7. Ajustes de los Indicadores mediante Información auxiliar

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

3.7.1. Modelos Multinivel

El análisis multinivel A. Multinivel 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 perteneciendo a 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 entre los diferentes grupos. Por ejemplo, los niños

aprenden en las clases, las condiciones de su grupo, tales como características de los maestros y la capacidad de otros niños en la clase, lo que puede influir en el logro educativo de un niño.
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. Más aún 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.
La misma situación de las escuelas se da en otras disciplinas como por ejemplo en salud cuando se lleva a cabo un estudio para investigar la relación entre el total de colesterol (CT) y edad.

$$\text{CT} = \beta_0 + \beta_1 * \text{edad} + \beta_2 * \text{género} + \varepsilon \tag{3.30}$$

Los pacientes en estudio en general pueden provenir de diferentes médicos consultantes (es como habitualmente se pueden hacer los estudios) con lo cual se puede pensar que existe una estructura jerárquica que responde al médico del cual 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. Mas aún si se desease comparar el modelo para un valor dado de edad en las mujeres por ejemplo, se tendría 11 modelos lineales univariados con diferentes interceptos pero una misma pendiente. La idea que está detrás del A. Multinivel es en lugar de considerar los interceptos por separado, considerar 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 (3.31).

$$\begin{aligned} \text{CT} = \beta_0 + \beta_1 * \text{edad} + \beta_2 * \text{D1} \ + ...\beta_m * \text{Dm-1} + \\ ... + \beta_{m+1} * \text{D1*edad} + ...\beta_{2m-1} * \text{D11*edad} + \epsilon \end{aligned} \tag{3.31}$$

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 A. Multinivel 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: o bien 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 2 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 mas recientemente como un caso particular de los "modelos mixtos" que se presentan en (McCulloch, 2001).

En general se puede ver el A. Multinivel como un caso de modelo mixto

$$\mathcal{Y} = \underbrace{\mathcal{X}\beta}_{\text{Coeficientes fijos}} + \underbrace{\mathcal{U}\delta}_{\text{Coeficientes aleatorios}} + \epsilon \tag{3.32}$$

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

$$y_i = \sum_{q=1}^{q=r} x_{qi}\beta_q + \sum_{s=1}^{s=p} u_{si}\delta_s + \epsilon_i \tag{3.33}$$

Una de las mayores ventajas 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), análisis multinivel de Poisson (AMP) para variables de conteo e incluso se puede desarrollar análisis de supervivencia de varios niveles.

También el A. Multinivel puede ser usado para el modelado de datos longitudinales, donde los datos están correlacionados, al ser por ejemplo los pa-

cientes los que se repiten y para los cuales se les mide una variable *tiempo-dependiente*. 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; en 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, 2016) 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 presentar al lector que pudiese no conocerlo, se opta por transcribir exactamente la definición que sus autores dan del mismo en su página web (`https://www.r-project.org/about.html`)

'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, ...) y técnicas gráficas, y 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 pueden producir gráficos de calidad de publicación bien diseñados, incluyendo 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 Licencia Pública General GNU de Free Software Foundation en forma de código fuente. Compila y se ejecuta en una amplia variedad de plataformas UNIX y sistemas

similares (incluyendo FreeBSD y Linux), Windows y MacOS.

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 a los usuarios agregar 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 de usadas en cada caso para la resolución de los diferentes 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 la complejidad

de las mismas 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 diferentes capítulos de la tesis así como las publicaciones que en el marco de la misma se desarrollan como artículos en revistas, documentos de trabajo de la serie Documentos de trabajo del IESTA y trabajos presentados en congresos nacionales, internacionales o jornadas académicas se usan fuentes de datos primarias que son que son 2 encuestas de base poblacional y otros 2 estudios que se presentan a continuación:

- **Primer Relevamiento Nacional de Salud Bucal en Población Joven y Adulta**, llevado adelante llevado adelante por la Facultad de Odontología de la Universidad de la República y supervisado por el servicio de Epidemiología y Estadística, de la cual el autor de la tesis 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 edad, de escuelas públicas y privadas. (Relevamiento de Facultad de Odontología con financiación de la ANII), ver sección 4.2.
- **Relevamiento en población que se asiste Facultad de Odontología durante 2015-2016**, que forma parte del proyecto CSIC I+D llevado adelante entre junio 2015 y junio 2016 , ver sección 4.3.
- **Determinación del perfil biológico de pacientes asistidos en Centro de Estudio y Diagnóstico de las Disgnacias del Uruguay IUCEDDU**, ver sección 4.4.

4.1. Primer Relevamiento Nacional de Salud Bucal en Población Joven y Adulta PRNSB2011

Es una encuesta donde se relevaron 1485 personas, cuya metodología se publica en (Lorenzo *et al.*, 2013a), (Olmos *et al.*, 2013), (Lorenzo *et al.*, 2013c), (Ourens *et al.*, 2013), (Casnati *et al.*, 2013) y se basó en los criterios sugeridos por la Organización Mundial de la Salud (OMS) para estudios poblacionales (1997), los que se adaptaron utilizando un diseño muestral complejo en 2 fases.

1. 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 o más habitantes 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, y 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 realiza cada dos meses. Así que eso es lo que se considera la primera fase de muestreo. Las 4 olas de la encuesta es la cantidad mínima necesaria para lograr el tamaño de muestra calculado para cada grupo de edad. Hay un total de 4.000 personas, de las cuales 1.500 pertenecen al grupo de edad de 15 a 24, los restantes 2.500 pertenecen a otros grupos de edad;
2. En la segunda fase de todas las personas pertenecientes al grupo de edad de 15 a 24 son seleccionados entre los 4.000 personas en la primera fase (alrededor de 1500). Los restantes 2.500 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 de $1 - \alpha$ 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 e Interior).

El sorteo de la muestra, lo realizó el Instituto Nacional de Estadística (INE), el que proporcionó los expansores asociados al diseño. Se relevaron las personas sorteadas de las poblaciones en 10 departamentos y 14 ciudades,Artigas; Canelones: Ciudad de la Costa, La Paz, Las Piedras; Colonia; Florida; Maldonado: San Carlos; Paysandú, Salto, San José,Rivera y Tacuarembó

4.1.1. Calibración de la muestra para PRNSB2011

El cálculo para el tamaño de muestra teniendo en cuenta los 6 dominios de estimación considerando 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 e Interior, con una precisión del $\delta = 5\,\%$, confianza del 95 %, Tasa de no respuesta del 20 % y un *deff* de 1.5, con lo cual al usar la ecuación 4.1, resultan los tamaños que se consignan en la Tabla 4.1

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

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.

Teniendo en cuenta la tasa de no respuesta se hizo un ajuste de calibración posterior mediante posestratificación siendo las variables utilizadas para la calibración sexo y edad. Se usó la librería *survey*, (Lumley, 2009).

4.2. Relevamiento y análisis de caries dental en adoloscentes escolarizados de 12 años de la ciudad de Montevideo,Uruguay RACA2012

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 edad, de escuelas públicas y privadas.
Para el cálculo del tamaño de la muestra, se utilizaron los siguientes parámetros: prevalencia de Caries de 60 % (22), el intervalo de confianza del 95 % (CI), un nivel de precisión de 4 % y un efecto de diseño (Deff) de 1.3, al que se añadió una tasa de no respuesta de 30 %. Por lo tanto, el tamaño de muestra necesario para este estudio fue 1.235 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 de edad, asistentes a estas escuelas fueron invitados a participar en el estudio, independientemente del año escolar en el que estuviesen.

Para la primera etapa se usa como marco muestral las escuelas de 5-6 años del sector privado y público, las que se estratifican en 3 estratos.

Se sortean 2 muestras independientes, una por estrato. Para el diseño usado las UPM, en este caso las escuelas se seleccionan con muestreo $\pi - ps$ (probabilidad proporcional al tamaño), es decir la variable total de niños matriculados que figuran en el marco muestral. Se usa 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 en base a la información disponible que es el total de niños por escuela, de los cuales una parte debe ser descartada ya que solo pueden ser incluídos los niños de 12 años, lo cual hace que se tenga *total de*

niños y *total de niños elegibles*. Luego dependiendo de la cantidad de niños relevados en la segunda etapa, lo que hace que se tenga un tamaño de muestra aleatorio (desconocido). 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 en base al total y no el total de elegibles, por lo cual para cada escuela se debe de 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 dió un valor de $1,45$, lo que debe de interpretarse como que hay en promedio un $45\,\%$ mas de niñas que de niños elegibles.

4.2.1. Calibración de la muestra RACA2012

Para la calibración de la muestra se deben de tener en cuenta 2 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 sobrerepresentación las mujeres, fue necesaria una calibración que se hizo usando la librería *survey*, (Lumley, 2009).

4.3. Relevamiento en población que se asiste Facultad de odontología RPAFO2015

Los datos provienen del estudio sobre personas que demandan atención en la Facultad de Odontología de la Universidad de la República, Uruguay y que son evaluados por los odontólogos del Servicio de registros de la Facultad,

desarrollado en el marco del proyecto 'I+D' de la CSIC 2014. Se aplica una muestra de 602 personas que consultan en el período que corresponde a mayo 2015-junio 2016, los que se seleccionan mediante muestreo sistemático, a los que se les aplica un cuestionario sociodemográfico y un examen completo de la boca, en donde se evalúa 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 %. Finalmente de los 640 originalmente calculados se obtuvieron 602, que representa una fracción de muestreo de alrededor del 15 % del total de personas que consultan anualmente.

4.4. 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

El DPBIO2009 es un Estudio descriptivo transversal donde se evaluaron 1006 modelos de yeso, pertenecientes a pacientes asistidos en la clínica de ortodoncia del IUCEDDU (Instituto Universitario Centro de Estudio y Diagnóstico de las Disgnacias del Uruguay), desde 2003 a 2009.

Las edades de los pacientes están en el rango 18 a 60 años, los modelos de yeso son de piezas de dentición permanente sana, incisivos y caninos totalmente erupcionados, sin restauraciones, aparatología ortodóncica, desgaste ni anomalías dentales, que manifestaron, libre y espontáneamente, mediante consentimiento informado, su aspiración y disposición de participar de la misma. En este trabajo de tesis solamente se consideran los datos que corresponden al maxilar superior.

Capítulo 5

Comparación de los modelos de regresión binaria y los modelos de conteo básicos aplicados a 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 es el trabajar con variables de respuesta en escala cuantitativa, las que se dicotomizan para un determinado umbral, y usar modelos predictivos, pero con la pérdida del gradiente de la enfermedad.

Se muestra entonces como 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 D. Poisson o la BN, que son las más usadas en la aplicaciones biomédicas frente a este tipo de problema. Sin embargo no siempre esas 2 distribuciones 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 R. Logística vs una regresión que conserve la escala original de medida preservando el gradiente de la enfermedad, para lo cual se debe 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 cual es la herramienta más adecuada dada 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 no 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, de sencillo uso, que la información esté disponible y que los especialistas en ciencias médicas lo entiendan y lo adopten.

Cuando la variable de respuesta es categórica (Agresti, 2005), se puede recurrir a la R. Logística, donde 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, prestando entonces 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*, con resultados posibles: *éxito* , *fracaso* codificados 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

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

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

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, las que se basan en la teoría de los MLG que se detalla en el apéndice A, en la sección A.3 del mismo. Con esas restricciones de ser variables con recorrido discreto, existen varias alternativas donde las más frecuentemente usadas son la D. Poisson 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 alguno supuestos básicos que están implícitos en la D. Poisson no se cumplen como es 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 D. Poisson la que tiene función de masa de probabilidad

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

que 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 $V(\mu) = \mu$. En la práctica esto a menudo no se verifica, con lo cual es 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 es adecuada al problema de sobredispersión, en el tratamiento de la varianza, (Zeileis, 2004, 2006).

Modelo Binomial Negativo

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

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

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 $V(\mu) = \mu + \frac{\mu^2}{\theta}$ (Hilbe, 2011), (Cameron y Trivedi, 1990).

Luego de haber presentado los detalles básicos de los MC y las diferentes alternativas de distribuciones el objetivo es poder 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), PRNSB2011

Tal como se aclaró en el capítulo 4, este estudio se caracteriza por tener un diseño muestral en 2 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 2 variables, pero genera un juego de pesos muestrales calibrados, que difieren de los originales, para lo cual se estudia la relación

que existe 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, siendo los hombres, los que más sus expansores aumentan

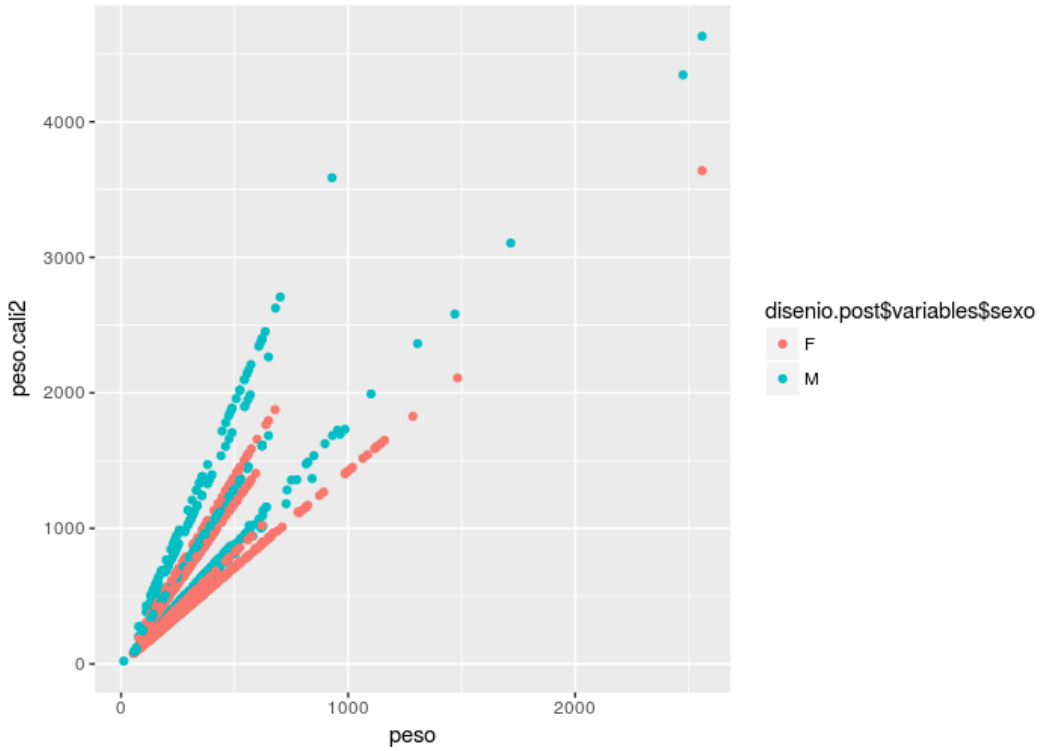

Figura 5.1: Comparación por sexo de los pesos muestrales originales vs pesos calibrados.

Mientras que para la edad, se puede ver que el incremento de los expansores se da en los 3 tramos de edad. A su vez en ambas figuras se ven que hay observaciones con pesos, luego del proceso de calibrado, que pueden considerarse extremas, que tendrían un impacto muy severo, por lo cual se decide truncar, fijando como límite inferior el percentil 10 del peso calibrado y como límite superior, el percentil 90 del peso calibrado.

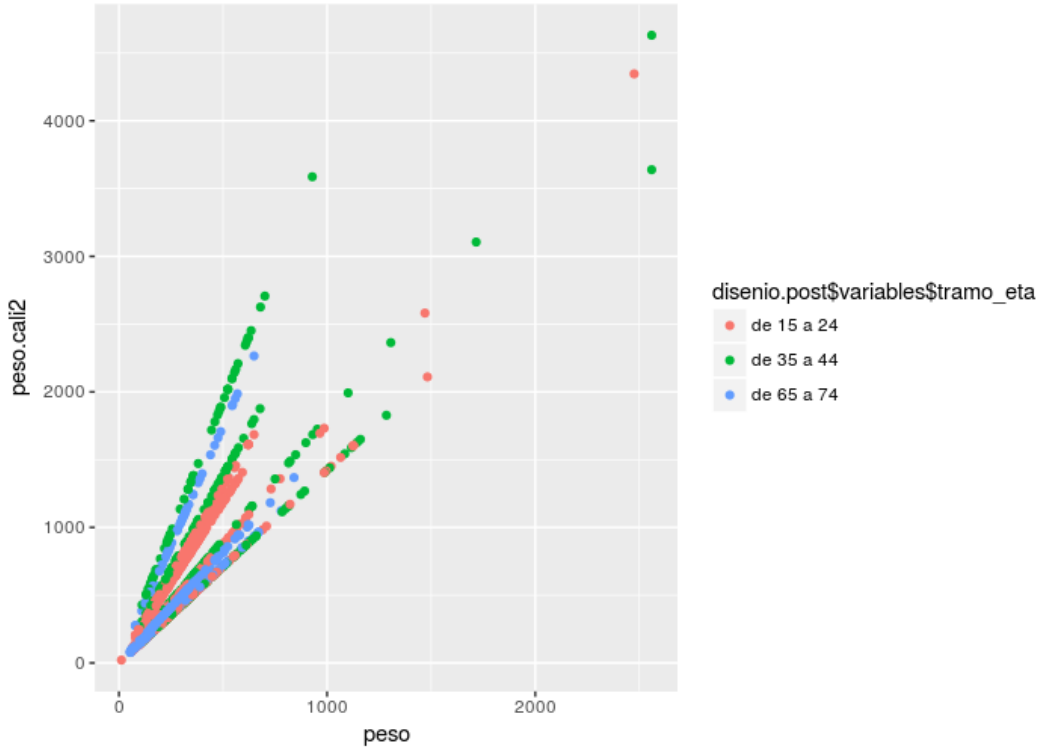

Figura 5.2: Comparación por edad de los pesos muestrales originales vs pesos calibrados.

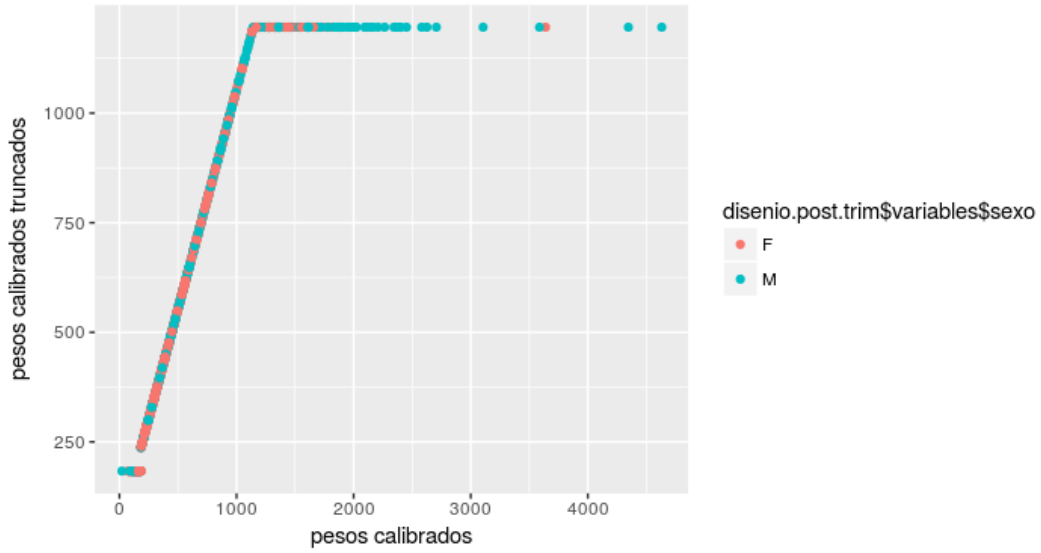

Figura 5.3: Comparación los pesos muestrales calibrados vs pesos calibrados truncados.

Se presenta a continuación las estimaciones puntuales del Componente C ya 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.

Descrición de Variables explicativas		
Variable	Nombre	Descripción
V(1)	CPO	(Nivel de CPO) (C)
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 de Interior)
V(5)	alcohol	Nivel de consumo de alcohol (2 niveles)
V(6)	Tabaco	Tabaco(2 niveles)
V(7)	INSE	INSE (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 resultados que siguen. En toda la sección se trabaja con el mismo conjunto de variables explicativas que aparecen en la Tabla 5.4

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: Prevalencias 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	0.0323	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.

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 provienen de una encuesta con diseño muestral complejo, con lo cual para elaborar la Sen 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 consideran este problema, el sistema R que tiene varias librerías para elaborar una curva ROC, tampoco lo resuelve.

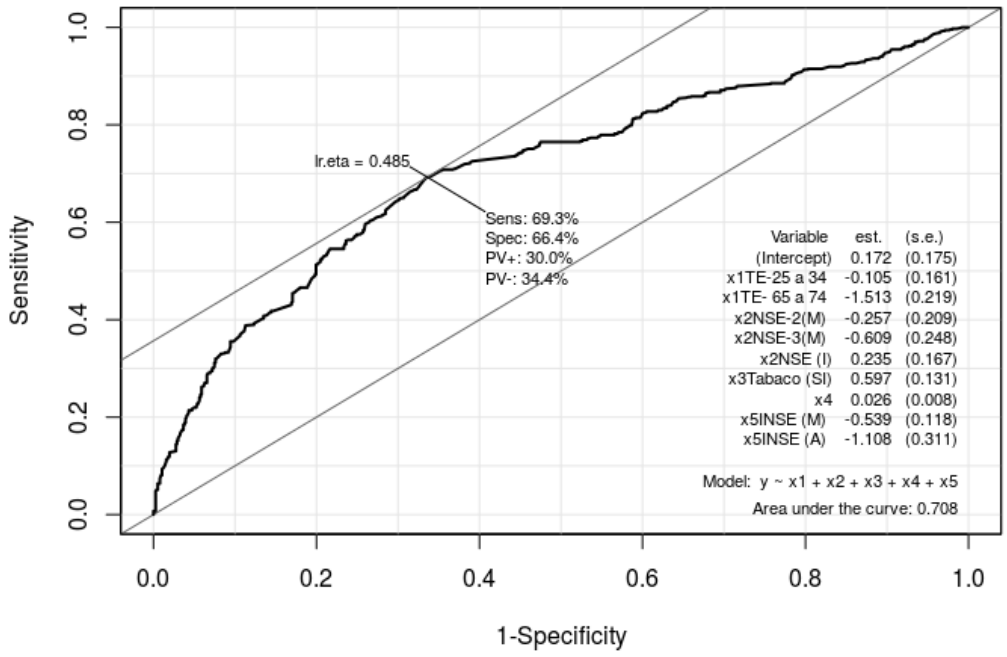

Figura 5.4: Curva Roc para modelo estimado en Tabla 5.4, sin considerar pesos muestrales.

Para superar el problema antes planteado se programa 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, estableciendo puntos de corte, creando una variable categórica sobre la que se totalizan los pesos muestrales)
Sobre la Tabla 5.5, se aplica luego una curva ROC que permite ver la perfor-

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.

mance 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 las Figuras 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.

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 cual es el comportamiento de C, para lo cual en la Tabla 5.6 se consigna la diferencia entre la distribución usando la exansión y que surge de considerar los datos crudos, que también se ve en la siguientes figuras

Observando la Figura y la Tabla 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.

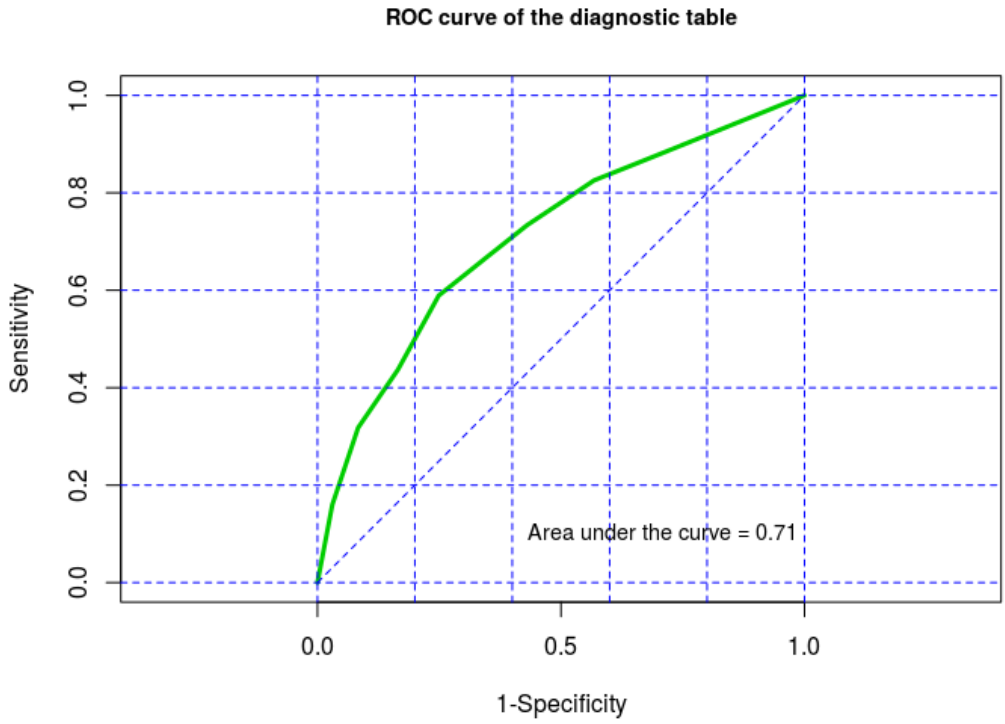

Figura 5.5: Curva Roc para modelo estimado en Tabla 5.4, considerando pesos muestrales.

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.

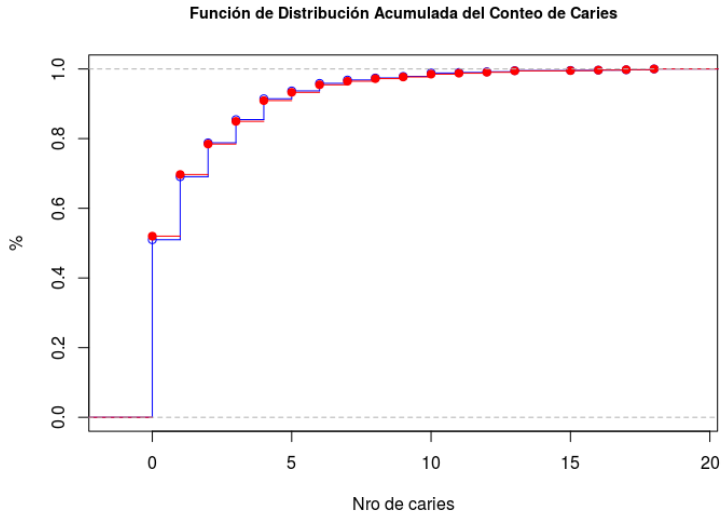

Figura 5.6: Curva de Distribución acumulada del conteo de caries, usando pesos expandidos y pesos crudos.

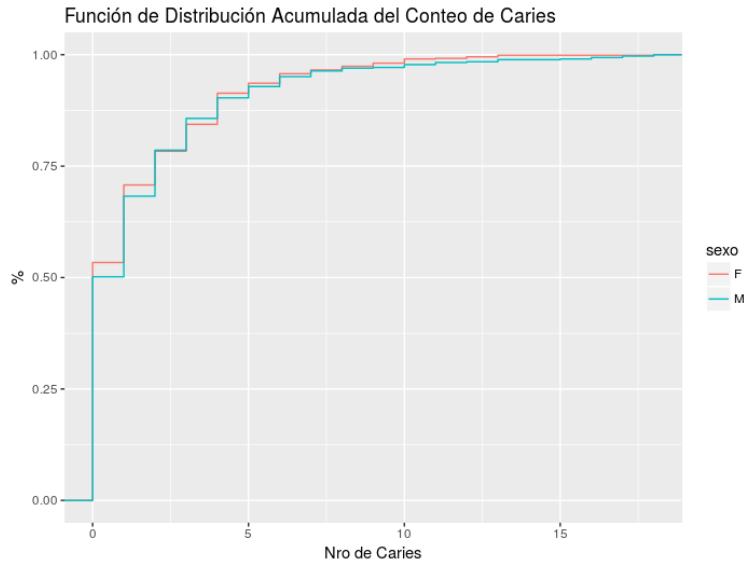

Figura 5.7: Curva de Distribución acumulada del conteo de caries, por sexo usando pesos sin expandir.

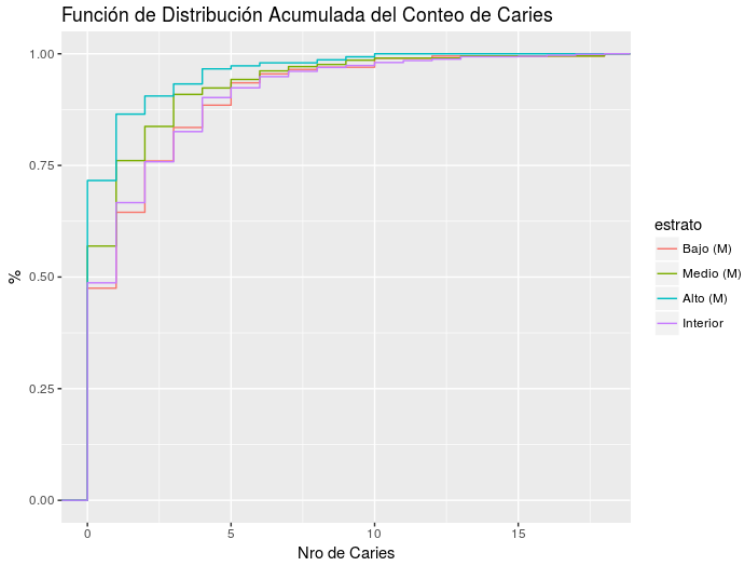

Figura 5.8: Curva de Distribución acumulada del conteo de caries, por tramo etario usando pesos sin expandir

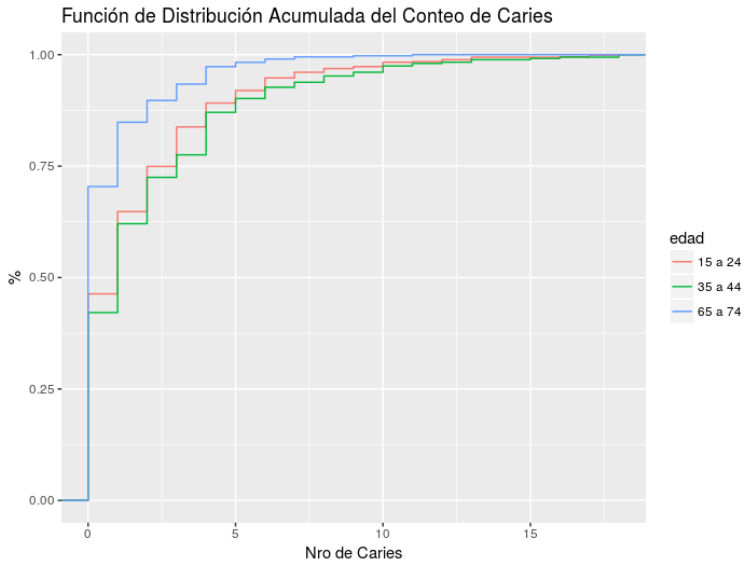

Figura 5.9: Curva de Distribución acumulada del conteo de caries, por estrato socioeconómico usando pesos sin expandir.

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 aparecen como 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 por ver la calidad del ajuste de este modelo, teniendo en cuenta 2 aspectos que ya a priori podrían hacer pensar que le distribución no es adecuada y es teniendo en cuenta el exceso de ceros y la sobre dispersión.

	Media	Varianza	Nro 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 %.

	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

Con esos 2 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 puede verse 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 y cuya justificación se presenta en la sección 5.7. El error consiste en

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 R. Logística y modelos R. Poisson.

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)

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 R. Poisson y Modelo de R. Poisson mal estimado.

5.7. Discusión

En la sección anterior se logró modelar el componente C del CPO, es decir la patología Caries usando 2 tipos de modelos: uno que da cuenta de la prevalencia de C a través de un modelo de R. Logística, que implica una transformación de la variable original, siendo un modelo 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 cual es la desventaja al usar un modelo más sencillo. Para eso en la Tabla 5.9 se comparan los coeficientes de ambos modelos donde 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.

Una mención aparte merece el mal uso de un modelo que trabaja como si fuera de conteo, pero con un recorrido que está truncado, ya que tiene solo 2 valores 0 (que significa que no tiene Caries) y 1 (que tiene cierta cantidad de Caries). Ese aspecto el método de estimación no lo toma en cuenta, si no se le advierte, lo cual lleva a estimar un modelo que no es correcto. Y 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 sencillo de interpretar que los Odds Ratios que surgen de la R. Logística. Eso se puede ver si para 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 la sección de Conclusiones 5.8, es importante retomar un aspecto antes mencionado para la evaluación del ajuste del modelo de R. Logística, presentado en la sección 5.5, donde se advertía de la inconveniencia de usar la curva ROC, no considerando 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. Hay varios autores que 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". Existe un artículo donde (Yao *et al.*, 2015), muestra mediante simulación Monte Carlo, que la varianza de la AUC, se calcula de forma diferente a como se hace habitualmente y una situación donde comparar 3 modelos de R. Logística a través de la AUC, puede llevar a resultados contradictorios optando 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, a priori podía ser más informativo al poder evaluar el impacto de las variables explicativas en el gradiente de enfermedad, la distribución de la variable de conteo es patológica, desde el punto de vista estadístico, al presentar sobre dispersión y exceso de 0. 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 sobre dispersión, ya que la librería *survey*, (Lumley, 2009), no permite trabajar con esa distribución.
Parece por lo tanto más honesto trabajar con la R. Logística, que no sería tan informativa y que muestra tener una performance global del 71 %, cuando se elabora usando los pesos muestrales y que tiene valores similares a la que se usan sin tener en cuenta este aspecto, que aparece reseñada en la Figura 5.4, pero donde algunos indicadores como el punto de corte óptimo $lr.\eta$, que estaría indicando la Sen, Esp y el Valor Predictivo Positivo (VP+) y Valor Predictivo Negativo (VP-), deberían ser tomados con precaución.

En cuanto a las variables relevantes que parecen modular la propensión a tener Caries aparecen 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 manejado 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 hace abuso de la técnica, tanto si es el quien lo hace u otro investigador con más formación en estadística.

Capítulo 6

Modelos de Conteo alternativos para los componentes C,P y O del CPO en estudio RPAFO2015

6.1. Introducción

La idea en este capítulo es comparar con otros tipos de MC, 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 existen alternativas como son la 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 rankeadas no se le presta mucha atención a estos aspectos, donde no queda muchas veces claro porque se opta por alternativas al M. de Poisson, sino que tampoco se trabaja la capacidad de ajuste (ver capacidad predictiva). Los autores muchas veces solamente se dedican a ver las variables y ajustar modelos, resolviéndose por aquellos donde aparece variables significativas pero que podrían ser muy pobres prediciendo. Este último aspecto es relevante ya que en base a esos modelos los investigadores luego en la discusión terminan elaborando teoría para explicar patologías en función de variables que no son buenas predictoras.

6.2. Diferentes Distribuciones de Probabilidad para Modelos de Conteo

Casi tan importante como poder modelar adecuadamente los modelos de conteo es fundamental en primer lugar 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 y donde queda por lo tanto determinada 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 PG las funciones de varianza son polinomios de grado 3 en μ (Hilbe, 2014).

Modelo	Media	Varianza
Poisson	μ	μ
Binomial Negativa (BN - tipo I)	μ	$\mu(1+\gamma) = \mu + \gamma\mu$
Binomial Negativa (BN - tipo 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.

Se puede comenzar por el caso más sencillo, con el Modelo de Poisson M. de Poisson, que tiene la siguiente función de cuantía:

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

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 cual una posible forma de variar el parámetro μ, de acuerdo a una distribución $Gamma(\alpha, \beta)$

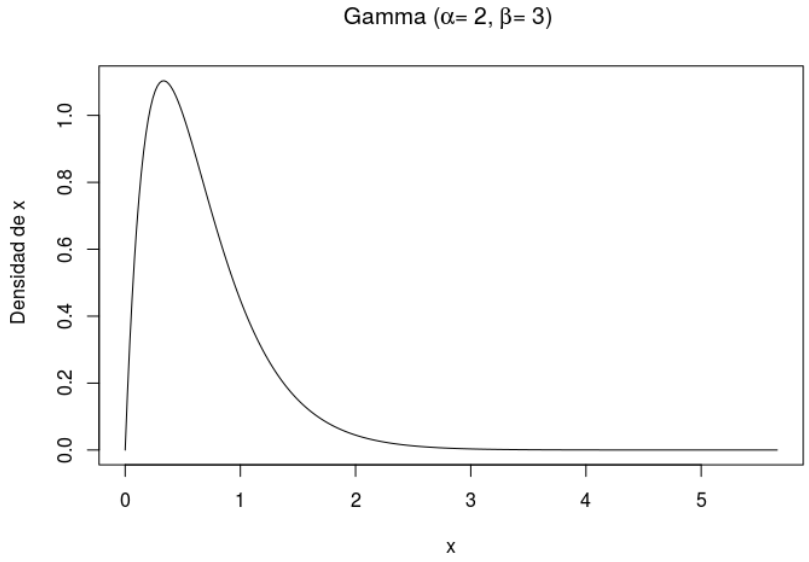

Figura 6.1: Densidad **Gamma**(α, β) para parámetro μ en distribución BN.

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

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

Otra distribución de probabilidad para la forma de variar μ es del tipo Inversa Gaussiana (IG)(μ, ϕ) . Es una distribución muy usada en Análisis de Sobrevida (Médico y Actuarial) caracterizada por ser asimétrica con parámetros de media μ y precisión ϕ. Esto da origen a la PIG, que se debe interpretar en 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).

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

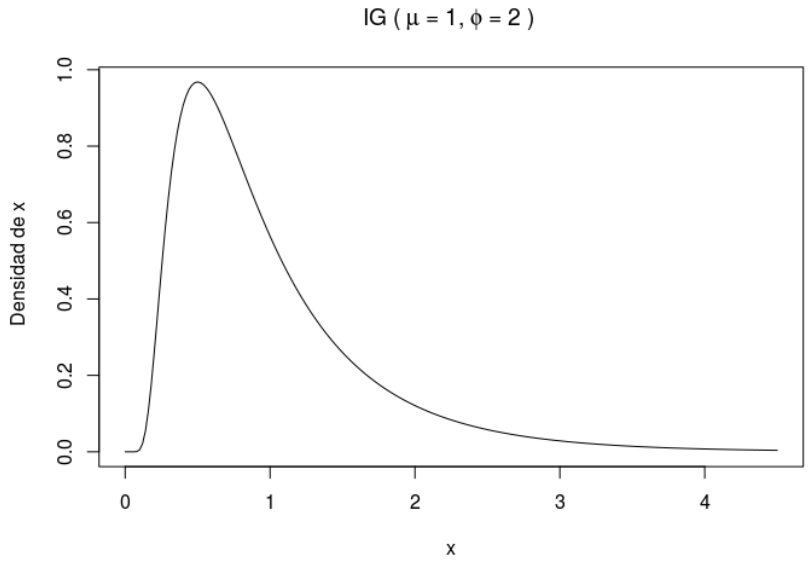

Figura 6.2: Densidad **IG**(μ, ϕ), para parámetro μ en distribución PIG.

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 Modelo Binomial Negativo-P (M. Binomial Negativo(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 era la siguiente:

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

es decir una potencia del término en $\gamma\mu$, donde ρ debe estimarse junto con el resto de los parámetros, (Hilbe, 2011).

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 tiempo en días de internación (TDI) de pacientes con patología coronaria que reciben 2 tipos de procedimientos quirúgicos, los que muestran tener una media para TDI de 8.8 y un desvío éstandar de 6.9 que muestra una sobre dispersión de 3.2. Sin embargo de los 3589 observaciones originales se intenta modelar el TDI de los que tienen menos de 8 días, siendo 1982 pacientes que verifican esa restricción mostrando una media de TDI de 4.4, con un desvío estándar de 2.30, con una dispersión de 0.79, lo que indica que no es conveniente en el contexto de regresión usar un M. de Poisson o un MBN, con lo cual una alternativa es precisamente usar la distribución PG.

Es importante entonces más allá de ver concretamente 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 éstos.

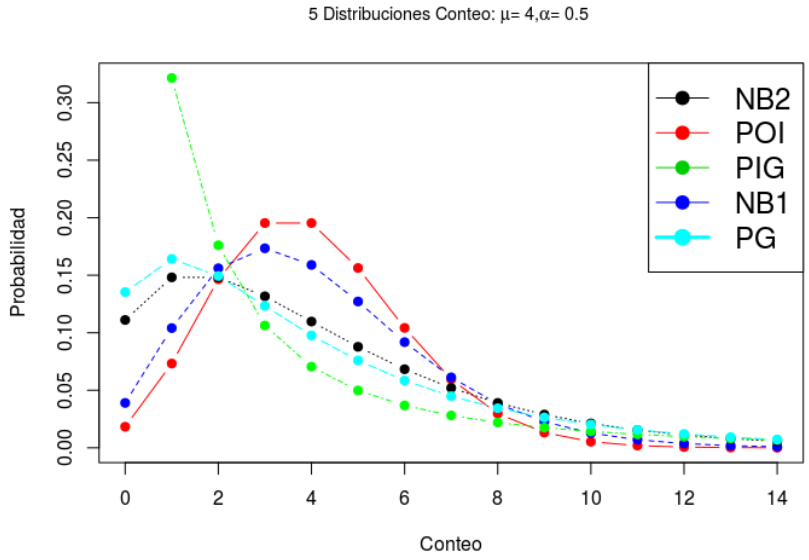

Figura 6.3: Comparación de diferentes distribuciones de Probabilidad para Modelos de Conteo.

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

6.2.1. Modelos Hurdle(MH)

Los MH o Hurdle Models, que podrían considerarse como modelos con *obstáculos*, son aquellos que combinan 2 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, (Cameron, 1998), (Mullahy, 1986).

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

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

separada.

La regresión sobre la media se da por la ecuación

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

6.2.2. Modelos con Exceso de Ceros (MEC)

Los Modelos con Exceso de Ceros (MEC) (de tipo Poisson (PEC), Binomial Negativa (BNEC)) son modelos de mezcla, que combinan un componente de conteo y una masa de probabilidad en cero, con el restante modelo para los conteos > 0,(Cameron, 1998) (Hilbe, 2014).

En este caso, hay 2 fuentes de 0 para el modelo, provenientes 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 $\pi = f_{\text{cero}}(0; z, \gamma)$

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

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 es

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

usando la función de enlace canónico.

A partir de los diferentes alternativas de modelos planteados hasta aquí se busca el objetivo de identificar los modelos de probabilidad que mejor ajustan al componente C del CPO, de los planteados en la Tabla 6.1, para luego en contextos de modelos de regresión ver cual es la mejor alternativa, que puede ser un modelo sencillo al trabajar con una sola distribución de probabilidad o un modelos más complejo de tipo combinado (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 la RPAFO2015 para los Componentes C, P, y O

Para evaluar como funcionan las distribuciones y los modelos antes presentados, se trabaja con los datos provenientes del estudio RPAFO2015 en el marco del proyecto de tipo 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, tratando de identificar su distribución, para luego estimar modelos de regresión usando las siguientes variables explicativas.

Como resumen de las diferentes parámetros de dispersión implícitos en los MC, se presenta las alternativas de librerías en (R Core Team, 2016), para su estimación

- $y \sim Poi(\mu) \longrightarrow$ Equidispersión *library(stats)*, (R Core Team, 2016);
- $y \sim BN(\mu, \sigma) \longrightarrow$ para el tratamiento de la Variabilidad de parámetros dispersión, $\lambda \sim Gamma(\mu, \lambda)$, *library(MASS)*, (Venables y Ripley, 2002a), *library(COUNT)*, (Hilbe, 2016);
- $y \sim BN(\mu, \sigma, \rho) \longrightarrow$ Variabilidad de parámetros dispersión a través de observaciones, $\lambda \sim IG(\mu, \phi)$, *library(gamlss)*, (Rigby y Stasinopoulos, 2005);
- $y \sim PIG \longrightarrow$ Variabilidad de parámetros dispersión, $\lambda \sim IG(\mu, \lambda)$, *library(gamlss)*, (Rigby y Stasinopoulos, 2005).

Se detallan las variables explicativas que se usarán para modelar los diferentes componentes del CPO en la Tabla 12.2

En principio se tiene la siguiente distribución para los 3 componentes y el CPO:

Si bien se estiman modelos para los 3 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 desde el punto de vista epidemiológico tiene.

Usando la librería *gamlss* (Rigby y Stasinopoulos, 2005), se puede estimar la mejor distribución paramétrica para ajustar la variable C. Si bien los resultados al aplicar la función *fitDist()* aparecen modelos que consideran inflación

Descrición 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.

Componentes	n	$\bar{x}$	sd	mediana	min	max
C	602.00	2.50	3.05	2.00	0.00	25.00
P	602.00	10.17	8.48	8.00	0.00	32.00
O	602.00	3.66	4.10	2.00	0.00	22.00
CPO	602.00	16.33	8.11	17.00	0.00	32.00

Tabla 6.3: Medidas de resumen de los componentes de CPO.

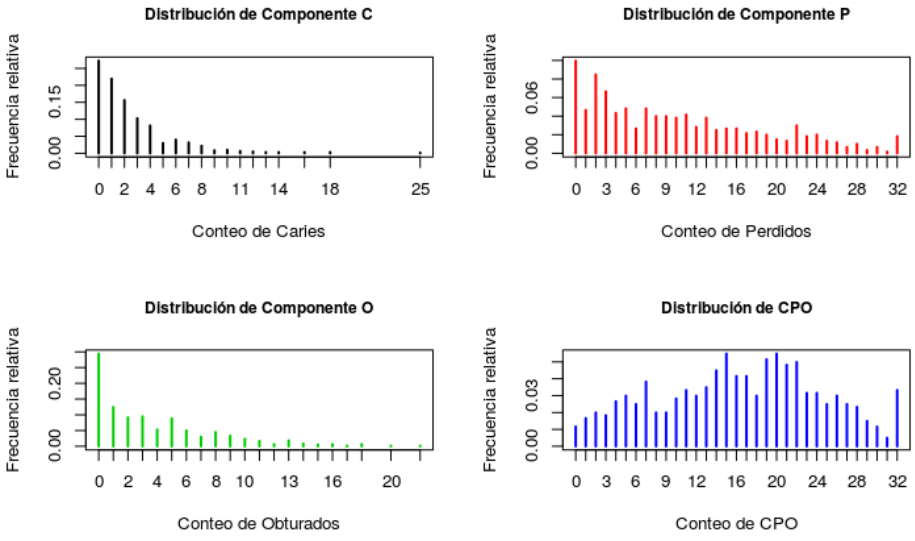

Figura 6.4: Distribución para el CPO y sus 3 componentes.

de 0, solamente se presentan los 5 modelos parámetricos vistos en la sección 6.2. En la Figura 6.5 se presentan los ajustes hechos 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 mejor alternativa, siendo que el PG es el 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.

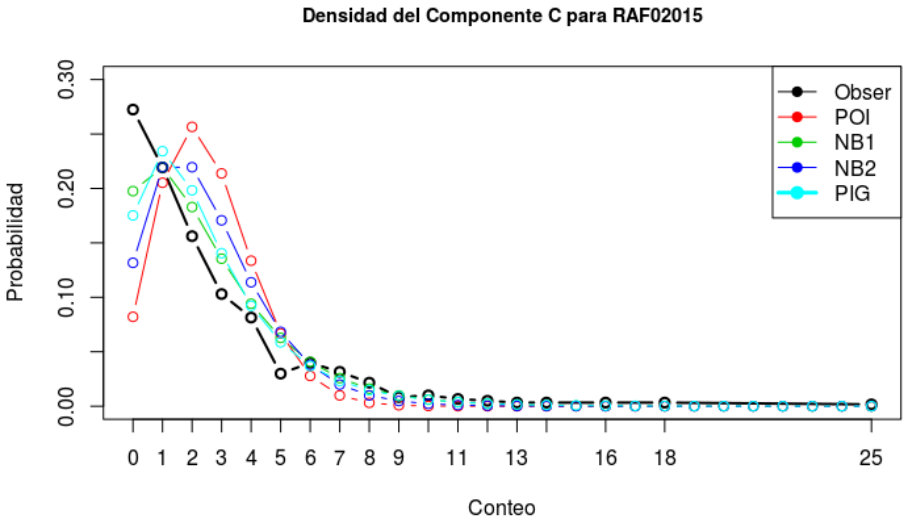

Figura 6.5: Distribución de diferentes MC para C, dado el valor de $\mu = 2.5$.

modelo ajustado para componente C			
tipo de MC	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.

En la Figura 6.7 puede verse 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.

En la Tabla 6.4 se muestra un resumen de los mejores modelos MC ajustados para el CPO y sus 3 componentes. En cada modelo ajustado se realiza

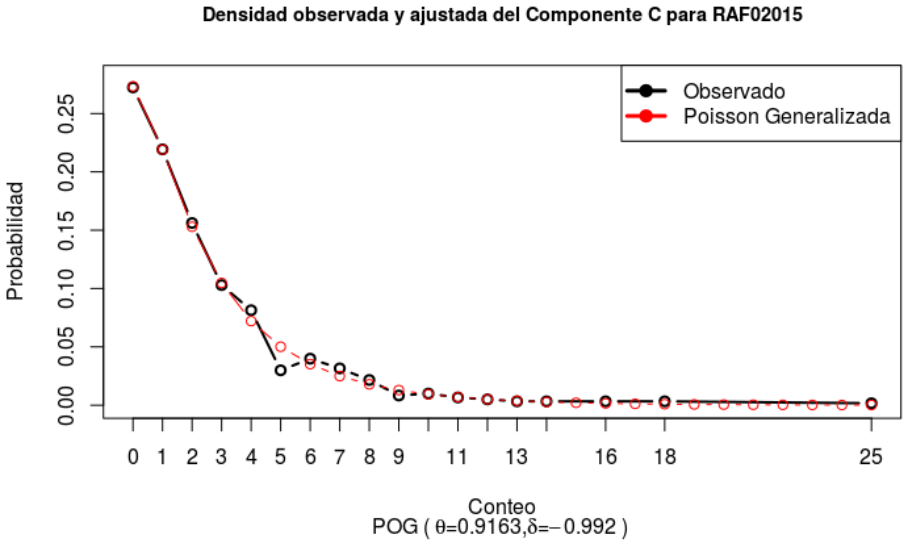

Figura 6.6: Ajuste del C, dado el valor de $\mu = 2.5$ para un modelo PG.

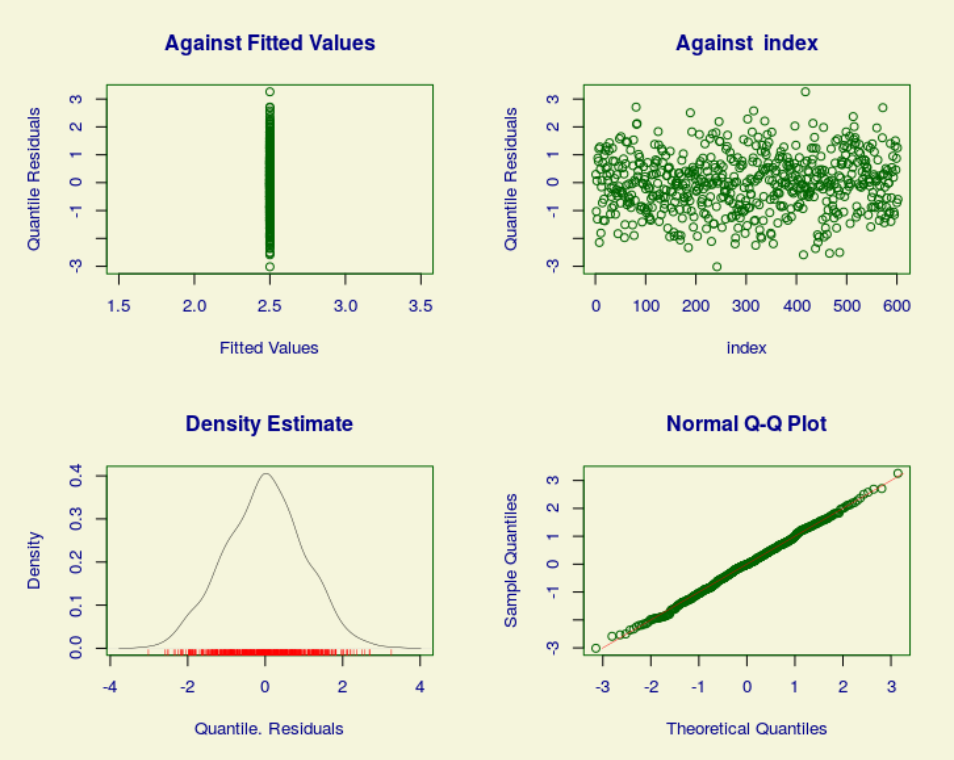

Figura 6.7: Gráficos de Bondad de ajuste para el modelo de conteo PG para C.

el mismo proceso iterativo para encontrar el modelo con mejores, indicadores de bondad como el Akaike Information Criteria (AIC), Bayesian information criterion (BIC) y la devianza global y a su vez se realiza el diagnóstico de ajuste a través de los residuos.

Para entender mejor los resultados encontrados se presenta en la siguiente tabla para cada componente del CPO cual es la bondad de ajuste y la jerarquía que queda para los diferentes MC presentados en 6.2 y donde finalmente en la Tabla 6.5 se consigna solamente los MC presentados en detalle previamente.

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 MC para CPO y sus 3 componentes.

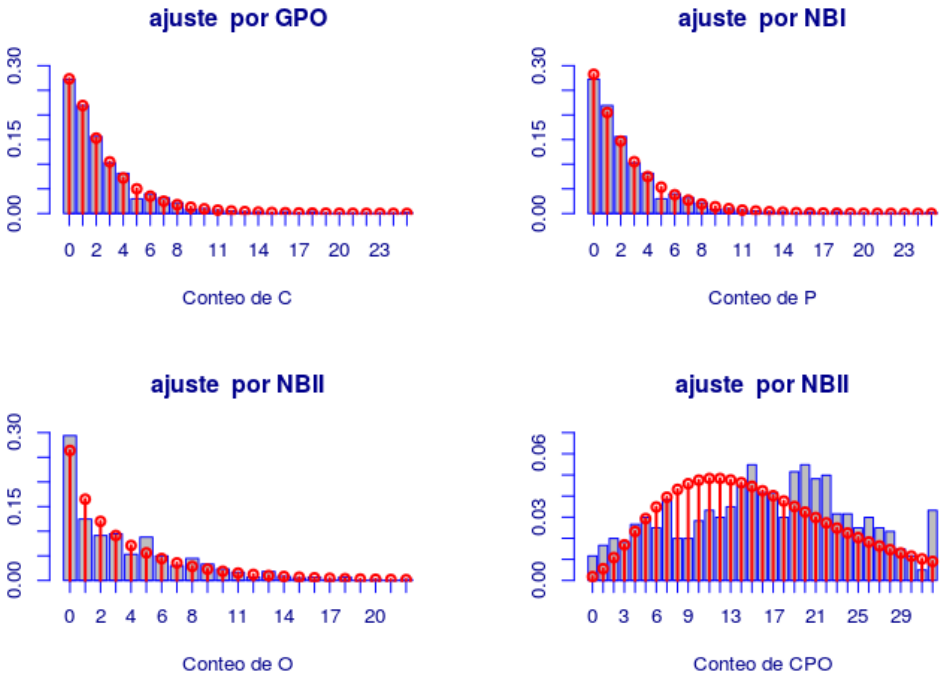

Figura 6.8: Representación gráfica de los modelos de probablidad ajustados para CPO y sus 3 componentes.

Un aspecto a tener en cuenta es que los coeficientes que se presentan 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 2 escenarios para poder manejar distribuciones que fueron 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.

Tipo	modelo ajustado para componente C parámetros			
PG	Coeficientes	E.E.	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	
Tipo	modelo ajustado para componente P parámetros			
BN - tipo I	Coeficientes	E.E.	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	
Tipo	modelo ajustado para componente O parámetros			
BN - tipo II	Coeficientes	E.E.	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	
Tipo	modelo ajustado para componente O parámetros			
BN - tipo II	Coeficientes	E.E.	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 3 componentes, (Escenario A).

Tal como se ve en la Tabla B.1 en el apéndice B aparecen 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 2 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.

Finalmente antes de pasar a la etapa de elaboración de modelos de pronóstico con las variables regresoras presentadas antes, en la Figura 6.8 se presentan los mejores modelos ajustados para el CPO y sus 3 componentes

A continuación se presentan los medidas de resumen para las variables regresoras del componente C, ya que solamente por ser de los 3 el más relevante desde el punto de vista epidemiológico y el más frecuentemente estudiado, será el único estudiado mediante modelos de regresión en este capítulo.

Medidas de resumen						
Variables	n	$\bar{x}$	sd	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	0.0000
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	0.0013
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 quasi-Poisson para componente C.

El modelo estimado presentado en la Tabla 6.8 toma en cuenta la sobre-dispersión que existe, que este caso es de 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 que ya se vió que ajusta mejor a los datos) como aparecen 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	0.0000
cedad	-0.0322	0.0039	-8.28	0.0000
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 Dispersion para quasipoisson =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
cedad	-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.

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.

Si en lugar de estimar un modelo MH con distribución Poisson para el componente truncado se usa la distribución BN, donde 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 estaría indicando que los cambios al usar el modelo con BN para el componente truncado no son relevantes

Finalmente en la Tabla 6.13 puede verse un resumen de la performance de los diferentes modelos, donde se presentan además indicadores de ajuste, donde 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

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, usando modelos parámetricos (MLG) y modelos con obstáculos.

Caries, lo que estaía marcando que el modelo hallado parece adecuarse a una 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 presentados los diferentes modelos de probabilidad en la sección 6.2, de los cuales se presentaba un ejemplo en la Figura 6.5 puede verse la forma de la distribución de los 4 componentes del estudio RPAFO2015, donde en primer lugar en esta sección se discute sobre modelado del componente C dada la importancia de éste desde el punto de vista epidemiológico y de la Salud Pública, para luego presentar el comportamiento observado para los restantes componentes del CPO.

Puede verse por lo tanto que si se trabaja solamente 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 MC asociado con las distribucines más sencillas, es el que corresponde a la BN-II, como se ve gráficamente en la Figura 6.5, pero sin perder de vista que para los 5 modelos paramétricos presentados en 6.2, el ajuste es pobre. Por eso motivo tal como se adelantó en la sección 6.3, la mejor alternativa es la PG, tal como se presenta en 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), y la PG está en el primer lugar. Entre todas esas distribuciones de conteo están las básicas presentadas en la sección 6.2, las que corresponden a 6.2.1 y 6.2.2, y varias variantes de las mismas que dada su complejidad matemática se dejan de lado. En la Figura 6.8, se presenta 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, el ajuste de estos a través de MC 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 ser

multimodal, parece difícil de responder a un modelo con distribución conocido, sino que parece adecuado considerar a un proceso de mezcla.

En cambio en un contexto de regresión en las Tablas 6.9 y 6.10 se presentan los resultados de ajustar por un M. de Poisson y un BN, dada la sobredispersión que existe y donde de las 8 variables regresoras previamente consideradas en la Tabla 6.3, solamente resultan significativas el CPO, la edad, el sexo, la cantidad de días en la semana donde se ingiere bebidas azucaradas y el ingresos de los individuos participantes del estudio. Para ambos modelos, donde los coeficientes muestran valores similares, se consideran las variables que son significativas y la asociación con el logaritmo del número medio de Caries muestra resultados esperable. La cantidad de Caries se asocia con un mayor nivel de CPO , con un aumento promedio de una Cariers por cada punto extra en el CPO, con un decremento de casi una Caries 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 a sexo masculino, donde el número medio de Caries es de casi 1.36 con respecto a las mujeres. Finalmente los ingresos muestran una asociación negativa donde las personas que están en el último tramo de ingresos tienen una reducción en el número de Caries importante con respecto a los 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 que para el investigador en biomedicina no advertido podría llevarlo a cometer errores, si no se toma en cuenta un aspecto que en general no se considera y es la capacidad predictiva del modelo estimado.

Para este caso donde el conteo muestra comportamiento patológico al tener sobredispersión y exceso importante de 0, es fundamental verificar el grado de ajuste y no alcanza por lo tanto que las variables sean significativas, ya que no tomar 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 que sucede con el conteo de 0. El componente C muestra que hay 164 personas sanas de Caries, es decir con un conteo=0, mientras que los 2 modelos básicos estimados para el caso de Poisson pronos-

tican 82 personas sin Caries y 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 la Tabla 6.11 y 6.12, muestran una mejor capacidad de detectar personas con C=0 y 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 modela mediante un logit y es el que permite saltar el obstáculo, es el que dice que perfil tiene la persona para tener o no Caries, que en el caso del MH Poisson 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, las variables que permiten saltar el obstáculo a través del modelo logit son CPO, edad, sexo e ingreso, mientras que 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 MC para el estudio RPAFO2015

De los resultados encontrados resulta fundamental recordar las siguientes 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

- Previamente examinar cuales 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 los cuales desarrollar teoría epidemiológica que tenga sentido para el investigador, ya que estaría basada en modelos que no son comparables con los datos bajo 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 2 submodelos
    - Uno que trabaja con personas que no tienen Caries (si está fuera la variable de conteo por ejemplo), aspecto que se modela con el componente 1 del modelo (MH);
    - Cuando tienen Caries, la cantidad de éstas se modelan con el componente 2 del modelo (MH).

Perfectamente podría suceder que las variables regresoras que se usan para salir del obstáculo no necesariamente sean las mismas que las que contribuyen a modelar el conteo truncado (> 0) e incluso siendo las mismas pueden cambiar el sentido de la asociación o la intensidad de las mismas 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 vió que las distribuciones no son identificables fácilmente 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 las limitaciones de los mismos y lograr un equilibrio entre modelo sencillo pero adecuado.

Capítulo 7

Uso de la Regresión Beta para la Creación de Indicadores Alternativos para la Vigilancia en salud bucal

7.1. Introducción

Tal como se manifestó en capítulos anteriores pueden existir limitaciones en los indicadores generalmente utilizados en la epidemiología y salud pública, 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 por lo tanto correctamente los fenómenos bajo estudio. Esta característica de uso de indicadores limitados (al no tomar 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 solamente en salud bucal, por lo menos en nuestro país.

7.1.1. Antecedentes

Para entender los aspectos antes mencionados, en este capítulo se consideran 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 en forma más detallada y gradualista, estadíos o niveles de enfermedad. Si se retoma como ejemplo el caso del CPO que tenía 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 que siendo un índice univariado enmascara muy diferentes configuraciones de los componentes aislados es necesario manejar alternativas, como utilizar los 3 componentes del CPO por separado, transformándolos en tasas, o proporciones o índices basados en medidas de entropía; es decir considerar la misma información pero analizándola 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 ińdices del 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 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 que puedan 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 convirtiéndolos en proporciones usando indistintamente los 3 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.

P	T	U	V	W	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE
Motivo de consulta	c18	c17	c16	c15	c33	c34	c35	c36	c37	c38	C	P	O	CPO	prop1	prop2
TRATAMIENTO	8	0	3	3	0	2	1	3	3	8	4	8	8	20	33%	20%
PROTESIS	0	2	3	3	0	0	0	3	3	8	2	12	0	14	100%	14%
PROTESIS	3	0	0	0	3	3	3	3	3	3	2	11	0	13	100%	15%
EXTRACCION	2	2	3	0	0	0	0	3	2	0	6	4	0	10	100%	60%
EXTRACCIONES REST	3	3	3	3	0	2	3	3	3	3	3	18	2	23	60%	13%
INFECCION	3	3	3	3	3	3	3	3	3	3	3	26	0	29	100%	10%
PROTESIS	2	3	3	3	0	0	3	3	3	3	6	19	0	25	100%	24%
EXTRACCIONES	2	3	3	3	0	3	0	3	3	3	3	20	0	23	100%	13%
REVISACION	8	0	3	3	0	3	0	3	3	3	1	18	0	19	100%	5%
EXTRACCION	2	3	3	2	0	0	0	0	3	2	6	11	0	17	100%	35%
ARREGLOS	8	2	2	2	0	0	0	2	3	3	15	3	0	18	100%	83%

Figura 7.1: Cálculo de las diferentes proporciones en CPO.

En la sección 7.3 se amplia y detalla el motivo para considerar 3 niveles de indicadores medidos con diferentes proporciones.

7.1.3. Formulación del modelo de probabilidad BETA

De los diferentes modelos que se podrían utilizar para modelar proporciones, se considera el modelo de probabilidad *BETA* que se presenta a continuación, para luego ser 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 y Neto, (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 se conoce como 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 7.2 se pueden ver diferentes ejemplos de distribución *BETA* al cambiar los parámetros de dispersión o precisión.

La formulación del modelo predictivo puede ser hecha a partir de los trabajos de, (Kieschnick y McCullough, 2003), (Salinas-Rodríguez *et al.*, 2009) donde

- 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 $\mu_i = g^{-1}(x_i^\top \beta)$ donde β, el vector de parámetros de regresión, se estima por máxima verosimilitud (ML).

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

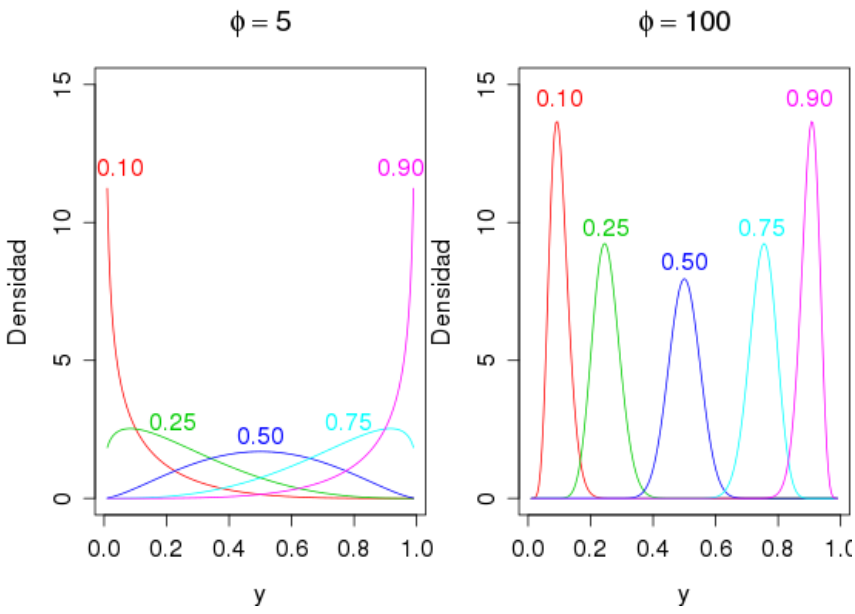

Figura 7.2: Densidades **BETA** en el intervalo $(0, 1)$ para diferentes valores de μ y ϕ.

El modelado de la dispersión se puede hacer a través de

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

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

La funciones de enlaces (link) para modelos de Regresión Beta pueden ser algunas de las que siguen

- 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 a través de todas las observaciones, con lo cual es necesario modelarlo, tal cual se hizo con la media.

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

$$g_1(\mu_i) = \eta_{1i} = x_i^\top \beta, \tag{7.7}$$
$$g_2(\phi_i) = \eta_{2i} = z_i^\top \gamma, \tag{7.8}$$

donde $\beta = (\beta_1, \ldots, \beta_k)^\top$, $\gamma = (\gamma_1, \ldots, \gamma_h)^\top$, $k + h < n$, son los *coeficientes de regresión* de ambas ecuaciones, η_{1i} and 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 $(0, 1)$ con lo cual, 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 necesario una transformación extra que trunca los extremos con 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, permitiendo poder estimar los parámetros.

7.2. Aplicación de modelos de Regresión Beta al estudio RPAFO2015

Los resultados que se presentan surgen del análisis de los componentes del CPO, considerando varias transformaciones en proporciones, con diferentes criterios de elaboración y que desde el punto de vista epidemiológico representan diferentes dimensiones del problema en el contexto del estudio **RPDAFO2015**. 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 7.3, donde se ve una gran asimetría para el componente de caries y para el componente de obturación, lo que habla de una población con una

carga de (enfermedad actual y/o pasada) media, pero con una mayor presencia del componente P 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.

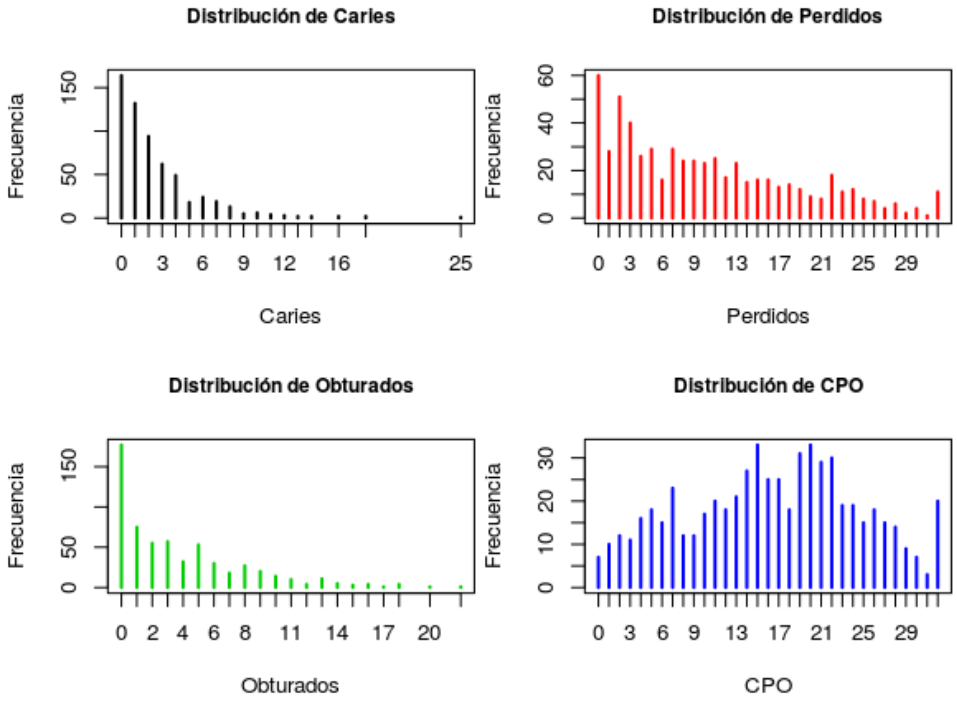

Figura 7.3: Densidades empíricas de los componentes del CPO.

También es importante ver como cambian los resultados al trabajar con los conteos re-expresándolos en proporciones, eligiendo diferentes formas de normalizar tal como se presentaron en 7.1.2

variables	n	$\bar{x}$	SE	mediana
prop1	560.00	0.54	0.38	0.60
prop2	560.00	0.46	0.38	0.40
prop3	602.00	0.48	0.25	0.47
prop4	602.00	0.32	0.27	0.28
prop5	595.00	0.19	0.23	0.11
prop6	595.00	0.54	0.30	0.57
prop7	595.00	0.27	0.28	0.19
prop8	602.00	0.08	0.10	0.06
prop9	602.00	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, PyS 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 puede obtener mejoras, a los resultados que se manejaron en el capítulo 6 se presentan varios modelos de regresión *Beta* transformando en proporciones los diferentes componentes del CPO. Si bien en los componentes no existen datos faltantes para sus componentes, como se ve en la tabla 7.2, al hacer la transformación en proporciones, según cual sea la transformación, aparecen datos faltantes al tener denominadores con conteos nulos. Teniendo en cuenta este detalle, se analiza si ese pérdida es aleatoria o por el contrario tiene un patrón definido.

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

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 7.3: Conjunto de variables regresoras usadas para los modelos de Regresión Beta en RPAFO2015.

Para cada modelo evaluado para las diferentes proporciones antes definidas en la sección 7.1.2 se toma de rutina considerar este bloque de 8 variables explicativas y se reporta el modelo final luego de que se eliminaron las variables no significativas y a su vez en cada caso se verificó la no existencia de heterocedasticidad, que daría lugar a un único parámetro ϕ de dispersión. En casa contrario se procede a modelar la dispersión asociada a cada variable tal como se detallaba en la ecuación (7.8) y el criterio para considerar si es pertinente la modelización de la 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 cuales 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 Phi 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 *prop1* en estudio RPAFO2015.

El **modelo 2** sería el que considera la relación entre *prop2* y el grupo de variables explicativas, pero dado que en realidad por como está construida, esta proporción es el complemento de *prop1*, por lo cual el modelo sería el mismo que para *prop1*.

Se cambia de grupo de proporciones del Nivel I, donde se trabaja con las 2 proporciones antes presentados y considerando solamente las piezas presentes, es decir dejando de lado el componente O, pasando al grupo Nivel II.

Se incorporan variantes a *prop1* y *prop2*, al usar el componente O, en el que el primer modelo que se presenta es para *prop3* consignado en la Tabla 7.5, y 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 y por otra parte es necesario modelar la heterocedasticidad, donde la edad aparece como 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 Phi 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 *prop3* 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

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

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	< .0001
(Test de Breusch-Pagan)				
BP = 61.69, df = 5, p-value = 5.436e-12				
(Parámetro Phi 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 *prop4* en estudio RPAFO2015.

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
sexoMasculino	-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 Phi 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.

Pasando a las proporciones del Nivel III, donde aparece por primera vez la dimensión de Salud, a través del componente S de piezas sanas, el indicador de Salud en el momento actual (*prop6*) tiene un modelo ajustado que presenta las siguientes relaciones:

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 Phi 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 *prop6* en estudio RPAFO2015.

En la Figura 7.4 se muestra como 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).

Para el indicador de necesidad de prótesis en el momento actual (*prop7*) el modelo ajustado es el que aparece en la Tabla 7.9

Finalmente para ordenar los resultados encontrados y facilitar la discusión de los mismos en la sección 7.3, se presenta la Tabla 7.12, que permite discernir la performance de cada modelo.

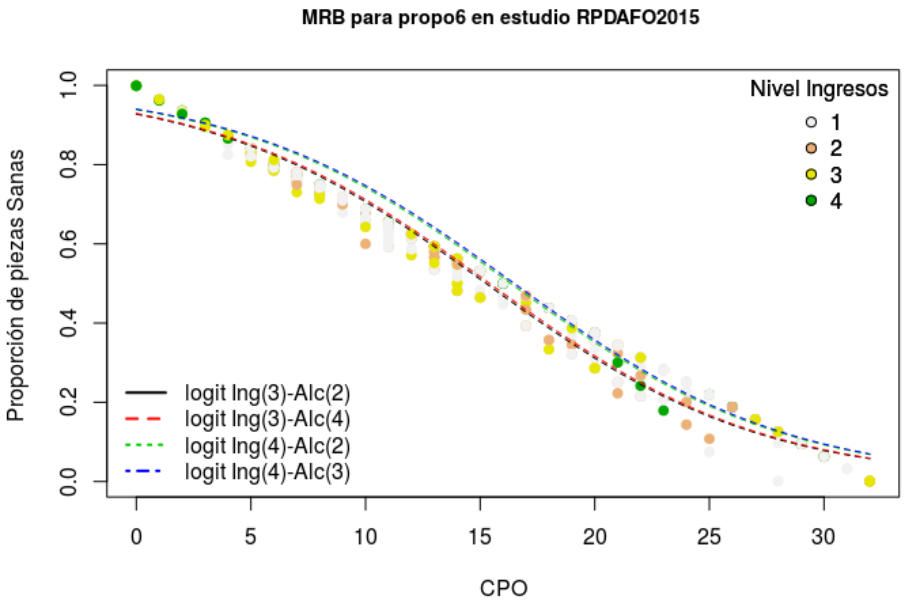

Figura 7.4: Relación entre *prop*6 y CPO para (**modelo 6**).

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 Phi 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 *prop7* en estudio RPAFO2015.

Coeficientes (Modelo 8 on 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 Phi 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.

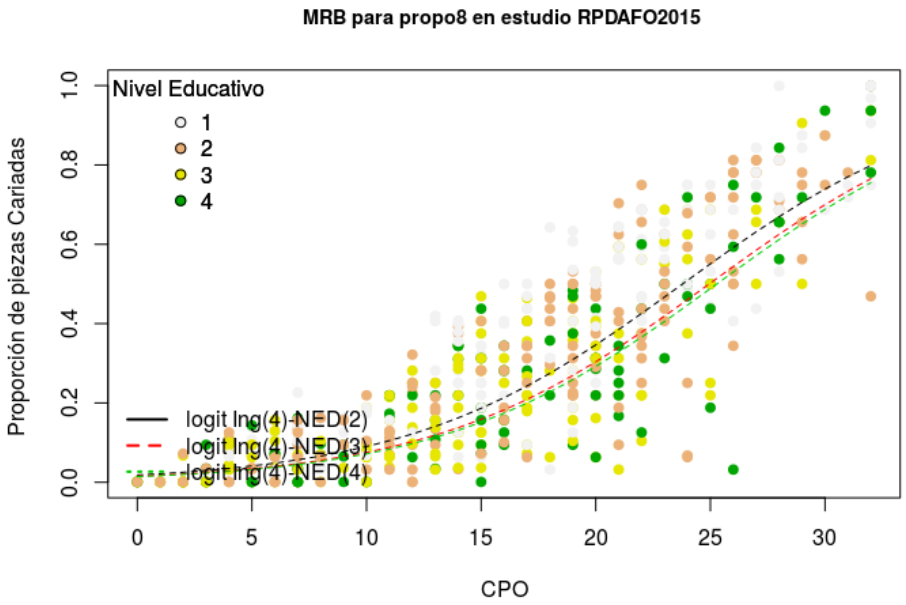

Figura 7.5: Relación entre *prop*8 y CPO para (**modelo 8**)

Coeficientes (Modelo 9 on 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(3)	0.621	0.117	5.29	< 0.0001
N.EDU(4)	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 Phi 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 *prop9* en estudio RPAFO2015.

Tipo y Calidad de ajuste para cada modelo estimado				
proporción	modelo	Pseudo-Rcuadrado	Heterocedasticidad	Grupo
prop1	modelo 1	0.171	No	Insuficiente
prop2	modelo 2	0.171	No	Insuficiente
prop3	modelo 3	0.158	Si	Insuficiente
prop4	modelo 4	0.684	Si	Suficiente
prop5	modelo 5	0.202	Si	Insuficiente
prop6	modelo 6	0.778	Si	Suficiente
prop7	modelo 7	0.115	Si	Insuficiente
prop8	modelo 8	0.687	Si	Suficiente
prop9	modelo 9	0.151	Si	Insuficiente

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

7.3. Discusión sobre los Modelos de Regresión Beta estimados

A lo largo de la sección 7.2 que contiene los resultados se fueron viendo diferentes modelos de Regresión Beta, siguiendo una lógica común de usar las mismas variables explicativas para todas las proporciones que previa transformaciones, podían dar cuenta de las diferentes dimensiones del CPO. Estas proporciones se agruparon en 3 niveles, siguiendo una jerarquía en términos de lo que conceptualmente pueden llegar a representar, donde las del Nivel I, son indicadores que muestran la enfermedad actual (C) o pasada (O) pero que dejan de lado el componente O que da de por sí representa una patología a ser revertida mediante la colocación de prótesis. Los 3 otros indicadores del grupo del Nivel II ya consideran la totalidad de los componentes de CPO, clásicamente utilizados en el análisis de epidemiología, con la ventaja que al considerar este tipo de modelado, muestra ventajas sobre los que se pueden alcanzar mediante el uso de MC, ya tratados en el capítulo 6, con mucho detalle. Finalmente los 4 últimos indicadores presentan una ventaja de incorporar una dimensión no siempre utilizada, como son las piezas sanas (es decir que aún no sufrieron patología cariogénica), y con la características que el componente S (sanos), pasa a formar parte del algoritmo de cálculo al estar en el denominador, pero a su vez de poder también introducir una proporción con la *prop6*, que es la única que muestra Salud (y no ausencia de ella), como si lo hacen los indicadores manejados en el grupo Nivel I y Nivel II.
Siguiendo por lo tanto este análisis jerárquico, a lo largo de la sección aparecen modelos que no siempre contienen las mismas variables explicativas y donde a su vez aparece una fuente de variabilidad extra, evaluada a través del componente ϕ de dispersión, donde también las variables que deben ser tenidas en cuenta, no son las mismas, lo que sugiere rápidamente una reflexión para el investigador en biomedicina que debiera trabajar con esta metodología. Para explicar dimensiones que están íntimamente relacionadas como son los componentes del CPO, el mecanismo indirecto (ya que no se puede hablar de causalidad), por el cual existe mas o menos patología es diferente, según sea la que se considere, y a su vez la heterogeneidad, que no debe ser vista como algo inadecuado, sino por el contrario como una forma de entender mejor la casuística observada, también parece estar regida por diferentes mecanismos al no ser siempre las mismas variables, por las cuales es necesario ajustar.

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

Tal como se dijo en el capítulo 3, lo ideal es que el especialista en ciencias médicas puede disponer de modelos estadísticos parsimoniosos, logrando un adecuado equilibrio entre sencillez y buenos resultados. Pero es ahí donde se vuelve sobre un aspecto ya mencionado a lo largo de los capítulos hasta ahora presentados de la parte II de la tesis, y que consiste en no olvidarse que en un contexto de modelización no puede el investigador quedarse en la etapa de identificación de las variables significativas, sino dadas las mismas evaluar su utilidad, en términos de la capacidad predictiva.

Es por eso que se puede decir que hay 2 grupos de modelos identificados en la Tabla 7.12 como suficientes y como insuficientes. En este aspecto se debe recalcar que cuando se habla de modelo suficiente, es en el sentido de que logra todos los cometidos antes planteados y como podrá advertise el motivo por el cual un modelo, donde para todos las variables explicativas son todas significativas, se considera suficiencia, si 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 9 modelos solo 3 pueden considerarse como suficientes, con lo cual serían recomendables y se pasan a comparar.

Si bien 2 de los 3 son los del grupo Nivel III y por lo tanto intentan modelar indicadores que tiene 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.

Para eso a modo de resumen se presenta la Tabla 7.13

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

Tabla 7.13: Síntesis de los modelos considerados suficientes para el estudio RPA-FO2015 (Het refiere a si es Heterocedástico).

La tabla anterior muestra que el nivel de CPO está presente en los 3 modelos, lo mismo sucede con la edad; para el caso del modelo 4, se ve que el componente P está asociado con el nivel educativo, mostrando que a mayor nivel menor proporción de caries y con la característica de que el hábito de fumar es relevante para marcar la heterogeneidad aumentando la proporción de P.

El modelo 6 que trabaja con la proporción de dientes sanos, muestra una asociación positiva con mayor ingreso y relación inversa con la ingesta de alcohol y el hábito de fumar, siendo la edad la variable que modula la heterogeneidad.

Finalmente el modelo 8, muestra que la proporción del componente C se asocia positivamente con la edad y el CPO, algo esperable, mientras que los ingresos y el nivel educativo se asocian negativamente, siendo nuevamente la edad y el CPO los que aparecen en la heterogeneidad.

Antes de concluir es importante, destacar que las otras proporciones para los que se estimaron modelos, no es que no 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 Regresión Beta , dan resultados satisfactorios y complementarios a los hallados en el capítulo 6 y resta poder 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 Modelo de Regresión Beta (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 $(0, 1)$, ya que los valores cambian en incrementos de $\frac{1}{32}$, lo que puede ser un inconveniente a la hora de evaluar la predicción.

En consecuencia este aspecto más 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 y pensar si también es posible obtener mejorías estratificando o segmentando los datos mediante el uso de técnicas estadísticas como los Árboles de Clasificación y Regresión.

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, en las que pueden agruparse enfermedades como las cardiovasculares, diabetes, cáncer y enfermedades respiratorias crónicas, son actualmente la causa de mortalidad a nivel mundial más importante con un peso de 63 % de las muertes globales, y con una característica extra y es que casi 40 % de estas muertes se producen entre los 30 y 70 años, con la consecuente carga social al ser el tramo de edad más relevante donde están 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, lo que se traduce en discapacidades por períodos más prolongados previo a que sobrevenga la muerte.

En este conjunto de enfermedades es fundamental el papel que juegan los

estilos de vida que se relacionan con aspectos como 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, consumo nocivo de alcohol, y el consumo de tabaco. Estos factores a su vez actúan en forma directa o indirecta, creando otros factores de riesgo como son la obesidad, los trastornos del metabolismo de los hidratos de carbono, la hipertensión arterial (HTA) o las dislipemias, (Skapino y Álvarez-Vaz, 2016). Una preocupación a nivel de la salud pública mundial es tratar de modificar los estilos de vida pasibles, mediante programas preventivos de manera de lograr una disminución importante en el número de muertes prematuras. Este tipo de problema de la salud pública tiene un impacto muy grande en el desarrollo macroeconómico de los países tal como consigna la OMS y el Foro Económico Mundial y donde se sostiene que en un escenario en que se mantengan estáticos los niveles de intervención, y las cifras de ENT continúen 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.
En el contexto de los estudios epidemiológicos donde se indaga por las ENT es práctica habitual trabajar con variables binarias que reflejan la presencia de determinadas enfermedades, las que a su vez se asocian con otro conjunto de enfermedades, denominadas comorbilidades, medidas también a través de variables binarias y que 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, donde ambos tipos de variables pueden ser intercambiables en cuanto a quien hace el rol de factor de riesgo.

En este capítulo el objetivo es obtener perfiles epidemiológicos bien diferenciados en base a los atributos binarios partiendo de un conjunto de variables, sin discriminar cuales son variables de respuesta, proponiendo la creación de grupos mediante 2 estrategias:

1. usar un método de clustering basado en el algoritmo *k-modes*;
2. a través del análisis de redes SNA, a partir de las mismas variables se construir 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 en que consiste el problema en estudio y los datos que se utilizan, para luego mostrar los resultados en la sección 8.5, discutir los hallazgos en 8.6, para terminar en la última sección 8.7, donde 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 mixto *Fuzzy C-modes*, y utilizada por (Álvarez-Vaz y Massa, 2012) para encontrar perfiles de infección parasitaria en escolares de Montevideo. En ambos trabajos cada individuo es previamente clasificado 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 un *k-prototipo* descrito por (Huang, 1997). En este caso el algoritmo tiene una lógica de funcionamiento similar a la del algoritmo *k-means*, y dada la naturaleza de las variables que son binarias, es necesario el uso de otras medidas de disimilaridad, usando un método basado en frecuencias para actualizar los modos, (Weihs *et al.*, 2005).

Por lo tanto del método mixto original antes planteado de (Tsekouras *et al.*, 2005) se trabaja solamente con el algorimo *k-modes* que aplica la siguiente disimilaridad, siendo x_i, y_i 2 individuos de los que se mide 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 los pasos:

1. selecciona k modos iniciales, uno para cada cluster con x, y variables categóricas binarias en este caso, actualizando el modo;
2. luego de que todos los individuos han sido asignados, reestima la disimilaridad de los objetos contra el actual modo 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;
3. 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 muy brevemente las diferentes métricas que se usan para la caracterización de las redes sociales, habitualmente usado en SNA. Para la presentación de las mismas 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 a seguir.

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 tá formada por 2 tipos de elementos: *nodos* y *enlaces*, donde los nodos pueden ser personas, variables o alguna otra entidad y 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, para um ejemplo simulado de 10 nodos, 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.

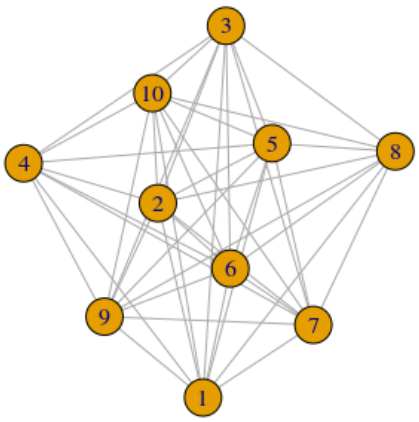

Figura 8.1: Ejemplo de red para 10 personas.

```
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, que en este caso son 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 sirve resumir la información, caracterizar la estructura de la red , a través de lo que se conoce como *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 simplemente 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 **(Betweenness centrality)**, (Freeman, 1979)

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

Donde $\sigma(s,t|v)$ 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), es dejar que la centralidad varíe inversamente con una medida de la distancia total de un vértice de todos los demás **(Closeness centrality)**

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

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

Finalmente, 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) y que es una extensión de la intermediación de nodos asignando 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 pueden consultar a (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 puede usarse triadas o componentes gigantes así como también 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 también (Kolaczyk y Csárdi, 2014) denomina *triangles*, de manera que al ir aumentado el tamaño de los cliques puede ver cual 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 al 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 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, la 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 refiere a la división de la misma en clases naturales tales que estas son disjuntas entre sí y a su vez la unión de ellas reproducen el conjunto de partida ($\bigcup_{k=1}^{K} C_k = S$). Pero a si vez es importante también evaluar si un subconjunto de nodos (algunas de esas clases) es "cohesivo" si para lo cual se entiende que es así si los nodos están bien conectados entre sí, y 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 al 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 $\mathcal{C}$, donde se 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 y que se miden con lo que se conoce como coeficiente de asortatividad (Assortativity coefficients) y que tiene una lógica muy similar a la de los coeficientes de correlación. Este concepto a veces también se conoce como *homofilia*, y expresa la tendencia de las personas a relacionarse con personas que se le parecen.

Cuando la característica que se estudia es de tipo categórico (nominal u ordinal) la medida 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 acotada 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-Udelar durante el período 2015-2016. En particular se consideran los siguientes atributos que conforman 3 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	Hipertesión	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).

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 *kmodes* (Weihs *et al.*, 2005), donde a través de la función *kmodes* 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

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	Hipertesión	**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.

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 este caso representa *Falta de dentición funcional* y por último la *Presencia de bolsa* que 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 consideró 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 150 minutos por semana de actividad física moderada
- Hipertensión arterial una medición de Presión arterial sistólica (PAS) mayor o igual a 140mmHg y/o Presión arterial diastólica (PAD) mayor o igual a 90 mmHg, o el autoreporte 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 autoreporte 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 en principio esta 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 o bien está enferma o tiene algunos de los factores de riesgo. En este caso de 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 restantes encuestados. Por último el no considerar esta observación no altera las prevalencias reportadas en la Tabla 8.2.

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.

En la Tabla 8.3 se puede ver como es el comportamiento cambiando de cantidad de clusters, para lo cual se definen 4 escenarios, en los que se presentan varias métricas.

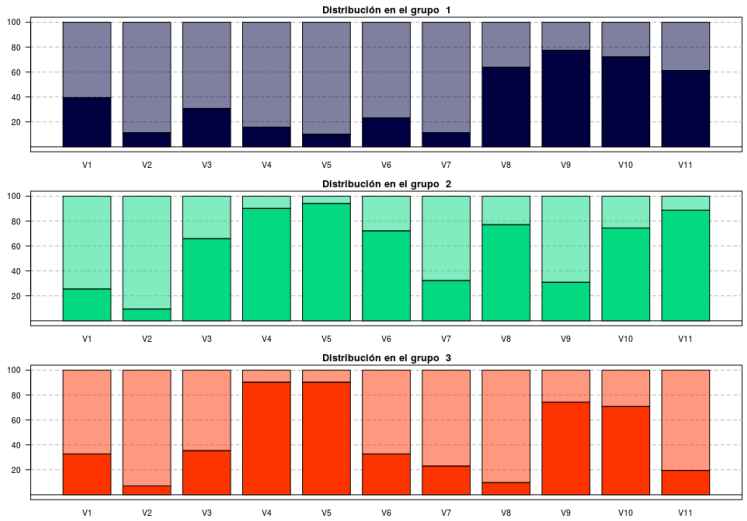

Figura 8.2: Prevalencias de las variables en cada grupo (3 grupos).

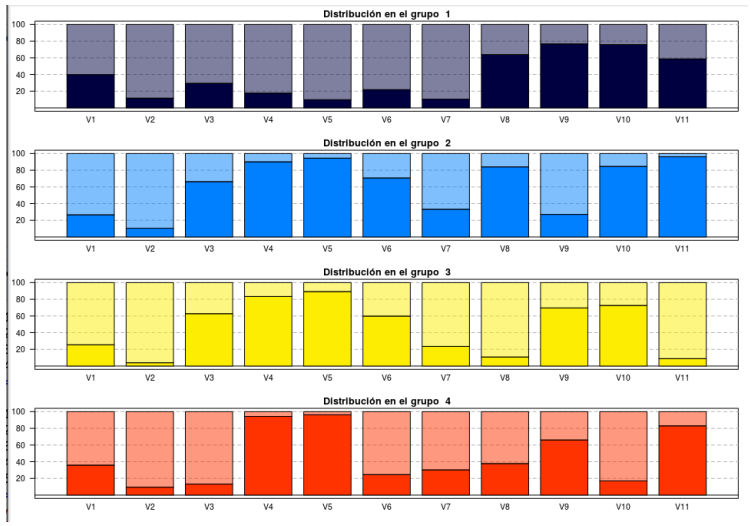

Figura 8.3: Prevalencias de las variables en cada grupo (4 grupos).

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 2 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 que puede verse como varía y decrece en algún cluster al cambiar la cantidad de grupos que se establece como restricción.

Analizando la densidad puede verse 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 como es la prevalencia de las variables a la interna de cada bloque para el escenario 2 y 3 respectivamente, siendo que la opción de 4 grupos la mejor, mostrando grupos que 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 cual es el perfil modal de cada grupo para las 4 soluciones planteadas en la Tabla 8.4

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.

En la Tabla 8.5 se presenta como es la prevalencia en cada grupo para los 3 bloques de variables y que se puede comparar con la prevalencia global de cada variable, para la opción de 3 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 intra-grupo (Cantidad de respuestas SI en las variables de los bloques 1, 2 y 3 de la Tabla 8.2)

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**
Hipertesión	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)	
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 kmodes 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 presenta los patrones de respuestas 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 y 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 asocia la matriz de adyacencias y sobre la que se crean las comunidades mediante el algoritmo de *Random Walk* y *Fast Greedy*. La matriz de adyacencias surge de considerar los nodos como las 601 personas del RPA-FO2015, y se establece que 2 nodos están conectados si comparten algunos de las 11 variables que se usaron la 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 usando para eso un *layout* que es aleatorio pero donde se usa la misma semilla para inicializar la visualización de manera de tener siempre en el mismo lugar los mismos nodos y poder hacer comparables las figuras.

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

Puede verse en la Tabla 8.8 que la configuración que más veces aparece 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, siendo un nodo tipo que tiene presente las 3 patologías del tipo ENT del bloque 2 y 2 de las 4 bucales. A este perfil de nodo se le antepone el nodo 97 (que es 1 de los 2 que aparece con esta configuración) y que está mucho menos conectado y se caracteriza por ser una persona que solamente presenta el hábito de fumar. El resto de las filas de la Tabla 8.8 muestras otras configuraciones que corresponden a nodos con mayor número de enlaces (mayor valor de la variable grados), pero que que aparecen menos veces, lo que se verifica que están asignados a diferentes clusters. Una característica que es común a los 6 nodos es que los individuos todos presentan las 4 patologías bucales.

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.

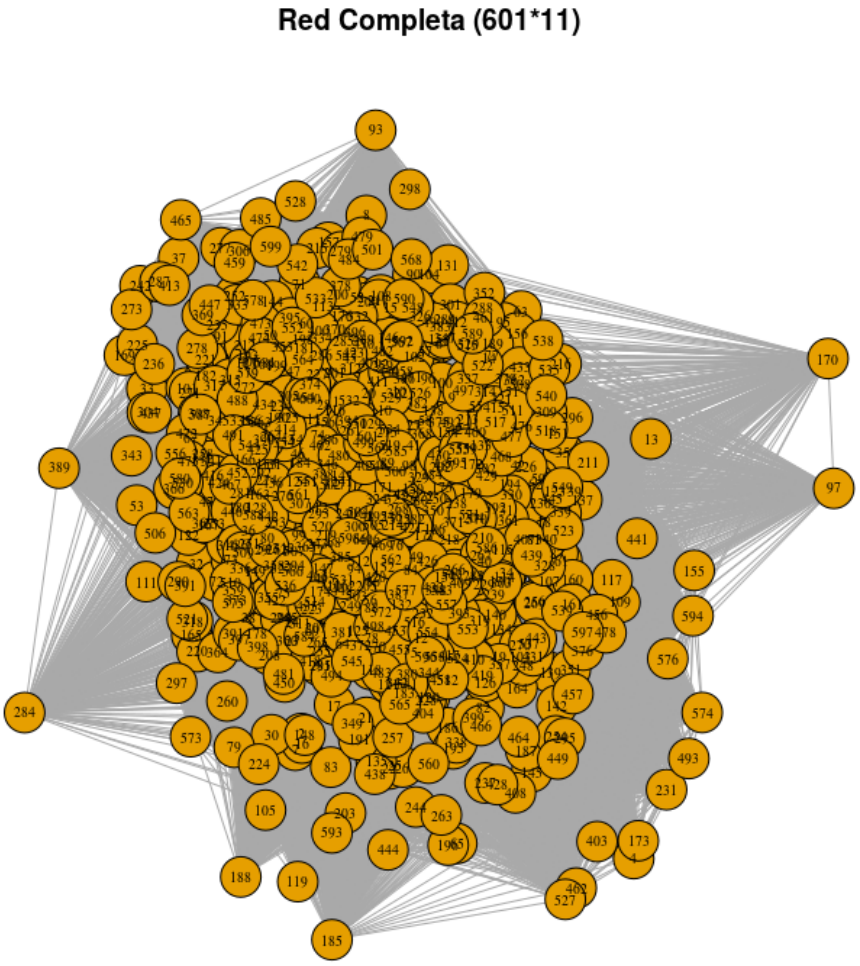

Figura 8.4: Red generada con 601 individuos analizados.

Sobre el grafo se aplican 2 algoritmos de detección de comunidades, donde la modularidad es mayor para la solución de 2 grupos que surge del algoritmo

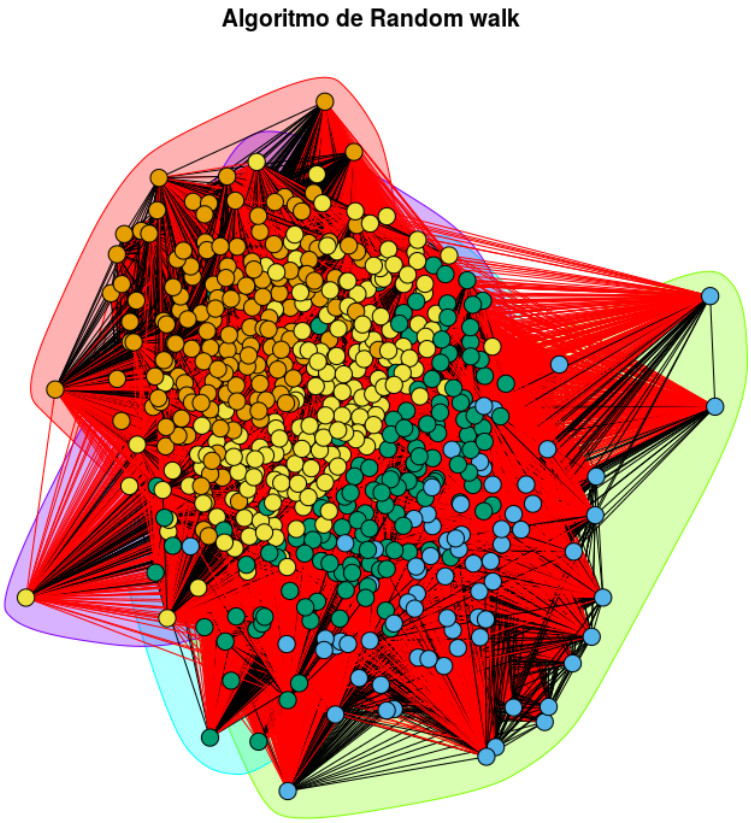

Figura 8.5: Comunidades identificadas con Algoritmo Random Walk.

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 Random Walk	Algoritmo (Cluster Fast Greedy) 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 2 métodos de detección de comunidades.

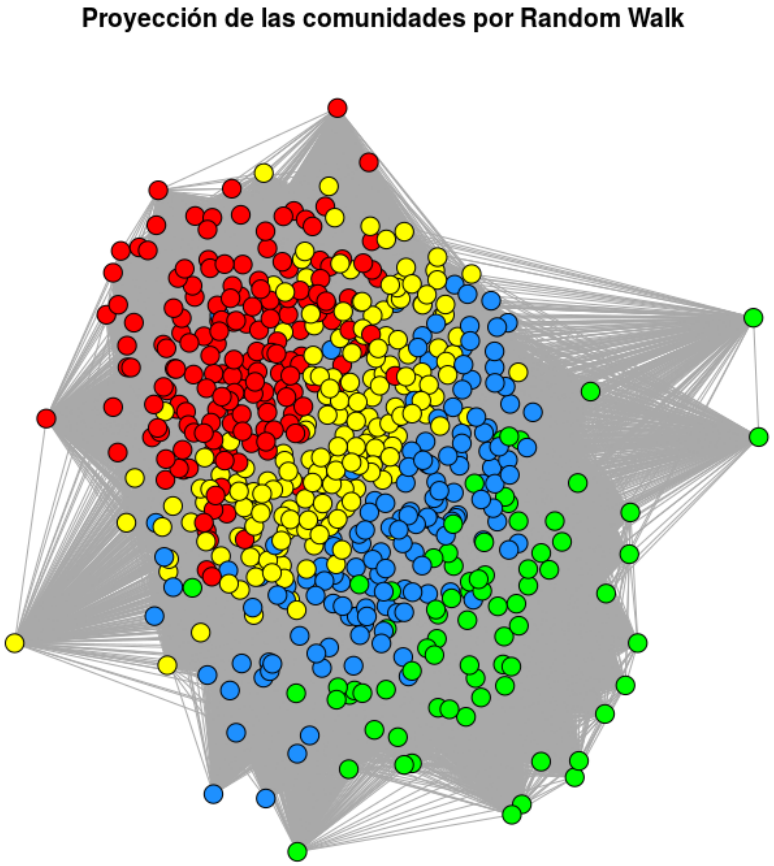

Figura 8.6: Proyección en la red de los grupos encontrados con algoritmo Random Walk.

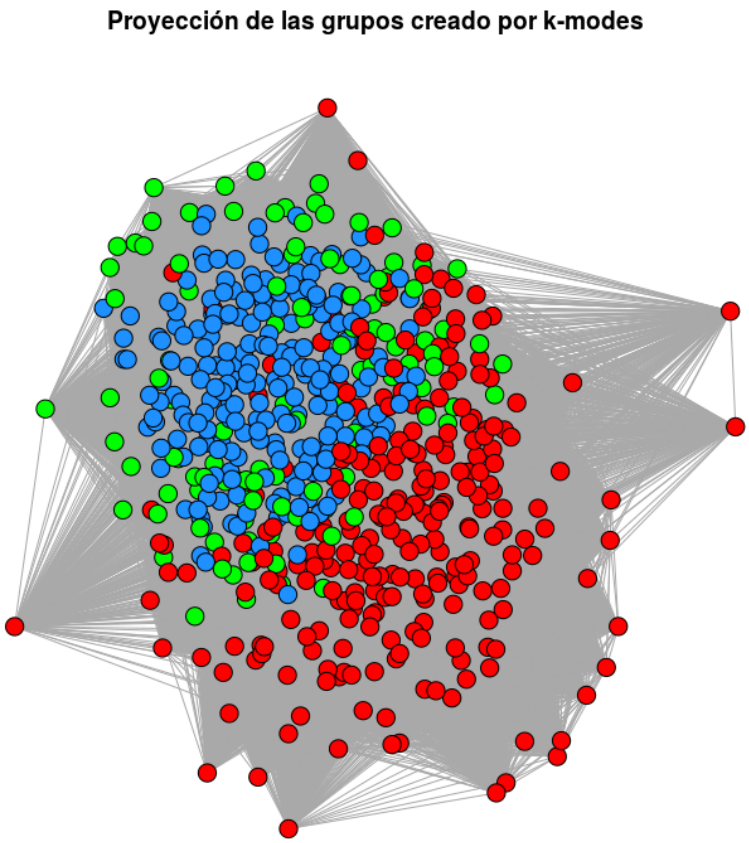

Figura 8.7: Proyección en la red de los grupos encontrados con algoritmo *k-modes*.

	(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 algoritmo Random Walk y algoritmo k-modes.

			(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.

8.6. Discusión

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

- G(1) con un total de (n=265) individuos que muestran 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 un grupo donde 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 con una baja prevalencia de pérdida dentaria (con prevalencia menor de caries y enf. periodontales), con a su vez prevalencias de las ENT muy aumentadas con respecto al promedio. En es un grupo donde la proporción de mujeres duplica la de hombres, con un edad promedio or debajo de la media global pero por encima de la del grupo 1. En cuanto al Ingreso, este muestra un perfil similar al promedio general;
- G(3)(n=223), donde los individuos muestran una alta proporción de conductas comportamentales negativas (que estaría dejándolo con un grupo con más perfil de riesgo) donde sin embargo las ENT estan fuertemente presentes, salvo la HTA y donde solamente 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 tomar en cuenta la participación o carga de variables con presencia de SI (que debe interpretarse como más factores de riesgo o más ENT o más patología Bucal). Un análisis de ese indicador muestra un trasiego de grupo con un aumento monótono, donde 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 con 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 puede comentar los siguientes hallazgos (con respecto al método de Fast greeedy):

- Grupo 1 con 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;
- Grupo 2 con más individuos (n=325), 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 es un grupo caracterizado por tener una alteración de las ENT muy por encima de los valores medios y con un comportamiento en cuanto a las variables odontológicas de menor prevalencia (están menos enfermos).

A su vez en las Figuras 8.5, 8.6 puede verse como es que quedaron las comunidades detectadas, claramente diferenciadas, donde 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 grupos creados con el método *k-modes* aparece el cluster 1 en rojo bastante diferenciado del cluster 2 en verde que aparece más solapado con el cluster 3, en azul.

Como todo problema de clustering, este no escapa a la situación donde no se sabe exactamente el número de clusters, 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* (homónimo del *k-means*), 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 estar más cerca de éstos que si estuviesen en otros clusters. Esta situación puede mejorarse extendiendo el método a un método mixto mediante un proceso de difusión donde cada individuo puede extenderse a otros clusters con diferentes grados de membresía, donde 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 acrecentada por la propia naturaleza de las variable binarias donde 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 con ambas metodologías, se puede decir ambas que detectan distintas cantidades de grupos, que si bien están diferenciados y tiene una explicación en el contexto del problema para el caso de la partición encontrada mediante *k-modes*, ésta es bastante difusa mientras que la partición creada mediante el SNA se muestra como más estable.

A su vez tal como se decía en la introducción el objetivo era 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 que existe en las variables al ponerlas a todas en los 3 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 bucales, las que puede ser explicadas por las variables comportamentales (bloque 1) las que configuran factores de riesgos y las variables ENT (bloque 2) que son de por si 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, los que se plantean en el capítulo 13, donde se resumen las principales conclusiones.

Capítulo 9

Evaluación de la salud bucal a través de la Teoría de la respuesta al Ítem en un estudio poblacional en Uruguay

9.1. Introducción

En los estudios epidemiológicos es práctica habitual trabajar con variables binarias que reflejan la presencia o ausencia de determinadas enfermedades, las que a su vez pueden estar asociadas con otro conjunto de variables, que en general se asumen como factores de riesgo de las primeras. El estudio de las patologías bucales más comunes (Caries, pérdida de inserción ósea, bolsa periodontal, dentición funcional) puede llevarse a cabo a través de distintos indicadores, siendo muy común el uso de variables binarias que representan la presencia/ausencia de cada patología, (Álvarez-Vaz y Massa, 2012). En el ámbito de las encuestas poblacionales, es práctica común que el análisis epidemiológico se encargue de indagar sobre los factores que propician la ocurrencia de dichas patologías valiéndose de MLG, (McCullagh, 1989). De esta manera es posible determinar cuales son los condicionantes de una determinada patología. Sin embargo, estos modelos relativamente sencillos y ampliamente usados no son capaces de considerar el análisis simultáneamente en varias patologías.

Por esta razón, se propone comparar los modelos de respuesta discreta con los que surgen de la Teoría de Respuesta al Ítem (TRI) (Baker, 2017) en el

ámbito epidemiológico debido a que:

- Los modelos TRI son capaces de analizar conjuntamente un grupo de variables de interés;
- son capaces de proporcionar una valoración de cada individuo.

9.1.1. Método de Regresión Logística

Cuando la variable de respuesta Y_i es una *variable aleatoria Bernoulli*, con resultados posibles: *éxito* , *fracaso* codificados 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

$$P(Y = 1|X) = \pi = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{9.1}$$

$$P(Y_i = 1|X_i) = \pi_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_p X_{pi}}}{1 + e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_p X_{pi}}} \tag{9.2}$$

9.1.2. Teoría de Respuesta al ítem

La forma que tiene el modelo TRI más sencillo es el que corresponde a una variable de tipo dicotómica θ_{ij} (manteniendo la notación que habitualmente se maneja en el ámbito de los TRI, (Baker, 2017)), donde se presentan en la Figura 9.1 un modelo teórico de Rasch y en la Figura 9.2 curvas para modelos teóricos de Rasch con diferentes parámetros de dificultad δ, donde Y_{ij} es la respuesta correcta para el ítem j del individuo i

$$\mathbb{P}(Y_{ij} = 1|\theta_i, \delta_j) = \begin{array}{l} \frac{e^{\theta_i-\delta_j}}{1+e^{\theta_i-\delta_j}} \; j = 1, 2, 3, 4 \\ i = 1, \ldots, n = 602 \end{array} \tag{9.3}$$

que se conoce como el modelo de Rasch o modelo de un sólo parámetro.

Sin embargo, los modelos TRI más usualmente utilizados proporcionan índices que describen el comportamiento de cada variable sin considerar el posible efecto de otras variables explicativas. Tratando de superar esta dificultad se propone modelar las variables de interés a través de un modelo de Rasch donde el comportamiento del parámetro individual esté determinado por una distribución normal cuya media sea modelada por un predictor lineal (ver la Figura 9.3).

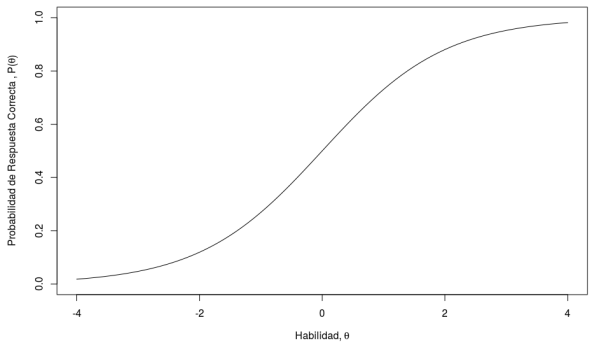

Figura 9.1: Modelo de un sólo parámetro o de Rasch.

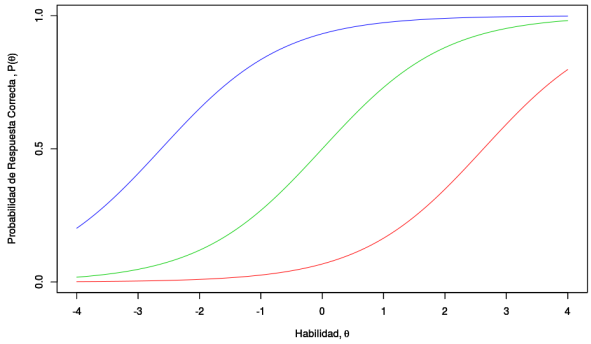

Figura 9.2: Comparación de Modelo de Rasch con diferentes parámetros de dificultad.

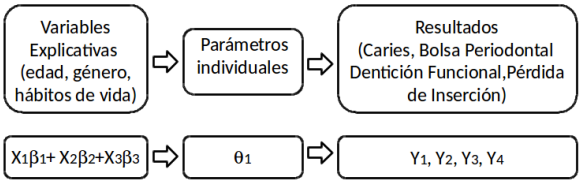

Figura 9.3: Diagrama de flujo del modelo considerado.

Como se mencionó en la sección anterior, dada la naturaleza binaria de las variables epidemiológicas consideradas, el modelo de Rasch surge como un punto de partida natural para el análisis. El modelo se generaliza al considerar un conjunto de predictores en la media del efecto aleatorio utilizado para modelar el comportamiento de cada individuo de un total de N, a los que se les evalúan k variables epidemiológicas de respuesta binaria.

$$\begin{array}{rcll} \mathbb{P}(Y_{ij}=1|\theta_i,\delta_j) & = & \frac{e^{\theta_i-\delta_j}}{1+e^{\theta_i-\delta_j}} & j=1,2,3,4 \\ \theta_i & \sim & N(X_i^T\beta,1) & i=1,\ldots,n=N \end{array} \tag{9.4}$$

donde en la Ecuación (9.4), Y_{ij} representa la ocurrencia de la enfermedad j en el participante i, θ_i es el parámetro individual (que en este contexto se puede interpretar como la propensión a la enfermedad (*sickness proneness*) de cada participante), δ_j es el parámetro de dificultad de cada variable (que aquí se relaciona con la prevalencia de cada patología), $X_i^T\beta$ es un predictor lineal que considera los efectos de las variables que habitualmente en otros contextos, donde se usan los modelos de variable de respuesta discreta, ofician de factores de riesgo, como por ejemplo en los modelos de regresión logística binaria, (Hilbe, 2017). La función de verosimilitud del modelo en este caso que considera 4 variables explicadas a través del TRI es la que aparece en la ecuación (9.5)

$$\begin{array}{c} \mathcal{L}(Y|X,\beta,\delta) \\ = \prod_{i=1}^{n}\int_{\mathbb{R}}\prod_{j=1}^{j=4}\mathbb{P}(\theta,\delta_j)^{Y_{ij}}(1-\mathbb{P}(\theta,\delta_j))^{(1-Y_{ij})}* \\ \frac{1}{\sqrt{2\pi}}e^{-\frac{(\theta-X_i\beta)^2}{2}}d\theta \end{array} \tag{9.5}$$

La maximización de dicha función de verosimilitud se hace numéricamente y a través de una aproximación de la matriz hessiana se obtiene una aproximación de la varianza de cada estimador. Todos los cálculos son llevados a cabo en el software R (R Core Team, 2017), para los cuales se programaron diferentes funciones (Rizopoulos, 2006).

9.2. Aplicación al estudio RPAFO2015

Se trabaja con los datos provenientes del estudio RPAFO2015 sobre personas que demandan atención en la Facultad de Odontología de la Universi-

dad de la República, Uruguay, ya presentados en la sección en otros capítulos de la tesis. A partir de la información que surge de este estudio se consideran para el análisis y el modelado los siguientes atributos que conforman 3 bloques de variables como se presentó en el capítulo 8. Parte del planteo teórico y algunos de los resultados preliminares están en el preprint en https://www.biorxiv.org/content/10.1101/611921v1

Variable	Descripción	Bloque
V1	Fuma a diario	1
V2	Consumo nocivo de alcohol	1
V3	Actividad física insuficiente	1
V4	IMC sobrepeso/obesidad,	2
V5	Razón de Cintura Cadera	2
V6	Hipertesión	2
V7	Diabetes	2
V8	Presencia bolsa	3
V9	Pérdida Dentaria	3
V10	Prevalencia de Caries	3
V11	Prevalencia de PIP	3

Tabla 9.1: Bloques de Variables utilizadas con modelo TRI.

Las primeras 3 variables constituyen factores de riesgo que están en el (Bloque 1), que son hábitos de vida o lo que se denomina variables conductuales, las que frecuentemente se asocian con las ENT que aparecen en el (Bloque 2), las que a su vez son patologías (enfermedades) y que también son factores de riesgo al mismo tiempo para las variables del bloque odontológico (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	Hipertesión	**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 9.2: Prevalencias de variables estudiadas.

Existen varias forma de trabajar con el bloque de variables explicativas (bloque 2) y (bloque 3), además de considerar, atributos como la edad y el sexo, por lo cual, para este la estrategia es trabajar con varios modelos, de los cuales se presentan los resultados al incluir sexo, actividad física y un spline para la edad (modelo m4), trabajada a escala continua.

En la Tabla 9.3 se puede observar que en este modelo, debido a que 3 de los parámetros de dificultad son negativos, esas patologías presentan prevalencias relativamente altas en la población estudiada. En cuanto al predictor lineal se refiere, no existen diferencias por sexo, mientras que si la actividad física en un factor relevante para explicar las prevalencias de las 4 variables bucales del bloque 3.

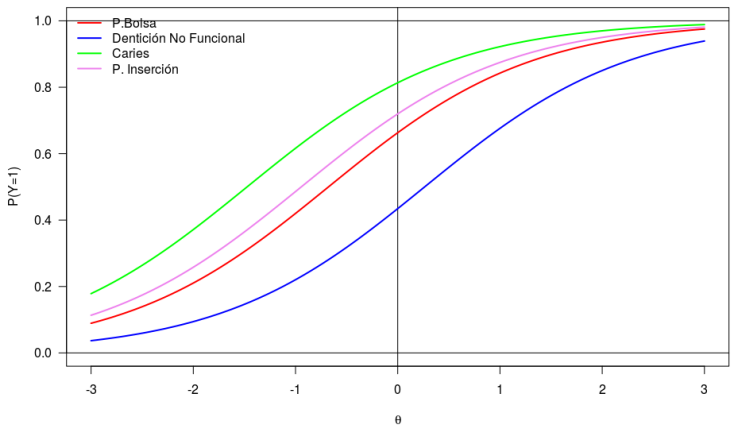

Figura 9.4: Curva característica del ítem para cada patología en modelo (m4).

	Estimación	Error Estandard	Z	p-valor
Parámetros de dificultad - modelo 4				
P. bolsa	-0.68	0.09	-7.51	≤ 0.001
P. dentaria	0.26	0.09	2.93	≤ 0.001
P. caries	-1.47	0.10	-15.21	≤ 0.001
P. inserción	-0.94	0.09	-10.28	≤ 0.001

Tabla 9.3: Coeficientes de los parámetros de propensión a la enfermedad.

En la Tabla 9.4 pueden verse los coeficientes para los 3 modelos estimados para los predictores lineales.

	Estimación	Error Estandard	Z	p-valor
Parámetros de regresión - modelo 4				
spl.edad0	1.01	0.11	9.57	≤ 0.001
spl.edad1	-0.46	0.10	-4.75	≤ 0.001
act.fis	0.25	0.10	2.41	0.02

Tabla 9.4: Coeficientes de los parámetros del predictor lineal.

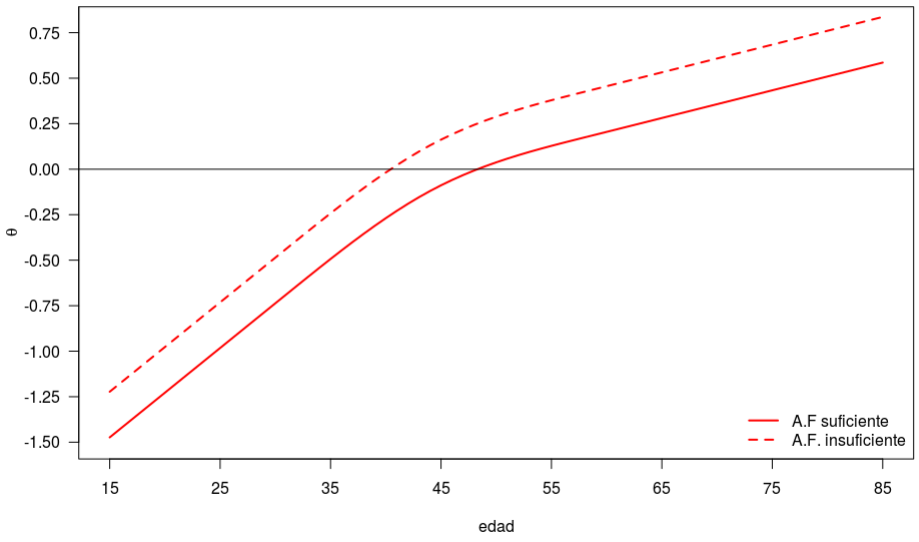

Figura 9.5: Efecto de la edad en la propensión a la enfermedad (modelo (m4).

Como ejemplo para el efecto de la edad, la Figura 9.5 muestra para el (modelo *m4*) el efecto no lineal de la misma , con un cambio en la velocidad de crecimiento de la *sickness proneness* a partir de los 45 años. Se puede observar que al aumentar la edad, la propensión a presentar alguna enfermedad bucal aumenta pero se desacelera. Sin embargo no se detectó un efecto significativo del hábito de fumar sobre la propensión a tener enfermedades bucales.

9.3. Conclusión

A través de los diferentes modelos propuestos se logró determinar el nivel de prevalencia de las patologías estudiadas en forma conjunta así como también el efecto de algunas covariables de interés. Se observa que el hábito de fumar no presenta un efecto significativo, así como tampoco el sexo, mientras que si la actividad física parece ser un atributo que en cierto manera modula la prevalencia de las enfermedades bucales aumentándolas. Sí se observó un efecto no lineal y creciente de la edad sobre la tendencia a padecer enfermedades bucales.

- Se propone comparar los resultados con los que surgirían de aplicar 4 regresiones logísticas por separado y el mismo juego de variables explicativas;
- Probar como quedan modelos de 2 o 3 parámetros para los TRI y tratar de ver como se interpretan en este contexto.

Capítulo 10

Visualización de la Estructura Multivariante de los Componentes del CPO a través del Análisis de Datos Composicionales.

10.1. Introducción

En los estudios epidemiológicos que analizan los componentes del CPO, es habitual hacerlo en forma separada, lo que lleva a una pérdida relevante de información que está contenida en la estructura multivariada de los datos. Como ya se manejó en capítulos anteriores 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 con 5 cariadas y 7 perdidas. En ambos casos, los niveles de enfermedad son importantes (tienen 12/28 % 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 enfermedad pasada. Resulta por lo tanto fundamental encontrar una forma de estudiar los 3 componentes del CPO (se deja de lado en principio el componente S), sin pérdida de información que refleje toda la esencia del CPO, por lo cual se propone una alternativa gráfica que se denominará *CPO-grama*, que permitirá explorar las relaciones 2 a 2 de cada componente del CPO pero también su valor global,

con la posibilidad de poder visualizar en forma simultánea cada componente, el CPO y otros atributos de cada persona analizada. Para eso es necesario previamente hacer una breve introducción a la metodología estadística que sustenta el método visual. Por lo tanto esta nueva forma de análisis es de tipo descriptivo a diferencia de otras técnicas que en capítulos anteriores intentaban modelizar los componentes del CPO.

10.2. Metodología de análisis de datos composicionales

Tradicionalmente, un conjunto de datos se llaman composicionales si éstos representan proporciones o partes de un total: porcentajes de trabajadores en diferentes sectores, porciones de los elementos químicos en un mineral, concentración de diferentes tipos de células en la sangre de un paciente, porciones de especies en un ecosistema o en una trampa, concentración de nutrientes en una bebida, porciones del tiempo de trabajo dedicado a diferentes tareas, porciones de tipos de fallas, porcentajes de votos para partidos políticos, etc.

El análisis sobre este tipo de datos, es un caso particular de lo que se denomina análisis de datos composicionales ADC. Las partes individuales de la composición se denominan componentes. Cada componente tiene una cantidad, representando su importancia dentro del conjunto. La suma sobre las cantidades de todos los componentes se llama la cantidad total. Las porciones son las cantidades individuales divididas por esta cantidad total. Es decir, las variables originales se transforman en porcentajes que tienen una suma constante de 100 % por individuo.

Mas formalmente desde el punto de vista estadístico un dato *composicional*, (Aitchison, 1982), (Aitchison, 1986), (van den Boogaart *et al.*, 2014) es un vector x cuyas componentes $(x_1, x_2, ..., x_D)$, estrictamente positivas, representan *partes* de un *todo*, por lo que x se encuentra sujeto a la siguiente restricción:

$$\sum_{i=1}^{D} x_i = k$$

- Al multiplicar una composición por una constante, la composición obtenida es la misma;

- Todos los vectores de D componentes positivas que son proporcionales; resultan equivalentes y representan la misma composición;
- Por lo general, se selecciona un *representante* de la composición.

- Se define un Operador *clausura* C - Correspondencia entre un vector $w = (w_1, w_2, ..., w_D)$ de componentes positivas y su dato *composicional* asociado;

$$x = (x_1, x_2, ..., x_D) \rightarrow C(w) = k(\frac{w_1}{\sum_{i=1}^{D} w_i}, \frac{w_2}{\sum_{i=1}^{D} w_i}, ..., \frac{w_D}{\sum_{i=1}^{D} w_i}) \tag{10.1}$$

- Las componentes del vector clausurado se denominan *partes*, sobre el *total* k, y definen el siguiente espacio (*simplex*):

$$S^D = \{(x_1, x_2, ..., x_D)/x_i > 0, i = 1, ..., D, \sum_{i=1}^{D} x_i = k\} \tag{10.2}$$

10.3. Visualización a través de GT

Los GT son un tipo de gráfico baricéntrico que permiten trabajar a la vez con 3 variables que tienen la característica de tener una suma constante por observación; son un caso particular (para 3 variables) de lo que ya se presentó.

En un GT que también se conoce como **ternary plot**, las proporciones de las tres variables a, b, y c deben sumar una constante, K. De esta manera hay solamente 2 variables que pueden fluctuar libremente debida a la restricción de que $a+b+c = K$ para todas las observaciones- sólo hay dos grados de libertad - es posible representar gráficamente la intersección de las tres variables en sólo dos dimensiones.

Si se analiza la posición que ocupa un punto cualquiera interior al triángulo equilátero que aparece en la figura 10.2, interesa ver cual es la relación entre las longitudes de los segmentos $\bar{HD}, \bar{GD}, \bar{FD}$; la misma relación prevalece entre las longitudes de los segmentos $\bar{AJ}, \bar{BI}, \bar{KB}$. Puede observarse por otra parte que las relaciones antes mencionados pueden verse en las proyecciones perpendiculares que de estas se hacen en el eje perpendicular a la base del triángulo y que está a la derecha y situado por fuera, en los que se representan los 3 segmentos $\bar{C_0C_1}, \bar{C_1C_2}, \bar{C_2C_3}$, que no son más que homotecias de los segmentos $\bar{AJ}, \bar{BI}, \bar{KB}$. Si los 3 lados del triángulo se usan para representar

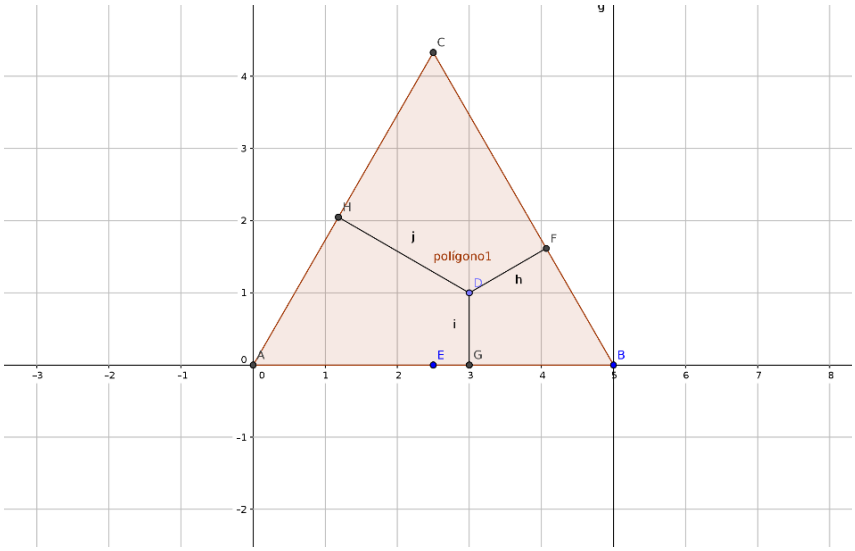

Figura 10.1: Ejemplo de gráfico triangular básico

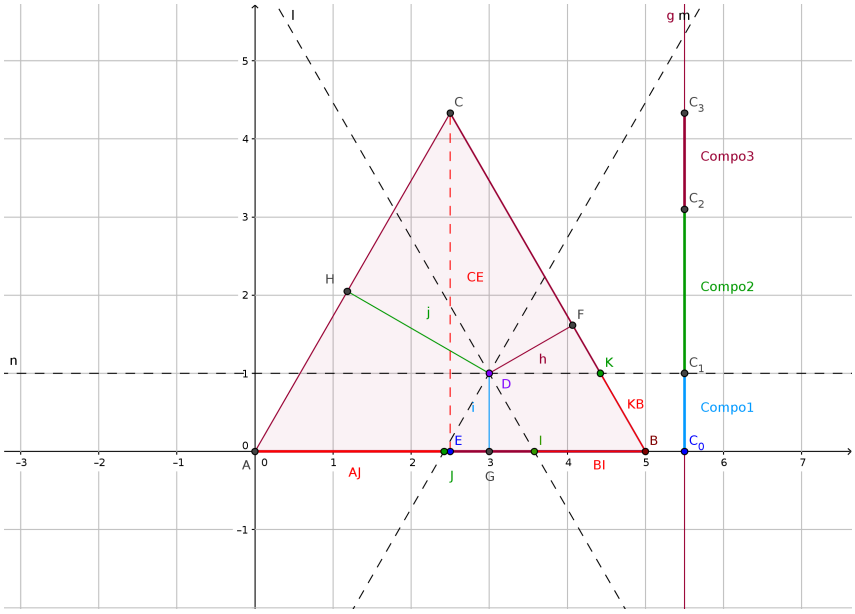

Figura 10.2: Ejemplo de gráfico triangular y sus componentes.

variables estadísticas, cuya suma tiene sentido (magnitudes económicas, físicas, etc) y es constante, para el caso de un triángulo cuyo lado tiene longitud 100, se está en presencia de lo que se llama GT. En este caso el punto interior al triángulo permite evaluar que % de cada variable estadística se tiene de la variable suma. Si por ejemplo se está analizando magnitudes continuas cada punto interior al triángulo indica que valor tiene en cada una de las 3 variables (son 3 coordenadas) las que a su vez representan que proporción de la variable suma tiene cada una de éstas.

Si por ejemplo se trata de caracterizar el comportamiento de los diferentes sectores de empleo (primario, secundario, terciario) por región o por país, se pueden usar tablas de contingencia de perfiles marginales fila, donde se está describiendo una variable con 3 categorías (en las columnas) vs una variable que podrían ser los países o grupos de ellos, con un GT puede tenerse una caracterización de los mismos, de acuerdo a su posición en el gráfico, con la ventaja de poder proyectar el baricentro de la tabla, que sería en este caso la distribución marginal de los sectores de empleo independiente de los países o grupos de estos. Se puede ver el ejemplo antes mencionado manejando la información del % de empleados del sector primario, secundario, terciario de 12 países europeos en el año 1978. Esas 3 columnas de la tabla de datos hacen un total de 100 % para cada pais. Por lo tanto puede ser muy interesante ver si el comportamiento de los 12 países es muy diferente entre sí, (datos extraídos de la librería ADE4 (Chessel *et al.*, 2004)).

Participación en cada sector del empleo			
Pais	primario	secundario	terciario
Bélgica	3.20	35.90	60.90
Dinamarca	7.90	31.90	60.20
España	20.60	37.20	42.20
Francia	9.20	36.80	54.00
Grecia	32.00	29.70	38.30
Irlanda	20.60	32.00	47.40
Italia	15.50	38.10	46.40
Luxemburgo	6.20	39.20	54.60
Holando	5.40	33.00	61.60
Portugal	31.30	34.80	33.90
Alemania	6.10	44.40	49.50
Reino Unido	2.80	39.00	58.20

Tabla 10.1: Participación en % de cada país en en los 3 sectores de empleo.

De acuerdo a la tabla anterior se puede observar el baricentro de la misma que es el Total (o lo que es lo mismo el perfil marginal es la última fila) que muestra la distribución de los sectores de empleo independientemente de los países.

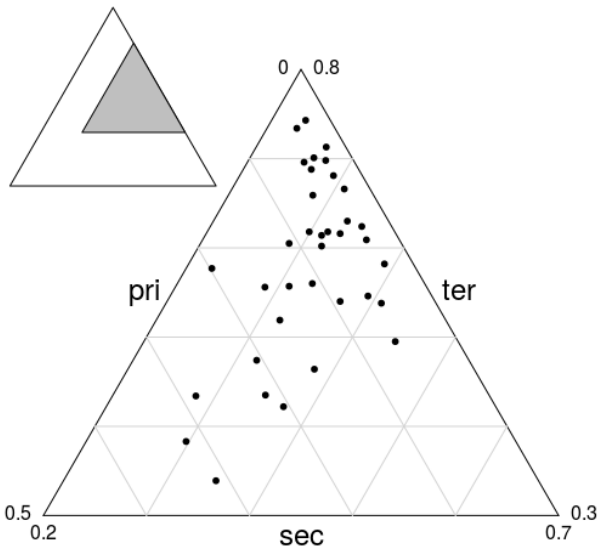

Figura 10.3: GT de Participación en % de cada país en en los 3 sectores de empleo.

Los GT tienen su origen en disciplinas como la geología, donde interesa ver que fraccción o proporción de un mineral componen un compuesto como puede ser la arena que tiene, u otros elementos. Tienen la ventaja de ser de fácil interpretación y no son muchas los paquetes estadísticos (Chessel *et al.*, 2004), (Meyer *et al.*, 2016) que tiene implementados este tipo de gráfico, salvo que el análisis se plantee desde un inicio desde la lógica del ADC pero supone un manejo de métodos cuantitativos casi desconocido en la biomedicina. La desventaja que puede tener este tipo de gráfico es que cuando la cantidad de observaciones es muy numerosa, puede ser difícil su interpretación, aunque si pueden mostrar esencialmente patrones de dispersión en los gráficos.

Por lo tanto una vez presentada la forma de construcción de un GT, a continuación en la seccción 10.4 se presenta una aplicación de los mismos como solución al problema presentado en la seccción 10.1.

10.4. Aplicación de ADC en estudio RPA-FO2015 para los componentes C, P, y O

Se presentan las distribuciones univariadas de los 3 componentes del CPO y el CPO global

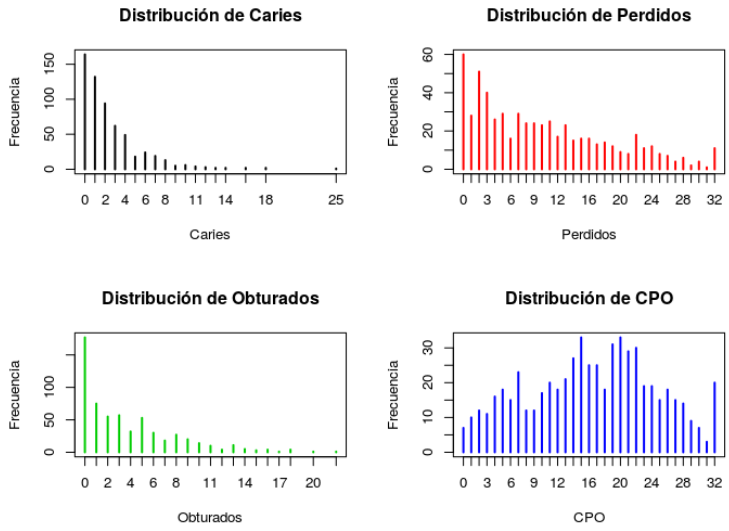

Figura 10.4: Distribución de cada componente del CPO en forma univariada.

Componente	n	$\bar{x}$	mediana	mínimo	máximo
Caries	602.00	2.50	2.00	0.00	25.00
Perdido	602.00	10.17	8.00	0.00	32.00
Obturado	602.00	3.66	2.00	0.00	22.00
CPO	602.00	16.33	17.00	0.00	32.00

En función de los valores para el mínimo del CPO y que se pueda operar como se presenta en (10.1), es necesario quitar las observaciones que corresponden a individuos totalmente sanos, es decir que tiene CPO=0. Quedan finalmente 595 encuestados, los que se analizan a continuación. Como la Figura 10.4, solo muestra la distribución univariada, sigue sin resolverse, el problema de no perder la estructura mutivariada, por ejemplo saber si la personas con bajo nivel de Caries, tienen bajo nivel de piezas perdidas o ambos atributos son independientes, por lo cual se da un paso más en la visualización y se obtiene la Figura 10.5.

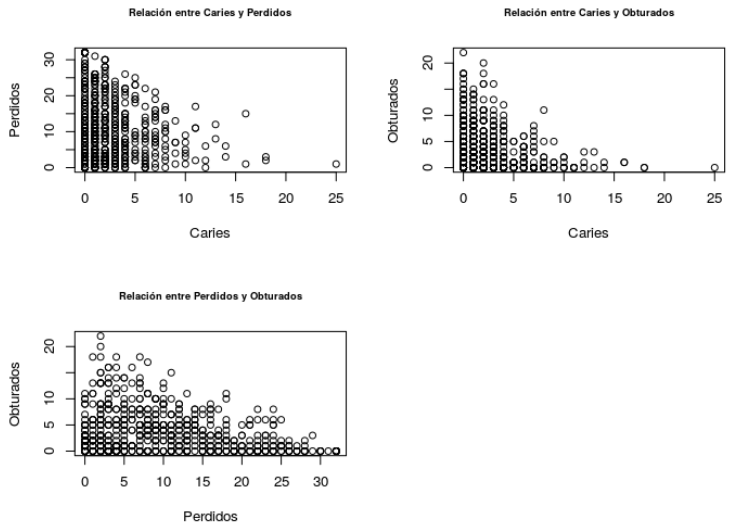

Figura 10.5: Relaciones 2 a 2 entre componentes del CPO por separado.

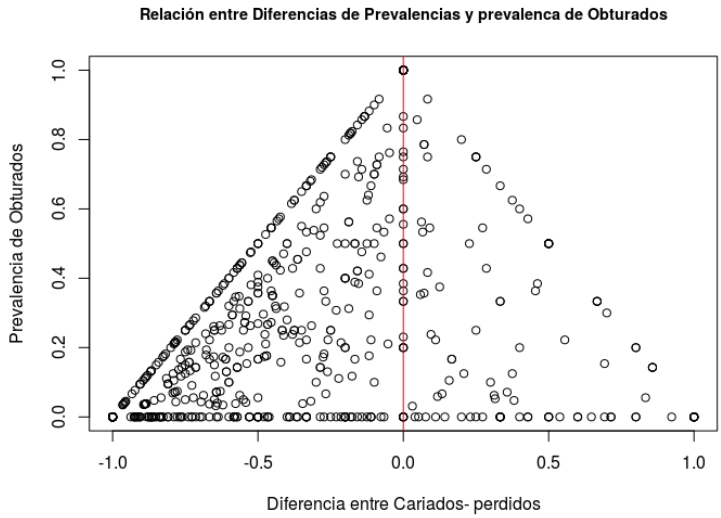

Figura 10.6: Relaciones entre diferencias entre prevalencias de los componentes del CPO.

La Figura 10.6 podría ayudar a ver la relación entre los 3 componentes, ya que el gráfico presentado es un gráfico de dispersión que debe interpretarse del siguiente modo: Al tener en el eje de las abcisas la diferencia entre prevalencias el eje de simetría pintado en rojo indica que los puntos que están por debajo de 0, son personas que tienen más proporción de dientes perdidos que de dientes cariados. Y luego de que sabe de que lado del eje están cuanto más elevado sea su coordenada en el eje vertical indica cuanta más proporción de dientes Obturados tiene, con un evidente descenso en la diferencia de prevalencias de perdidos y cariados. Es claro que hay un patrón de los puntos pero resulta de difícil interpretación. Incluso los puntos que caen sobre los lados del triángulo que se formaron son de más compleja interpretación, por lo cual se opta por trabajar con GT, definidos antes, los que por ser gráficos que revelan el patrón de asociación de los componentes del CPO, se denominarán CPO-grama, de ahora en adelante.

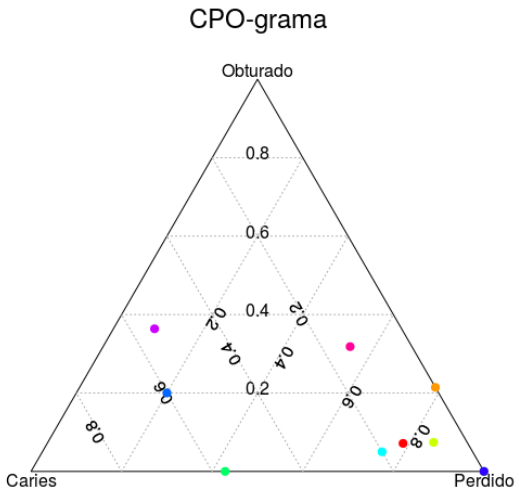

Figura 10.7: Ejemplo de CPO-grama para 10 individuos.

Se opta entonces por hacer un análisis gradualista tratando de entender lo que representan los CPO-gramas. La Figura 10.7 presenta como es la relación por ejemplo de 10 individuos para los 3 componentes y en función de la explicación dada en la Figura 10.2, para cada componente se traza una línea parela a la base opuesta al vértice del componente estudiado, por ejemplo Perdido, con lo cual de los 10 puntos hay una persona que tiene 60 % de Caries, 20 % de obturado y por construcción (las 3 proporciones deben sumar 100), tiene

20 %. Por otra parte si se observa el individuo que está en el vértice derecho abajo donde aparece Perdido en color azul oscuro, se puede aseverar que es una persona que tiene el 100 % de sus piezas cariadas. Cuando las observaciones se encuentran en algunos de los lados del tríangulo equilátero en el que se basa el CPO-grama, se puede decir que éstos tiene 0 % del componente opuesto al cateto donde se encuentran, tal es el caso de una observación en color naranja que tiene 0 % de Caries , 20 % de Obturados y 80 % de piezas perdidas, mientras que la observación en verde más intenso se caracteriza por no tener piezas obturadas repartiendo el CPO entre 40 % de piezas pedidas y el resto en piezas con Caries.

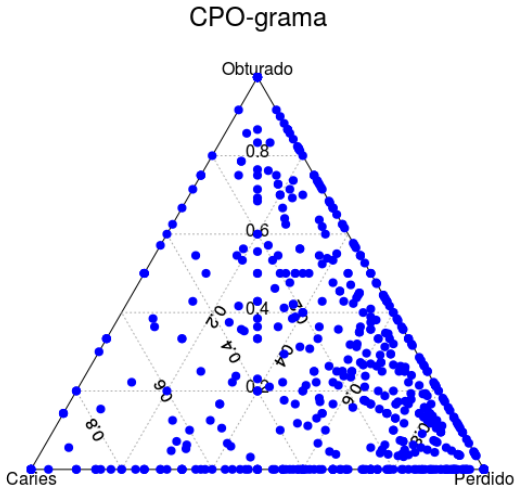

Figura 10.8: CPO-grama completo.

En las Figuras 10.9 y 10.10 aparecen los resultados a través de los CPO-gramas, que permiten encontrar patrones de comportamiento asociadas a otros atributos como son el sexo y por ejemplo el nivel de CPO, ya que una restricción que tienen los GT, que al ser representaciones gráficas de datos composicionales, mantienen la invarianza y pueden haber 2 CPO-grama iguales a pesar de que los niveles de CPO sea uno la mitad que el otro.

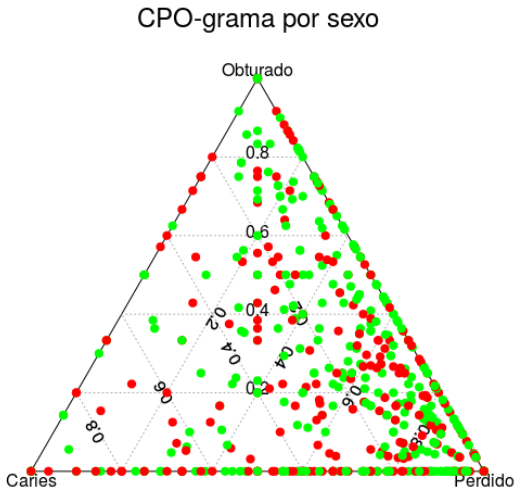

Figura 10.9: CPO-grama por sexo.

Para poder contrastar la asociación de C, P y O, se elabora una nueva variable que consiste en categorizar en CPO en 4 categorías que podrían asimilarse a cuartiles y que se muestran en la Tabla 10.2.

Sexo	[1,11]	(11,17]	(17,22]	(22,32]	Total
Masculino	75	66	50	54	245
Femenino	91	83	91	85	350
Total	166	149	141	139	595

Tabla 10.2: Distribución de personas por categoría de CPO según sexo.

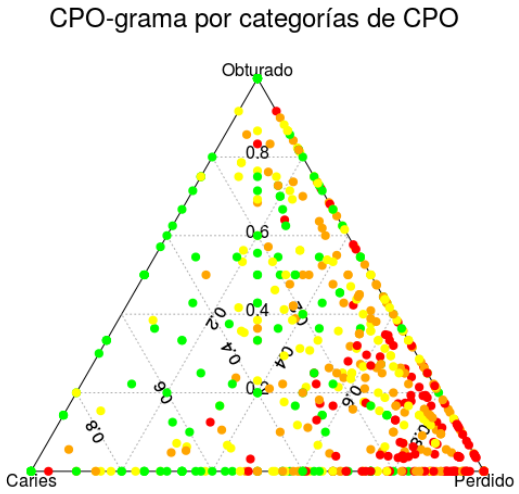

Figura 10.10: CPO-grama por nivel de CPO.

10.5. Discusión sobre el CPO-grama

Los resultados se presentan mediante una estrategia gradualista permitiendo al investigador en biomedicina ir captando cada vez con más detalle el fenómeno del CPO, partiendo de distribuciones univariadas, donde se nota un gran exceso de 0, aspecto ya conocido e identificado en los anteriores capítulos donde de trabajó con el estudio RPAFO2015, pero donde queda claro que las distribuciones univariadas son insuficientes para comprender exactamente el problema, del mismo modo que las distribuciones 2 a 2 donde siempre resta por ver que sucede con el tercer componente del CPO que no se manejó en ese gráfico bivariado. El esfuerzo a través del gráfico en forma de embudo invertido que aparece en la Figura 10.6 para el investigador que no conoce los GT, no alcanza aún a contestar un fenómeno que es trivariado representado en una imagen plana, algo que si se logra con el GT.

Si ahora se presta atención a la Figura 10.8, puede verse que existe un patrón muy marcado en las relaciones de los componentes del CPO, donde hay un predominio del porcentaje de piezas perdidas, con una zona muy cargada en el vértice derecho inferior; por otra parte si se examinan los lados del GT, se observa que la densidad es menor en el lado que corresponde al que une el componente Caries con Obturado, lo que significa que son pocas las personas del estudio que solo tiene piezas cariadas y Obturadas. Por último antes de pasar a un análisis más detallado que tome en cuenta otros atributos, puede decirse que la densidad en la parte central de la Figura 10.8 es baja, lo que debe interpretarse que son pocas las personas que reparten por igual Perdidos y Cariados y dientes Obturados.

Si ahora se evalúan atributos extras como el sexo la Figura 10.9 muestra que no hay un patrón claro de que algún componente concentre más individuos de un sexo que de otro.

Cuando ahora el contraste de la asociación de componentes del CPO se hace en función del nivel del mismo (aspecto hasta ahora no se ha tenido en cuenta), surge un claro patrón de que no solamente predomina el componente de diente perdidos, sino que además los puntos en colores naranja (CPO en el tramo $(17, 22]$) y rojo (CPO en el tramo $(22, 32]$), son los que allí aparecen dando una idea de la carga de enfermedad donde está concentrada. En oposición a este realidad, surge que los puntos verdes (CPO=$(17, 22]$) se dan en los lados del triángulo opuesto al vértice de Perdidos, indicando que son personas

con Caries o piezas Obturadas y bajo nivel de CPO, es decir muchas piezas sanas. Algunos otros puntos verdes se dan en general lejanos al componente de perdidos y en la zonas más baricéntricas del gráfico.

10.6. Conclusiones y futuros pasos

En este capítulo se presenta una alternativa gráfica para la visualización del CPO sin perder la esencia de la estructura multivariada del mismo. Se hace a través de un tipo de gráfico llamado GT, el que se denominará CPO-grama, y que no es más que un caso particular del ADC. En este nuevo tipo de herramienta visual (nueva para el trabajador del área biomédica pero ya muy usada en otras disciplinas), se logra resolver el objetivo, sin pedir información extra, y no perder la estructura multivariada del fenómeno. Se logran identificar patrones de comportamiento, que permiten decir que las personas del estudio RPAFO2015, se caracterizan por tener un fuerte predominio del componente de dientes perdidos, que se asocia con un elevado nivel de CPO. Es muy probable que este patrón pueda no estar asociado al sexo aunque si lo pueda estar con la edad o por ejemplo el ingreso de las personas. Esto se podría corroborar en lugar de contrastar por CPO (en términos de colores se haga por nivel de ingreso), pero nuevamente hay una restricción al quitar el atributo de nivel de CPO para sustituirlo por otro.

Queda entonces como desafío lograr en una imagen plana que maneja 3 variables (con la restricción de que tienen suma constante), lograr incorporar más información de la estructura multivariada. Ya el contraste por colores se logra y se propone tratar por ejemplo que una quinta variable como el ingreso, se pueda considerar al variar la forma del punto graficado; por otro lado si se quiere considerar además una variable de tipo cuantitativa se podría solapar al CPO-grama, curvas de nivel en función de la variable cuantitativa, para poner de manifiesto un patrón, si existiese. Resta entonces lograr que en este caso ese planteo se logre resolver mediante alguna subrutina de graficación más potente. Actualmente hay una serie de librerías del R que funcionan en combinación con la librería *ggplot*, (Wickham, 2009), que elaboran gráficos de alto nivel y que podrían tal vez resolver ese problema de combinar múliples atributos y curvas de nivel. Por último, considerando que se trata de datos composicionales, es decisión de las y los investigadores em biomedicomna decidir si trabajar desde la perspectiva de estadística multivariante clásica que prescinde del hecho de

que el ADC quita grados de libertad a la fluctuación de las variables y algo tan sencillo como una matriz de correlación en ADC puede mostrar resultados hasta contrapuestos a los que se obtienen en estadística multivariante clásica y como más rezón evaluar como sería el procedimiento de clustering, buscando crear grupos, sabiendo de las restricciones en los datos establecidas en la ecuación (10.1) y (10.2).

Capítulo 11

Medición y Caracterización de las Desigualdades en salud bucal para escolares de 12 años de Montevideo, Uruguay

11.1. Introducción

La reducción de las desigualdades socioeconómicas en salud bucal para el 2020 dentro los países es una meta de la OMS. Por lo tanto, el monitoreo de las desigualdades socioeconómicas en salud bucal es importante para evaluación de esta meta. Las desigualdades en salud son reconocidas universalmente como un importante problema (CSDH, 2008). Existe un gradiente de riesgo que a traviesa toda la población, siendo las personas más pobres y más desfavorecidas las que tienen más riesgo a enfermar y sufren peores condiciones de salud (Sanders, 2007).

Para (Graham, 2004a), (Graham, 2004b) las desigualdades en salud se refieren a las desigualdades socioeconómicas en salud . En su tipología, las desigualdades en salud distinguen: la mala condición de salud de las personas en situación socioeconómica desfavorable, las brechas en salud entre los diferentes grupos, y los gradientes sociales en toda la población. Es en este sentido que desigualdad / igualdad en salud se consideran como inequidad / equidad en salud , estas últimas siendo conceptos políticos que expresan compromiso moral con la justicia social, (Whitehead, 1992). Por otro lado, analizando el proceso

de transición epidemiológica en la población infantil de los países industrializados, se observa una dramática disminución de la Caries dental, el índice CPO-D a los 12 años se redujo de 6 a 1 en 25 años, (Petersen, 2009). Al mismo tiempo hay una redistribución de la carga de la enfermedad Caries, un gran número de casos están concentrados en un grupo pequeño de la población, este fenómeno es conocido como polarización, (Narvai *et al.*, 2006). Investigaciones recientes han demostrado que este fenómeno es producido por la existencia de brechas sociales claras y consistentes en la salud bucal en varios países, (Aida *et al.*, 2011), (Do *et al.*, 2010), (Elani *et al.*, 2012), (Tsakos *et al.*, 2011).

A pesar de los compromisos explícitos de luchar contra las desigualdades en salud consagrados internacionalmente, (World Health Organization, 2000),(World Health Organization, 2005), las desigualdades en salud y salud bucal muestran pocas señales de estrechamiento. Para avanzar en este aspecto, es necesario medir la magnitud de las desigualdades en salud para identificar los aspectos más desiguales dentro de una población, las áreas prioritarias de intervención, y además permitir la comparación entre poblaciones, (Mackenbach y Kunst, 1997).

En el contexto de la evaluación de las desigualdades en salud se propone contar con una metodología de análisis alternativa a la usada habitualmente en los estudios epidemiológicos en odontología, que permita crear indicadores fáciles de ser medidos e interpretados para la toma de decisiones clínicas y de gestión en servicios de salud bucal, ayudando a reorientar la toma de decisiones de políticas en salud púbica hacia un abordaje más equitativo.

Los objetivos que se buscan son desarrollar y sistematizar una metodología de análisis que permita caracterizar la distribución de diferentes indicadores habitualmente usados como el CPO, CPI entre otros y determinar las desigualdades que existen entre grupos poblacionales creados en base a variables geodemográficas (por ejemplo centros de estudio, centros de atención en salud, barrios, secciones censales, departamentos o grupos de éstos). Al trabajar con estas variables geodemográficas se crean nuevas unidades que surgen de agregar las originales y sobre las nuevas unidades agregadas es que se aplican y calculan los índices de desigualdad que se presentan en la sección 11.2.

Otra posibilidad es segmentar la población es estudio creando en forma artificial grupos mediante el análisis de cluster o conglomerados. Sobre esta nueva partición se agregan las variables bajo estudio y se aplican las medidas de desigualdad, pudiéndose entonces comparar ambas estrategias de agregación

y evaluar como se modifican las desigualdades.

A partir de una batería de medidas de desigualdad habitualmente usadas en el campo de la economía, como las medidas de entropía y divergencia estadística basadas en la teoría de la información, se presentan una serie de índices basados en rangos, índices de disparidad, índices de concentración.

Esta metodología de análisis alternativa a la usada habitualmente en los estudios epidemiológicos en odontología, permitiría crear indicadores fáciles de ser medidos e interpretados para la toma de decisiones clínicas y de gestión en servicios de salud bucal, ayudando a reorientar la toma de decisiones de políticas en salud púbica hacia un abordaje más equitativo.

11.2. Medidas de Desigualdad e Indices basados en Teoría de la Información

Todas los índices que en forma convencional se calculan para medir Caries, Enfermedad Peridontal y Erosión (CPO-D, CPO-S, CPI) tienen las características de que mediante alguna transformación pueden ser expresarse como tasas; en este caso sobre esta reexpresión de las variables bajo estudio se pueden aplicar, para poder discriminar el comportamiento en la población de los diferentes parámetros a evaluar en la salud bucal, una serie de índices capaces de mostrar desigualdad entre individuos o grupos de individuos agregados a nivel de las Unidades Geodemográficas (UG).

Existen diversas formas de presentar y ordenar los índices que dan cuenta de las desigualdades, no habiendo por lo tanto un consenso, por lo cual en este capítulo se sigue la lógica desarrollada por (Santiago Pérez *et al.*, 2010), donde éste presenta una compilación de los mismos, y donde previamente se detallan algunas propiedades que debieran tener éstos.

- Que reflejen la dimensión social de las desigualdades en salud : de otro modo no sería útil para medir desigualdades sociales en salud , sino sólo desigualdades sanitarias. El vínculo con la dimensión social no es siempre explícito: en ocasiones se consigue indirectamente al caracterizar los objetos de medición mediante variables sociales. Por ejemplo: hay un vínculo explícito cuando se miden desigualdades entre grupos dados por deciles de ingreso; pero hay un vínculo indirecto cuando los municipios

de un país se ordenan o clasifican según el índice de desarrollo humano o el porcentaje de población en estado de pobreza;

- Que utilicen toda la información disponible y no sólo la correspondiente a determinados grupos: algunos índices utilizan sólo algunos grupos de la jerarquía o clasificación social, por ejemplo, los grupos extremos, o algún grupo especial que se toma como referente. Estos índices ocultan espacios de desigualdad que pueden tener especial relevancia. En general, es preferible utilizar índices que abarquen a todos los grupos poblacionales definidos por el indicador social de interés;
- Que sean sensibles a cambios en la distribución de la población: algunos índices se calculan sólo a partir del indicador de salud , sin tomar en cuenta el tamaño de los grupos a los que se aplica. Si la composición de la población cambia, la relación entre la cantidad de personas expuestas y las no expuestas a determinadas condiciones desfavorables puede variar en un sentido positivo o negativo, y el índice debería ser sensible a la magnitud y a la dirección del cambio. Por ejemplo, el riesgo relativo de desnutrición crónica entre hijos de familias pobres con respecto a los hijos de familias no pobres, podría mantenerse invariante (o incluso aumentar), pese a que tanto la prevalencia de desnutrición crónica como los porcentajes de pobreza hayan disminuido. Hay índices que toman en cuenta la distribución de la población y otros que la ignoran. En general los primeros son preferibles, pero nuevamente en dicha preferencia intervienen juicios de valor;
- Que sean sencillos de calcular y fáciles de interpretar: puesto que normalmente las medidas de desigualdad se calculan para quienes deben tomar decisiones, el uso de medidas complejas y de difícil interpretación no es aconsejable, independientemente de sus propiedades analíticas. Algunas medidas que se usan con frecuencia en el campo de la economía son menos populares en el ámbito de la salud porque son difíciles de interpretar y de representar mediante los recursos gráficos y descriptivos más comunes.

11.2.1. Índices basados en rangos

En el artículo de (Wagstaff *et al.*, 1991), los autores hablan de índices basados en rangos refiriéndose a un grupo de indicadores que se basan en

evaluar la relación del indicador de salud entre grupos extremos de una jerarquía poblacional, la que previamente fue ordenada en función de un indicador socioeconómico o en función del mismo indicador de salud bajo análisis.

En general buena parte del trabajo empírico sobre desigualdades se ha efectuado en términos relativos, sin embargo no debe perderse de vista, que grandes diferencias relativas pueden coexistir con diferencias absolutas pequeñas, que sin embargo tienen un escaso significado en términos de impacto sobre la salud de la población, mientras que en otras situaciones algunas diferencias relativas pueden interpretarse como injustas o discriminatorias, a pesar de estar reflejando pequeñas diferencias absolutas.

Por lo tanto se pueden manejar 4 índices que hacen comparaciones 2 a 2, para categorías que están ordenadas

Concepto	Expresión
Índices absolutos	
Diferencia de Tasas Extremas (DT)	$\mathrm{DT} = \mathrm{T}_{(1)} - T_{(N)}$ (11.1)
Riesgo Atribuible Poblacional (RAP)	$\mathrm{RAP} = \mathrm{T}_{(Total)} - T_{(N)}$ (11.2)
Índices relativos	
Cociente de Tasas Extremas (CTE)	$\mathrm{CTE}=\mathrm{T}_{(1)}/T_{(N)}$ (11.3)
Riesgo Atribuible Poblacional Porcentual (RAPr)	$\mathrm{RAP}_r = \frac{T_{(Total)}-T_{(N)}}{T_{(Total)}}$ (11.4)

Tabla 11.1: Índices basados en rangos.

donde para las ecuaciones antes presentadas se define cada elemento que la compone

- N es el número de UG
- $T_{total} = \sum_{i=1}^{i=N} W_i T_i$
- $W_i = \frac{n_i}{n}$es el tamaño relativo de la i-ésima unidad, (i=1,...,N)
- n_i es la población de la i-ésima unidad, i=1,...,N,
- $n = \sum_{i=1}^{i=N} n_i$ es el tamaño total de la población,
- T_i es la variable de salud de la i-ésima unidad geodemográfica, (i=1,...,N)
- $T_{(i)}$ es la variable de salud de la unidad geodemográfica que ocupa la i-ésima posición, (i=1,...,N) después de haber sido ordenadas por la variable de salud o por la variable socioeconómica.

La ecuación (11.1) establece una brecha absoluta entre 2 unidades geodemográficas mientras que (11.3) expresa la brecha relativa entre unidades geogemográficas.

Por otra parte (11.2) expresa el riesgo atribuible poblacional. El RAP es un índice que en general se denomina como "efecto absoluto poblacional" y se define como la diferencia entre la tasa poblacional del indicador de salud y su valor en el grupo con la mejor condición socioeconómica. (Schneider *et al.*, 2002) habla del RAP como el índice que mide el exceso de eventos (eg. muertes o enfermedades) por cada mil sujetos que experimenta la población general con respecto al mejor de los grupos, lo que también podría o alternativamente entenderse como la reducción que debería experimentar la población para igualarse con el grupo que está en mejores condiciones socioeconómicas. Este mismo concepto puede reexpresarse a través del RAPr, que es una versión relativa y porcentual del RAP y que consiste en expresarlo como porcentaje, para poder entonces evaluar la reducción relativa que debe experimentar la población para igualarse con el grupo en mejores condiciones socioeconómicas.

11.2.2. Índices basados en medidas de concentración

La Curva de Lorenz (CL) es una forma gráfica de mostrar la distribución de la renta en una población. En ella se relacionan los porcentajes acumulados de población con porcentajes acumulados de la renta que esta población recibe. En el eje de las abcisas se representa la población 'ordenada' de forma que los percentiles de renta más baja quedan a la izquierda y los de renta más alta quedan a la derecha. El eje de ordenadas representa las rentas.

A partir de la curva de Lorenz se puede evaluar la concentración del ingreso, usando el área que se encuentra entre la curva y la diagonal. Esa superficie se llama área de concentración. Cuanto mayor sea esta área más concentrada estará la riqueza; cuanto más pequeña sea esta área, más equitativa será la distribución de la renta del país representado.

Para poder cuantificar la concentración se puede recurrrir al Índice de Gini (I de Gini), que es un índice de concentración de la riqueza calculado de la siguiente manera

$$I_G = \frac{\sum_{i=1}^{i=n-1} (F(x'_i) - T(x'_i))}{\sum_{i=1}^{i=n-1} F(x'_i)} \tag{11.5}$$

Cuando la renta esta repartida por igual, es decir que la concentración es mínima $F(x'_i) = T(x'_i)$, entonces tenemos que

$$I_G = \frac{0}{\sum_{i=1}^{i=n-1} F(x_i')} = 0$$

Cuando hay un solo individuo o grupo que concentra todo el ingreso queda

$$I_G = \frac{\sum_{i=1}^{i=n-1} F(x_i') - 0}{\sum_{i=1}^{i=n-1} F(x_i')} = 1$$

Por lo tanto para I_G su valor estará entre cero y uno.

Otra forma alternativa de calcular el I_G es en función de la CL

$$IG = 2 * ACL \tag{11.6}$$

Si la concentración es mínima la CL coincide con la diagonal y el área del numerador es O y por lo tanto $I_G = 0$.

Un aspecto relevante a tener en cuenta es que es práctica habitual en el ámbito económico que al medir las desigualdades sociales, en general se ordene de acuerdo al ingreso. Hay que tener en cuenta que si en lugar de ordenar la población según ingreso en forma ascendente se ordena según una variable de salud también en forma ascendente (del más enfermo al más salud able), se puede obtener una variante a la curva de Lorenz, que se llamará Curva de concentración (CC) pero que dependiendo de si el indicador es un indicador positivo (cobertura de inmunizaciones o acceso a los servicios de salud), se ubicará por debajo de la diagonal o línea de equidistribución, mientras que si es de naturaleza negativa (como la mortalidad o la morbilidad) la curva se ubicará por encima de la diagonal.

Se define un índice analítico similar al índice de Gini, que dependen como se localice la curva, tendrá un rango de variación $(-1; 1)$, donde valores próximos a 0 siguen siendo expresión de poca desigualdad.

Teniendo en cuenta estos 2 índices de concentración que logran evidenciar diferentes aspectos de la desigualdad, no debe perderse de vista que es necesario que éstos satisfagan atributos esenciales como que reflejen el componente socioeconómico de las desigualdades en salud , produciendo una medición de la desigualdad que pueda vincularse en algún sentido a un gradiente socioeconómico. A su vez usando toda la información contenida en toda la jerar-

quía social, bajo estudio, es decir no usar solamente grupos extremos o pares de grupos elegidos como referentes. Y por otra parte que sean sensibles a los cambios en la composición de la población en los grupos que integran la escala socioeconómica, así como en la distribución del indicador de salud , tomando en cuenta el tamaño de los grupos, ya que no es lo mismo por ejemplo, que el grupo con las peores condiciones de salud esté compuesto por el 60 %, que por el 50 % de la población, con independencia del valor promedio de las condiciones de salud que existan en él. Es por eso que es muy importante la curva de concentración y su respectivo índice, ya que logra cumplir con todos los atributos requeridos, mientras que el IG no cumple el de producir una medida vinculada al gradiente socioeconómico (Wagstaff *et al.*, 1991), (Bacallao *et al.*, 2002).

11.2.3. Índices basados en el concepto de disparidad

Los indicadores basados en rangos presentados en la sección 11.2.1 resultan los más fáciles de calcular y a su vez los más intuitivos y fáciles de interpretar, sin embargo son más limitados, ya que dejan de lado un aspecto clave como es la disparidad, por lo cual es necesario incorporar al índice la noción de variabilidad, que resulta básica en estadística, ya que ponen de manifiesto la desemejanza, desigualdad o diferencia entre grupos, teniendo en cuenta el término anglosajón *disparity* cuando se hace referencia a las desigualdades socioeconómicas o a las desigualdades sociales en salud . Por otra parte una limitación de los índices basados en rangos cuando hay más de 2 grupos ignoran la variabilidad debida a los grupos no extremos en el ordenamiento de las clases según condición socioeconómica, mientras que los de disparidad sí la toman en cuenta, (Santiago Pérez *et al.*, 2010).

Para eso se presentan en la tabla 11.2 2 tipos de Índices de disparidad o dispersión como son el Índice de Pearcy-Keppel (IPK) y su variante ponderada, (Keppel *et al.*, 2004), (Keppel *et al.*, 2005), por un lado y los de Índice de Varianza entre grupos (VEG) y su variante relativa, (Bacallao *et al.*, 2002),(Bacallao, 2013).

Concepto	Expresión
Índices de Pearcy-Keppel	
IPK	$\mathrm{I}_{PK} = \frac{\left(\frac{1}{N}\sum_{i=1}^{i=N}\left\|T_i - T_{ref}\right\|\right)}{T_{ref}}$ (11.7)
Índice de Pearcy-Keppel Ponderado (IPKp) ponderado	$\mathrm{I}^{*}_{PK} = \sum_{i=1}^{i=N} W_i \left\|T_i - T_{ref}\right\|$ (11.8)
Índices Varianzas entre grupos	
VEG	$\mathrm{VEG}=\sum_{i=1}^{i=N} W_i (T_i - T_{total})^2$ (11.9)
Índice de Varianza relativa entre grupos (VEGr)	$\mathrm{VEG}^{*} = \frac{\sum_{i=1}^{i=N} W_i (T_i - T_{total})^2}{T_{total}}$ (11.10)

Tabla 11.2: Índices basados en disparidad o dispersión.

- N es el número de UG
- T_i es la variable de salud de la i-ésima unidad geodemográfica, $(i = 1, ..., N)$
- T_{ref} es el valor de referencia de la variable de salud : mínimo $(\min T_i, i = 1, ..., N)$, máximo $(\max T_i, i = 1, ..., N)$ o valor definido por el usuario
- $T_{total} = \sum_{i=1}^{i=N} W_i T_i$
- $W_i = \frac{n_i}{n}$es el tamaño relativo de la i-ésima unidad, (i=1,...,N)
- n_i es la población de la i-ésima unidad, i=1,...,N,
- $n = \sum_{i=1}^{i=N} n_i$ es el tamaño total de la población,

11.2.4. Índices basados en Distribuciones de Probabilidad y medidas de entropía

Otra clase de índices para evaluar la posible desigualdad en los procesos de morbimortalidad, es la que surge de comparar distribuciones empíricas de probabilidad con respecto a un mismo dominio de clases sociales o unidades geodemográficas. Se manejan las dos distribuciones que corresponden a la población total y a la población de casos.

Clases	Población		Variable de salud	
	N	W	C	ε
1	N_1	$W_1 = \frac{N_1}{N}$	C_1	$\varepsilon_1 = \frac{C_1}{C}$
2	N_2	$W_2 = \frac{N_2}{N}$	C_2	$\varepsilon_2 = \frac{C_2}{C}$
...	...	...	...	
k	N_k	$W_k = \frac{N_k}{N}$	C_k	$\varepsilon_k = \frac{C_k}{C}$

Tabla 11.3: Tabla de comparación de distribuciones de probabilidad de población y variable de salud.

Concepto	Expresión
Índices	
Discrepancia de Kullback-Liebler (DKL)	$\mathrm{R}_{KL} = \frac{1}{2}\sum_{i=1}^{N} d_i ln(u_i)$ (11.11)
Índice de Hoover (I de Hoover)	$\mathrm{R}_{H} = \frac{1}{2}\sum_{i=1}^{N} \lvert d_i \rvert$ (11.12)
Índice de Theil (I de Theil)	$\mathrm{R}_{T} = \sum_{i=1}^{N} \varepsilon_i ln\left(\frac{\varepsilon_i}{p_i}\right)$ (11.13)

Tabla 11.4: Índices para comparar distribuciones de probabilidad.

- N es el número de UG,
- $d_i = \varepsilon_i - P_i$
- $\varepsilon_i = \frac{C_i}{C}$
- $C_i = n_i T_i$ numero de casos de la iesima unidad
- $C = \sum_{i=1}^{i=N} C_i$ numero de casos
- $P_i = \frac{n_i}{n}$ es la proporcion de la poblacion de la i-ésima unidad, (i=1,...,N)
- n_i es la población de la i-ésima unidad, i=1,...,N,
- $n = \sum_{i=1}^{i=N} n_i$ es el tamaño total de la población,
- T_i es la variable de salud de la i-ésima unidad geodemográfica, (i=1,...,N)
- $u_i = \frac{C_i}{n_i}$

El I de Hoover que tiene la expresión de (11.12) es un índice de disimilaridad que describe (Wagstaff *et al.*, 1991) el que se puede interpretar como la proporción de casos en la población necesarios que deben ser redistribuidos para alcanzar la total igualdad. Se puede representar gráficamente como la distancia vertical más larga entre la curva de Lorenz, o la parte acumulada de los ingresos totales celebrada por debajo de cierto percentil de ingresos, y la línea de igualdad perfecta. Es inmediata la utilidad de este índice, para la redistribución de recursos (eg. camas hospitalarias o inmunizaciones)

Una forma de medir la equidad en la distribución de la carga de enfermedad es a través de medidas de *entropía*, como por ejemplo el I de Theil , que puede ser visto como una forma de medir el grado de desorden o de uniformidad en una distribución, tal como se hacen con magnitudes económicas .
Si por ejemplo hay n personas que tiene cada uno $x_1, x_2, x_3, ..., x_n$ rentas (podrían ser sueldos) se tiene

$$\sum_{i=1}^{i=n} x_i = X \tag{11.14}$$

donde X es la renta acumulada o *total*. Se puede calcular el % de renta que le corresponde a cada uno como

$$p_i = \frac{x_i}{X} \tag{11.15}$$

De ese modo

$$\sum_{i=1}^{i=n} p_i = 1 = 100\,\%$$

En este caso la *entropía* de esa series de rentas se puede calcular como

$$H_N(x) = -\sum_{i=1}^{i=n} p_i log(p_i) = \sum_{i=1}^{i=n} p_i log(\frac{1}{p_i}) \tag{11.16}$$

Un análisis de este índice además de lo ya dicho permite ver que

- $H_n(x)$ es siempre positivo ya que $log(p_i) \leq 0$
- Cuando las rentas se reparten por igual se verifica $p_i = \frac{1}{n}$ con lo cual

$$\sum_{i=1}^{i=n} p_i log(\frac{1}{p_i}) = \sum_{i=1}^{i=n} \frac{1}{n} log(\frac{1}{1/n}) = log(n)$$

- Cuando hay máxima concentración, es decir que hay un $p_i = 1$ con lo cual los restantes p_j con $i \neq j$ se puede probar que

$$H_n(x) = \sum_{i=1}^{i=n} p_i log(\frac{1}{p_i}) = 0 + 0 + ... + 1.0.. + 0 = 0$$

- Teniendo en cuentas las 2 situaciones anteriores se puede decir que el rango de variación es

$$0 < H_n(x) \leq log(n)$$

A partir de este índice, se puede ver la entropía como una medida de la equidad en la distribución de magnitudes económicas. Del mismo modo puede usarse este concepto para evaluar la carga de enfermedad entre diferentes grupos. Sin embargo este indicador depende de la cantidad de observaciones n.

Se pueden construir 2 medidas de desigualdad pero de manera que cuando sea mínima valga 0 y sea máxima para $log(n)$

Este nuevo índice se llama Índice de redundancia (I de Redundancia) y es opuesto a la entropía. Para subsanar el problema de la dependencia del

valor de la entropía en función de n, se puede relativizar, con lo cual queda la *redundancia relativa* que es

$$T = log(n) - H_n(x) \Rightarrow T_r = \frac{(log(n) - H_n(x))}{log(n)} = 1 - \frac{H_n(x)}{log(n)},\ 0 \leq T_r\ < 1 \tag{11.17}$$

que permite comparar 2 situaciones con cantidades de rentas diferentes. Igualmente la redundancia relativa puede usarse para evaluar situaciones con diferentes cantidad de grupos donde medir desigualdad en la carga de enfermedad.

El I de Theil puede descomponerse en 2 fuentes, una que corresponde a lo que se puede llamar "Theil dentro de grupos" y por otra parte el componente que corresponde a "Theil entre grupos". Al igual que la varianza total que puede descomponerse en una varianza dentro de los grupos o (within) y otras que es la varianza entre grupos (between), el I de Theil tiene esta propiedad que permite evaluar qué magnitud de la desigualdad total se debe a los grupos y por lo tanto que parte de la desigualdad es residual o se debe a otros factores ajenos al factor de clasificación, siendo además que funciona a cualquier nivel de agregación. Es decir que si se está trabajando a nivel de barrrios que pertenecen a diferentes departamentos, la desigualdad total entre barrios se puede establecer como la suma de la desigualdad entre departamentos y dentro de los departamentos.

El último índice DKL se conoce como discrepancia de Kullback-Liebler y que se expresa como (11.11) a diferencia del I de Hoover tiene la propiedad de ponderar cada diferencia d_i, entre las dos distribuciones por el logaritmo de la tasa (en base 2) de la clase o unidad correspondiente. Esta hace que las clases con tasas más altas contribuyen más a la desigualdad, lo que se entiende como principio de 'aversión a la desigualdad', siendo además que es una disimilaridad no simétrica, (Wagstaff, 2002).

Cualquiera de los 3 índices de la Tabla 11.4 antes mencionados, a los efectos de ser comparables pueden estandarizarse a la escala $[0, 1]$, mediante $Z = 1 - exp(-R)$ y la transformación de equivalencia de entropía del índice Z, (Bacallao *et al.*, 2002)

$$p = \frac{1}{\pi}[arcsen((1 - Z)^{(0.06Z+0.6)})] \tag{11.18}$$

11.3. Aplicación de Medidas de Desigualdad en el estudio RACA2012

Los datos sobre los que se hace la aplicación de los diferentes índices presentados en la sección 11.2 provienen de la encuesta nacional de base poblacional en escolares "Relevamiento y análisis de Caries dental en adolescentes de 12 años de la ciudad de Montevideo", presentada en la sección 4.2.

En particular las medidas planteadas en la sección 11.2 se aplican sobre el componente de Caries con la pauta de que la presencia de cavidad en esmalte, se considera como presencia de Caries. Para poder trabajar se definen nuevas unidades de análisis que son la UG, que toman en cuenta 2 características: por un lado la clasificación de las escuelas y por otro lado la ubicación de las mismas en la geografía de Montevideo, dando lugar entonces a 12 posibles unidades geodemográficas, ya que hay 4 tipos de escuelas: (Escuela de Contexto Socioeconómico Crítico (ECSC), Escuela Pública Común Urbana (ECU), Escuela Pública de Tiempo Completo (ETC), Escuela Privada (EP)) y 3 zonas de Montevideo (1='Este',2='Centro',3='Oeste').

Para evaluar como funcionan las medidas de desigualdad para la prevalencia de Caries y dado que al no existir datos longitudinales en Uruguay en escolares con respecto a esta patología, se trabaja con los datos del estudio RACA2012, donde a partir de los datos que allí surgen, se toman los mismos como los del escenario basal, a partir del cual se compara con otros 3 posibles escenarios, con las siguientes características:

Escenario 1 - Corresponde a los valores de prevalencia para las unidades geodemográficas usadas en la 11.5, es decir escenario basal;

Escenario 2 - Se evalúa el impacto en la desigualdad luego de hacer una intervención donde se propone un descenso en 15 puntos porcentuales en el valor de la prevalencia en cada UG;

Escenario 3 - Se intenta ver el impacto en la desigualdad al hacer una intervención donde el descenso es el mismo en términos porcentuales en cada UG y es del 10 %;

Escenario 4 - Se evalúa el impacto al hacer una intervención diferencial, donde en las primeras 4 UG con menor prevalencia, el descenso es del 5 %, las siguientes 4 es del 10 %, mientras que las 3 últimas UG con mayor prevalencia se intenta hacer un descenso del 15 %.

Escolares				
	Zonas de Montevideo			
Tipo de Escuela	1 (Este)	2 (Centro)	3 (Oeste)	Total
1 (ECSC)	4347	817	2418	7583
2 (ECU)	4329	54809	752	10562
3 (ETC)	–	1158	1357	2515
4 (EP)	2874	1458	3220	7554
Total	11551	8916	7748	28216
Escuelas				
	Zonas de Montevideo			
Tipo de Escuela	1 (Este)	2 (Centro)	3 (Oeste)	Total
1 (ECSC)	5	2	4	11
2 (ECU)	6	8	4	18
3 (ETC)	–	2	1	3
4 (EP)	6	3	3	12
Total	17	15	12	44

Tabla 11.5: Total de escolares y de escuelas por tipos de unidades geodemográficas.

Para entender la codificación al construir las UG, debe tenerse en cuenta que el primer dígito refiere al tipo de escuela y el segundo a la zona de Montevideo. Una vez construida las unidades geodemográficas, que son 11, se ordenan las mismas de acuerdo a la variable socieconómica seleccionada, que en este caso es el Índice de Nivel Socioeconómico (INSE), en forma descendente. Hay que recordar que el INSE es un índice que permite ordenar a los hogares y sus individuos de acuerdo a las ciertas características de tipo socio económico, resultando en un score que está en $[0, 93]$, (en este caso se rescala a 100), de manera que un valor bajo indica menor nivel socioeconómico.

Si se trabaja agregando la prevalencia de Caries en cada UG, usando la media como representante en cada una (es decir la prevalencia) y teniendo en cuenta que se trata de una muestra con diseño complejo, donde cada escolar tiene un factor de expansión, se puede evaluar las medidas de desigualdad vistas en la sección anterior.

En la Tabla 11.7, se presentan los valores de prevalencia para los 4 escenarios definidos en el párrafo anterior. A su vez para poder comparar se presentan las 4 medidas basados en rangos seleccionadas en sección 11.2.1.

UG	Total de Escolares	INSE	% de C(OMS)	Total de escolares con Caries	W_i
43	3221	60,0	49, 8 %	1603	11, 41 %
42	1459	60,0	42, 9 %	626	5, 17 %
41	2874	53,8	57, 5 %	1653	10, 19 %
21	4329	46,7	58, 9 %	2549	15, 34 %
23	752	42,4	57, 5 %	433	2, 67 %
22	5481	42,4	65, 4 %	3584	19, 42 %
33	1357	39,0	64, 7 %	878	4, 81 %
13	2418	33,8	68, 2 %	1649	8, 57 %
11	4348	32,4	72, 6 %	3155	15, 41 %
12	818	32,4	73, 3 %	599	2, 90 %
32	1159	31,8	60, 2 %	698	4, 11 %
Total	28216	43.9	61, 8 %	17427	100 %

Índices	
Diferencias de tasas extremas	30.4 %
Cociente de tasas extremas	1.7
Riesgo atribuible Poblacional	18.9 %
Riesgo atribuible Poblacional relativo	30.6 %

Tabla 11.6: Índices en rangos para Prevalencia de C según unidades geodemográficas para Escenario 1.

		Escenarios			
Unidad Geodemográfica	Nro escolares	Escenario 1	Escenario 2	Escenario 3	Escenario 4
43	3221	42,9	27,9	38,6	36,5
42	1459	49,8	34,8	44,8	42,3
41	2874	57,5	42,5	51,7	48,9
21	4329	57,5	42,5	51,7	48,9
23	752	58,9	43,9	53,0	53,0
22	5481	60,2	45,2	54,1	54,2
33	1357	64,7	49,7	58,2	58,2
13	2418	65,4	50,4	58,8	58,9
11	4348	68,2	53,2	61,3	61,4
12	818	72,6	57,6	65,3	69,0
32	1159	73,3	58,3	65,9	69,6
Global	28216	59,7	44,8	53,8	52,9

Índices	Escenario 1	Escenario 2	Escenario 3	Escenario 4
Diferencias de tasas extremas	30.4 %	30.4 %	27.4 %	21.6
Cociente de tasas extremas	1.7	2.1	1.7	1.5
Riesgo atribuible Poblacional	16.8 %	16.8 %	15.2 %	13.3 %
Riesgo atribuible Poblacional relativo	28.2 %	37.7 %	28.2 %	24.7

Tabla 11.7: Índices en rangos para Prevalencia de C según unidades geodemográficas para los 4 Escenarios.

En la Tabla 11.8 se presentan las diferentes medidas vistas en la sección 11.2.3, para los valores de prevalencia de Caries correspondientes a los 4 escenarios y para los cuales a su vez se evalúan 3 umbrales de comparación que son el mínimo, el promedio y el máximo para cada escenario.

		Escenarios			
Unidad Geodemográfica	Nro escolares	Escenario 1	Escenario 2	Escenario 3	Escenario 4
43	3221	42,9	27,9	38,6	36,5
42	1459	49,8	34,8	44,8	42,3
41	2874	57,5	42,5	51,7	48,9
21	4329	57,5	42,5	51,7	48,9
23	752	58,9	43,9	53,0	53,0
22	5481	60,2	45,2	54,1	54,2
33	1357	64,7	49,7	58,2	58,2
13	2418	65,4	50,4	58,8	58,9
11	4348	68,2	53,2	61,3	61,4
12	818	72,6	57,6	65,3	69,0
32	1159	73,3	58,3	65,9	69,6
Global	28216	59,7	44,8	53,8	52,9
Referencia= Prevalencia mínima					
Índices		Escenario 1	Escenario 2	Escenario 3	Escenario 4
Tasa de referencia		42,9	27,9	38,6	40.8
Pearcy-Keppel		0,42	0,64	0,42	0,34
Pearcy-Keppel Ponderado		16,8	16,8	15,1	13,3
Varianza entre grupos		65,4	65,4	52,9	31,9
Índice de varianza relativa entre grupos		1,09	1,46	0,98	0,59
Referencia= Prevalencia Global					
Índices		Escenario 1	Escenario 2	Escenario 3	Escenario 4
Tasa de referencia		59,7	44,7	53,8	52,9
Pearcy-Keppel		0,36	0,15	0,11	0,15
Pearcy-Keppel Ponderado		15,4	6,0	5,4	6,9
Varianza entre grupos		65,4	65,4	52,9	74,6
Índice de varianza relativa entre grupos		1,1	1,5	0,98	1,41
Referencia= Prevalencia máxima					
Índices		Escenario 1	Escenario 2	Escenario 3	Escenario 4
Tasa de rteferemcia		73,3	58,3	65,9	69,6
Pearcy-Keppel		0,16	0,21	0,16	0,21
Pearcy-Keppel Ponderado		13,5	13,5	12,2	16,7
Varianza entre grupos		65,4	65,4	52,9	74,6
Índice de varianza relativa entre grupos		1,1	1,5	0,98	1,4

Tabla 11.8: Medidas de Disparidad para 4 Escenarios para prevalencia de C según unidades geodemográficas.

Con respecto a las medidas de concentración se plantean para los 4 escenarios la CL y la CC, que se presentan en las Figuras 11.1 y 11.2 respectivamente.

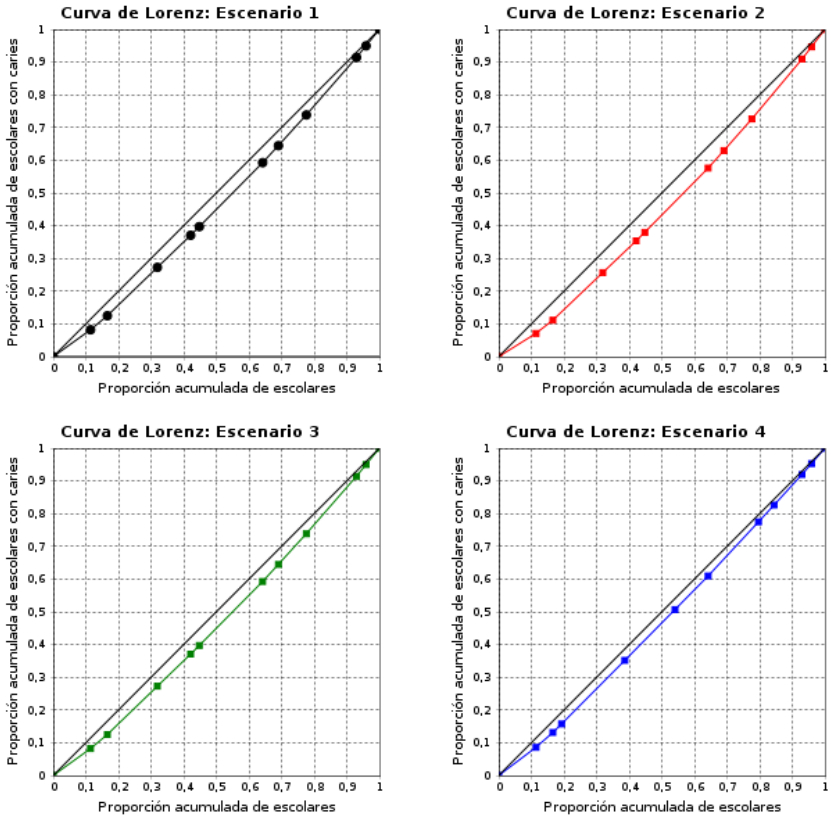

Figura 11.1: Curva de Lorenz para Prevalencia de Caries en las 11 UG.

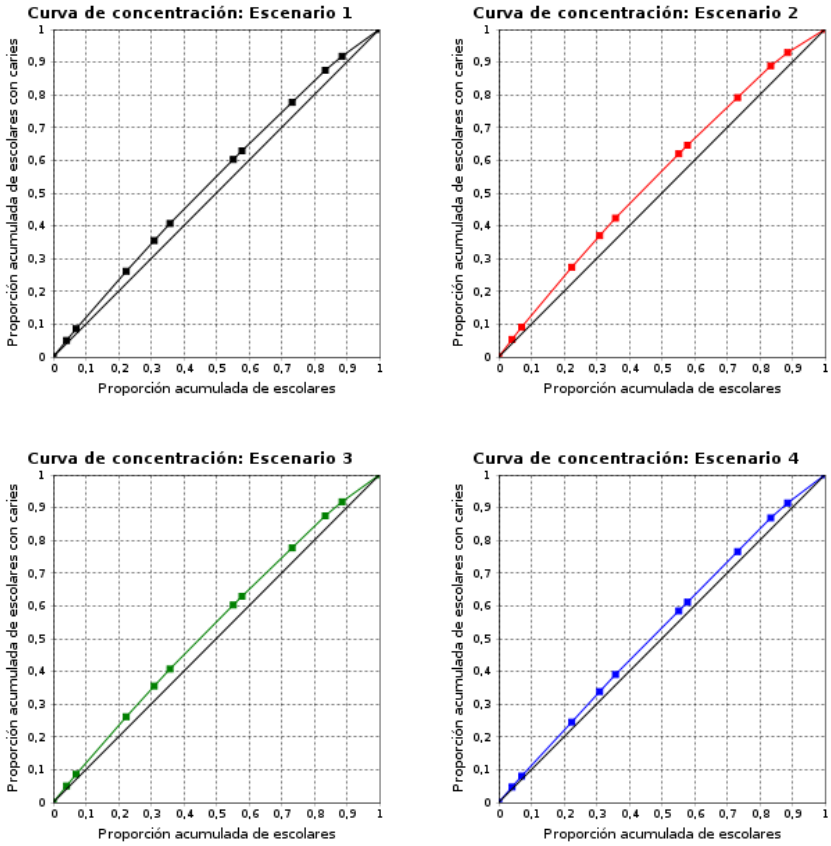

Figura 11.2: Curva de Concentración para Prevalencia de Caries en las 11 UG

En cuanto a los Índices basados en distribuciones de Probabilidad la Tabla 11.9 permite ver 3 de las medidas presentadas en la sección 11.2.4, con sus correspondientes transformaciones para una mejor interpretación de los resultados.

	Escenarios			
	Escenario 1	Escenario 2	Escenario 3	Escenario 4
Índices originales				
Índice de Kullback-Liebler	0,01	0,018	0,01	0,006
Índice de Theil	0,01	0,017	0,01	0,006
Índices estandarizados a escala $[0, 1]$				
Índice de Kullback-Lieber	0,01	0,018	0,01	0,006
Índice de Theil	0,010	0,017	0,010	0,006
Transformación de equivalencia de entropía				
Índice de Kullback-Lieber	0,465	0,453	0,465	0,473
Índice de Theil	0,466	0,454	0,466	0,473

Tabla 11.9: Índices de Entropía para las UG para prevalencia de Caries.

Por otra parte puede trabajarse a nivel individual, es decir de cada escolar y evaluar un nuevo concepto que se denominará Curva de Carga (CCC) de Caries, pero que podría extenderse para cualquier otra patología y que dará cuenta de la carga de la enfermedad a nivel individual, pero que luego al agregar los individuos de acuerdo a la escuela a la que pertenecen, se evidenciará si existe desigualdad de la carga de enfermedad entre UG. Para una mejor comprensión visual se presentan las CCC por zonas de las escuelas y por tipo de escuela.

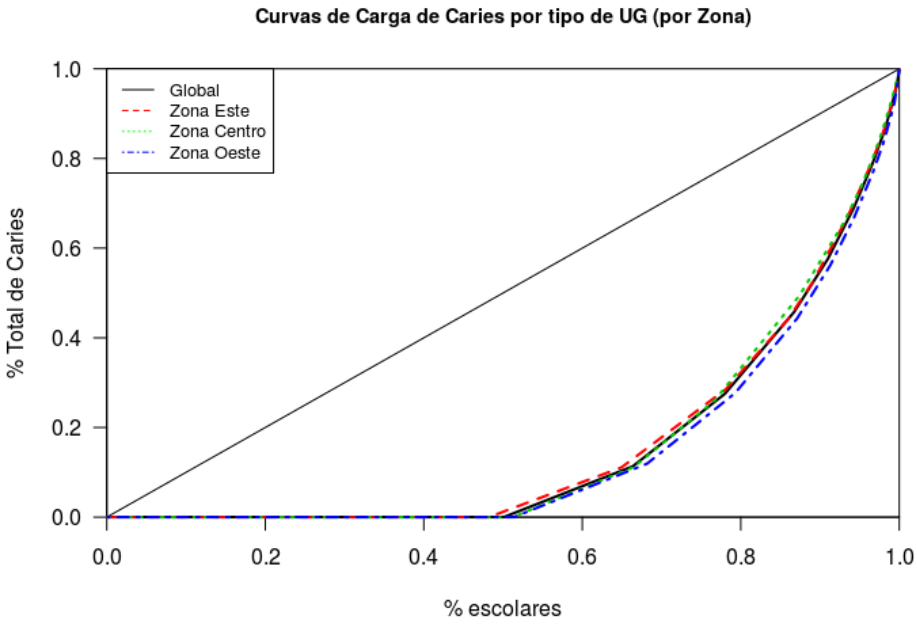

Figura 11.3: Curva de Carga de Caries por Zona de la ciudad.

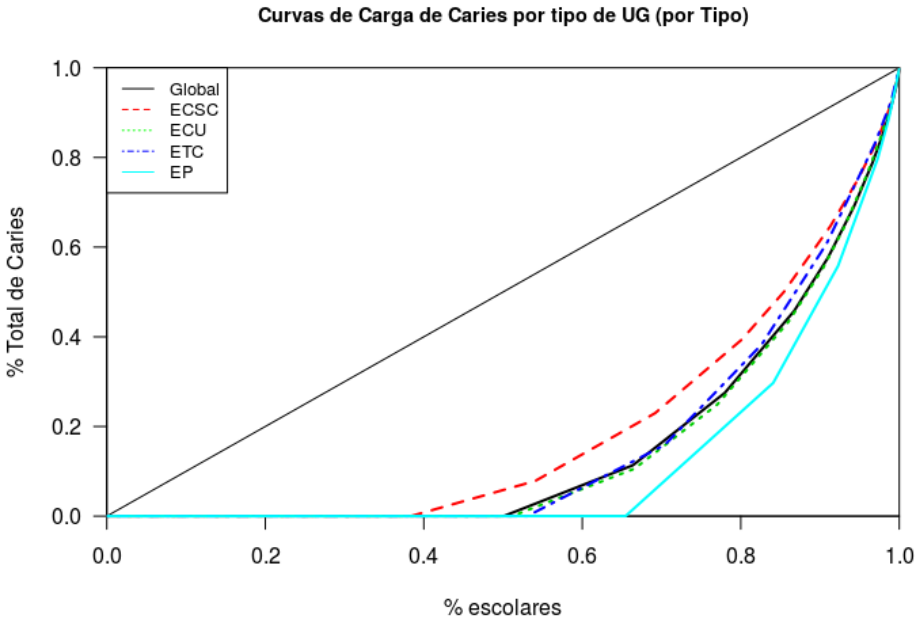

Figura 11.4: Curva de Carga de Caries por Tipo de escuela.

En la Tabla siguiente se muestran 5 medidas de desigualdad entre UG pero desde la perspectiva de la carga de enfermedad individual agregada a nivel de las UG.

U.G	Gini	Theil	Entropía	Atkinson
Zona Este	0.68	0.26	1.30	0.55
Centro	0.69	0.21	1.35	0.56
Oeste	0.71	0.29	1.40	0.58
ECSC	0.62	0.25	1.05	0.46
ECU	0.70	0.24	1.37	0.57
ETC	0.68	0.17	1.38	0.57
EP	0.76	0.16	1.74	0.68
Global	0.69	0.25	1.35	0.56

Tabla 11.10: Índices de Desigualdad para las UG aplicadas a nivel individual para el componente C en estudio RACA2012.

11.4. Discusión de las diferentes medidas de Desigualdad

Considerando los 4 escenarios y las diferentes medidas de desigualdad, se encuentran resultados muy interesantes y que representan un ejemplo como, en el ámbito de la biomedicina pero desde una perspectiva epidemiológica, es importante tener en cuenta el impacto que pueden provocar determinadas intervenciones hechas para provocar descenso en la patología pero que pueden alterar y ampliar las brechas existentes entre determinados grupos de individuos, en este caso escolares pertenecientes a la tipología construida a través de las UG. Por un lado al manejar las más sencillas de las medidas que son las basadas en rangos presentadas en la sección 11.2.1, una intervención con un descenso general absoluto de 15 puntos porcentuales pensado para el escenario 2 en la prevalencia, muestra que la brecha absoluta se mantiene, mientras que la brecha relativa se incrementa en 23 % al pasar de 1.7 a 2.1, manteniéndose constante el RAP al pasar del escenario basal al escenario 2, donde el valor indica que si le se lograse en Montevideo alcanzar las condiciones que existen en el grupo de escuelas con la menor prevalencia (UG=43), se evitarían en promedio 17 escolares menos con Caries por cada 100 escolares. Esta ganancia en términos del descenso de escolares con Caries representaría un descenso de

casi 38 %. Al considerar ahora el descenso de la prevalencia de Caries en un 10 % por igual para todas las UG, correspondiente al escenario 3, se verifica que hay un descenso del mismo nivel en la brecha absoluta mientras que la brecha relativa se mantiene en 1.7. El RAP muestra un descenso del también del 10 %, es decir que se lograría ganar bajando en casi 15 escolares menos la presencia de Caries en promedio en Montevideo con respecto a los escolares de la (UG=43), pero con un descenso promedio relativo de casi 33 %e, al pasar de un RAPr de 37.7 % a 28.2 %. Por último el escenario 4 con una intervención diferencial es el que parece adecuar mejor las métricas al bajar la brecha absoluta 21.6 %, lo que representa un descenso del 40 % con respecto al escenario1, la brecha relativa muestra el valor más bajo de los 4 escenarios con un descenso del 13 % con respecto al escenario 1; el RAP se achica mostrando que la ganancia en términos de escolares libres de Caries al igualar a Montevideo con la mejor UG baja a 13.3 %, que representa en términos relativos un 24.7 % de descenso con respecto a la prevalencia global. Otro aspecto que se da en este caso, es que si bien hay un aumento monótono de la prevalencia al descender el INSE (la variable socio-económica), solo una UG incumple la regla, donde el INSE que es el más bajo no se acompaña del valor más alto de la patología.

Si ahora el foco del análisis se da en el estudio de la disparidad entre UG, prescindiendo de la variable socio-económica, se encuentra que si la referencia de comparación es el mínimo, el IPK, que se puede interpretar como la desviación absoluta media con respecto al mínimo en términos porcentuales es del 40 %, valor que se incrementa al pasar al escenario 2, cuando se intentaba una reducción pareja de 15 puntos porcentuales para todos las UG, mostrando entonces un aumento de la disparidad. Sin embargo el escenario 2 muestra que el IPKp que si toma en cuenta el peso de cada UG y la VEG son invariantes a este tipo de reducción pareja para las 11 UG. Por lo tanto en resumen al compararse con el valor de referencia mínimo el escenario que parece más adecuado es el 4 ya que muestra un descenso de la disparidad medida por el IPK así como la varianza entre grupos, mostrando un disminución de la brecha siempre.

Si ahora se tiene en cuenta los resultados de las medidas de entropía presentadas en la Tabla 11.9, de los 2 índices para su mejor interpretación, interesa evaluar los que ya fueron transformados a a equivalencia de entropía. Para el escenario basal se puede decir que a nivel agregado (en las UG) la distribución de la variable Caries, corresponde a la de una población de 2 clases, en la

que el el 46.5 % de una de ellas soporta el 53.5 % de la carga de enfermedad, mientras que la otra clase el 53.5 % soporta una carga del 46.5 %. Usando este razonamiento para el resto de los escenarios, puede decirse que los cambios en las prevalencia de Caries en las UG para los diferentes escenarios, apenas impacta en la escala de equivalencia de entropía. En cualquiera de los 4 escenarios surge que a nivel agregado de las UG, no hay una disparidad de la carga de enfermedad.

Un índice que en este caso no se puede presentar es el I de Hoover, ya que una gran utilidad que brindaría este, es para la redistribución de recursos y no existe esa información. Se dispone de la prevalencia de Caries en cada UG, pero no la cantidad de odontólogos disponibles en esa zona geográfica. Si se conociese ese total se podría reasignar recursos de modo de poder lograr la mayor equidad, para combatir esa patología. Incluso en este caso se debería separar las UG de acuerdo al nivel sociocultural que se usa para clasificarlas, ya que es de esperar que el nivel socioeconómico de las escuelas privadas, fuese un diferencial en el acceso a la atención de las patologías. Si por otra parte en el contexto de un programa de atención en salud bucal para escolares, donde los derechos a la atención fuesen los mismos con independencia del contexto de la escuela, la reasignación de recursos en función del índice de I de Hoover, incluiría a todas.

Por último antes de pasar a la sección 11.5 de conclusiones, resta por analizar y discutir los resultados, considerando las desigualdades a nivel individual (para cada escolar), en lugar de a nivel de las UG, para luego si estratificar por éstas. Si se observan las CCC, en general muestran una concentración de la patología Caries muy importante a nivel global, lo que es un indicio de que si en lugar de considerar la prevalencia, se considera la extensión a través del conteo, como se hizo en varios capítulos de la tesis (5, 6), la situación cambia drásticamente, ya que ponen de manifiesto una situación epidemiológica que tiene 2 lecturas. Que exista concentración a nivel global o incluso estratificando por tipo de UG, muestra que hay pocos niños que llevan la mayor carga de Caries, es decir que hay pocos escolares con muchas Caries. Tal como se dijo antes esto tiene una lectura por un lado de concentración del fenómeno, que puede ser revertido actuando sobre pocos escolares, que puede verse en las Figuras 11.3 y 11.4. La otra lectura que puede hacerse, es que al trabajar a nivel individual aparece que las UG que contienen escuelas del sector privado se diferencian, de las otras, mostrando que el problema es a 2 niveles; por un

lado la patología está igualmente repartida entre todos los escolares de las escuelas de tipo ECSC, es decir es un problema más díficil de solucionar donde solamente el 40 %, mientras que para las escuelas de las UG de escuelas privadas la intervención es más sencilla, ya que son solamente el 40 % los escolares que presentan patología pero con una diferenciación de las escuelas de tipo ECSC, ya que solo el 20 % de los escolares de EP tiene el 80 % de a carga de enfermedad, contra un 40 % que aparece como carga para el 20 % más enfermo de los escolares de ECSC. Si se observan las CCC, cuando se estratifican solamente por zona geográfica, el fenómeno de extensión de Caries no presenta diferenciación.

Estos mismos análisis hechos en función de las CCC, pueden ser complementados con los índices presentados en la Tabla 11.10, donde observando la columna del I de Gini, aparece la brecha entre las UG de ECSC y de EP. Esa misma brecha se verifica para el I de Theil y se amplifica al considerar la entropía y el índice de Atkinson.

11.5. Conclusiones y futuros pasos

Habiendo hecho el análisis de la patología Caries en términos de desigualdad, usando diferentes herramientas, surgen resultados, muy importantes en cuanto al diagnóstico epidemiológico y la planificación que permite realizar a partir de éstos.

Para hacer la comparación se optó por crear UG, naturales combinando 2 atributos como son la ubicación geográfica y la tipología de escuelas, mostrando que en muchos casos las brechas para el escenario basal no son muy importantes, tanto que se considere el nivel socioeconómico, como que se prescinda del mismo. Eso puede ser un buen indicio de que tal vez en lugar de trabajar con UG naturales, se podría establecer una tipología de éstas, creando una clasificación de escuelas a través de algún método de clustering (ver sección 3.3.2), donde las nuevas UG, estarían formadas por escuelas mezcladas en la geografía de Montevideo y en cuanto a su pertenencia a la tipología de escuelas que hace ANEP. En esa nueva partición de las escuelas, estarían diferenciadas en términos de patología y nivel socioeconómico y tal vez mostrarían mayores brechas en términos de desigualdad.

Por último es importante destacar que el trabajar a nivel agregado, es decir con las UG, las brechas que parecían no existir aparecen, mostrando resultados

muy diferentes, que además son buenos trazadores para fijar las políticas de intervención. Sería muy importante como trabajo a futuro, hacer el ejercicio similar al que se presentó para los 4 escenarios (1 basal y los otros 3), trabajando a nivel individual, donde mediante Simulación Monte Carlo, evaluar como impactan las variaciones hechas para cada escenario, donde sería necesario establecer como funciona un descenso del 10 % a nivel de la UG, que es a nivel promedio pero que a nivel individual debería funcionar de forma diferente. Esta nueva perspectiva de análisis, implica en lugar de considerar un descenso de 10 % (si fuera el caso), una distribución de probabilidad, donde el 10 % sería la media, pero faltarían otros parámetros para caracterizar el comportamiento, como por ejemplo la varianza. También recordando que la prevalencia, puede ser modelada como una proporción a través de la distribución *Beta*, tal como se vió en 3.4.2, se podría usar para simular a nivel individual la patología en cada escenario.

Capítulo 12

Indice Canino Maxilar: Identificación del Sexo en odontología forense mediante Técnicas de Clasificación supervisada

12.1. Introducción

La identificación, una de las más trascendentes tareas del perito legista, estriba en un dinámico proceso técnico-científico tendiente a establecer la identidad de una persona, es decir, a individualizarla en la sociedad de la cual forma parte (Campos Neto y Paulete Vanrell, 2014). Es indudable que dicha tarea resulta sumamente engorrosa en presencia de restos humanos mutilados, esqueletizados, carbonizados o en avanzado estado de descomposición (Srivastava, 2010), requiriendo de la participación de un competente equipo interdisciplinario (Clark, 1994), para lograr la tan anhelada y necesaria reconstrucción del perfil biológico de la o las víctimas, por medio de sus cuatro principales componentes: edad, sexo, estatura y ancestralidad, (Prabhu y Acharya, 2009), (Pereira *et al.*, 2010). Los órganos dentales constituyen un excelente material para la ejecución de procedimientos comparativos y reconstructivos postmortem, en virtud de su extraordinaria dureza y resistencia a la acción de los más diversos agentes físicos, químicos y biológicos (Harvey, 1975), (Acharya y

Mainali, 2009), (Acharya *et al.*, 2011) y sus características anatómicas y volumétricas, que pueden expresar cierto grado de dimorfismo sexual (Harvey, 1975), (Rai y Anand, 2007), (Prabhu y Acharya, 2009), siendo pasibles de un pormenorizado análisis odontométrico (Rao *et al.*, 1989), (Rai y Anand, 2007), (Zorba *et al.*, 2011). La odontología, desde que se separó de la medicina, ha sufrido numerosos cambios, obteniendo tasas de éxito sucesivas con los tratamientos propuestos. Sin embargo, la odontología forense ha buscado muchos puntos (áreas) de concordancia con la medicina forense, desarrollando modelos estadísticos y software siempre con el objetivo de contribuir, mejorar y hacer más fiable el proceso de identificación, proceso técnico-científico tendiente a establecer la identidad de una persona, una de las más trascendentes tareas del perito legista, (Campos Neto y Paulete Vanrell, 2014). Siempre frente a desastres de grandes proporciones, hay un gran número de cadáveres, huesos y / o restos a ser identificados, si el antropólogo forense y / o dentista forense determinan el género, lograrán reducir y dividir la muestra en aproximadamente un 50 %, lo que facilitará el proceso de identificación, (Reverté Coma, 1999). Es incuestionable que la participación de un equipo interdisciplinario es imprescindible para lograr la reconstrucción del perfil biológico de la o las víctimas, basados en de sus cuatro principales componentes: edad, sexo, estatura y ancestralidad, (Prabhu y Acharya, 2009),(Pereira *et al.*, 2010). Para (Harvey, 1975), (Rai y Anand, 2007) la dentición se considera un complemento útil en la determinación del sexo de los esqueletos no identificados, principalmente porque los dientes son resistentes a la destrucción y la fragmentación postmortem, y por presentar características anatómicas y volumétricas, que pueden expresar cierto grado de dimorfismo sexual, que puede estudiarse por un análisis odontométrico. Los órganos dentales constituyen un excelente material para la ejecución de procedimientos comparativos y reconstructivos, en virtud de su extraordinaria dureza y resistencia a la acción de los más diversos agentes físicos, (Rao *et al.*, 1989), (Rai y Anand, 2007), (Zorba *et al.*, 2011). Varias investigaciones han tenido en cuenta el análisis odontométrico, la mayoría de ellas hacen referencia a los dientes caninos, es valido mencionar aquellas en las cuales se obtuvieron resultados significativos para el dimorfismo. Sherfudhin et al. (1996) analizaron, 301 estudiantes de secundaria (151 mujeres y 150 hombres) de la India; de 14 a 17 años de edad; calcularon el índice maxilar canino y obteniendo un índice de acierto de 88 % en hombres y 86.8 % en mujeres y global de 87.4 %. Kalia (2006) hizo lo propio con 504 individuos (252

hombres y 252 mujeres) de la India; de 15 a 21 años de edad; índice de acierto de 77.38 % en hombres y 74.21 % en mujeres y global de 75.79 %. (Parekh *et al.*, 2012) trabajaron con 368 estudiantes (216 hombres y 152 mujeres) de un colegio médico del oeste de la India; de 18 a 24 años de edad; concluyen que el ICMax mostró diferencias significativas entre hombres y mujeres. (Zirahei *et al.*, 2013), llevaron a cabo su investigación en 231 estudiantes nigerianos (127 hombres y 104 mujeres); de 18 a 24 años de edad; arribaron a valores del ICMax que presentaron diferencias significativas entre hombres y mujeres ($p - valor < 0.0001$). Otros estudios cuantitativos que realizaron medidas odontométricas involucrando a caninos superiores, no obtuvieron valores significativos para el dimorfismo entre ellos destacan (Sharma y Gorea, 2010): 117 voluntarios del noroeste de la India; de 17 a 50 años de edad;(Eboh y Etetafia, 2010): 101 sujetos (51 mujeres y 50 hombres) nigerianos (población del delta del Níger, región sureste de Nigeria); de 17 a 25 años de edad; (Picapedra *et al.*, 2012): 118 pacientes (59 hombres y 59 mujeres) uruguayos; de 21 a 60 años de edad; índice de a acierto de 52.5 % para mujeres y 45.7 % para hombres y global de 49 %; valor padrón 0.231, (Nahidh *et al.*, 2013): 200 sujetos (100 hombres y 100 mujeres) iraquíes; de 17 a 23 años de edad; índice de acierto de 74 % para mujeres y 44 % para hombres y global de 59 %; ,(Bakkannavar *et al.*, 2014) : 500 estudiantes universitarios (250 hombres y 250 mujeres) del sur de la India; de 15 a 25 años de edad; índice de acierto global de 48.6 %.

Teniendo en cuenta lo antes expuesto, el presente capítulo procura comprobar si los índices caninos mandibulares y maxilares estándares, constituyen instrumentos confiables y válidos para la determinación del sexo, en un estudio uruguayo en base a modelos de yeso usando también otros algoritmos que se basan en técnicas estadísticas multivariantes como la regresión logística, el análisis discriminante y los métodos CART, que se presentan en las secciones 12.2.2, 12.2.3 y 12.3.4 y que forman parte de un conjunto mas grande de técnicas estadísticas que se conocen como *Clasificación Supervisada*.

12.2. Técnicas de Clasificación propuestas

En esta sección se presentan las 4 técnicas que luego se compararán en la sección de aplicación, siendo la primera la que surge del artículo de (Rao *et al.*, 1989) y las otras tres diferentes técnicas estadísticas de clasificación supervisada.

12.2.1. Modelo de discriminación de Rao

Las indicadores necesarios a ser relevados para el Índice de Rao aparecen en las figuras 12.1 y 12.2

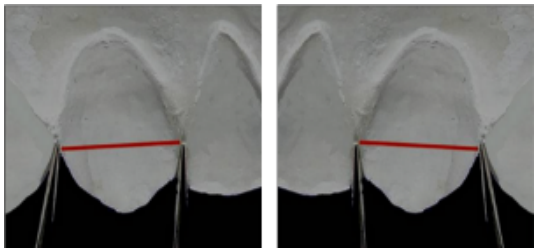

Figura 12.1: Medida del Ancho Mesiodistal.

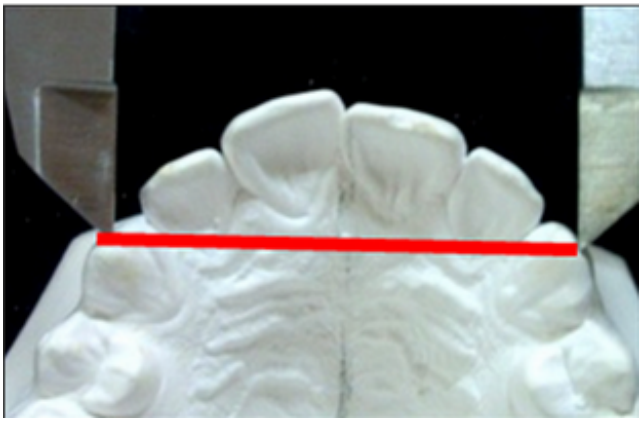

Figura 12.2: Medida de la Distancia Intercanina.

Rao define el Índice Mandibular Canino (I.M. Canino) como

$$\text{IMC} = \frac{\text{ancho de corona mesiodistal de caninos mandibulares}}{\text{ancho arco canino mandibular}} \tag{12.1}$$

a partir del cual computa el Índice de Rao (I de Rao)

$$\text{IMC estándar} = \frac{(IMC_1 - DE_1) + (IMC_2 + DE_2)}{2} \tag{12.2}$$

donde IMC_1 es el Indice Promedio Mandibular Canino para hombres y el IMC_2 es el Indice Promedio Mandibular Canino para mujeres. A su vez DE_1 y DE_2 representan los desvíos respectivamente de hombres y mujeres. A partir de la ecuación (12.2) se fija un punto de corte en 0.274, de manera que si el IMC estándar$_i \leq 0.274$ el individuo *-iésimo* se clasifica como hombre y si IMC estándar$_i > 0.274$ la persona se clasifica como del sexo femenino.

12.2.2. Método de Regresión Logística

Cuando la variable de respuesta Y_i es una *variable aleatoria Bernoulli*, con resultados posibles: *éxito* , *fracaso* codificados 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 sección 3.2.2 del capítulo 3

$$P(Y = 1|X) = \pi = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{12.3}$$

$$P(Y_i = 1|X_i) = \pi_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_p X_{pi}}}{1 + e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_p X_{pi}}} \tag{12.4}$$

12.2.3. Método de Análisis Discriminante

Al igual que lo visto en la sección 3.2.2, el Análisis Discriminante Probabilístico (AD) se trata de una técnica estadística multivariante de clasificación supervisada pero que tiene como finalidad la *Descripción*, donde interesa analizar si existen diferencias entre una serie de grupos en los que se divide una población con respecto a un conjunto de variables y, en caso afirmativo, averiguar a qué se deben. Por otro el AD busca hacer *Predicción*, a través de un procedimiento sistemático de clasificación de nuevas observaciones de origen desconocido en algunos de los grupos considerados.

Se dispone de k muestras de tamaño n_g $(g = 1, 2, \ldots k)$ que provienen de k poblaciones de las que se miden p características cuantitativas. Usando dicha información se desea determinar de cual de esas k poblaciones una nueva observación es más probable que haya sido seleccionada aleatoriamente.

Se asigna cada elemento de la población a uno de los grupos según una determinada regla de decisión, procurando cometer el menor error posible.

El análisis tiene determinado poder "predictivo" pues de algún modo los criterios usados para clasificar una población actual, pueden ser utilizados para nuevos elementos que se incorporen a ella.

La desventaja de este método es que se limita a variables explicativas de tipo cuantitativo. En el apéndice A, en la sección A.4 se dan detalles en profundidad del funcionamiento de la técnica.

12.2.4. Métodos CART (Classification and Regression Trees)

Esta métodología permite la construcción de modelos basados en técnicas no paramétricas, lo que supone muchas menos restricciones de distribuciones de probabilidad en las variables consideradas, permitiendo encontrar las variables que mejor discriminan el comportamiento de una variable de respuesta o dependiente de tipo categórica; la aplicación immediata es sobre variables que clasifican en ausencia o presencia de una patología, en diferentes niveles de patología (maloclusión, enfermedad periodontal) por ejemplo 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 facilmente usables e interpretables por los no especialistas en estadística, (Abernathy *et al.*, 1987).

El método CART (Classification and Regression Trees) fue introducido en 1984 por (Breiman *et al.*, 1984). Es una herramienta de análisis exploratorio de datos con el objetivo de encontrar reglas de clasificación y de predicción. Es uno de los métodos de aprendizaje inductivo supervisado no paramétrico más utilizado. Se aplican en diversas disciplinas, y se destacan por su fácil interpretación y aplicabilidad. Los Árboles de Clasificación y Regresión puede ser utilizados tanto en variables independientes categóricas como continuas. Si la variable de respuesta es continua, se obtienen árboles de regresión y son de clasificación cuando la variable de respuesta es categórica. Se caracteriza por no requerir una selección a priori de variables, presencia de interacciones o transformaciones de variables y es de más fácil interpretación que los modelos de regresión tradicionales.

El método es relativamente sencillo y consiste en la construcción de un árbol de decisiones, muy similar al que se usa en las guías clínicas. Se comienza con un nodo raíz que contiene todas las observaciones, este es dividido en subgrupos determinados por la partición de la variable elegida, generando nodos descendentes. Estos subgrupos son divididos 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, obteniéndose la mayor representación de una clase. Se consideran todas las particiones posibles, siendo cada partición 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.

Resumiendo, entre las ventajas de usar Árboles de Clasificación y Regresión se tiene:

- puede ser aplicado a cualquier tipo de variables predictoras: contínuas y categóricas;
- los resultados son fáciles de entender e interpretar;
- puede trabajar con datos faltantes;
- hace automáticamente la selección de variables;
- es invariante a transformaciones de las variables predictoras;
- es robusto a la presencia de *outliers*;
- es un clasificador no paramétrico, es decir que no requiere suposiciones de distribución de probabilidades;
- toma en cuenta las interacciones que pueden existir entre las variables predictoras;
- es rápido de calcular;
- muy fácil de representarse visualmente para los no especialistas.

Entre las desventajas se encuentran:

- el proceso de selección de variables es sesgado hacia las variables con más valores diferentes;
- dificultad para elegir el árbol óptimo;
- requiere un gran número de datos para asegurarse que la cantidad de observaciones en los nodos terminales sea importante.

12.3. Aplicación al estudio DPBIO2009

Este trabajo forma parte del estudio DPBIO2009 en el que se evaluaron varias medidas entre dientes en ambos maxilares, del lado derecho e izquierdo de los mismos en modelos de yeso.

12.3.1. Medidas y cálculos efectuados

Las variables que están en la tabla de datos y que se describen a continuación son las que se obtuvieron mediante el proceso de medición o que se

Maxilar	Masculino	Femenino	Total
Inferior	264	261	525
Superior	243	238	481
Total	507	499	1006

Tabla 12.1: Nro de maxilares evaluados por sexo según tipo de maxilar.

calcularon a partir de éstas. Para la aplicación que sigue solamente se presentan los resultados que corresponde al maxilar superior, para poder hacer la comparación con el trabajo original de Rao y Rao (1986).

Descrición de Variables		
Variable	Nombre	Descripción
V(1)	DMDd	diámetro mesiodistal derecho
V(2)	DMDi	diámetro mesiodistal izquierdo
V(3)	Biotipo	Biotipo (2 niveles)
V(4)	Tipo	Tipo de apiñamiento (3 niveles)
V(5)	AGId	Altura gingivo incisal derecho
V(6)	AGIi	Altura gingivo incisal izquierdo
V(7)	DIC	distancia intercanina
Índices calculados		
V(8)	ICMaxd	Indice Maxilar Canino derecho
V(9)	ICMaxi	Indice Maxilar Canino izquierdo
V(10)	IGId	Indice gingivo incisal derecho
V(11)	IGIi	Indice gingivo incisal izquierdo

Tabla 12.2: Conjunto de variables relevadas.

El Diámetro Mesiodistal (DMD) fue definido como la máxima distancia lineal entre las superficies proximales de los caninos (13, 23, 33 y 43), siendo medido a nivel de los correspondientes puntos de contacto, valiéndose de un compás de punta seca de Korhaus (Dentaurum), posicionado perpendicularmente al eje mayor dental, como se ve en la Figura 12.1. La Distancia Intercanina (DIC), también denominado ancho del arco canino, fue concebido como el segmento lineal delimitado por las puntas cuspídeas de los caninos (13 y 23 y 33 y 43, respectivamente), midiéndose con el auxilio de un calibre digital de puntas finas con una resolución de 0.01 mm. (150 mm. - Digimess, San Pablo, Brasil), tal cual se aprecia en la Figura 12.2.

Para cumplir con el debido proceso de calibración, por parte de un único operador, todas las medidas se tomaron en tres ocasiones diferentes, con un intervalo no menor a una semana entre ellas, sobre 25 modelos de trabajo

seleccionados aleatoriamente de la muestra principal.

Se determinaron los Índices Caninos Maxilares Estándares (ICM) del lado derecho ICMaxd e izquierdo ICMaxi para cada sujeto, con sus correspondientes medias y desvíos estándares, que se consignan en la Tabla 12.3). A partir de estos valores se establecen los puntos de corte siguiendo la metodología de Rao (ecuación (12.5)).

Índice Canino Maxilar derecho por sexo							
Sexo	Min	Q_1	Mediana	$\bar{x}$	Q_3	Max.	NA's
M	0.1983	0.2215	0.2323	0.2341	0.2430	0.3187	16
F	0.1392	0.2213	0.2325	0.2335	0.2433	0.3096	12
Índice Canino Maxilar izquierdo por sexo							
M	0.1988	0.2209	0.2310	0.2332	0.2423	0.3889	16
F	0.1390	0.2205	0.2313	0.2320	0.2420	0.3377	12

Tabla 12.3: Índices Caninos Maxilares Estándares.

$$\text{ICMDE} = (0.233 + 0.0203 + 0.234 - 0.019)/2 = 0.234 \tag{12.5}$$

$$\text{ICMIE} = (0.232 + 0.0295 + 0.233 - 0.025)/2 = 0.2345 \tag{12.6}$$

Clasificación según algoritmo de Rao para maxilar superior derecho			
	M	F	Total
	M	F	Total
Pronostica (M)	128	122	282
Pronostica (F)	99	104	203
Total	227	226	453
Clasificación según algoritmo de Rao para maxilar superior izquierdo			
	M	F	Total
Pronostica (M)	135	126	261
Pronostica (F)	92	100	192
Total	227	226	453

Tabla 12.4: Tabla de clasificación para sexo usando algoritmo de Rao.

En las figuras que siguen se puede observar cual es la densidad condicional del sexo de acuerdo a 2 variables que son el DMDi y el DIC variables originales

y por otra parte el ICM creado por Rao. En este caso se muestra para el ICMi (es decir para lado izquierdo)

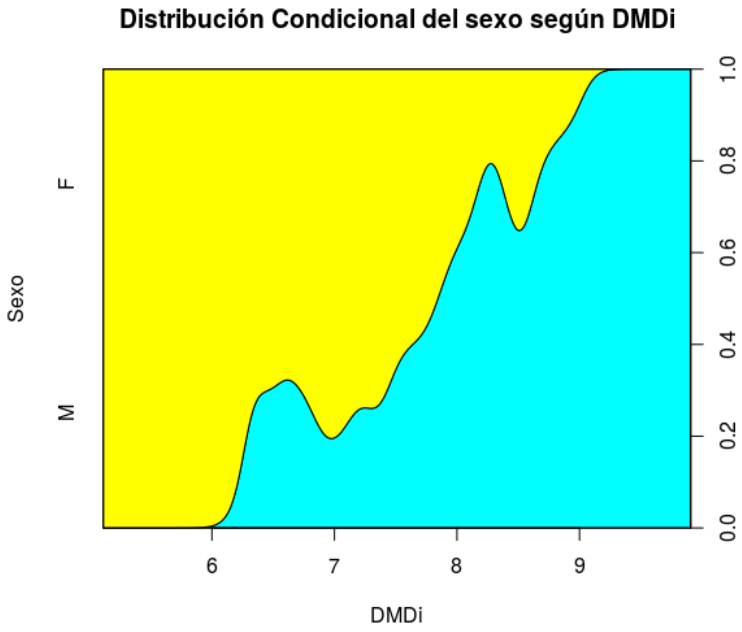

Figura 12.3: Proporción de M y F según DMDi para maxilar superior.

Por otra parte se muestra como es la relación bivariada de los 2 medidas usadas por Rao para la elaboración del ICM

Observando las Figuras 12.3 y 12.4 ya se puede ver que la distribución condicional del sexo al variar el IC y el ICMi de Rao el comportamiento no es monótono creciente lo que sería un indicio de que no hay una buena diferenciación por sexo en esas 2 variables, mientras que si la proporción de mujeres parece decrecer al aumentar el DMDi. Los mismos aspectos se pueden evaluar cuando se analiza el comportamiento multivariado de las variables y el Índice de Rao, lo que justifica el análisis que aparece en las próximas secciones, donde solamente se usa las mismas variables utilizadas por Rao pero con otro herramental estadístico.

12.3.2. Performance de la Regresión Logística

Para evaluar como funciona el método de RL se considera las variables usadas por Rao.

Para ver luego la calidad del ajuste se presenta una curva ROC en función

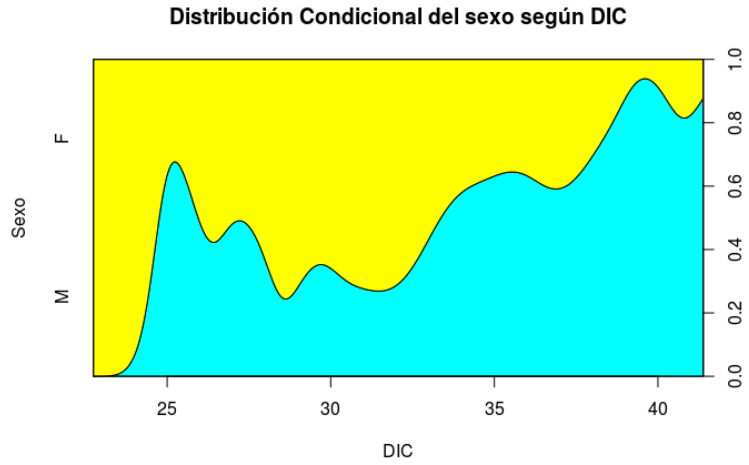

Figura 12.4: Proporción de M y F según DIC, para maxilar superior.

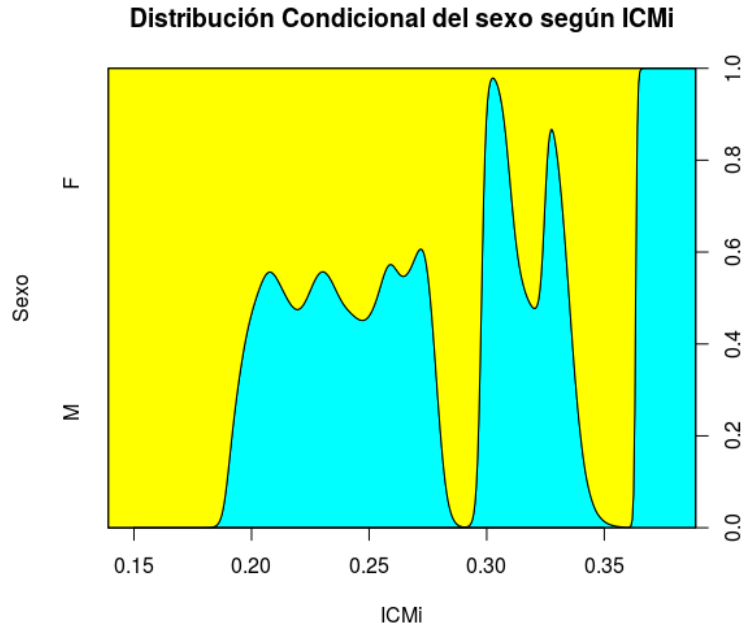

Figura 12.5: Proporción de M y F según ICMi, para maxilar superior.

Variables	Coeficiente	EE	valor Z	Pr(> |z|)
(Intercepto)	-17.7130	2.2489	-7.88	0.0000
DMDi	1.5946	0.2596	6.14	0.0000
DIC	0.1569	0.0457	3.43	0.0006

Tabla 12.5: Modelo de Regresión Logística para predicción del sexo.

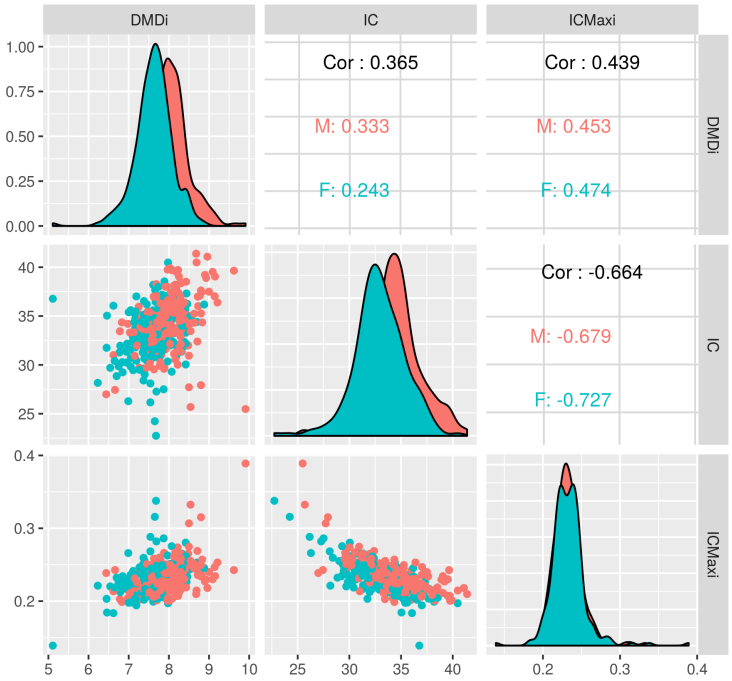

Figura 12.6: Relación para caninos izquierdos para maxilar superior.

Variables	coef	OR	LIIC_OR	LSIC_OR
DMDi	1.59	4.93	3.02	8.35
DIC	0.16	1.17	1.07	1.28

Tabla 12.6: Coeficientes y OR para Modelo de Regresión Logística.

del modelo ajustado, donde se puede ver el punto de inflexión de la curva donde estaría el óptimo de la Sen y de Esp y el área bajo la curva (AUC) como calidad global.

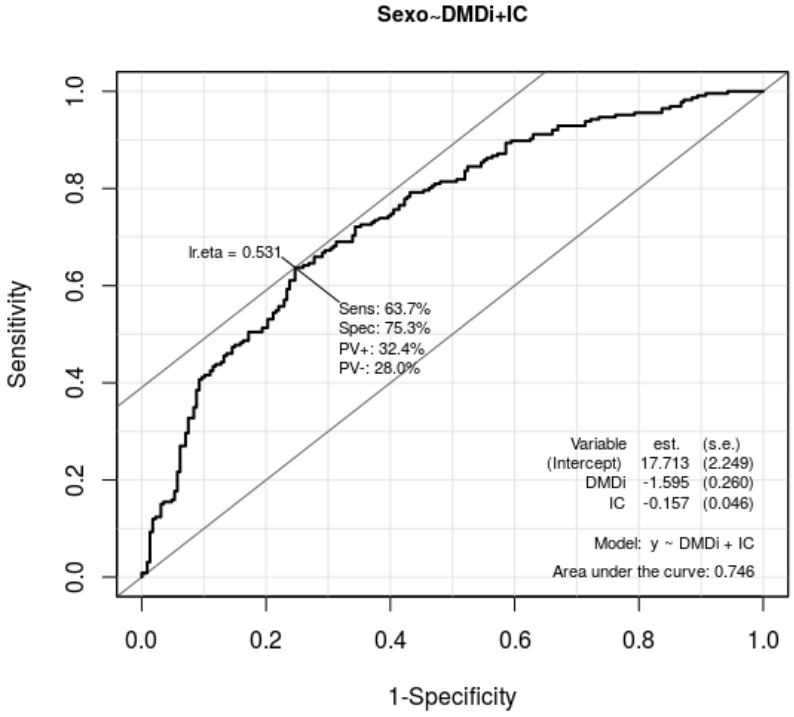

Figura 12.7: Curva ROC del Modelo de Regresión Logística para Maxilar Superior.

En la Figura 12.8 pueden observarse los valores de Sen, Esp junto con los valores predictivos

12.3.3. Performance del AD

Tomando en cuenta las 2 variables también propuestas para el modelo de RL, se estima una función discriminante lineal, para lo cual es necesario en primer lugar evaluar si realmente existe diferencias entre grupos para las 2 medias.

Se verifica a través de una prueba de igualdad de medias de las variables DMDi y de DIC por sexo que se rechaza, lo que es un buen indicio de que pueden ser buenas variables para discriminación y por otro lado se puede aceptar la igualdad de varianzas, que garantiza una buena performance del AD lineal, que resulta más sencillo que el que surge al aplicar un AD cuadrático que

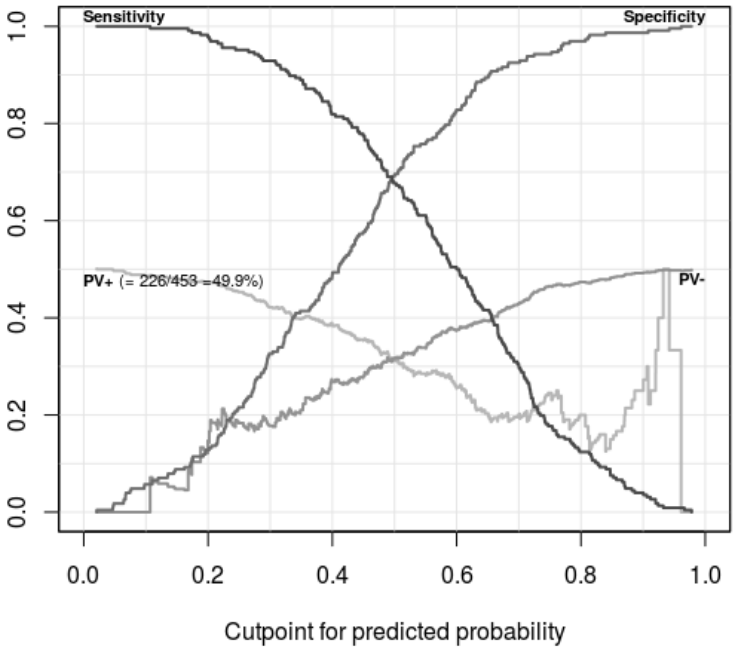

Figura 12.8: Sensibilidad, Especificidad y Valores predictivos de Curva ROC para Maxilar Superior.

además de tener más términos tiene como base que las matrices de covarianzas en ambos grupos no sea la misma en cada grupo.

Probabilidades a priori para los grupos		
M	F	
50.1	49.9	
Media de los grupos		
Sexo	DMDi	DIC
M	7.9	34.4
F	7.6	32.9
Coeficientes de la función discriminante lineal		
LD1	DMDi	DIC
	-0.164	-0.170

Tabla 12.7: Función discriminante lineal estimada (LD1) para predicción del sexo.

En la Tabla 12.7 quedan presentados los coeficientes que componen la función discriminante que en este caso es lineal.

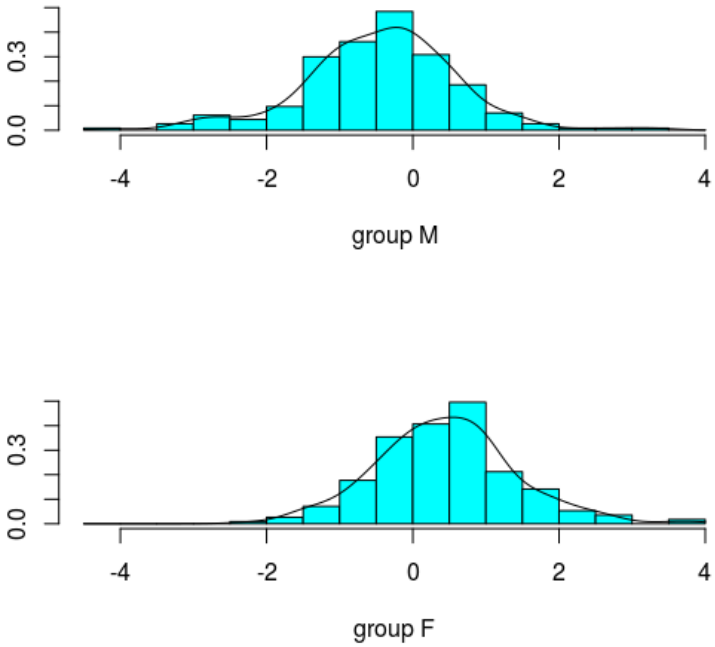

Figura 12.9: Score discriminante según grupos obervados.

12.3.4. Performance del CART

En este caso se estima una versión convencional de Árboles de Clasificación y Regresión donde en base a criterios de impureza de los nodos y una función que evalúa la calidad de la partición se elaboran los nodos y se debe decidir donde *podar* (que refiere donde cortar el árbol aislando grupos de individuos o nodos intermedios) bien diferenciados internamente y no llegar a los nodos terminales que serían todos los individuos por separado, (Therneau y Atkinson, 2018). Por otra parte se utiliza una versión más nueva y complementaria de CART que permite hacer un particionado recursivo haciendo inferencia condicional en los nodos. Es decir que a pesar de ser un modelo no parámetrico, permite en cada rama del árbol establecer si existe una diferencia significativa en cada nodo donde se establece un umbral de corte para las variables que en este caso son continuas; es decir que a la vez que construye el árbol con mayor capacidad de discriminación se está elaborando una prueba de hipótesis en forma secuencial. Este tipo de análisis de logra a través del uso de la librería *party* (Hothorn *et al.*, 2006). En la sección A.5 del apéndice A aparece

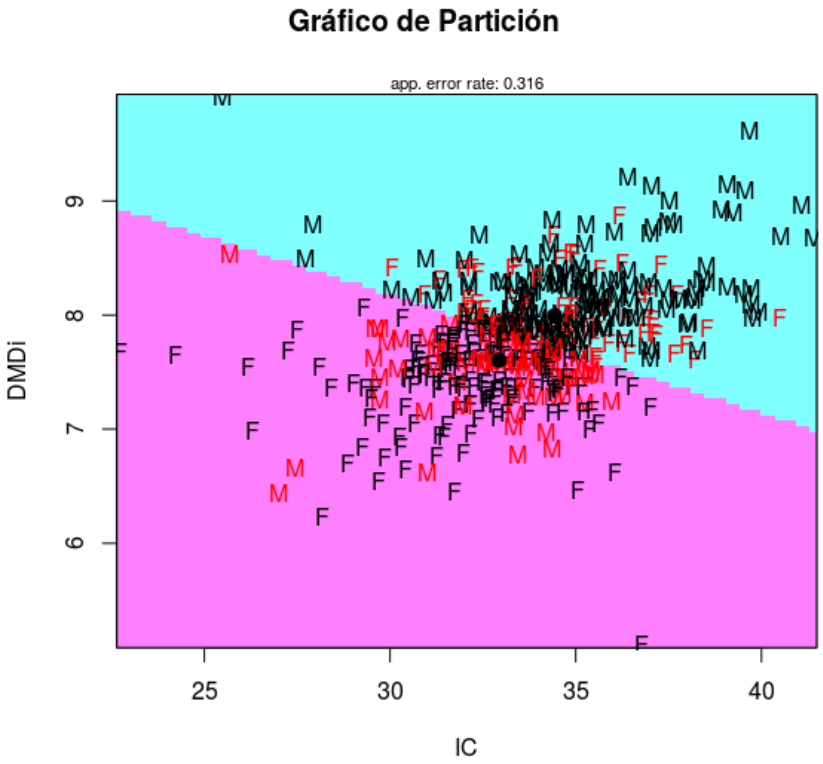

Figura 12.10: Gráfico de partición para AD.

un detalle más elaborado los aspectos más tecnicos de los CART.

De la Tabla 12.8 debe considerarse que se establece un valor de $CP = 0.02$, para establecer donde efectuar la poda y se observa que hay un cambio en la importancia relativa de la variable DMDi con respecto a DIC. Luego de la poda se obtienen 5 nodos, como aparecen en la Figura 12.11, donde los nodos terminales aparecen como rectángulos mientras que los nodos intermedios aparecen como óvalos. La idea de como interpretar cada nodo es comparar la proporción de F que se logra obtener diferenciándose lo más posible de la proporción original del nodo *raíz*.

fórmula = Sexo.rec DMDi + DIC			
Root node error: 226/453 = 0.4989	n= 453		
Costo Complejidad (CP)	nsplit	error relativo	
0.376	0	1.00	
0.020	1	0.62	
0.020	4	0.56	
0.010	8	0.51	
0.010	13	0.45	
Importancia de las variables			
DMDi	DIC	Punto de corte (CP)	Árbol
56	44	-	Original
64	36	0.02	Podado

Tabla 12.8: Indicadores de calidad de ajuste para 'poda' del árbol estimado.

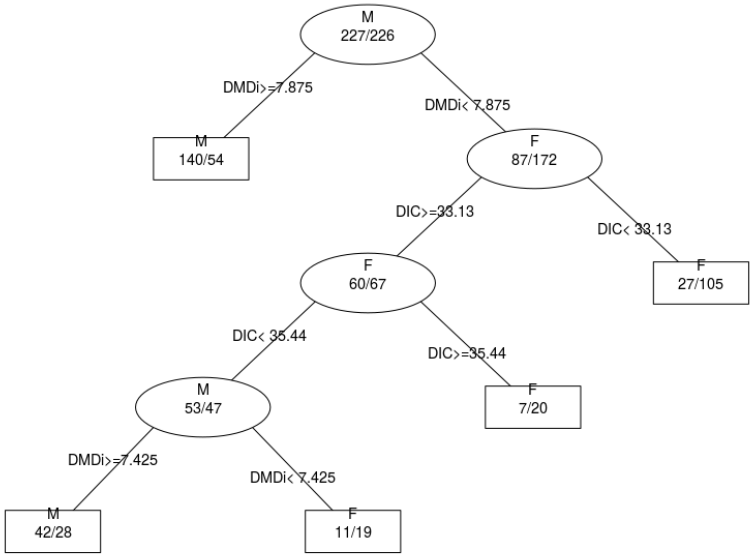

Figura 12.11: Árbol de Clasificación para sexo.

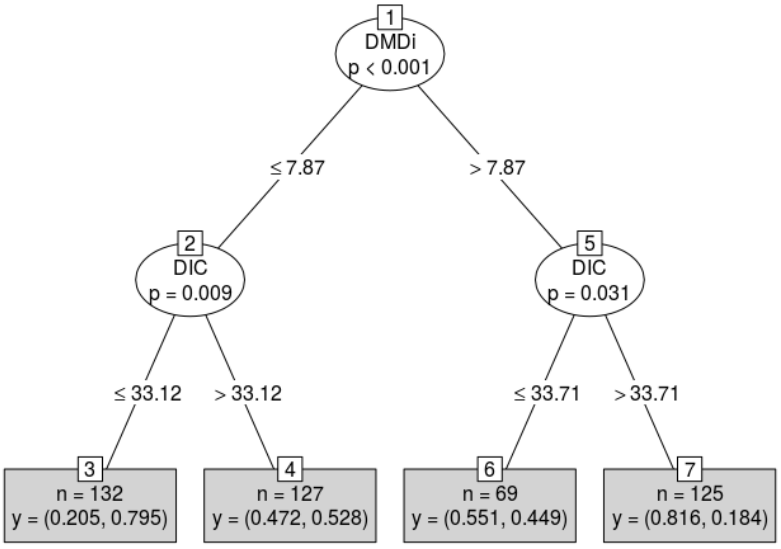

Figura 12.12: Árbol de Clasificación condicional para sexo.

En el reporte que sigue puede verse cual es el árbol que resulta de considerar la alternativa de inferencia condicional

```
Conditional inference tree with 4 terminal nodes

Response:  Sexo.rec
Inputs:  DMDi, IC
Number of observations:  453

1) DMDi <= 7.87; criterion = 1, statistic = 64.382
2) DIC <= 33.12; criterion = 0.991, statistic = 8.023
3)*  weights = 132
2) DIC > 33.12
4)*  weights = 127
1) DMDi > 7.87
5) DIC <= 33.71; criterion = 0.969, statistic = 5.852
6)*  weights = 69
5) IC > 33.71
7)*  weights = 125
```

La interpretación para el árbol mediante particionado recursivo en cuanto a los nodos es la misma que para el caso del CART.

12.4. Discusión

Luego de haber aplicado los 4 métodos de clasificación surgen 4 Índices que se denominarán Índice Canino Estándar de Rao (ICMDE,ICMIE), Índice canino por Análisis Discriminante, (ICAD), Índice canino por Regresión Logística (ICLOG) e Índice canino por árbol de clasificación (ICACa, ICACb), de los cuales se hace un resumen en cuanto a su performance

Comparación de los diferentes métodos de clasificación					
Índice	Método	Tasa de acierto Masculina	Tasa de acierto Femenino	Tasa de acierto Global	Tipo
ICMDE	Algoritmo de Rao derecho	56.3	45.9	51.2	M
ICMIE	Algoritmo de Rao izquierdo	59.4	46.0	51.9	M
ICAD	A. Discriminante	69.2	67.6	68.4	P
ICLOG	R.Logística	63.7	75.3	74.6	P
ICACa	Arbol Clasificación	75.3	79.6	77.4	NP
ICACb	Arbol particionado recursivo	61.6	76.1	68.8	NP

Tabla 12.9: Performance de los diferentes Índices de clasificación.

De los 4 índices es indudable que a pesar de ser el ICMDE o el ICMIE de Rao las formas más usadas, a partir de su artículo seminal, (Rao y Rao, 1986), para el caso del estudio en Uruguay éstos muestran una capacidad predictiva muy pobre. Por lo tanto los resultados más relevantes en este trabajo es que usando la mismas 2 variables el DMDi y el DIC con la RL, el AD y los árboles se incrementa notoriamente la capacidad predictiva, tal como se ve en la Tabla 12.9, donde es necesario aclarar que la columna Tipo refiere a si el índice es Manual (M), Paramétrico (P) y No paramétrico (NP).

Observando la Figura 12.6, ya se puede ver que las densidades de ambas subpoblaciones (F y M) son muy similares para el ICMaxi, lo que luego se refleja en una pobre performance del Índice canino de Rao.

Las ventajas de los Índices que surgen del AD y de la RL es que permiten ver el *peso* que tiene cada variable en la capacidad de discriminación, lo que se refleja los coeficientes de la función discriminante para el AD o en el impacto a través de los OR para el caso de la RL. La RL es un modelo paramétrico, ampliamente usado en biomedicina y para el cual se puede evaluar a través de la curva ROC en forma más fina su performance pudiendo encontrar el punto óptimo de corte para maximizar Sen y Esp, pudiendo además incorporar atributos de tipo categórico. En este caso y como complemento se muestra los resultados del mejor modelo de RL que partiendo de las 6 posibles variables

regresoras relevadas, aparece el apiñamiento que resulta relevante y que se consigna en la Tabla B.2, en el apéndice B .

Por otra parte los Índices que surgen de la aplicación de los métodos de CART y de particionado recursivo, son muy fáciles de visualizar e interpretar ya que son en base a umbrales de corte que surgen de las variables observadas y no consideran el tener que elaborar cálculos engorrosos posteriores como es el caso de los modelos paramétricos de RL y AD. En los 2 casos de las metodologías de CART y su variantes los resultados son muy similares mostrando ambos las mismas variables como las más importantes y con los mismos puntos de corte para la variable DMDi, que es la más importante, seguida por la DIC. Para el caso del método de CART convencional se obtienen 5 nodos terminales, donde la variable DIC sirve para con otros umbrales de corte generar nodos y por último vuelve a aparecer nuevamente DMDi. Por el otro lado para el caso del árbol construido por el particionado recursivo, se obtiene 4 nodos, donde no es necesario hacer proceso de poda. Ambos resultados se pueden ver en un gráfico de dispersión al ser solamente 2 las variables que interviene en el particionado.

12.5. Conclusiones

Luego de presentados los índices que surgen de aplicar, un método convencional y 4 basados en técnicas estadísticas presentadas mínimamente, se puede decir que para los datos del estudio DPBIO2009, se logra una mejora muy importante en la capacidad predictiva, siendo entonces el caso de lo que se definieron como Indicadores clásicos en sección 3.2, en este caso el ICLOG o como Indicadores del tipo alternativo como en sección 3.4 como lo son el ICAD, el ICACa y ICACb. Son una buena alternativa donde se recomienda en primer término por considerar una técnica estadística paramétrica ampliamente conocida, el ICLOG; y dado el potencial y facilidad de interpretación, el ICACa que resulta en un árbol convencional. Por otra parte las 3 técnicas de clasificación supervisadas tienen ya mucho tiempo de desarrolladas y han sido ampliamente utilizadas, siendo que en la actualidad, existen dentro de la misma familia otras, como por ejemplo las *redes neuronales*, que con el adelanto de la potencia de las computadoras cada vez se usan más en la biomedicina, formando parte de los que se conoce como *Machine Learning* y que han mostrado en algunos problemas de alta dimensión, una mejor performance.

Para finalizar, no obstante la mejoría encontrada, es necesario no perder de vista que se pueden mejorar la performance de los índices recomendados, tal como se plantea más adelante en el capítulo 13, tomando una serie de recaudos.

Parte III

Discusión y Conclusiones Generales

Capítulo 13

Conclusiones

Este trabajo lleva como nombre "*Sistematización y creación de Indicadores e Índices para la vigilancia epidemioógica en salud bucal: Uso de técnicas estadísticas multivariantes y de análisis espacio-temporal*" y es en el capítulo 1 donde en la sección 1.2 y 1.3 se propone la elaboración de una tipología de indicadores a través de la presentación y desarrollo de un conjunto de metodologías y técnicas estadísticas específicas, por lo cual el autor de esta tesis considera que se han alcanzado los objetivos.

13.1. Consideraciones sobre la parte 1 de la Tesis

Fueron necesarios 3 capítulos que forman la parte I para introducir al especialista en biomedicina, en la justificación, y la necesidad de elaborar esa sistematización en el ámbito de la salud bucal, como subdisciplina de la biomedicina y finalmente la metodología necesaria para lograrlo.
El capítulo 1 presenta una introducción acerca de los motivos en salud pública por los cuales para poder intervenir y modificar situaciones es necesario conocer a través de diferentes fuentes de datos y con diferentes estrategias metodológicas, información sistematizable.
A partir de esa información sistematizada es que se puede plantear hacer VE, donde juegan un rol fundamental los indicadores que resumen la misma, 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 la operacionalización de los

mismos en un intento de facilitar sus cálculos. Es esencialmente éste el disparador del objetivo general y los objetivos específicos que se plantean al final del capítulo, buscando elaborar una tipología de 3 tipos de indicadores: los I. Alternativos, los I. Combinados y los I. Espacio-Temporales. Esa tipología sin embargo a lo largo del trabajo se irá presentando al especialista en biomedicina, preservando un aspecto relevante y que consiste en aprovechar al máximo toda la información que cada uno de éstos contenga, es decir nunca perder la estructura multivariada aunque eso implique tener que trabajar con indicadores e índices más elaborados, pero con otras características y que consiste en no pedir más información, sino que tratarla de forma diferente, usando técnicas estadísticas tal vez un poco más sofisticadas y en muchas casos simplemente desconocidas en este ámbito.

El capítulo 2 presenta una síntesis de las patologías en salud oral más relevantes para ser tratadas desde la vigilancia epidemiológica y que son la Caries, la enfermedad Periodontal, la Erosión, la Maloclusión, los Trastornos Temporomandibulares y las Lesiones de Mucosa. En ese capítulo se resume para el investigador en biomedicina no especialista en salud bucal el alcance y la importancia de hacer vigilancia sobre las mismas, presentando a su vez los indicadores tradicionales y los algoritmos de cálculo de los mismos y que tal como se presentan son justamente los que justifican la necesidad antes planteada de darles otra mirada, con la ayuda de técnicas estadísticas más modernas y abarcativas y que potencian el análisis de las patologías.

Finalmente es el capítulo 3, el más extenso de esta primera parte pero que permite a juicio del autor de la tesis poner en conocimiento del investigador biomédico, el conjunto de técnicas estadísticas mínimas que debería no dominar pero si conocer su existencia, 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 3 tipos de indicadores consignados en la secciones 1.2 y 1.3 y abarcan los modelos de regresión (de tipo discreto y continuo), ampliamente usados en biomedicina. Aparecen entonces 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 son los modelos de conteo MC, los cuales se introducen muy brevemente para que el lector de familiarice, 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 I. Combinados, es que se presentan técnicas estadísticas que funcionan combinando muchos indicadores básicos (de ahí su nombre) y donde surge el Análisis Factorial, el Análisis o métodos de Clustering, que se presentan mínimamente y se profundizan si es necesario en los apéndices A.

Surge luego frente a la necesidad de medir más adecuadamente el mismo fenómeno la necesidad de generar I. Alternativos, con una simple transformación de los datos, trabajando con tasas o proporciones, teniendo que considerar una clase de modelos no tan extendidos en el ámbito de la biomedicina, como son los modelos de Regresión Beta, 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 Regresión Lineal (R. Lin).

Como última situación de indicadores alternativos, se 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, que muestran tener una gran utilidad para el análisis de la desigualdad, que se presentan con mucho de detalle en la parte II, en el capítulo 11, con una importante aplicación en una encuesta poblacional.

Para terminar la presentación de la tipología de indicadores se completa ésta con la que aparece en la sección 4, donde esta clase de indicadores I. Espacio-Temporales se divide en 3 de acuerdo a las dimensiones que intentan considerar

Agregaciones espaciales - donde interesa ver si en una determinada localización geográfica aparece o un patrón espacial que pueda considerarse como no aleatorio

Agregaciones temporales - donde se puede 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 espacio-temporales - donde se busca encontrar si existen clusters en el espacio, por un lado y en el tiempo, en forma simultánea, 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 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 esta tesis considera como aspecto complementario el herramental básico a ser conocido y tomado en cuenta por el investigador en biomedicina, dos tópicos que 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 epidemología muchas veces se debe tener en cuenta.

Por un lado se presentan los recaudos a tener en cuenta para poder trabajar adecuadamente en función del mecanismo de generación de los datos, habitualmente mediante muestras probabilísticas que garantizan la validez de los métodos estadísticos que habitualmente puedan usarse. Es así que surgen de ese modo 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 abundantemente con gran cantidad de referencias bibliográficas.

Finalmente como último aspecto a tener en cuenta por el lector es el que tiene que ver cuando es necesario hacer ajustes a los indicadores usando información auxiliar, cuando en la práctica cualquiera de los indices 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, donde no se verifica la independencia entre observaciones. Para eso una forma de corregir esos aspectos es trabajar con A. Multinivel.

13.2. Consideraciones sobre la parte 2 de la Tesis

Esta segunda parte de la tesis es a juicio del autor 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 en coautoría en varios artículos de revistas que se referencian, no se ve, dado la especificidad de la mismas (revistas clínicas o epidemiológicas del ámbito de la odontología), mostrando por lo tanto un énfasis en los detalles más técnicos estadísticos.

Esta segunda parte se divide en 9 capítulos, uno de los cuales presenta con mucho detalle las 4 fuentes de datos a ser tratadas en los 8 capítulos siguientes. El capítulo 4 presenta 4 estudios, 3 de los cuales fueron llevados adelante por docentes de la Facultad de Odontología de la Udelar y constituyen actualmente la información epidemiológica más importante a nivel poblacional de las patologías bucales presentadas en el capítulo 1, donde el autor participó en todas ellas teniendo 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, y que por lo tanto justifican su uso en solamente 2 de las 9 aplicaciones que se detallan a continuación.

Antes de resumir los hallazgos más relevantes en cada capítulo de esta parte II, el lector debe recordar que solamente se trataron algunas de las patologías presentadas en el capítulo 1, como son el CPO, y sus componentes así como la enfermedad periodontal. Por otra parte la elección de la patología y sobre que 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, la que se dicotomiza para un determinado umbral y trabajando 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, de que 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 R. Poisson. Por otra parte el autor muestra como contraejemplo, lo que sucede con una incorrecta aplicación 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.

En el capítulo 6, se trabaja con el estudio RPAFO2015, que tiene la ventaja de tener un diseño autoponderado, donde para el especialista en biomedicina, el manejo de modelos de conteo más complicados, puede hacerse con total tran-

quilidad. 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, buscando ajustar los componentes C, P y O del CPO.
Dado que persiste la patología en el comportamiento de esas variables se opta por presentar otra serie de modelos que si son conocidos y usado en varias publicaciones internacionales como son el MH o el MEC, que son modelos paramétricos más complicados al ser modelos combinados. Los aspectos 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, que debe entenderse como que la persona tiene Caries, no son las mismas si la parte del modelo de conteo del método combinado MH es Poisson o Binomial Negativa.

En el capítulo 7, se continua 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, mediante el uso de Regresión Beta, con una presentación detallada, logrando 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, siendo la edad 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, siendo nuevamente la edad y el CPO los que aparecen en la heterogeneidad.
Los demás modelos estimados si bien muestran variables significativas, el autor de la tesis si bien los presenta, no recomienda su uso, dado el pobre ajuste de los datos.

En estos 3 capítulos presentados hasta el momento, siempre se fue muy estricto en considerar que un modelo es bueno no solamente si tiene variables sigificativas, sino si muestra una buena capacidad de ajuste.

El capítulo 8 tiene un enfoque totalmente diferente y donde se usa una lógica de indicador combinado, pero con una diferencia metodológica muy im-

portante, donde 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 presencia de determinadas enfermedades, las que a su vez se asocian con otro conjunto de enfermedades, denominadas comorbilidades, medidas también a a través de variables binarias y 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 en base a los atributos binarios partiendo de un conjunto de variables, sin discriminar cuales son variables de respuesta, proponiendo la creación de grupos mediante 2 estrategias:

1. usando un método de clustering basado en el algoritmo *k-modes*
2. a través del análisis de redes SNA, construyendo 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.

Con ambas metodologías los resultados encontrados permiten decir que ambas detectan distintas cantidades de grupos, que si bien están diferenciados y tiene una explicación en el contexto del problema para el caso de la partición encontrada mediante *k-modes*, ésta es bastante difusa, mientras que la partición creada mediante el SNA se muestra como más estable.

El capítulo 9 vuelve a retomar las mismas 10 variables en formato binario pero donde se sigue una lógica modelizante y en esencia la forma de trabajo interna corresponde al tipo de indicador combinado, ya que usa muchos indicadores que de por sí tienen sentidos como son las prevalencias de las ENT, 4 vinculadas a la salud bucal y donde se logra desarrollar un modelo paramétrico, totalmente novedoso que combina la teoría de TRI, modelos de Teoría de Respuesta al Ítem (específicamente el modelo de Rasch) con los MLG, permitiendo explotar al máximo la estructura multivariada de la información y poniendo de manifiesto una variable latente, ya que estos permiten el análisis conjunto de varias variables obteniendo como subproducto una valoración individual, que en este caso se interpreta como "sickness proneness" (propensión a

la enfermedad). Adicionalmente, el análisis presentado aquí extiende el modelo de Rasch incluyendo un predictor lineal que permite indagar sobre algunos los factores determinantes de la tendencia de los individuos a padecer las distintas patologías.
También se obtienen 4 curvas características para cada patología y surge como resultado más relevante que el nivel de prevalencia de las patologías estudiadas en forma conjunta están moduladas por el efecto de algunas covariables de interés, donde ni el hábito de fumar ni el sexo presentan un efecto significativo, mientras que la actividad física parece ser un atributo que en cierto manera modula la prevalencia de las enfermedades bucales aumentándolas, observando un efecto no lineal y creciente de la edad sobre la tendencia a padecer enfermedades bucales.

Como última aplicación sobre el estudio RPAFO2015, se presenta en el capítulo 10, un indicador combinado cuya aspecto más relevante, es el poner en evidencia la estructura multivariante de los componentes del CPO, pero desde una perspectiva visual exclusivamente, la que se presenta con un enfoque gradualista.
Este nueva forma de análisis es de tipo descriptivo a diferencia de otras técnicas que en capítulos anteriores intentaban modelizar los componentes del CPO y se basa en la metodología de ADC, donde las partes individuales de la composición se denominan componentes. Cada componente tiene una cantidad, representando su importancia dentro del conjunto y la suma es constante. Como caso particular del análisis de datos de composicionales, el CPO al tener solamente 3 componentes permiten una visualización plana a través de los gráficos triangulares.
Esta alternativa gráfica denominada *CPO-grama*, permitirá explorar las relaciones 2 a 2 de cada componente del CPO pero también su valor global, con la posibilidad de poder visualizar en forma simultánea cada componente, el CPO y otros atributos de cada persona analizada.

El capítulo 11 es un claro ejemplo de como a partir de indicadores básicos, se crean indicadores combinados y alternativos al mismo tiempo, que permiten evaluar desigualdades en salud sobre algún tipo de patología bucal. En este caso se trabaja con el estudio RACA2012, que tiene diseño muestral complejo, que podría representar una complejidad extra, con lo relativo a los pesos muestrales, aspecto que se soluciona trabajando a nivel agregado de las escuelas, expandiendo la información, agrupando esta últimas en UG.

Sobre este estudio se trabaja con la prevalencia de Caries, mostrando 4 tipos de Índices de desigualdad, creados a partir de indicadores básicos y que muestran que trabajar a nivel agregado de tipologías de escuelas (sociodemográfica y geográfica) en Montevideo, no es la mejor alternativa ya que tiende atenuar las desigualdades. Cuando se trabaja a nivel individual y respetando el diseño muestral surge una clara diferenciación, mostrando que hay una gran concentración, incluso en tipo de escuelas más favorables como las privadas, lo que es un muy buen trazador de políticas de intervención a nivel de programas de salud y racionalización de recursos.

El capítulo 12 es una clara aplicación de un problema, donde a partir de un indicador validado por la comunidad que trabaja en Antropología Forense para la identificación del sexo a partir de algunas medidas de las piezas dentales en restos humanos, usando indicadores de tipo combinados, se logra una mejora sustancial con la clasificación supervisada, usando 2 modelos de tipo paramétricos y 2 de tipo no paramétricos.
En este caso los 4 modelos presentados dan origen a 4 índices que solamente usan la misma información del algoritmo original de Rao, dándole a la odontología forense por lo menos 2 soluciones (un modelo no paramétrico basado en aŕboles de clasificación), muy fácil de ser evaluado, sin necesidad de especialistas en estadística, con una representación visual muy similar al de las guías clínicas, que suelen representarse mediante árboles de decisión.
El otro modelo que mejora la capacidad predictiva sobre el algoritmo de Rao es un modelo paramétrico (en este caso modelo de regresión logística R. Logística), que tiene la ventaja de poder complementarse con otros instrumentos como la curva ROC.

Finalmente antes de pasar a la sección 13.3, donde se dejarían planteadas algunas futuras líneas de acción, es importante mostrar al lector de la tesis el debe con respecto a unos de los 3 tipos de indicadores desarrollados y presentados en la sección 3.5 y para los cuales no hay en esta segunda parte de la tesis ninguna aplicación.
El motivo de que no se puedan 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 tanto en el capítulo 1, como en el anteproyecto de tesis que se adjunta en el apéndice C, lo esperable era trabajar con el sistema Rediente, sin embargo este sistema que tiene mucho dato relativo a las personas que a lo largo de muchos años, se han atendido en la Facultad de Odontología de la Udelar, es

un sistema cerrado que solo genera indicadores agregados y no permite trabajar con los microdatos, donde aplicar cualquiera de las metodologías planteadas en la sección 3.5.

Lamentablemente el resto de las fuentes de datos trabajadas, que son de corte transversal, tampoco fueron concebidas en su diseño muestral, para contemplar el uso de ese herramental y tal vez solamente en el caso del estudio RACA2012, se podrían ensayar el uso de algunos indicadores para mostrar agregaciones espaciales, pero siendo cuidadoso de no abusar de las mismas. Tal como se consignó oportunamente el manejo de matrices de contigüidad, aspecto clave en estas técnicas es lo más difícil de establecer. De cualquier modo, en el capítulo 11, surgen indicios de que existen agregaciones espaciales para la prevalencia de Caries, a través de las medidas de desigualdad reportadas.

Resta advertir al lector de la tesis que si bien no hay aplicaciones de este tipo, es fundamental que el investigador en biomedicina sepa que también el manejo del tiempo y espacio es un aspecto relevante en el mecanismo de generación de los datos, tanto como el saber si éstos tienen asociado un diseño muestral que pueda condicionar el análisis y de ahí el motivo de la inclusión en el capítulo 3 a juicio del autor de este trabajo.

En la Tabla 13.1 se resentan un resumen de los diferentes tipos de Indicadores sistematizados y en que capítulo de presentan, con la particularidad que en varios aplicaciones se combinan éstos.

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
Capítulo 9	Si	Si	Si
Capítulo 10		Si	Si
Capítulo 11			Si
Capítulo 12	Si		Si

Tabla 13.1: Resumen de los diferentes tipos de Indicadores según capítulo donde aparece una aplicación en parte II de la tesis

13.3. Consideraciones generales y planes a futuro

En las 2 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, pero donde en cada una de éstas, cuando se termina se dejan planteadas las limitaciones encontradas y también eventuales desarrollos complementarios.

Un primer desafío para el autor es tratar de socializar los metodologías y resultados encontrados entre el grupo de investigadores del área de la biomedicina, con el que 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. Que el investigador en biomedicina en Uruguay, cualquiera sea su área de trabajo vea este trabajo como una aproximación a poder usar el herramental estadístico que acá se presenta y potenciar su trabajo con el uso de éstos para sentirse competitivo a nivel internacional.

Por otra parte el autor de este trabajo considera que el tiempo que fue necesario para llegar a este producto era el adecuado y donde los cursos tomados en diferentes programas de posgrado (Maestría en Economía, Maestría en Ingeniería Matemática, Maestría en Bioinformática) sirvieron para ir adaptando cada aplicación y redefinirlas, aspecto que se ve comparando las metodologías usadas y las que se propusieron en el anteproyecto.

Otra cuestión que interesa dejar planteada es que si bien cada aplicación se trabajó siendo extremadamente riguroso en no abusar de la técnica estadística, se preservaron 2 aspectos primordiales: no perder la estructura multivariada pero solamente usar modelos que tengan buen nivel de ajuste, ya que de otro modo el investigador en biomedicina estaría desarrollando teoría clínica o epidemiológica (cualquiera sea el problema en estudio) que estaría describiendo otra realidad.

Por otra parte en todos los casos no se trabajó, particionando los datos con muestras de aprendizajes para los modelos de regresión, donde 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 enteder más aún, partiendo del supuesto que en general en mucha revistas de biomedicina, ese aspecto metodológico no se sabe si se lleva a cabo al no consignarlo.

Para terminar y dejar algunas líneas de trabajo (en las que actualmente se está trabajando) se propone para:

Modelos de conteo Tratar de solucionar las limitaciones del uso de todos estos, en el contexto de estudios con diseños muestrales complejos, que en el trabajo de tesis, dadas las limitaciones del software usado permitió trabajar solamente con un tipo, con lo cual la solución puede ser desarrollar subrutinas que contemplen adecuadamente los aspectos requeridos o cambiar de paquete estadístico, ya que 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 como realiza los procesos de estimación

Extensión sobre Modelos Probabilísticos para ajustar tasas Profundizar en el uso de los mismos 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 Regresión Beta) 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 Regresión Beta, (Ospina y Ferrari, 2012),(Grün *et al.*, 2011),(Zeileis *et al.*, 2008).

Extensión de los modelos de TRI Extender los modelos analizando sus propiedades y los procedimientos estadísticos necesarios para realizar inferencias sobre sus parámetros. Evaluar resultados de la aplicación errónea de modelos TRI clásicos, tratando de profundizar los aspectos computacionales y posibles variantes o mejoras en el ajuste e inferencia (opcional). Evaluar las modificaciones análogas a las realizadas en los modelos de variables binarias pero al caso de variables categóricas o de tipo ordinal. Finalmente sería deseable poder abordar el problema con un enfoque bayesiano de los modelos TRI.

Análisis de Redes Incorporar cada vez más este tipo de análisis. En realidad el problema bajo 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. También dado que hay además de las ENT, otros atributos de las personas que pueden justificar su inclusión en el análisis se puede pensar en elaborar una red donde 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 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 le 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 aparece como 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. Proponer un análisis de redes validando modelos estadísticos (recordar que esto es solo 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 espacio-temporales Dado que actualmente el Uruguay no dispone de datos disponibles a investigadores en salud bucal (ni con el sistema Rediente ni con la nueva HIFO (Historia clínica de la Facultad de Odontología), pensada para la gestión y no como sistema de información estadístico), se propone trabajar con las lesiones por accidente de tránsito e internaciones que surge diferentes fuentes en 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 3 fuentes de datos que tiene 15 años de existencia, con frecuencia mensual, permitiría aplicar muchas de las herramientas presentadas en la sección 3.5. Actualmente el autor de la tesis 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 posteriormente ser complementada

con la metodología de Age-period-cohort Analysis (Age-period-cohort Analysis), (Yang y Land, 2013).

Referencias bibliográficas

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.

Acharya, A. B. y Mainali, S. (2009). Limitations of the mandibular canine index in sex assessment. *J. Forensic Leg. Med.*, 16:67–69.

Acharya, A. B., Prabhu, S., y Muddapur, M. V. (2011). Odontometric sex assessment from logistic regression analysis. *International Journal of Legal Medicine*, 125(2):199–204.

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.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological).*, 44(2):139 – 177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Springer Netherlands.

Alvarez Loureiro, L., Fabruccini Fager, A., Alves, L., Álvarez-Vaz, R., y Maltz, M. (2015). Erosive tooth wear among 12-year-old schoolchildren: A population-based cross-sectional study in montevideo, uruguay. *Caries Research*, 49:216–225.

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. (2019). Caracterización de las desigualdades de salud bucal entre escolares de 12 años de montevideo, uruguay. Documento de Trabajo Serie DT (19 / 02) - ISSN : 1688-6453, IESTA.

Álvarez-Vaz, R. y Dibarboure, H. (2008). Programa de vacunación antineumocócica al alta hospitalaria. Módulo Epidemiología Aplicada a los Servicios de Salud: Programa destinado a pacientes usuarios del Sistema de Salud del Sub-Sector Público, Adultos, Montevideo, 2008.

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. (2014). Distribución bernoulli multivariada. una aplicación a la salud oral. Documento de Trabajo Serie DT (14 / 03) - ISSN : 1688-6453, IESTA.

Álvarez-Vaz, R. y Massa, F. (2017). Elaboración de perfiles epidemiológicos en estudios sanitarios mediante técnicas de clustering difuso. En *VII Jornadas Académicas de la Facultad de Ciencias Económicas y de Administración, Udelar*.

Álvarez-Vaz, R. y Massa, F. (2018a). Elaboración de perfiles epidemiológicos en estudios sanitarios mediante técnicas de clustering difuso. *Saberes*. En referato.

Álvarez-Vaz, R. y Massa, F. (2018b). Evaluación De La Salud Bucal a Través De La Teoría De La Respuesta Al Ítem en Un Estudio Poblacional en Uruguay. En *XI Semana Internacional de la Estadística y la Probabilidad,FCFM-BUAP*, Aplicaciones en Estadística y la Probabilidad. Benemérita Universidad Autónoma De Puebla.

Álvarez-Vaz, R. y Massa, F. (2018c). Item response theory modelling assessment of oral health in a uruguayan population study'. En *33rd International Workshop on Statistical Modelling*, volumen II, Bristol.

Álvarez-Vaz, R. y Massa, F. (2018d). *Uso de la Distribución Bernoulli Multivariada en salud bucal*. Aplicaciones en Estadística y la Probabilidad. Benemérita Universidad Autónoma De Puebla-Dirección General de Fomento Editorial.

Álvarez-Vaz, R. y Massa, F. (2018e). Uso de la distribución bernoulli multivariada en salud bucal. Poster para XI Semana Internacional de la Estadística y la Probabilidad,FCFM-BUAP.

Álvarez-Vaz, R., Massa, F., y Lorenzo-Erro, S. (2019a). Aplicación de análisis de redes para la elaboración de perfiles epidemiológicos en estudios sanitarios. En *XII Semana Internacional de la Estadística y la Probabilidad*, Trabajos Extensos. Facultad de Ciencias Físico-Matemáticas, Universidad de Puebla.

Álvarez-Vaz, R., Massa, F., Lorenzo-Erro, S., y Fabruccini Fager, A. (2018a). Comparing 2 methods of statistical modeling of oral health in a population study in uruguay. En *International Association for Dental Research, VII Congreso de la Región Latinoamericana*.

Álvarez-Vaz, R., Massa, F., y Muñoz Woolf, M. (2018b). Visualization for the multivariate structure of the components of the DMFT using analysis of compositional data. En *International Association for Dental Research, VII Congreso de la Región Latinoamericana*.

Álvarez-Vaz, R., Massa, F., y Vernazza, E. (2019b). Modelos de conteo alternativos para los componentes del cpo en un estudio poblacional. *Revista de la Facultad de CIencias de la UNiversidad Nacional de Colombia*. En referato.

Álvarez-Vaz, R., Massa, F., Vernazza, E., y Fabruccini Fager, A. (2019c). Creación de indicadores alternativos para la vigilancia en salud oral mediante Regresión Beta. En FCFM, editor, *XII Semana Internacional de la Estadística y la Probabilidad*, Trabajos Extensos. Facultad de Física-Matemática, Universidad de Puebla.

Álvarez-Vaz, R., Picapedra, A., y Hugo-Neves, F. (2018c). Characterization of morbidity and mortality from automobile accidents using temporal spatial epidemiological tools in Brazil between 2000 and 2015. En *International Association for Dental Research, VII Congreso de la Región Latinoamericana*.

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. y Vernazza, E. (2012). Creación De Indicadores Alternativos Para La Vigilancia en Salud Oral, Mediante Regresión Beta. En *III Jornadas Académicas de la Facultad de Ciencias Económicas y de Administración, 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., 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.

Baker, F. (2017). *The basics of item response theory using R*. Springer, Cham, Switzerland.

Bakkannavar, S. M., Manjunath, S., Nayak, V. C., y Pradeep Kumar, G. (2014). Canine index - a tool for sex determination. *Egyptian Journal of Forensic Sciences*, pp. 8–12.

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 en Lecture Notes in Computer Science. Springer, Berlin ; New York.

Brathall, D. (2000). Introducing the significant caries index together with a proposal for new global oral health goal for 12-year-olds. *International Dental Journal*, 50(6):378–384.

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*, capítulo 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 Pública*, 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.

Campos Neto, M. y Paulete Vanrell, J. (2014). *Atlas de medicina legal. Guia prático para médicos e operadores do direito*, volumen tomo 1. LEUD.

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(362):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(1):47–9.

Chen, R., Mantel, N., Connelly, R., e Isacson, P. (1982). A monitoring system for chronic diseases. *Meth Inform Med.*, 21(2):86–90.

Chen, R., McDowell, M., Terzian, E., y Weatherall, J. (1983). Eurocat guide to monitoring methods for malformation registers. *EEC Concerted Action Project Eurocat.*

Chessel, D., Dufour, A. B., y Thioulouse, J. (2004). The ade4 package — I: One-table methods. *R News*, 4(1):5–10.

Clark, D. (1994). An analysis of the value of forensic odontology in ten mass disasters. *International Dental Journal*, 44:241–250.

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.

Eboh, D. y Etetafia, M. (2010). Maxillary canine teeth as supplement tool in sex determination. *Annals of Biomedical Sciences*, 9(1).

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). Socio-economic 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.

Fabruccini, A., Alves, L., Alvarez, L., Álvarez-Vaz, R., Susin, C., y Maltz, M. (2016). Comparative effectiveness of water and salt community-based fluoridation methods in preventing dental caries among schoolchildren. *Community Dentistry and Oral Epidemiology*, 44(6):577–585.

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.

Goettems, M., Ourens, M., Cosetti, L., Lorenzo, S., Álvarez-Vaz, R., y Celeste, R. K. (2018). Nível socioeconômico na primeira infancia e oclusopatia em adolescentes e adultos jovens no Uruguai. *Cad. Saúde Pública*, 34(3).

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(2):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. *American Journal of Epidemiology*, 1(132):532– 42.

Harvey, J. (1975). *Dental identification and Forensic Odontology*, pp. 140–157. Legal aspects of dental Practice. John Wright &Sons, Bristol.

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.

Hilbe, J. (2017). *Logistic Regression Models*. CRC Press.

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.

Hosmer, D. y Lemeshow, S. (1988). *Logistic Regression*. John Wiley & Sons, New York.

Hothorn, T., Hornik, K., y Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

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(1):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.

Kreiner, M., Álvarez-Vaz, R., Waldenström, A., Michelis, V., Muñiz, R., e Isberg, A. (2014). Craniofacial pain of cardiac origin is associated with inferior wall ischemia. *Journal of Oral Facial Pain Headache*, 28(4):317–321.

Kreiner, M., Zaffaroni, A., Álvarez-Vaz, R., y Clark, G. (2010). Validation of a simplified sham acupuncture technique for its use in clinical research: A randomised, single blind, crossover study. *Acupuncture in Medicine*, 28(1):33–36.

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., Andrade, E., Piccardo, V., Francia, A., Massa, F., Correa, M. B., y Peres, M. (2015). Periodontal conditions and associated factors among adults and the elderly: findings from the first National Oral Health Survey in Uruguay. *Cadernos de Saúde Pública*, 31:2425 – 2436.

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 Álvarez-Vaz, R. (2013b). Enfermedad Periodontal en la población joven y adulta uruguaya del Interior del pais: Relevamiento Nacional 2010-2011. *Odontoestomatología*, 15:35 – 46.

Lorenzo, S., Piccardo, V., Alvarez, F., Massa, F., y Alvarez-Vaz, R. (2013c). Enfermedad Periodontal en la población joven y adulta uruguaya del Interior del país: Relevamiento Nacional 2010-2011. *Odontoestomatología*, 15:35 – 46.

Lorenzo-Erro, S. (2003). Caries and socio-cultural factors in 12 year old pupils attending state owned schools. Tesis de máster, London: Kings' College. Dept. of Oral Health Services Research and Dental Public Health.

Lorenzo-Erro, S. M., Massa, F., Álvarez-Vaz, R., Schuch, H. S., Corrrea, M. B., y Peres, M. A. (2018). The role of contextual and individual factors on periodontal disease in Uruguayan adults. *Brazilian Oral Research*, 32.

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.

McCullagh, P. (1989). *Generalized linear models*. Chapman and Hall, London New York.

McCulloch, C. (2001). *Generalized, linear, and mixed models*. John Wiley & Sons, New York.

Meyer, D., Zeileis, A., y Hornik, K. (2016). *vcd: Visualizing Categorical Data*. R package version 1.4-3.

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*. División Epidemiología, Ministerio de Salud Pública.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365.

Nahidh, M., Ali Ahmed, H., Mahmoud, A., Murad, A., y Mehdi, B. (2013). The role of maxillary canines in forensic odontology. *Journal Bangh. Coll. Dentistry*, 25(4):109–113.

Narvai, P., Frazo, P., Roncalli, A., y Antunes, J. (2006). Caries dentaria no brasil: declínio, polarizacão, iniqüidade e exclusão social. *Rev. Panam. Salud Pública*, 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(310):532–38.

Naus, J. (1982). Approximations for distributions of scan statistics. *Journal American Statistical Association*, 77(377):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(3):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*, capítulo 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.

Parekh, D., Patel, S., Zalawadia, A., y Patel, S. (2012). Odontometric study of maxillary canine teeth to establish sexual dimorphism in gujarat population . *Int J Biol Med Res*, 3(3):1935–1937.

Pereira, C., Bernardo, M., Pestana, D., Santos, J., y Mendonca de, M. (2010). Contribution of teeth in human forensic identification-discriminant function sexing odontometrical techniques in portuguese population. *J. Forensic and Legal Med*, 17:105–110.

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.

Picapedra, A., Sassi, C., Massa, F., Francesquin Jr, L., Daruge, E., y Daruge Jr, E. (2012). Odontometric analysis of maxillas: a device for sex determination. *Inter J Dental Anthropol*, 21:01–16.

Pineault, R. (1988). *La Planificación sanitaria. Conceptos Métodos y Estrategias*. Masson.

Prabhu, S. y Acharya, A. B. (2009). Odontometric sex assessment in indians. *Forensic Science International*, 192(1):129.e1 – 129.e5.

Proffit, W. (1990). *Reactor paper: risk assessment for developmental problems-where are we now?*, pp. 162–163. Bader, J D. Ed. Risk assessment in dentistry. Chapell Hill, NC: University of North Carolina,.

R Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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.

Rai, B. y Anand, S. (2007). Gender determination by diagonal distances of teeth. *The Internet Journal of Biological Anthropology*, 1(110).

Raj, D. (1968). *Sampling Theory*. McGraw-Hill, Inc., New York.

Rao, C. y Miller, J. (2008). *Handbook of Statistics:Epidemiology and Medical Statistics*. Elsevier.

Rao, N., Rao, M., Pai, L., y Kotian, S. (1989). Mandibular Canine Index. *Forensic Science International*.

Rao, N. y Rao, N. (1986). Mandibular Canines in Establishing Sex Identity. *Journal Indian Forensic Medicine*, 8(1-2):5–12.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. John Wiley & Sons.

Reverté Coma, J. (1999). *Antropología Forense*. 2da edición edici

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.

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, 17(5):1–25.

Rothman, K. J. y Greenland, S. (1998). *Modern Epidemiology*. Lippincott Williams & Wilkins, 2 edición.

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.

Sharma, M. y Gorea, R. (2010). Importance of mandibular and maxillary canines in sex determination. *Journal of Punjab Academy of Forensic Medicine & Toxicology*, 10:27–30.

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.

Srivastava, P. (2010). Correlation of odontometric measures in sex determination. *Journal for Indian Academic Forensic Medicine*, 32(1):56–61.

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(1):199–213.

Tango, T. . (2010). *Statistical Methods for Disease Clustering*. Statistics for Biology and Health. Springer.

Therneau, T. y Atkinson, B. (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.

Thomas, J. C. y Weber, D. J. (2001). *Epidemiologic Methods for the Study of Infectious Diseases*. Oxford University Press, USA, 1 edición. ISBN: 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.

van den Boogaart, K. G., Tolosana, R., y Bren, M. (2014). *compositions: Compositional Data Analysis*. R package version 1.40-1.

Venables, W. y Ripley, B. (2002a). *Modern Applied Statistics with S*. Springer-Verlag, New York, 4th edición.

Venables, W. N. y Ripley, B. D. (2002b). *Modern Applied Statistics with S*. Springer, New York, 4 edici[U+FFFD]ISBN 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 den- tistry. 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 edición. 344.

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(1):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.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

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*, capítulo 13, p. 119. 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.

Zirahei, J., Sambo Amaza, D., Hamman, L., Jacks, T., Kwabwugge, Y., Quagar, J., y Kamal, S. (2013). Sexual Dimorphism in Maxillary Canine Teeth among Students of Kogi State Polytechnic, Nigeria.

Zorba, E., Moraitis, K., y Manolis, S. K. (2011). Sexual dimorphism in permanent teeth of modern greeks. *Forensic Sci. Int.*, 210(1-3):74–81.

APÉNDICES

Apéndice A

Aspectos metodológicos estadísticos

A continuación se presenta una suscinta descripción teórica de las herramientas estadísticas utilizadas en la tesis.

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 entonces, reducir las dimensiones de la tabla de datos, con una pérdida mínima de información.

Existen diversos métodos de análisis factorial, el *análisis de componentes principales (ACP)* que trata las matrices cruzadas de individuos por variables cuantitativas, el *análisis de correspondencias simple (ACS)*, que manipula tablas de contingencia, el *análisis de correspondencias múltiples* que trabaja con las llamadas tablas disjuntas, en donde se cruzan individuos por variables cualitativas, entre otros.

El *ACM* 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.[1],

La tabla de datos puede representarse de dos maneras, la primera es mediante la *tabla disyuntiva completa (TDC)*, en donde las filas representan a

[1]La presente Sección se encuentra basada en los textos de B. Escofier y J. Pags

los individuos y las columnas a las modalidades de las variables. Se definen la siguientes 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 sólo "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 tales tablas. Esta tabla es análoga a una matriz de correlaciones, en el sentido de que 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 basarse en una noción de semejanza tal que dos individuos están tanto más próximos cuanto mayor es el número de modalidades que poseen 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, o sea, obtener un número pequeño de variables sintéticas numéricas.
3. **Estudio de las modalidades**: una modalidad puede ser considerada según dos puntos de vista:
   - Como variable indicadora definida sobre el conjunto de los individuos. Esto es una columna de la *TDC*.
   - Como clase de individuos de la que se conoce la distribución sobre el conjunto de las modalidades. Sea 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 es 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 es su asociación con las mismas modalidades.

A.2. Análisis Factorial Múltiple *(AFM)*

El *AFM* es una herramienta desarrollada por B. Escofier y J. Pags (Escofier y Pagés, 2008), que permite el análisis simultáneo de variables que se encuentran 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 se encuentren, más determinan los principales factores.

Una de las formas de ponderar a cada variable dentro de un grupo puede determinarse por el inverso del valor propio asociado al primer eje factorial del *ACM* o *ACP* realizado en cada uno de ellos, denotado como λ_j^1.

El *AFM* permite estudiar:

- relaciones entre los 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, se encuentran situados 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 se encuentran 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 en el sentido de 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;
- 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;
- 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 aún equivalente a un ACM. Obsérvese que un ACM es equivalente a realizar 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

el describe 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, donde 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 cual miembro de la famila se usará, como por ejemplo distribución normal, binomial o de Poisson. En este caso la media condicional y la varianza de y_i vienen dadas por $E[y_i \,|\, x_i] = \mu_i = b'(\lambda_i)$ and $VAR[y_i \,|\, x_i] = \phi \cdot b''(\lambda_i)$.

Teniendo en cuenta el parámetro de dispersión ϕ, la distribución of 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 generalmente 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 par ajustar el modelo, se derivan de una familia particular.

A.4. Análisis Discriminante

Recordando que en el AD se dispone de k muestras de tamaño n_g ($g = 1, 2, \ldots k$) que provienen de k poblaciones de las que se miden p características y que se buscarán funciones que discriminen de la mejor manera entre grupos. es necesario definir reglas de decisión, procurando cometer el menor error posible, es necesario definir entonces:

- Modelo que se considere adecuado.
- Criterio de decisión o de clasificación.

- Modos de medir el error al utilizar los distintos criterios de asignación.
- Funciones discriminantes.

A la hora de clasificar se distinguen tres tipos de distancia que son de interés:

- Distancia entre unidades
- Distancia entre poblaciones
- Distancia entre unidad y población

A.4.1. Distancia entre individuos

La distancia de Mahalanobis entre los individuos i y l es:

$$D_{il} = \sqrt{(\mathbf{x_i} - \mathbf{x_l})'\Sigma^{-1}(\mathbf{x_i} - \mathbf{x_l})} \; \forall i \neq l \tag{A.6}$$

donde $\mathbf{x_i'} = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ vector observado del individuo i, Σ matriz de varianzas y covarianzas.

A.4.2. Distancia entre poblaciones o grupos

$$D_{gg'} = \sqrt{(\mu_\mathbf{g} - \mu_{g'})'\Sigma^{-1}(\mu_g - \mu_\mathbf{g'})} \; \forall g' \neq g \tag{A.7}$$

donde $\mu_\mathbf{g}' = (\mu_{1g}, \ldots, \mu_{pg})'$ centroide, vector de medias de las p variables. Σ matriz de varianzas y covarianzas.

A.4.3. Distancia entre individuo i y centroide de grupo

$$D_{ig} = \sqrt{(\mathbf{x_i} - \mu_\mathbf{g})'\Sigma_g^{-1}(\mathbf{x_i} - \mu_\mathbf{g})} \tag{A.8}$$

donde $\mathbf{x_i'} = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ vector observado del individuo i. Σ_g matriz de varianzas y covarianzas del grupo g.

La observación i es clasificada en el grupo g si $D_{ig} < D_{ig'} \; \forall g' \neq g$.

Se pueden cometer distintos tipos de error, llamamos $P(2|1)$ a la probabilidad de error de clasificar en el *grupo* 2 una observación proveniente del *grupo* 1, $P(3|2)$ a la probabilidad de error de clasificar en el *grupo* 3 una observación proveniente del *grupo* 2, etc. En forma genérica llamaremos $P(g'|g)$ al error de clasificar en el grupo g' una observación proveniente del grupo g.

La probabilidad de error de clasificar en el grupo g una observación proveniente del grupo g' es:

$$P(g'|g) = \int_{R_{g'}} f_g(x)dx \tag{A.9}$$

La probabilidad de clasificar erróneamente todas las observaciones pertenecientes al grupo g:

$$P(g) = \sum_{g'=1,\ g'\neq g}^{k} P(g'|g) = 1 - P(g|g) \tag{A.10}$$

La probabilidad total de clasificación errónea es:

$$P(R, f) = \sum_{g=1}^{k} \pi_g P(g) \tag{A.11}$$

La regla de decisión es la que minimiza la probabilidad total de error.

A.4.4. Principio de máxima verosimilitud

El principio consiste en asignar la observación i a la población donde el vector observado $\mathbf{x_i'} = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ tenga mayor verosimilitud de ocurrir, esto es, se asigna i al grupo g si:

$$\begin{aligned} f(\mathbf{x_i}|g) > f(\mathbf{x_i}|g')\ \forall g' \neq g \\ P(\mathbf{x_i}|g) > P(\mathbf{x_i}|g')\ \forall g' \neq g \end{aligned} \tag{A.12}$$

$$\Rightarrow \frac{f(x_i|g)}{f(x_i|g')} > 1 \tag{A.13}$$

A.4.5. Principio de probabilidad a posteriori

El principio consiste en asignar la observación i a la población con mayor probabilidad a posteriori (probabilidad de que i pertenezca a g condicionado al vector observado x_i).

La probabilidad a posteriori, según el Teorema de Bayes, se plantea de la siguiente forma:

$$P(i \in g|\mathbf{x} = \mathbf{x_i}) = \frac{\pi_g P(\mathbf{x_i}|g)}{P(\mathbf{x_i})} = \frac{\pi_g P(\mathbf{x_i}|g)}{\sum_{g'=1}^{k} \pi'_g P(\mathbf{x_i}|g')}$$
$$P(i \in g|\mathbf{x} = \mathbf{x_i}) = \frac{\pi_g f(\mathbf{x_i}|g)}{\sum_{g'=1}^{k} \pi_{g'} f(\mathbf{x_i}|g')} \tag{A.14}$$

donde $\pi_g = P(i \in g)$ es la probabilidad a priori de pertenecer al grupo g. La observación i se clasifica en el grupo g si:

$$P(i \in g|\mathbf{x} = \mathbf{x_i}) > P(i \in g'|\mathbf{x} = \mathbf{x_i}) \ \forall \ g' \neq g \tag{A.15}$$

Las reglas planteadas son óptimas si la distribución conjunta es conocida, con todos los parámetros conocidos.

A.4.6. Reglas de Clasificación, aplicadas a 2 grupos.

Dado $\mathbf{x_i} \sim \mathbf{N}_p(\mu, \Sigma)$ la función de densidad es:

$$f(\mathbf{x_i}|\mathbf{g}) = \frac{1}{(2\pi)^{p|2} \left|\Sigma_{\mathbf{g}}\right|^{1|2}} \exp\left\{-\frac{1}{2}(\mathbf{x_i} - \mu_{\mathbf{g}})' \Sigma_{\mathbf{g}}^{-\mathbf{1}} (\mathbf{x_i} - \mu_{\mathbf{g}})\right\} \tag{A.16}$$

$$\hat{P}(i \in g|\mathbf{x_i}) = \frac{\hat{\pi}_g \hat{f}(\mathbf{x_i}|g)}{\sum_{g'=1}^{k} \hat{\pi}_{g'} \hat{f}(\mathbf{x_i}|g')} = \frac{\hat{\pi}_g \left|\mathbf{S_g}\right|^{-1|2} \exp\left\{-\frac{1}{2}\hat{D}^2_{ig}\right\}}{\sum_{g'=1}^{k} \hat{\pi}_{g'} \left|\mathbf{S_{g'}}\right|^{-1|2} \exp\left\{-\frac{1}{2}\hat{D}^2_{ig'}\right\}} \tag{A.17}$$

Se distinguen dos casos:

1. En el caso de igualdad de matrices de varianzas la observación i se asigna al grupo g si:

$$\hat{\pi}_g \exp\left\{-\frac{1}{2}\hat{D}^2_{ig}\right\} > \hat{\pi}_{g'} \exp\left\{-\frac{1}{2}\hat{D}^2_{ig'}\right\} \ \forall g' \neq g \tag{A.18}$$

2. En el caso de matrices de varianzas diferentes la observación i se asigna al grupo g si:

$$\hat{\pi}_g \left|\mathbf{S_g}\right|^{-1|2} \exp\left\{-\frac{1}{2}\hat{D}^2_{ig}\right\} > \hat{\pi}_{g'} \left|\mathbf{S_{g'}}\right|^{-1|2} \exp\left\{-\frac{1}{2}\hat{D}^2_{ig'}\right\} \ \forall g' \neq g \quad \text{(A.19)}$$

Igualdad de costos Asignar i al *grupo* 1 si: $\pi_1 f(\mathbf{x_i}|\mathbf{1}) > \pi_\mathbf{2}\mathbf{f}(\mathbf{x_i}|\mathbf{2})$, la regla de clasificación suele presentarse de la siguiente forma

$$\frac{f(\mathbf{x_i}|\mathbf{1})}{f(\mathbf{x_i}|\mathbf{2})} > \frac{\pi_2}{\pi_1} \quad \text{(A.20)}$$

Con respecto a las probabilidades a priori se suelen estimar en base a la proporción que representa cada grupo en la muestra.**Igualdad de costos e igualdad de probabilidades a priori** Asignar i al *grupo* 1 si:

$$f(\mathbf{x_i}|\mathbf{1}) > \mathbf{f}(\mathbf{x_i}|\mathbf{2})$$

$$\frac{f(\mathbf{x_i}|\mathbf{1})}{f(\mathbf{x_i}|\mathbf{2})} > 1 \quad \text{(A.21)}$$

Costos y probabilidades a priori diferentes para cada grupo Asignar i al *grupo* 1 si: $\frac{f(\mathbf{x_i}|\mathbf{1})}{f(\mathbf{x_i}|\mathbf{2})} > \frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}$ Siendo $C(1|2)$ el costo de clasificar una unidad en el grupo 1 cuando proviene del grupo 2 y $C(2|1)$ el costo de clasificar una unidad como 2 cuando proviene del grupo 1.

A.4.7. Errores de clasificación en AD

Asociado a cada regla de decisión existen diversos modos de medir la probabilidad del error cometido. Se utilizan los n datos para construir la función de clasificación y luego se los clasifica en función de la misma. Una vez clasificados se cuentan los errores cometidos.

Para la Tasa de error aparente en el caso de 2 grupos($k = 2$), si π_1 y π_2 son desconocidos y las $n = n_1 + n_2$ observaciones son una muestra aleatoria de la población P, tenemos intuitivamente el estimador:

$$\hat{\pi}_i = \frac{n_i}{n_1 + n_2} \quad \text{(A.22)}$$

Y el estimador llamado "error aparente total" por algunos autores es:

$$\hat{e}_{app} = \hat{\pi}_1 e_{1,app} + \hat{\pi}_2 e_{2,app} = \frac{n_{12} + n_{21}}{n_1 + n_2} \quad \text{(A.23)}$$

De las n_1 observaciones del *grupo* 1, n_{12} fueron mal clasificadas al *grupo* 2 y n_{11} fueron bien clasificadas en el *grupo* 1 De las n_2 observaciones del *grupo* 2, n_{21} fueron mal clasificadas al *grupo* 1 y n_{22} fueron bien clasificadas en el *grupo* 2

Lo planteado anteriormente puede resumirse en la siguiente tabla de clasificación:

	g 1 (predicho)	g 2 (predicho)	Total
g 1 (observado)	n_{11}	n_{12}	n_1
g 2 (observado)	n_{21}	n_{22}	n_2
			n

A.4.8. Función discriminante para Análisis Discriminante Lineal

labelsec:ADL Las reglas de clasificación de las que se deriva la función discriminante lineal tienen como hipótesis:

- Distribución conjunta normal en cada grupo
- Igual de matriz de varianzas y covarianzas ($\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$).

Acorde a lo planteado en A.4.5 maximizar $P(i \in g|x = x_i)$ es equivalente a maximizar $\pi_g \exp\{-\frac{1}{2}D_{ig}^2\}$. Aplicando logaritmo

$$\begin{aligned} \log_e \pi_g - \frac{1}{2}D_{ig}^2 = \log_e \pi_g - \frac{1}{2}\{(x_i - \mu_g)'\Sigma^{-1}(x_i - \mu_g)\} = \\ \log_e \pi_g - \frac{1}{2}x_i'\Sigma^{-1}x_i + \frac{1}{2}\mu_g'\Sigma^{-1}x_i + \frac{1}{2}x_i'\Sigma^{-1}\mu_g - \frac{1}{2}\mu_g'\Sigma^{-1}\mu_g \end{aligned} \tag{A.24}$$

sustituyendo μ por $\bar{x}$ y Σ por S:

$$\begin{aligned} \log_e \hat{\pi}_g - \frac{1}{2}\hat{D}_{ig}^2 = \log_e \hat{\pi}_g - \frac{1}{2}\{(x_i - \bar{x}_g)'S^{-1}(x_i - \bar{x}_g)\} = \\ \log_e \hat{\pi}_g - \frac{1}{2}x_i'S^{-1}x_i + \frac{1}{2}\bar{x}_g'S^{-1}x_i + \frac{1}{2}x_i'S^{-1}\bar{x}_g - \frac{1}{2}\bar{x}_g'S^{-1}\bar{x}_g \end{aligned} \tag{A.25}$$

El *"score"* de la observación i en la **función discriminante lineal** estimada para el *grupo* g es:

$$L_{ig} = \log_e \hat{\pi}_g + \bar{x}_g'S^{-1}x_i - \frac{1}{2}\bar{x}_g'S^{-1}x_g \quad g = 1 \ldots k \tag{A.26}$$

O de otra manera la función discriminante es $L_g = Xb + C$

$$L_{ig} = b'_g x_i + c_g = b_{1g}x_{1ig} + b_{2g}x_{2ig} + \ldots + b_{pg}x_{pig} + c_g \tag{A.27}$$

con $b'_g = \bar{x}'_g S^{-1}$ y $c_g = \frac{1}{2}\bar{x}'_g S^{-1}\bar{x}_g + \log \hat{\pi}_g$

La *función de clasificación* estimada de la observación i en el *grupo* g es:

$$\begin{aligned} L_{igg'} = \frac{1}{2}\bar{x}'_g S^{-1}x_i + \frac{1}{2}x'_i S^{-1}\bar{x}_g - \frac{1}{2}\bar{x}'_g S^{-1}\bar{x}_g - \frac{1}{2}\bar{x}'_{g'} S^{-1}x_i - \frac{1}{2}x'_i S^{-1}\bar{x}_{g'} \\ = (\bar{x}_g - \bar{x}_{g'})S^{-1}[x_i - \frac{1}{2}(\bar{x}_g + \bar{x}_{g'})] \end{aligned} \tag{A.28}$$

La Región de clasificación es

$$R_{gg'} : L_{igg'} > \log(\frac{\hat{\pi}_{g'}}{\hat{\pi}_g}) \tag{A.29}$$

En el caso de probabilidades a priori iguales la regla se reduce a $L_{igg'} > 0$.

otra forma de plantearlo es: la observación i se clasifica en el *grupo* g si $L_{ig} > L_{ig'}\ \forall g' \neq g$.

A.5. Árboles de clasificación (CART)

La construcción del árbol se basa en tomar, en forma sucesiva distintas decisiones. Se comienza con un nodo raíz que contiene todas las observaciones, este es dividido en subgrupos determinados por la partición de la variable elegida, generando nodos descendentes. Estos subgrupos son divididos 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, obteniéndose la mayor representación de una clase. Se consideran todas las particiones posibles, siendo cada partición 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 existen un conjunto de preguntas binarias $\mathcal{Q}$. Si la variable es cuantitativa la pregunta será del tipo: $X_m \leq v$ y si la variable es cualitativa la pregunta será del tipo: $x_m \in S$ siendo S un subconjunto de $(b_1, b_2, ..., b_L)$ existiendo L clases posibles para la variable X . Si la respuesta

es positiva va a la izquierda y si la respuesta es negativa va a la derecha. Se generan así, 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, reduciendo los nodos terminales hasta llegar a un árbol óptimo.

Las particiones tienen como objetivo incrementar la homogeneidad de los subconjuntos resultantes de la misma. Existe asociada a cada partición 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

$$i(t) = \Phi(p(1/t), p(2/t), ..., p(J/t)) \tag{A.30}$$

con $p(j/t) = \frac{N_j(t)}{N(t)}$.

Como medidas de impureza se tienen:

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 sub-árbol se realiza mediante estimaciones de las tasas de clasificación incorrecta (por sustitución, por muestra de validación o por validación cruzada).

A.6. Diseños en fases

Hay situaciones donde 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 donde la precisión se logra minimizar en base a 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 costo aun grande ocasionado por la construcción del nuevo marco muestral.

La alternativa que queda finalmente 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 esos s_a se reune 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 bajo 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),(Breslow y Beckwith, 1982). Este procedimiento presenta además una ventaja extra , 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 originalmente pensada, puede ser vista como la primer fase de selección y el conjunto r/s como la submuestra de la fase 2.

A.7. Evaluación de la necesidad del análisis Multinivel

Un aspecto importante a considerar y 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 o para el ejemplo del estudio sobre colesterol los -médicos-hospitales-ciudades) (Finch, 2014).

$$ICC = \rho = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{A.31}$$

donde τ^2 representa la varianza entre clusters y σ^2 es la varianza dentro de los clusters. Un valor elevado de ρ estaría indicando que la variable de respuesta está muy 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 $\hat{\sigma^2}$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(n_j - 1)S_j^2}{N - C} \tag{A.32}$$

donde S_j^2es la varianza intraclase

$$S_j^2 = \frac{\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)^2}{n_j - 1} \tag{A.33}$$

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 simplemente el promedio ponderado de la varianza intraclusters.

Para estimar $\hat{\tau^2}$ se debe calcular la varianza ponderada interclase

$$\hat{S}_B^2 = \frac{\sum_{i=1}^{c}(\overline{y}_j - \overline{y})^2}{\tilde{n}(C - 1)} \tag{A.34}$$

donde $\overline{y}_j$ es la media de la variable de respuesta en el cluster j y $\overline{y}$ es la media global

$$\tilde{n} = \frac{1}{C - 1}\left[N - \frac{\sum_{i=1}^{c} n_j^2}{N}\right] \tag{A.35}$$

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

$$\hat{\tau}^2 = \hat{S}_B^2 - \frac{\hat{\sigma}^2}{\tilde{n}} \tag{A.36}$$

de donde puede calcularse

$$\rho_I = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{S}_B^2} \tag{A.37}$$

Apéndice B

Resultados estadísticos complementarios de las aplicaciones

En este apéndice se presentan resultados intermedios de las diferentes aplicaciones, que pueden ser tablas con modelos y resultados que por criterios de espacio se optan por que estén a disposición del lector interesado en profundizar. También figuras que sean de utilidad para complementar los resultados.

B.1. Aplicación 2

Tipo	modelo ajustado para componente C parámetros			
PG	Coeficientes	E.E.	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	
Tipo	modelo ajustado para componente P parámetros			
DP	Coeficientes	E.E.	t	Pr($>\|t\|$)
	$\mu = 2.28$	0.041	54.45	< 0.001
	$\gamma = 2.13$	0.0789	-12.56	< 0.001
	Devianza Global =4008.5	AIC=4012.5	SBC=4021.3	
Tipo	modelo ajustado para componente O parámetros			
DP	Coeficientes	E.E.	t	Pr($>\|t\|$)
	$\mu = 0.965$	0.109	8.80	< 0.001
	$\gamma = 2.03$	0.125	16.06	< 0.001
	Devianza Global =2880.1	AIC=2884.1	SBC=2892.1	
Tipo	modelo ajustado para componente O parámetros			
DP	Coeficientes	E.E.	t	Pr($>\|t\|$)
	$\mu = 2.80$	0.021	133.80	< 0.001
	$\gamma = 1.47$	0.05	26.8	< 0.001
	Devianza Global =4227.1	AIC=4231.1	SBC=4239.1	

Tabla B.1: Ajuste de la distribución de CPO y sus 3 componentes (Escenario B)

B.2. Aplicación 3

Variables	Coeficiente	Std. Error	z value	Pr(> |z|)
(Intercept)	19.1341	2.5178	7.60	0.0000
DMDi	-1.4657	0.2724	-5.38	0.0000
AGIi	-0.4377	0.1302	-3.36	0.0008
IC	-0.1485	0.0467	-3.18	0.0015
Tipo.rec=A	1.5946	0.6611	2.41	0.0159
Tipo.rec=N	1.2570	0.6527	1.93	0.0541

Tabla B.2: Modelo de Regresión Logística

s

Apéndice C

Anteproyecto de Tesis

Creación de indicadores alternativos para la vigilancia en salud oral, mediante técnicas estadísticas multivariantes, de aprendizaje supervisado y de análisis temporoespacial

Ramón Álvarez Vaz

Propuesta de Investigación de Doctorado para
PROGRAMA DE MAESTRIAS Y DOCTORADOS
EN CIENCIAS MEDICAS PRO.IN.BIO.
Escuela de Graduados (Facultad de Medicina)

RESUMEN

En el ámbito de la Salud Pública, existe la necesidad de conocer en profundidad las características de las poblaciones y los problemas de salud. De esa manera se puede intervenir para mejorarlos. Significa que sea necesario por lo menos tener una idea de la situación de partida y para eso se recurre a las fuentes de datos existentes. Entre ellas se destacan las estadísticas vitales; los registros de problemas específicos de salud (los registros de cancer que son registros de base poblacional), que permiten entre otras cosas establecer la incidencia de la enfermedad; registros de enfermedades de etiología infecciosa, con notificación obligatoria en los que se basan los sistemas de vigilancia epidemiológica.

Cuando la información que el epidemiólogo necesita no está disponible, se debe recurrir a diferentes mecanismos de generación a través del método científico de los diferentes diseño s de estudios sanitarios.

Sin embargo pueden existir limitaciones en los indicadores generalmente utilizados en la epidemiología y salud pública, 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 por lo tanto correctamente los fenómenos bajo estudio.

Teniendo en cuenta los antecedentes antes planteados con respecto a las fuentes de información en salud y en salud oral en particular, se propone construir un conjunto de indicadores alternativos y complementarios a los que ya existen. Se reformulará la forma de considerar la información que ya se viene recogiendo y para los cuales existen ya varias índices recomendados de la Organización Mundial de la Salud (CPO, CPI, SIC, IHOS) y otros índices epidemiológicos sobre estado de la salud oral. Para eso la propuesta a desarrollar consiste en estudiar la factibilidad de elaborar este conjunto de indicadores epidemiológicos a través de técnicas estadísticas multivariantes (*Análisis Factorial*, *Análisis de Conglomerados*, *Análisis Multiway*) y de aprendizaje supervisado como son los métodos CART *(Classification and Regression Trees)*, técnicas que no son muy usadas en el ámbito de la epidemiología. Con ellas se espera poder construir *tipologías* o grupos de poblaciones con perfiles epidemiológicos bien diferenciados de acuerdo a las patologías y los factores de riesgos asociados. A su vez se incorporará la dimensión temporoespacial, indispensable en la vigilancia epidemiológica actual y se sistematizaránlas fuentes de información disponibles para poder en una segunda instancia proponer los nuevos indicadores y evaluar finalmente la aplicabilidad de los mismos y la sustentantibilidad de los sistemas de vigilancia integrados por estos indicadores, en el tiempo y distribución territorial del país.

C.1. Antecedentes

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 en (Oriol, 1997), existen diferentes fuentes de datos para generar indicadores. Entre ellas encontramos las estadísticas vitales; 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; registros de enfermedades de etiología infecciosa, con notificación obligatoria en los que se basan los sistemas de vigilancia epidemiológica. Cuando la información que el epidemiólogo necesita no está disponible a través de algunas de las fuentes antes mencionadas, se debe recurrir a diferentes 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. Así entonces 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),(Martínez *et al.*, 1997),(Rothman y Greenland, 1998),(Särndal *et al.*, 1992),(Silva, 2000),(Vittinghoff *et al.*, 2005),(Clayton y Hills, 1993). Toda la información recolectada, se puede sistematizar y clasificar en forma protocolizada a través de la CIE10 (décima versión de la Clasificación Internacional de las Enfermedades) (www.who.int/classifications/en/).

Esta sistematización de las diferentes fuentes de información permitiría en rigor construir un sistema de información (Oriol, 1997), que es 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.

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 y 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, la administración, apoyándose a su vez en la estadística para construir una gran diversidad de indicadores que deben cumplir algunas características tal como presenta por ejemplo el autor Silva (Silva, 1998)

- validez de aspecto
- validez de contenido
- validez de criterio o concurrencia
- capacidad predictiva
- fiabilidad o reproducibilidad

Todos estos conceptos son válidos para la vigilancia en salud pública 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), y que se usan en encuestas de base poblacional como por ejemplo en Brasil (Ministério da Saúde, 2003)

- **CPO**: Indice de piezas con caries, perdidas y obturadas. Es un índice univariado que sirve para marcar historia de enfermedad (revertida o aún sin tratar)

$$CPO = \sum_1^n C_i + P_i + Oi$$

- **CPI**: Indice Periodontal Comunitario
- **SIC**: Indice significativo de caries (que corresponde al promedio de los valores de CPO que son mayores al **cuantil** 70 % de la distribución del CPO.(Brathall, 2000)

Hasta la fecha los antecedentes de estudios a nivel nacional consideran la aplicación de algunos de esos índices.(Beca *et al.*, 1996),(Bianco *et al.*, 1997). Sin embargo estos indicadores son inadecuados a la hora de ser usados en la toma de decisiones para generar acciones concretas con el propósito de mejorar la salud oral de la población. Esto se debe a que no considera toda la información necesaria a ser usada 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 georeferenciar la información y 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; el usar técnicas estadísticas multivariantes recientes pueden 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 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 solamente en salud oral, por lo menos en nuestro país.

Otro aspecto a ser considerado 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 al provenir de diferentes fuentes de información. En Uruguay se pueden destacar 2 tipos de fuentes en salud oral y que tienen niveles de existencia en el tiempo y desarrollo o profundidad muy heterogéneos.

- Registros de atención en el primer nivel de atención en el ámbito de atención del sector público y privado
- Encuestas de base poblacional (Son muy pocas las que se han desarrollado con alcance nacional) y donde más se ha estudiado, es en la población de escolares.(Lorenzo-Erro, 2003)

Estas 2 grandes fuentes de información hacen que en realidad la población sobre la que se pretende hacer vigilancia en salud oral puede ser dividida en 3 grupos de personas

- población de las personas que consultan
- población de las personas que tienen cobertura de algún tipo y no consultan
- población a nivel general

Teniendo en cuenta 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 necesariamente deban trabajar

sobre marcos muestrales con multiplicidad, con diferentes probabilidades de inclusión y con sesgos de selección,(Ministerio de Salud Pública, 2009),(Nithila *et al.*, 1998),(Thomas y Weber, 2001),(Särndal *et al.*, 1992),(Kim y Dailey, 2008),(Oakes y Kaufman, 2006),(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 las enfermedades no transmisibles (ENT) de la que forman parte las patologías orales. Las enfermedades transmisibles (ET) como por ejemplo los virus respiratorios, la Hepatitis A, el VIH, el Dengue, 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 "temporoespacial" 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, justifican esta línea de trabajo, en la que se plantean los siguientes objetivos.

C.2. Objetivos

Teniendo en cuenta los antecedentes antes planteados con respecto a las fuentes de información en salud en general y en salud oral en particular, se propone 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.

1. Por un lado se construirá un conjunto de indicadores epidemiológicos combinando los clásicos CPO, SIC, CPI, con los que surgen de considerar la estructurada multivariada de la información independientemente de la fuente de datos, usando técnicas estadísticas de no uso habitual en investigación epidemiológica en nuestro país (ver sección C.3.4).
2. Construir un conjunto de indicadores epidemiológicos alternativos a los clásicos CPO, SIC, usando la misma información pero considerándola de forma diferente, que se pueden resumir como tasas a ser pronosticadas en función de características epidemiológicas de las personas usando métodos estadísticos de aprendizaje supervisado (ver sección C.3.5).

3. Desarrollar indicadores que den cuenta de la distribución temporoespacial de las patologías orales en estudio (ver sección C.3.6).

C.3. Metodología

Las diferentes etapas a recorrer en el proceso de construcción y aplicabilidad de los indicadores propuestos consiste, en una primera revisión de la literatura especializada (en este caso salud pública, salud oral, epidemiología) en revistas, bases de datos como Pubmed y Bireme y una posterior sistematización de la mismas, con la propuesta que se presenta a continuación.

C.3.1. Estrategia de búsqueda bibliográfica

Los términos MESH en ingés para búsqueda podrían ser

- Health-chronic disease surveillance system (sistemas de vigilancia en enfermedades crónicas)
- DMFT (Decayed, Missing due to caries, and Filled Teeth)(CPO)
- CPI (Community Periodontal Index)
- Dental health surveys (Encuestas de Salud Oral)
- Health status indicators (Indicadores de Salud)
- Risk factors (Factores de riesgo)

Otra estrategia de búsqueda cruzada será desde las técnicas estadísticas que están en los 3 tipos de indicadores propuestos, hacia revistas especializadas en Salud Oral.

Una propuesta inicial (pero no exhaustiva) es

Tipo de Fuente	Nombre	Dirección electrónica
Electrónica	Pubmed	www.ncbi.nlm.nih.gov/pubmed
Electrónica	Scielo	www.scielo.org
Electrónica	Lilacs	bases.bvsalud.org
Electrónica	Oxford Journals	www.oxfordjournals.org
Electrónica	Ebsco	search.ebscohost.com
Rev. especializadas	Community Dentistry and Oral Epidemiology	http://onlinelibrary.wiley.com
en salud oral	J.of Public Health Dentistry	http://onlinelibrary.wiley.com
	J. of Periodontal Research	http://onlinelibrary.wiley.com
Rev. especializadas	American J. of Epidemiology	www.oxfordjournals.org
en epidemiología	Int. J. of Epidemiology	www.oxfordjournals.org
y / o salud pública	Biostatistics	www.oxfordjournals.org
	Epireview	www.publish.csiro.au

Tabla C.1: Tipo de fuente bibliográfica a buscar

La aplicación se propone hacer sobre los 2 tipos de fuentes de información manejadas en la sección C.1 para los que se hará una recopilación exhaustiva pero que incluye por lo menos las que se describen a continuación. Los diferentes algoritmos de cálculos se explican en las secciones C.3.4 y C.3.4.

C.3.2. Encuestas de base poblacional

Se trabajará con 'Primer Relevamiento Nacional en Salud Bucal en la Población Adolescente y Adulta del Uruguay' actualmente en curso y que lleva adelante la Facultad de Odontología de la Universidad de la República. A su vez se evaluará el módulo de salud oral que se incluirá por primera vez en la 2da Encuesta de Factores de Riesgo STEPS (World Health Organization, 2006) a realizarse en 2012, que depende del MSP (las patologías de salud oral son un subconjunto de las enfermedades no transmisibles (ENT) y comparten los mismos factores de riesgos, que son básicamente hábitos de vida nocivos modificables mediante programas preventivos adecuados).

C.3.3. Sistemas de registros

Por un lado se evaluarán los diferentes registros que se articulan en el primer nivel de atención del sector de salud del ámbito privado y público (incluyendo la atención que se hace a nivel de Salud Pública, o en ámbitos municipales).

Se evaluará también la información que se sistematiza a través del sistema Rediente (http://www.rediente.org).

Rediente según consta en su página web "*es un sistema que facilita el registro y la evaluación de la salud bucal. El uso de rediente inicia con el llenado de la historia clínica al lado del paciente, que se lleva su propio carnet rediente y continua con el ingreso de datos a la base de datos. Esta base 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, la administración de servicios de salud (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*". En función de la descripción que antecede es muy importante tomarla como ejemplo sistema de información estadístico-epidemiológico donde aplicar eventualmente los indicadores propuestos.

Sobre esta sistematización de los sistemas de registros se podrá construir un **marco muestral**, sobre el cual implementar mediante diferentes diseños muestrales adecuados el monitoreo o vigilancia. Esta opción es la que puede hacer realizable la vigilancia 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 actualmente impulsando el equipo técnico de vigilancia de enfermedades no transmisibles ENT, del MSP, estableciendo vigilancia por muestreo en las emergencias hospitalarias.

Una vez hecho el diagnóstico de las fuentes de información se elaborarán una serie de indicadores detallados a continuación para evaluar luego en una tercera etapa como aplicarlos en cada fuente de información, para finalmente analizar la sustentabilidad del sistema de vigilancia.

C.3.4. Indicadores combinados

Se usarán técnicas **descriptivas multivariantes** como lo son el Análisis Factorial y el Análisis de Grupos o Conglomerados, técnicas que no son muy usadas en el ámbito de la epidemiología. El Análisis Factorial permitiría descartar las variables que son menos 'importantes' (usando Análisis de Componentes Principales o Análisis Factorial de Correspondencias). Sobre el resultado de estas técnicas que consisten en la creación de **Índices** (combinaciones lineales de las variables originales, donde se le da un peso o importancia diferente a cada una) se puede a través del Análisis de Grupos o Cluster, crear **tipologías** o grupos de poblaciones con perfiles epidemiológicos bien diferenciados de acuerdo a las patologías y los factores de riesgos asociados. (Silva, 1995),(Hosmer y Lemeshow, 1988),(Rencher, 2002),(Agresti, 2005),(Blanco, 2006),(Cuadras, 2007),(Rao y Miller, 2008)

Para el caso de los datos para los que se agrega la dimensión temporo-espacial (datos longitudinales, y/o por región geogáfica), se aplicará el Análisis Multiway (que consiste en yuxtaponer las tablas de datos multivariadas de individuos por variables) de manera de incorporar una tercera dimensión de tiempo y/o espacio.

Se presentará un nuevo enfoque (poco usado en la epidemiología de salud oral) basado en técnicas estadísticas de **aprendizaje supervisado**, como son los métodos CART (Classification and Regression Trees). Esta metodología permite la construcción de modelos basados en técnicas no paramétricas, lo que supone muchas menos restricciones de distribuciones de probabilidad en las variables consideradas, permitiendo encontrar las variables que mejor discriminan el comportamiento de una variable de respuesta o dependiente de tipo categórica; la aplicación immediata es sobre variables que clasifican en ausencia o presencia de una patología, en diferentes niveles de patología (maloclusión, enfermedad periodontal) (ver ejemplo de 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 facilmente usables e interpretables por los no especialistas en estadística.(Breiman *et al.*, 1984),(Abernathy *et al.*, 1987)

C.3.5. Indicadores alternativos

Se proponen nuevos indicadores construídos a partir del CPO, para lo que se utilizarán distintos tipos de *modelos lineales generalizados* (GLM) adaptados para modelar proporciones - familia Binomial (logit y probit) y modelado de variables de conteo - familia Poisson.

En primer lugar se utilizarán modelos de regresión beta (mediante una reparametrización adecuada), propuesta por Ferrari y Cribari-Neto (2004) para modelar variables de respuesta continua a valores en el intervalo $(0, 1)$. Este modelo tiene gran flexibilidad para adaptarse a distribuciones asimétricas y la posibilidad de interpretar las estimaciones en términos de la variable de interés y no de una transformación de la misma.(Cribari-Neto y Zeileis, 2010)
La idea 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 de rango reducido adaptados para variables de respuestas no gaussianas. (Yee y Hastie, 2003),(Yee, 2010)

En ambos casos las variables explicativas tendrán que ver con aspectos sociodemográficos individuales (edad, sexo, educación), de contexto (región, barrio) y con la historia de salud bucal de cada individuo (el motivo de su consulta, la cantidad de prótesis, el tiempo sin concurrir al dentista, etc).

También se podrán usar, y ver como se pueden adaptar, para poder discriminar el comportamiento en la población de los diferentes parámetros a evaluar en la salud oral, índices basados en la teoría de la información de Shanon, ampliamente usado en economía y demografía económica, para medir la distribución del ingreso. Entre estos se pueden encontrar los Indices de Theil, Indices de Atkinson, Indice de Hoover, que pueden ser vistos como índices de entropía. Otros índices para evaluar otros aspectos de los perfiles epidemiológicos en Salud Oral pueden adaptarse desde disciplinas como la ecología, donde se aplican índices de abundancia de especies, de riqueza de especies.(Shannon

y Weaver, 1949)

C.3.6. Indicadores temporo-espaciales

Los indicadores que se proponen aplicar son de 3 tipos y buscan tener un manejo en el tiempo y espacio de los fenómenos morbosos de la salud oral, tal como se trabaja en la epidemiología de las *enfermedades transmisibles*.

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 ciudades, barrios o conjuntos de manzanas).
2. **Agregaciones temporales**- Para la detección de agregaciones temporales se utilizarán indicadores buscando 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. Algunos de los métodos pueden ser Método Chen, Método Sets, Método Scan, Método Texas, Método Poisson, Método Cusum, Método Ederer-Myers-Mantel, Índice de Tango.
3. **Agregaciones Temporoespaciales**- En este caso se buscarán encontrar si existen clusters en el espacio, por un lado y en el tiempo, en forma simultánea, 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 considerar que la localización de un caso depende de la localización del caso que lo precede. Una metodología adecuada para modelizar ambos componentes es a través de un modelo estadístico que fuera temporalmente dinámico y espacialmente descriptivo (Método Knox, Método Kulldorf, Test de Mantel, Test de Baker, Test k-NN de Jacquez, Test de Diggle)

Estos indicadores pueden considerarse como herramientas epidemiológicas de cálculo relativamente sencillo y que están pensadas para la descripción.

Además de esta batería de indicadores que permitiría guiar una acción de intervención para modificar la situación epidemiológica fuera de la común, se intentará 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 al tener que manejar procesos estocásticos, con los cuales hay que estimar **variogramas** (funciones de autocorrelación espacial) sobre los cuales hacer Interpolación por **método de krigging** sobre los puntos de muestreo georeferenciados. A partir de esta interpolación direccional se pueden obtener curvas de nivel que se pueden representar gráficamente en mapas. Para evaluar su aplicabilidad se propondrán diferentes herramientas de GIS que permiten generar esos indicadores y cuales son los algoritmos que manejan y los supuestos que están detrás de éstos.(Waller y Gotway, 2004),(Lawson y Kleinman, 2005),(Bivand *et al.*, 2008),(Pfeiffer *et al.*, 2008),(Lawson, 2009),(Tango, 2010)

En una primera instancia se trabajará con el sistema **R** (R Core Team, 2010) 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. Actualmente es la herramienta mas usada en el campo de la estadística aplicada a la epidemiología en la enseñanza e investigación en las universidades más importantes. Es un lenguaje matricial basado en subrutinas (llamadas librerías) desarrolladas por la propia comunidad usuaria de R.

C.4. Plan de Trabajo

En primera instancia se hará la revisión de la literatura especializada siguiendo la estrategia metodológica planteada en C.3. Posteriormente se hará una revisión de las fuentes de información disponibles. Sobre las diferentes fuentes identificadas y disponibles se aplicarán los diferentes grupos de indicadores propuestos en C.3.4, 3.4 y C.3.6. Se propondrán a su vez diferentes formas de como combinar los indicadores provenientes de diferentes tipos de fuentes de información (registros vs encuestas poblacionales).

Además de los objetivos principales planteados en la sección C.2 se evaluarán otros 2 aspectos fundamentales como son

- el grado de completitud de la información disponible en las diferentes fuentes de información relevadas, necesarias para estos nuevos indicadores.
- grado de aplicabilidad de este nuevo sistema de vigilancia dependiente de los nuevos indicadores propuestos (a través de un análisis de fortalezas, oportunidades, debilidades, amenazas *FODA*).

C.5. Resultados esperados

Se espera poder desarrollar una serie de metodologías de análisis de la información en salud oral, indispensables en la vigilancia epidemiológica y que consisten en identificar las componentes del sistema de vigilancia.

1. Sistematización en forma exhaustiva de las diferentes fuentes de información disponibles en salud oral para Uruguay
2. Creación de grupos de indicadores para evaluación permanente de la situación epidemiológica, cobertura, acceso y gestión en salud oral
3. Protocolizar la forma de cálculo y la información necesaria para cada grupo de indicadores
4. Reseña de los diferentes programas de cálculo necesarios para le estimación de los indicadores propuestos

Con respecto a los aspectos vinculados a la completitud de la información es necesario tener claro que la forma de hacerlo rigurosamente es con la ayuda del herramental de la administración (Pineault, 1988) y de la planificación estratégica. La creación de nuevos indicadores epidemiológicos necesariamente implicará la capacitación en su uso, en como comunicarlos, creación de una serie de documentos similares a las guías clínicas (que protocolizan el manejo de la información y las decisiones a tomar). Se creará de esta manera una serie de nuevas actividades que involucrarán recursos logísticos y humanos, por lo cual se propone incorporar al sistema de vigilancia herramientas como la matriz programática (Álvarez-Vaz y Dibarboure, 2008), que entre otras cosas estaría compuesta por una serie de indicadores del tipo

- Indicadores de estructura - Una de las fuentes de información del sistema de vigilancia estaría compuesto por una red de puntos o nodos de generación de información (puestos centinelas) que serían los especialistas en salud oral en los diferentes niveles de atención. Sobre esos nodos se construirán indicadores relativos a la accesibilidad en un sentido bidireccional (hacia los nodos y desde los nodos)
- Indicadores de Proceso - Para esta dimensión del Sistema de Vigilancia se construirán indicadores para evaluar la forma en que se llevan a cabo las actividades de registro de la información, los tiempos de espera en hacer la notificación, el porcentaje de nodos que siguen los procedimientos de rutina establecidos para el reporte de la información.
- Indicadores de Resultados -Se construirán indicadores para medir los efectos a corto plazo y los impactos de largo plazo en el uso de los nuevos indicadores epidemiológicos, creados tal como se planteó en la sección C.3.5, C.3.4 y C.3.6, por parte de los potenciales usuarios del sistema de vigilancia. Se deberán evaluar los resultados concluídos y aplicados, los resultados concluídos y no aplicados que a su vez pueden servir para la génesis de proyectos de innovación.

El producto de este conjunto de investigaciones puede transformarse en un apoyo muy importante para el actual Programa de Salud Bucal del MSP, ya que que en el año 2006 el Programa de Salud Bucal a cargo, estableció como parte de sus líneamientos estratégicos, el n^o3 en particular, lo siguiente:

Promover la implementación de un sistema de vigilancia de la salud bucal en la población.

1. *-Implementar un Sistema Unico de Registro Nacional de Salud Bucal.*
2. *-Implementar la Encuesta Nacional de Salud Bucal.*
3. *-Promover la formación de un equipo técnico permanente de referencia en Vigilancia a la salud (DIGESA-MSP).*
4. *-Consolidar la vigilancia epidemiológica correspondiente al Programa de Fluoración de la Sal.*

El lineamiento estratégico solo propone que debiera de hacerse pero no como, por lo cual esta propuesta de vigilancia (compuesta por las fuentes de información, varios grupos de indicadores y la pertinencia y aplicabilidad de cada uno) es fundamental y podría ser tomada por el Programa de Salud

Bucal, con ayuda de la Sala de Situación en Salud del MSP, dependiente de Epidemiología, aprovechando la descentralización que se está desarrollando, a través de las Unidades Departamentales de Epidemiología (UDE).