CREACIÓN DE INDICADORES ALTERNATIVOS PARA LA VIGILANCIA EN SALUD ORAL, MEDIANTE REGRESIÓN BETA

Ramón Alvarez[1]; Elena Vernazza[1]

RESUMEN

En el ámbito de la Salud Pública, en particular en los sistemas de vigilancia epidemiológica pueden existir limitaciones en los indicadores generalmente utilizados, 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.

Se propone comenzar a construir un conjunto de indicadores alternativos y complementarios a los que ya existen en salud oral. Se reformularán algunos de los índices recomendados por la Organización Mundial de la Salud en el estudio de la salud oral de la población como es el CPO (número total de piezas cariadas, perdidas y obturadas).

La aplicación se hace con información proveniente del relevamiento en necesidades de tratamiento y demanda de servicios de salud bucal, de la población de Trabajo por Uruguay (TPU) a agosto de 2007.

Se presentan las limitaciones que tiene trabajar con un *índice agregado univariado*. Como alternativa se propone, por un lado, estimar los componentes del (CPO) mediante modelos para variables de conteo como es la *Regresión de Poisson* y,por otro lado,modelar proporciones construídas a partir del CPO, utilizando modelos de *Regresión Beta*, mediante una reparametrización adecuada, propuesta por Ferrari y Cribari-Neto (2004) para evaluar variables de respuesta continua a valores en el intervalo $(0, 1)$. Se presentan avances comparando ambos enfoques, usando en este caso variables explicativas que tienen que ver con aspectos sociodemográficos individuales (edad, sexo, educación), de contexto (región, barrio) y con la historia de salud bucal (motivo de consulta, cantidad de prótesis, el tiempo sin concurrir al dentista, etc). Se proponen modificaciones a los modelos al existir problemas con la información disponible.

Palabras clave: *CPO,Regresión Beta,Salud Oral,variables de conteo,vigilancia epidemiológica*

[1]INSTITUTO DE ESTADÍSTICA

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 esto 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 (Ramis 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 (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. 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 (Ramis Oriol 1997), que es uno de los pilares necesarios para la verdadera vigilancia epidemiológica (VE), permitirá luego hacer intervenciones en salud, para poder modificar la situación.

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. 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.

1.1. Diferentes aspectos que deben ser medidos en Salud Oral

A nivel de la salud oral 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)

Tal como se presentó en la introducción existen diferentes fuentes de datos donde debe medirse en Salud Oral, 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 m?s pequeñas pero debidamente identificadas.

Sin intentar ser exhaustivos, existen varios indicadores de las diferentes dimensiones que se determinan a nivel individual en salud oral y que pueden ser considerados a nivel colectivo desde una perspectiva epidemiológica. Se consideran los indicadores **ceo, CPO, ICDASII** de los que más adelante se dan detalles, también teniendo en cuenta los que se construyen a partir de la CIE10 capítulo **K** (es importante destacar que la CIE10 es para la salud lo que puede ser la Clasificacón CIU, para clasificar empresas).

Por lo tanto, para este trabajo se presentan los índices que dan cuenta del estado de las piezas dentales, para lo cual es necesario hacer definiciones y establecer una nomenclatura de los diferentes unidades de observación.

- i individuo, j diente, k cuadrante, l superficie, g grupo o subpoblación (podríamos 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 $s^g_{i,j,l}$

1.2. Antecedentes

Para entender los aspectos antes mencionados, en este trabajo se consideran aquellos 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 evaluúa en forma más detallada y gradualista, estados o niveles de enfermedad.

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 la *caries* de un conjunto de individuos. Los valores bajos de (CPO) indican un buen 'status' de salud oral, mostrando que las piezas dentales tienen poca hstoria 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 lo que 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:

$$CPO^g_{i,j,k} = \sum_j^n C^g_{i,j,k} + \sum_j^n P^g_{i,j,k} + \sum_j^n O^g_{i,j,k} \tag{1}$$

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, o 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 oral, 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.

2. Metodología

¿Qué tipos de modelos usar? Desde el punto de vista **epidemiológico** es deseable recurrir a modelos *parsimoniosos* pero adecuados, en el sentido de que sean fáciles de estimar, de sencillo uso (que la información esté disponible) y que los salubristas lo entiendan, lo adopten y lo difundan entre sus colegas odontólogos.

Las alternativas pueden ser:

1. $CPO_i = f(X_{ij})$ (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 **(VGAM)**(Yee & Hastie 2003),(Yee 2010) **tipo2**

En este trabajo solo se consideran modelos del tipo 1. Lo ideal sería un modelo de regresión lineal **(MRL)**, pero lamentablemente los MRL exigen supuestos necesarios que no se cumplen, ya que la variable de respuesta (CPO) o cualquiera de sus componentes como variables aleatorias tiene recorrido discreto, por lo cual es necesario usar otros que se detallan a continuación.

2.1. Modelos de conteo

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

Tipo	Distribución	Descripción
MLG	Poisson	regresión Poisson estimado por máxima verosimilitud (MV)
		regresión "quasi-Poisson": ajustes para Inferencia
		regresión"Poisson ajustado ": ajustes para Inferencia
	BN	regresión BN :
	(B Negativa)	estimado por MV incluye parámetro de forma

Cuadro 1: Diferentes alternativas a los modelos de Conteo. Los MLG usan la misma función lineal para la media ($\log(\mu) = x^\top \beta$) .

2.2. Características de estos tipos de modelos de conteo

Modelo de Poisson (MP)

$$f(y;\mu) \quad = \quad \frac{\exp(-\mu)\cdot\mu^y}{y!}, \tag{2}$$

Cuando existe sobredispersión tenemos un modelo Binomial Negativo (BN)

$$f(y;\mu,\theta) = \int_0^\infty \frac{\exp(-\mu)\cdot\mu^y}{y!} f_\Gamma(\mu)d\mu \tag{3}$$

$$= \quad \frac{\Gamma(y+\theta)}{\Gamma(\theta)\cdot y!}\cdot\frac{\mu^y\cdot\theta^\theta}{(\mu+\theta)^{y+\theta}} \tag{4}$$

con media μ y parámetro de forma θ

2.3. 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 CPO o algunos de sus componentes relativizados contra diferentes totales: 32 (máximo número de piezas, número de piezas presente,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 *inferencias*, considerando que:

- Las proporciones a estimar están en el rango $(0,1)$
- Otras distribuciones 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 trabajo se relativizan los componentes del CPO convirtiéndolos en tasas o proporciones usando los 3 componentes del CPO del siguiente modo

1. $\frac{\sum O_i}{\sum O_i+\sum C_i}$ nivel de cobertura de la enfermedad previa a la entrada al programa (*prop1*)
2. $\frac{\sum C_i}{\sum O_i+\sum C_i}$ indicador de estadio de la enfermedad en el momento actual (*prop2*)
3. $\frac{\sum S_i}{\sum O_i+\sum C_i+\sum P_i+\sum S_i}$ indicador de salud en caries en el momento actual (*prop3*)
4. $\frac{\sum P_i}{\sum O_i+\sum C_i+\sum P_i+\sum S_i}$ indicador de necesidad de prótesis en el momento actual (*prop4*)

2.4. Formulación del modelo de probabilidad BETA

De los diferentes modelos que se podrían utilizar para modelar tasas o 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).

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0<y<1, \tag{5}$$

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

  Se puede escribir $y \sim \mathscr{B}(\mu,\phi)$. 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.

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0<y<1, \tag{6}$$

En la figura 1 se pueden ver diferentes ejemplos de distribución BETA al cambiar los parámetros de dispersión o precisión.

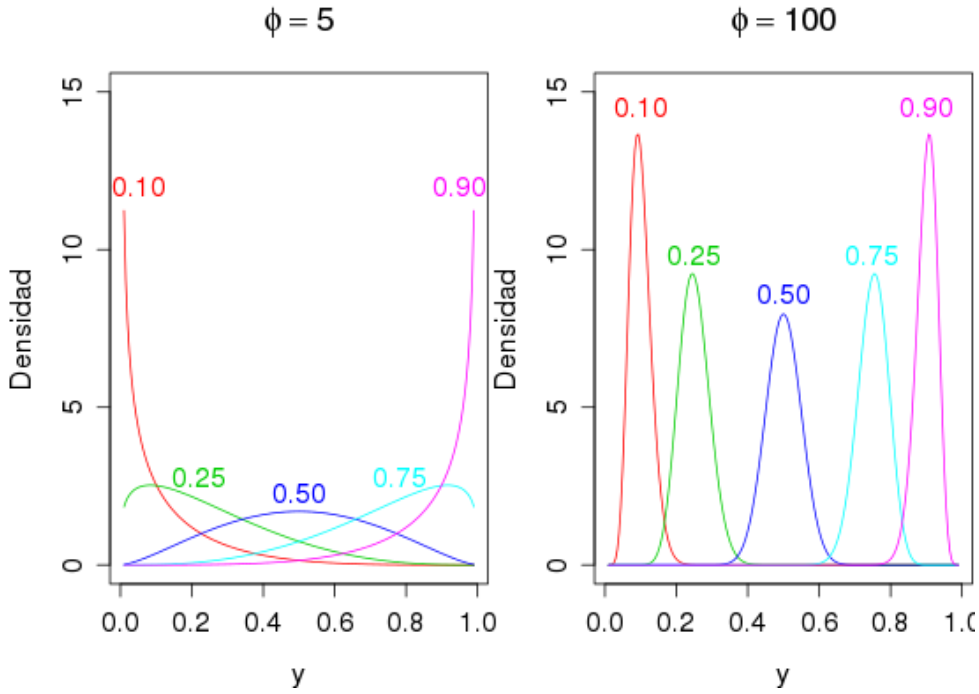

Figura 1: densidades **BETA** en el intervalo $(0,1)$ al cambiar μ y ϕ

La formulación del modelo predictivo puede ser hecha a partir de los trabajos de (Kieschnick & McCullough 2003), (Salinas-Rodríguez, Manrique-Espinoza & Sosa-Rubí 2009) donde

- Si se tiene $y_1,\ldots,y_n$ una muestra aleatoria tal que $y_i \sim \mathscr{B}(\mu_i,\phi)$, $i=1,\ldots,n$.
- 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).

$$g(\mu_i) = x_i^\top \beta = \eta_i, \tag{7}$$

El modelado de la dispersión se peude hacer a través de

$$VAR(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi} = \frac{g^{-1}(x_i^\top\beta)[1-g^{-1}(x_i^\top\beta)]}{1+\phi}. \tag{8}$$

$$\begin{aligned} \ell_i(\mu_i,\phi) &= \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) + (\mu_i\phi-1)\log y_i \\ &\quad +\{(1-\mu_i)\phi-1\}\log(1-y_i). \end{aligned} \tag{9}$$

La funciones de enlaces (link) para (MRB) 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 $y_i \sim \mathscr{B}(\mu_i, \phi_i)$, $i = 1, \ldots, n$, y

$$g_1(\mu_i) = \eta_{1i} = x_i^\top \beta, \tag{10}$$

$$g_2(\phi_i) = \eta_{2i} = z_i^\top \gamma, \tag{11}$$

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 9.

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 tasa 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 & 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.

3. Aplicación

La aplicación de la metodología antes presentada es para una Encuesta para Trabajo por Uruguay (TPU) en el marco del Plan Nacional de Atención a la Emergencia Social (PANES), desarrollado a partir de 2005, con el principal programa para atender la situación de la población adulta en situación de extrema pobreza. En el PANES se desarrollaron varios programas de asistencia, como por ejemplo el TPU que daba trabajo a los beneficiarios, brindándole a su vez asistencia en salud bucal, a través de un convenio con la IMM, la Facultad de Odontología, asistentes sociales y una red de ONGs. De esta manera, se realizó un relevamiento epidemiológico, de necesidades de tratamiento y demanda de servicios de salud bucal para una muestra de la población de TPU del 2007, donde participó la Cátedra de Odontología Social de la Facultad de Odontología de la Udelar.

La población de estudio se compone de 1185 personas mayores de 18 años correspondiente al tercer llamado de TPU en Montevideo. Se obtuvo una muestra aleatoria mediante un muestreo multietápico (n=308) de individuos provenientes de tres regiones de Montevideo. La encuesta fue realizada por docentes de la Cátedra de Odontología Social e incluyó un examen clínico y un cuestionario personal que relevaba características sociodemográficas de los participantes.

Algunas de las variables estudiadas fueron la **subregión** de la ciudad de Montevideo, **edad**, **sexo**, **nivel educativo alcanzado**, fecha de la **última consulta** y **motivo** de consulta.

4. Resultados

Los resultados que se presentan, tanto para los gráficos como para los modelos, fueron hechos con el lenguaje *R* (R Core Team 2012), en particular para los modelos de conteo se

trabajó con la librería *MASS* (Venables & Ripley 2002). La distribución de los componentes del CPO es la que se muestra en la figura 2, donde se ve una gran asimetría para el componente de caries y para el componente de obturación, lo que habla de una población muy enferma (enfermedad actual y/o pasada), resultado esperable teniendo en cuenta los antecedentes antes mencionados en la sección 3

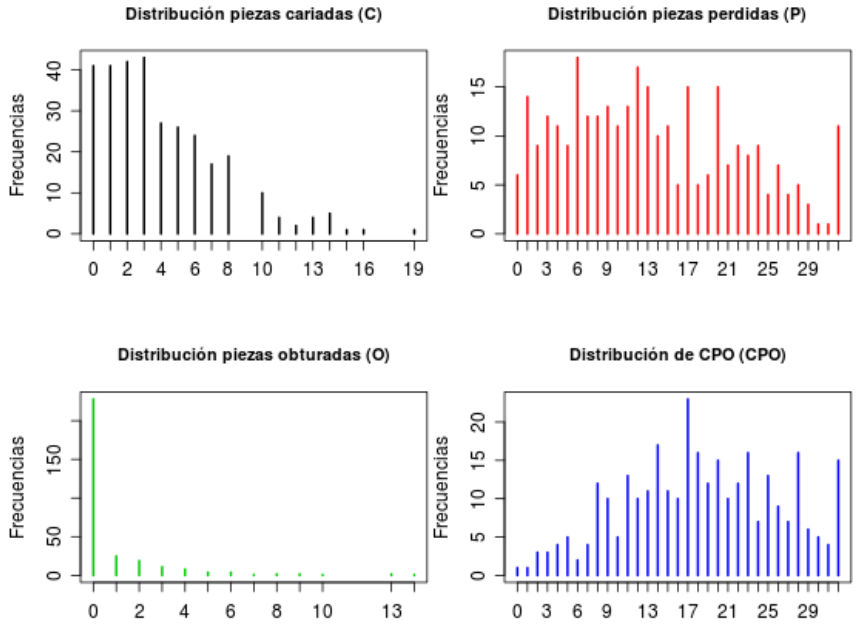

Figura 2: Densidades empíricas de los **componentes del CPO**

También es importante ver como cambian los resultados si trabajamos los conteos reexpresándolos en proporciones, eligiendo diferentes formas de normalizar tal como se presentaron en 2.3

La asociación que existe entre los componentes del CPO y las respectivas proporciones medida a través de la matriz de correlación lineal se muestra a continuación:

	C	P	O	CPO		prop2(C)	prop4(P)	prop1(O)	prop.CPO
C	1.000	-0.307	-0.150	0.079	prop2(C)	1.000	0.189	-1.000	-0.138
P	-0.307	1.000	-0.288	0.894	prop4(P)	0.189	1.000	-0.189	-0.879
O	-0.150	-0.288	1.000	-0.118	prop1(O)	-1.000	-0.189	1.000	0.138
CPO	0.079	0.894	-0.118	1.000	prop.CPO	-0.138	-0.879	0.138	1.000

Para la variable de conteo *CPO* transformada en tasa se tiene las siguientes medidas de resumen

Vemos como cambia el CPO expresado como tasa luego de hacer la transformación (Verkuilen & Smithson 2012).

Los modelos que se ajustaron son varios pero se presenta, como avance de resultados, los que evalúan el número de Caries (C), y que son de tipo *quassi-poisson*, ya que existe una sobredispersión muy grande, con una varianza mucho mayor a la media. Las variables regresoras fueron la edad, la región, sexo y CPO, y se presentan resultados de las que dieron significativas.

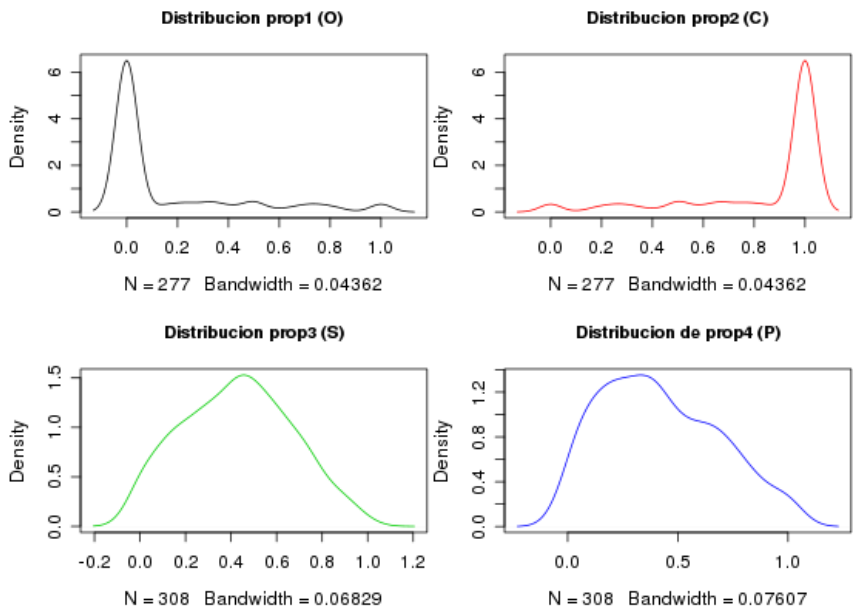

Figura 3: densidades de los **componentes del CPO** en proporciones

	CPO/32		prop.CPO
Mín	0.000	Mín	0.002
Q1	0.406	Q1	0.407
Mediana	0.562	Mediana	0.562
Media	0.572	Media	0.572
Q3	0.750	Q3	0.749
Máx	1.000	Máx	0.998

	Coeficiente	Error Std.	valor t	Pr(>|t|)
(Intercept)	1.1553	0.1228	9.41	0.0000
edad.recDe 35 a 44	-0.4280	0.1103	-3.88	0.0001
edad.recMayor a 45	-0.9667	0.1469	-6.58	0.0000
CPO	0.0279	0.0065	4.30	0.0000

Cuadro 2: Modelo de Poisson (MP) para *C* vs edad y CPO

```
glm(formula = C ~ edad.rec + CPO, family = quasipoisson, data = tpu)
Signif. codes:  0 '***' 0.001 '*' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for quasipoisson family taken to be 2.50656)
    Null deviance: 950.37  on 307  degrees of freedom
Residual deviance: 812.95  on 304  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 5
```

Para afinar las estimaciones, se estimó también un modelo de regresión para C de tipo Binomial-Negativo(BN)

	Coeficiente	Error Std.	valor t	Pr(>|t|)
(Intercept)	1.1106	0.1213	9.16	0.0000
edad.recDe 35 a 44	-0.4342	0.1113	-3.90	0.0001
edad.recMayor a 45	-0.9559	0.1348	-7.09	0.0000
CPO	0.0303	0.0066	4.59	0.0000

Cuadro 3: Modelo Binomial Negativo (MBN) para C vs edad y CPO

```
glm.nb(formula = C ~ edad.rec + CPO, data = tpu, init.theta = 2.37183983,link = log)
(Dispersion parameter for Negative Binomial(2.3718) family taken to be 1)
    Null deviance: 406.80  on 307  degrees of freedom
Residual deviance: 354.38  on 304  degrees of freedom
AIC: 1486.6
Number of Fisher Scoring iterations: 1
              Theta:  2.372
          Std. Err.:  0.337
 2 x log-likelihood:  -1476.624
```

Para los modelos que aparecen en los cuadros 2 y 3 las variables son significativas pero el ajuste es muy pobre, ya que tal como se vió en la figura 2, hay un exceso de 0 muy importantes, que los 2 tipos de modelos ajustados no logran captar ya que, por ejemplo, el número de personas que tienen $C = 0$ en la muestra son 61 mientras que ambos modelos solo reproducen 32, es decir, subestiman muchísimo. Por otra parte, en los valores extremos los modelos funcionan mal, ya que solo predicen hasta un máximo de 11 caries cuando hay personas con hasta 19 caries.

Se presenta como quedan finalmente construídas las proporciones a partir de los conteos absolutos y cual es su distribución

```
prop1<-O/(O+C)
prop2<-C/(O+C)
prop3<-S/(S+C+O+P)
prop4<-P/(S+C+O+P)
prop5<-O/(S+C+O+P)
prop6<-C/(S+C+O+P)
```

	prop1	prop2	prop5	prop6
Mínimo	0.000	0.000	0.000	0.000
Q_1	0.0000	0.8000	0.00000	0.03125
Mediana	0.0000	1	0.00000	0.09375
Media	0.1513	0.8487	0.02668	0.12652
Q_3	0.2000	1	0.03125	0.18750
Máximo	1	1	0.43750	0.59375
NA's	31	31		

Para evaluar si se puede obtener una mejora, se presenta el modelo de regresión de C transformado en tasa (MRB), donde previamente se redefinieron las tasas o proporciones, normalizando tal como se ve en la figura que sigue, lo que para el caso de $prop1 = \frac{O}{O+C}$ y $prop2 = \frac{C}{O+C}$ introduce un problema extra al tener 31 individuos con datos faltantes para esas 2 proporciones. Analizando más en detalle porqué se produce esto, se ve que son 31 personas (casi un 10 %) de la muestra que se caracterizan por sólo tener piezas perdidas componente (P), lo que hace que los cocientes $prop1 = \frac{O}{O+C}$ y $prop2 = \frac{C}{O+C}$ no estén definidos. Teniendo en cuenta este detalle, se analiza si ese 10 % es una sub-muestra aleatoria de la muestra original o por el contrario tiene un patrón definido. Como una de las variables que se usan como regresoras es la edad, se analiza si existe asociación entre esa característica y el hecho de tener solamente piezas perdidas o no, y se ve que sí hay un patrón claro donde la distribución marginal de la edad cambia notoriamente, lo que es entendible ya que es natural que (P) aumente con la edad.

P. obturadas	De 18 a 34	De 35 a 44	Mayor a 45	$valor - p$
0	54.80	24.40	20.80	
≥ 0	50.00	37.50	12.50	
Sin C o O	29.00	19.40	51.60	8.594e-05
P. cariadas	De 18 a 34	De 35 a 44	Mayor a 45	
0	30.00	50.00	20.0	
≥ 0	54.30	27.30	18.40	
Sin C o O	29.00	19.40	51.60	0.0002998

Cuadro 4: Distribución % de Personas por edad para (C) y (O)

Teniendo en cuenta que la $prop6 = \frac{C}{(S+C+O+P)}$ está definida sobre el total, se ajusta la proporción de caries con un modelo (MRB)

	Coeficiente	Error Std.	valor t	Pr(>|t|)
(Intercept)	1.1106	0.1213	9.16	0.0000
edad.recDe 35 a 44	-0.4342	0.1113	-3.90	0.0001
edad.recMayor a 45	-0.9559	0.1348	-7.09	0.0000
CPO	0.0303	0.0066	4.59	0.0000
	Coeficiente	Error Std.	valor z	Pr(>|z|)
De Dispersión(ϕ)	7.2103	0.6104	11.81	$< 2e-16$

Cuadro 5: Modelo De Regresión Beta (MRB) para C vs edad y CPO

```
betareg(formula = prop.trans6 ~ edad.rec + CPO, data = tpu)
Type of estimator: ML (maximum likelihood)
Log-likelihood: 353.1 on 5 Df
Pseudo R-squared: 0.09512
Number of iterations: 18 (BFGS) + 1 (Fisher scoring)
```

5. Conclusiones

¿Cómo seguir? Hasta el momento los resultados dan, tanto para los modelos de conteo (MC) como para el (MRB), significativos pero el ajuste es malo, por lo cual hay que probar con otras variables explicativas. Teniendo en cuenta los artículos estudiados, sobre todo para los MRB, las variables explicativas usadas, eran de tipo cuantitativos y algunas cualitativas, lo que puede ser una explicación del ajuste tan pobre. Además, para el caso de la reparametrización que se usa para el MRB donde la variable de respuesta se transforma en una proporción, aparece una dificultad extra, que podemos llamar un 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 $\frac{1}{32}$. Por otra parte, para este ejemplo en particular, la población considerada es muy especial, con una distribución del *CPO* con una 'simetría' que sólo se explica al ser ésta una población muy enferma. A nivel general se esperaría un comportamiento diferente con una **concentración** del *CPO* en valores más bajos y con una distribución más asimétrica.

Para finalizar, antes de presentar futuros pasos, hay otras situaciones donde es posible aplicar en particular el modelo **MRB**

- En el ámbito educativo o desempeño académico, para explicar la proporción de aciertos en pruebas o tests.
- En economía, los factores que influyen en la proporción de hogares que se suscriben a la televisión por cable.
- En epidemiología, para estudiar la cantidad de personas en los hogares que padecen una misma patología.
- En Economía de la salud, en particular salud pública, para estudiar el nivel de cobertura que tiene una población con beneficios de una canasta de prestaciones.

6. Futuros pasos

Existen varias alternativas metodológicas para trabajar con exceso de ceros en las variables de conteo

6.1. Modelos Hurdle (MH)

Los modelos Hurdle (Hurdle Models), que podrían considerarse como modelos con *obstáculos*, son modelos que combinan 2 procesos de conteo, uno para los 0 , con $f_{\text{zero}}(y;z,\gamma)$ (censurado por la derecha en $y=1$) y otro para conteos >0 $f_{\text{count}}(y;x,\beta)$ (truncado por la izquierda en $y=1$), que puede ser de tipo Poisson, geométrico o binomial negativo.

$$f_{\text{hurdle}}(y;x,z,\beta,\gamma)=\begin{cases} f_{\text{zero}}(0;z,\gamma) & \text{si } y=0, \\ (1-f_{\text{zero}}(0;z,\gamma))\cdot f_{\text{count}}(y;x,\beta)/(1-f_{\text{count}}(0;x,\beta)) & \text{si } y>0 \end{cases} \quad (12)$$

Los parámetros del modelo β, γ, y potenciales parámetros de dispersión θ (si f_{count} o f_{zero} o ambos con dendidad negativa binomial) se estiman por MVL, donde la especificación de la

verosimilitud tiene la ventaja de que los componentes del conteo y de hurdle pueden maximizarse en forma separada.

La correspondiente regresión sobre la media se da por la ecuación

$$\log(\mu_i) \quad = \quad x_i^\top \beta + \log(1 - f_{\text{zero}}(0; z_i, \gamma)) - \log(1 - f_{\text{count}}(0; x_i, \beta)), \tag{13}$$

6.2. Modelos con Exceso de Ceros (MEC)

Los Modelos con Excesos de ceros (de tipo Poisson (PEC), (BNEC) que permiten tener ceros) 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.

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 $f_{\text{count}}(y; x, \beta)$. La probabilidad de observar un conteo de 0 se incrementa con probabilidad $\pi = f_{\text{zero}}(0; z, \gamma)$

$$f_{\text{zeroinfl}}(y; x, z, \beta, \gamma) \quad = \quad f_{\text{zero}}(0; z, \gamma) \cdot I_{\{0\}}(y) \; + \; (1 - f_{\text{zero}}(0; z, \gamma)) \cdot f_{\text{count}}(y; x, \beta), \tag{14}$$

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

$$\mu_i \quad = \quad \pi_i \cdot 0 \; + \; (1 - \pi_i) \cdot \exp(x_i^\top \beta), \tag{15}$$

usando la función de enlace canónico.

De ambos tipos de modelo de conteo los autores de este trabajo consideran que los del tipo **hurdle** son más sencillos de interpretar ya que de alguna manera están modelando lo siguiente:

- Hay personas que no tienen Caries (si esta 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)

La idea entonces es poder evaluar el funcionamiento de los modelos que contemplan los excesos de 0 en el conteo y trabajar, a su vez, reexpresando los componentes del CPO en proporciones, a través de los (MRB), pero sabiendo que el problema de excesos de 0 persistirá con lo cual es necesario condiderar (MRB) adecuados a esta situación. Para esto hay que tomar en cuenta los desarrollos recientes de autores como (Ospina & Ferrari 2012),(Grün, Kosmidis & Zeileis 2011),(Zeileis, Kleiber & Jackman 2008) quienes plantean:

- Corrección y reducción del sesgo
- Estimación de mejores modelos mediante árboles de regresión (Combinación de los métodos CART y los (MRB))
- (MRB) con variable de clase latente que explican mejor la *heterogeneidad*

Por último, es importante dejar en claro que esta metodología en Uruguay no se ha utilizado en el campo de la salud oral hasta la fecha y que cuando se pueda replicar para otra encuesta de base poblacional donde la carga de enfermedad no sea tan grande y exista a su vez mas variabilidad, los autores esperan mejorar la performance.

7. Bibliografía

Cribari-Neto, F. & Zeileis, A. (2010), 'Beta regression in r', *Journal of Statistical Software* **34**(2), 1–24.
URL: *http://www.jstatsoft.org/v34/i02*

Grün, B., Kosmidis, I. & 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.
URL: *http://EconPapers.RePEc.org/RePEc:inn:wpaper:2011-22*

Kieschnick, R. & McCullough, B. D. (2003), 'Regression analysis of variates observed on (0, 1): percentages, proportions and fractions.', *Statistical Modelling: An International Journal* **3**(3), 193.

Mullahy, J. (1986), 'Specification and testing of some modified count data models', *Journal of Econometrics* **33**, 341–365.

Ospina, R. & Ferrari, S. L. (2012), 'A general class of zero-or-one inflated beta regression models', *Computational Statistics & Data Analysis* **Volume 56**(Issue 6), 1609â1623.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Ramis Oriol, J. (1997), *Salud Pública*, number Cap 8, Mc Graw Hill-Interamericana.

Salinas-Rodríguez, A., Manrique-Espinoza, B. & 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.
URL: *http://www.scielosp.org/scielo.php?script=sci$_a$rttext pid = S0036 − 36342009000500007nrm = iso*

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*

Verkuilen, J. & 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.
URL: *http://jeb.sagepub.com/content/37/1/82.abstract*

Yee, T. & 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**, 1–34.
URL: *http://www.jstatsoft.org/v32/i10/*

Zeileis, A. (2006), 'Object-oriented computation of sandwich estimators', *Journal of Statistical Software* **16**(9), 1–16.
URL: *http://www.jstatsoft.org/v16/i09/*

Zeileis, A. & Kleiber, C. (2008), *AER: Applied Econometrics with R*. R package version 0.9-0.
URL: *http://CRAN.R-project.org/package=AER*

Zeileis, A., Kleiber, C. & Jackman, S. (2008), 'Regression models for count data in r', *Journal of Statistical Software* **27**(8), 1–25.
URL: *http://www.jstatsoft.org/v27/i08*