CIMAT

Centro de Investigación en Matemáticas,A.C

Trabajo Recepcional de Tesis

Comparación de algunos diseños πps con el diseño Binomial negativo

que para obtener el grado de

Mastría en Ciencias en Estadística Oficial presenta:

Ana Miriam Romo Anaya

Dr. Víctor Alfredo Bustos y de la Tijera

(Director de tesis)

Guanajuato,Gto. Méx. **2009**

La tesis de Ana Miriam Romo Anaya es aprobada.

Dr. Jorge Domínguez y Domínguez (Presidente)

Dr. José Elías Rodríguez Muñoz(Secretario)

Dr. Víctor Alfredo Bustos y de la Tijera (Vocal) , Director de la tesis

Centro de Investigación en Matemáticas,A.C

2009

A Dana por elegir que yo sea parte de su vida y darme la oportunidad de seguir aprendiendo los procesos de mi vida.

Agradecimientos

Agradezco a cada uno de mis guias y maestros que he tenido a lo largo de mi vida.

Agradezo a mi primera maestra: mi madre Maria Jesus Anaya por darme el apoyo en mi superación profesional.

A mis 8 hermanos que con su ejemplo y su personalidad han aportado en mí enseñanzas para forjar mi camino en esta vida. Especialmente a David y Gabriel por ayudarme económicamente y moralmente en mi carrera profesional.

Gracias a Carlos Alejandro Betancourt por ser padre y madre durante largas horas y por sus palabras de aliento.

A todas aquellas personas que me compartieron sus conocimientos para llegar concluir esta tésis. En especial al Dr Alfredo Bustos por proporcionar ideas y recomendaciones en este proyecto.

Contenido

Lista de Figuras

Lista de Tablas

Resumen de la Tesis

Comparación de algunos diseños πps con el diseño Binomial negativo

por

Ana Miriam Romo Anaya

Centro de Investigación en Matemáticas,A.C-, (2009)

Dr. Víctor Alfredo Bustos y de la Tijera (Vocal) , Director de la Tesis

Esta tesis está basada en el documento **On the Distances Between Some πps Sampling Desings**, de Anders Lundqvist publicado el 5 abril del 2007, el cual trata sobre la comparación de algunos diseños de muestreo con propiedades deseables cuando se quiere utilizar el estimador Horvitz and Thompson (diseños πps también llamados proporcionales al tamaño). La comparación es realizada utilizando las distancias entre distribuciones de probabilidad de los diseños pro-puestos por medio de ejercicios numéricos de poblaciones pequeñas. El artículo compara los siguientes diseños: Poisson condicional(Cp), Poisson condicional ajustado (CPA), Sampford (S), Paretto (Par), Paretto ajustado(ParA). Se concluye que los diseños que producen probabilidades de inclusión deseadas, están muy cerca entre sí. Cada uno tiene ciertas propiedades que pueden ser importantes en un diseño de muestreo y son elegidos según la preferencia del usuario.

El principal propósito de este trabajo de tesis, es describir y analizar el diseño de muestreo al que llamaremos Binomial Negativo y realizar su comparación con los anteriores diseños verificando que éste también se encuentra dentro del grupo con propiedades deseables convirtíendose en una opción que puede ser elegido según las ventajas y desventajas que cada diseño presente.

CAPÍTULO 1

Introducción

Cada vez es mayor la necesidad contar con información estadística que proporcione datos confiables sobre un conjunto de elementos de cierta población, dicha información puede ser obtenida estudiando toda la población de interés o bien, seleccionar por algún criterio solo una parte de la población y obtener inferencias acerca de ella. Las ventajas de realizar una muestra son conocidas, reduce costos y tiempo, etc. Sin embargo cuando se realiza este tipo de estudio es necesario contar con métodos o procedimientos para obtener una muestra que sea realmente representativa o para obtener un buen diseño de muestreo que controle y mida confiabilidad de la información estadística. Es por ello que los actuales grupos de investigación están constantemente estudiando las nuevas técnicas y metodologías efectivas para lograrlo.

El estudio de un muestreo probabilístico es importante pues proporciona técnicas para generalizar los resultados que se obtienen a partir de la muestra a toda la población. El muestreo probabilístico introduce efectos aleatorios para seleccionar una muestra y proporciona información útil para evaluar la precisión de un estimador calculado a partir de ella. Existen una gran variedad de procedimientos probabilísticos para escoger una muestra, estos tiene que satisfacer las siguientes condiciones. [$Särndal, p$,8]

Si denotamos a una población finita como un conjunto $U = \{u_1, u_2, \ldots, u_N\}$.

1. Podemos extraer todos los subconjuntos posibles de U por medio de un procedimiento de muestreo.

2. Proporcionar una probabilidad de selección P asociada a cada posible muestra(llamada s).

3. La probabilidad de seleccionar cada unidad de la población es distinta de cero

4. Seleccionar una muestra con un procedimiento aleatorio bajo el cual cada posible muestra reciba exactamente la probabilidad $P(s)$.

Una muestra bajo estos requerimientos es llamada *muestra probabilística*.

El procedimiento aleatorio de selección de la muestra mencionado en el punto 1, es llamado esquema de muestreo.
La función P mencionada en el punto 2 define una distribución de probabilidad de muestras y es llamada diseño de muestreo. Esta función juega un papel importante debido a que determina las propiedades estadísticas esenciales (distribución de muestras, valor esperado, varianza) de ciertas cantidades calculadas a partir de la muestra tal como la media de la muestra y la varianza de la muestra. [$Särndal, p$,27].
La probabilidad a la que se refiere el punto 3 es llamada probabilidad de inclusión, y está asociada a la probabilidad de que alguna unidad de la población sea elegida para estar en la muestra.

Existe una gran variedad de procedimientos de muestreo que cumplen dichas propiedades pero siempre se busca encontrar uno óptimo en el sentido de que

produzca una buena precisión del estimador, y que sea fácil de implementar.

En esta tesis hacemos referencia a algunos diseños que son llamados πps o también conocidos como diseños de muestreo proporcionales al tamaño, propiedad que aumenta la precisión del estimador de Horvitz and Thompson (HT) por medio de la reducción su varianza estimada. Algunos diseños clásicos que se encuentran en la literatura y que son presentados en esta tesis son: Poisson Condicional Ajustado, Sampford, Paretto Ajustado.
En este trabajo de tesis se propone como una alternativa de los diseños conocidos: el diseño que llamaremos Binomial Negativo. Se explica el procedimiento de selección de la muestra y la función resultante de probabilidad y es comparado con los métodos que aquí se planean usando sus funciones de distribución de probabilidad.

En las siguiente secciones, se describe como una inducción, el estimador de HT y las propiedades que se requieren de un diseño para aumentar su precisión entre ellas la definición de un diseño πps, y la importancia de un tamaño fijo de muestra. Se plantea el objetivo y delimitación de la tesis y una sección sobre cómo está constituido este trabajo.

1.1. Estimador de Horvitz and Thompson

Usualmente, cuando deseamos estudiar una característica y de una población finita de tamaño N representada como $U = (u_1, u_2, u_3, \ldots, u_N)$ se utiliza el si-

guiente parámetro que indica la población total de la característica y.

$$t = \sum_{U} y_i$$

Donde y_i indica el valor de la variable para el i-ésimo elemento de la población. Bajo una muestra $s \subset U$, este parámetro puede ser estimado como:

$$\widehat{t} = \sum_{s} w_i y_i$$

Donde w_i es un valor que indica el peso del diseño.

Si denotamos a una variable aleatoria I_i que toma valor 1 si la unidad u_i esta en la muestra aleatoria S y el valor 0 si no lo está para toda $i = 1, 2, \ldots, N$ entonces podemos definir a π_i como la probabilidad de que la unidad u_i sea elegida en cualquier muestra de U, en otras palabras $\pi_i = P(I_i = 1)$. Entonces $w_i = 1/\pi_i$, y se obtiene:

$$\widehat{t_y} = \sum_{u_i \in s} \frac{y_i}{\pi_i} \tag{1.1}$$

alternativamente como

$$\widehat{t_y} = \sum_{U} I_i \frac{y_i}{\pi_i} \tag{1.2}$$

Este estimador es llamado Horvitz and Thompson (HT) es un estimador insesgado con varianza dada por:

$$V(\widehat{t_y}) = \sum_{i=1}^{N} (1 - \pi_i) \frac{y_i^2}{\pi_i} + \sum_{i=1}^{N} \sum_{i \neq j} (\pi_{ij} - \pi_i \pi_j) \frac{y_i y_j}{\pi_i \pi_j} \tag{1.3}$$

y puede estimarse por medio de

$$\widehat{V(\widehat{t_y})} = \sum_{i=1}^{N}(1-\pi_i)\frac{y_i^2}{\pi_i^2} + \sum_{i=1}^{N}\sum_{i\neq j}(\pi_{ij}-\pi_i\pi_j)\frac{y_i y_j}{\pi_{ij}\pi_i\pi_j} \tag{1.4}$$

Donde π_{ij} es la probabilidad de que la unidad u_i y u_j estén en cualquier muestra obtenida de U o bien $\pi_{ij} = P(I_i = 1, I_j = 1)$, la cantidad $(\pi_{ij} - \pi_i\pi_j)$ representa la $COV(I_i, I_j)$.

La estimación de la varianza es un indicador de la calidad de la estimación de cierta característica en una muestra.

Para un diseño que proporcione un tamaño de muestra fijo, se tiene una alternativa para la expresión de la varianza [10]:

$$V(\widehat{t_y}) = -\frac{1}{2}\sum_{i}^{N}\sum_{j\neq i}(\pi_{ij}-\pi_i\pi_j)(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 \tag{1.5}$$

Una estimación de la varianza de $\widehat{t_y}$ propuesta por Yates y Grundy (1953) cuando n es fijo:

$$\widehat{V}(\widehat{t_y}) = -\frac{1}{2}\sum_{i}^{n}\sum_{j\neq i}(\pi_{ij}-\pi_i\pi_j)(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2\frac{1}{\pi_{ij}} \tag{1.6}$$

Remark 1.1.1 *Si n es fijo y las variables I_i y I_j son independientes, entonces $\pi_{ij} = \pi_i\pi_j$ y por lo tanto la varianza se ve reducida a cero*

Remark 1.1.2 *Si n es fijo y las probabilidades de inclusión son proporcionales a la variable de interes $\pi_k \propto y_k$, esto es que $\frac{y_k}{\pi_i} = c$ para algun $c \in R$ e implica que $\widehat{t} = nc$ y que por lo tanto el estimador no tiene variabilidad y su varianza se reduce a cero.*

Obtener los valores y_k no es siempre posible, ya que esto implicaría tener todos los valores de la variable para cada elemento de la población. Pero si se existe una variable auxiliar cuyos valores son conocidos para toda la población z_i obtenidas en censos o en encuestras en estudios anteriores y que tenga una correlación preferentemente alta con cada y_i para toda $i = 1, 2, \ldots, N$ se conseguiría no anular la varianza pero si reducirla.

En tal caso, la proporcionalidad de π_i con los valores conocidos z_i está dada bajo la relación:

$$\pi_i = \frac{nz_i}{\sum_U z_j} \quad para \quad i = 1, 2, \ldots, N \tag{1.7}$$

Donde n es el tamaño fijo de la muestra y N tamaño de la población y $\pi_i \propto z_i$. Y se asume que $z_i \leq \sum_U z_i/n$ para toda i, si $z_i > \sum_U z_i/n$, entonces se toma $\pi_i = 1$.

Los diseños de muestreo que proporcionan probabilidades π_i proporcionales a valores positivos z_i son llamados diseños de probabilidad proporcionales al tamaño (πps).

Por otro lado, Hansen and Hurwitz (1943) mostraron que si se usan probabilidades de inclusión distintas para cada elemento de la población, incrementa la precisión del estimador HT comparado con probabilidades de inclusión iguales [1].

En resumen, las propiedades que aumentan la precisión del estimador HT.

- Tamaño fijo de muestra n y diseños que sean πps, es decir que las probabilidades de inclusión del diseño sean iguales que se expresan en 1.7 o al menos muy aproximadas.

- Tamaño fijo de muestra n y variables indicadoras independientes es decir

$COV(I_i, I_j) = 0$

- Probabilidades de inclusión distintas (desiguales) para cada unidad de la población.

De manera adicional se utiliza el concepto de entropía que es interpretada como el grado de aleatoriedad existente en el diseño de muestreo, es decir, la máxima entropía es cuando todas las muestras posibles son igualmente probables y el resultados es poco predecible, lo que indica que no influye el diseño en las estimaciones y la independencia de las variables indicadoras es mayor. Se mostrará en los ejercicios numéricos que los diseños que aquí se presentan que son muy cercanos a ser πps tienen entropía máxima.

1.2. Descripción del problema

Idealmente, las características que se pedirían para obtener un diseño óptimo son:

- Que el procedimiento de selección sea sencillo, es decir, que no implique demasiados cálculos numéricos
- Que maximize entropía
- Que el diseño sea de tamaño fijo y proporcione variables de inclusión independientes
- Que diseño sea de tamaño fijo y πps

Sin embargo estas propiedades no son realizables simultáneamente. En el caso del diseño Poisson (se detalla más adelante) se obtienen variables indicadoras

independientes sin embargo, se aceptan todos las posibles tamaños de muestras al aplicar la rutina de obtención de la muestra, por lo que el tamaño de muestra es aleatorio y en consecuencia no se logra reducir la varianza del estimador de HT.

Por otro lado, existen algunos métodos que aunque tengan un tamaño de muestra fijo no proporcionan exactamente probabilidades de inclusión iguales a la expresión 1.7.

En este trabajo de tesis hacemos referencia a los diseños con tamaño de muestra fijo, que aunque no son propiamente diseños πps (excepto Sampford) pueden modificarse con el propósito de aproximarlos a serlos:

Poisson Condicional (CP), Condicional Ajustado (CPA), Sampford (S) Pareto (Par), Pareto Ajustado (ParA), Binomial negativo (BN).

Como ya se mencionó en un diseño Poisson cada unidad de la población es seleccionada independiente de las demás con probabilidad de inclusión igual a la expresión 1.7, sin embargo el tamaño de muestra es una variable aleatoria, este diseño puede ser modificado para obtener un tamaño de muestra fijo n, es decir rechazar todas las muestras que no proporcionen un tamaño n, lo que implica que el esquema Poisson está condicionado a n, es decir $\sum_{i=1}^{N} I_i = n$. Este último es llamado Condicional Poisson se obtiene un tamaño de muestra fijo, pero no es propiamente un πps. En el caso de Pareto, obtener una muestra en el proceso de selección es relativamente más sencillo que los demás pero tampoco es un πps. En las secciones correspondientes de cada método se verá una forma de aproximar estos diseños a que sean πps. El diseño Sampford se considera muy cercano al

óptimo, pues es un diseño s y el tamaño de muestra es fijo, sin embargo obtener una muestra puede consumir mucho tiempo.

Se mostrará que el método Binomial Negativo (BN) al igual que los diseños clásicos tiene propiedades deseables y puede ser tomado como una opción alternativa sobre los diseños ya propuestos y conocidos en la literatura.

Se realiza una comparación de sus funciones de probabilidades por medio de ejercicios numéricos con probabilidades proporcionales al tamaño desiguales. Para ilustrarlo gráficamente se utiliza la técnica de escalamiento multi-dimensional (multidimensional scaling) en particular el análisis de coordenadas principales (principal coordinate analysis, PCO). Se realiza una comparación numérica de los tiempos de ejecución en cada esquema de muestreo, es decir, el tiempo de obtención de una muestra para cada método. Se obtienen conclusiones.

En este trabajo de tesis no se trata las probabilidades de inclusión de segundo orden π_{ij}, ni se realiza aproximaciones de su covarianza asociada ya que además de no ser parte del objetivo principal la estimación de la varianza correspondiente necesita más estudios comparativos.

No se utiliza la varianza del estimador asociada a cada uno de los métodos para realizar una comparación de diseños.

1.3. Objetivos

El objetivo principal de esta tesis es mostrar que el diseño propuesto, es decir el Binomial negativo(BN) está muy cerca de los diseños clasicos que aquí se detallan, para mostrarlo se comparan las distancias entre la función de distribución de probabilidad para cada uno de ellos.

Probamos con algunos ejercicios numéricos que los diseños que se encuentran muy cerca entre sí, son losIncremental Search (forward) s, y que además comparten ventajas y desventajas similares. Y que por tanto el diseño propuesto se encuentra dentro de esa clasificación.

Como objetivo secundario es presentar el esquema y el diseño de cada método, mostrando ventajas y desventajas. Los detalles teóricos y sustentos matemáticos no son parte de este objetivo, aun así se proporciona algunos anexos y se señalan los documentos donde se puede encontrar dicha información.

1.4. Resumen general

El capítulo siguiente trata sobre definiciones y conceptos básicos de muestreo, además se definen algunos parámetros y la notación que se utilizará en expresiones algebraicas posteriores. Se presenta el diseño Poisson como base e introducción de los demás.

En el capítulo 3, se describen cada uno de los diseños conocidos por la literatura clásica.

Para lograr los objetivos, la descripción de cada método planteado, es estructurado utilizando los siguientes puntos.

Introducción Proporciona una breve introducción del método y el algoritmo para la obtención de la muestra del método.

Función de probabilidad Se especifica la función de distribucion de probabilidades de las muestras del método.

Probabilidades de inclusión Se describen las ecuaciones para obtener las probabilidades de inclusión.

Conclusiones Se señalan algunas ventajas y desventajas del método planteado.

En el capítulo 4, se presenta el método propuesto (BN), se detalla el procedimiento, comenzando por la obtención de la muestra y a partir este algoritmo se construye la distribución de probabilidad y las probabilidades de inclusión deseadas.

En la parte final se presentan resultados de ejemplos con poblaciones pequeñas que ilustran y comprueban numéricamente las suposiciones iniciales. Para ilustrarlo gráficamente se utiliza la técnica de escalamiento multidimensional (multidimensional scaling) en particular el análisis de coordenadas principales (principal coordinate analysis, PCO).

Finalmente se realizan comentarios sobre los resultados obtenidos y se presentan conclusiones y propuestas de trabajos futuros.

CAPÍTULO 2

Notación y preliminares

Este capítulo trata sobre las definiciones básicas en el muestreo y la notación estándar que se utilizarán en los siguientes capítulos. Se describe el diseño Poisson como uno de los más conocidos y estudiados, y nos servirá como introducción a los diseños que se expondrán posteriormente.

Definición 2.0.1 *Una población finita se considera como un conjunto de unidades o elementos que se denota*

$$U = \{u_1, u_2, u_3, \ldots, u_N\}. \tag{2.1}$$

Cada valor es identificable. El valor N es el tamaño de la población.

Definición 2.0.2 *Una muestra s es un subconjunto cualquiera de U*

Notacion 2.0.1 *Se denotará como Ω a un conjunto de muestras de la población U.*

Un conjunto natural de muestras Ω es el llamado conjunto potencia con cardinalidad 2^N que contiene todos las muestras posibles obtenidas de U.

Notacion 2.0.2 *Se denotará como S a una muestra que es elegida aleatoriamente que toma valores en Ω de acuerdo a una probabilidad $P(S = s)$.*

Definición 2.0.3 *Una variable aleatoria I_i que nos indica si la unidad u_i está en la muestra, definida bajo la siguiente función $I_i : U \longrightarrow \{0, 1\}$.*

$$I_i = I_{(u_i)} = \begin{cases} 1 & si \quad u_i \in s \\ 0 & si \quad u_i \notin s \end{cases} \tag{2.2}$$

para toda $i = 1, 2 \ldots N$ es llamada variable indicadora.

A partir de esta última expresión, es posible definir al vector columna de tamaño N que define una muestra aleatoria como:

$$\mathbf{I} = (I_1, I_2 \ldots I_N)^t \tag{2.3}$$

Al conjunto de vectores que definen muestras aleatorias será denotado por Q, el cual puede ser identificado como $Q = \{0, 1\}^N$.

Al subconjunto de Q que contenga muestras de tamaño fijo n lo denotaremos como Q_n está definido como:

$$Q_n = \{\mathbf{x} \in Q | \sum_{k=1}^{N} x_k = n\} \tag{2.4}$$

Este conjunto tiene cardinalidad $\frac{N!}{n!(N-n)!}$.

En general, en este trabajo de tesis estaremos haciendo referencia a este último conjunto como el dominio de las funciones de probabilidad de muestras para cada método.

Para seleccionar los elementos que conforman una muestra, es necesario un *procedimiento de selección*, que llamaremos esquema de muestreo el cual se define

como un conjunto de procedimientos cuyo objetivo es seleccionar elementos de U para integrar la muestra. Dichos procedimientos se señalan con los algoritmos que se expondrán en cada método de muestreo aquí planteados.

La función de probabilidad resultante del procedimiento de selección de muestra se define a continuación.

Definición 2.0.4 *Un diseño de muestra f es una distribución de probabilidad multivariada $f : Q \longrightarrow (0, 1]$ tal que para cualquier muestra $\boldsymbol{x}$ cumpla con:*

$$P(\boldsymbol{I} = \boldsymbol{x}) = f(\boldsymbol{x}) > 0 \tag{2.5}$$

y

$$\sum_{\forall \boldsymbol{x} \in Q} f(\boldsymbol{x}) = 1 \tag{2.6}$$

En particular si el muestreo es sin reemplazo con tamaño fijo n, entonces el conjunto Q que es el dominio de f es sustituido por el conjunto Q_n.

Definición 2.0.5 *Probabilidades de inclusión de primer orden. Es la probabilidad de que la unidad u_i esté en la muestra*

$$\pi_i = E(I_i = 1) = P(I_i = 1) = P(u_i \in s) \tag{2.7}$$

Las probabilidad de inclusión para cada unidad en U es denotado por un vector π:

$$\pi = (\pi_1, \pi_2, \ldots, \pi_I, \ldots, \pi_N)^t \tag{2.8}$$

Definición 2.0.6 *La probabilidad de inclusión conjunta de orden 2 es la probabilidad de que la unidad u_i y u_j aparezcan en la muestra.*

$$\pi_{i,j} = P(S = s | u_i, u_j \in s) = P(\boldsymbol{I} = \boldsymbol{x} \, | x_i = 1, x_j = 1)$$

Además, la probabilidad de que una unidad u_i pertenezca a la muestra, es tal que

$$\pi_i = \sum_{\forall \mathbf{x} \in Q | x_i = 1} f(\mathbf{x}) \tag{2.9}$$

y la probabilidad de que dos elementos en particular pertenezcan a la muestra

$$\pi_{i,j} = \sum_{\forall \mathbf{x} \in Q | (x_i = 1, x_j = 1)} f(\mathbf{x}). \tag{2.10}$$

Definición 2.0.7 *La entropía sobre todas las posibles muestras de U de tamaño fijo n se define como una función que depende del diseño de muestra f*

$$Ent(f) = -\sum_{\forall \boldsymbol{x} \in Q} f(\boldsymbol{x}) log f(\boldsymbol{x}) \tag{2.11}$$

La entropía es interpretada como el grado de aleatoriedad existente en el diseño de muestreo, es decir, la máxima entropía es cuando todas las muestras posibles son igualmente probables y el resultados es poco predecible, lo que indica que no influye el diseño en las estimaciones.

Notacion 2.0.3 *Las probabilidades que se expresaron en 1.7 en capítulo 1 bajo una variable auxiliar z serán representadas como p_i, y serán llamadas probabilidades objetivo ya que se pretende que en los diseños que aquí se muestran exista la relación $p_i = \pi_i$, donde π_i son las probabilidades de inclusión para cada diseño. En resumen*

$$p_i = \frac{n z_i}{\sum_1^N z_j} \quad \forall i : 1, 2, 3, , , N \qquad tal \quad que \qquad \sum_i^N p_i = n. \tag{2.12}$$

Las siguientes cantidades serán de utilidad en la simplificación de las ecuaciones y funciones planteadas en cada diseño. $\forall i : 1, 2, 3, , , N$

$$q_i = 1 - p_i, \qquad r_i = p_i/q_i, \qquad d = \sum_{i=1}^{N} q_i p_i, \qquad b_i = q_i p_i (p_i - \frac{1}{2}). \tag{2.13}$$

2.1. Diseño Poisson

En un diseño Poisson la selección de unidades para conformar una muestra es simple de implementar.

La descripción se muestra en el siguiente algoritmo

Algoritmo 1 Muestra Poisson

Requiere: $p_1, p_2, \ldots, p_N \in [0,1]$ tal que $\sum p_i = n$; $n \in \mathbb{N}$ y $0 < n \leq N$

Devuelve: Un vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ que representa una muestra de unidades aleatorias con diseño Poisson.

1: $x \Leftarrow (x_1, x_2, \ldots, x_N)$ tal que $x_i = 0\ \forall i$
2: $v \Leftarrow (v_1, v_2, \ldots, v_N)$ Donde $v_i \sim U(0,1)$
3: **for** i:1 to N **do**
4: **if** $v_i < p_i$ **then**
5: $x_i \Leftarrow 1$
6: **else**
7: $x_i \Leftarrow 0$
8: **end if**
9: **end for**
10: **print** x

El diseño de una muestra Poisson se define como:

Definición 2.1.1 *La distribución de probabilidad de un esquema Poisson está dada por una función:* $f_p : Q_n \longrightarrow (0,1]$

$$f_p(\boldsymbol{x}) = \prod p_i^{x_i} q_i^{1-x_i} = C_p \prod r_i^{x_i}; \qquad C_p = \prod q_i. \tag{2.14}$$

r_i es como se define en la expresión 2.13.

Sin embargo, el diseño Poisson tiene una gran desventaja. Notamos en el Paso 3 del algoritmo 1 que el número de unidades que entran a la muestra no es siempre n, por lo tanto el tamaño de muestra es una variable aleatoria $\mathbf{n}$ variando entre $[0, N]$ con

$$E(\mathbf{n}) = \sum_i \pi_i. \qquad V(\mathbf{n}) = \sum \pi_i(1-\pi_i) \tag{2.15}$$

La varianza del estimador $\widehat{t}$ es:

$$V(\widehat{t}) = \sum_U \pi_i(1-\pi_i)\frac{y_i}{\pi_i}^2 \tag{2.16}$$

y un estimador insesgado de la varianza está dado por:

$$\widehat{V}(\widehat{t}) = \sum_s (1-\pi_i)\frac{y_i}{\pi_i}^2 \tag{2.17}$$

La varianza expresada en 2.16 puede ser muy grande debido a la variabilidad de la muestra. Minimizar la varianza para un tamaño esperado $n = \sum_U \pi_i$ se logra cuando se tienen la condición: $\pi_i = \propto z_i$ para $i = 1, 2, , , N$, donde z_i son los valores de una variable auxiliar z. Es decir $\pi_i = \frac{nz_i}{\sum_1^N z_j}$ para $i = 1, 2, , , N$.

Este diseño proporciona probabilidades de inclusión exactamente iguales a las probabilidades objetivo, es decir es un diseño πps. Además tiene la ventaja de que la selección de unidades son independientes, por lo que las probabilidades de inclusión de segundo orden π_{ij} son iguales a $\pi_i\pi_j$. Sin embargo, el proceso de seleccionar una muestra, puede repetirse una gran cantidad de veces rechazando todas las muestras que no cumplan el tamaño n. Esta repetición causa ciertos inconvenientes que resultan obvios, incluso las probabilidades de inclusión pueden llegar a ser afectadas. Además, la variabilidad de n no permite una reducción de

varianza del estimador HT.

En las dos primeras secciones del siguiente capítulo mostraremos las modificaciones de este diseño, con el propósito de que se obtengan las ventajas de un Poisson simple pero condicionado a tener un n fijo, y un ajuste de probabilidades iniciales para obtener probabilidades de inclusión deseadas.

CAPÍTULO 3

Diseños Clásicos

3.1. Resumen Capitular

En este capítulo se exponen los métodos tradicionales que existen en la literatura clásica sobre los diseños de muestreo muy próximos a ser πps y que minimizan la varianza del estimador de HT.

Se describe la forma en que un diseño Poisson simple y un diseño Pareto pueden ser modificados bajo ciertas condiciones, con la finalidad de obtener probabilidades de inclusión muy aproximadas a las probabilidades objetivo. Estas modificaciones se basan en el hecho de que Sampford es un diseño πps y entre más cercanas sean la funciónes de probabilidad de estos métodos a la de Sampford las probabilidades de inclusión de primer orden serán las deseadas.

3.2. Poisson Condicional

Este diseño de muestreo es obtenido seleccionando muestras Poisson hasta obtener el tamaño de muestra deseado. Es decir, rechazar todas las muestras que no cumplan con el tamaño especifico n. Este se presenta en el algoritmo 2.

Este proceso puede llegar a ser tardado sobre todo si N y n son grandes. Además, como veremos más adelante, las probabilidades de inclusión bajo este

Algoritmo 2 Muestra Poisson Condicional

Requiere: $p_1, p_2, \ldots, p_N \in [0,1]$ tal que $\sum p_i = n$; $n \in \mathbb{N}$ y $0 < n \leq N$

Devuelve: Un vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ que representa una muestra con n unidades aleatorias con diseño Poisson Condicional

1: **repeat**
2: $x \Leftarrow (x_1, x_2, \ldots, x_N)$ tal que $x_i = 0$; $\forall i$
3: $v \Leftarrow (v_i, v_2, \ldots, v_N)$ Donde $v_i \sim U(0,1)$
4: **for** i=1 to N **do**
5: **if** $(v_i < p_i)$ **then**
6: $x_i \Leftarrow 1$
7: **else**
8: $x_i \Leftarrow 0$
9: **end if**
10: **end for**
11: **until** $sum(x) = n$
12: **print** x

esquema, solo aproximan a las objetivo p_i.

Hajek(1981) dedicó gran parte de sus estudios de investigación en este diseño. Propuso varias aproximaciones en las probabilidades de inclusión a fin de implementar el diseño.

Chen *et.al* (1994) vinculó el diseño Poisson condicional con la teoría de familia de exponenciales que ha allanado el camino para varias implementaciones rápidas del método. Es por ello que este método puede ser llamado diseño exponencial con tamaño de muestra fijo.

3.2.1. Función de probabilidad

Sea Q_n el conjunto de vectores aleatorios de tamaño n tal como se definió en 2.4. La función de distribución de probabilidad de un esquema Poisson denotada por $f_{cp} : Q_n \longrightarrow (0, 1]$ está definida:

$$f_{cp}(\mathbf{x}) = P(\mathbf{I} = \mathbf{x}) = C_{cp} \prod p_i^{x_i} (1 - p_i)^{1-x_i} = C_{cp} \prod r_i^{x_i} \tag{3.1}$$

sujeta a $|\mathbf{x}| = n$.

La función de probabilidad proporciona las probabilidades de todas la muestras seleccionadas en un diseño condicional Poisson. Esta expresión es bien conocida por la literatura (Hajek, 1981).

La constante C_{cp} es llamada constante de normalización, y es obtenida utilizando el hecho de que la suma de la función de probabilidad sobre todas las muestras posibles es 1 como se describe en la definición 2.0.4. Se obtiene que el valor de esta constante es:

$$C_{cp}^{-1} = \sum_{|x|=n} \prod r_i^{x_i}. \tag{3.2}$$

3.2.2. Probabilidades de inclusión

Las probabilidades de inclusión en este diseño, están condicionadas a un tamaño de muestra fijo n.

Como ejemplo si $N = 5$ y $n = 2$ y las probabilidades iniciales de selección p_i son tales que $\sum p_i = n$, la probabilidad de que esta unidad 1 aparezca en la muestra es tal que:

$$\pi_1 = P(I_{u_1} = 1|n = 2) = \frac{P(I_{u_1} = 1, n = 2)}{P(n = 2)}$$

Donde la probabilidad $P(n = 2) = \frac{1}{2}\sum_i^N P(I_{u_i} = 1, n = 2)$.

De manera general si se denota a Q_n como las muestras posibles de tamaño n y Q_n^i el conjunto de muestras que contienen al elemento i, entonces la probabilidad de inclusión de primer orden para la unidad u_i en un proceso CP es:

$$\pi_i = P(I_{u_i} = 1|n) = \frac{p_i \sum_{Q_n^i} \prod_{j\neq i} p_j^{x_j}(1-p_j)^{1-x_j}}{\sum_{Q_n} \prod_j p_j^{x_j}(1-p_j)^{1-x_j}} \tag{3.3}$$

.

Entonces para este ejemplo en particular, la $P(I_{u_1} = 1, n = 2)$ está dada por

$$p_1[p_2(1-p_3)(1-p_4)(1-p_5) + p_3(1-p_2)(1-p_4)(1-p_5)+$$

$$p_4((1-p_2)(1-p_3)(1-p_5) + p_5(1-p_2)(1-p_3)(1-p_4)]$$

Similarmente si $Q_n^{i,j}$ denota el conjunto de I_n que contiene las unidades i y j entonces las probabilidades de inclusión de segundo orden

$$\pi_{u_i,u_j} = P(I_{u_i} = 1, I_{u_j} = 1|n) = \frac{p_i \sum_{Q_n^{ij}} \prod_{j \neq i} p_j^{x_j}(1-p_j)^{1-x_j}}{\sum_{Q_n} \prod_j p_j^{x_j}(1-p_j)^{1-x_j}} \tag{3.4}$$

Diferentes autores (Chen *et al.,* 1994; Chen & Liu, 1997; Aires, 1999, 2000, 2004; Chen, 2000; Deville, 2000) han dado distintas rutinas para calcular las probabilidades de inclusión de todos los ordenes. Por ejemplo si se utiliza n como súper índice para indicar el tamaño de muestra, para $n = 1, 2, 3, , N$ tenemos que:

$$\pi_i^{(n)} = \frac{n r_i (1 - \pi_i^{(n-1)})}{\sum r_k (1 - \pi_k^{(n-1)})} \quad y \quad \pi_{ij}^{(n)} = \pi_i^{(n)} \frac{\pi_j^{(n-1)} - \pi_{ij}^{(n-1)}}{1 - \pi_i^{(n-1)}}, \quad i \neq j$$

π_{ij} denotan las probabilidades de inclusión de segundo orden, y $r_i = \frac{p_i}{1-p_i}$

Reemplazando j en la segunda fórmula por jk,jkl... obtenemos probabilidades de inclusión de ordenes mayores.

Incluso las de segundo orden también pueden ser obtenidas dado las de primer orden. Si $p_i \neq p_j$, entonces

$$\pi_{ij}^{cp} = \frac{r_j}{r_j - r_i} \pi_i^{cp} + \frac{r_i}{r_i - r_j} \pi_j^{cp}. \tag{3.5}$$

Si $p_i = p_j$ se usa el hecho de que

$$(N_i - 1)\pi_{ij}^{CP} = n\pi_i^{CP} - \sum_{\{k: p_k \neq p_i\}} \pi_{ik}^{CP}$$

donde $N_i = \#\{j : p_j = p_i\}$. Estas fórmulas son derivadas por Aires(1999) [4]

Existe un algoritmo que se presenta como opción que facilita la obtención de probabilidades de inclusión utilizando herramientas de algebra lineal (L. Bondesson y I. Traat 2005)[3] se puede obtener las probabilidades de inclusión de primer orden para diferentes tamaños de muestras, utilizando el hecho de que en general,$\sum_{j:j\neq i}\pi_{ij} = (n-1)\pi_i^{CP}$.

Sustituyendo en la expresión 3.5

$$(\sum_{j:j\neq i}\frac{r_j}{r_j - r_i})\pi_i + \sum_{j:j\neq i}\frac{r_i}{r_i - r_j}\pi_j = (n-1)\pi_i \;\; para\; i = 1, 2, , N.$$

En sistema matricial tenemos que

$$A\pi = (n-i)\pi \tag{3.6}$$

donde

$$A = diag(\mathbf{1}^T\mathbf{C}) + \mathbf{C}: \quad NxN \tag{3.7}$$

con

$$c_{ij} = \frac{r_i}{r_i - r_j},\; c_{ii} = 0, \quad i \neq j \tag{3.8}$$

El vector **1** es de unos, π es el eigenvector derecho de A correspondiente a el eigenvalor con valor $n-1$.

Si descomponemos a la matriz A en sus valores singulares tenemos $A = P\triangle Q$, donde $\triangle$ es una matriz diagonal que contine los eigenvalores de A y P es una matriz cuadrada, ortogonal y sus columnas contiene los eigenvectores izquierdos de A y $Q = P^{-1}$ y normalizamos cada columna de P, tal que la suma de cada columna sea igual a 1, y cada valor de la primera columna es multiplicada por

N, segunda columna por $N-1$ asi sucesivamente. Obtenemos una matriz P^* con probabilidades de inclusión de cada diferentes tamaños de muestras, primera columna tamaño de muestra N, segunda con tamaño de muestra $N-1$, etc.

3.2.3. Poisson Condicional Ajustado

En un diseño CP las probabilidades de inclusión π_i no son muy aproximadas a las probabilidades p_i .

Un método para obtener las probabilidades de inclusión deseadas es usar una aproximación asintótica propuesta por Hakej (1981, p. 72) entre mayor sea la cantidad $d = \sum p_i(1-p_i)$ la conexión con la función de probabilidad Sampford (que se describirá en la siguiente sección) será más estrecha y por lo tanto $\pi_i \approx p_i$ Las probabilidades objetivo p_i son tales $\sum p_i = n$. Realizamos un ajuste de estas últimas. Se utiliza las siguientes referencias para el desarrollo del ajuste [4] y [11].

Sean $p_i^{'}$ las probabilidades ajustadas que cumplan $\sum p_i^{'} = n$ y que $\forall i$ se cumple que

$$\frac{p_i^{'}}{1-p_i^{'}} \propto \frac{p_i}{1-p_i} exp(\frac{1-p_i}{d}) \tag{3.9}$$

Para resolver este ultimo sistema y obtener los valores $p_i^{'}$, se busca una constante λ tal que

$$\frac{p_i^{'}}{1-p_i^{'}} = \lambda \frac{p_i}{1-p_i} exp(\frac{1-p_i}{d}) \tag{3.10}$$

y que además

$$n = \sum_i^N \frac{p_i}{p_i + \frac{1}{\lambda}(1-p_i)exp(-\frac{(1-p_i)}{d})} \tag{3.11}$$

Sistema no lineal que es resuelto numéricamente utilizando herramientas computacionales, de esta manera se obtienen p_i ajustadas, ahora llamadas $p_i^{'}$ tal que son sustituidas en 3.1 se obtiene un diseño con probabilidades de inclusión muy próximas a las deseadas.

Por otro lado, si se sustituye en la expresión 3.1 notamos que la función de probabilidad ahora es proporcional a

$$C_{cp}\prod(\frac{p_i}{1-p_i})^{x_i}exp(\sum\frac{1-p_i}{d}x_i)$$

y el último factor es proporcional a

$$exp(\sum\frac{1-p_i}{d}(x_i-p_i))\approx 1+\sum\frac{1-p_i}{d}(x_i-p_i)=\frac{1}{d}\sum(1-p_i)x_i.$$

entonces se obtiene un diseño muy cercano a Sampford expresado en la ecuación (3.18):

$$C_{cp}\prod(\frac{p_i}{1-p_i})^{x_i}\frac{1}{d}\sum(1-p_i)x_i\approx C_s\prod(\frac{p_i}{1-p_i})^{x_i}\sum(1-p_i)x_i$$

Debido a lo anterior, se define un diseño Poisson Condicional Ajustado (CPA) con dominio Q_n

$$f_{CPA}(\mathbf{x})=P(\mathbf{I}=\mathbf{x})=C_{cpa}\prod r_i^{x_i}exp(\frac{\sum q_k x_k}{d}) \tag{3.12}$$

sujeta a $|x|=n$ y $r_i=p_i/(1-p_i)$, $q=1-p$, proporciona probabilidades de inclusión aproximadas a las p_i.

Para obtener la constante C_{cpa} de la expresión 3.2 multiplicada por 1 tenemos

$$\sum_{|x|=n} \prod r_i^{x_i} = C_{cpa}^{-1} \sum_{|x|=n} \left[C_{cpa} \prod r_i^{x_i} exp(\frac{\sum q_k x_k}{d}) exp(-\frac{\sum q_k x_k}{d}) \right]$$

El último factor que es una suma sobre $|\mathbf{x}| = n$ es identificado como la esperanza de $exp(\frac{-\sum q_k x_k}{d})$ de la distribución CPA. De esto tenemos

$$C_{cp}^{-1} = C_{cpa}^{-1} E_{cpa} exp(-\frac{\sum q_k x_k}{d})$$

por el hecho de que $d = \sum q_k p_k$ la anterior ecuación es igual a

$$C_{cp}^{-1} = C_{cpa}^{-1} \frac{1}{e} E_{cpa} exp(-\frac{\sum q_k (x_k - p_k)}{d})$$

Considerando aproximación de Gauss , $E(g(Z)) \approx g(E(Z))$, sobre el valor Z igual a $Z = -d^{-1} \sum q_k (x_k - p_k)$ y $g(Z) = e^z$ junto con el hecho de que $E_{cpa}(x_k) = p_k$, se tiene que

$$E_{cpa} exp(-\frac{\sum q_k (x_k - p_k)}{d}) \approx exp[-E_{cpa} \frac{\sum q_k (x_k - p_k)}{d}] = 1$$

Así se tiene una aproximación para la constante de normalización en la función de CPA

$$C_{cp}^{-1} \approx (e C_{cpa})^{-1} \tag{3.13}$$

3.2.4. Conclusiones

En un diseño Poisson Condicional, la muestra, se obtiene seleccionando una muestra Poisson hasta obtener una del tamaño n. Si N y n son relativamente

grandes, entonces este proceso de selección puede llegar a ser muy tardado, ya que se pueden rechazar un gran número de muestras antes de obtener la deseada.

Además, las probabilidades de inclusión no son iguales a las inciales p_i, solo son aproximadas, es decir, no es un diseño πps. Varios investigadores han dado propuestas con la finalidad de aproximarlas (Hajek, Chen.al, Livi 1997, Aires 1999) Proporcionando algoritmos iterativos que simplifican los cálculos.

Estas propuestas se basan en el hecho de que si la función de distribución de las muestras de un diseño Poisson condicional es cercana a la de Sampford, entonces se obtendrá $\pi_i \approx p_i$. Dicha aproximación es llamada asintótica, pues se hace la suposición que la cantidad $d = \sum_1^N p_i(1 - p_i) \longrightarrow \infty$. Lo anterior indica que la aproximación de π_i a p_i depende del valor de d.

Si la diferencia entre las funciones de probabilidad de Sampford y CPA es muy pequeña, entonces la aproximación a ser un diseños πps es muy cercana.

Para que que diseño Condicional Poisson sea πps sin usar a aproximación de Sampford, es necesario resolver un sistema de ecuaciones no lineales para encontrar los valores p_i que cumplan $\sum_{i=1}^N p_i = n$ y la igualdad $\pi_i = p_i$ utilizando la expresion 3.3.

Una de las grandes ventajas de un diseño Poisson condicional ajustado es que maximiza la aleatoriedad en la selección de la muestra, en otras palabras, maximiza entropia. Es decir, el diseño CPA es tal, que la función de entropía mencionada en la definición (2.0.7) tiene un máximo cuando $f = f_{cpa}$. Propiedad

que se considera deseable y robustece el diseño [8].

En cuanto a la simplicidad analítica de la función de distribución de muestras se dice que es poco explícita, ya que está constituida por una familia de exponenciales [4].

3.3. Sampford

El diseño Sampford (1967) se considera uno de los más ingeniosos [10] y tiene la gran ventaja de obtener probabilidades de inclusión iguales a las probabilidades p_i iniciales.

Aunque existen diferentes formas de obtener una muestra Sampford, en la práctica se utiliza el método con reemplazo que consiste en obtener la primera unidad de la población que sea proporcional a p_i/n y las $n-1$ unidades restantes proporcionales a $p_i/(1-p_i)$. Una de las formas de realizar esto es como se detalla en el algoritmo 3.

3.3.1. Función de probabilidad

Dadas las probabilidades p_i tal que $\sum p_i = n$ y $r_i = \frac{p_i}{1-p_i}$ entonces el diseño Sampford [10] es definido como una funcion $f_s : Q_n \longrightarrow (0, 1]$

$$f_s(\mathbf{x}) = C_n \sum_{i=1}^{N} p_i x_i \prod_{k=1|k\neq i}^{N} r_k^{x_k} \tag{3.14}$$

o equivalentemente a

$$f_s(\mathbf{x}) = C_n \left(\prod_{1}^{N} r_i^{x_i} \right) \left(n - \sum_{1}^{N} p_i x_i \right) \tag{3.15}$$

Donde

$$C_n = \left(\sum_{t=1}^{n} t D_{n-t} \right)^{-1} \tag{3.16}$$

y

$$D_z = \sum_{x \in Q_z} \prod_{k=1}^{N} r_k^{x_k} \tag{3.17}$$

Tal que $D_0 = 1$ y el conjunto Q_z indica las muestras posibles de tamaño z en una población de tamaño N.

Bajo esta definición, se puede mostrar que

$$\sum_{x \in Q_n} f_s(\mathbf{x}) = 1$$

y que

$$\sum_{x \in Q_n | x_i = 1} f_s(\mathbf{x}) = \pi_i$$

Los detalles se presentan en el Apéndice A.

Alternativamente [8] presenta otra forma equivalente de escribir la función 3.14 que es muy conocida Sampford (1967).

Algoritmo 3 Para obtener una muestra Sampford

Requiere: $p_1, p_2, \ldots, p_N \in [0,1]$ tal que $\sum p_i = n$; $n \in \mathbb{N}$ y $0 < n \leq N$

Devuelve: Un vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ que representa una muestra con n unidades aleatorias con diseño Sampford.

1: $m \Leftarrow \sum_{i=1}^{N} \frac{p_i}{(1-p_i)}$

2: **for** k:1 to N **do**

3: $\quad \rho_k \Leftarrow \sum_{i=1}^{k} p_i/n$

4: **end for**

5: **for** l:1 to N **do**

6: $\quad \beta_l \Leftarrow \frac{1}{m} \sum_{i=1}^{l} \frac{p_i}{1-p_i}$

7: **end for**

8: $x \Leftarrow (x_1, x_2, \ldots, x_n)$; con $x_i = 0 \ \forall i$

9: $\varepsilon 1 \Leftarrow u$, para $u \sim U(0,1)$

10: Encontrar k para el cual $\rho_{k-1} < \varepsilon 1 \leq \rho_k$

11: $x_k \Leftarrow 1$

12: **repeat**

13: $\quad \varepsilon 2 \Leftarrow u$, para $u \sim U(0,1)$

14: $\quad$ Encontrar l tal que $\beta_{l-1} < \varepsilon 2 \leq \beta_l$

15: $\quad$ **if** $x_l = 0$ **then**

16: $\quad\quad x_l \Leftarrow 1$

17: $\quad$ **else**

18: $\quad\quad$ Ir a 8

19: $\quad$ **end if**

20: **until** $(sum(x) = n)$

21: **print** x.

$$f_s(\mathbf{x}) = C_s \prod_{i=1}^{N} r_i^{x_i} \sum_{k=1}^{N} q_k x_k \tag{3.18}$$

El valor de la constante C_s es encontrada suponiendo que d definida en la expresión dada por 2.13 es suficientemente grande, y utilizando el hecho de que la suma de la función de distribución sobre todas las muestras posibles de tamaño n es igual a 1, es decir, $\sum_{x:|x|=n} f(x) = 1$.

Si sabemos que $C_{cp}^{-1} = \sum_{|x|=n} \prod r_i^{x_i}$ y utilizando los supuestos anteriores no es dificil comprobar que:

$$C_s^{-1} = \sum_{|x|=n} \prod r_i^{x_i} \sum q_k x_k = \sum q_k \sum_{|x|=n} x_k \prod r_i^{x_i} = C_{cp}^{-1} \sum q_k \pi_k^{cp} \approx C_{cp}^{-1} d.$$

Lo que implica que $C_s \approx C_{cp}/d$.
Numericamente es facil ver que $C_n \approx C_s$. Y que además

$$\sum_{i=1}^{N} p_i x_i \prod_{k=1|k\neq i}^{N} r_k^{x_k} = \left(\prod_{1}^{N} r_i^{x_i} \right) \left(n - \sum_{1}^{N} p_i x_i \right) = \prod_{i=1}^{N} r_i^{x_i} \sum_{k=1}^{N} q_k x_k$$

Se utilizará la función expresada en 3.18 para realizar la comparación de diseños en los ejercicios ilustrativos en el capítulo 5.

3.3.2. Probabilidades de inclusión

En el esquema Sampford se cumple que $\pi_i = p_i$ probado por (Sampford (1967) y Hajek (1981, p. 86).

Es decir se cumple que:

$$\sum_{\forall x \in Q_n | x_i = 1} f_s(x) = \pi_i$$

Esto se demuestra en el apéndice A

Una prueba corta proporcionado por [4] es: Dado que se cumple que

$$\sum (1 - p_k) x_k = \sum p_k (1 - x_k)$$

manipulando esta relación se encuentra que

$$(1 - p_i)\pi_i^S = p_i(1 - \pi_i^S)$$

entonces se tiene $\pi_i = p_i$.

Existe una relación entre probabilidades de inclusión de primer y segundo orden de un diseño Poisson condicional y un diseño Sampford:

$$\pi_i^S = \frac{\sum_k (1 - \pi_k)\pi_{ik}^{CP}}{\sum_k (1 - \pi_k)\pi_k^{CP}} \tag{3.19}$$

y

$$\pi_{ij}^S = \frac{\sum_k (1 - \pi_k)\pi_{ijk}^{CP}}{\sum_k (1 - \pi_k)\pi_k^{CP}} \tag{3.20}$$

Esta relación es útil, cuando por algún método alternativo, es más fácil obtener las probabilidades de inclusión de un Poisson condicional.

3.3.3. Conclusiones

Este método tiene la gran ventaja de que se obtiene probabilidades deseadas, es decir se cumple $\pi_i = p_i$.

Debido a esta gran ventaja, Hajek (1981,p 72,88) propone que para obtener probabilidades de inclusión deseadas en Poisson Condicional y Pareto (este último lo describimos en la siguiente sección) se realiza una aproximación de sus diseños a la de un Sampford.

La forma analítica de la función de distribución de muestras, es explicita y sencilla, por lo que se considera una ventaja adicional.

La desventaja consiste en el tiempo que puede tomar en seleccionar la muestra, puede llegar a ser un proceso tardado y tedioso sobre todo si N y n son grandes. Al aplicar el algoritmo 3 se percibe que se puede presentar un gran número de muestras rechazadas antes de obtener la deseada. En el capítulo 5 donde se realizan ejercicios numéricos se muestra el tiempo promedio que toma el software para obtener una muestra bajo este esquema, concluyendo que el esquema Sampford toma más tiempo que los demás.

Aunque existen distintas formas de obtener una muestra Sampford, por ejemplo, en [12] muestra algunas opciones y sus respectivos procedimientos para seleccionar la muestra que puede ser sin reemplazo, con reemplazo, secuencial, via Paretto etcétera. También se realiza una simulación del tiempo que puede tomar cada método hasta obtener una muestra Sampford y aunque se demuestra que existen métodos más rápidos que otros, aun así la rapidez no es alta, y sigue dependiendo de los valores N y n.

3.4. Paretto

El diseño de Parreto fue introducido por Rósen(1997a,b) Usando la idea de Ohlsson(1990,1998). La selección de unidades para conformar una muestra se describe en el siguiente algoritmo.

Algoritmo 4 Para obtener una muestra Pareto

Requiere: $p_1, p_2, \ldots, p_N \in [0,1]$ tal que $\sum p_i = n$; $n \in \mathbb{N}$ y $0 < n \leq N$

Devuelve: Un vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ que representa una muestra con n unidades aleotorias con diseño Pareto.

1: $x \Leftarrow (x_1, x_2, \ldots, x_n)$ con $x_i = 0$ para $\forall i$

2: $Q \Leftarrow \phi$

3: $v \Leftarrow (v_1, v_2, , v_N)$ Donde $v_i \sim U(0,1)$

4: **for** $i : 1$ to N **do**

5: $\quad Q_i \Leftarrow \frac{v_i/(1-v_i)}{p_i/(1-p_i)}$

6: **end for**

7: $Q' \Leftarrow sort(Q)$

8: $x \Leftarrow Q'$

9: **print** x

Bajo este esquema las probabilidades de inclusión solo aproximan a las probalidades p_i, sin embargo como en el caso Poisson condicional puede encontrarse un método de ajuste para tener las probabilidades deseadas. Esta adaptación del diseño es llamado Paretto ajustado (Rosén y Saavedra (1995)).

3.4.1. Función de probabilidad

La función de probabilidad del diseño Pareto $f_{Par} : Q_n \longrightarrow (0, 1]$ es de la siguiente forma

$$f_{Par}(\mathbf{x}) = C_{par} \prod r_i^{x_i} \sum c_k^{'} x_k. \qquad |x| = n \tag{3.21}$$

Donde $r_i = p_i/q_i$ para toda $i = 1, 2, .., N$. Utilizando la notación Bondesson *et al* [12] los valores $c_k^{'}$ son representados como:

$$c_k^{'} = \frac{\int_0^\infty x^{n-1}(\prod \frac{1+r_j}{1+r_j x}) \frac{1}{1+r_k x} dx}{\int_0^\infty x^{n-1}(\prod \frac{1+r_j}{1+r_j x}) dx} \tag{3.22}$$

La integral puede ser resulta por medio de métodos numéricos, o analíticamente evaluada por descomposición parcial.

Si llamamos c_o a la integral

$$c_o = \int_0^\infty x^{n-1}(\prod \frac{1+r_j}{1+r_j x}) dx$$

y

$$c_k = \int_0^\infty x^{n-1}(\prod \frac{1+r_j}{1+r_j x}) \frac{1}{1+r_k x} dx$$

entonces

$$c_k^{'} = c_k/c_o$$

Utilizando aproximación de Laplace se puede llegar a demostrar que

$$c_o \approx \sqrt{\frac{2\pi}{d}}$$

y que

$$c_k^{'} = \frac{c_k}{c_o} = \frac{1}{1+r_k} + r_k \frac{\partial log c_o}{\partial r_k} \approx (1 - p_i)(1 + \frac{p_i(p_i - \frac{1}{2})}{d}) \tag{3.23}$$

entonces para d no pequeña se cumple,

$$c_k^{'} \approx (1 - p_i) \tag{3.24}$$

Los detalles se muestran en el Apéndice B.1

Si sustituimos la expresión 3.24 en 3.21 notamos que si d tienda a ∞ entonces los diseños Sampford y Pareto son muy cercanas.

Ademas, si lo anterior sucede, se puede ver que la constante C_{Par} es aproximada a:

$$C_{Par} \approx C_S \approx C_{Cp}/d \tag{3.25}$$

Por otro lado si n es pequeña o n es muy cercana a N entonces d es pequeña, e implica que las distribuciones de Sampford y Pareto difieren en la siguiente cantidad:

$$J_k = \frac{c_k}{c_o(1 - p_k)} \tag{3.26}$$

Para obtener los valores $c_k^{'}$ se utilizan aproximaciones. Una primera aproximación está dada por:

$$c_k^{'} = (1 - p_k)J_k^* \quad con \quad J_k^* = (1 + \frac{p_i(p_i - \frac{1}{2})}{d}) \tag{3.27}$$

Aproximaciones mas exactas pueden ser obtenidas por medio de analísis numérico y que son relativamente fáciles de implementar computacionalmente. Aunque en este apartado no se da el sustento teórico, la ecuación que aproxima dichos valores como segunda aproximación es la siguiente y es llamada calibrada

$$(\frac{c_k}{c_o})^{cal} = (1 - p_k)\frac{(N - n)\sigma_k exp\{\sigma_k^2 p_k^2/2\}}{\sum(1 - p_i)\sigma_i exp\{\sigma_i^2 p_i^2/2\}} \tag{3.28}$$

donde $\sigma_k^2 = \frac{1}{d + p_i(1-p_i)}$

3.4.2. Probabilidades de inclusión

De una manera intuitiva, las probabilidades de inclusión bajo este esquema son obtenidas usando el hecho de que al tener una secuencia de variables

$$Q_1, Q_2, .., Q_N$$

y su correspondiente secuencia de variables ordenadas llamada estadístico de orden

$$Q_{(1)}, Q_{(2)}, .., Q_{(N)}$$

es decir $Q_{(1)} = min(Q_i)$ y $Q_{(N)} = max(Q_i)$.
Entonces, la probabilidad de que la unidad i esté en la muestra de tamaño n para $i = 1, 2, , N$ implicará obtener la probabilidad de que el valor Q_i sea menor que $Q_{(n)}$.

El planteamiento formal para encontrar dicha probabilidad se basa en que, las cantidades Q_i son variables que tienen una función de distribución Pareto de la forma $F_i(t) = \frac{r_i t}{1+r_i t}$ y que la secuencia de las N variables ordenadas $Q_{(i)}^N$ tienen como función de distribución $F_{(i)}^N$ con $i = 1, 2, , N$ y parámetros $r_i > 0$, $r_i = p_i/(1 - p_i)$.

La solución se reduce a encontrar las funciones $F_{(i)}^N$, y la resolución de una integral, por medio de métodos numéricos.

Por ejemplo, si $i = N$, la probabilidad de que N pertenezca a la muestra. está dada por:

$\pi_N = P(N \in S) = P(Q_{(n)}^{N-1} > Q_N) = \int_0^\infty (1 - F_n^{N-1}(t)) f_N(t) dt$(3.29)Donde f_N es la función de densidad de Q_N, es sencillo verificar que

$$f_N = \theta/(1+\theta_N)^2$$

Para encontrar cualquier otro valor π_i, bastará con reordenar las variables Q_i

$$Q_1, Q_2, ...Q_{i-1}, Q_{i+1}, ..., Q_N, Q_i$$

y se procede de la misma manera que antes.

Para encontrar las probabilidades de segundo orden, se usa una forma similar

$$\pi_{N-1,N} = P(N-1 \in S, N \in S) = P(Q_{(n-1)}^{N-2} > max(Q_{N-1}, Q_N))$$

que es igual a

$$\int_0^\infty (1 - F_{n-1}^{N-2}(t)) f_{max(Q_{N-1},Q_N)}(t) dt \tag{3.30}$$

Los detalles se muestran en el Apendice B.2

Como en el caso Sampford existe otro camino para obtener dichas probabilidades por medio del diseño CP y se basa en el siguiente teorema. La demostración se encuentra en [4].

Teorema 1 *Para un diseño Pareto con probabilidades p_i para cada i,se cumple que*

$$\pi_i^{Par} = \frac{\sum_k c_k \pi_{ik}^{cp}}{\sum_k c_k \pi_k^{cp}} \qquad \pi_{ij}^{Par} = \frac{\sum_k c_k \pi_{ijk}^{cp}}{\sum_k c_k \pi_k^{cp}} \tag{3.31}$$

Sin embargo en un diseño Poisson solo produce probabilidades de inclusión aproximadas a las deseadas, es decir no es un diseño πps.

Para lograrlo, bastará ajustar las probabilidades p_i como se realizó en el caso Poisson condicional. Tratando de acercar su función de distribución a una Sampford.

3.4.3. Paretto Ajustado

Se Realiza una aproximación asintótica como en el caso Poisson condicional ajustado. Buscamos las probabilidades $p^{'}$ tal que $\sum p_i^{'} = n$ y

$$\frac{p_i^{'}}{1-p_i^{'}} \propto \frac{p_i}{1-p_i} exp(-\frac{p_i(1-p_i)(p_i-1/2)}{d^2}) \tag{3.32}$$

Se encuentra λ por métodos numéricos computacionales, tal que
$\sum p_i^{'} = n$ y

$$\frac{p_i^{'}}{1-p_i^{'}} = \frac{1}{\lambda}\frac{p_i}{1-p_i} exp(-\frac{p_i(1-p_i)(p_i-1/2)}{d^2})$$

Entonces, para obtener probabilidades de inclusión deseadas, las probabilidades resultantes $p^{'}$ son reemplazadas por p_i $\forall i$ y sustituidas en 3.21.

Por otro lado, si se sustituye la expresión 3.32 en la ecuación (3.21) se obtiene:

$$f_{Par}(\mathbf{x}) = C_{par} \prod (\frac{p_i}{1-p_i})^{x_i} exp(-\sum \frac{p_i(1-p_i)(p_i-1/2)x_i}{d^2}) \sum c_i^{'} x_i \tag{3.33}$$

Donde $b_i = p_i(1-p_i)(p_i-1/2)$ y $d = \sum_i^N p_i(1-p_i)$. Si este último es suficientemente grande, entonces se tiene:

$$exp(-\sum \frac{b_i x_i}{d^2}) \approx 1 - \sum \frac{b_i}{d^2} \approx 1$$

Por lo tanto (3.33) aproxima al diseño Sampford bajo la expresion 3.18 de la seccion 3.3, cuando $C_S = C_{par}$.

Por lo tanto, dadas las probabilidades p_i se define un diseño Pareto Ajustado, con función de distribución $f_{ParA} : Q_n \longrightarrow (0, 1]$

$$f_{ParA}(\mathbf{x}) = C_{ParA} \prod r_i^{x_i} exp(-\frac{\sum b_k x_i}{d^2}) \sum c_i^{'} x_i, \qquad |x| = n. \tag{3.34}$$

y $C_{parA} = C_{par}$, en las que se obtiene probabilidades de inclusión muy aproximadas a las deseadas.

Remark 3.4.1 *Se puede usar estos resultados para ajustar las variables Q_i usadas en un diseño Pareto. Estas variables ajustadas son*

$$Q_i = \frac{U_i/(1-U_i)}{p_i/(1-p_i)} exp(\frac{p_i(1-p_i)(p_i - 1/2)}{d^2})$$

3.4.4. Conclusiones

Saavedra(1995) sugiere el hecho de que este diseño es muy eficiente y muy cercano a ser un diseño optimo. Una de las grandes ventajas de este método, es que la selección de una muestra Pareto es muy sencilla de obtener, es decir, no se requieren cálculos elaborados o gran cantidad de rechazos de muestras como en los esquemas hasta ahora presentados. Debido a lo anterior, algunos autores han implementado procedimientos para obtener muestras Sampford, o Poisson

condicional por medio de una muestra Pareto[4].

Bajo un diseño Pareto no se obtienen probabilidades $\pi_i = p_i$, pero si se realiza un ajuste de probabilidades objetivo p_i se obtendrán probabilidades muy aproximadas a las deseadas. Como en el caso Poisson Condicional esta aproximación depende del valor de $d = \sum_1^N p_i(1-p_i)$. Si éste último valor no es grande, entonces se aproximan numéricamente los parámetros $c_k^{'}$ y se resuelve un sistema de ecuaciones no lineales para encontrar las nuevas probabilidades ajustadas de tal manera que que se cumpla $\pi_i = p_i \ \forall i = 1, 2, , N$ usando la expresión 3.29.
La función de probabilidad de un diseño Pareto ajustado involucra una familia de funciones exponenciales, y no es sencilla desde el punto de vista analítico.

CAPÍTULO 4

Binomial negativo

4.1. Resumen Capitular

Como hemos visto en los anteriores esquemas, cada uno tiene un marco teórico sustentado, y han sido probados por expertos que continuamente han propuesto mejoras en cada uno de ellos.

En este capítulo se detalla la metodología de un método de muestreo con ciertas propiedades deseables: Tamaño fijo de muestra y πps y máxima entropía. El procedimiento es obtenido por (Bustos 2007,[14])

Como introducción, se describe el algoritmo para obtener una muestra definiendo primeramente las probabilidades que serán tomadas para este procedimiento, luego se muestran una serie de implicaciones que se derivan del algoritmo que son las bases para construir la función de distribución de las muestras y las probabilidades de inclusión. Por último se describe la forma de obtener un diseño πps por medio de un ajuste de parámetros, como en el caso CPA y ParA.

En el capítulo 5 veremos por medio de ejercicios numéricos, que desde el punto de vista probabilístico el BN es muy cercano a los diseños (Sampford y Paretto) y que además la estrecha cercanía con CPA nos llevará a comprobar que se en-

cuentra dentro de los diseños con alta entropía.

4.2. Procedimiento para obtener una muestra

Las probabilidades que se usan para obtener una muestra bajo este esquema serán denotadas por α_i para cada unidad u_i con $i = 1, 2, , , N$ tal que $\sum_{i=1}^{N} \alpha_i = 1$ (La forma de obtenerlas se detallan en una sección posterior).

En el diagrama de flujo 4.1, se representa la forma de obtener n unidades distintas.

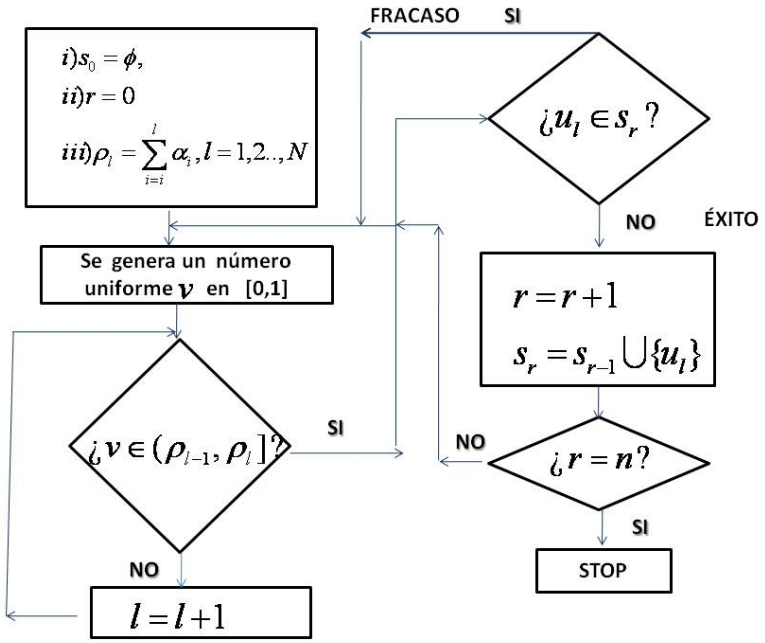

Figura 4.1: Diagrama de flujo para obtener una muestra BN

En forma de algoritmo tenemos:

Algoritmo 5 Obtener una muestra BN

Requiere: $\alpha_1, \alpha_2, \ldots, \alpha_N$ $\alpha_i \in [0,1]$ tal que $\sum \alpha_i = 1$; $n \in \mathbb{N}$ y $0 < n \leq N$

Devuelve: Un vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ que representa una muestra con n unidades aleatorias con diseño binomial negativo.

1: **for** k:1 to N **do**
2: $\rho_k \Leftarrow \sum_{i=1}^{k} \alpha_i$
3: **end for**
4: $x \Leftarrow (x_1, x_2, \ldots, x_N)$ tal que $\forall i$ $x_i = 0$
5: **repeat**
6: $\varepsilon \Leftarrow u$, para $u \sim U(0,1)$
7: Encontrar k para el cual $\rho_{k-1} < \varepsilon \leq \rho_k$
8: **if** $x_k = 0$ **then**
9: $x_k \Leftarrow 1$
10: **end if**
11: **until** $(sum(x) = n)$
12: **print** x

En términos descriptivos, el algoritmo 5 se basa en elegir las unidades de la población por medio de extracciones con reemplazo hasta obtener exactamente n unidades distintas para conformar la muestra. Cuando ocurre que una unidad aparece como candidata para entrar a la muestra y ésta ya fue seleccionada en un proceso anterior entonces decimos que ocurre un fracaso y la unidad no es extraída. Cuando la unidad es distinta de las que ya fueron seleccionadas entonces la unidad es elegida y decimos que se obtuvo una extracción exitosa.

Una implicación obvia cuando se implementa este algoritmo, es que la pri-

mera extracción siempre resulta un éxito, es decir, será seleccionada una unidad siempre en la primera extracción.

Si suponemos que en la primera extracción fué seleccionada la unidad u_i y la segunda extracción exitosa resultó u_j con $i \neq j$, entonces pudieron darse alguno de los siguientes sucesos:

1. La unidad u_j aparece en la segunda extracción, que fué exitosa, es decir no hay extracciones de fracaso, aparece u_j inmediatamente después de elegir a u_i. Este evento tiene probabilidad $\alpha_i\alpha_j$.

2. La unidad u_j aparece en la segunda extracción exitosa después de una extracción de fracaso, es decir, después de elegir u_i en la primera extracción la segunda resultó otra vez u_i. Este evento tiene probabilidad $\alpha_i[\alpha_i]\alpha_j$

3. La unidad u_j aparece en la segunda extracción exitosa después de dos extracciones de fracaso, es decir u_i fué extraida dos veces nuevamente antes de u_j. Este evento tiene probabilidad $\alpha_i[\alpha_i\alpha_i]\alpha_j$

4. La unidad u_j aparece en la segunda extracción exitosa después de $i-1$ extracciones de fracaso. Con probabilidad $\alpha_i[\alpha_k^{i-1}]\alpha_j$

5. Etc

Por lo tanto existe una infinidad de acontecimientos antes de tener una segunda extracción de éxito. Este hecho, presenta una forma muy natural de obtener la probabilidad de elegir u_i en la primera extraccion y u_j en la segunda.

$$\alpha_i \left[\sum_{r=0}^{\infty}(\alpha_i^r)\right] \alpha_j = \alpha_i \left(\frac{1}{1-\alpha_i}\right) \alpha_j \tag{4.1}$$

En el caso de la tercera extracción exitosa, la aparición de las unidades u_i y u_j en la tercera extracción exitosa resultan ser fracasos con probabilidad $(\alpha_k + \alpha_j)$. Es fácil probar que la probabilidad de obtener u_i en la primera extracción, u_j en la segunda extracción exitosa y u_l en la tercera extracción exitosa, está dada por:

$$\alpha_i \left[\sum_{r=0}^{\infty}(\alpha_i^r)\right] \alpha_j \left[\sum_{r=0}^{\infty}(\alpha_i + \alpha_j)^r\right] \alpha_l = \alpha_i \left(\frac{1}{1-\alpha_i}\right) \alpha_j \left(\frac{1}{1-\alpha_i-\alpha_j}\right) \alpha_l \quad (4.2)$$

Sin pérdida de generalidad, supongamos que las n unidades seleccionadas son representadas como un vector

$$\mathbf{w} = (u_{v1}^1, u_{v2}^2, \ldots, u_{vn}^k) \quad (4.3)$$

donde u_{vi}^k representa que la unidad vi fué seleccionada en la k-ésima extracción exitosa,para $i = 1, 2, , N$ y $k = 1, 2, , ,$.

La probabilidad de haber seleccionado la unidad vn en la k-ésima extracción exitosa, depende de las unidades seleccionadas anteriormente bajo la siguiente relación.

$$P(u_{vn}^k | u_{v1}^1, u_{v2}^2, \ldots, u_{v(n-1)}^{k-1}) = \frac{\alpha_{vn}}{1-\alpha_{v1}-\alpha_{v2}-\ldots-\alpha_{v(n-1)}} \quad (4.4)$$

Y la probabilidad de seleccionar exactamente los elementos de $\mathbf{w}$ está dada por:

$$P(\mathbf{w}) = P(u_{v1}^1)P(u_{v2}^2|u_{v1}^1)\cdots P(u_{vn}^k|u_{v1}^1, u_{v2}^2, \ldots, u_{v(n-1)}^{k-1}) \quad (4.5)$$

Equivalente a:

$$P(\mathbf{w}) = \frac{\alpha_{v1}\alpha_{v2}\cdots\alpha_{vn}}{(1-\alpha_{v1})\cdots(1-\alpha_{v1}-\alpha_{v2}-\cdots-\alpha_{v(n-1)})} \quad (4.6)$$

Si decimos que $\mathbf{w}$ es una posible muestra de U y definimos un nuevo vector $\mathbf{w}'$ con las mismas unidades seleccionadas en $\mathbf{w}$ pero cambiando el orden de aparición de las unidades, entonces en términos de una muestra, decimos que $\mathbf{w} = \mathbf{w}'$, pues no importa cómo fueron apareciendo, sino cuáles unidades fueron seleccionadas. Es claro que (4.6) no representa la función de distribución de muestras.

En términos de variables indicadoras, el vector resultante después de aplicar algoritmo 5 es la muestra obtenida por un diseño binomial negativo. Este vector $\mathbf{x}$ corresponde a las unidades seleccionadas sin importar el orden de aparición.

Para obtener la función de probabilidad es necesario considerar las formas posibles de ordenamiento o permutaciones de las unidades de $\mathbf{w}$ y la suma de las probabilidades de cada vector permutado indicará la probabilidad de obtener una muestra posible de U. La ecuación que la representa se especifica a continuación.

4.3. Función de probabilidad

Sea $\mathbf{w}$ el vector como se definió en la expresión 4.3 que indica las unidades seleccionadas respetando el orden de aparición. Y el conjunto W, con cardinalidad $n!$ que contiene todas las permutaciones posibles de elementos del vector $\mathbf{w}$.

$$W = \{(u_{v1}^1, u_{v2}^2, \ldots, u_{vn}^k), (u_{v1}^2, u_{v2}^1, \ldots, u_{vn}^k), \ldots, (u_{v1}^k, u_{v2}^{(k-1)}, \ldots, u_{vn}^1)\}$$

La relación del conjunto W con una muestra de variables indicadoras x_i es muy estrecha.

Para cualquier vector $\mathbf{w} \in W$ con $i = 1, 2, 3, , , N$

$$x_i = \begin{cases} 1 & \text{si } u_i \in \mathbf{w} \\ 0 & O.C. \end{cases}$$

Definición 4.3.1 *La función de probabilidad muestras la denotaremos por* f_{bn} : $Q_n \longrightarrow (0,1]$ *esta definida como:*

$$f_{bn}(\boldsymbol{x}) = f(\boldsymbol{I} = \boldsymbol{x}) = P(\bigcup_{W \ni w} w) = \sum_{W \ni w} P(\boldsymbol{w}) \tag{4.7}$$

La probabilidad $P(\mathbf{w})$ es obtenida por medio de la expresión dada por 4.6 que dependen de la probabilidad α_i.

Como ejemplo, si se considera la población $U = (u_1, u_2, u_3, u_4)$ y una muestra posible de tamaño 3, aplicando el algoritmo BN la muestra resultante es $(1, 0, 1, 1)$ correspondiente a las unidades seleccionadas $\{u_1, u_3, u_4\}$. La probabilidad de seleccionar esta muestra está dada por $f_{bn}(\mathbf{x}) = f((1, 0, 1, 1))$.

$$\frac{\alpha_1\alpha_3\alpha_4}{(1-\alpha_1)(1-\alpha_1-\alpha_3)} + \frac{\alpha_1\alpha_4\alpha_3}{(1-\alpha_1)(1-\alpha_1-\alpha_4)} + \ldots + \frac{\alpha_3\alpha_1\alpha_4}{(1-\alpha_3)(1-\alpha_3-\alpha_1)}$$

4.4. Probabilidades de selección incondicionales

Haremos referencia a una probabilidad de selección incondicional, como aquella probabilidad de obtener la unidad u_j en k-ésima extracción exitosa, sin importar cuales unidades son seleccionadas en las $k-1$ ésimas extracciones anteriores.

La probabilidad (4.2) nos indica la probabilidad de obtener u_l en la tercera extracción exitosa después de haber seleccionado las unidades u_k y u_j.

Sin embargo, para obtener la probabilidad de selección de la unidad u_l en la tercera extraccion exitosa se debe considerar todas las posibles unidades diferentes de u_i que pudieran ser seleccionadas en extracciones exitosas anteriores.

Llamamos $\alpha_j^{(k)}$ como la probabilidad de que aparezca la unidad u_j en la k esima extracción, para cada $k : 1, , ,$ y $j : 1, 2, , N$ y realizando algunos cálculos algebraicos tenemos lo siguiente:

- Para $k = 1$

$$\alpha_j^{(1)} = \alpha_j \tag{4.8}$$

- Para $k = 2$

$$\alpha_j^{(2)} = \alpha_j \sum_{i \neq j} \frac{\alpha_i}{1 - \alpha_i} \tag{4.9}$$

- Para $k = 3$

$$\alpha_j^{(3)} = \alpha_j \sum_{i \neq j} \left[\frac{\alpha_i}{1 - \alpha_i} \left(\sum_{i' \neq i} \frac{\alpha_{i'}}{1 - \alpha_{i'} - \alpha_i} \right) \right] \tag{4.10}$$

- Para $k = 4$

$$\alpha_j^{(4)} = \alpha_j \sum_{i \neq j} \left[\frac{\alpha_i}{1 - \alpha_i} \left(\sum_{i' \neq i} \frac{\alpha_{i'}}{1 - \alpha_{i'} - \alpha_i} (\sum_{s \neq i'} \frac{\alpha_s}{1 - \alpha_{i'} - \alpha_i - \alpha_s}) \right) \right] \tag{4.11}$$

Notamos que todas las probabilidades $\alpha_i^{(k)}$ dependen de las probabilidades iniciales α_i, que son iguales a las de la primera extracción. Además son siempre positivas, menores a uno y para cada k se cumple que $\sum_i^N \alpha_i^{(k)} = 1$.

4.5. Probabilidades de selección condicional

Para obtener probabilidades de selección condicionadas a unidades que son seleccionadas en extracciones anteriores tal como se muestra en la expresión (4.4), haremos uso de una muestra es decir un vector $\mathbf{x} \in Q_n$.

Llamamos $\alpha_j^{(k)}|x$ como la probabilidad de que aparezca la unidad u_j en la k esima extracción condicionada a las unidades que aparecen en la muestra $\mathbf{x}$, para cada $k : 1, , , n$ y $j : 1, 2, , N$ tenemos:

- Para $k = 1$

$$\alpha_j^{(1)}|x = x_j \alpha_j \tag{4.12}$$

- Para $k = 2$

$$\alpha_j^{(2)}|x = x_j \alpha_j \sum_{i \neq j} \frac{x_i \alpha_i}{1 - \alpha_i} \tag{4.13}$$

- Para $k = 3$

$$\alpha_j^{(3)}|x = x_j \alpha_j \sum_{i \neq j} \left[\frac{x_i \alpha_i}{1 - \alpha_i} \left(\sum_{i' \neq i} \frac{x_{i'} \alpha_{i'}}{1 - \alpha_{i'} - \alpha_i} \right) \right] \tag{4.14}$$

- Para $k = 4$

$$\alpha_j^{(4)}|x = x_j \alpha_j \sum_{i \neq j} \left[\frac{x_i \alpha_i}{1 - \alpha_i} \left(\sum_{i' \neq i} \frac{x_{i'} \alpha_{i'}}{1 - \alpha_{i'} - \alpha_i} (\sum_{s \neq i'} \frac{x_s \alpha_s}{1 - \alpha_{i'} - \alpha_i - \alpha_s}) \right) \right] \tag{4.15}$$

$\vdots$

Estas probabilidades condicionales nos ayudan para representar la función de distribución de probabilidad en forma equivalente a la función expresada en 4.7.

$$f_{bn}(\mathbf{x}) = f(\mathbf{I} = \mathbf{x}) = \sum_{i=1}^{N} \alpha_i^{(n)}|x \tag{4.16}$$

Donde n es fijo y representa el tamaño de la muestra.

4.6. Probabilidades de inclusión

Hacemos uso de la definición 2.9 donde se dice que las probabilidades de inclusión de primer orden es la probabilidad de que la unidad u_i se encuentre en la muestra. En otras palabras, que la unidad u_i haya sido seleccionada en la primera extracción exitosa, o en la segunda o en la tercera,,,, o en la n ésima. Cada una de éstas probabilidades son mutuamente excluyentes así que se puede expresar, para $i = 1, 2, \ldots, N$ como:

$$\pi_i = \alpha_i^{(1)} + \alpha_i^{(2)} + \ldots + \alpha_i^{(n)} \tag{4.17}$$

Donde $\alpha_i^{(k)}$ que son obtenidas como se detallan en las expresiones 4.8 a 4.11, éstas dependen de las probabilidades iniciales α_i, en la siguiente sección se muestra cómo obtenerlas para la realización de un diseño πps.

4.7. Binomial negativo Ajustado

Uno de los propósitos principales de este diseño propuesto, es obtener probabilidades de inclusión deseadas, la expresión (4.17) muestra que las probabilidades de inclusión de primer orden dependen de las iniciales $\alpha_i^{(k)}$ a través de las relaciones 4.8 a 4.11.

Es momento de encontrar las probabilidades α_i para la realización de las π_i deseadas.

Se resuelve un sistema de ecuaciones no lineales para obtener dicha relación,

en este caso tenemos:

Dadas las probabilidades p_i objetivo tal que $\sum p_i = n$ buscamos las α_i tal que se obtenga $\pi_i = p_i$, dando solución al siguiente sistema.

$$\begin{cases} \alpha_i^{(1)} + \alpha_i^{(2)} + .. + \alpha_i^{(n)} - p_i = 0 \quad \text{i=1,2,,,N} \\ \sum_i^N p_i - n = 0 \end{cases} \tag{4.18}$$

Se emplea conceptos de análisis numérico para la solución del sistema, comúnmente utilizando herramientas computacionales es relativamente sencillo obtener la solución de este sistema, en el cual obtenemos las α_i tal que $\sum \alpha_i = 1$ y por consiguiente obtener $\alpha_i^{(k)}$ que cumplen con $\sum_i^N \alpha_i^k = 1$ para cada $k = 1, 2, , , n$.

De esta manera obteniendo los valores α_i, que son la base para obtener una muestra bajo este esquema, se garantiza que obtendremos probabilidades de inclusión deseadas.

4.8. Conclusiones

Bajo este esquema de muestreo, dadas las probabilidades iniciales p_i tal que $\sum p_i = n$, el paso inicial será resolver el sistema de ecuaciones no lineal descrito en 4.18 con la finalidad de encontrar las probabilidades α_i que son los parámetros necesarios para seleccionar la muestra y por consiguiente se obtendrán las probabilidades de inclusión deseadas. Realizar lo anterior puede automatizarse por medio de herramientas computacionales.

Una característica importante del esquema es que es relativamente sencillo

obtener una muestra. El algoritmo se centra en obtener numeros aleatorios, seleccionando las unidades que entraran en la muestra, sin rechazar muestras completas que no cumplan con el tamaño especificado, además las sumas parciales de probabilidades son fijas, es decir, no son recalculadas en posteriores procesos.

En la parte final del siguiente capítulo, se expone los resultados de tiempos de procesos computacionales cuando se realizan simulaciones para obtener una muestra de cada esquema de muestreo que se expusieron en esta tesis, notando que el tiempo de ejecución para obtener una muestra bajo este esquema es menor comparado con los demás.

Otra característica, es que no se hacen supuestos de algunos parámetros para obtener probabilidades de inclusión iguales a las iniciales, como en el caso de Poisson y Paretto.

La expresión 4.16 es de gran utilidad cuando se quiere programar computacionalmente la función de probabilidad.

Aunque no se demuestra analíticamente, la cercanía de este diseño con los ya descritos es muy estrecha, además se encuentra dentro de los diseños con entropía máxima. Estas propiedades las verificaremos en los ejercicios numéricos propuestos y resueltos en el siguiente capítulo.

CAPÍTULO 5

Presentación de ejemplos resueltos

En este capítulo se presentan ejemplos numéricos para ilustrar las distancias entre funciones de distibucion de muestras entre los distintos esquemas de muestreo que ya hemos presentado. Para verlo gráficamente se utilizó la técnica MDS (multidimensional scaling), en particular la herramienta PCO (Principal coordinate analysis).

Aunque existen diferentes métricas que definen una distancia, la que emplearemos aquí es la distancia Hellinger, ya que se considera una métrica simétrica, euclidiana que satisface la desigualdad del tríangulo.

Definición 5.0.1 *Sea f_1 y f_2 dos funciones entonces la distancia Hellinger se define como.*

$$D_H^2(f_1, f_2) = 2 * \sum_x \left(\sqrt{f_1(x)} - \sqrt{f_2(x)}\right)^2$$

el factor 2 puede ser omitido (Gibbs and Su [6])

El análisis PCO consiste en obtener una matriz D simétrica con valores positivos y la diagonal principal con ceros, entonces esta matriz es llamada matriz de disimilaridad pues contiene las diferencias entre objetos (funciones de probabilidad).

El objetivo de PCO es buscar una representación Euclidiana para cada objeto tal que, las distancias Euclidiana entre los objetos sea la misma como la disimilaridad entre objetos.
Intuitivamante, PCO representa k objetos en un espacio de dimensión $k-1$ sin violar la propiedad de la desigualdad del triangulo. Otra forma de entender el objetivo de PCO es que encuentra la dirección de mayor variabilidad (1 coordenada principal) la segunda dirección de mayor variabilidad (2 coordenada principal), sucesivamente hasta la $k-1$ coordenada, muy similar a componentes principales (PCA) con el diferencia de que la matriz es Euclidiana, y solo proporciona información sobre unidades o variables no de ambas.
Con esta herramienta podemos representar la posición de las distribuciones de cada método, en un espacio de dimensión 2. Lo que nos indicaría qué tan lejos o cerca se encuentran las funciones de probabilidad desde el punto de vista probabilístico.

El procedimiento para realizar la comparación de funciones de probabilidad de muestras para cada diseño se indica a continuación.

- Dado los parámetros n y N se obtienen todas las muestras posibles de una población U representadas por medio de variables indicadoras, este conjunto es llamado el dominio de la función y denotado como Q_n.

- Para cada método de muestreo, se obtiene la probabilidad de todas los elementos de Q_n, por medio de su respectiva función de probabilidad. Obteniendo los valores de $f(\mathbf{I}=\mathbf{x})\ \forall \mathbf{x} \in Q_n$

- Se encuentra la distancia Hellinger aplicando la definición 5.0.1 para cada par de funciones obtenidos en el paso anterior, y por consiguiente se obtie-

ne una matriz de distancias D llamada de disimilaridad que contiene las distancias entre los diseños.

En el caso del diseño binomial negativo, fué necesario primeramente encontrar las probabilidades α_i resolviendo el sistema de ecuaciones no lineales 4.18.

Para encontrar los valores $c_k^{'}$ en el caso Paretto, y Pareto ajustado, se hizo uso de la hipótesis de que el valor d es grande y por consiguiente $c_k^{'} \approx (1 - p_i)$.

La automatización de los procesos en los ejercicios planteados fueron realizados en el software estadístico R aprovechando las ventajas que éste proporciona por medio de las librerías correspondientes para la obtención gráficas, solución de sistemas de ecuaciones no lineales etc.

Excluimos el esquema Poisson simple, pues bajo la distancia Hellinger queda muy alejado de las demás. Ademas, el objetivo principal es comparar los diseños con propiedades similares, entre ellas el tamaño fijo.

Ejercicio 1 *(Población TMB).* $N = 6$, $n = 3$ *y* $P_1 = P_2 = P_3 = \frac{1}{3}$ $P_4 = P_5 = P_6 = \frac{2}{3}$

Podemos notar que la suma de probabilidades es exactamente el tamaño de muestra $n = 3$, además $d = 3/4$. Las probabilidades α_i para cada unidad de la poblacion se presenta en la tabla siguiente.

i	1	2	3	4	5	6
α_i	0.0945	0.0945	0.0945	0.2387	0.2387	0.2387

Tabla 5.1: Probabilidades α_i (Población TMB)

Las distancias entre distribuciones se presentan en la tabla 5.2

	Sampford	CP	CPA	Par	ParA	BN
Sampford	0.00000	0.11709	0.01236	0.0189	0.00497	0.01209
CP	0.11709	0.00000	0.11437	0.09490	0.11377	0.11562
CPA	0.01236	0.11437	0.00000	0.02184	0.00826	0.00142
Par	0.0189	0.09490	0.02184	0.00000	0.01903	0.02296
ParA	0.00497	0.11377	0.00826	0.01903	0.00000	0.00839
BN	0.01209	0.11562	0.00142	0.02296	0.00839	0.00000

Tabla 5.2: Matriz de disimilaridad (Población TMB)

En la figura 5.1 podemos notar claramente que se forma un un grupo de cuatro esquemas compuesto por (CPA,S,ParA,BN). Sampford y Pareto Ajustado se encuentran muy cercanas, lo mismo sucede con CPA y BN. El diseño Poisson condicional está muy alejado de los otros y la función Pareto aunque está cerca

del grupo, existe una pequeña separación notable entre las demás.

Figura 5.1: Población TMB refleja la posición de las distribuciones

En la tabla 5.3 se muestra las probabilidades de inclusión y la entropía para cada uno de los diseños. Podemos notar que en efecto, el diseño Sampford y el Binomial negativo proporcionan probabilidades de inclusion de primer orden exactas a las probabilidades iniciales p_i, los otros diseños tienen una buena aproximación excepto CP.

Los tres diseños con mayor entropía son Sampford, BN, CPA. Aunque habiamos mencionado que el diseño CPA es teoricamente el que posee la mayor entropía, en este caso particular, al no ser así, puede deberse a que el valor $d = 3/4$ no es suficientemente grande.

π_i	Sampford	CP	CPA	Par	ParA	BN
1	0.3333333	0.2979592	0.3328908	0.3266922	0.3324879	0.3333333
2	0.3333333	0.2979592	0.3328908	0.3266922	0.3324879	0.3333333
3	0.3333333	0.2979592	0.3328908	0.3266922	0.3324879	0.3333333
4	0.6666667	0.7020408	0.6671092	0.6733078	0.6675121	0.6666666
5	0.6666667	0.7020408	0.6671092	0.6733078	0.6675121	0.6666666
6	0.6666667	0.7020408	0.6671092	0.6733078	0.6675121	0.6666666
Suma	3.0000000	3.0000000	3.0000000	3.0000000	3.0000000	3.0000000
Entropía	2.7150262	2.5815525	2.7136699	2.6920663	2.7122285	2.7141166

Tabla 5.3: Probabilidades de inclusión y entropía para cada diseño. Población TMB

Ejercicio 2 *(Población Sampford-Hajek. Considerada por Sampford y Hajek)* $N = 10$ *y* $n = 5$ *y* $P_1 = 0{,}2$, $P_2 = 0{,}25$, $P_3 = 0{,}35$, $P_4 = 0{,}4$, $P_5 = 0{,}5$, $P_6 = 0{,}5$, $P_7 = 0{,}55$, $P_8 = 0{,}65$, $P_9 = 0{,}7$, $P_{10} = 0{,}9$

La suma de probabilidades p_i $\forall i$ suman exactamente 5, el valor d es igual a 2.09 los valores obtenidos α_i para este ejemplo fueron como se muestra en la siguiente tabla:

i	1	2	3	4	5	6	7	8	9	10
α_i	0.028	0.036	0.053	0.062	0.083	0.083	0.095	0.124	0.143	0.289

Tabla 5.4: Probabilidades α_i (Población Sampford-Hajek)

La matriz de disimilaridad está representada por la siguiente tabla

	Sampford	CP	CPA	Par	ParA	BN
Sampfod	0.00000	0.08542	0.00714	0.0079	0.00117	0.01215
CP	0.08542	0.00000	0.08717	0.07735	0.08463	0.08498
CPA	0.00714	0.08717	0.00000	0.01145	0.00637	0.00729
Par	0.00117	0.07735	0.01145	0.00000	0.00756	0.01382
ParA	0.00135	0.08463	0.00637	0.00756	0.00000	0.01115
BN	0.01215	0.08498	0.00729	0.01382	0.01115	0.00000

Tabla 5.5: Distancia Hellinger entre funciones de distribución población Sampford

En el grafico 5.2 podemos ver que al igual que en ejercicio anterior sigue existiendo una tendencia de cercania de las 4 funciones (CPA,S,ParA,BN), la distribución Paretto se acerca más al grupo y la función CP sigue apareciendo alejada de las demás.

Figura 5.2: Población Sampford-Hajek refleja la posición de las distribuciones

Podemos notar en la tabla 5.6, que el diseño CPA tiene ahora la mayor entropía a diferencia del ejercicio anterior debido a que el valor de d es mayor . Aún así BN y Sampford tambien tienen entropía alta.

Lo anterior confirma el hecho de que los diseños con probabilidades de inclusión iguales a las objetivo y con máxima entropía son cercanos entre si desde el punto de vista probabilístico.

	Sampford	CP	CPA	Par	ParA	BN
1	0.2000000	0.1776002	0.2007922	0.1980615	0.1998530	0.2000000
2	0.2500000	0.2273531	0.2504537	0.2477464	0.2498215	0.2500000
3	0.3500000	0.3329162	0.3499266	0.3479722	0.3497989	0.3500000
4	0.4000000	0.3882836	0.3998571	0.3985333	0.3998396	0.4000000
5	0.5000000	0.5020487	0.5000728	0.5002460	0.5000255	0.5000001
6	0.5000000	0.5020487	0.5000728	0.5002460	0.5000255	0.5000001
7	0.5500000	0.5591474	0.5502002	0.5511540	0.5501255	0.5500000
8	0.6500000	0.6704621	0.6500789	0.6524351	0.6502175	0.6500000
9	0.7000000	0.7237542	0.6997972	0.7026506	0.7002034	0.7000000
10	0.9000000	0.9163859	0.8987484	0.9009550	0.9000897	0.9000001
suma	5.0000000	5.0000000	5.0000000	5.0000000	5.0000000	5.0000000
Entropía	4.7269498	4.5924153	4.7309224	4.7151764	4.7259615	4.726948

Tabla 5.6: Probabilidades de inclusión y entropía para cada diseño. Población Sampford

Ejercicio 3 *Utilizando cada algoritmo para la obtencion de una muestra de cada esquema de muestreo (Poisson condicional,Sampford,Binomial negativo) se simulan 100 muestras de tamaño* $n = 5$ *de una población de tamaño* $N = 10$ *y se registra el tiempo promedio de ejecución.*

La siguiente tabla muestra un intervalo del 95 % de confianza del tiempo promedio en segundos en la obtención de 100 muestras, para cada esquema.

Segun los resultados, podemos notar que toma menos tiempo obtener una muestra por el método Binomial, seguido por Poisson condicional. El algoritmo

Método	Intervalo (seg)
BN	(0.042,0.045)*
CP	(0.046,0.047)
SAMP	(2.30,2.39)

Tabla 5.7: Simulación del tiempo promedio para la obtención de muestras

Paretto fué descartado en esta simulación ya que es el mas sencillo comparado con los demás.

CAPÍTULO 6

Conclusiones

Cada uno de los diseños presentados tiene propiedades deseables en un diseño de muestreo, podemos resumir sus caracteristicas y propiedades de cada uno de ellos.

El diseño Poisson simple tiene probabilidades de inclusión identicas a las probabilidades de selección, es decir es un diseño πps. El procedimiento para obtener una muestra es relativamente fácil, sin embargo tiene la desventaja de que el tamaño de muestra es aleatorio y esto, puede ocacionar un número alto de re-chazos cuando se quiere tener una muestra de tamaño fijo n.

Se propone una adaptación de éste último diseño, llamado Poisson condicional(CP) el cuál está condicionado a obtener una muestra de tamaño fijo n, realizando el mismo procedimiento que un Poisson simple pero rechazando todas las muestras que no sean de tamaño n. Esta adaptación no produce probabilidades de inclusión idénticas a las deseadas.

El diseño Poisson Condicional Ajustado tiene propiedades iguales a un CP pero con la gran ventaja de obtener probabilidades muy cercanas a las deseadas, y es el diseño que tiene entropía máxima, en otras palabras minimiza la KL-divergencia entre un diseño Poisson y los diseño de tamaño fijo.

El diseño Sampford, es considerado como uno de los óptimos pues se obtienen exactamente las probabilidades requeridas, y la expresión análitica de su función de distribución es más simple que CPA ya que esta última implica una familia

de exponenciales. Sin embargo, obtener una muestra bajo este esquema no es un proceso sencillo.
Pareto, lleva la delantera en la simplicidad de la obtención de una muestra, pero para obtener probabilidades de inclusión deseadas, es necesario realizar un ajuste de parámetros iniciales, tratando de que esta función de distribución se acerque a una función de distribución Sampford. Proceso que es llamado Pareto Ajustado.

Los ajustes que se realizan en un Poisson condicional y Paretto, dependen de ciertas condiciones para que la cercanía a la función de distribución de Sampford sea muy aproximada, y por ende, obtener probabilidades de inclusión deseadas.

Para que el método Binomial negativo sea un diseño πps, se realiza un ajuste de parámetros iniciales pero a diferencia de los otros métodos ajustados, éste no requiere de supuestos basta con resolver un sistema de ecuaciones no lineales, que al igual que los otros no es computacionalmente demandante. Tiene la ventaja de que obtener una muestra bajo este esquema no es complicado, pues sólo se rechazan unidades, no muestras completas como en los otros métodos excluyendo a Pareto.

En los ejemplos presentados, se mostró gráficamente que existe un compotamiento similar entre 4 diseños (Sampford, Pareto Ajustado, CPA, y BN) que bien podríamos encasillarlos en un grupo o cluster que comparten algo en común: Producen probabilidades iguales a las deseadas y entropía máxima. Hecho que confirma uno de los objetivos de este trabajo, pues el método propuesto está dentro de esta clasificación.

Los ejemplos presentados confirman nuestra sospecha. El método binomial

negativo se ecuentra dentro de los diseños óptimos lo cual se convierte en una alternativa más de los diseños πps.

Apéndice A

Anexos

A.1. Demostración sobre las probabilidades de inclusión en un diseño Sampford

Dadas las probabilidades p_i tal que $\sum p_i = n$ y $r_i = \frac{p_i}{1-p_i}$ entonces el diseño Sampford es definido como [10]

$$f_s(x) = C_n \sum_{i=1}^{N} p_i x_i \prod_{k=1|k\neq i}^{N} r_k^{x_k}$$

o equivalentemente a

$$f_s(x) = C_n \left(\prod_{1}^{N} r_i^{x_i} \right) \left(n - \sum_{1}^{N} p_i x_i \right)$$

Donde

$$C_n = \left(\sum_{t=1}^{n} t D_{n-t} \right)^{-1}$$

y

$$D_z = \sum_{x \in Q_z} \prod_{k=1}^{N} r_k^{x_k}$$

Tal que $D_0 = 1$ y el conjunto Q_z indica las muestras posibles de tamaño z en una población de tamaño N.

Bajo esta definición, se puede mostrar que

$$\sum_{x \in Q_n} f_s(x) = 1$$

y que

$$\sum_{x \in Q_n | x_i = 1} f_s(x) = \pi_i$$

Para probar lo anterior se demuestra primero el siguiente Lema

Lema 1 *(Sampford 1967) Sea π_i las probabilidades de inclusión de un diseño Sampford tal que $\sum \pi_i = n$ y $r_i = \frac{\pi_i}{1-\pi_i}$, si*

$$g(n, j, k) = \sum_{x \in Q_{n-j}(U \backslash k)} \left(\prod_{l \in U} r_l^{x_l} \right) \left(n - j r_k - \sum_{i=1}^{N} r_i x_i \right)$$

donde

$$Q_{n-j}(U \backslash k) = \{ x \in Q_{n-j} | x_k = 0 \}$$

entonces

$$g(n, j, k) = (1 - \pi_k) \sum_{t=i}^{n} t D_{n-t} \tag{A.1}$$

Proof.

Cuando la cardinalidad de Q es igual a (n-j) se tiene que

$$n - j\pi_k - \sum_{1}^{N} \pi_i x_i = j(1 - \pi_k) + \sum_{1}^{N} (1 - \pi_i) x_i$$

en consecuencia

$$g(n, j, k) = \left[j(1 - \pi_k) D_{n-i}(\overline{k}) + h(n, j, k) \right] \tag{A.2}$$

donde

$$h(n, j, k) = \sum_{x \in Q_{n-j}(U \backslash k)} \left(\prod_{l \in U} r_l^{x_l} \right) \sum_{1}^{N} (1 - \pi_i) x_i \tag{A.3}$$

y

$$D_z(\overline{k}) = \sum_{x \in Q_{n-j}(U \setminus k)} \prod_{l=1}^{N} r_l^{x_l}$$

Note que $1 - \pi_i = \pi_i / r_i$. Si en la expresión (A.3), reemplazamos $1 - \pi_i$ tendremos

$$\begin{aligned} h(n,j,k) &= \sum_{x \in Q_{n-j}(U \setminus k)} \left(\prod_{l \in U} r_l^{x_l} \right) \sum_{1}^{N} \frac{\pi_i}{r_i} x_i \\ &= \sum_{x \in Q_{n-j}(U \setminus k)} \left(\prod_{l \in U} r_l^{x_l} \right) \left[\sum_{i=1 | i \neq k}^{N} \pi_i (1 - x_i) \right] \\ &= \sum_{x \in Q_{n-j}(U \setminus k)} \left(\prod_{l \in U} r_l^{x_l} \right) \left(1 - \pi_k - \sum_{i=1 | i \neq k}^{N} \pi_i x_i \right) \\ &= g(n, j+1, k) + j \pi_k D_{n-j-1}(\overline{k}) \end{aligned} \tag{A.4}$$

Debido a que

$$D_m(\overline{k}) = D_m - r_k D_{m-1}(\overline{k})$$

de la expresión (A.2) y de (A.4) obtenemos la siguiente relación recursiva:

$$g(n,j,k) = j(1 - \pi_k) D_{n-j} + g(n, j+1, k),$$

con la condición inicial de $g(n,n,k) = 1 - \pi_k$.
Asi que la expresión (A.1) satisface esta ultima relación. ■

Bajo este lema podemos realizar la modificación para cuando $\pi_k = 1$

$$g(n,j,k) = \pi_k \sum_{t=j}^{n-1} D_{n-t-1}(\overline{k})$$

Los siguientes resultados pueden ser derivados de la siguiente proposición.

Proposicion 1 *Para un diseño Sampford se tiene*

1. $\sum_{\forall x \in Q_n} f(x) = 1.$
2. $\sum_{\forall x \in Q_n} x_k f(x) = \pi_k$

Proof.

i) Supóngase una unidad ficticia z tiene una probabilidad de inclusión nula, esto es, que $\pi_z = r_z = 0$, y $C_n = \left(\sum_{t=1}^{n} tD_{n-t}\right)^{-1}$ bajo el lema

$$\sum_{\forall x \in Q_n} f(x) = C_n g(n, 0, z) = C_n \sum_{t=1}^{n} tD_{n-t} = 1$$

ii)

$$\sum_{\forall x \in Q_n} x_k f(x) = C_n r_k g(n, 1, k) = \pi_k C_n \sum_{t=1}^{n} tD_{n-t} = \pi_k.$$

■

Apéndice B

Anexos

B.1. Función de probabilidad de un diseño Pareto

Lema 2 *considere la secuencia de variables* $Q_1, Q_2, .., Q_N$ *con función de probabilidad* $F_1, F_2, .., F_N$*, y sea* $Q^N_{(n)}$ *el n ésimo orden estadístico de las cantidades* $Q_{(1)}, Q_{(2)}, .., Q_{(N)}$ *con función de distribución* F^N_n*. Entonces para* $N = 1, 2, , , n = 1, 2, , N$*, la función* F^N_n *cumple la ecuación recursiva:*

$$F^N_n((t) = F^{N-1}_n(t) + F_N(t)[F^{N-1}_{n-1}(t) - F^{N-1}_n(t)]. \tag{B.1}$$

y las cantidades definidas $F^0_0(t) = 1$*,* $F^N_0(t) = 1$*,* $F^{N-1}_N = 0$*, para* $N = 1, 2, , , ,$ *y* $0 \leq t < \infty$*.*

Proof. Observe que.

$$\{Q^N_{(n)} \leq t\} \Leftrightarrow \{Q^N_{(n)} \leq t\} \bigcup \left(\{Q^{N-1}_{(n-1)} \leq t \leq Q^{N-1}_{(n)}\} \bigcap \{Q_N \leq t\} \right)$$

De modo que

$$P(Q^N_{(n)} \leq t) = P(Q^{N-1}_{(n)} \leq t) + (P(Q^{N-1}_{(n-1)} \leq t) - P(Q^{N-1}_{(n)} \leq t))P(Q_N \leq t)$$

Esto, es equivalente a B.7 ■

La probabilidad de que el elemento N pertenesca a la muestra S es

$$\pi_N = P(N \in S) = P(Q^{N-1}_{(n)} > Q_N) = \int_0^\infty (1 - F^{N-1}_n(t)) f_N(t) dt \tag{B.2}$$

Donde f_N es la función de densidad de Q_N, es sencillo veficar que $f_N = \theta/(1+\theta_N)^2$

Para encontrar cualquier otro valor π_i, bastará con reordenar las variables Q_i

$$Q_1, Q_2, ...Q_{i-1}, Q_{i+1}, ..., Q_N, Q_i$$

Para las de segundo orden, se utiliza una forma similar

$$\pi_{N-1,N} = P(N-1 \in S, N \in S) = P(Q_{(n-1)}^{N-2} > max(Q_{N-1}, Q_N))$$

que es igual a

$$\int_0^\infty (1 - F_{n-1}^{N-2}(t)) f_{max(Q_{N-1},Q_N)}(t)dt \tag{B.3}$$

B.2. Aproximación de Laplace

Utilizando las siguientes integrales

$$c_k^{'} = \frac{\int_0^\infty x^{n-1}(\prod \frac{1+r_j}{1+r_j x})\frac{1}{1+r_k x}dx}{\int_0^\infty x^{n-1}(\prod \frac{1+r_j}{1+r_j x})dx}$$

$$c_o = \int_0^\infty x^{n-1}(\prod \frac{1+r_j}{1+r_j x})dx$$

$$c_k = \int_0^\infty x^{n-1}(\prod \frac{1+r_j}{1+r_j x})\frac{1}{1+r_k x}dx$$

La siguientes relaciones puede ser derivadas

1. $c_k = \frac{c_o}{1+r_k} + r_k \frac{\partial c_o}{\partial r_k}$
2. $\sum c_k = (N-n)c_o.$

Bajo la condición de que $\sum p_k = n$. Se evalua $r_k \frac{\partial c_o}{\partial r_k}$ y usando la relación $r_k \frac{x}{1+r_k x} = 1 - \frac{1}{1+r_k x}$ y reorganizando terminos se llega a 1.

Para llegar a 2 usamos el hecho de que $\frac{1}{1+r_k x} = 1 - r_k \frac{x}{1+r_k x}$ y obtenemos que

$$\sum c_k = N c_o - \int_0^\infty x^{n-1} \prod \frac{1+r_i}{1+r_i x} \sum \frac{r_k x}{1+r_k x} dx$$

Por medio de una integración por partes se tiene

$$\sum c_k = N c_o + \int_0^\infty x^n \frac{d}{dx} \prod \frac{1+r_i}{1+r_i x} dx = (N-n) c_o$$

c_o puede ser aproximada por medio de una aproximación de Laplace (ver e.g Kass, 1997). Sustituyendo $x = e^y$ reescribimos a c_o como

$$c_o = \int_{-\infty}^\infty exp\{g_o(y)\} dy$$

donde

$$g_o(y) = ny + \sum log(\frac{1+r_i}{1+r_i e^y})$$

Con $g_o(0) = 0$ y $g_o(y)$ tiene un máximo en $y = 0$. Y

$$g_o'(y) = n - \sum \frac{r_i e^y}{1+r_i e^y}$$

entonces

$$g_o'(0) = n - \sum \frac{r_i}{1+r_i} = n - \sum p_i = 0.$$

Usando la expansión de Taylor de segundo orden alrededor de y=0

$$g_o''(0) = -\sum \frac{r_i}{(1+r_i)^2} = -\sum p_i(1-p_i) = -d$$

Poniendo a $\sigma_o^2 = 1/d$, tenemos que

$$c_o \approx c_o^* = \int_{-\infty}^\infty exp\{-\frac{y^2}{2\sigma_o^2}\} dy = \sqrt{2\pi}\sigma_o = \sqrt{\frac{2\pi}{d}}$$

Dividiendo c_k por c_o usando esta última expresión en la forma $c_o \approx log(\sqrt{2\pi/d})$ y derivando con respecto a r_k, cambiando despues por terminos de p_i y simplificando tenemos:

$$\frac{c_k}{c_o} = \frac{1}{1+r_i} + r_k \frac{\partial log c_o}{\partial r_k} \approx (1-p_i)(1 + \frac{p_k(p_k - 1/2)}{d}) \tag{B.4}$$

Para d no tan pequeño se tiene que $c_k \propto 1 - p_i$. Lo que implica que Pareto es muy cercano al diseño Sampford.

Para d pequeño, puede significar que n es pequeño o cercano a N, entonces Pareto y Sampford difieren un radio de

$$J_k = \frac{c_k}{c_o(1-p_i)}$$

Utilizando esto último y B.4 tenemos:

Aproximación 1.

$$\frac{c_k}{c_o} \approx (1-p_k)J_k^*$$

con

$$J_k^* = 1 + 1 + \frac{p_k(p_k - 1/2)}{d}$$

Hay otras aproximaciones, las constantes c_k pueden ser evaluadas directamente por aproximación de Laplace. Tenemos:

$$c_k = \int_{-\infty}^{\infty} exp\{g_o(y) - log(1 + r_k e^y)\}dy \tag{B.5}$$

Expandiendo por Taylor alrededor de $y = 0$, y sustituyendo $\sigma_k^2 = \frac{1}{(d+p_k(1-p_k))}$ y procediendo como antes tenemos que:

$$c_k \approx c_k^* = \int_{-\infty}^{\infty} exp\{-log(1+r_k) - y\frac{r_k}{1+r_k} - \frac{y^2}{2\sigma_k^2}\}dy$$

Utilizando la función generatriz de una distribución normal y el hecho de que $\frac{r_k}{1+r_k} = p_k$ se tiene que:

$$c_k^* = (1-p_k)\sqrt{2\pi}\sigma_k exp\{\frac{\sigma_k p_k^2}{2}\} \tag{B.6}$$

Notamos que se cumple $c_k \propto 1-p_i$ aproximadamente si $\sigma_k^2 s$ son pequeños (i.e d grande). La ecuación (7.6) puede ser calibrado para mejorar por medio de la relación $\sum c_k = (N-n)c_o$.

$$c_k^{*cal} = \frac{(N-n)c_k^*}{\sum c_i^*}c_o$$

Esta última expresión puede ser escrita como una aproximacion 2

Aproximación 2

$$(\frac{c_k}{c_o})^{*(cal)} = (1-p_k)\frac{(N-n)\sigma_k exp\{\sigma_k^2 p_k^2/2\}}{\sum(1-p_i)\sigma_i exp\{\sigma_i * 2p_i^2/2\}}$$

Una aproximación puede ser obtenida expandiendo la función $g_k(y) = g_o(y) - log(1+r_k e^y)$ en (7.6) alrededor de y_{max} que es cercano a $-p_k\sigma_k^2$. Se obtiene

$$c_k^{**} = \sqrt{2\pi}\widetilde{\sigma}_k exp\{g_k(y_{max}\}$$

donde $\widetilde{\sigma}_k^2 = -1/g_k^{''}(y_{max})$.

B.3. Función de probabilidad de un diseño Pareto

Lema 3 *considere la secuencia de variables $Q_1, Q_2, .., Q_N$ con función de probabilidad $F_1, F_2, .., F_N$, y sea $Q_{(n)}^N$ el n ésimo orden estadístico de las cantidades*

$Q_{(1)}, Q_{(2)}, .., Q_{(N)}$ *con función de distribución* F_n^N. *Entonces para* $N = 1, 2, , , n = 1, 2, , N$, *la función* F_n^N *cumple la ecuación recursiva:*

$$F_n^N((t) = F_n^{N-1}(t) + F_N(t)[F_{n-1}^{N-1}(t) - F_n^{N-1}(t)]. \tag{B.7}$$

y las cantidades definidas $F_0^0(t) = 1$, $F_0^N(t) = 1$, $F_N^{N-1} = 0$, *para* $N = 1, 2, , , ,$ *y* $0 \leq t < \infty$.

Proof. Observe que.

$$\{Q_{(n)}^N \leq t\} \Leftrightarrow \{Q_{(n)}^N \leq t\} \bigcup \left(\{Q_{(n-1)}^{N-1} \leq t \leq Q_{(n)}^{N-1}\} \bigcap \{Q_N \leq t\} \right)$$

De modo que

$$P(Q_{(n)}^N \leq t) = P(Q_{(n)}^{N-1} \leq t) + (P(Q_{(n-1)}^{N-1} \leq t) - P(Q_{(n)}^{N-1} \leq t))P(Q_N \leq t)$$

Esto, es equivalente a B.7 ■

La probabilidad de que el elemento N pertenezca a la muestra S es

$$\pi_N = P(N \in S) = P(Q_{(n)}^{N-1} > Q_N) = \int_0^\infty (1 - F_n^{N-1}(t)) f_N(t) dt \tag{B.8}$$

Donde f_N es la función de densidad de Q_N, es sencillo verificar que $f_N = \theta/(1 + \theta_N)^2$

Para encontrar cualquier otro valor π_i, bastará con reordenar las variables Q_i

$$Q_1, Q_2, ...Q_{i-1}, Q_{i+1}, ..., Q_N, Q_i$$

Para las de segundo orden, se utiliza una forma similar

$$\pi_{N-1,N} = P(N - 1 \in S, N \in S) = P(Q_{(n-1)}^{N-2} > max(Q_{N-1}, Q_N))$$

que es igual a

$$\int_0^\infty (1 - F_{n-1}^{N-2}(t)) f_{max(Q_{N-1}, Q_N)}(t) dt \tag{B.9}$$

Referencias

[1] Aires,N.(2000). Techniques to calculate exact inclusion probabilities for condicional Poisson sampling and Pareto πps sampling designs. Department of Mathematical Statistics, Göteborg University,Sweden.

[2] Aires, N., Jonasson, J. Nerman, O.(2002). Order sampling design with prescribed inclusion probabilities. Scand J. Statist. 29, 183-187.

[3] Bondesson,L.,Traat,I.(2005). On a matrix with integer eigenvalues and its relation to conditional Poisson sampling. Res. Lett. Inf. Math. Sci. 8, 155-163.

[4] Bondesson,L.,Traat,I.,Lundqvist,A.(2006). Pareto sampling versus Sampford and condicional Poisson sampling. Scand,J .Stat. 33, 699-720.

[5] Chen, X., Dempster, A. P., Liu, J. S. (1994). Weighted finite population sampling to maximize entropy, Biometrika, Vol. 81, No. 3. (Aug., 1994), pp. 457-469.

[6] Ghosh,D. Vogt,A. : A fixed sample size variant of Poisson sampling.

[7] Hajek, J.(1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann. Math. Statist. 35, 1491-1523.

[8] Lundqvist,A.(2007). On the distance between some πps sampling designs. Springer, New York.

[9] Sampford,M.R.(1967). On samplingn without remplacement with unequal probabilities of selection.Biometrika 54,499-513.

[10] Tillé, Y.(2006), Sampling Algorithms,Springer-Verlag, New York.

[11] Lundqvist,A.,Lennart Bondesson, On sampling with desired inclusion probabilities of first and second order,Umea University.

[12] Anton Grafstrom , Comparisons of methods for generating conditional Poisson samples and Sampford samples,Masterś thesis in Mathematical Statistics, June 2005.

[13] Carl-Erik Srndal, Model Assisted Survey Sampling, Springel Series in Statistics ed. 2003

[14] Bustos (2007), Exact expressions for simple and joint inclusion probabilities in sampling without replacement, contribucion a ISI (International Statistical Institute, 2007)