Métodos de Monte Carlo

Entrega 4

(Unidad 3, Sesión 7) Ejercicio 7.1

Integrantes

Nombre	CI	Correo
Germán Ouviña	4.823.566-1	german.ouvina@fing.edu.uy

Índice

1. Problema 3

2. Solución 4

2.1. Descripción 4

2.2. Pseudocódigo 5

2.3. Código 6

3. Experimentación 6

3.1. Especificaciones 6

3.2. Resultados 7

1. Problema

Entrega 4: Ejercicio 7.1 (INDIVIDUAL):

Para diseñar un Sistema Nacional de Areas Protegidas, uno de los modelos empleados tiene en cuenta por un lado un conjunto $E = 1, .., e$ de especies que se desea preservar, y por otra parte un conjunto Z de zonas donde es posible establecer una reserva. La relación entre ambos conjuntos está dada por una matriz $P = ((p_{ij}))$, con $i \in Z$ y $j \in E$, tal que $z_{ij} = 1$ ssi en la zona i se ha observado la presencia de la especie j.

Es interesante elegir un conjunto de M zonas, tales que todas las especies estén representadas en al menos una zona. La determinación del menor M que hace posible esta propiedad es un problema de "set covering" (NP-difícil), que escapa el alcance de este curso.

Suponiendo que el valor de M ya fue elegido, un segundo nivel es ver cuántas formas distintas hay de elegir este conjunto de zonas.

Para esto, es posible aplicar el método Monte Carlo para, dado el cardinal de E, el cardinal de Z, la matriz $P = ((p_{ij}))$, y un valor de M predeterminado, estimar cuántas combinaciones de M zonas distintas cumplen la propiedad requerida (representar todas las especies).

Se debe recibir en entrada el número de replicaciones a realizar, y el nivel de confianza; en salida, se debe dar la estimación del número de combinaciones $N_C(M)$, así como la desviación estándar y un intervalo de confianza (del nivel especificado) calculado en base al criterio de Agresti-Coull.

1. Escribir un programa para hacer el cálculo previamente descrito. Entregar seudocódigo y código.
2. Sea el siguiente caso: $z = 15$ zonas, $e = 8$ especies, y la matriz provista en la consigna. Usando el programa anterior, y empleando 1000 replicaciones de Monte Carlo, estimar los valores de $N_C(M)$ para $M = 5$ y para $M = 6$ con intervalos de confianza de nivel 95 %.

2. Solución

2.1. Descripción

Para resolver este ejercicio, se buscó adaptar el método de Monte Carlo utilizado en ejercicios anteriores a la realidad concreta del conteo de subconjuntos. Dado el conjunto base X y el conjunto de objetos $F = S_1, ..., S_k$ definido en el teórico, el objetivo primordial antes de resolver el problema fue plantear la consigna en estos términos.

Se desea obtener la cantidad de combinaciones de M zonas (de un total de Z zonas) que, en conjunto, logran cubrir las E especies. Teniendo esto en cuenta, el conjunto base X tiene sentido que incluya toda combinación de M zonas tomadas de las Z zonas que hay en total. En otras palabras, el cardinal del conjunto base queda limitaod por $|X| = C^Z_M$, es decir, combinaciones de Z tomadas de M.

Adaptando el algoritmo ya existente, el primer cambio que se hace resulta en el sorteo que se realiza iteración a iteración. De las Z zonas, se extraen M zonas distintas entre sí con probabilidad $\frac{1}{Z}$. Para optimizar el sorteo y evitar repetidos, se creó un conjunto de zonas Z', del cual se extraen las zonas a medida que se sortean. Por ejemplo, si la zona z_1 sale sorteada en el primer paso, se actualiza $Z' = Z - \{z_1\}$ y en el próximo paso, se sortea una zona z_2 de Z' con probabilidad $\frac{1}{|Z'|} = \frac{1}{Z-1}$. Así sucesivamente hasta obtener las M zonas distintas.

Una vez hecho esto, el proceso es sencillo: se chequea si para las M zonas se cubren las E especies y en caso afirmativo, se acumula (como en el caso del volumen, por ejemplo). Finalmente, las operaciones con los estimadores se modifican levemente para incluir la información de $|X|$.

En la sección 2.2 se entra en mayor detalle sobre las características del algoritmo y del código en general. En la sección 3.2 se detallan los resultados obtenidos para la parte b y demás comparativas de experimentación.

2.2. Pseudocódigo

Procedimiento EstimacionMonteCarloConteo (int n, int E, int Z, int M, real δ)
Parámetros de entrada:

- n - tamaño de la muestra
- E - cantidad de especies
- Z - cantidad de zonas
- M - cantidad de zonas a elegir
- δ - nivel de confianza

Parámetros de salida:

- $\hat{X}$ - estimador del conteo
- $\hat{V}$ - estimador de la desviacion del estimador puntual $\hat{X}$
- ω_1, ω_2 - intervalo de confianza para $(1-\delta)$

Pseudocódigo:

1. $S, \hat{X}, \hat{V} = 0$ // Inicializar acumulador y estimadores
2. Calcular $X = C_M^Z$ // Calcular cardinal total de conjunto base X
3. For $i = 1, ..., n+1$ do
   a) $Z' = Z$ // Inicializar conjunto auxiliar de zonas para sorteo
   b) $M', E' = \{\}$ // Inicializar conjuntos auxiliares para acumular zonas y especies
   c) For $m = 0, ..., M$ do
      1) Sortear zona $z \in Z'$ con probabilidad $\frac{1}{Z-m}$
      2) Actualizar $M' = M' \cup \{z\}$
      3) Actualizar $Z' = Z' - \{z\}$
   d) For $z = 0, ..., M$ do // Iterar según zonas sorteadas
      1) For $e = 0, .., E$ do // Iterar según total de especies
         a' Si $P[z][e] = 1 \to E' = E' \cup \{e\}$ // Si la especie e se encuentra en la zona z, agregarla al conjunto de especies auxiliar
4. $\hat{X} = X * S/n$
5. $\sqrt{\hat{V}} = \sqrt{\hat{X} * (X - \hat{X})/(n-1)}$
6. $\omega_1, \omega_2 = I_1(S, X, n, \delta), I_2(S, X, n, \delta)$

El anterior pseudocódigo bosqueja un método de monte carlo para el conteo de combinaciones de conjuntos, siguiendo las pautas del pseudocódigo presentado en clase. Además de los cambios mencionados en la sección anterior en relación al sorteo de zonas, se modifica el calculo de estimadores $\hat{X}$ y $\hat{V}$, para incluir el cardinal del conjunto base X como parámetro. Cabe también destacar que el cálculo del intervalo de confianza en esta instancia vuelve a utilizar el valor del acumulador S.

2.3. Código

El código para solucionar el problema se encuentra adjunto en un archivo *.zip* junto al informe, y se encuentra disponible en el repositorio `https://github.com/gouvina-fing/fing-mmc2025`.

El mismo fue desarrollado en *Python*, cuenta con instrucciones de como instalarlo y correrlo, y ofrece la posibilidad de generar tablas y gráficas comparativas siguiendo la consigna del ejercicio, automáticamente ejecutando el método para los distintos valores de n y M mencionados.

3. Experimentación

3.1. Especificaciones

A continuación, se desglosan las especificaciones técnicas utilizadas durante la ejecución:

- **Procesador:** 11th Gen Intel(R) Core(TM) i5-11400F @ 2.60GHz
- **Memoria RAM:** 16.0 GB
- **Sistema Operativo:** Windows 10
- **Semilla:** 42
- **Tamaños de muestra:** 10^3, 10^4, 10^5, 10^6

3.2. Resultados

Se ejecutó el algoritmo para los valores de n y M pedidos, utilizando los valores de $Z = 15$ y $E = 8$, así como la matriz provista por la consigna. Para estudiar el comportamiento asintótico de los estimadores, se ejecutó el algoritmo para tamaños de muestra n mayores y se graficaron los resultados. Los valores resultaron esperados como de costumbre.

A continuación, se adjuntan las tablas comparativas para $M = 5$ y $M = 6$ marcando en rojo los valores de interés según la consigna:

N° de iteraciones (n)	**Estimador ($\hat{X}$)**	**Desviación ($\sqrt{\hat{V}}$)**	**IdeC (Agresti-Coull)**	**Tiempo (s)**
10^3	2792.7900	24.2417	2733.6097 2830.5994	0.02
10^4	2800.5978	7.5293	2784.7326 2814.3124	0.22
10^5	2806.0932	2.3506	2801.3773 2810.5937	2.15
10^6	2810.4536	0.7356	2809.0010 2811.8846	19.95

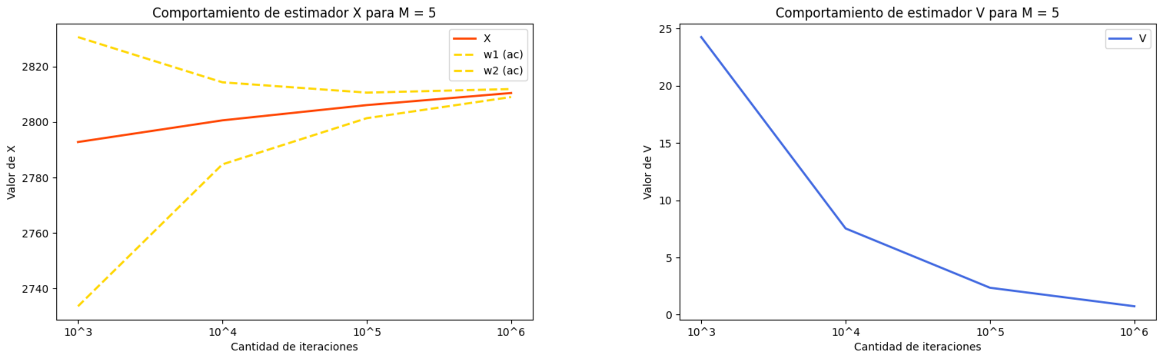

Figura 3.1: Comportamiento de los estimadores X y V con M = 5

N° de iteraciones (n)	**Estimador ($\hat{X}$)**	**Desviación ($\sqrt{\hat{V}}$)**	**IdeC (Agresti-Coull)**	**Tiempo (s)**
10^3	4909.9050	21.6188	4844.9492 4937.2865	0.02
10^4	4906.9020	6.9383	4891.2950 4918.7408	0.24
10^5	4910.4555	2.1546	4906.0398 4914.4940	2.45
10^6	4913.0531	0.6721	4911.7168 4914.3517	23.60

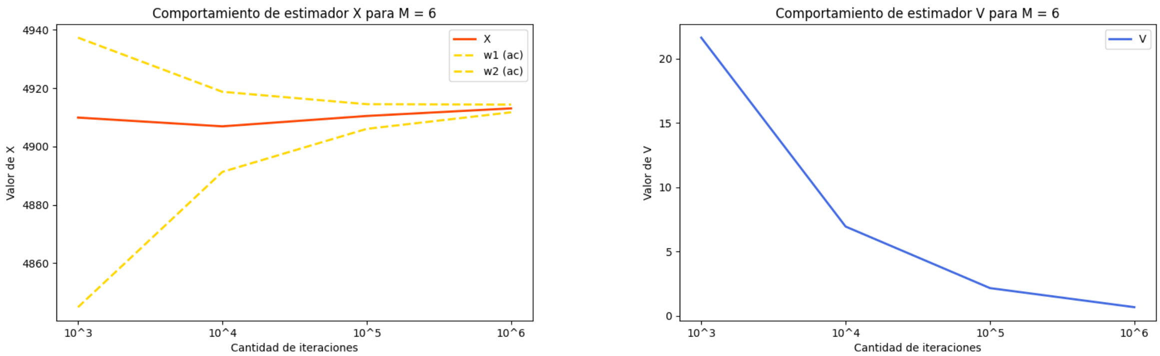

Figura 3.2: Comportamiento de los estimadores X y V con M = 6