# Estimadores

Rodrigo Zepeda

ENTREGA: 13 OCTUBRE

Se presentan 4 opciones de proyecto. Sólo puede haber tres equipos por opción según se rellene en la encuesta en https://forms.gle/DvQG5rU9LX7oLgoj9. Las opciones que siguen disponibles pueden verse acá: https://shorturl.at/dCEI9 Dentro de cada opción lo que tienen que hacer es diseñar estrategias de muestreo. Para ello en un `pdf` ó `html` deben:

1. Establecer el problema que se está resolviendo.
2. Resolver un problema por persona. Hay un encargado de cada inciso (son 3 incisos) y los demás deben de revisarlo pero el encargado es la persona que, de estar mal el inciso, pierde/gana toda la puntuación.
3. Reportar todo a computadora. Sin que haya pasos mágicos.
4. Pueden ayudarse de la librería caracas de `R` o de `sympy` (Python), `Sage`, `Wolfram` ó `Symbolab` para hacer la talacha.
5. Si se usa información de otra fuente debe ir citada.
6. Argumentar con cuál estimador se quedan y bajo

qué criterios. Aquí se puede hacer de dos opciones:

   (a) Calculando todas sus propiedades a mano. Por ejemplo calculando el sesgo como hemos hecho en clase.

   (b) Mostrando que a mano llegan a un punto a partir del cuál no se puede de manera exacta (por ejemplo llegan a una integral que no tiene forma cerrada como la acumulada de la normal) y entonces realizando simulaciones para ver bajo distintos escenarios cómo se comporta el estimador.

7. Este documento habla de código en R pero si prefieren pueden usar Python, Julia, Javascript, Java, Matlab, Mathematica, Go, C++, C#, Octave, Stata, LISP (mientras sea ANSI) o Bash. En caso de que elijan un programa que necesite compilación yo requiero las opciones que especificaron en el compilador.

8. En cada inciso no hay una única solución. Lo que para algunas personas puede ser un buen estimador *por sus criterios* puede ser un mal estimador para otras personas *que usaron otro criterio*.

9. Todos los muestreos realizados son aleatorios simples a menos que se diga lo contrario.

10. A menos que se especifique en el inciso, no se puede suponer independencia entre las variables.

**TODOS LOS EQUIPOS DE MÁS DE UNA PERSONA DEBEN INDICAR DE CUÁL INCISO SE ENCARGÓ QUIEN (mínimo un inciso por persona), QUIEN SE ENCARGÓ DE LA ESCRITURA DEL TEXTO, QUIÉN SE ENCARGÓ DE LA EDICIÓN Y ENTREGA DEL TEXTO Y QUIÉN SE ENCARGÓ DE LA PROGRAMACIÓN DE QUÉ**

# 1 Encuestas panel

## 1.1 Problema

Interesa cuantificar cómo cambia la estatura de los niños mexicanos desde que tienen 2 hasta que tienen 10 años. Para ello se tomará una muestra de niños que nacieron en 2019 y se les seguirá en el tiempo midiéndolos año con año hasta que tengan 10.

Considera una población finita donde las variables cambian con el tiempo (por ejemplo la estatura de los niños, el peso de las personas, la duración de las baterías de litio)

$$U = \{x_1(0), x_2(0), \ldots, x_N(0)\}$$

de la cual se obtiene una muestra aleatoria $S_0$ mediante muestreo simple con $\pi_k = \mathbb{P}(x_k(t) \in S_0)$ en el tiempo $t_0 = 0$. Supongamos, además en cada momento $t = \tau$ se obtiene una muestra $S_\tau$ de tamaño $n$ dada por:

$$S = \{x_1(\tau), x_2(\tau), \ldots, x_N(\tau)\}$$

Se desea estimar la función $\mu(t)$ dada por la media de la estatura en el momento $t$:

$$\mu(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t) \qquad t \in [0, 8]$$

así como

$$\Delta_\mu(t) = \frac{1}{N}\sum_{i=1}^{N} \big(x_i(t) - x_i(t-1)\big) \qquad t \in [1, 8]$$

el cambio promedio en estatura de cada niño.

**Sólo hay presupuesto para medir $n = 2000$ niños cada año.**

Para ello se proponen distintos métodos:

1. Se seleccionan aleatoriamente sin reemplazo $n$ niños de 2 años en el primer momento (en el tiempo 0 equivalente al año 2021) y cada año se repite la medición de exactamente los mismos niños hasta que cumplen 10 años. **Ojo** Se sabe que cada año una proporción $p \in (0, 1)$ de los niños se pierden al seguimiento (es decir sus papás ya no los vuelven a llevar a que los midan los investigadores)

2. Se seleccionan aleatoriamente sin reemplazo $n$ niños de 2 años en el primer año (en el tiempo 0 equivalente al año 2021) y a partir del segundo y cada medio año se seleccionan aleatoriamente con reemplazo $\lfloor n/2 \rfloor$ niños de esa muestra inicial. Se repite el proceso de medición cada seis meses seleccionándose una nueva muestra aleatoria sin reemplazo hasta que cumplen 10 años. Se sabe que cada año una proporción $p \in (0, 1)$ de los niños se pierden al seguimiento (es decir sus papás ya no los vuelven a llevar a que los midan los investigadores).

3. Se seleccionan $n$ niños de 2 años en el primer momento sin reemplazo. Al año siguiente se seleccionan otros $n$ niños de la población (posible pero no necesariamente distintos) sin reemplazo. Y así sucesivamente cada año.

**Proyecto** Construye los estimadores para $\mu(t)$ y $\Delta_\mu(t)$ y determina cuál es el mejor usando como criterios: funciones de pérdida, varianza, consistencia (de Fisher), sesgo. Genera sus intervalos de confianza para cada $t$. Recuerda que $\mu(t)$ y $\Delta_\mu(t)$ tienen la $t \in \mathbb{R}$ pero las mediciones sólo se hacen en momentos determinados entonces también vas a tener que ver cómo extrapolas esos valores. Finalmente la función de `R` tiene que ser así:

```
#Funci n para estimar mu y Delta
#INPUT
#datos es una lista de data.frames de cada una de las mediciones que se hicieron
panel(datos)
```

lo que devuelve la función `panel` es una lista de dos funciones una para $\mu$ y una para $\Delta_\mu$

# 2 Encuestas censuradas

## 2.1 Problema

Interesa cuantificar el tiempo de sobrevida de un paciente a partir de su diagnóstico con una enfermedad. Se sabe que dicha enfermedad reduce el tiempo de vida de las personas (respecto a la media de los no enfermos) e interesa determinar cuánto tiempo vivirá después de su diagnóstico. Finalmente suponemos que todas las personas eventualmente mueren (independientemente de la causa).

Considera una población finita de pacientes $x_i$ los cuales tienen asociados su tiempo de sobrevida (*i.e.* cuánto tiempo tardan en fallecer a partir del momento de diagnóstico)

$$U = \{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\}$$

De esta población se obtiene una muestra aleatoria $S$ mediante muestreo simple con $\pi_k = \mathbb{P}((x_k, t_k) \in S)$. Estas personas son los recién diagnosticados. Cada día se les llama a los cuidadores de los pacientes para ver cómo han evolucionado en su enfermedad. En el caso de que se reporte que el paciente $i$

haya fallecido se registra $t_i = \tau_i$ el día en el que falleció (medido a partir del tiempo $t_0$ de diagnóstico). Por ejemplo, si un paciente fue diagnosticado hace veinte días y en la llamada de hoy nos dicen que murió entonces su $t_i$ es 20. En teoría obtendríamos la muestra conjunta de los pacientes y el tiempo que tardaron en fallecer:

$$S = \{(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)\}$$

donde un ejemplo sería $x_i =$ Rod y $t_i = 10$ si el paciente Rod se murió a los 10 días de su diagnóstico.

**PERO...** La vida real no es tan bonita. En particular, cada día los cuidadores tienen una probabilidad $p \in [0, 1)$ (constante desconocida) de dejar de contestar el teléfono para siempre incluso antes de que el paciente muera por lo que para $n - \ell$ pacientes se registra no el momento en el que murieron sino el último momento en el que se supo que estaban vivos antes de que el cuidador dejara de contestar el teléfono. En este escenario representamos la muestra como:

$$S = \{(x_1, t_1), (x_2, t_2), \ldots, (x_\ell, t_\ell), (x_{\ell+1}, \nu_{\ell+1}), (x_{\ell+2}, \nu_{\ell+2}), \ldots, (x_n, \nu_n)\}$$

donde $\nu_j$ representa el último momento en el que se contactó al cuidador del paciente (y se sabe que $t_j > \nu_j$).

1. Una opción para estimar el tiempo de sobrevida de los pacientes (*definido como* $\mu = \frac{1}{N}\sum_{i=1}^{N} t_i$ es quitar a los se perdieron en el seguimiento y estimar la media sólo con los que sí se registraron completos:

   $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{\ell} t_i$$

   (recuerda que $\ell$ es aleatorio pues depende de $p$). Determina su error cuadrático medio, sesgo y varianza así como su consistencia. Recuerda que van a depender de $p$ no de $\ell$.

2. Se construye la función de supervivencia empírica dada por:

   $$S_n(t) = 1 - F_n(t)$$

   donde

   $$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[\tau_i, \infty)}(t)$$

   con $\tau_i = t_i$ ó $\tau_i = \nu_i$ según el que se haya medido. Verifica la consistencia (Fisher), sesgo y varianza de $S_n(t)$.

3. Finalmente, utiliza que a partir de la función $S_n(t)$ se puede calcular la media (recuerda que hay formas de usar la distribución acumulada para calcular la media *sin derivar porque esto no se puede derivar en todo su dominio y en donde sí es cero*) para construir un estimador de $\mu$.

**Proyecto** Construye los estimadores para $\mu$ y determina cuál es el mejor usando como criterios: funciones de pérdida, varianza, consistencia (de Fisher), sesgo ya sea de manera directa o mediante experimentos numéricos. Genera sus intervalos de confianza (bootstrap). Finalmente construye una función de `R` para realizar la estimación. La base de datos que se le pasaría sería así donde que no haya dato en `Tiempo de fallecimiento` significa que se perdió durante el seguimiento.

| Persona | Tiempo de fallecimiento (días) | Último tiempo registrado |
|---|---|---|
| 1 | 20 | 20 |
| 2 | 12 | 12 |
| 3 | | 87 |
| 4 | 30 | 30 |
| 5 | 100 | 100 |
| 6 | | 20 |
| 7 | | 8 |
| 8 | 15 | 15 |

Table 1: Para encuestas censuradas

# 3 Encuestas espaciales

## 3.1 Problema

Se desea estimar la cantidad promedio de contaminantes en una ciudad. Por simplicidad consideraremos que la ciudad puede representarse como un rectángulo de lados $a$ y $b$. La pregunta es dónde realizar las mediciones (suponiendo que cualquier punto en el rectángulo $[0, a] \times [0, b]$ está disponible). Por presupuesto, a lo más se pueden poner $n = 100$ instrumentos de medición de contaminantes.

Para cada punto $(x, y)$ en el intervalo $[0, a] \times [0, b]$ se tiene $c_{x,y}$ la concentración de contaminantes en ese punto. Estamos en el caso de un universo infinito donde:

$$U = \{c_{x,y} : (x, y) \in [0, a] \times [0, b]\}$$

Se define la cantidad promedio de contaminantes como el promedio de una uniforme en el rectángulo:

$$\mu = \int_0^a \int_0^b c(x, y) \frac{1}{a} \frac{1}{b} dy dx$$

Se proponen las siguientes formas para decidir dónde (qué coordenadas $x, y$ muestrear):

1. **Partición equitativa** Se divide al intervalo $[0, a]$ en 10 pedazos: $A_{10} = \{0, \frac{a}{9}, \frac{2a}{9}, \frac{3a}{9}, \ldots, \frac{9a}{9}\}$ y lo mismo para el intervalo $b$: $B_{10} = \{0, \frac{b}{9}, \frac{2b}{9}, \frac{3b}{9}, \ldots, \frac{9b}{9}\}$. Se registran los promedios de contaminantes en todas las $c$ con coordenadas en $A_{10} \times B_{10}$; es decir la muestra no es aleatoria y es:

$$S = \{c_{x,y} : (x, y) \in [0, a] \times [0, b]\}$$

   Se obtiene el promedio de dichas $c$ y con eso se estima $\mu$. Construye una función de R que asigne la muestra aleatoria bajo este esquema.

2. **Partición Uniforme** Para obtener un punto donde medir los contaminantes se simula $X \sim \text{Uniforme}(0, a)$ y $Y \sim \text{Uniforme}(0, b)$ obteniéndose así el punto $(X, Y)$ en el cual se coloca $c_{X,Y}$. Se obtiene el promedio de dichas $c$ y con eso se estima $\mu$. Construye una función de R que asigne la muestra aleatoria bajo este esquema.

3. **Partición Uniforme Polar** Por facilidad se considera el rectángulo alternativo con vértices en $(\pm a/2, \pm b/2)$ el cual es sólo un desplazamiento del $[0, a] \times [0, b]$. Se simula el ángulo de manera uniforme entre 0 y $2\pi$. Finalmente se estima el radio (como función del ángulo). Se transforman las coordenadas a cartesianas y se coloca $c_{x,y}$ en los puntos resultantes. Se obtiene el promedio de dichas $c$ y con eso se estima $\mu$. Construye una función de R que asigne la muestra aleatoria bajo este esquema.

**Proyecto** Construye los estimadores para $\mu$ y determina cuál es el mejor esquema de muestreo usando como criterios: funciones de pérdida, varianza, consistencia (de Fisher), sesgo ya sea de manera directa o mediante experimentos numéricos. Finalmente construye una función de `R` para realizar el muestreo en cada escenario. La función de R debe tener como inputs: 1) las coordenadas del rectángulo donde está la ciudad, 2) el método que se desea implementar y 3) la $n$. Debe devolver una matriz o data frame de los $n$ puntos en coordenadas $x, y$.

```r
#Todas estas deben de funcionar tu funci n debe llamarse muestreo.
muestreo(largo = c(0,5), ancho = c(0,6), metodo = "Polar", n = 100)
muestreo(largo = c(2,8), ancho = c(-2,5), metodo = "Uniforme", n = 302)
muestreo(largo = c(-1,22), ancho = c(10,24), metodo = "Equitativa", n = 821)
```

# 4 Encuestas con ruido

## 4.1 Problema

Se realizó la medición de ciclos que ejecuta un CPU por segundo (*i.e.* cuántas operaciones hace). Para ello se utiliza un instrumento que no mide los Gigahertz ($g$) de manera exacta sino con un error $\epsilon$ . Es decir nuestro instrumento nos devuelve $x = g + \epsilon$ donde $g$ son los verdaderos Gigahertz y $\epsilon$ es el error de medición. Lo que realmente nos interesa es conocer $g$.

Considera una población:

$$U = \{(g_1, \epsilon_1), (g_2, \epsilon_2), \ldots, (g_N, \epsilon_N)\}$$

con $\bar{\epsilon} = \frac{1}{N}\sum_{i=1}^{N} \epsilon_i$ y $\bar{g} = \frac{1}{N}\sum_{i=1}^{N} g_i$. Se obtiene una muestra aleatoria de $x_i = g_i + \epsilon_i$ tamaño $n$:

$$S = \{x_1, x_2, \ldots, x_n\}$$

Interesa estimar $g$.

1. En particular si suponemos que $\bar{\epsilon} = m$ es una constante conocida, obtén un estimador insesgado de $\bar{g}$. ¿Cómo se ve su varianza? Con $N = 1000$ y $n = 100$ realiza un ejemplo en `R` de cómo se vería la estimación en este caso.

2. Si suponemos que un cierto porcentaje $p < 0.5$ de las $\epsilon_i$ cumplen que $|\epsilon_i| > \max_{g \in U}\{|g|\}$ y el resto $(1-p)$ de las $\epsilon_i$ son cero entonces se propone el estimador:

   $$\bar{g}_{\text{Trunc}} = \frac{1}{\lfloor (1-p)n \rfloor - \lceil pn \rceil} \sum_{i=\lceil pn \rceil}^{\lfloor (1-p)n \rfloor} g_{(i)}$$

   siempre y cuando $\lfloor (1-p)n \rfloor > \lceil pn \rceil$. ¿Es insesgado? ¿Cómo se ve su varianza (si no te sale a mano justifica por qué no y realiza experimentos numéricos)? Con $N = 1000$ y $n = 100$ realiza un ejemplo en `R` de cómo se vería la estimación en este caso.

3. Si suponemos que las $\epsilon_i$ son realizaciones de una variable aleatoria $E$ simétrica en torno a 0 y que las $g_i$ son simétricas en torno a $m$ se propone el estimador:

   $$\hat{g}_{\text{Med}} = \text{Mediana}(S)$$

   ¿Es insesgado? ¿Cómo se ve su varianza (si no te sale a mano justifica por qué no y realiza experimentos numéricos)? Con $N = 1000$ y $n = 100$ realiza un ejemplo en `R` de cómo se vería la estimación en este caso.

**Proyecto** Construye los estimadores para $\bar{g}$ y determina cuál es el mejor esquema de muestreo usando como criterios: funciones de pérdida, varianza, consistencia (de Fisher), sesgo ya sea de manera directa o mediante experimentos numéricos (según el caso, hay 3 casos).

# 5 Encuestas y medida

## 5.1 Problema

Este problema no se sugiere hacerlo. Es sólo para las personas que les gusta mucho la teoría matemática (muy abstracta) detrás de las cosas de

estadística. Se sugiere saber, además, conceptos de convergencia de variables aleatorias (proba 2).

**Proyecto** En máximo 5 páginas explicar qué es la derivada de Radon-Nikodym y cómo surge en muestreo por importancia (*importance sampling*).

### 5.2 EVALUACIÓN ESPECIFICA PARA ESTE PROYECTO

1. Explicar qué es muestreo por importancia y para qué se usa **1 pts**
2. Explicar qué es una medida de probabilidad (esto es básicamente que cumpla axiomas de Kolmogorov). **0.5 pts**
3. Explicar continuidad absoluta de una medida respecto a otra. **1 pts**
4. Explicar qué es el movimiento browniano. **0.5 pts**
5. Explicar qué es la derivada de Radon-Nikodym y cómo se interpreta **1 pts**
6. Explicar cómo aparece en muestreo por importancia y cómo se usa.**2 pts**
7. No necesitas demostrar nada basta con referir a los teoremas apropiados enunciarlos y citarlos. Lo que sí necesitas es explicar qué significa lo que dice. **1 pts**
8. Realiza un ejemplo con dos medidas de probabilidad famosonas (por ejemplo una normal y una exponencial). **3 pts**

### 5.3 Referencia sugerida

Aquí http://web.math.ku.dk/ rolf/teaching/mfe03/EE.impsamp.pdf explican cómo sale y dónde se usa sin embargo no explican nada de Radon-Nikodym ni de browniano ni continuidad absoluta. Para esos conceptos te sugiero *A first look at rigorous probability* o si quieres algo más pesado *Measure Theory* de Bogachev.

**¡NO OLVIDES CITAR! TRADUCIR ALGO TAMBIÉN IMPLICA CITARLO**

## 6 EVALUACIÓN

Esto es cómo se evalúan los proyectos del 1 al 4. El 5o es distinto

1. 50% individual: Estimación a mano de las propiedades de los estimadores (al menos dos) o bien justificación clara de por qué a mano no se pudo y código reproducible de simulaciones que justifiquen el proceso.
   (a) 5% orden y presentación.
   (b) 10% que las matemáticas se puedan seguir (narración del cálculo de qué se está haciendo y con qué hipótesis).

(c) 20% por estimación correcta de los estimadores (se divide el 20 entre total de estimadores) o por una justificación adecuada de por qué no se pudo estimar y un código reproducible que haga sentido para el problema.

(d) 15% Código de `R` asociado puede correr y estima lo que debería estimar.

2. 50% grupal:

   (a) 20% las estimaciones correctas de los estimadores de los compañeros (inciso 3 del apartado individual referido a los demás). Esto es para verificar que revisaron el trabajo del resto del equipo y que los demás lo hubieran hecho bien.

   (b) 10% Redacción y presentación de la nota.

   (c) 20% Justificación de por qué el estimador que eligieron es el mejor para el problema