# Trabajo Práctico 2: Conservando el anonimato

Melanie Sclar, Sebastián Cherny

22 de septiembre de 2019

## Ejercicio 1: Estimador insesgado para el procedimiento

$\hat{p}_A = 3\overline{X} - 1$

Luego, $E[\hat{p}_A] = 3 * E[\overline{X}] - 1$ Ahora, $\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ , donde las $X_i$ son como dice el enunciado. Como cada respuesta es independiente (una tirada del dado o de la moneda no afecta otras respuestas más que la propia), podemos calcular $E[\overline{X}]$ como $\frac{1}{n} \cdot \sum_{i=1}^{n} E[X_i]$. Y la esperanza de $X_i$ no es más que la probabilidad de que la i-ésima respuesta sea 'Sí', ya que $X_i$ toma valor 1 en este caso y 0 en caso contrario.
Ahora, la probabilidad de que la respuesta sea 'sí' con este procedimiento es la probabilidad de que deba responder la verdad (dado = 1 ó 2) Y la verdad sea que probó drogas duras, sumada a la probabilidad de que NO deba respondar la verdad, pero que la moneda indique que debe responder 'sí'. Luego, $E[X_i] = \frac{2}{6} * p + \frac{4}{6} * \frac{1}{2} = \frac{p}{3} + \frac{1}{3}$ para todo $1 \leq i \leq n$.

Finalmente, $E[\hat{p}_A] = \frac{3}{n} \cdot n \cdot \left(\frac{p}{3} + \frac{1}{3}\right) - 1 = p$, por lo que es un estimador insesgado.

El valor máximo que puede tomar el estimador es 2, si todas las personas respondieran 'sí', mientras que el valor mínimo es $-1$, si todas respondieran 'no'. Los valores que puede tomar $p$, que es la probabilidad de que alguien realmente haya probado drogas duras en Capital Federal, son entre 0 y 1, como toda probabilidad, por lo que no coincide con el rango del estimador. Es decir, qué significa que nuestro estimador de como valor 2, que el 200 % de las personas probaron drogas duras?

## Ejercicio 2: ECM de distintos estimadores

Sea $Y_1, \ldots, Y_n$ una muestra de las respuestas ante la pregunta directa, es decir sin todo el procedimiento del dado y eso.
Luego, $Y_i$ tiene una distribución de Bernoulli de parámetro $p$. Entonces, $\hat{p}_{EMV} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{Y}{n}$ llamando $Y$ a la cantidad de respuestas 'sí' obtenidas , por lo que $ECM_{\hat{p}_{EMV}} = E[(\frac{Y}{n} - p)^2] = E[(\frac{Y}{n})^2] - E[2 \cdot \frac{Y}{n} \cdot p] + E[p^2] = \frac{1}{n^2} E[Y^2] - \frac{2 \cdot p}{n} \cdot E[Y] + p^2$ .
Si ahora pensamos en $E[Y^2]$, podemos usar que la varianza de una binomial es conocida, $Var(Y) = n \cdot p \cdot (1-p)$, y usando la fórmula de la varianza $Var(Y) = E[Y^2] - E[Y]^2$, obtenemos que $E[Y^2] = n \cdot p \cdot (1-p) + (n \cdot p)^2 = n \cdot p \cdot (n \cdot p + 1 - p)$. Entonces finalmente, $ECM_{\hat{p}_{EMV}} = \frac{1}{n^2} * n * p * (n * p + 1 - p) - 2p^2 + p^2 = p \cdot (\frac{1}{n} * (n * p + 1 - p) - p)$
Ahora, veamos el otro estimador: $\hat{p}_A = 3\overline{X} - 1$ , donde $\overline{X}$ es la cantidad de respuestas 'sí' obtenidas, dividida por $n$.
Pero si pensamos que las $X_i$ son Bernoullis, por el hecho que tienen únicamente dos respuestas, y notando que son independientes (una tirada de dado o moneda no afecta a ninguna otra persona), $\overline{X}$ es simplemente una Binomial, con $n$ repeticiones, y donde ahora la probabilidad de 'éxito' (es decir, que la respuesta sea 'sí') es $q = \frac{p+1}{3}$ como vimos antes.

Entonces el ECM será parecido al anterior: $ECM_{\hat{p}_A} = E[(3 \cdot X - 1 - p)^2]$ $ECM_{\hat{p}_A} = E[(3 \cdot X - 1)^2] - E[2 \cdot (3 \cdot X - 1) \cdot p] + E[p^2]$ $ECM_{\hat{p}_A} = 9 * E[X^2] - 2 * 3 * E[X] + E[1] - E[2 \cdot (3 \cdot X - 1) \cdot p] + E[p^2]$ $ECM_{\hat{p}_A} = 9 * \frac{q * (n * q + 1 - q)}{n} - 6 \cdot q + 1 - 2 \cdot p^2 + p^2$

Si analizamos la situación cuando $n = 1000$ y $p = 0{,}2$, obtenemos $ECM_{\hat{p}_{EMV}} = 0{,}00015$ y $ECM_{\hat{p}_A} = 0{,}00216$, por lo que $ECM_{\hat{p}_A} - ECM_{\hat{p}_{EMV}} = 0{,}002$

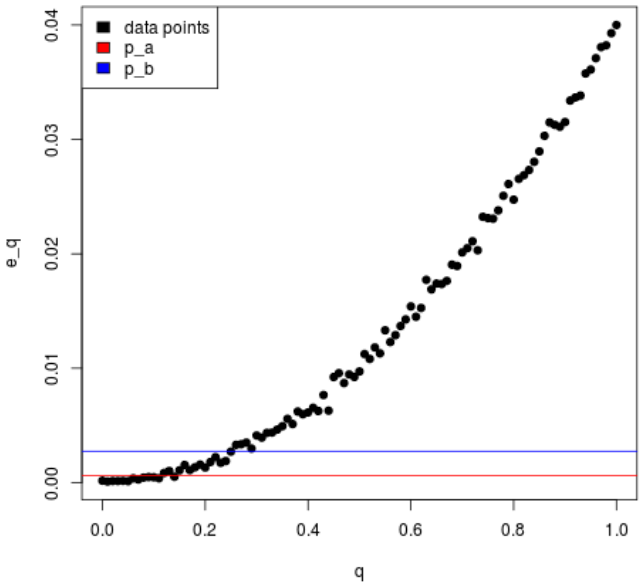
## Ejercicio 3: Nuevo estimador: truncado

El sentido del nuevo estimador $\hat{p}_B$ es truncar el estimador anterior, que viene a resolver el inconveniente mencionado en la primera parte, que el rango de valores de $\hat{p}_A$ se va de una probabilidad. Podemos ver que si $\hat{p}_A < 0$, entonces ambas indicadoras dan 0 y por lo tanto $\hat{p}_B = 0$; si $0 < \hat{p}_A < 1$ solamente la primera indicadora da 1 por lo que $\hat{p}_B = \hat{p}_A$ ; y si $\hat{p}_A > 1$ la segunda indicadora da 1 valiendo $\hat{p}_B = 1$.
La simulación Monte Carlo (con seed 1) indica que el ECM de $\hat{p}_A$ es aproximadamente 0,000613 mientras que el ECM de $\hat{p}_B$ es aproximadamente 0,00275, por lo que empíricamente vemos que el estimador truncado $\hat{p}_B$ es incluso peor que $\hat{p}_A$.

## Ejercicio 4: Simulación cuando con probabilidad $q$ mienten

Se hizo la función, modificando una función anterior para que reciba el parámetro $q$ y en base a este, cuando la respuesta de alguien sería 'sí', tirar una moneda cargada para que con proba $q$ cambie su respuesta.

## Ejercicio 5: Gráfico, ECM del EMV según probabilidad de mentir




Entendiendo por 'conviene usar un estimador frente a otro', que el estimador que conviene tiene menor ECM que el otro, podemos ver en el gráfico que a partir de $q = 0{,}14$ aproximadamente, conviene usar el estimador $\hat{p}_A$ con el procedimiento explicado antes que preguntar por sí o por no y tomar $\hat{p}_{EMV}$ como estimador.

## Ejercicio 6: Nuevo estimador propuesto

¿Qué pasaría si usamos un nuevo estimador, $\hat{p}_C$, que venga del $\hat{p}_A$ pero que en vez de truncar, establezca una relación lineal entre los rangos de $\hat{p}_A$ y $p$? Es decir, $\hat{p}_C = \frac{1}{3} * \hat{p}_A + \frac{1}{3}$.

Creamos una función para que aproxime el ECM de este nuevo estimador con una simulación Monte Carlo y obtuvimos 0,04, que es casi 20 veces más grande que el ECM de $\hat{p}_B$.