Revista Digital: Matemática, Educación e Internet
ISSN: 1659-0643
revistadigitalmatematica@itcr.ac.cr
Instituto Tecnológico de Costa Rica
Costa Rica

Benavides Murillo, Francisco; Jiménez Oviedo, Byron
¿Por qué en ocasiones los resultados de las encuestas de opinión se alejan de la realidad? Un análisis con escenarios en Python
Revista Digital: Matemática, Educación e Internet, vol. 21, núm. 2, 2021, Marzo-, pp. 1-13
Instituto Tecnológico de Costa Rica
Costa Rica

DOI: https://doi.org/10.18845/rdmei.v21i2.5604

Disponible en: https://www.redalyc.org/articulo.oa?id=607964424003

Revista digital
Matemática, Educación e Internet
https://tecdigital.tec.ac.cr/revistamatematica/
Vol 21, No 2. Marzo – Agosto, 2021 ISSN 1659 -0643

¿Por qué en ocasiones los resultados de las encuestas de opinión se alejan de la realidad? Un análisis con escenarios en Python

| Why are opinion polls sometimes far from reality? An analysis with scenarios in Python |

Francisco Benavides Murillo
francisco.benavides.murillo@una.ac.cr
Universidad Nacional
Costa Rica

Byron Jiménez Oviedo
byron.jimenez.oviedo@una.ac.cr
Universidad Nacional
Costa Rica

Recibido: 4 Abril 2020 Aceptado: 20 noviembre 2020

Resumen: Antes de la segunda ronda de las elecciones presidenciales costarricenses del año 2018 en Costa Rica, se realizaron varias encuestas de opinión con el fin de evaluar las preferencias del electorado. Las encuestadoras en Costa Rica indicaban, en algunos casos, un empate técnico entre el candidato de Restauración Nacional y el del Partido Acción Ciudadana, y en otros casos una clara preferencia por el primero. El resultado final de las elecciones fue muy diferente al de estos estudios, y el triunfo del candidato del partido Acción Ciudadana fue claro y totalmente alejado de los márgenes de error asociados con la mayoría de los resultados obtenidos por las empresas encuestadoras. A raíz de esto, una pregunta natural surge: ¿por qué fallan las encuestas electorales? Uno de los argumentos utilizados es que las encuestas fueron realizadas con al menos una semana de antelación a las votaciones, y la opinión del electorado cambió en ese tiempo. En el presente trabajo se plantean situaciones hipotéticas de posibles escenarios en los que el sesgo de las encuestas se puede presentar, así como una metodología para su detección. Se analiza, en particular, la tasa de respuesta de los encuestados y su influencia en las estimaciones de la encuesta. El hecho es que si los que apoyan a un determinado partido tienen una tendencia generalizada de no responder una encuesta, entonces los resultados de ésta serán sesgados, sin importar el momento en que se hagan las entrevistas. En este trabajo, los escenarios sintéticos son simulados computacionalmente utilizando Python por su interés didáctico. También se simulan métodos de ponderación para compensar estas deficiencias. Este artículo pretende abrir un espacio didáctico de discusión de los aspectos técnicos, matemáticos y computacionales de las encuestas de opinión nacionales, así como producir un escenario sintético computacional que pueda derribar algunos mitos populares frecuentes, como el mito de que sólo mil personas encuestadas no son suficientes para obtener una representación significativa de la opinión del electorado.

Palabras Clave: Voto, escenarios, elecciones, encuestas, estadística, correlación

Abstract: Before the runoff election that took place in Costa Rica in 2018, several opinion public polls were executed in order to survey the electorate tendencies. The poll companies that made these surveys showed a technical draw between both candidates or, in other cases, a clear advantage for the "Restauración Nacional" candidate over the "Acción Ciudadana" candidate. Opposed to this, the final

election result gave the "Acción Ciudadana" candidate a clear win, with an advantage that was considerably larger than the error percentage that all polls estimated. Considering this, a natural question can be asked: why do polls fail? One of the answers provided by the poll's companies was that they were not allowed to make surveys unit one week before the election, and the electorate opinion changes with time. In the present work we give a series of hypothetical scenarios in which polls may fails, and a methodology to detect these biased results. Particularly, we analyze the electorate's polls answer rate and its influence in the polls estimates. In fact, if the supporters of a given candidate don't want to answer a poll survey, the poll's results will be biased, independently of the time in which it was made. In this work, we create these scenarios using Python and we also propose alternate weighing methods to compensate these biases. We pretend to open a discussion in poll's results that take into account technical, mathematical and computer simulation aspects that are not generally taken into account in Costa Rican's public opinion surveys and avoid myths such as the sample size is unable to predict the entire population preferences.

Keywords: Vote, elections, polls, statistics

1. Introducción

Aunque los fundamentos matemáticos de las encuestas de opinión pública están bien establecidos, en la práctica los resultados de una encuesta aún pueden contener sesgos que dan pie a predicciones muy alejadas de la realidad. En Costa Rica ocurrió precisamente eso en la segunda ronda de las elecciones presidenciales del año 2018, en el que prácticamente todas las encuestas mostraban, ya sea un virtual empate entre los candidatos, o una ventaja importante por parte del candidato, hoy perdedor, del Partido Restauración Nacional. Podemos identificar al menos tres causas para estos errores:

1. La selección de la población no es uniforme. La hipótesis de que cada miembro de la población total (denotada por Ω) tiene la misma probabilidad de ser encuestado es fundamental. En la práctica, garantizar la uniformidad de la distribución que selecciona a los miembros de la subpoblación entrevistada es más difícil de lo que parece. Diversos criterios son utilizados por las encuestadoras para garantizar la uniformidad de esta distribución ([5]). Hemos identificado al menos dos criterios que se practican en Costa Rica:

   a) Hacer entrevistas personalizadas de acuerdo con los estratos en los que se divide la población (partiendo de la información de censos nacionales). Este método requiere de un conocimiento previo y un cuidadoso diseño de la muestra.

   b) Elección aleatoria de números de teléfono (lo que cubre la mayor parte de la población). Este método requiere de una validación de la distribución obtenida, a partir del conocimiento de la estratificación de la población total.

2. La opinión de los electores puede cambiar en el tiempo. Es por eso que se considera que una encuesta de opinión será más exacta cuanto más cerca se encuentre de la fecha de las elecciones.

3. También es importante tomar en cuenta la intención de respuesta del entrevistado. El encuestado puede mentir, o dar respuestas rápidas que lo libren del encuestador rápidamente. En general, se asume que la no respuesta es una variable que se distribuye uniformemente en la población o, en otras palabras, que la probabilidad de que alguien no desee responder la encuesta es idéntica en los simpatizantes de todos los diferentes partidos políticos. No existen razones para asumir eso. Por el contrario; el comportamiento de los que no responden puede ser radicalmente diferente, tanto como sus preferencias políticas ([2]), las cuales pueden verse afectadas inclusive por la empresa que realiza la encuesta. En el caso de Costa Rica, en las elecciones del 2018, se vieron importantes diferencias entre las encuestas realizadas por la empresa OPOL ([1]) y las

de la Universidad de Costa Rica ([8]), ambas diferenciadas del resultado final de las elecciones nacionales por un margen de error superior al 3 % (el error podría ser de hasta 362 000 votos vea ([10])) . La encuesta de la UCR daba como resultado un virtual empate entre los candidatos, mientras que la encuesta OPOL le daba la victoria al candidato que en la elección final resultó perdedor. Si la muestra de ambas encuestas fue tomada de manera uniforme (como se hace sintéticamente en los ejemplos que siguen), y, dado que los sondeos fueron realizados en fechas similares, entonces la única forma de explicar este fenómeno es un comportamiento diferente de la población indecisa en función del encuestador.

En este trabajo suponemos que la selección de la subpoblación se hace con una distribución uniforme y que la encuesta mide, en un lapso de tiempo específico, la intención del electorado. La evolución de la decisión del electorado a lo largo del tiempo depende de muchos factores tales como: escándalos recientes, variaciones en las estrategias de campaña o incluso eventos que no tengan ninguna relación, a primera vista, con las elecciones presidenciales. Por esta razón, tal evolución se considera muy difícil de medir. Los resultados de una encuesta deben visualizarse como una medición instantánea de la intención de voto en un momento específico, y como un posible escenario futuro en caso de que las decisiones de los electores no cambien dramáticamente con el tiempo.

En este trabajo se crean escenarios computacionales con condiciones similares a las de elecciones presidenciales costarricenses y se analiza el papel que en el resultado electoral final pueden jugar los indecisos, quienes no desean o no pueden responder a los encuestadores. De hecho, se verifica, mediante estos escenarios sintéticos, que la tasa de respuesta puede introducir sesgos importantes en las encuestas de opinión, similares a las predicciones erróneas de las pasadas elecciones costarricenses. Con ayuda de Python, se plantean escenarios computacionales que reproducen una situación de error en una encuesta, así como un mecanismo para detectarla. De este modo, la distribución subsecuente de la metodología es la siguiente:

1. Escenario de verificación. Se crea una población grande de individuos con el voto repartido entre los candidatos A y B. Se toman varias muestras de unos 1000 individuos con el objeto de verificar el marco teórico descrito en esta sección.
2. Escenario de respuesta. Se introduce una nueva variable aleatoria de respuesta en la población anterior. Esta variable establece los individuos que responden y los que no responden a la encuesta. Se supone que los partidarios de uno de los candidatos tienen una marcada tendencia a no responder la encuesta y se verifica que esto puede conducir a predicciones erróneas.
3. Escenario de control. Se introduce una variable de control correlacionada con uno de los candidatos y cuya distribución se supone que es conocida en la población entera. Con esta variable de control se observa un sesgo en la subpoblación de muestra y, con base en ella, se corrigen las decisiones de cada individuo mediante ponderación. De esta manera, se obtiene una encuesta corregida en que el voto de cada encuestado puede tener un valor distinto de 1, pero que en principio está más apegada a la realidad

El objetivo de estos escenarios computacionales es ofrecer una herramienta didáctica que permita entender los aspectos técnicos de las encuestas de opinión. Todo el código fuente, programado por los autores para el presente artículo, se encuentra disponible en `https://github.com/fbenavDEV/EcnuestasPYTHON`. Las rutinas mostradas no son sólo muy fáciles de programar, entender y utilizar, sino que, a juicio de los autores, quien esté interesado en cambiar algunos parámetros (tales como el tamaño de la muestra o el número de candidatos) puede hacerlo y adquirir una sencilla experiencia empírica acerca de los postulados que sustentan una encuesta de opinión. Los autores consideran que este tipo de experiencias ayudarían al público en general a no asumir mitos tales como la idea de que es imposible medir la opinión de todo un país a partir únicamente de una muestra de mil personas o de que las encuestas son, en general, instrumentos de campaña política fundamentados en ciencia espuria sujeta a intereses ilegítimos.

2. Marco Teórico

En general, los fundamentos matemáticos de una encuesta de opinión se encuentran en la adecuada utilización del Teorema del Límite Central y algunas ideas de estadística básica. Supóngase que en una población Ω, un total de N_A personas votan por el candidato A y un total de N_{NA} personas no lo hacen. El tamaño de la población total N está dado por $N = N_A + N_{NA}$. Por lo general, el valor de N es muy grande y en una encuesta es imposible entrevistar a todos los miembros de la población. Por esa razón, se utiliza un subconjunto de tamaño n (donde n es mucho menor que N) de miembros de la población. A esta subpoblación se le denomina "muestra". Consideramos que $M = \{X_1, \ldots, X_n\}$ es una colección de variables aleatorias que representan las decisiones de cada uno de los miembros de la muestra. Los valores posibles de cada X_i son 1 o 0, de modo que $X_i = 1$ si X_i apoya a A y $X_i = 0$ si no. Si $p = \dfrac{N_A}{N}$ entonces $P(X_i = 1) = p$ y $P(X_i = 0) = 1 - p = q$. Es decir, la probabilidad de que un individuo en Ω se incline por el candidato A es p y la probabilidad de que no lo haga es q. La distribución de probabilidad de la decisión de cada entrevistado se corresponde con una distribución de Bernoulli ([7]), cuyo valor medio es p y cuya desviación estándar es $\sigma = \sqrt{pq}$. De acuerdo con el teorema del límite central ([7]), la variable aleatoria Z sigue una distribución normal estándar:

$$Z = \frac{\hat{X} - p}{\sigma/\sqrt{n}}\, N(0,1)\,,$$

donde $\hat{X} = \frac{1}{n}(X_1 + \ldots + X_n)$. Es importante notar que tanto el valor de p como el de σ son desconocidos en la encuesta. Sin embargo, por propiedades del área bajo la curva normal, es posible establecer que, para $\alpha = 1.96 \approx 2$:

$$P\left(-\alpha \leq \frac{\hat{X} - p}{\sigma/\sqrt{n}} \leq \alpha\right) = 0.95.$$

Es decir, el valor de p está aproximado dentro del intervalo $\left[\hat{X} - \frac{2\sigma}{\sqrt{n}}, \hat{X} + \frac{2\sigma}{\sqrt{n}}\right]$ en el 95 % de las ocasiones en que se tome una muestra de tamaño n. Este porcentaje no es otra cosa que el nivel de confianza de la encuesta. Significa que la probabilidad de estimar el valor de p con un error mayor a $\epsilon = 2\sigma/\sqrt{n}$ es una en veinte, es decir, el 5 %. Esta confianza del 95 % es universalmente aceptada en las encuestas de opinión y es la que utilizaron todas las empresas encuestadoras de las elecciones costarricenses ([9]).

Para determinar el tamaño de la muestra n asociado al error $\epsilon = \alpha\sigma/\sqrt{n}$ de la encuesta se considera el valor máximo que puede tomar σ^2. La maximización de la función parabólica $f(p) = p(1-p)$ corresponde a $p = \frac{1}{2}$ de modo que el valor máximo de la varianza es $\sigma^2 = p(1-p) \leq 1/4$. De esta manera, el error ϵ con el que $\hat{X}$ aproxima el valor verdadero de p está limitado por:

$$\epsilon \leq \frac{2}{2\sqrt{n}} = \frac{1}{\sqrt{n}}\,.$$

Por lo general, el número de encuestados es de mil personas, lo que corresponde a un error de $\epsilon = \frac{1}{\sqrt{1000}} = 0.032$ o, aproximadamente, un 3 %. Estos valores coinciden con el intervalo de confianza y el error que habitualmente es aceptado en las encuestas de opinión por parte de las encuestadores (aunque, vale la pena considerar que, de manera sistemática, la empresa OPOL empleaba márgenes de error más estrechos, producto de una muestra mayor). Se puede observar que la cota superior para el error de la desviación estándar σ es la misma para todos los candidatos.

3. Resultados

El propósito del presente trabajo es reproducir, de manera sintética, un escenario similar al que se dio en las elecciones presidenciales costarricenses del 2018. La idea es plantear una situación en la que uno de los candidatos gana con el 60 % de los votos, pero que, en las mediciones realizadas con encuestas, se muestra una situación de empate o de victoria del otro candidato. Al ser esta una situación totalmente controlada, se propone una metodología, basada en ponderaciones para detectar este sesgo y se observa su efectividad. La velocidad con la que se puede sintetizar una encuesta computacionalmente permite observar el comportamiento de un cierto tamaño de muestra a lo largo de varias encuestas, para observar su distribución empírica en un histograma. Los autores consideran que esto resulta interesante desde un punto de vista didáctico, pues permite reproducir, con un costo computacional reducido y una programación muy sencilla, los escenarios medidos en una encuesta de opinión.

3.1. Escenario de verificación

En este escenario únicamente se verifica, con un ejemplo práctico, la efectividad de los métodos expuestos en el preámbulo anterior. La idea es crear una matriz P cuyo número de filas es igual a la cantidad de individuos de la población. El número de columnas de P es igual al número de candidatos y las únicas entradas de la matriz son 1 o 0. Cada fila tiene un único valor diferente de 0, indicando el apoyo del individuo de la fila por el candidato asociado a la columna. La creación de la población se realiza con el comando "pob" definido en el código. Por ejemplo, si se quiere crear una población de 3 millones de habitantes, cuya distribución entre dos candidatos sea 6 a 4, se utiliza el siguiente comando:

```
P = pob(3000000, numpy.array([6,4])) ,
```

el cual retorna una matriz P con esas características. Para obtener el porcentaje de apoyo al primer candidato, por ejemplo, se utiliza el siguiente comando:

```
sum(P[:,0])/len(P) ,
```

que deberá retornar un valor similar a 0.6 en el caso presente. Nótese que una diferencia de 6 a 4 es aproximadamente la diferencia entre los resultados de las elecciones presidenciales costarricenses en el año 2018 ([3]).

Ahora bien, para simular el proceso de entrevistas de una encuesta, lo único que se debe hacer es seleccionar una subpoblación aleatoria y distribuida uniformemente y obtener su preferencia. Esto se hace mediante la rutina "subpop" que selecciona, arbitrariamente un cierto número de individuos y los retorna en una nueva matriz. Así, por ejemplo, el comando:

```
subpob(P,1000)
```

retorna una subpoblación de 1000 individuos seleccionados de la población P. Los resultados de una "encuesta" ejecutada en esa subpoblación se obtienen, también, mediante conteo simple, usando el comando "sum" como se hizo anteriormente.

Con el objeto de visualizar mejor el concepto de error asociado con una encuesta, puede ser interesante realizar este proceso varias veces y seleccionar, a lo largo de varias iteraciones, poblaciones diferentes de 1000 individuos. Esto no puede hacerse en la práctica, por los altos costos que conlleva, pero en un escenario sintético como el presente, el proceso se lleva a cabo en sólo unos segundos. De hecho, se ha creado la rutina "subpob_s" para ejecutar esta función. La rutina devuelve una matriz V que contiene los porcentajes (del 0 al 100) de apoyo a los diferentes candidatos en las diferentes iteraciones. Tales resultados pueden ser tabulados en un histograma, como se muestra en la Figura 1. En la ilustración

resulta interesante observar la distribución, aproximadamente normal, de los diferentes resultados que oscilan dentro del margen de error.

Cabe destacar que, hasta donde sabemos, el análisis de las encuestas que se realizan en Costa Rica llega a este punto, al menos en la divulgación de sus resultados. Se asume la uniformidad de la muestra (garantizada probablemente por algún método de estratificación ([6])) y a partir de ahí se hace un conteo de las decisiones de cada uno de los entrevistados. No se realiza ningún análisis a posteriori a partir de la distribución de la "no respuesta", aunque ésta sea considerable. Inclusive, en general, las encuestadoras agrupan la no-respuesta bajo la categoría común "No sabe/ No responde".

A partir de ahora, tal y como se muestra en figura 1, todos los ejemplos giran en torno a este escenario: el candidato "naranja" gana con un 60 % de los votos sobre el candidato azul.

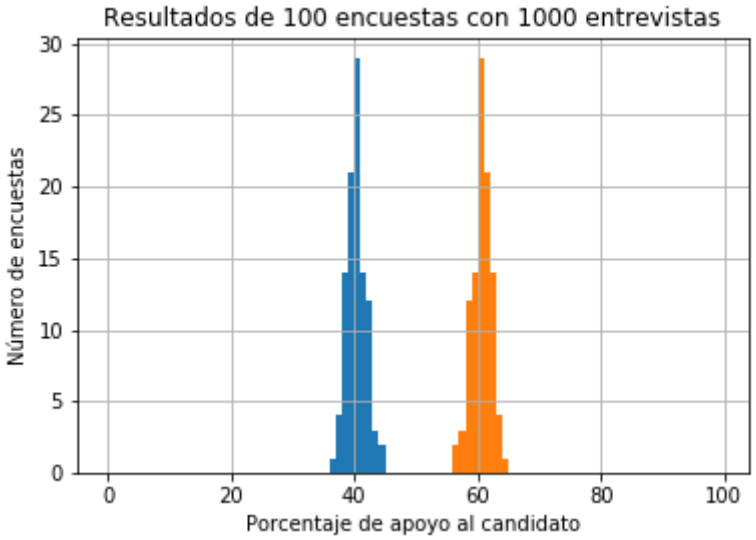

Figura 1: Histograma de resultados de "encuestas" virtuales, tabulando los resultados de 100 encuestas con 1000 individuos cada una. Los histogramas asociados a la preferencia de ambos candidatos tienen un valor medio igual a las preferencias reales de la población, con un intervalo de error de alrededor del 3 %

3.2. Escenario de respuesta

En la práctica, es posible imaginar que el papel de los indecisos (o los que no responden) en una encuesta de opinión es esconder su intención de voto para el encuestador. Su rol es fundamental en el resultado final de la elección pues, aunque es posible suponer que los indecisos se podrán distribuir de manera más o menos uniforme entre los candidatos, también puede ocurrir que sus inclinaciones se encuentren tan desbalanceadas que cambien completamente el resultado de la elección. Una metodología encuestadora que únicamente pregunte por la intención de voto de los entrevistados carece de mecanismos para detectar un posible sesgo general en el grupo de los indecisos.

Esto se ilustra, con el código desarrollado en Python para el presente artículo, de la siguiente manera: se supone que en la población de individuos existe un vector de respuesta con los valores de 1 o 0. Cuando el valor del vector es 1, entonces el individuo responde la encuesta, pero cuando es 0, su voto futuro es invisible para el encuestador. La decisión de la población indecisa puede ser escondida mediante la rutina "esc" cuyo uso se plantea en el siguiente comando:

```
PI = esc(P,numpy.array([65,90])).
```

La población PI consiste en el conjunto de individuos de P que tienen un voto igual a 0 para todos los

candidatos (su decisión está oculta). El primer parámetro es la población entera P y el segundo es la forma en que se distribuye la indecisión entre los candidatos. Cabe recordar en este punto que la idea de estas rutinas es crear escenarios sesgados pero controlados, en los que las encuestas puedan fallar. En la práctica es imposible establecer cómo se distribuye la indecisión entre los candidatos.

En el comando descrito en la llamada de rutina anterior, los individuos que apoyan al candidato naranja afirman estar decididos en el 65 % de los casos, mientras que, para el candidato azul, el porcentaje de decididos es de un 90 %. Esta distribución le quita peso al apoyo del candidato naranja, que queda oculto en el rubro de los indecisos (marcado en verde en la figura 2) produciendo un empate técnico entre los candidatos y una indecisión de alrededor del 23 %. Como el escenario es controlado mediante código, el número de indecisos es la cantidad de individuos encuestados que tienen un 0 en la columna de decisión. En otras palabras, la columna de indecisión PI puede interpretarse como un tercer candidato, que retira votos de los otros dos, con su correspondiente valor medio y margen de error. En la Figura 2 es claro que las distribuciones entre ambos candidatos se solapan, produciendo un empate técnico. Se han tomado 100 grupos de 1000 individuos cada uno para producir estos histogramas. En verde se despliega la distribución de los indecisos

Figura 2: Distribución del apoyo de los candidatos de la Ilustración 1, pero tomando en cuenta que apenas el 65 % de los individuos que apoyan al candidato naranja afirman tener decisión en cuanto a su voto.

3.3. Escenario de control

Uno de los problemas más serios concernientes a la presencia de un alto porcentaje de indecisos en la encuesta es que, aunque la muestra de 1000 individuos pudo haber sido obtenida de manera uniforme a lo largo de toda la población, la subpoblación de alrededor de 800 individuos que efectivamente responde la encuesta puede no ser representativa de todo el espacio muestral. Como ya se dijo, existe la posibilidad de que los indecisos estén decididos en realidad y simplemente escondan su intención de voto para el encuestador hasta el último momento ([4]).

El hecho es que, en la práctica, aunque se hayan tenido los mayores cuidados en que la muestra de 1000 individuos esté equilibrada y represente adecuadamente la totalidad de la población, el número efectivo de individuos que responden la encuesta puede estar sesgado y por tanto los resultados obtenidos con base en ellos deberían ser compensados. Esto se puede lograr, por ejemplo, con una variable de control cuya distribución en la totalidad de población sea conocida de antemano mediante un censo. Si la distribución de esa variable no es igual en la población de individuos que respondieron la encuesta, entonces es posible concluir que existe un sesgo en el resultado del sondeo. La forma

de compensar ese sesgo consiste en aumentar, de manera proporcional, el peso de las decisiones de los individuos que están subrepresentados, mientras se disminuye el peso de las decisiones de los individuos que están sobrerrepresentados. Es claro que encontrar estas variables de control requiere de un mayor número de preguntas en la encuesta que puedan ubicar al individuo según su estrato económico, nivel de educación, edad, o cualquier otro dato que permita comparar la distribución de estas variables con la de la población total.

De manera más precisa, supongamos que existe una variable de control V_C con m categorías cuya proporción en la población está dada por la distribución discreta $\{p_1, .., p_m\}$ con la propiedad de que:

$$\sum_{j=1}^{m} p_j = 1.$$

Las diferentes categorías de esta variable dependen del tipo de variable de control utilizada: rango de edad, sexo, estado civil, etc. Se supone también que esta distribución es conocida en la población total, por medio de censos (por ejemplo, proporción de hombres y mujeres, porcentaje de población correspondiente a una determinada faja etaria, etc). Supóngase también que la representación de esa variable en el subconjunto de individuos de la muestra que respondieron la encuesta está dada por los pesos: $\{w_1, \ldots, w_m\}$. El ajuste de las decisiones de cada individuo consiste en modificar su valor unitario de acuerdo con la categoría a la que pertenece, otorgándole el valor de la representatividad que tiene dentro de la muestra i.e. p_i/w_i para $i = 1, \ldots, m$. De este modo, el individuo sobrerrepresentado dentro de la muestra tendrá un peso menor en el conteo de los votos, mientras que el subrepresentado tendrá un peso mayor.

Como se desprende de lo anterior, el proceso es simple siempre que sea posible encontrar la variable V_C. Para simular este escenario en Python, se procede a fabricar, de manera general, una variable que tenga una cierta correlación c con la variable X que determina el apoyo a un cierto candidato en la población. X se representa mediante un vector columna con valores de 1 o 0 (aunque en principio, este enfoque tan simplista podría alterarse) mientras que V_C es un vector con valores discretos $\{1, \ldots, m\}$ de acuerdo con el número de categorías m.

El proceso de creación de V_C es como sigue: primero se construye un vector columna V con la misma cantidad de individuos de la población, con elementos generados a partir de una distribución uniforme con media 0 y varianza 1. Seguidamente se calcula:

$$V_{CC} = cX + V\sqrt{1 - c^2}.$$

Con esto, la correlación entre V_{CC} y X es c. Mediante adecuadas operaciones de translación (restando a cada elemento de V_{CC} el valor medio de la variable), normalización (división entre el la desviación estandar para cada elemento deV_{CC}) y discretización (redondeo de los resultados obtenidos en un número fijo de categorías), es posible obtener V_C con la distribución deseada $\{p_1, \ldots, p_m\}$ y una correlación c_d con respecto a X que se encuentre cerca del valor deseado c. El resultado no es exacto por la pérdida de precisión durante la discretización.

Cabe destacar que la variable así producida es sintética y sus condiciones pueden ser manipuladas libremente en el escenario computacional. En la práctica, durante una encuesta de opinión, la correlación de V_C con la intención de voto para un candidato puede no ser tan simple de calcular, pues la única información de la que se dispone es la respuesta de los encuestados. En otras palabras, la correlación real entre V_C y la intención de voto no puede ser obtenida en toda la población y apenas puede ser estimada por medio de los 1000 individuos encuestados. La presencia de indecisos dificulta aún más esta estimación. Por ejemplo, la correlación real de 0.7 mostrada en la figura 3 es estimada con una correlación de alrededor de 0.4 entre los individuos que responden la encuesta (con el porcentaje de indecisos mostrado en la figura 2) tal y como se muestra en la figura 4.

En la figura 3 se han realizado 100 encuestas con 1000 individuos cada una, ponderando los pesos de las decisiones individuales de acuerdo con la variable V_C Las distribuciones se han separado completamente, mostrando una clara ventaja del candidato naranja con respecto al resultado de la figura 6.

Además, aunque la correlación varía de un ensayo a otro, su valor siempre oscila alrededor de 0.4 y siempre es menor al valor de correlación real, dado por 0.7 (ver la fig 4).

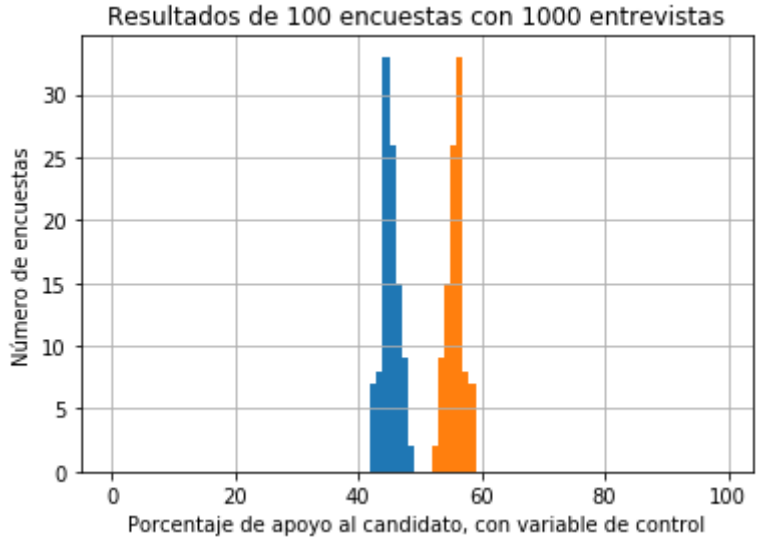

Figura 3: Efecto de la variable de control sobre el ajuste de la respuesta de los encuestados para una correlación entre la intención de voto de uno de los candidatos y V_C igual a 0.7.

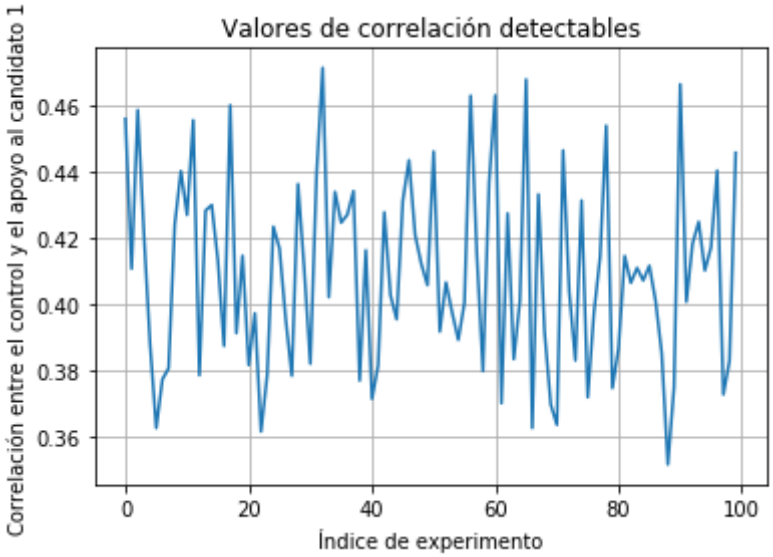

Figura 4: Correlación entre la intención de voto hacia uno de los candidatos y la variable de control V_C, medida en cada uno de los 100 grupos de 1000 individuos en que se realiza la encuesta.

Sin embargo, como se muestra en las figura 3, 5 y 6 , inclusive ponderaciones basadas en correlaciones bajas pueden mejorar la estimación de los resultados de la encuesta, mostrando una tendencia hacia el alta por parte del candidato “naranja”.

Las rutinas que producen estos resultados se basan en la utilización de la biblioteca NumPY de Python y el código desarrollado para el presente artículo. Se invocan de la siguiente manera:

```
R = correlac(V,corr_val,pesos).
```

Este comando produce una variable R con categorías numeradas desde 1 hasta m donde m es la longitud del vector de pesos que contiene la distribución discreta de cada categoría. V es el vector con el que R se correlaciona con el valor dado en corr.

Figura 5: . Efecto de la variable de control sobre el ajuste de la respuesta de los encuestados para una correlación entre la intención de voto de uno de los candidatos y V_C igual a 0.2.

En la figura 5 se han realizado 100 encuestas con 1000 individuos cada una, ponderando los pesos de las decisiones individuales de acuerdo con la variable V_C. Observe que las distribuciones se han separado ligeramente con respecto al resultado de la figura 2. Por otro lado, en la figura 5 se han ponderado los pesos de las decisiones individuales de acuerdo con la variable V_C Las distribuciones se han separado ligeramente con respecto al resultado de la figura 5.

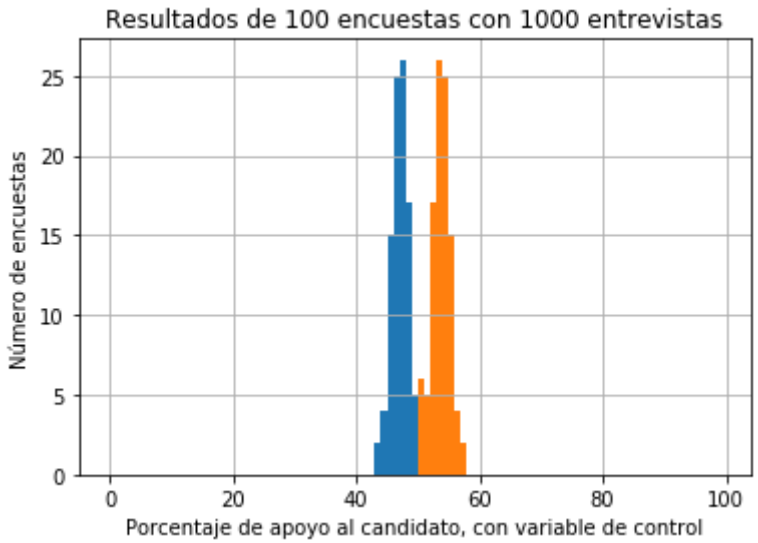

Figura 6: Efecto de la variable de control sobre el ajuste de la respuesta de los encuestados para una correlación entre la intención de voto de uno de los candidatos y V_C igual a 0.5.

La ponderación de los resultados de la encuesta parece ser efectiva incluso en los casos en los que el porcentaje de indecisos en el candidato naranja es tan alto, que en la encuesta da la impresión de que

el candidato azul tiene la ventaja. Por ejemplo en la figura 7 el candidato azul tiene una ventaja en una proporción de 6 a 4 con respecto al número de individuos decididos (lo cual es similar al escenario de OPOL) y un 35 % de individuos que no responden.

Figura 7: Resultado de la "encuesta" al suponer que el candidato naranja gana con un 60 %, pero apenas el 45 % de esos votantes responde la encuesta.

El resultado de la ponderación se muestra en la figura 8. No sólo la ventaja del azul es menos evidente, sino que el candidato naranja parece tener una ligera ventaja, aunque ésta se encuentre dentro del margen de error de la encuesta. Es claro que uno de los problemas de esta técnica de ponderación, es que puede reinterpretar el resultado obtenido a uno totalmente opuesto, basándose en una correlación que no puede ser estimada correctamente. Por eso es importante valorar el escenario de la no-respuesta (individuos que afirman estar indecisos o inclusive que no van a votar) y estudiar cuidadosamente la estratificación de la población que sí ha respondido la encuesta. Este estudio podría hacerse a posteriori, con base en resultados anteriores y en las preguntas que se le han hecho a los encuestados.

Figura 8: Resultado de la "encuesta" de la figura 7, ponderando los votos de acuerdo con una variable de control que se correlacional con la intención de voto en un valor de 0.7.

4. Conclusión

En este artículo se ha descrito la forma de producir un escenario artificial, obtenido con un código en Python sencillo, basado en NumPy y con el que se pueden ejecutar varios experimentos. La idea del proceso es generar una situación en la que se presenten resultados fallidos similares a los que se predijeron en las encuestas electorales del 2018 en Costa Rica. Como se ha visto a lo largo del artículo, no es imposible imaginar un comportamiento sesgado de los indecisos y por tanto predicciones equivocadas por parte de las encuestas, inclusive sin tomar en cuenta el factor tiempo de la encuesta con la elección real. El problema aquí se ha planteado de forma inversa: se crea una población de individuos que, bajo ciertas circunstancias, producen un resultado de encuesta equivocado. Luego se fabrica una variable de control que permita compensar ese resultado. A pesar de esta artificialidad, no parece imposible suponer la obtención de una variable así durante el cuestionario de la encuesta. A cada individuo se le pueden hacer preguntas sobre su edad, sexo u orientación religiosa, entre otras variables, que puedan ser ponderadas con la población total. Aunque parece que estas preguntas sí se realizan en algunas de las encuestas que se llevan a cabo en Costa Rica, no parece existir ningún estudio sobre el efecto de la no-respuesta y la representatividad de cada estrato de la población entre los que sí responden el cuestionario.

El problema, en la práctica, consiste en producir esta clase de escenarios a partir de la información que proporciona la encuesta realizada en mil individuos. La obtención de una variable de control es un tema complejo, pues su correlación con el apoyo a alguno de los candidatos no parece ser detectable con precisión en una muestra limitada y esto puede inducir errores de valoración. La variable de control sólo puede ser obtenida a través de preguntas adicionales en el cuestionario y con base en censos conocidos. La elección de estas preguntas se corresponde con aspectos sociales y económicos de población sondeada, lo que en la práctica puede producir también controversias. Su aplicación en una encuesta real podría estar siempre sujeta al escrutinio y cuestionamiento del público pues, como ya se vio, el resultado de la encuesta sin ponderación puede ser radicalmente diferente del resultado ponderado.

El presente desarrollo proporciona una herramienta fácil de utilizar y programar para crear escenarios que proporcionen una experiencia empírica sobre el funcionamiento estadístico de las encuestas de opinión. Pese a esta simplicidad, estos escenarios proporcionan, también, una justificación alternativa, pero plausible, de las diferencias de la encuesta de opinión y los resultados reales de las elecciones. Cabe, por tanto, imaginar un escenario en que la encuesta se realiza en un momento cercano a las elecciones, pero, debido al comportamiento asimétrico de los encuestados con relación a la exposición de su opinión a la empresa encuestadora, produce resultados totalmente alejados de la realidad.

Referencias

[1] Angulo, Y. (2018). Encuesta: Fabricio Alvarado57, 35 % ? Carlos Alvarado 42, 65 %. ELMUNDO.CR. Retrieved from https://www.elmundo.cr/costa-rica/encuesta-fabricio-alvarado-5735-carlos-alvarado-4265/

[2] Crane, H., & Martin, R. (2017). Rethinking Probabilistic Prediction in the Wake of the 2016 U.S. Presidential Election. Ssrn, 1-19. https://doi.org/10.2139/ssrn.2953278

[3] Garcia, David Alire; Pretel, E. A. (2018). Costa Rica center-left easily wins presidency in vote fought on gay rights. Reuters. Retrieved from https://www.reuters.com/article/us-costarica-election/costa-rica-center-left-easily-wins-presidency-in-vote-fought-on-gay-rights-idUSKBN1H80

[4] Gelman, A., Goel, S., Rivers, D., & Rothschild, D. (2016). The mythical swing voter. Quarterly Journal of Political Science, 11(1), 103-130. https://doi.org/10.1561/100.00015031

[5] Kenett, R. S., Pfeffermann, D., & Steinberg, D. M. (2018). Election Polls - A Survey, A Critique, and Proposals. Annual Review of Statistics and Its Application, 5, 1-24. https://doi.org/10.1146/annurev-statistics-031017-100204

[6] Little, R. J. A. (1993). Post-Stratification: A Modeler's Perspective. Journal of the American Statistical Association, 88(423), 1001-1012. https://doi.org/10.1080/01621459.1993.10476368

[7] Montgomery, D. C. (2003). Applied Statistics and Probability for Engineers Third Edition. In Phoenix Usa (Vol. 37). https://doi.org/10.2307/1269738

[8] Murillo, A. (2018). Candidatos arrancan parejo hacia segunda ronda. Semanario Universidad. Retrieved from https://semanariouniversidad.com/pais/candidatos-arrancan-parejo-hacia-segunda-ronda/

[9] Redondo, R. A., Rodríguez, F. A., Cascante, M. J., & Castillo, J. G. (2018). ENCUESTA DE OPINIÓN SOCIOPOLÍTICA REALIZADA EN FEBRERO DE 2018. Retrieved from PROYECTO "ESTUDIOS DE OPINIÓN PÚBLICA" , UNIVERSIDAD DE COSTA RICA website: https://ciep.ucr.ac.cr/index.php/proyectos/encuestas-de-opinion

[10] Elecciones presidenciales de Costa Rica de 2018. En Wikipedia. Recuperado el 24 de septiembre de 2020 de https://es.wikipedia.org/wiki/Elecciones_presidenciales_de_Costa_Rica_de_2018