Reglas del TP:

- Este trabajo debe hacerse de forma individual o de a dos.
- Deben enviarse el script de R y un informe que contenga las respuestas a todas las preguntas y los gráficos pedidos. No hace falta explicar en el informe que es lo que hace cada una de las funciones del script.
- El script debe estar prolijo. Esto en particular implica que las variables tienen que tener un nombre descriptivo (es decir, no llamar `a`, `b`, `c` a las variables).

Estimando probabilidades preservando el anonimato

Supongamos que nos interesa saber cuál es la proporción de la población entre 18 y 40 años en Capital Federal que probó drogas duras. Llamamos p a dicha proporción. Si toda la población dijera la verdad, entonces sería natural estimar a p usando el estimador de máxima verosimilitud $\widehat{p}_{\text{EMV}}$ (donde $\widehat{p}_{\text{EMV}} = s/n$ donde s es la cantidad de personas que contestaron **sí** y n es la cantidad total de encuestados). Sin embargo, es común que en este tipo de preguntas el encuestado mienta al responder.

Vamos a considerar una forma alternativa de estimar p de modo que el encuestador no pueda saber con certeza si una determinada persona probó o no drogas duras. Esto garantiza el anonimato de los encuestados.

El procedimiento de encuesta es el siguiente:

- Se le pide a la persona que tire un dado y no le muestre el resultado al encuestador.
- Si en el dado sale 1 o 2, el encuestado debe contestar la verdad respecto de si probó o no drogas duras.

- Si en el dado sale 3, 4, 5 o 6, el encuestado debe contestar al azar de la siguiente forma: tira una moneda equilibrada, si sale cara responde **sí** y si sale ceca responde **no**.
- El encuestador no ve los resultados del procedimiento del encuestado. Sólamente sabe si el encuestado responde **sí** o **no** (ni siquiera sabe si fue necesaria la tirada de la moneda).

1. Sea $X_i = \mathbb{I}\{\text{la persona } i \text{ respondió sí}\}$. Probar que $\widehat{p}_A = 3\overline{X} - 1$ es un estimador insesgado de p. ¿Qué valores máximo y mínimo puede tomar este estimador? ¿Coinciden con el rango de valores que puede tomar p?

2. Suponiendo que todos dicen la verdad siempre, hallar el error cuadrático medio de $\widehat{p}_{\text{EMV}}$ y $\widehat{p}_A$. ¿Cuál de ellos es mayor? Calcular esta diferencia cuando $n = 1000$ y $p = 0.2$.

3. Considerar el estimador $\widehat{p}_B = \widehat{p}_A\, \mathbb{I}\{\widehat{p}_A \in [0,1]\} + \mathbb{I}\{\widehat{p}_A > 1\}$. ¿Qué sentido intuitivo le encuentra a este estimador? Si $p = 0.2$ y $n = 1000$, aproximar, mediante una simulación de Monte Carlo, el error cuadrático medio de $\widehat{p}_B$ suponiendo que todos los encuestados dicen la verdad. ¿Resulta una mejora sustancial respecto al estimador $\widehat{p}_A$?

4. Supongamos que $p = 0.2$ y se encuestan $n = 1000$ personas. Soupongamos también que se decide usar el $\widehat{p}_{\text{EMV}}$ y una proporción q $(0 < q < 1)$ de los que probaron drogas duras mienten cuando se les hace la pregunta. Hacer una función que permita aproximar/calcular, mediante una simulación de Monte Carlo o analíticamente, el error cuadrático medio de $\widehat{p}_{\text{EMV}}$ en este caso en función de q. Llamemos e_q a esta cantidad.

5. En el contexto del item anterior, graficar a q contra e_q. ¿Cuán grande debe ser q para que convenga usar el estimador $\widehat{p}_A$? ¿Cuán grande debe ser q para que convenga usar el estimador $\widehat{p}_B$?
   Aclaración: se supone que si se usan los estimadores $\widehat{p}_A$ o $\widehat{p}_B$, la gente siempre sigue el procedimiento y no miente nunca (porque se garantiza su anonimato).

6. (Opcional) Extender este trabajo en la dirección que le parezca natural.

7. (Opcional) Inventar otra forma de estimar p que conserve el anonimato de los encuestados. Calcular analíticamente o aproximar con una simulación de Monte Carlo el error cuadrático medio del nuevo estimador.