Guía 1

1. Considerar el caso de un médico residente de un hospital que cuando empieza la temporada anual de gripe sabe que aproximadamente el 40% de los pacientes tendrán gripe estacional. El hospital cuenta con un kit de diagnóstico rápido con las siguientes propiedades: da un resultado positivo en el 95% de las personas que tienen gripe (sensibilidad) y da negativo en el 45% de las personas que no tienen gripe (especificidad).

   a). Un paciente que dio positivo con este kit de diagnóstico rápido. ¿Cuál es la probabilidad de que el paciente realmente tenga la gripe estacional?

   b). Después de ver a varios pacientes, el médico se da cuenta de que puede hacer un mejor trabajo estimando aproximadamente la probabilidad de que un paciente tenga gripe estacional antes de realizar el test Supongan que, después de escuchar la historia clínica de un paciente y examinar sus síntomas, encuentra que es extremadamente improbable que tenga la gripe: su estimación aproximada (la "probabilidad previa al test") es que la probabilidad de que el paciente tenga gripe estacional es de 0.1. ¿Cuál es la probabilidad de que este paciente tenga gripe estacional si da positivo (la "probabilidad posterior al test")?

   c). Hacer un gráfico de la probabilidad posterior (a hacer el test) de tener la gripe estacional en función de la probabilidad previa al test, suponiendo que el resultado del test es positivo.

   d). Al ver el gráfico de la pregunta anterior, el médico supervisor le dijo al residente que no tiene que realizar tests rápidos en algunos casos. ¿Tiene sentido? Si es así, ¿En qué casos un test positivo es poco informativo (digamos, probabilidad menos al 50%) de que el paciente tiene gripe?

2. Se tiene un algoritmo para detectar imágenes de gatos. El algoritmo identifica correctamente el 80% de las imágenes de gatos como gatos, pero identifica incorrectamente el 50% de las imágenes que no son gatos como gatos. Prueban el algoritmo con un nuevo conjunto de imágenes, de las cuales el 8% son gatos. ¿Cuál es la probabilidad de que una imagen sea realmente un gato si el algoritmo la identifica como un gato? Responder a simulando datos para 10.000 imágenes. Para esto, primero crear un conjunto de datos simulados en el que cada observación corresponde una imagen y que tenga dos variables binarias: si la imagen tiene un gato y si fue identificado correctamente o no. Luego, quedarse sólo con las observaciones que fueron clasificadas como gatos por el algoritmo y contar la proporción de estas observaciones que son efectivamente imagenes de gatos. Comparar el resultado de la simulación con el resultado analítico usando el Teorema de Bayes.

3. Considerar dos variables binarias aleatorias e independientes, X_1 y X_2, tales que $P(X_i = 1) = 1/2$ y $P(X_i = -1) = 1/2$. Se define $Y_i = \theta + X_i$, donde $\theta \in \mathbb{R}$ es un número fijo desconocido. Suponiendo que sólo se observan Y_1 e Y_2 y que se quiere estimar θ, responder las siguientes preguntas.

   a. Comprobar que el conjunto C definido a continuación es un "intervalo de confianza" del 75%. Es decir, comprobar que $P(\theta \in C) = 3/4$.

   $$C = \begin{cases} Y_1 - 1 & \text{si} \quad Y_1 = Y_2 \\ (Y_1 + Y_2)/2 & \text{si} \quad Y_1 \neq Y_2 \end{cases}$$

   b. Supongan que se observó $Y_1 = 15$ e $Y_2 = 17$, Dar un intervalo de confianza C del 75% para θ?

c. Encuentren el valor de θ (es decir, con 100% de certeza)

d. ¿Hay alguna contradicción entre lo encontrado en ii y en iii? Discutir la diferencia entre afirmaciones probabilísticas sobre θ (bayesianas) y afirmaciones sobre un intervalo aleatorio (frecuentistas).

4. Un procedimiento aleatorio para estimar la proporción de la Tierra cubierta por agua es el siguiente: lanzar el globo por el aire y, cuando se lo vuelve a atajar, registrar si la superficie debajo del dedo índice de la mano derecha es “agua” (A) o “tierra” (T). Se lanzó el globo 11 veces y se obtuvo la siguientes observaciones:

$$\{A, A, T, A, T, T, T, T, T, T, T\}$$

a. Asumiendo un prior uniforme para la proporción de la Tierra cubierta por agua, θ, obtener y graficar la distribución posterior para θ a medida que se van obteniendo las observaciones.

b. Usando los datos de las 14 observaciones, dar una estimación puntual de la superficie cubierta por agua y un intervalo de credibilidad del 95%.

c. Hacer un histograma de la predicción del número de “A” que se obtendrían en 5 nuevas observaciones. Es decir, la *posterior predictive* para una nueva muestra de 5 puntos. Comparar con la distribución que se obtiene usando el estimador de máxima verosimilitud para θ.

d. Las observaciones muestran una racha de 7 T’s. Simular secuencias de 11 observaciones y hacer un histograma de la distribución del largo de la racha más larga de T’s. Comparar con lo observado y discutir si la capacidad del modelo para reproducir esta observación.

5. Con el mismos datos que el ejercicio 4, obtener y graficar la distribución posterior para θ partiendo de un prior que asume que la mayoría de la tierra está cubierta por agua:

$$p(\theta) = \begin{cases} 0 & \text{si} \quad \theta < 0.5 \\ 2 & \text{si} \quad \theta \geq 0.5 \end{cases}$$

Comparar gráficamente las distribuciones posteriores obtenidas con las que se obtuvieron con los dos priors. ¿A medida que se acumulan observaciones la influencia del prior disminuye?

6. Con el mismo procedimiento que el ejercicio 4, supongan que se obtienen A = 5 puntos de agua, pero se olvidaron de anotar el tamaño de la muestra (es decir, no se sabe T). Suponiendo que $\theta = 0.7$, calcular la distribución posterior del tamaño de muestra, N. Ayuda: Usar la distribución binomial.

7. Dos jugadores (A y B) juegan una partida de azar que consiste en tirar una moneda al aire sucesivas veces y sumar un punto para A si sale cara o un punto para B si sale ceca. El primero que llega a 6 puntos gana la partida.

a. Si la partida se interrumpe cuando va ganando A por 4-3.¿Cuál es la probabilidad de que B gane la partida?

b. Estimar la probabilidad que se pregunta en a) haciendo una simulación de 500 partidas interrumpidas con el puntaje 4-3. Graficar la proporción de partidas que gana B en función del número de partidas simuladas y comparar con el resultado analítico de la pregunta anterior.

8. Considerar una partida de destreza entre dos jugadores (una partida de pool, un set de tenis, ...) en la que el jugador A tiene 3 puntos y el jugador B tiene 5 puntos. El primero en llegar a 6 gana la partida. Se quiere estimar la probabilidad de que la partida la gane A.

    a. Partiendo de un prior uniforme para la probabilidad de que B gane un punto, θ, estimar la probabilidad de que A gane la partida.

    b. Usando el estimador de máxima verosimilitud para θ, encontrar la probabilidad de que A gane la partida. ¿A qué se debe la diferencia con lo encontrado en a)?

    c. Estimar la probabilidad de que A gane la partida haciendo simulaciones. Para esto, elegir un valor de θ al azar entre 0 y 1, dado que se desconoce cuál es su valor. Luego simular 8 puntos y, si el resultado es 5-3 favorable a B, continuar la partida para ver quién gana. Finalmente, estimar la probabilidad de que A gane la partida calculando la fracción de partidas simuladas en las que B ganaba por 5 a 3 y luego A ganó la partida. ¿El resultado se parece a lo encontrado en a) o en b)?

    d. ¿Podría encontrarse el mismo resultado que en a) frecuentistamente?

9. El jugador de fútbol Gonzalo Montiel convirtió los 12 penales que pateó, al día de la fecha (20 de Agosto de 2024), en su carrera profesional.

    a. Usando un prior beta de la probabilidad que tiene Montiel de convertir un penal, es decir $\theta \sim \text{beta}(\alpha, \beta)$ encontrar la distribución posterior para θ y graficarla (definir a gusto los parámetros α y β de la distribución).

    b. ¿Cuál es la probabilidad de que convierta el penal número 13? ¿Cómo se compara con la estimación frecuentista?

    c. ¿Qué supuestos estamos haciendo sobre el proceso que generó los datos?

    d. Haciendo simulaciones, crear un histograma de la distribución predicha de penales convertidos en los próximos 10 penales que ejecute Montiel (*posterior predictive distribution*).

    e. Estimar la probabilidad de que Montiel meta al menos 8 de los próximos 10 penales que patee.

10. Una máquina funciona perfectamente mientras tiene una sustancia que la protege. Sin embargo, esta sustancia se va consumiendo y cuando se agota, después de un tiempo θ, puede fallar en algún momento aleatorio que sigue una distribución exponencial. El tiempo que pasa hasta que produce la falla,x, sigue una distribución exponencial truncada dada por:

$$f(x|\theta) = \begin{cases} 0 & x < \theta \\ e^{-(x-\theta)} & x > \theta \end{cases}$$

    Se mide el tiempo de falla de tres máquinas obteniendo: datos $= \{10, 12, 15\}$. El objetivo es, a partir de estos datos, inferir θ. Obtener un intervalo de confianza frecuentista y un intervalo de credibilidad bayesiano para θ.