En Busqueda de la Prevalencia

Iván Agustín Bravo, Paz Lavenia, Pedro Bergaglio
bravoivanagustin@gmail.com, pazlavenia@gmail.com, pedrobergaglio04@gmail.com
Universidad de Buenos Aires

November 2025

1. Introducción

El presente trabajo se enmarca en el problema de estimar la prevalencia de una enfermedad, es decir, la proporción de individuos que tienen una enfermedad en una población determinada en un momento específico.

Este problema es ampliamente investigado en el área de epidemología, ya que permite describir un fenómeno de salud, identificar la frecuencia poblacional del mismo y generar hipótesis explicatorias. Para analizar este problema utilizaremos distintas técnicas estadísticas y experimentación por medio de simulaciones.

Los conceptos teoricos utilizados serán previamente presentados y explicados según corresponda.

1.1. Definiciones

A lo largo de este trabajo se utilizará la siguiente notación:

- T: variable aleatoria que representa el resultado del test diagnóstico. Vale 1 en caso de que el resultado sea positivo y 0 en caso contrario.
- Y: variable aleatoria que representa si una individuo está enfermo o no. Vale 1 en caso de que lo esté y 0 en caso contrario.
- Se: sensibilidad del test, se define como $P(T = 1 \mid Y = 1)$.
- Sp: especificidad del test, se define como $P(T = 0 \mid Y = 0)$.
- θ: prevalencia de la enfermedad, se define como $P(Y = 1)$.

2. Test Perfecto

Se llama test diagnóstico perfecto a aquel que arroja un resultado positivo cuando el individuo al cual se evalúa está enfermo y arroja un resultado negativo en caso contrario, por lo que cumple se que $Se = Sp = 1$.

En esta primera sección del trabajo se trabajará con un test diagnóstico perfecto, a partir de una muestra y utilizando este test se intentará estudiar una estimación de θ.

Para poder realizar una estimación de θ se tomará una muestra de n individuos de la población de interés. Se nota T_{per} a la variable aleatoria que cuenta la cantidad de individuos enfermos en la muestra. Para cada $i \in \{1, \dots, n\}$ se define X_i como la variable aleatoria que representa si el i-ésimo individuo de la muestra está enferma, vale 1 en caso de que lo esté y 0 caso contrario.

El primer supuesto que se tendrá en cuenta es que X_i es independiente de X_j para $i \neq j$, con $i, j \in \{1, \dots, n\}$. Esto quiere decir que si un individuo está enfermo es independiente de si otro lo está.

Se puede observar que $\forall i \in \{1, \dots, n\},\ X_i \sim Bi(1, \theta)$. Esto se debe a que el rango de X_i es $\{0, 1\}$ y se tiene que $P(X_i = 1) = \theta$ por la definición de θ. Entonces tenemos una muestra aleatoria $X_1, \dots, X_n$ i.i.d. proveniente de una distribución $Bi(1, \theta)$

Por como se definió $\{X_i\}_{i=1}^n$ y utilizando que son independientes vale que:

$$T_{per} = \sum_{i=1}^{n} X_i \sim Bi(n, \theta) \tag{1}$$

Para obtener el estimador de máxima verosimilitud de θ primero es necesario definir la verosimilitud de θ sobre la muestra $X_1, \dots, X_n$. Como primer paso escribimos la función de probabilidad de una distribución $Bi(1, \theta)$ como:

$$f_\theta(x) = \theta^x \cdot (1 - \theta)^{(1-x)} \cdot \mathbb{1}_{\{0,1\}}(x) \tag{2}$$

Ahora que tenemos la función de probabilidad de cada X_i se puede definir la verosimilitud de θ. Como esta misma se calcula en función de una realización de la muestra y la indicadora de la función de probabilidad no depende de θ podemos suponer que la realización de cada muestra ya pertenece al $\{0, 1\}$ podemos escribirla utilizando la formula (2) como:

$$L(\theta) = \hat{f}_\theta(x_1, \dots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta^{x_i} \cdot (1 - \theta)^{1-x_i} \tag{3}$$

Donde $\hat{f}_\theta$ denota la función de probabilidad conjunta de la muestra, usando la independencia de la muestra se puede separar en el producto de la función de probabilidad de cada muestra. A partir de la formula (3) podemos calcular la log-verosimilitud de θ como:

$$l(\theta) = ln(L(\theta)) = \sum_{i=1}^{n} x_i \cdot ln(\theta) + (1 - x_i) \cdot ln(1 - \theta) \tag{4}$$

Debido a que el logaritmo natural es una función estrictamente monótona vale que un θ realiza el máximo absoluto de $ln(L(\theta))$ tambien realiza el máximo absoluto de $L(\theta)$.

El siguiente paso para encontrar el estimador de máxima verosimilitud es encontrar el máximo absoluto de $l(\theta)$. Derivando la expresión (4) obtenemos que:

$$l'(\theta) = \sum_{i=1}^{n} \frac{x_i}{\theta} - \frac{1 - x_i}{1 - \theta} = \frac{1}{\theta} \cdot \left(\sum_{i=1}^{n} x_i \right) - \frac{1}{1 - \theta} \left(n - \sum_{i=1}^{n} x_i \right) \tag{5}$$

Entonces buscando los θ donde $l'(\theta) = 0$ se obtiene que:

$$l'(\theta) = 0 \iff \theta = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \bar{x} \tag{6}$$

Además derivando la expresión (5) se obtiene:

$$l''(\theta) = -\frac{1}{\theta^2} \cdot \left(\sum_{i=1}^{n} x_i \right) - \frac{1}{(1 - \theta)^2} \left(n - \sum_{i=1}^{n} x_i \right) \tag{7}$$

Y ademas como $x_i \in \{0, 1\}$:

$$0 \leq \sum_{i=1}^{n} x_i \leq n \Longrightarrow \sum_{i=1}^{n} x_i \geq 0\ ,\ n - \sum_{i=1}^{n} x_i \geq 0 \tag{8}$$

Como además θ^2 y $(1 - \theta)^2$ son positivos entonces $l''(\theta) \leq 0$. En particular $l''(\bar{x}) \leq 0$, esto sumado a que $l(\theta)$ es una función continua y que $\bar{x}$ es el único punto crítico se tiene que $\bar{x}$ es el unico punto donde se realiza un absoluto de $l(\theta)$.

Comentario: Preguntar si hay que tener salvedad sobre los extremos muy estrictamente. Es un quilombo si nos queremos poner a hablar de que pasa ahi, no están bien definidas las funciones pero pueden ser un estimador valido.

Entonces $\hat{\theta}_{obs} = \bar{x}$, este mismo es una estimación de θ utilizando el estimador de máxima verosimilitud, dada por una realización de la muestra $X_1, \ldots, X_n$. Por lo tanto se puede generalizar al estimador de máxima verosimilitud de θ como:

$$\hat{\theta}_{per} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} = \frac{T_{per}}{n} \tag{9}$$

2.1. Propiedades de $\hat{\theta}_{per}$

Estudiemos un par de características de este estimador. Primero el sesgo:

$$\mathbb{E}(\hat{\theta}_{per}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mathbb{E}(X_1) = \theta \Longrightarrow \mathbb{B}(\hat{\theta}_{per}) = 0 \tag{10}$$

Para la varianza usamos fuertemente la independencia entre las muestras para poder separar la suma, tenemos que:

$$\mathbb{V}(\hat{\theta}_{per}) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{1}{n}\mathbb{V}(X_1) = \frac{\theta\cdot(1-\theta)}{n} \tag{11}$$

Como el estimador es insesgado se obtiene que:

$$ECM(\hat{\theta}_{per}) = \mathbb{V}(\hat{\theta}_{per}) + \left(\mathbb{B}(\hat{\theta}_{per})\right)^2 = \frac{\theta\cdot(1-\theta)}{n} \tag{12}$$

Por Ley Fuerte de los Grandes Numeros se tiene que:

$$\overline{X} \xrightarrow{c.s} \mathbb{E}(X_1) \Longrightarrow \hat{\theta}_{per} \xrightarrow{c.s} \theta \tag{13}$$

Sea Z una variable aleatoria tal que $Z \sim N(0,1)$. Por Teorema Central del Limite se tiene:

$$\sqrt{n}\cdot\frac{\overline{X} - \mathbb{E}(X_1)}{\sqrt{\mathbb{V}(X_1)}} \xrightarrow{D} Z \tag{14}$$

Entonces utilizando lo obtenido en (10), (11) y (14) se tiene que:

$$\sqrt{n}\cdot\frac{\hat{\theta}_{per} - \theta}{\sqrt{\theta\cdot(1-\theta)}} \xrightarrow{D} Z \Longrightarrow \sqrt{n}\cdot\left(\hat{\theta}_{per} - \theta\right) \xrightarrow{D} \sqrt{\theta\cdot(1-\theta)}\cdot Z = W \tag{15}$$

Donde W es una variable aleatoria que se distribuye como $N(0, \theta\cdot(1-\theta))$. Entonces la distribución asintotica de $\hat{\theta}_{per}$ es $N(0, \theta\cdot(1-\theta))$.

Para construir un intervalo de confianza se puede utilizar la primera parte de la implicación (15) junto con el Teorema de Slutsky, con estos mismo se obtiene que:

$$\sqrt{n}\cdot\frac{\hat{\theta}_{per} - \theta}{\sqrt{\hat{\theta}_{per}\cdot(1-\hat{\theta}_{per})}} = \sqrt{n}\cdot\frac{\hat{\theta}_{per} - \theta}{\sqrt{\theta\cdot(1-\theta)}}\cdot\frac{\sqrt{\theta\cdot(1-\theta)}}{\sqrt{\hat{\theta}_{per}\cdot(1-\hat{\theta}_{per})}} \xrightarrow{D} Z \tag{16}$$

Entonces utilizando la definición de convergencia en distribución:

$$P\left(-z_{\frac{\alpha}{2}} \leq \sqrt{n}\cdot\frac{\hat{\theta}_{per} - \theta}{\sqrt{\hat{\theta}_{per}\cdot(1-\hat{\theta}_{per})}} \leq z_{\frac{\alpha}{2}}\right) \longrightarrow 1-\alpha \tag{17}$$

De lo que se puede deducir que un intervalo de nivel asintótico $1-\alpha$ para θ se puede obtener con:

$$IC_\alpha(\theta) = \left[\hat{\theta}_{per} - \sqrt{\frac{\hat{\theta}_{per}\cdot(1-\hat{\theta}_{per})}{n}}\cdot z_{\frac{\alpha}{2}}, \hat{\theta}_{per} + \sqrt{\frac{\hat{\theta}_{per}\cdot(1-\hat{\theta}_{per})}{n}}\cdot z_{\frac{\alpha}{2}}\right] \tag{18}$$

3. Test Imperfecto

A partir de este momento en el trabajo no se supondrá que el test con el cual se trabaja es perfecto. Se define p como la probabilidad de que el test sea positivo, o sea, $p = P(T=1)$. El objetivo de esta sección es estudiar una estimación sobre p.

Veasé que:

$$\begin{aligned} p = P(T=1) &= P(T=1 \mid Y=1)\cdot P(Y=1) + P(T=1 \mid Y=0)\cdot P(Y=0) \\ &= Se\cdot\theta + (1-Sp)\cdot(1-\theta) \end{aligned} \tag{19}$$

Entonces si se quiere utilizar lo estudiado en la sección previa para estimar p se puede utilizar el estimador encontrado para θ:

$$\hat{p} = \hat{\theta}_{per}\cdot(Se+Sp-1) + (1-Sp) = \frac{T_{per}}{n}\cdot(Se+Sp-1) + (1-Sp) \tag{20}$$

Valores particulares de Se, Sp y θ se pueden reemplazar en la formula (3) para obtener el valor del p asociado. Para los valores $Se = 0{,}9$, $Sp = 0{,}95$ y $\theta = 0{,}25$ se tiene que $p = 0{,}2625$. ~~Observese que ocurre cuando varia uno de los parámetros mientras que el resto se mantiene constante.~~

Comentario: armar graficos y subirlos. Ponemos p como función de Sp, Se y θ?

3.1. Estimador de Momentos

En el trabajo se supondra que $Se+Sp-1>0$, la cual es una condición mínima indispensable para que el test se considere una herramienta de medición válida. Para observar porque tiene sentido esta suposición se puede reescribir la desigualdad como $Se > 1-Sp$, si ahora se reemplazan los terminos Se y Sp por sus definiciones se obtiene que:

$$P(T=1 \mid Y=1) > 1 - P(T=0 \mid Y=0) = P(T=1 \mid Y=0) \tag{21}$$

Esta relación implica que la probabilidad de obtener un resultado positivo en el test es mayor cuando se evalua a un individuo enfermo que cuando se evalua a uno sano. Análogamente, esto asegura que $Sp > 1-Se$, es decir, la probabilidad de obtener un resultado negativo en el test es mayor cuando se evalua un individuo sano que cuando se evalua a uno enfermo. En definitiva, la asunción garantiza que el test posee una capacidad discriminativa mejor que la de elegir de manera aleatoria quien está enfermo y quien no.

~~**Comentario**: Agregamos todo esto?~~

Teniendo en cuenta lo ~~explicado~~ anteriormente se puede despejar θ como:

$$\theta = \frac{p+Sp-1}{Sp+Se-1} \tag{22}$$

Si se tiene una muestra de tamaño n sobre una población de interés, sobre la cual se aplica el test diagnóstico, y se supone que el resultado del test es independiente entre dos individuos, entonces se obtiene una muestra aleatoria $T_1,\ldots,T_n$. Donde T_i se define como la variable aleatoria que representa el resultado del test aplicado al i-ésimo individuo, que vale 1 si el resultado es positivo y 0 caso contrario. Se cumple que $\forall i \in \{1,\ldots,n\}$, $P(T_i=1)=p$ y que su rango es $\{0,1\}$, por lo tanto $\forall i \in \{1,\ldots,n\}$, $T_i \sim Bi(1,p)$.

Debido a los resultados anteriores se obtiene que $\mathbb{E}(T_1) = p$, y por el método de momentos generalizados, utilizando la ecuación (22) se cumple que:

$$q(\theta) = \theta,\ g(T_1) = \frac{T_1 + Sp - 1}{Sp + Se - 1} \Longrightarrow \widehat{q(\theta)} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i + Sp - 1}{Sp + Se - 1} \tag{23}$$

Entonces el estimador de momentos de θ es:

$$\hat{\theta}_{MoM} = \frac{\overline{T} + Sp - 1}{Sp + Se - 1} \tag{24}$$

Veamos un par de características sobre este estimador. Primero el sesgo:

$$\mathbb{E}(\hat{\theta}_{MoM}) = \mathbb{E}\left(\frac{\overline{T} + Sp - 1}{Sp + Se - 1}\right) = \frac{\mathbb{E}(\overline{T}) + Sp - 1}{Sp + Se - 1} = \frac{p + Sp - 1}{Sp + Se - 1} = \theta \Longrightarrow \mathbb{B}(\hat{\theta}_{per}) = 0 \tag{25}$$

Para la varianza usamos fuertemente la independencia entre las muestras para poder separar la suma, tenemos que:

$$\mathbb{V}(\hat{\theta}_{MoM}) = \mathbb{V}\left(\frac{\overline{T} + Sp - 1}{Sp + Se - 1}\right) = \frac{\mathbb{V}(\overline{T})}{(Sp + Se - 1)^2} = \frac{\frac{1}{n}\mathbb{V}(T_1)}{(Sp + Se - 1)^2} = \frac{p \cdot (1-p)}{n \cdot (Sp + Se - 1)^2} \tag{26}$$

Como el estimador es insesgado se obtiene que:

$$ECM(\hat{\theta}_{per}) = \mathbb{V}(\hat{\theta}_{per}) + \left(\mathbb{B}(\hat{\theta}_{per})\right)^2 = \frac{p \cdot (1-p)}{n \cdot (Sp + Se - 1)^2} \tag{27}$$

Por Ley Fuerte de los Grandes Numeros se tiene que:

$$\overline{T} \xrightarrow{c.s} \mathbb{E}(T_1) = p \Longrightarrow \frac{\overline{T} + Sp - 1}{Sp + Se - 1} \xrightarrow{c.s} \frac{p + Sp - 1}{Sp + Se - 1} \Longrightarrow \hat{\theta}_{MoM} \xrightarrow{c.s} \theta \tag{28}$$

Comentario: Hacer graficos de ECM.
Comentario: Hacer simulacion.
Comentario: Hacer bootstrap.

3.2. Intervalos de Confianza

Comentario: Hacer bootstrap.

Sea Z una variable aleatoria tal que $Z \sim N(0,1)$ por Teorema Central del Limite y Teorema de Slutsky vale que:

$$\sqrt{n} \cdot \frac{\overline{T} - p}{\sqrt{p \cdot (1-p)}} \xrightarrow{D} Z \Longrightarrow \sqrt{n} \cdot (\overline{T} - p) \xrightarrow{D} \sqrt{p \cdot (1-p)} \cdot Z = W \tag{29}$$

Con $W \sim N(0, p \cdot (1-p))$. Entonces definiendo la función:

$$q(x) = \frac{x + Sp - 1}{Se + Sp - 1} \tag{30}$$

Utilizando que la función q de (30) es de clase C^1 y su derivada no se anula, vale por Método Delta que:

$$\sqrt{n} \cdot (q(\overline{T}) - q(p)) \xrightarrow{D} \frac{1}{Se + Sp - 1} \cdot W = U \Longrightarrow \sqrt{n} \cdot (\hat{\theta}_{MoM} - \theta) \xrightarrow{D} U$$

$$\text{Tal que } U \sim N(0, \frac{p \cdot (1-p)}{(Se + Sp - 1)^2}) \tag{31}$$

Nuevamente se puede estandarizar la normal obtenida y a partir de utilizar el Teorema de Slutsky, sabiendo que $\overline{T} \xrightarrow{c.s.} p$, se pueden formular intervalos de confianza asintóticos para θ. Por ende se tiene que:

$$\sqrt{n}\frac{(Se+Sp-1)\cdot(\hat{\theta}_{MoM}-\theta)}{\sqrt{p\cdot(1-p)}} \xrightarrow{D} Z \Longrightarrow Z_n = \sqrt{n}\frac{(Se+Sp-1)\cdot(\hat{\theta}_{MoM}-\theta)}{\sqrt{\overline{T}\cdot(1-\overline{T})}} \xrightarrow{D} Z \tag{32}$$

Entonces por la definición de convergencia en distribución se tiene que:

$$P(-z_{\frac{\alpha}{2}} \leq Z_n \leq z_{\frac{\alpha}{2}}) \longrightarrow 1-\alpha \tag{33}$$

Se puede deducir por ende que:

$$IC_\alpha(\theta) = \left[\hat{\theta}_{MoM} - z_{\frac{\alpha}{2}}\cdot\frac{\sqrt{\overline{T}\cdot(1-\overline{T})}}{\sqrt{n}\cdot(Se+Sp-1)},\ \hat{\theta}_{MoM} + z_{\frac{\alpha}{2}}\cdot\frac{\sqrt{\overline{T}\cdot(1-\overline{T})}}{\sqrt{n}\cdot(Se+Sp-1)}\right] \tag{34}$$

Es un intervalo de nivel asintótico $1-\alpha$ para θ.

3.3. Estimador Truncado

Observese que:

$$0 \leq \hat{\theta}_{MoM} \leq 1 \Longleftrightarrow 1-Sp \leq \overline{T} \leq Se \tag{35}$$

Entonces aunque teóricamente $\theta \in [0,1]$, debido a la variabilidad muestral es posible que $\overline{T}$ caiga fuera del intervalo $[1-Sp, Se]$. En esos casos, el estimador de momentos arroja valores fuera del espacio paramétrico lógico.

La desigualdad $\overline{T} < 1-Sp$ implica que en la muestra se obtuvo una proporción de positivos inferior a la tasa de positivos que se esperaría tener únicamente por error. Es decir, la cantidad de positivos que se obtuvieron es menor a la cantidad de positivos que debería haber en una población completamente sana. Matemáticamente, el estimador interpreta esto como que tiene que haber probabilidad negativa de que un individuo esté enfermo.

La desigualdad $\overline{T} > Se$ implica que en la muestra se obtuvo una proporción de positivos superior a la tasa de verdaderos positivos que se esperaría tener realmente. Es decir, la cantidad de positivos es mayor a la cantidad de positivos que habría aunque la población estuviera completamente infectada. Matemáticamente, el estimador interpreta esto como que tiene que haber probabilidad mayor a 1 de que un individuo esté enfermo.

Para esto definimos el estimador truncado de θ como:

$$\hat{\theta}_{trunc} = \begin{cases} \hat{\theta}_{MoM} & \text{si } 0 \leq \hat{\theta}_{MoM} \leq 1 \\ 0 & \text{si } \hat{\theta}_{MoM} < 0 \\ 1 & \text{si } \hat{\theta}_{MoM} > 1 \end{cases} \tag{36}$$

Comentario: Hacer simulación

4. Dos muestras

¿Que pasa si la población pasa por una vacunación? ¿Cambia la prevalencia?

Esta sección se basará en decidir si la aplicación de una vacuna surge efecto en la prevalencia de la enfermedad. Para esto se obtienen dos muestras, una previa y una posterior a la vacunación. Se utilizarán parte de los resultados obtenidos hasta el momento.

Para esta última sección se utiliza la siguiente notación:

- X_A: variable aleatoria que representa la cantidad de individuos para los cuales el test arrojo un resultado positivo en la etapa A. Donde A puede ser pre o post, con pre refiriendose a la etapa previa a la vacunación y post a la etapa posterior.
- n_A: tamaño de la muestra en la etapa A.
- θ_A: prevalencia de la enfermedad en la etapa A.
- X_A^i: variable aleatoria que representa si el resultado del test arroja un resultado positivo cuando se evalua al i-ésimo individuo de la muestra en la etapa A, vale 1 en caso de que arroje positivo y 0 caso contrario.
- p_A: probabilidad de que el test arroje un resultado positivo.

Utilizando las definiciones anteriores se tiene que el rango de X_A^i es $\{0,1\}$ y además $\forall i \in \{1,\dots,n_A\}$, $P(X_A^i = 1) = p_A$, por lo cual vale que $\forall i \in \{1,\dots,n_A\}$, $X_A^i \sim Bi(1,p_A)$. Además se tendrá como supuesto que X_A^i es independiente de X_A^j para $i \neq j$, con $i,j \in \{1,\dots,n_A\}$. Con esto vale que $X_{pre}^1,\dots X_{pre}^{n_{pre}}$ y $X_{post}^1,\dots X_{post}^{n_{post}}$ son dos muestras aleatorias independientes entre si.

Por como se definió $\{X_A^i\}_{i=1}^{n_A}$ y utilizando que son independientes vale que:

$$X_A = \sum_{i_1}^{n_A} X_A^i \sim Bi(n_A, p_A) \tag{37}$$

4.1. Test de Hipotesis

Al querer predecir si cambió la prevalencia después de la etapa de vacunación se puede estudiar la diferencia entre la prevalencia previa y la posterior, por lo cual se define:

$$\Delta = \theta_{post} - \theta_{pre} \tag{38}$$

Entonces se definen las hipotesis:

$$H_0 : \Delta = 0 \quad \text{vs} \quad H_1 : \Delta \neq 0 \tag{39}$$

Es necesario encontrar un estadístico para obtener un test de nivel aproximado (o asintótico) α. Veamos como construirlo utilizando el Metodo del Pivote.

Utilizando lo estudiado en la sección previa se puede estimar θ_A como:

$$\hat{\theta}_A = \frac{\overline{X}_A + Sp - 1}{Se + Sp - 1} \tag{40}$$

Donde:

$$\overline{X}_A = \frac{1}{n}\sum_{i=1}^{n_A} X_A^i = \frac{1}{n} X_A \tag{41}$$

Entonces utilizando el resultado (31) tenemos que:

$$\sqrt{n_A} \cdot (\hat{\theta}_A - \theta_A) \xrightarrow{D} U_A \sim N(0, \sigma_A^2) \quad \text{con } \sigma_A^2 = \frac{p_A \cdot (1-p_A)}{(Sp + Se - 1)^2} \tag{42}$$

Todo esto se realiza con la motivacion de encontrar la distribucion asintótica de $\hat{\theta}_{post} - \hat{\theta}_{pre}$. Entonces se define $N = n_{post} + n_{pre}$. Además para poder trabajar con esta distribución se tendrá como supuesto que existen constantes λ_{post} y λ_{pre} mayores a 0 tal que:

$$\lim_{N\to\infty} \frac{n_{post}}{N} = \lambda_{post} \quad \text{y} \quad \lim_{N\to\infty} \frac{n_{pre}}{N} = \lambda_{pre} \tag{43}$$

Cuando se nota $N \to \infty$, tanto n_{post} como n_{pre} tienden a infinito. Este supuesto representa que la diferencia entre la cantidad de muestras de un grupo y el otro no sea demasiado grande.

Como vale que:

$$\mathbb{E}(\hat{\theta}_{post} - \hat{\theta}_{pre}) = \mathbb{E}(\hat{\theta}_{post}) - \mathbb{E}(\hat{\theta}_{pre}) = \theta_{post} - \theta_{pre} = \Delta \tag{44}$$

Se quiere estudiar cual es la distribución de:

$$\sqrt{N}(\hat{\theta}_{post} - \hat{\theta}_{pre} - \Delta) = \sqrt{N}(\hat{\theta}_{post} - \theta_{post}) - \sqrt{N}(\hat{\theta}_{pre} - \theta_{pre}) \tag{45}$$

Utilizando que las muestras previas y posteriores a la vacunación son independientes entre si vale que $\overline{X}_{pre}$ y $\overline{X}_{post}$ son independientes, por lo tanto $\hat{\theta}_{pre}$ y $\hat{\theta}_{post}$ son independientes. Utilizando esto y el Teorema de Slutsky se obtiene que:

$$\begin{pmatrix} \sqrt{n_{post}} \cdot (\hat{\theta}_{post} - \theta_{post}) \\ \sqrt{n_{pre}} \cdot (\hat{\theta}_{pre} - \theta_{pre}) \end{pmatrix} \xrightarrow{D} \begin{pmatrix} U_{post} \\ U_{pre} \end{pmatrix} \Longrightarrow \begin{pmatrix} \sqrt{N} \cdot (\hat{\theta}_{post} - \theta_{post}) \\ \sqrt{N} \cdot (\hat{\theta}_{pre} - \theta_{pre}) \end{pmatrix} \xrightarrow{D} \begin{pmatrix} \sqrt{\frac{1}{\lambda_{post}}} U_{post} \\ \sqrt{\frac{1}{\lambda_{pre}}} U_{pre} \end{pmatrix} \tag{46}$$

Utilizando el Teorema de la Aplicación Continua (van der Vaart, 1998, Teorema 2.3), dado que la función $g(u, v) = u - v$ es continua en todo $\mathbb{R}^2$ y el vector conjunto converge en distribución, se tiene que:

$$g(\sqrt{N} \cdot (\hat{\theta}_{post} - \theta_{post}), \sqrt{N} \cdot (\hat{\theta}_{pre} - \theta_{pre})) \xrightarrow{D} g(\sqrt{\frac{1}{\lambda_{post}}} U_{post}, \sqrt{\frac{1}{\lambda_{pre}}} U_{pre}) \tag{47}$$

Entonces:

$$\sqrt{N}(\hat{\theta}_{post} - \hat{\theta}_{pre}) = \sqrt{N}(\hat{\theta}_{post} - \theta_{post}) - \sqrt{N}(\hat{\theta}_{pre} - \theta_{pre}) \xrightarrow{D} \sqrt{\frac{1}{\lambda_{post}}} U_{post} - \sqrt{\frac{1}{\lambda_{pre}}} U_{pre} \tag{48}$$

Utilizando la distribución de U_A vale que:

$$\sqrt{\frac{1}{\lambda_A}} U_A \sim N(0, \frac{\sigma_A^2}{\lambda_A}) \tag{49}$$

Entonces para la diferencia se tiene que:

$$U = \sqrt{\frac{1}{\lambda_{post}}} U_{post} - \sqrt{\frac{1}{\lambda_{pre}}} U_{pre} \sim N(0, \frac{\sigma_{post}^2}{\lambda_{post}} + \frac{\sigma_{pre}^2}{\lambda_{pre}}) \tag{50}$$

Si se estandariza la varianza de esta normal se obtiene que:

$$\sqrt{N} \frac{(\hat{\theta}_{post} - \hat{\theta}_{pre} - \Delta)}{\sqrt{\frac{\sigma_{post}^2}{\lambda_{post}} + \frac{\sigma_{pre}^2}{\lambda_{pre}}}} \xrightarrow{D} Z \sim N(0, 1) \tag{51}$$

Se podria aplicar el Metodo del Pivote aqui mismo, por la salvedad de que los σ^2 y los λ son desconocidos. Vease que si tomamos:

$$\hat{\sigma}_A^2 = \frac{\hat{p}_A \cdot (1 - \hat{p}_A)}{(Sp + Se - 1)^2}, \quad \text{con } \hat{p}_A = (Se + Sp - 1)\hat{\theta}_A + (1 - sp) \tag{52}$$

Y sabiendo que $\hat{\theta}_A \xrightarrow{c.s} \theta_A$ vale que:

$$\hat{\sigma}_A^2 \xrightarrow{c.s} \sigma_A^2 \quad y \quad \frac{n_A}{N} \xrightarrow{c.s} \lambda_A \tag{53}$$

Entonces:

$$\sqrt{\frac{N\cdot\hat{\sigma}^2_{post}}{n_{post}}+\frac{N\cdot\hat{\sigma}^2_{pre}}{n_{pre}}}\xrightarrow{c.s}\sqrt{\frac{\sigma^2_{post}}{\lambda_{post}}+\frac{\sigma^2_{pre}}{\lambda_{pre}}} \tag{54}$$

Por lo cual aplicando el Teorema de Slutsky se obtiene que:

$$W_{n_{post},n_{pre}}=\frac{\hat{\theta}_{post}-\hat{\theta}_{pre}-\Delta}{\sqrt{\frac{\hat{\sigma}^2_{post}}{n_{post}}+\frac{\hat{\sigma}^2_{pre}}{n_{pre}}}}=\sqrt{N}\frac{(\hat{\theta}_{post}-\hat{\theta}_{pre}-\Delta)}{\sqrt{\frac{N\cdot\hat{\sigma}^2_{post}}{n_{post}}+\frac{N\cdot\hat{\sigma}^2_{pre}}{n_{pre}}}}\xrightarrow{D}Z\sim N(0,1) \tag{55}$$

Con esto se logró obtener un pivote, el cual es decreciente en Δ y converge a una distribución que no depende de Δ.

Por lo tanto se puede obtener un estadístico reemplazando el parámetro Δ por $\Delta_0=0$, de lo cuál resulta que el estadístico es:

$$W=\frac{\hat{\theta}_{post}-\hat{\theta}_{pre}}{\sqrt{\frac{\hat{\sigma}^2_{post}}{n_{post}}+\frac{\hat{\sigma}^2_{pre}}{n_{pre}}}} \tag{56}$$

Entonces según el Metodo del Pivote tenemos que el test ϕ con región de rechazo:

$$\frac{|\hat{\theta}_{post}-\hat{\theta}_{pre}|}{\sqrt{\frac{\hat{\sigma}^2_{post}}{n_{post}}+\frac{\hat{\sigma}^2_{pre}}{n_{pre}}}}>z_{\frac{\alpha}{2}} \tag{57}$$

Es un test de nivel asintótico α para las hipotesis (39). Para que sea de nivel 0,05 reemplazamos α por ese valor, de lo cual se obtiene que la región de rechazo es:

$$\frac{|\hat{\theta}_{post}-\hat{\theta}_{pre}|}{\sqrt{\frac{\hat{\sigma}^2_{post}}{n_{post}}+\frac{\hat{\sigma}^2_{pre}}{n_{pre}}}}>1{,}96 \tag{58}$$

Comentario: revisar cuentas de esta partecita

Para el caso particular en que se tienen los datos $n_{pre}=n_{post}=100$, $Se=0{,}90$, $Sp=0{,}95$, $\hat{\theta}_{pre}=0{,}2$, $\hat{\theta}_{post}=0{,}15$ y $\alpha=0{,}05$. De estos mismos se obtiene que $\hat{p}_{pre}=0{,}22$ y $\hat{p}_{post}=0{,}1775$, por lo tanto $\hat{\sigma}^2_{pre}=0{,}2375$ y $\hat{\sigma}^2_{post}=0{,}2021$. Si se nota W_{obs} a la observación del estadístico se tiene que:

$$|W_{obs}|=|-0{,}754|=0{,}754<1{,}96 \tag{59}$$

Por el test planteado y los datos observados se puede decir que no hay suficiente evidencia estadística para rechazar la hipotesis nula. Esto quiere decir que no hay razón para creer que cambió la prevalencia luego de la vacunación.

Cuando se disminuyen los tamaños de las muestra aumenta el error estándar de $\hat{\theta}_{post}-\hat{\theta}_{pre}$, por lo tanto se tiene mas incertidumbre acerca de las mediciones, esto quiere decir que es mas dificil rechazar H_0. Por ejemplos en el caso de $n_{pre}=n_{post}=25$ y mantieniendo el resto de los datos iguales se obtiene que:

$$|W_{obs}|=|-0{,}377|=0{,}377<0{,}754<1{,}96 \tag{60}$$

4.2. Intervalo de Confianza

En este apartado se buscará un intervalo de confianza de nivel asintótico $1-\alpha$ para el parametro Δ. Se partirá del resultado obtenido en (55), por la definición de convergencia en distribución vale que:

$$P\big(-z_{\frac{\alpha}{2}} \leq W_{n_{post},n_{pre}} \leq z_{\frac{\alpha}{2}}\big) \longrightarrow 1-\alpha \tag{61}$$

Por lo tanto despejando Δ en esta desigualdad se puede deducir que:

$$\left[\hat{\theta}_{post} - \hat{\theta}_{pre} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{N \cdot \hat{\sigma}^2_{post}}{n_{post}} + \frac{N \cdot \hat{\sigma}^2_{pre}}{n_{pre}}}, \hat{\theta}_{post} - \hat{\theta}_{pre} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{N \cdot \hat{\sigma}^2_{post}}{n_{post}} + \frac{N \cdot \hat{\sigma}^2_{pre}}{n_{pre}}}\right] \tag{62}$$

Es un intervalo de confianza de nivel asintótico $1-\alpha$ para Δ.
Comentario: falta hacer simulación

Referencias

[1] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press

¿otras?