LECCIÓN 5: INFERENCIA ESTADÍSTICA

SAAVEDRA, P.

1. El problema de la inferencia estadística

El ***inductivismo*** es un método científico que elabora conclusiones generales a partir de enunciados observacionales particulares y por tanto, lleva de lo ***particular*** a lo ***general***. Bertrand Russell fundamenta la inducción en el concepto de *uniformidad de la naturaleza* merced al cual el curso de ésta se mantiene constante, de modo que el futuro ha de parecerse al pasado dado que a causas semejantes siguen efectos semejantes. *Las uniformidades pasadas causan expectativas con respecto al futuro* (Russell, 1973).

Tomamos este debate como punto de partida para hacer una aproximación al concepto de ***inferencia estadística*** y lo haremos considerando el problema de establecer si un nuevo tratamiento terapéutico para una enfermedad realmente la curará, o cual es la probabilidad de que lo haga (*tasa de respuesta favorable al tratamiento*). La *complejidad* de los seres vivos conduce a que normalmente sólo podamos hablar de *probabilidad de curación*. La necesidad de predecir en el futuro la evolución de la enfermedad justifica la necesidad de conocer esta probabilidad, la cual sería una ley que nos dice algo acerca de las expectativas de curación de cualquier sujeto de la población de enfermos que han recibido el tratamiento de estudio. Las cantidades que nos informan de determinados aspectos de tales poblaciones, en este caso la tasa de respuesta al tratamiento, reciben el nombre de ***parámetros***.

Ahora bien, ¿cómo podríamos conocer el verdadero valor de los parámetros? Para el problema de la tasa de respuesta a un tratamiento, casi todo el mundo pensaría inmediatamente que mediante la aplicación del tratamiento a un subconjunto de pacientes (muestra) y ***estimando*** el valor desconocido del parámetro por la frecuencia de respuestas favorables. Esto es, ***extrapolando*** los resultados de la muestra a toda la población. Ahora bien, si el estudio se repitiese con diferentes muestras, previsiblemente los estimaciones variarían entre los estudios. Por tanto, ¿qué valor tienen realmente las estimaciones de los parámetros? El propósito de esta lección es describir métodos que nos permitan hacer ***estimaciones fiables*** de parámetros.

Los elementos por tanto de un problema de ***inferencia estadística*** son:

- ***Universo*** o población de estudio.
- ***Parámetros*** que son cantidades que nos dicen algo sobre la población de estudio.
- ***Muestra*** o subconjunto de la población de estudio, normalmente extraída de forma aleatoria.
- ***Datos*** observados en los elementos de la muestra y que constituyen la base de la inferencia.
- ***Estadísticos*** o funciones de los datos, cuya finalidad es inferir a la población lo observado en la muestra (método inductivo).

2. Estimación puntual

En esta sección se considera el problema de ***estimar*** un parámetro desconocido θ a través de un conjunto de datos $\mathcal{X}$ obtenidos de una muestra aleatoria de la población de estudio. Haremos en primer lugar una aproximación heurística al concepto de estimador puntual a través de un estudio de simulación. El estudio pondrá de manifiesto que el estadístico ***estimador*** es una variable aleatoria y por tanto, que su fiablidad depende de su distribución de probabilidad. Se formalizará entonces el concepto de estimador puntual como un estadístico cuyo propósito es hacer buenas aproximaciones al verdadero valor de θ. Finalmente, se analizarán los estimadores de diversos parámetros simples.

Estudio de simulación. Sea un parámetro π que representa la tasa de respuesta favorable a un tratamiento. Su estimador natural es la *proporción de respuestas favorables* observadas en una muestra aleatoria de n pacientes que han recibido el tratamiento de estudio y al cual denotaremos por $\hat{\pi}_n$. Obviamente el valor de $\hat{\pi}_n$ depende de la muestra aleatoriamente seleccionada, lo que supone que es una variable aleatoria. Resumimos ahora el algoritmo de simulación en los siguientes pasos:

1. Supóngase que la tasa de respuestas es $\pi = 0.70$, lo que significa que cada vez que se aplica el tratamiento a un paciente, hay una probabilidad del 70% de que responda favorablemente.

2. Se aplica el tratamiento a una muestra aleatoria de n pacientes y se obtiene la proporción de respuestas favorables al mismo $\hat{\pi}_n$. Se considerarán los tamaños muestrales: $n = 20$, 100, 384 y 1000.

3. El paso anterior se repite 10,000 veces, lo que supone que se dispone de 10,000 observaciones de la variable aleatoria $\hat{\pi}_n$.

4. Representamos finalmente las 10,000 observaciones de $\hat{\pi}_n$ mediante un histograma para cada uno de los tamaños muestrales n.

En la figura 1 se muestran los histogramas correspondientes a las 10,000 observaciones del estimador $\hat{\pi}_n$ para los diferentes tamaños muestrales considerados. Nótese que para $n = 20$, una buena parte de las estimaciones están próximas al verdadero valor $\pi = 0.70$, pero algunas bajan hasta 0.40 lo que supone que el estimador carece de fiabilidad para este tamaño muestral. Para $n = 100$ raramente las estimaciones son inferiores a 0.60 o superiores a 0.80 lo que implica una mayor fiabilidad que en el caso anterior. El aumento a $n = 384$ conduce a una mayor fiabilidad (pocas estimaciones se alejan en más de 0.05 del verdadero valor de π). Nótese por último que para $n = 1000$, se consigue aún un cierto aumento de la fiabilidad (pocas estimaciones se alejan más de 0.03 del verdadero valor de π); sin embargo pasar de $n = 384$ a $n = 1000$ significa aumentar muchísimo el tamaño de la muestra para conseguir un ligero aumento de la fiabilidad. La idea a extraer del estudio es, por tanto, que el estimador $\hat{\pi}_n$ es una variable aleatoria y que su fiabilidad depende de su distribución de probabilidad. Más concretamente, que la distribución debe *concentrarse* alrededor del verdadero valor del parámetro.

Definición de estimador puntual. La estimación de cualquier parámetro desconocido θ se basa en general en un conjunto de datos $\mathcal{X}$ obtenidos aleatoriamente de la población de estudio, los cuales deben contener información acerca del verdadero valor de dicho parámetro.

En general, un ***estimador puntual*** para un parámetro θ basado en un conjunto de datos $\mathcal{X}$, es un estadístico (una función de los datos) que representaremos por: $\hat{\theta} = \hat{\theta}(\mathcal{X})$ y cuya finalidad es *hacer buenas aproximaciones* al verdadero valor de θ. La naturaleza aleatoria del conjunto de datos $\mathcal{X}$ da lugar a que el estimador $\hat{\theta}$ sea una ***variable aleatoria***, y de esta forma, su fiabilidad dependerá de su distribución de probabilidad. Dos propiedades deseables de cualquier estimador son:

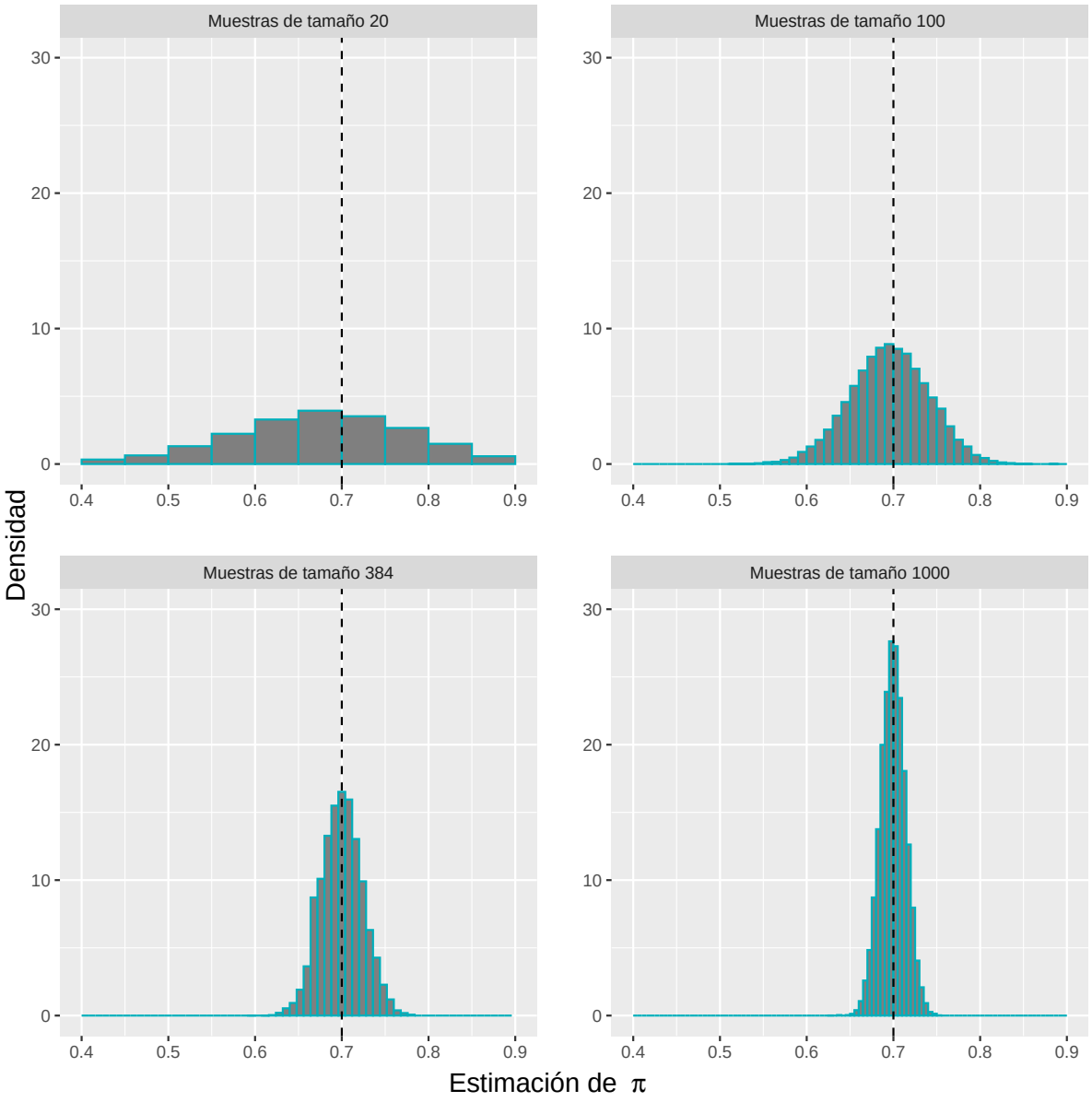

FIGURA 1. Esta simulación se basa en el hecho de que el verdadero valor del parámetro es $\pi = 0.7$. Para cada uno de los tamaños muestrales se han realizado 10,000 estimaciones. Este estudio pone de manifiesto que las estimaciones tienen como *centro de gravedad* el verdadero valor de π y que son más fiables (menor varianza) cuando el tamaño muestral es mayor.

- Que sea ***centrado***; esto es: que $E\left[\hat{\theta}\right] = \theta$. Esta propiedad supone que todas las estimaciones posibles tendrán como centro de gravedad el verdadero valor del parámetro θ.

- Que tenga poca varianza. En este sentido, se llama ***error estándar*** de un estimador a su desviación estándar; esto es: $\text{sd}\left(\hat{\theta}\right)$.

En aquellos casos en los que los datos consistan en una única muestra aleatoria de tamaño n, podemos expresar el estimador usando la notación $\hat{\theta}_n$. En este caso, una propiedad deseable del estimador es que $\text{sd}\left(\hat{\theta}_n\right) \longrightarrow 0$, según $n \longrightarrow \infty$.

3. Estimación de parámetros elementales

En esta sección se proponen estimadores para los parámetros más usuales y se obtienen sus esperanzas y varianzas. Asimismo se examinará el problema de determinar el tamaño muestral necesario para que el estimador tenga una fiabilidad predeterminada. Para ello utilizaremos el siguiente resultado (ver también la figura figura 1):

Proposición 1: Si $Z \cong N(0,1)$, entonces: $\Pr\left(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}\right) = 1-\alpha$.

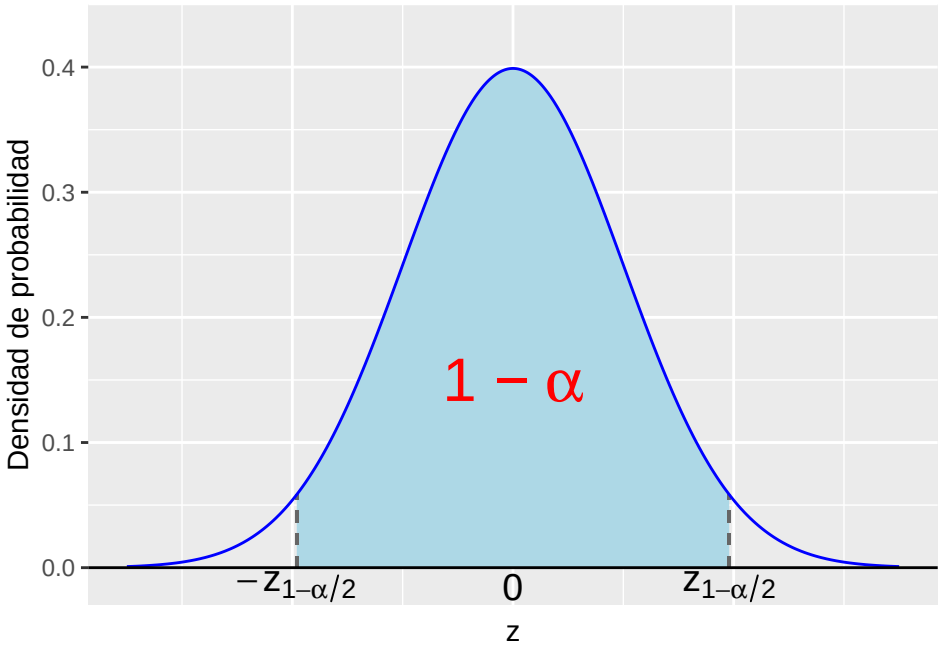

Figura 2. Distribución de probabilidad normal estándar

Recuérdese que z_q representa el cuantil q de la distribución normal estándar; esto es, si $Z \cong N(0,1)$ y se fija un valor $0 < q < 1$, z_q satisface: $q = \Pr\left(Z \leq z_q\right)$.

Probabilidades. Considérese un experimento aleatorio en el que la probabilidad de que ocurra un cierto suceso es π. Por ejemplo, π puede ser la probabilidad de respuesta favorable a un tratamiento terapéutico (*tasa de respuesta favorable al tratamiento*) o la probabilidad de que al seleccionar aleatoriamente un sujeto de una cierta población tenga una enfermedad (*prevalencia de esa enfermedad*).

En orden a estimar π el experimento aleatorio se repite n veces en las mismas condiciones de tal forma que los resultados de las sucesivas repeticiones sean independientes.

Asociada a cada una de las repeticiones del experimento aleatorio se definen las variables aleatorias X_i como 1 ó 0 según en la i-ésima repetición ocurra o no el suceso considerado (por ejemplo, que el paciente i-ésimo responda favorablemente al tratamiento considerado). De esta forma, $X_1, ..., X_n$ son variables aleatorias independientes y con distribución de probabilidad $b(1, \pi)$. Nótese que $E[X_i] = \pi$ y $\text{var}(X_i) = \pi(1-\pi)$.

Un estimador natural de π basado en $X_1, ..., X_n$ es la ***proporción muestral*** de veces que ocurre el suceso de interés, la cual se obtiene como:

$$\hat{\pi}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Veamos ahora que el estimador es centrado y calculemos su varianza. En efecto:

$$E[\hat{\pi}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n\pi = \pi$$

$$\text{var}(\hat{\pi}_n) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i) = \frac{\pi(1-\pi)}{n}$$

Supóngase ahora que se desea determinar el tamaño muestral n que se precisa para estimar el parámetro π a través de $\hat{\pi}_n$ con una precisión dada por la expresión:

$$\Pr(\pi - B \leq \hat{\pi}_n \leq \pi + B) = 1 - \alpha$$

Aquí, B recibe el nombre de ***cota de error*** y $1-\alpha$ de ***confianza***. Para resolver esta ecuación de precisión se requiere disponer de información acerca de la distribución de probabilidad de $\hat{\pi}_n$. Para ello tendremos en cuenta que $\hat{\pi}_n$ es una suma de variables

aleatorias independientes, y de ahí, la siguiente aproximación basada en el teorema central del límite:

$$\frac{\hat{\pi}_n - \pi}{\sqrt{\pi\left(1-\pi\right)/n}} \approx N\left(0,1\right)$$

Entonces, tipificando en la ecuación de precisión se obtiene:

$$\Pr\left(-\frac{B}{\sqrt{\pi\left(1-\pi\right)/n}} \leq \frac{\hat{\pi}_n - \pi}{\sqrt{\pi\left(1-\pi\right)/n}} \leq \frac{B}{\sqrt{\pi\left(1-\pi\right)/n}}\right) = 1-\alpha$$

Teniéndose en cuenta la proposición 1 se llega a:

$$\frac{B}{\sqrt{\pi\left(1-\pi\right)/n}} = z_{1-\alpha}$$

Despejando el tamaño muestral n queda:

$$n = \frac{z_{1-\alpha}^2}{B^2} \cdot \pi\left(1-\pi\right)$$

Nótese que la solución depende del parámetro que se desea estimar (π), cuyo valor no se conoce. La solución práctica puede aproximarse de dos formas alternativas, a saber:

1. Si se tiene una idea aproximada acerca del posible valor π, sustituir tal valor en la expresión de n.
2. Utilizar la desigualdad $\pi\left(1-\pi\right) \leq 1/4$ (es fácil su comprobación). En tal caso, la solución queda acotada en la forma: $n \leq z_{1-\alpha}^2/\left(4B^2\right)$.

Seleccionando los valores $\alpha = B = 0.05$, puede comprobarse que $n \leq 384$.

Esperanzas. Considérese una magnitud aleatoria cuya distribución de probabilidad $\mathcal{P}$ tiene esperanza μ y varianza σ^2. En orden a estimar μ se selecciona una muestra aleatoria $X_1, \ldots, X_n$ de $\mathcal{P}$ y se considera el estimador ***media muestral***, definido como:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Nótese que $E\left[X_i\right] = \mu$ y $\text{var}\left(X_i\right) = \sigma^2$. Veamos que $\hat{\mu}_n$ es centrado para μ. En efecto:

$$E\left[\hat{\mu}_n\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n} n\mu = \mu$$

La varianza del estimador es:

$$\mathrm{var}\left(\hat{\mu}_n\right) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}\left(X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

De lo anterior resulta que el error estándar del estimador es: $\sigma/\sqrt{n}$.

Supóngase ahora que se desea estimar la esperanza de una distribución de probabilidad μ con una cota de error B y una confianza $1-\alpha$. El tamaño muestral se obtendrá entonces como solución de la ecuación:

$$\Pr\left(\mu - B \leq \hat{\mu}_n \leq \mu + B\right) = 1 - \alpha$$

Para resolver la ecuación anterior hacemos uso de la siguiente aproximación basada en el teorema central del límite:

$$\sqrt{n} \cdot \frac{\hat{\mu}_n - \mu}{\sigma} \approx N\left(0, 1\right)$$

Tipificando en la ecuación de precisión queda:

$$\Pr\left(-\frac{B}{\sigma}\sqrt{n} \leq \sqrt{n} \cdot \frac{\hat{\mu}_n - \mu}{\sigma} \leq \frac{B}{\sigma}\sqrt{n}\right) = 1 - \alpha$$

Ello supone que: $B\sqrt{n}/\sigma = z_{1-\alpha/2}$, lo que lleva finalmente a:

$$n = \frac{z_{1-\alpha/2}^2 \cdot \sigma^2}{B^2}$$

Nótese que para poder fijar el tamaño muestral n, se precisa hacer uso de alguna información que se pueda disponer del verdadero valor de σ^2.

Varianzas. Considérese nuevamente una magnitud aleatoria cuya distribución de probabilidad $\mathcal{P}$ tiene esperanza μ y varianza σ^2. Un estimador natural para σ^2 basado en una muestra aleatoria $X_1, \ldots, X_n$ de la correspondiente distribución de probabilidad sería:

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$$

Sin embargo, el parámetro μ no se conoce en la práctica, por lo cual procedería sustituirlo por la estimación $\hat{\mu}_n$ dada en la subsección anterior. En orden a que el estimador sea centrado, proponemos finalmente el estimador llamado varianza muestral, el cual se define por:

$$\hat{\sigma}_n^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\hat{\mu}_n\right)^2$$

Teorema *(de Fisher) Para el supuesto en el que la distribución de probabilidad fuese* $\mathcal{P}=N\left(\mu,\sigma\right)$, *se satisface:*

$$\left(n-1\right)\frac{\hat{\sigma}_n^2}{\sigma^2}\cong\chi^2\left(n-1\right)$$

Del teorema anterior se sigue que $\hat{\sigma}_n^2$ es un estimador centrado para σ^2. En efecto:

$$E\left[\left(n-1\right)\frac{\hat{\sigma}_n^2}{\sigma^2}\right]=n-1$$

Teniendo en cuenta las propiedades de la esperanza se sigue que $E\left[\hat{\sigma}_n^2\right]=\sigma^2$.

Para la varianza se tiene:

$$\mathrm{var}\left(\left(n-1\right)\frac{\hat{\sigma}_n^2}{\sigma^2}\right)=2\left(n-1\right)$$

Aplicando las propiedades de la varianza se llega finalmente a:

$$\mathrm{var}\left(\hat{\sigma}_n^2\right)=\frac{2\sigma^4}{n-1}$$

Riesgo relativo y odds-ratio. Considérense las poblaciones E y C y sean π_E y π_C las probabilidades de que los elementos de estas poblaciones posean un atributo D; esto es: $\pi_E=\Pr\left(D\mid E\right)$ y $\pi_C=\Pr\left(D\mid C\right)$.

El *riesgo relativo* es entonces:

$$\rho = \frac{\pi_E}{\pi_C}$$

y la *odds-ratio*:

$$\omega = \frac{\pi_E\,(1-\pi_C)}{\pi_C\,(1-\pi_E)}$$

Para la estimación de π_E se observa una muestra aleatoria $X_{E,1},...,X_{E,n_E}$ de la distribución de probabilidad $b\,(1,\pi_E)$ y se estima π_E por la proporción muestral $\hat{\pi}_E = (1/n_E)\sum_{j=1}^{n_E} X_{E,j}$. Análogamente, para la estimación de π_C se observa una muestra aleatoria $X_{C,1},...,X_{C,n_C}$ de la distribución de probabilidad $b\,(1,\pi_C)$ y se estima π_C por la proporción muestral $\hat{\pi}_C = (1/n_C)\sum_{j=1}^{n_C} X_{C,j}$. Las propiedades de los estimadores $\hat{\pi}_E$ y $\hat{\pi}_C$ se dan en 3.1. Finalmente, se consideran los siguiente estimadores:

Para el *riesgo relativo* ρ

$$\hat{\rho} = \frac{\hat{\pi}_E}{\hat{\pi}_C}$$

y para la *odds-ratio* ω:

$$\hat{\omega} = \frac{\hat{\pi}_E\,(1-\hat{\pi}_C)}{\hat{\pi}_C\,(1-\hat{\pi}_E)}$$

Los cálculos de las esperanzas y varianzas de ambos estimadores resultan más complicados que en los casos anteriores. Por tal motivo consideramos inicialmente el estadístico $\ln\,\hat{\rho}$ como estimador del parámetro $\ln\,\rho$ mediante la siguiente aproximación :

$$\ln\,\hat{\rho} \approx \ln\,\rho + \frac{\hat{\pi}_E}{\pi_E} - \frac{\hat{\pi}_C}{\pi_C}$$

Veamos que el estimador $\ln\hat{\rho}$ es centrado para $\ln\rho$.

En efecto:

$$E\left[\ln\hat{\rho}\right] \approx \ln\rho + \frac{1}{\pi_E}E\left[\hat{\pi}_E\right] - \frac{1}{\pi_C}E\left[\hat{\pi}_C\right] = \ln\rho$$

Análogamente, la varianza de $\ln\hat{\rho}$ es:

$$\text{var}\left(\ln\,\hat{\rho}\right) \approx \frac{1}{\pi_E^2}\text{var}\left(\hat{\pi}_E\right) + \left(-1\right)^2 \frac{1}{\pi_C^2}\text{var}\left(\hat{\pi}_C\right) =$$

$$\frac{1}{\pi_E^2}\frac{\pi_E\left(1-\pi_E\right)}{n_E} + \frac{1}{\pi_C^2}\frac{\pi_C\left(1-\pi_C\right)}{n_C\pi_C} = \frac{1-\pi_E}{n_E\pi_E} + \frac{1-\pi_C}{n_C\pi_C}$$

Para la odds-ratio ω consideramos la aproximación:

$$\ln\,\hat{\omega} \approx \ln\,\omega + \frac{\hat{\pi}_E - \pi_E}{\pi_E\left(1-\pi_E\right)} - \frac{\hat{\pi}_C - \pi_C}{\pi_C\left(1-\pi_C\right)}$$

De aquí puede deducirse también que $\ln\,\hat{\omega}$ es centrado para $\log\,\omega$ y que la varianza tiene la forma:

$$\text{var}\left(\ln\hat{\omega}\right) \approx \frac{1}{n_E\pi_E\left(1-\pi_E\right)} + \frac{1}{n_C\pi_C\left(1-\pi_C\right)}$$

4. Intervalos de confianza

Considérese un parámetro desconocido θ y un conjunto de datos $\mathcal{X}$ conteniendo información sobre θ. En este escenario, un ***intervalo de confianza al nivel*** $1-\alpha$ para θ, es un intervalo cuyos extremos son estadísticos $\theta_L = \theta_L\left(\mathcal{X}\right)$ y $\theta_U = \theta_U\left(\mathcal{X}\right)$ tales que:

$$\Pr\left(\theta_L \leq \theta \leq \theta_U\right) = 1-\alpha$$

Nótese que θ, aunque desconocido, es un valor fijo, pero los extremos al depender de los datos son de naturaleza aleatoria. Ello significa que la cobertura de θ por el intervalo $[\theta_L\,;\,\theta_U]$ es incierta. Que el intervalo sea al nivel $1-\alpha$ significa que la probabilidad de cobertura es $1-\alpha$.

Sea μ el valor esperado de una distribución de probabilidad $N\left(\mu,1\right)$. En orden a determinar un intervalo de confianza al 95% para μ, se selecciona una muestra aleatoria $X_1,...,X_n$. Proponemos entonces el siguiente intervalo de confianza para μ :

$$\left[\hat{\mu}_n - \frac{1.96}{\sqrt{n}}\,;\,\hat{\mu}_n + \frac{1.96}{\sqrt{n}}\right]$$

Aquí, $\hat{\mu}_n = (1/n)\sum_{i=1}^{n} X_i$. El hecho de que este sea un intervalo de confianza al 95% significa que tiene un probabilidad del 95% de cubrir al verdadero valor de μ. Ello supone que se tomásemos un número muy elevado de muestras $X_1, ..., X_n$ y de cada una de ellas obtuviésemos el correspondiente intervalo de confianza, aproximadamente el 95% de las veces cubriría el verdadero valor de μ.

Proposición 2. Considérese un estimador centrado $\hat{\theta}$ para un parámetro θ tal que:

$$\frac{\hat{\theta} - \theta}{\text{sd}\left(\hat{\theta}\right)} \approx N(0,1)$$

En tales condiciones, un intervalo de confianza al nivel $1-\alpha$ para θ tiene la forma:

$$\left[\hat{\theta} - z_{1-\alpha/2} \cdot \text{sd}\left(\hat{\theta}\right) \ ; \ \hat{\theta} + z_{1-\alpha/2} \cdot \text{sd}\left(\hat{\theta}\right)\right]$$

teniendo en cuenta que el estadístico $\left(\hat{\theta} - \theta\right)/\text{sd}\left(\hat{\theta}\right) \approx N(0,1)$, podemos entonces escribir de acuerdo con la proposición 2:

$$\Pr\left(-z_{1-\alpha/2} \leq \frac{\hat{\theta} - \theta}{\text{sd}\left(\hat{\theta}\right)} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

Si la inecuación se multiplica por (-1) cambia de sentido la desigualdad y queda:

$$\Pr\left(-z_{1-\alpha/2} \leq \frac{\theta - \hat{\theta}}{\text{sd}\left(\hat{\theta}\right)} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

Finalmente, despejando θ queda:

$$\Pr\left(\hat{\theta} - z_{1-\alpha/2} \cdot \text{sd}\left(\hat{\theta}\right) \leq \theta \leq \hat{\theta} + z_{1-\alpha/2} \cdot \text{sd}\left(\hat{\theta}\right)\right) = 1 - \alpha$$

Lo anterior completa la demostración.

Intervalo de confianza para una probabilidad. Aplicando la propiedad dada en esta proposición 2, podemos obtener intervalos de confianza para los parámetros elementales estudiados en la sección anterior. Así por ejemplo, para el parámetro π (probabilidad) analizado en la sección 3.1 vimos que el estimador $\hat{\pi}_n$ estandarizado tenía

aproximadamente, en virtud del teorema central del límite, una distribución normal estándar. De esta forma un intervalo de confianza al nivel $1-\alpha$ para π es:

$$\left[\hat{\pi}_n - z_{1-\alpha/2} \cdot \sqrt{\frac{\pi\left(1-\pi\right)}{n}}\ ; \hat{\pi}_n + z_{1-\alpha/2} \cdot \sqrt{\frac{\pi\left(1-\pi\right)}{n}}\right]$$

Este intervalo de confianza en la práctica no puede obtenerse dado que el parámetro π es desconocido. Este problema puede obviarse sustituyendo π por su estimación $\hat{\pi}_n$ (método *plug-in*) y entonces se obtiene el intervalo aproximado:

$$\left[\hat{\pi}_n - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\pi}_n\left(1-\hat{\pi}_n\right)}{n}}\ ; \hat{\pi}_n + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\pi}_n\left(1-\hat{\pi}_n\right)}{n}}\right]$$

Intervalo de confianza para la esperanza. Considérese ahora una muestra aleatoria $X_1, ..., X_n$ de una distribución de probabilidad $N\left(\mu, \sigma\right)$. En orden a obtener un intervalo de confianza para μ, cosideramos inicialmente el estadístico $\sqrt{n}\left(\hat{\mu}_n - \mu\right)/\sigma$ el cual tiene una distribución de probabilidad normal estándar. Sin embargo, el intervalo de confianza deducido de este estadístico depende del parámetro desconocido σ. Para obviar este problema, consideramos el estadístico alternativo que resulta de sustituir σ por su estimador $\hat{\sigma}_n$. El estadístico que resulta tiene entonces una distribución de probabilidad t *de Student* con $n-1$ grados de libertad; esto es:

$$\sqrt{n} \cdot \frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n} \cong t\left(n-1\right)$$

De esta forma, puede deducirse fácilmente que un intervalo de confianza al nivel $1-\alpha$ para μ es:

$$\left[\hat{\mu}_n - t_{1-\alpha/2}\left(n-1\right) \cdot \frac{\hat{\sigma}_n}{\sqrt{n}}\ ; \hat{\mu}_n + t_{1-\alpha/2}\left(n-1\right) \cdot \frac{\hat{\sigma}_n}{\sqrt{n}}\right]$$

siendo $t_{1-\alpha/2}\left(n-1\right)$ el cuantil $1-\alpha/2$ de la distribución de probabilidad t *de Student* con $n-1$ grados de libertad.

Intervalo de confianza para el riesgo relativo. En 3.4 se dio (sin demostración) la aproximación:

$$\ln \hat{\rho} \approx \ln \rho + \frac{\hat{\pi}_E}{\pi_E} - \frac{\hat{\pi}_C}{\pi_C}$$

Nótese que $\hat{\pi}_E = (1/n_E) \sum_{i=1}^{n_E} X_{E,i}$ y $\hat{\pi}_C = (1/n_C) \sum_{i=1}^{n_C} X_{C,i}$ y de esta forma, $\ln \hat{\rho} - \ln \rho$ es una suma de variables aleatorias independientes. Dado que además $E[\ln \hat{\rho}] \approx \ln \rho$, el teorema central del límite justifica la aproximación:

$$\frac{\ln \hat{\rho} - \ln \rho}{\text{sd}(\ln \hat{\rho})} \approx N(0,1)$$

De esta forma, un intervalo de confianza al nivel $1-\alpha$ para $\ln \rho$ es:

$$\left[\ln \hat{\rho} - z_{1-\alpha/2} \cdot \text{sd}(\ln \hat{\rho}) \ ; \ln \hat{\rho} + z_{1-\alpha/2} \cdot \text{sd}(\ln \hat{\rho})\right]$$

donde según se vio, $\text{sd}(\ln \hat{\rho}) \approx \sqrt{(1-\pi_E)/(n_E \pi_E) + (1-\pi_C)/(n_C \pi_C)}$. Dado que los parámetros π_E y π_C son desconocidos, aproximaremos $\text{sd}(\ln \hat{\rho})$ mediante la sustitución (plug-in) de estos parámetros por sus estimadores $\hat{\pi}_E$ y $\hat{\pi}_C$.

Finalmente, el intervalo de confianza al nivel $1-\alpha$ para ρ es.

$$\left[\hat{\rho} \cdot \exp\left(-z_{1-\alpha/2} \cdot \text{sd}(\ln \hat{\rho})\right) \ ; \ \hat{\rho} \cdot \exp\left(z_{1-\alpha/2} \cdot \text{sd}(\ln \hat{\rho})\right)\right]$$

Téngase en cuenta que en la expresión del error estándar de $\ln \hat{\rho}$ aparecen los parámetros desconocidos π_E y π_C lo que supone que deben ser sustituidos por sus estimaciones $\hat{\pi}_E$ y $\hat{\pi}_C$.

Ejercicios

1. Sean μ y σ el valor esperado y la desviación estándar respectivamente de la HDL (mg/dL) en la población de Telde no diabética de más de 30 años. En el estudio de Telde se incluyeron 902 sujetos no diabéticos y se obtuvieron las siguientes estimaciones: $\hat{\mu}_{902} = 54.9$ y $\hat{\sigma}_{902} = 12.3$. Hallar un intervalo de confianza al 95% para μ.

2. Considérese el mismo ejercicio anterior pero para la población diabética de Telde con más de 30 años. El número de diabéticos incluido fue de 128 y las estimaciones obtenidas para μ y σ fueron respectivamente $\widehat{\mu}_{128} = 50.0$ y $\widehat{\sigma}_{128} = 12.7$. Hallar un intervalo de confianza al 95% para μ. ¿Qué conclusión se obtiene comparando los resultados de éste y el anterior ejercicio?

3. A partir de una muestra aleatoria $X_1, ..., X_n$ de una distribución de probabilidad $N(\mu, \sigma)$, con μ y σ desconocidos, obtener un intervalo de confianza para σ^2. [Hacer uso del teorema de Fisher: $(n-1)\,\sigma_n^2/\sigma^2 \cong \widehat{\chi}^2\,(n-1)$]

4. Sean $X_{E,1}, ..., X_{E,n_E}$ y $X_{C,1}, ..., X_{C,n_C}$ muestras aleatorias de las distribuciones de probabilidad $N(\mu_E, \sigma_E)$ y $N(\mu_C, \sigma_C)$ respectivamente. Obtener un intervalo de confianza al nivel $1-\alpha$ para el parámetro $\delta = \mu_E - \mu_C$.

5. Kelly *et al* (2020) llevaron a efecto un ensayo clínico aleatorizado y a doble ciego, que consistió en un período de tratamiento de 56 semanas y un período de seguimiento de 26 semanas. Se inscribieron adolescentes (de 12 a 18 años de edad) con obesidad y una mala respuesta a la terapia de estilo de vida solamente. Los participantes fueron asignados aleatoriamente (1:1) a recibir liraglutida (3,0 mg) o placebo por vía subcutánea una vez al día, además de la terapia de estilo de vida. El criterio de valoración primario fue el cambio en la puntuación del índice de masa corporal en la semana 56 con respecto al valor inicial. Los resultados del ensayo para el cambio del IMC (Kg/m^2) se resumen en la siguiente tabla (los datos que se muestran son $medias \pm Desv.estándar$).

	Liraglutida ($n = 125$)	Placebo ($n = 126$)
Cambio en el IMC (Kg/m^2)	-0.23 ± 0.05	-0.00 ± 0.05

Obtener un intervalo de confianza al 95% para la diferencia entre las variaciones de los índices de masa corporal. ¿hay evidencias que la liraglutida es más eficaz en la reducción del IMC que el placebo?

6. El estudio 4*S* (*scandinavian simvastatin survival study*) es un ensayo clínico aleatorizado con dos grupos paralelos, diseñado para evaluar el efecto de la simvastatina frente a placebo en la reducción de la morbi-mortalidad cardiovascular en pacientes que habían sufrido previamente un infarto ó angina de pecho y que tenían niveles de colesterol total superiores a 212 mg/dL. Tras un seguimiento

de 5,4 años en mediana, se observó que de los 2221 pacientes del grupo simvastatina (E) murieron por causas cardiovasculares 111, mientras que en los 2223 del grupo placebo (C) lo hicieron 189. Si representamos por π_E y π_C las probabilidades de muerte cardiovascular en los brazos E y C, hallar un intervalo de confianza al 95% para el riesgo relativo $\rho = \pi_E/\pi_C$.

7. El gen UCP2 regula la secreción de insulina y juega un importante papel en la relación entre obesidad y diabetes mellitus de tipo 2 (DM2). Bulotta *et al* (2005) analizaron el polimorfismo - 866G/A correspondiente al referido gen con la finalidad de evaluar su asociación con la DM2. De esta forma, los genotipos asociados los representamos por GG, GA y AA. Dado que el alelo G es el de mayor prevalencia, entenderemos que el genotipo GG es el de referencia, considerándose por tanto que las variantes GA y AA son mutaciones del gen ($G \longrightarrow A$). Los autores incluyeron en el estudio 746 personas con DM2 y 327 controles. Los resultados del estudio se resumen en la siguiente tabla:

Tabla 2. Asociación DM2 - presencia del alelo A en el UCP2

	DM2 $n = 746$	Control $n = 327$
$G \longrightarrow A$	362	185

Estimar la odds-ratio que mide la asociación entre la mutación $G \longrightarrow A$ y la DM2.

Referencias

1. Russell, Bertrand, Joaquín Xirau, and Emilio Lledó Iñigo. Los problemas de la filosofía. Labor, 1973. Versión en inglés en el Proyecto Gutenberg

2. Boronat, M., Varillas VF, Saavedra P, Suárez V, Bosch E, Carrillo A, Nóvoa FJ. Diabetes mellitus and impaired glucose regulation in the Canary Islands (Spain): prevalence and associated factors in the adult population of Telde, Gran Canaria. Diabet Med. 2006 Feb;23(2):148-55. doi: 10.1111/j.1464-5491.2005.01739.x. PMID: 16433712.

3. Kelly AS, Auerbach P, Barrientos-Perez M, Gies I, Hale PM, Marcus C, Mastrandrea LD, Prabhu N, Arslanian S; NN8022-4180 Trial Investigators. A Randomized, Controlled Trial of Liraglutide for Adolescents with Obesity. N Engl

J Med. 2020 May 28;382(22):2117-2128. doi: 10.1056/NEJMoa1916038. Epub 2020 Mar 31. PMID: 32233338. Ver pdf

4. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet. 1994 Nov 19;344(8934):1383-9. PMID: 7968073.

5. Bulotta A, Ludovico O, Coco A, Di Paola R, Quattrone A, Carella M, Pellegrini F, Prudente S, Trischitta V. The common -866G/A polymorphism in the promoter region of the UCP-2 gene is associated with reduced risk of type 2 diabetes in Caucasians from Italy. J Clin Endocrinol Metab. 2005 Feb;90(2):1176-80. doi: 10.1210/jc.2004-1072. Epub 2004 Nov 23. PMID: 15562023. Ver pdf