Análisis Estadístico **Grado en Ingeniería Matemática, UAX**

Tema 5: Estimación puntual, por intervalo y contraste de hipótesis

Salvador Robles *curso 2020/2021*

Resumen: *Estudiamos la estimación puntual, la estimación por intervalo y el contraste de hipótesis.*

Índice

5.1. Estimación puntual

Como ya hemos dicho en temas anteriores, el objetivo de la inferencia estadística es el de inferir las características de la distribución poblacional a partir de los datos proporcionados por una o más muestras. Suponiendo que la relación funcional de la distribución poblacional es conocida, la cuestión es entonces la de estimar los parámetros de dicha distribución a partir de lo que llamaremos *estimadores*, es decir, estadísticos que sirven para realizar la estimación de los parámetros poblacionales que deseamos estimar. Por ejemplo, para estimar el valor de la media poblacional, μ, podemos utilizar el estimador *media muestral*, $\overline{X}$. Sin embargo, para un mismo parámetro poblacional se pueden definir diferentes estimadores, unos mejores y otros

[0] Estas notas constituyen los apuntes informales para el seguimiento de la asignatura Análisis Estadístico del grado de Ingeniería matemática. No pretenden (y no lo hacen) sustituir al material bibliográfico recomendado para la asignatura. Deben considerarse por tanto como una guía de estudio de la asignatura que hay que trabajar conjuntamente con la bibliografía y con las hojas de ejercicios propuestos. Además, son una primera versión de los mismos así que pueden contener pequeños errores tipográficos o de otro tipo, cuya corrección se irá incorporando en sucesivas versiones.

peores en el sentido de que tendrán (o no) unas ciertas propiedades que en principio parecen deseables para un estimador. Estas propiedades son las siguientes:

- Decimos que el estimador T del parámetro pobacional θ es *insesgado* (o centrado) si su esperanza matemática coincide con el parámetro poblacional, es decir,
$$E[T] = \theta. \tag{5.1}$$
- Dados dos estimadores, T_1 y T_2, del parámetro poblacional θ, se dice que T_1 es más eficiente que T_2 si su varianza es menor, es decir, si
$$\sigma_{T_1} < \sigma T_2. \tag{5.2}$$
- Por último, diremos que un estimador T del parámetro poblacional θ es consistente si en el límite $n \to \infty$ converge al parámetro poblacional y si su varianza se hace nula, es decir,
$$\lim_{n\to\infty} T = \theta \quad , \quad \lim_{n\to\infty} \sigma_T^2 = 0. \tag{5.3}$$

Lo ideal será, siempre que sea posible, trabajar con un estimador insesgado, de máxima eficiencia y por supuesto consistente. Por ejemplo, para estimar la media y varianza poblacionales de una distribución normal utilizaremos la media y la cuasivarianza muestrales, $\overline{X}$ y S^2, respectivamente, que en este caso reúnen las tres propiedades deseadas: son estimadores insesgados, de máxima eficiencia u óptimos y consistentes. Para estimar la probabilidad de 'éxito' en una distribución binomial utilizaremos como estimador el estadístico $\overline{P}$, que igualmente cumple los tres requisitos, y para estimar el parámetro poblacional λ de una distribución de Poisson, utilizaremos la media muestral
$$\overline{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{5.4}$$

Sin embargo, no siempre está claro qué estimador utilizar para la estimación de un cierto parámetro poblacional. Un método útil para encontrar estimadores puntuales es el método de máxima verosimilitud, que consiste en determinar el estadístico que hace máxima la probabilidad conjunta de todas las muestras posibles, que en general dependerá del parámetro que se pretende estimar. Por ejemplo, supongamos que la distribución poblacional viene dada por la función $f(x, \theta)$, donde x es el valor de la variable aleatoria X y θ es el valor del parámetro que se desea estimar. En tal caso, para una muestra de tamaño n, representada por las n variables aleatorias $X_1, X_2, \ldots, X_n$, todas ellas supuestas independientes y con distribución $f(x_i, \theta)$, la distribución conjunta resulta
$$L(X_1, X_2, \ldots, X_n; \theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta). \tag{5.5}$$
A la función L se le llama *función de verosimilitud*, y, para una muestra concreta dada $(x_1, x_2, \ldots, x_n)$, solo depende del parámetro θ. El método de máxima verosimilitud consiste en determinar la combinación de los valores de la muestra, es deicr, el estadístico, que haga máximo el valor de la función de verosimilitud. Esto determinará el valor del parámetro θ que hace que más probable el valor de la muestra $(x_1, x_2, \ldots, x_n)$, genérico pero fijo.

Para los estimadores que hemos visto antes, se puede fácilmente demostrar que son los estimadores de máxima verosimilitud de los parámetros poblacionales que estiman. Pero además, el método de máxima erosimilitud nos proporciona una forma de obtener buenos estimadores puntuales en otros casos en los que la determinación del estadístico a utilizar no resulta tan obvia.

Ejemplo 5.1 *Ejemplo resuelto III-7 en Ref. [1]*

Ejemplo 5.2 *Determinar el estadístico de máxima verosimilitud para el parámetro θ si la distribución de la variable aleatoria X es*
$$f(x, \theta) = \frac{x}{\theta^2} e^{-x/\theta}$$

5.2. Intervalos de confianza

5.2.1. Introducción

La estimación puntual tiene al menos dos problemas. Por un lado, dado que los estimadores son variables aleatorias, cualquier estimación puntual que hagamos a partir de un estadístico para determinar el valor de un parámetro poblacional no nos proporcionará, por regla general, su valor exacto. Además, tampoco nos proporciona un criterio sobre la fiabilidad del resultado obtenido. Por estos motivos, es en general más interesante la estimación de los parámetros poblacionales por *intervalos de confianza*, que son intervalos de valores en los que existe una cierta probabilidad de que esté incluido el parámetro poblacional buscado. Más concretamente, la idea es buscar el intervalo de valores, $[L_1, L_2]$, tal que la probabilidad de que el parámetro θ esté en dicho intervalo sea $1-\alpha$, valor que llamaremos *nivel de confianza*. Es decir, queremos encontrar los valores L_1 y L_2 para los que se cumple

$$P(L_1 < \theta < L_2) = 1 - \alpha. \tag{5.6}$$

Al intervalo $[L_1, L_2]$ se le denomina intervalo de confianza del $(1-\alpha)100\,\%$. La determinación de los límites del intervalo, L_1 y L_2, dependerá de la distribución de probabilidad del estadístico que utilicemos para estimar el parámetro poblacional θ. Por eso, veremos como calcularlo en los casos concretos que vamos a estudiar este curso, que básicamente son dos: la estimación de la media poblacional, μ, y la estimación de la varianza poblacional, σ. En ambos casos, supondremos que la población sigue una distribución normal $N(\mu, \sigma)$.

5.2.2. Intervalo de confianza para la media

- **Si la varianza poblacional, σ^2, es conocida:**

  Para estimar el valor de la media poblacional, μ, se usa el estimador media muestral, $\overline{X}$. Empezaremos suponiendo que la desviación típica poblacional, σ, es conocida. En tal caso, sabemos que el estadístico,

  $$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}, \tag{5.7}$$

  sigue una distribución normal estándar, $N(0,1)$, a partir de la que podemos obtener el valor $z_{\alpha/2}$ para el que se cumple,

  $$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha. \tag{5.8}$$

  El hecho de llamar $z_{\alpha/2}$ al valor de la abscisa es porque la condición (5.8) es equivalente a la condición,

  $$P(Z > z_{\alpha/2}) = \frac{\alpha}{2}, \tag{5.9}$$

  por eso es habitual en inferencia estadística utilizar la tabla de la normal en la que se dan los valores de la cola derecha de la distribución. Una vez encontrado el valor $z_{\alpha/2}$ que cumple (5.9) (que se obtiene directamente de la tabla a partir del valor del nivel de significación $1-\alpha$, o más concretamente a partir del valor de $\alpha/2$), entonces, como

  $$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), \tag{5.10}$$

  tendremos que el intervalo para el que la probabilidad de que contenga a la media poblacional es $1-\alpha$ es

  $$I = [L_1, L_2] = \left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{5.11}$$

Hay que destacar varias cuestiones. Primera, distintas muestras darán distintos valores de la media muestral, $\overline{X}$, y por tanto distintos valores del intervalo $[L_1, L_2]$. Como decíamos anteriormente, que la probabilidad de que el parámetro poblacional esté en el intervalo dado por (5.11) sea $(1-\alpha)$ hay que interpretarlo, en términos frecuentistas, como que $(1-\alpha)$ es precisamente la proporción (o frecuencia relativa) de muestras para las que el intervalo dado en (5.11) contiene el valor de la media poblacional, μ. Por otro lado, para cualquier valor de la media muestral, la longitud del intervalo es la misma y viene dada por,

$$\Delta L = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{5.12}$$

Por tanto, fijados α y σ, la longitud del intervalo es menor para valores mayores del tamaño muestral. Es decir, con muestras de mayor tamaño la precisión de la estimación es mayor.

- **Si la varianza poblacional, σ^2, es desconocida pero el tamaño muestral es grande ($n > 30$)**

  Por regla general, la desviación típica poblacional σ es desconocida y por tanto no podríamos aplicar (5.11) para obtener el intervalo de confianza. Sin embargo, si el tamaño de la muestra es muy grande (en la práctica, $n > 30$), la desviación típica muestral (obtenida a partir de la cuasivarianza muestral!), $S = \sqrt{S^2}$, es un buen estimador de la varianza poblacional. En tal caso, podemos utilizar (5.11) sustituyendo σ por S, es decir,

$$I = \left[\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]. \tag{5.13}$$

  Ejemplo 5.3 *Ejemplo resuelto III-8 en Ref. [1]*

- **Si la varianza poblacional, σ^2, es desconocida y el tamaño poblacional es pequeño**

  Si el tamaño de la muestra no es muy grande la aproximación anterior puede no ser válida. En ese caso, podemos utilizar el hecho de que el estadístico,

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}, \tag{5.14}$$

  sigue una distribución t de Student con $n-1$ grados de libertad, la cual es también simétrica respecto al origen, como vimos en los temas anteriores. Por tanto, siguiendo un razonamiento similar al realizado en los apartados anteriores llegamos a que,

$$P\left(-t_{\alpha/2,n-1} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1-\alpha, \tag{5.15}$$

  o lo que es lo mismo,

$$P\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha. \tag{5.16}$$

  Por tanto, el intervalo de confianza de nivel $(1-\alpha)$ será

$$I = \left[\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]. \tag{5.17}$$

Ejemplo 5.4 *Ejemplo resuelto III-9 en Ref. [1]*

Ejemplo 5.5 *Ejemplo resuelto p. 49 (1^a parte) en Ref. [2]*

5.2.2.1. Intervalo de confianza para una proporción y para el parámetro λ de una distribución de Poisson.

Como vimos anteriormente, el estimador $\overline{P}$ para estimar la proporción de éxitos o probabilidad p en un ensayo de Bernoulli es un caso particular del estimador media muestral. Sabemos, además, que cuando la muestra es grande, la distribución muestral de $\overline{P}$ se aproxima a una distribución normal de media, $\mu_{\overline{P}} = p$, y desviación típica, $\sigma_{\overline{P}} = \sqrt{p(1-p)/n}$. Por tanto, es un caso concreto de lo visto anteriormente y podemos escribir,

$$P\left(\overline{P} - z_{\alpha/2}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}} < p < \overline{P} + z_{\alpha/2}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right) = 1-\alpha, \tag{5.18}$$

donde hemos sustituido p por $\overline{P}$ en el valor de la varianza, que es una buena aproximación si desconocemos la desviación típica muestral y el tamaño muestral es grande[1]. De este modo, el intervalo de confianza de nivel $(1-\alpha)$ resulta

$$I = \left[\overline{P} \pm z_{\alpha/2}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right]. \tag{5.19}$$

En el caso de una distribución de Poisson, para la estimación del parámetro λ utilizamos el estadístico $\overline{\lambda}$, que es un caso particular de la media muestral. Además, si la muestra es grande y $\overline{\lambda} > 5$, podemos suponer que $\overline{\lambda}$ sigue una distribución normal, $N(\overline{\lambda}, \sqrt{\overline{\lambda}/n})$. Por tanto, el intervalo de confianza de nivel $(1-\alpha)$ para el parámetro λ resulta

$$I = \left[\overline{\lambda} \pm z_{\alpha/2}\sqrt{\frac{\overline{\lambda}}{n}}\right]. \tag{5.20}$$

Ejemplo 5.6 *Ejemplo resuelto III-10 en Ref. [1]*

5.2.3. Intervalo de confianza para la varianza

Para calcular el intervalo de confianza para la varianza seguiremos un razonamiento análogo al realizado para la estimación por intervalo de la media. Sin embargo, ahora haremos uso del hecho de que el estadístico,

$$\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2}, \tag{5.21}$$

sigue una distribución χ^2 con $n-1$ grados de libertad. También hay otra diferencia, la distribución χ^2 no es simétrica respecto al origen y por tanto no podemos utilizar los valores de las abscisas, $\chi^2_{\alpha/2,n-1}$ y $-\chi^2_{\alpha/2,n-1}$, puesto que este último no existe (la variable χ^2 es siempre positiva). Seguiremos utilizando los valores de las abscisas que dejan a su izquierda y a su derecha una probabilidad $\alpha/2$, con lo que el nivel de confianza (la región central) resulta $1-\alpha$, solo que ahora esos valores vendrán dados por $\chi_{1-\alpha/2,n-1}$, para el límite inferior, y por $\chi_{\alpha/2,n-1}$ para el valor por la derecha, es decir, ahora tendremos

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha. \tag{5.22}$$

[1] En general, si conocemos la desviación típica poblacional, σ, es mejor utilizarla frente a las estimaciones. Si la desconocemos y el tamaño muestral es grande, podemos sustituirla por la desviación típica muestral, y si el tamaño es pequeño utilizar la distribución t de Student en vez de la normal. Si no conocemos la desviación típica muestral y el tamaño muestral es grande podemos utilizar la aproximación hecha en (5.18).

Pero operando con los términos de la desigualdad, resulta que la condición (5.22) es equivalente a

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} > \sigma^2 > \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right) = 1-\alpha. \tag{5.23}$$

Es decir, que el intervalo de confianza de nivel $(1-\alpha)$ para la varianza de una distribución normal con cuasivarianza muestral dada por S^2 es

$$I = \left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right], \tag{5.24}$$

y por tanto para la desviación típica tenemos

$$I = \left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}\right]. \tag{5.25}$$

Ejemplo 5.7 *Ejemplo resuelto p. 49 (2^a parte) en Ref. [2]*

5.2.4. Determinación del tamaño de la muestra

Ya hemos visto en las secciones anteriores que el intervalo de confianza para una cierta estimación depende del tamaño de la muestra, n. Por regla general, fijados el resto de los parámetros de la estimación, el intervalo de confianza es más estrecho para valores mayores de n. Que la longitud del intervalo de confianza sea pequeña es interesante porque de alguna forma su tamaño determina el error que estaríamos cometiendo al elegir uno u otro valor dentro de dicho intervalo, y además, dicho valor me acota el rango de posibles valores. De hecho, en muchos casos lo que queremos es precisamente determinar el tamaño que tiene que tener una muestra para que ese error esté por debajo de una cierta cota. Por ejemplo, supongamos que quiero determinar la media poblacional de una distribución normal, μ, a partir del valor de la media muestral, $\overline{X}$. Queremos ahora encontrar el valor del tamaño muestral, n, para el cual el valor de la media poblacional μ está, con un nivel de confianza $(1-\alpha)$, entre los valores $\overline{X}-\epsilon$ y $\overline{X}+\epsilon$, es decir,

$$P(\overline{X}-\epsilon < \mu < \overline{X}+\epsilon) = 1-\alpha, \tag{5.26}$$

en donde sabemos que, en el caso de que la desviación típica poblacional sea conocida,

$$\epsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{5.27}$$

Si desconocemos la desviación típica poblacional, σ, podemos estimarla a partir de la cuasivarianza muestral, S^2, si el tamaño de la muestra es lo suficientemente grande ($n > 30$), o utilizar el valor $t_{\alpha/2}$ de la distribución t de Student en vez del valor $z_{\alpha/2}$ de la distribución normal. Si ahora queremos que el error (5.27) esté por debajo de un cierto valor ϵ_0, necesitaremos que el tamaño de la muestra n cumpla

$$n > n_0 \equiv z^2_{\alpha/2}\frac{\sigma^2}{\epsilon_0}. \tag{5.28}$$

Ejemplo 5.8 *Ejemplo resuelto III-17 en Ref. [1]*

5.3. Contraste de hipótesis

5.3.1. Conceptos previos

Un contraste de hipótesis es un problema en el que lo que se plantea es decidir, a la vista de los datos de una muestra, sobre dos hipótesis: la hipótesis de partida o hipótesis nula, H_0, y la hipótesis alternativa o H_1. Es importante destacar que no se trata de determinar si la hipótesis nula (o la alternativa) son más o menos coherentes o parecen más o menos plausibles. La cuestión es mucho más aséptica, se trata de decidir si en términos de probabilidad los datos de la muestra son suficientemente significativos como para rechazar la hipótesis de partida. Este es un planteamiento conservador en el sentido de que por defecto vamos a aceptar que la hipótesis de partida es cierta, y solo en el caso en el que los datos de la muestra evidencien una clara contradicción con la hipótesis de partida es cuando rechazaremos la hipótesis. Es decir, en caso de duda, nos quedamos con la hipótesis de partida.

Para tomar tal decisión necesitamos evaluar la probabilidad de que un cierto estadístico, que consideraremos apropiado para cada caso, tome un rango de valores que vamos a denominar *región de aceptación*. Y si el valor concreto del estadístico para una muestra dada está en la región de aceptación aceptaremos como cierta la hipótesis de partida, y si por el contrario cae fuera de la región de aceptación, que llamaremos *región crítica*, entonces, rechazaremos la hipótesis nula para aceptar como cierta la hipótesis alternativa, H_1. De esta forma, resolver un contraste de hipótesis se reduce a:

1. Elegir el estadístico apropiado en cada caso.
2. Determinar la región de aceptación como la región en la que, aceptando como cierta la hipótesis nula, la probabilidad de encontrar al estadístico es alta, $P = 1 - \alpha$. Es decir, la región de aceptación es la región donde esperaríamos encontrar el valor del estadístico con un nivel de confianza del $(1-\alpha)100\,\%$.
3. Calcular el valor concreto del estadístico para la muestra dada, y ver si cae en la región de aceptación o fuera de ella, es decir, en la región crítica. En el primer caso, mantenemos la hipótesis nula como cierta. En el segundo caso, la rechazaremos para aceptar como cierta la hipótesis alternativa.

Sin embargo, por muy probable que sea que el valor del estadístico caiga en la región de aceptación siempre existe la posibilidad de cometer el error de aceptar la hipótesis nula siendo realmente falsa. Tenemos que darnos cuenta de que hay dos motivos por los que dicho valor puede caer fuera de la región de aceptación: uno puede ser porque efectivamente la hipótesis nula es falsa y por tanto estaríamos acertando en nuestra decisión (es decir, estaríamos rechazando H_0 cuando realmente es falsa), pero también puede ser (en principio con menos probabilidad) que caiga fuera de la región de aceptación por puro azar. De hecho, sabemos que siendo H_0 verdadera y teniendo la región de aceptación un nivel de confianza $(1 - \alpha)$, en el $\alpha 100\,\%$ de las muestras el valor del estadístico caer fuera de la región de aceptación, es decir, en la región crítica. En tal caso, siguiendo el procedimiento que hemos descrito acabaríamos rechazando la hipótesis nula H_0 cuando en realidad es cierta. Este es el error que llamamos de Tipo I, y al valor de α, que da la probabilidad de cometer este error, se le denomina *nivel de significación*.

Podemos también cometer otro tipo de error, y es aceptar H_0 cuando en realidad es falsa. Este es el error que llamamos de Tipo II, y la probabilidad de cometer este tipo de error viene dada por una variable, que llamaremos β. Por regla general al disminuir un tipo de error se aumenta el otro ya que disminuir por ejemplo el valor del nivel de significación α (que es la probabilidad de cometer el error de Tipo I) implica aumentar la región de aceptación, pero cuanto más grande es esta región mayor será también la posibilidad de que estemos aceptando H_0 incluso en los casos en los que esta hipótesis sea falsa (que es el error de Tipo II, dado por β). En nuestro caso trabajaremos solo con el nivel de significación, α, que para la gran mayoría de los casos vendrá dado por regla general por los valores, $\alpha = 0,05$ o $\alpha = 0,01$.

Por último, veremos en este apartado dos tipos de contrastes de hipótesis: el contraste bilateral y el contraste unilateral. En el primero, la región crítica se divide en dos partes, serán contrastes de hipótesis de la forma:

$$\begin{cases} H_0: & \mu = \mu_0 \\ \\ H_1: & \mu \neq \mu_0 \end{cases} \tag{5.29}$$

mientras que los contrastes unilaterales la región crítica está formada por un solo conjunto de valores. Serán contrastes del estilo

$$\begin{cases} H_0: & \mu \leq \mu_0 \\ \\ H_1: & \mu > \mu_0 \end{cases} \quad \text{o} \quad \begin{cases} H_0: & \mu \geq \mu_0 \\ \\ H_1: & \mu < \mu_0 \end{cases} \tag{5.30}$$

5.3.2. Contraste de la media de una población normal

- Varianza σ^2 conocida

  a) Contraste bilateral

  El test de hipótesis planteado es

$$\begin{cases} H_0: & \mu = \mu_0 \\ \\ H_1: & \mu \neq \mu_0 \end{cases} \tag{5.31}$$

  En este caso, si asumimos como cierta la hipótesis nula, H_0, que es la hipótesis de partida, es decir, si asumimos que la media de la distribución poblacional es, $\mu = \mu_0$, entonces, sabemos que el estadístico,

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}, \tag{5.32}$$

  sigue una distribución normal estándar, $N(0,1)$. En tal caso, tendríamos

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, \tag{5.33}$$

  y esta condición nos determina la región crítica, C, y la región de aceptación, A,

$$C = \{z : |z| > z_{\alpha/2}\} \quad , \quad A = \{z : |z| \leq z_{\alpha/2}\}. \tag{5.34}$$

  Es decir, aceptaremos la hipótesis nula si el valor de la muestra es tal que z está en la región de aceptación, A, y la rechazaremos si z cae en la región crítica, C. Es decir, aceptaremos la hipótesis nula si

$$\overline{x} \in \left[\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \equiv A \tag{5.35}$$

  y la rechazaremos (al nivel de significación α) si

$$\overline{x} \in \left(-\infty, \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \cup \left(\mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \infty\right) \equiv C \tag{5.36}$$

  Todos los datos en (5.35) y (5.36) son conocidos. Por tanto, lo que tenemos que hacer el calcular el valor concreto de la media muestral, $\overline{x}$, para la muestra dada y comprobar si cae en la región de aceptación o en la región crítica.

Ejemplo 5.9 *Ejemplo resuelto IV-5 en Ref. [1]*

b) Contraste unilateral

En este caso, el test de hipótesis planteado puede ser o bien

$$\begin{cases} H_0: & \mu \leq \mu_0 \\ H_1: & \mu > \mu_0 \end{cases} \tag{5.37}$$

o bien

$$\begin{cases} H_0: & \mu \geq \mu_0 \\ H_1: & \mu < \mu_0 \end{cases} \tag{5.38}$$

En el primer caso, la región de aceptación A y la región crítica C son,

$$A = \left(-\infty, \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right] \quad , \quad C = \left(\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right) \tag{5.39}$$

y, de forma análoga, para el segundo caso

$$C = \left(-\infty, \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}\right) \quad , \quad A = \left[\mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right) \tag{5.40}$$

Hay que darse cuenta que para calcular los intervalos (5.39) y (5.40) no utilizamos el valor $z_{\alpha/2}$ sino el valor de z_α, ya que al quedarnos con una parte entera de la distribución, la izquierda en el caso del test (5.37) o la derecha para el test (5.38), el valor de z que deja a su derecha o a su izquierda, respectivamente, una probabilidad α es z_α (el valor de $z_{\alpha/2}$ venía de dejar una cola izquierda y una cola derecha cada una con probabilidad $\frac{\alpha}{2}$, para sumar entre las dos colas una probabilidad total, α).

- Varianza σ^2 desconocida y $n > 30$

  Ya vimos al calcular los intervalos de confianza que si el tamaño de la muestra es relativamente grande, asumiendo en la práctica un valor $n > 30$, el valor de la cuasivarianza S^2 suponía una buena estimación de la varianza poblacional, σ^2. Por tanto, los intervalos de aceptación y crítico para la media de una distribución normal en la que desconocemos el valor de la desviación típica se puede calcular, si $n > 30$, con los mismos valores del apartado anterior sustituyendo σ por el valor de S que proporciona la muestra.

- Varianza σ^2 desconocida y $n \leq 30$

  En el caso de que el tamaño de la muestra no sea grande, $n \leq 30$, la estimación de σ por S puede no ser muy precisa. Pero sabemos del tema anterior que en tal caso el estadístico,

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}, \tag{5.41}$$

  sigue una distribución t de Student con $n-1$ grados de libertad, y por tanto, podemos realizar un análisis análogo al realizado en el primer apartado de esta sección para concluir que las regiones crítica y de aceptación serán las mismas que en aquel apartado pero sustituyendo el valor de $z_{\alpha/2}$ o z_α por los correspondientes valores de la distribución t de Student (que, recordemos, también es simétrica respecto al origen como la distribución normal), $t_{\alpha/2,n-1}$ y $t_{\alpha,n-1}$, respectivamente. Es decir, para el test bilateral tendremos que la región de aceptación A y la región crítica C vendrán dadas por los intervalos,

$$A = \left[\mu_0 - t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}, \mu_0 + t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}\right] \tag{5.42}$$

y

$$C = \left(-\infty, \mu_0 - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) \cup \left(\mu_0 + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \infty\right), \tag{5.43}$$

y para los test unilaterales,

$$A = \left(-\infty, \mu_0 + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right] \quad , \quad C = \left(\mu_0 + t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty\right), \tag{5.44}$$

para el test (5.37), y

$$C = \left(-\infty, \mu_0 - t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right) \quad , \quad A = \left[\mu_0 - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty\right), \tag{5.45}$$

para el test (5.38). Al igual que en las regiones dadas por (5.39) y (5.40), en las regiones (5.44) y (5.45) utilizamos el valor $t_{\alpha,n-1}$ y no el valor $t_{\alpha/2,n-1}$, por los mismos motivos que se expusieron entonces.

- Contraste de una proporción

  También vimos en el tema anterior que estimar el valor de la probabilidad p de un proceso de Bernoulli a partir de una muestra de tamaño n era un caso particular de la estimación de la media muestral, y utilizamos el estimador $\overline{P}$, que contabiliza la proporción de 'éxitos' del total n de pruebas del proceso de Bernoulli. Si suponemos que la muestra es lo suficientemente grande la distribución de este estimador puede aproximarse a una distribución normal centrada en el valor p y con una varianza, $p(1-p)/n$.

  Por tanto, las regiones de aceptación y crítica para el contraste bilateral,

$$\begin{cases} H_0: & p = p_0 \\ H_1: & p \neq p_0 \end{cases} \tag{5.46}$$

  serán

$$A = \left[p_0 - z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}, p_0 + z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right] \tag{5.47}$$

  y

$$C = \left(-\infty, p_0 - z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right) \cup \left(p_0 + z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}, \infty\right), \tag{5.48}$$

  respectivamente, donde, como ya hicimos en la Sec. 5.2.2.1, hemos estimado la varianza poblacional por, $\frac{\overline{p}(1-\overline{p})}{n}$. Y para los contrastes unilaterales

$$\begin{cases} H_0: & p \leq p_0 \\ H_1: & p > p_0 \end{cases} \tag{5.49}$$

  y,

$$\begin{cases} H_0: & p \geq p_0 \\ H_1: & p < p_0 \end{cases} \tag{5.50}$$

  tendremos,

$$A = \left(-\infty, p_0 + z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right] \quad , \quad C = \left(p_0 + z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}, \infty\right), \tag{5.51}$$

  y

$$C = \left(-\infty, p_0 - z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right) \quad , \quad A = \left[p_0 - z_{\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}, \infty\right), \tag{5.52}$$

  respectivamente.

Ejemplo 5.10 *Ejemplo resuelto IV-8 en Ref. [1]*

5.3.3. Contraste de la varianza de una población normal

Para estimar el valor de la varianza a partir de una muestra utilizábamos el estadístico,

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}, \tag{5.53}$$

cuya distribución de probabilidad sabemos que es la distribución χ^2 de $n-1$ grados de libertad. Por tanto, las regiones de aceptación y crítica de los test de hipótesis construidos para el valor de la varianza de una población normal dependerán de los valores de dicha distribución.

a) Contraste bilateral

El contraste de hipótesis que planteamos en este caso es el siguiente,

$$\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 \neq \sigma_0^2 \end{cases} \tag{5.54}$$

Si aceptamos la hipótesis nula, H_0, esperamos que los valores de χ^2 estén, con un nivel de confianza del $(1-\alpha)100\,\%$, en la región que se encuentra entre el valor $\chi^2_{1-\alpha/2,n-1}$ y el valor $\chi^2_{\alpha/2,n-1}$ de la distribución. Hay que recordar que, a diferencia de la distribución normal y la distribución t de Student, la distribución χ^2 no es simétrica respecto al origen, y de hecho su rango de valores es $[0,\infty)$. Por eso hemos cogido estos valores, que son los que dejan a su izquierda y a su derecha, respectivamente, un valor $\alpha/2$ de la probabilidad. Por tanto, con ese mismo nivel de confianza esperamos que el valor del estadístico S^2 se encuentre en el intervalo (región de aceptación)

$$A = \left[\chi^2_{1-\alpha/2,n-1}\frac{\sigma_0^2}{n-1}, \chi^2_{\alpha/2,n-1}\frac{\sigma_0^2}{n-1}\right], \tag{5.55}$$

y rechazaremos la hipótesis nula, H_0, a nivel de significación α si el valor de la cuasivarianza S^2 cae en la región crítica,

$$C = \left[0, \chi^2_{1-\alpha/2,n-1}\frac{\sigma_0^2}{n-1}\right) \cup \left(\chi^2_{\alpha/2,n-1}\frac{\sigma_0^2}{n-1}, \infty\right). \tag{5.56}$$

b) Contraste unilateral

Los test de hipótesis se plantean ahora son,

$$\begin{cases} H_0: & \sigma^2 \leq \sigma_0^2 \\ H_1: & \sigma^2 > \sigma_0^2 \end{cases} \tag{5.57}$$

y,

$$\begin{cases} H_0: & \sigma^2 \geq \sigma_0^2 \\ H_1: & \sigma^2 < \sigma_0^2 \end{cases} \tag{5.58}$$

Fijémonos en el primero de ellos. La región de aceptación será la unión de todos los posibles intervalos que pueda construir del estilo de (5.55) para todos los valores, $\sigma^2 \in [0, \sigma_0^2]$. Por tanto, las regiones de

aceptación y crítica resultan

$$A = \left[0, \chi^2_{\alpha,n-1}\frac{\sigma_0^2}{n-1}\right] \quad , \quad C = \left(\chi^2_{\alpha,n-1}\frac{\sigma_0^2}{n-1}, \infty\right) \tag{5.59}$$

y con un razonamiento análogo para el test (5.58) obtenemos,

$$C = \left[0, \chi^2_{1-\alpha,n-1}\frac{\sigma_0^2}{n-1}\right) \quad , \quad A = \left[\chi^2_{1-\alpha,n-1}\frac{\sigma_0^2}{n-1}, \infty\right). \tag{5.60}$$

Ejemplo 5.11 *Ejemplo resuelto IV-9 en Ref. [1]*

5.3.4. Prueba de la bondad del ajuste

Hasta ahora, en todas las estimaciones que hemos estudiado hemos supuesto una distribución de partida para la variable poblacional, de la que desconocíamos algún parámetro que era precisamente el que pretendíamos estimar o sobre el que formulábamos un test de hipótesis. Esto es lo que llamamos al inicio del tema *estimación paramétrica*, es decir, estimar los valores de los parámetros de una distribución dada a partir de los valores de una muestra que, se asume, siguen la distribución de partida. En la mayoría de los casos hemos además supuesto que la distribución poblacional era una distribución normal y la estimación se realizaba sobre la media, μ, o sobre la varianza, σ^2. La cuestión ahora es decidir si a la vista de los datos de una muestra concreta podemos concluir o no que la variable poblacional sigue una distribución dada. En líneas generales, lo que vamos a decidir es si la diferencia entre los datos de la muestra y los datos que se obtendrían de la distribución teórica son tan pequeños que podemos suponer que dichas diferencias son debidas al azar o, por el contrario, son demasiado grandes como para aceptar que los datos siguen la distribución propuesta.

Es decir, dada una distribución teórica hacemos la hipótesis (nula, es decir, H_0) de suponer que los datos de la muestra siguen dicha distribución. Entonces, supongamos que tenemos una muestra de tamaño n, y supongamos también que la variable poblacional puede tomar los k valores $X_1, X_2, \ldots, X_k$. En el caso de que la variable X sea continua (o si el número de valores es muy grande) los k valores anteriores pueden ser las marcas de clase de k intervalos en los que podemos dividir el conjunto de valores de X. En cualquier caso, los n valores de la muestra se distribuirán entre los k posibles valores $X_1, X_2, \ldots, X_k$, bien porque la variable aleatoria sea discreta y los valores de la muestra coincidan necesariamente con los X_i, para algún valor de i, o bien porque los valores de la muestra pertenezcan a alguno de los k intervalos en los que hemos dividido el dominio de la variable. Entonces, podemos definimos el siguiente estadístico,

$$D = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}, \tag{5.61}$$

donde, N_i es la frecuencia en la muestra del valor $X = x_i$, es decir, es el número de valores de la muestra que coinciden con el valor x_i (o dicho de otro modo, el número de 'x_i's que hay en la muestra), y p_i es el valor que la distribución teórica de probabilidad asigna al valor $X = x_i$. De esta forma, lo que esencialmente mide el estadístico (5.61) son las diferencias cuadráticas que hay entre el número de x_i's que hay en la muestra, N_i, y el número de x_i's que debería haber si las variables de la muestra siguiesen efectivamente la distribución teórica, en tal caso dado por, $np_i \equiv n_i$, ya que recordemos que n es el número total de datos en la muestra y p_i la frecuencia relativa teórica para x_i.

Por otro lado, aunque la demostración está fuera del alcance de este curso, se puede demostrar que el estadístico D sigue una distribución χ^2 con $k-1$ grados de libertad. La idea ahora es estudiar si los valores de D son los que esperaríamos con un nivel de confianza $1-\alpha$ en la distribución χ^2_{k-1}. En ese caso,

aceptaremos la hipótesis de que los datos de la muestra siguen la distribución teórica de partida o, en caso contrario, la rechazaremos. Es decir, lo que tenemos que hacer es calcular, para la muestra dada, el valor del estadítico,

$$D = \sum_{i=1}^{k} \frac{(N_i - n_i)^2}{n_i} = \sum_{i=1}^{k} \frac{N_i^2}{n_i} - n, \tag{5.62}$$

y ver si cae en la región de aceptación,

$$A = \left[0, \chi^2_{\alpha,k-1}\right], \tag{5.63}$$

que es la región en la que esperaríamos que estuviese el valor de D, con un nivel de confianza $(1-\alpha)100\,\%$, si los datos de la muestra siguiesen la distribución propuesta.

Para que este ajuste sea válido necesitamos que el tamaño de la muestra n sea suficientemente grande como para que la frecuencia absoluta de cada uno de los intervalos sea también significativa. Se suele exigir los valores, $n > 30$, y $n_i > 5$ para todos los intervalos. A veces, para que esta última condición se cumpla podemos reagrupar los valores de X_i en intervalos diferentes de manera que todos los intervalos incluyan el número requerido de datos. Por otro lado, si para calcular las frecuencias esperadas tenemos que usar estimaciones de los parámetros poblacionales a partir de los datos de la muestra, entonces, el número de grados de libertad de la distribución χ^2 hay que reducirlo a $k-m-1$, donde m es el número de parámetros estimados, es decir, hay que utilizar el valor de $\chi^2_{\alpha,k-m-1}$.

Ejemplo 5.12 *Ejemplo resuelto IV-14 en Ref. [1]*

Ejemplo 5.13 *En el estudio de cierta especie de árbol en una determinada región se mide el número de árboles de esta especie en 400 cuadrados de muestreo, tomados al azar en la región bajo estudio, obteniéndose los siguientes datos:*

Nº árboles	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Cuadrados de la muestra	*56*	*104*	*80*	*62*	*42*	*27*	*9*	*9*	*5*	*3*	*2*	*0*	*1*

Comprobar la hipótesis de que los datos obtenidos se adaptan a una distribución de Poisson.

Referencias

[1] J. Gorgas García, N. Cardiel López, and J. Zamorano Calvo. *Estadística básica para estudiantes de ciencias*. Editorial UCM, 2011.

[2] J. Olarrea Busto and M. Cordero Gracia. *Inferencia estadística. 20 problemas útiles*. García-Maroto Editores S. L., 2007.