UDLA
UNIVERSIDAD DE LAS AMÉRICAS

INSTITUTO DE MATEMÁTICA, FÍSICA Y ESTADÍSTICA

INFERENCIA ESTADISTICA

CLASE: ESTIMACIÓN PUNTUAL Y ESTIMACIÓN POR INTERVALO

Introducción

La teoría de Inferencia Estadística consiste en aquellos métodos con los cuales se pueden realizar inferencias (deducciones) o generalizaciones acerca de una población a partir de una muestra extraída de la misma, y tomar decisiones acerca de toda la población de estudio. La idea es estimar parámetros de la población por medio del cual las inferencias se basan en información que se obtiene de una muestra aleatoria seleccionada de una población.

La inferencia estadística puede dividirse en 2 áreas

- Estimación de parámetros
- Pruebas de Hipótesis

Uno de los objetivos principales a estudiar en esta unidad, es estimar los parámetros de la población desconocidos como la media, la varianza y la proporción poblacional mediante el cálculo de estadísticas de muestras aleatorias.

Estimación de parámetros

Tenemos que un **parámetro** es un número que describe algún aspecto de la población en estudio. En la mayoría de los casos el valor del parámetro es desconocido. Un **estadístico** es un número que se calcula a partir de los datos muestrales. Si se utiliza para estimar un parámetro, se le conoce como **estimador**.

Se denomina estimador de un parámetro a cualquier variable aleatoria que se exprese en función de la muestra aleatoria y que tenga por objetivo aproximar el valor del parámetro. Los parámetros a estudiar son los parámetros poblacionales como la media y la varianza. Existen dos formas de entregar el valor estimado del parámetro, con: un **Estimador Puntual** o **Estimación por Intervalo.**

Estimación Puntual

Definición 1: Sea θ un parámetro. A $\hat{\theta}$ se le llama **Estimador Puntual** de θ y se considera como una aproximación del valor de θ.

Existen métodos para obtener el estimador puntual de un parámetro, como el Método de los Momentos y Método de máxima Verosimilitud. A través de ellos se puede obtener por ejemplo, que el valor $\bar{x}$ (media muestral) es un estimador puntual del parámetro μ (media poblacional) y que s^2 (varianza muestral) es un estimador de la varianza poblacional σ^2.

Los estimadores tienen ciertas propiedades, como se enuncia a continuación:

1. **Insesgado:** Se dice que un estadístico $\hat{\theta}$ es un estimador insesgado del parámetro θ si $E(\hat{\theta}) = \theta$, es dceir, la esperanza del estimador es igual al valor del parámetro.
2. **Eficiente:** Si se tienen dos estimadores insesgados, $\hat{\theta}_1$ y $\hat{\theta}_2$, del parámetro θ, el de menor varianza se llama estimador más eficiente de θ.

Estimación por intervalo

Puede ser que no siempre el estimador puntual más eficiente estime el parámetro poblacional con exactitud, la precisión aumenta con muestras grandes, pero no hay razón de esperar que un estimador puntual de una muestra sea exactamente igual al parámetro poblacional que estima, es por ello que a veces es preferible determinar un intervalo dentro del cual esperaríamos encontrar el valor del parámetro, tal intervalo se llama **Intervalo de confianza**.

Definición 2: (Intervalo de Confianza)
Sea $x_1, x_2, \dots, x_n$ una muestra aleatoria de una población, cuya distribución de probabilidad es la función $f(x, \theta)$. Una estimación por intervalo de un parámetro poblacional θ es un intervalo de la forma $\hat{\theta}_1 < \theta < \hat{\theta}_2$, donde $\hat{\theta}_1 < \hat{\theta}_2$, dependen del valor de $\hat{\theta}$ para una muestra particular y también de la distribución muestral de $\hat{\theta}$.

Basado en la distribución muestral de $\hat{\theta}$ se puede determinar si el intervalo $(\hat{\theta}_1, \hat{\theta}_2)$ con una probabilidad dada contiene realmente al parámetro θ que se supone que va a estimar. Esto es:

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha, \text{ donde } 0 < \alpha < 1.$$

El intervalo $(\hat{\theta}_1, \hat{\theta}_2)$ calculado de una muestra particular se llama **intervalo de confianza** del $(1 - \alpha)100\%$ de confianza para θ.

La fracción $1 - \alpha$ se denomina **coeficiente de confianza** o **nivel de confianza** y los puntos $\hat{\theta}_1$ y $\hat{\theta}_2$ se llaman **límites de confianza**.

Por ejemplo
Si $\alpha = 0{,}05$, entonces se tiene un intervalo de confianza del 95% para θ.
Esto quiere decir, de las muestras que podemos obtener, cerca del 95% contendrán al verdadero valor de θ.

Lo ideal, es tener un intervalo corto (angosto) con un alto grado de confianza. Por ejemplo, es mejor tener una confianza del 95% de que la vida promedio de un transistor eléctrico este entre 5 y 6 años, que temer una confianza de 99% de que este entre 4 y 10 años.

Observación: la elección del nivel de confianza depende del investigador.

El objetivo acá será de determinar un posible “rango” de valores o “intervalo”, en el que pueda precisarse, con determinada probabilidad, que el verdadero valor del parámetro se encuentra dentro de esos límites. Se revisará como llega a constituirse un intervalo de confianza, y los posibles intervalos de confianza para los parámetros más usuales como la media, la varianza y proporciones. Para esto, trabajaremos bajo el supuesto de que la variable en estudio es una variable aleatoria que sigue una distribución. A continuación, se presentan los intervalos de confianza a trabajar aquí.

Dada una variable aleatoria de distribución normal $X \sim N(\mu\,, \sigma^2)$, nos interesa calcular el intervalo de confianza para el parámetro μ. Se tiene dos casos:

- **Intervalo de confianza para la media (μ), con varianza conocida**

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de tamaño n que distribuye normal con media μ y varianza σ^2, es decir $x \sim N(\mu\,, \sigma^2)$. Entonces la media muestral, $\bar{x}$, también distribuye normal con media μ y varianza σ^2/n, es decir $\bar{x} \sim N\left(\mu\,, \frac{\sigma^2}{n}\right)$.

Como $\bar{x} \sim N\left(\mu\,, \frac{\sigma^2}{n}\right)$, entonces $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0\,, 1)$

Basándonos en esta fórmula podemos determinar el intervalo de confianza para μ, eligiendo un nivel de confianza $1-\alpha$, podemos determinar dos valores Z_1 y Z_2, talque

$$P(Z_1 < Z < Z_2) = 1 - \alpha$$

hay infinitas formas de escoger Z_1 y Z_2 que cumplan tal condición, el más simple es escoger que $Z_2 = -Z_1 = Z_0$. Por lo que

$$P(-Z_0 < Z < Z_0) = 1 - \alpha$$

$$P\left(-Z_0 < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < Z_0\right) = 1 - \alpha$$

$$P\left(-Z_0 \cdot \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < Z_0 \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(-\bar{x} - Z_0 \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + Z_0 \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\bar{x} - Z_0 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_0 \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Donde: $Z_0 =$ valor crítico de la distribución

Por simetría de la curva normal, se tiene $Z_0 = Z_{1-\frac{\alpha}{2}}$, donde $Z_{1-\frac{\alpha}{2}}$ se obtiene utilizando la tabla de la curva normal tipificada.

Por lo que el intervalo para la media (μ) con varianza conocida, queda

$$P\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

o

$$IC_\mu = \left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\,;\, \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Ejemplo 1

Si una muestra aleatoria de tamaño 20 de una población normal con varianza 225 tiene una media muestral de 64,3. Construya un intervalo de confianza del 95% de confianza para μ.

Desarrollo

$(1-\alpha)100\% = 95\% \Longrightarrow \alpha = 0{,}05$

$n = 20$

$\sigma^2 = 225 \Longrightarrow \sigma = 15$

$\bar{x} = 64{,}3$

$$P\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

$$P\left(64{,}3 - Z_{1-\frac{0{,}05}{2}} \cdot \frac{15}{\sqrt{20}} \leq \mu \leq 64{,}3 + Z_{1-\frac{0{,}05}{2}} \cdot \frac{15}{\sqrt{20}}\right) = 0{,}95$$

$$IC_\mu = \left(64{,}3 - Z_{0{,}975} \cdot \frac{15}{\sqrt{20}} \leq \mu \leq 64{,}3 + Z_{0{,}975} \cdot \frac{15}{\sqrt{20}}\right)$$

$$IC_\mu = \left(64{,}3 - (1{,}96) \cdot \frac{15}{\sqrt{20}} \,;\, 64{,}3 + (1{,}96) \cdot \frac{15}{\sqrt{20}}\right)$$

$$IC_\mu = (57{,}7 \,;\, 70{,}9)$$

Existe una probabilidad del 95% de que la media, se encuentre contenida entre 57,7 y 70,9.

Ejemplo

Un médico interesado en conocer la media del colesterol en la población toma una muestra de tamaño 225. Determine un intervalo de confianza del 99% para el promedio de colesterol, si en la muestra se encontró un promedio de 190 mg/dL y una desviación estándar de 15 mg/dL.

Desarrollo

$(1-\alpha)100\% = 99\% \Longrightarrow Z_{1-\frac{0{,}01}{2}} = Z_{0{,}995} = 2{,}57$

$n = 225 \qquad \bar{x} = 190 \qquad \sigma = 15$

$$P\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

$$P\left(190 - 2{,}57 \cdot \frac{15}{\sqrt{225}} \leq \mu \leq 190 + 2{,}57 \cdot \frac{15}{\sqrt{225}}\right) = 0{,}99$$

$$IC_\mu = (187{,}43 \leq \mu \leq 192{,}57)$$

$$IC_\mu = (187{,}43 \,;\, 192{,}57)$$

Existe una probabilidad del 99% de que la media del colesterol se encuentre contenida entre 187,43 y 192,57 mg/dL

- **Intervalo de confianza para la media (μ), con varianza desconocida**

Sea $x_1, x_2, \dots, x_n$ una muestra aleatoria de $X \sim N(\mu, \sigma^2)$ con σ^2 desconocida. Sea la variable T, donde

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

luego T tiene una distribución $t - student$ con $(n-1)$ grados de libertad, donde s es la desviación estándar de la muestra. Para esta situación, en la que no se conoce σ, se puede utilizar T para construir un intervalo de confianza para μ.

La función de densidad de la $t - student$ gráficamente es similar a la función de densidad de la normal y su función de distribución acumulada se encuentra tabulada. Donde el parámetro que caracteriza a la $t - student$ se conoce como **grados de libertad.**

El procedimiento es igual que cuando se conoce σ, solo que se reemplaza σ por s, y la distribución normal estándar se reemplaza con la distribución $t - student$.

Podemos determinar el intervalo de confianza para μ, eligiendo un nivel de confianza $1 - \alpha$,

$$P(t_1 < T < t_2) = 1 - \alpha$$

eligiendo a $-t_1 = t_2 = t_0$, y que $t_0 = t_{(n-1,1-\alpha/2)}$, se tiene

$$P(-t_0 < T < t_0) = 1 - \alpha$$

$$P\left(-t_{\left(n-1,1-\frac{\alpha}{2}\right)} < \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} < t_{\left(n-1,1-\frac{\alpha}{2}\right)}\right) = 1 - \alpha$$

$$P\left(-t_{\left(n-1,1-\frac{\alpha}{2}\right)} \cdot \frac{s}{\sqrt{n}} < \bar{x} - \mu < t_{\left(n-1,1-\frac{\alpha}{2}\right)} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\bar{x} - t_{\left(n-1,1-\frac{\alpha}{2}\right)} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\left(n-1,1-\frac{\alpha}{2}\right)} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

si $\bar{x}$ y s son la media y desviación estándar de una muestra aleatoria de tamaño n de una población normal con varianza σ^2 desconocida, el intervalo de confianza de $(1-\alpha)100\%$ para μ es:

$$IC_\mu = \left(\bar{x} - t_{\left(n-1,1-\frac{\alpha}{2}\right)} \cdot \frac{s}{\sqrt{n}} \; ; \; \bar{x} + t_{\left(n-1,1-\frac{\alpha}{2}\right)} \cdot \frac{s}{\sqrt{n}}\right)$$

donde $t_{\left(n-1,1-\frac{\alpha}{2}\right)}$ es el valor t con $n-1$ grados de libertad de la tabla.

Ejemplo 2

Un fabricante de pintura quiere determinar el tiempo de secado promedio para una nueva pintura para pared interior. Si para una prueba de 12 áreas de igual tamaño obtiene un tiempo medio de secado de 66,3 minutos y una desviación estándar de 8,4 minutos. Construya un intervalo del 95% de confianza para μ si el tiempo de secado tiene distribución normal.

Desarrollo

$n = 12 \Longrightarrow n - 1 = 11$

$\alpha = 0{,}05 \ \rightarrow \ t_{\left(n-1,1-\frac{\alpha}{2}\right)} = t_{(11,0{,}975)} \Longrightarrow t_{(11,0{,}975)} = 2{,}201$

$\bar{x} = 66{,}3$

$s = 8{,}4$

$$P\left(\bar{x} - t_{\left(n-1,1-\frac{\alpha}{2}\right)} \cdot \frac{s}{\sqrt{n}} < \mu < \ \bar{x} + t_{\left(n-1,1-\frac{\alpha}{2}\right)} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(66{,}3 - t_{(11,0{,}975)} \cdot \frac{8{,}4}{\sqrt{12}} < \mu < \ \bar{x} + t_{(11,0{,}975)} \cdot \frac{8{,}4}{\sqrt{12}}\right) = 0{,}95$$

$$IC_{\mu} = \left(66{,}3 - (2{,}201) \cdot \frac{8{,}4}{\sqrt{12}} \ ; \ 66{,}3 + (2{,}201) \cdot \frac{8{,}4}{\sqrt{12}}\right)$$

$$IC_{\mu} = (61 \ ; \ 71{,}6)$$

Existe una probabilidad del 95% de que la media, se encuentre contenida entre 61 y 71,6 minutos.

Ejercicio 3

Se encuentra que la concentración promedio de zinc que se saca del agua a partir de una muestra de mediciones de zinc de 36 sitios diferentes es 2,6 gramos por mililitro. Encuentre el intervalo de confianza del 99% para la concentración media de zinc en el rio, considere que la desviación estándar de la población es 0,3 gramos por mililitro.

Desarrollo

$(1 - \alpha)100\% = 99\% \Longrightarrow \ \alpha = 0{,}01 \ \rightarrow Z_{1-\frac{0{,}01}{2}} = 2{,}57$

$n = 36 \qquad \bar{x} = 2{,}6 \qquad \sigma = 0{,}3$

$$P\left(\bar{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \ \leq \ \mu \ \leq \ \bar{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

$$P\left(2{,}6 - \cdot \frac{0{,}3}{\sqrt{36}} \leq \mu \leq \ \ 2{,}6 + Z_{1-\frac{0{,}01}{2}} \cdot \frac{0{,}3}{\sqrt{36}}\right) = 0{,}95$$

$$IC_{\mu} = \left(2{,}6 - 2{,}57 \cdot \frac{0{,}3}{6} \leq \mu \leq \ \ 2{,}6 + 2{,}57 \cdot \frac{0{,}3}{6}\right)$$

$$IC_{\mu} = (2{,}5 \ ; \ 2{,}7)$$

Existe una probabilidad del 95% de que la concentración media de zinc en el rio, se encuentre contenida entre 2,5 y 2,7 gramos por mililitro.

- **Intervalo de confianza para la proporción poblacional**

Cuando en una población de interés se está estudiando un rasgo particular y cada miembro de la población puede clasificarse según que posea o no ese rasgo, definimos como p a la proporción (porcentaje) de la población que tiene el rasgo.

El estimador puntual para p, lo podemos obtener extrayendo una muestra aleatoria de la población de interés y determinar la proporción de objetos con el rasgo en la muestra y utilizar esta "proporción muestral" como estimador de la proporción p. Es decir,

$$\hat{p} = \frac{número\ de\ objetos\ en\ la\ muestra\ con\ el\ rasgo}{tamaño\ de\ la\ muestra\ (n)}$$

Si n es suficientemente grande, $\hat{p}$ tiene una distribución aproximadamente normal con media p y desviación estándar $\sqrt{\frac{p(1-p)}{n}}$. Es decir $\hat{p}$ es aproximadamente normal, $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Con el mismo razonamiento que empleamos en la estimación de la media poblacional μ, el planteo inicial para estimar la proporción poblacional es determinar p_1 y p_2 que verifiquen

$$P(p_1 < \hat{p} < p_2) = 1 - \alpha$$

Y eligiendo a $-p_1 = p_2 = p_0$, considerando que $p_0 = Z_{1-\frac{\alpha}{2}}$, tenemos

$$P(-p_0 < \hat{p} < p_0) = 1 - \alpha$$

$$P\left(-Z_{1-\frac{\alpha}{2}} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < Z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(\hat{p} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

El problema es que no conocemos el parámetro, por lo que no conocemos la desviación estándar, asa que se estima la desviación estándar con el error estándar de $\hat{p}$, luego $S(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

El intervalo de confianza para la proporción poblacional para un nivel de confianza $(1-\alpha)100\%$ es:

$$IC_{\mu} = \left(\hat{p} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ;\ \hat{p} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Ejemplo 4

Una empresa de cable desea conocer qué proporción de sus clientes se informan de las noticias a través de los noticiarios que difunden. Para ello seleccionó una muestra aleatoria de 200 clientes. De las 200 personas, 110 respondieron que se informan a través de los noticieros televisivos.

a) Calcule una estimación puntual de la proporción de personas que se informan a través de los noticieros
b) Calcule un intervalo de confianza, al 95% de confianza, para la proporción anterior.

Desarrollo

a) Se tiene $n = 200$ clientes
De ellos 110, se informan a través de los noticieros
Luego, la proporción queda $\hat{p} = \frac{110}{200} = 0{,}55$

b) Intervalo de confianza, al 95% de confianza
$(1-\alpha)100\% = 95\% \Longrightarrow \alpha = 0{,}05$

$$IC_{\mu} = \left(\hat{p} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \; ; \; \hat{p} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) = 1-\alpha$$

$$P\left(0{,}55 - Z_{0{,}975} \cdot \sqrt{\frac{0{,}55(1-0{,}55)}{200}} \leq \hat{p} \leq \; 0{,}55 + Z_{0{,}975} \cdot \sqrt{\frac{0{,}55(1-0{,}55)}{200}} \right) = 0{,}95$$

$$IC_{\hat{p}} = (0{,}55 - 1{,}96 \cdot 0{,}04 \; \leq \hat{p} \leq \; 0{,}55 + 1{,}96 \cdot 0{,}04)$$

$$IC_{\hat{p}} = (0{,}4716 \; ; \; 0{,}6284)$$

Existe una probabilidad del 95% de que la proporción de clientes que se informan a través de los noticieros se encuentra entre el 47% y el 62,8%.

Ejercicio propuesto

1. Se quiere conocer el tiempo promedio de permanencia de los pacientes en un hospital, con el fin de estudiar una posible ampliación del mismo. Se tienen datos referidos a la permanencia, en días, de 42 pacientes, obteniéndose los siguientes que la media es de 9 días y la desviación estándar $s = 5$ días. Determine un intervalo de confianza del 95%, para el tiempo promedio de permanencia e interpretar el valor obtenido.

2. Se preguntó a 80 pacientes si habían sufrido algún efecto secundario tras seguir un tratamiento, de los cuales 60 dijeron que no. Calcule un intervalo de confianza, al 95% de confianza, para la proporción de pacientes que sufre efectos secundarios tras el tratamiento.