Intervalos de Confianza

Estimación puntual y por intervalo

Sergio Nava

18/4/2023

Estimadores Puntuales

El objetivo de la estadística inferencial consiste en hacer inferencias (estimaciones) acerca de los parámetros de una población teniendo en cuenta la información contenida en la muestra.

La inferencia estadística comprende una serie de técnicas de uso imprescindible para tomar decisiones con respecto a la cuestión planteada por el investigador al comienzo de su tarea de análisis de datos.

Es necesario destacar que, las decisiones que debe tomar el investigador, ante la situación de incertidumbre, así como, inferir sobre casos particulares a la generalidad, dichas decisiones deben de estar basadas en razonamientos que garanticen probabilidades pequeñas de equivocarse.

Estimación por intervalos

La estimación por intervalos es un procedimiento mediante el cual se puede afirmar, con una determinada confianza, que el intervalo (a, b) encierra el verdadero valor del parámetro. Para realizar una estimación por intervalos se hace la siguiente afirmación:

$$P(a \leq \theta \leq b) = 1 - \alpha$$

El intervalo (a, b) se llama intervalo de confianza, $b - a$ es una medida de la amplitud de dicho intervalo y $1 - \alpha$ es una medida de la confianza con la que contamos para efectuar la estimación.

Nota: α por sí sola se denomina nivel de significancia.

Intervalos de confianza para medias con varianza conocida

Hemos explicado que $\bar{x}$ es el mejor estimador puntual de μ, pero nos sorprendería realmente que la media muestral fuera exactamente igual a μ. Resultaría más comprensible pensar que el valor aportado por un estimador se ubica en las cercanías del parámetro.

Esta situación sugiere que puede ser más apropiado efectuar un intervalo alrededor de $\bar{x}$ y establecer una cierta confianza de que μ esté comprendido en dicho intervalo.

Importante

- En la estimación por intervalos de confianza de un parámetro poblacional siempre hablaremos de la confianza de que el intervalo contenga al parámetro.
- El parámetro es una cantidad desconocida, pero fija.

Ya sabemos que el mejor estimador puntual de μ es $\bar{x}$, la media muestral y, en consecuencia, lo utilizaremos para la construcción del Intervalo de confianza. Basándonos en el teorema central del límite podemos establecer:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Para poder utilizar la tabla de probabilidades normales debemos estandarizar esta variable aleatoria.

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

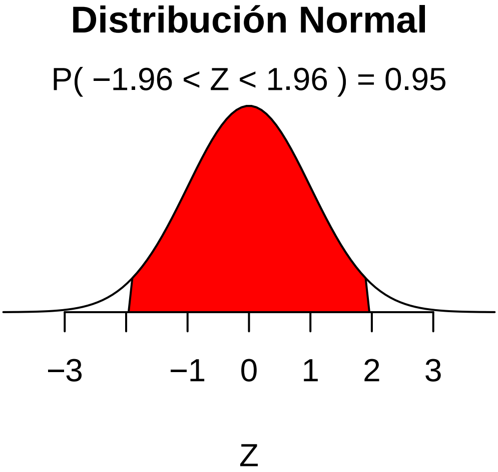
Siendo Z una variable normal estandarizada, se deberán buscar dos valores z_1 y z_2 tales que

$$P(z_1 \leq Z \leq z_2) = 1 - \alpha$$

o lo que es lo mismo

$$P\left(z_1 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_2\right) = 1 - \alpha$$

Donde $z_1 = Z_{1-\frac{\alpha}{2}}$ y $z_2 = Z_{\frac{\alpha}{2}}$. Por ejemplo para $\alpha = 0.05$

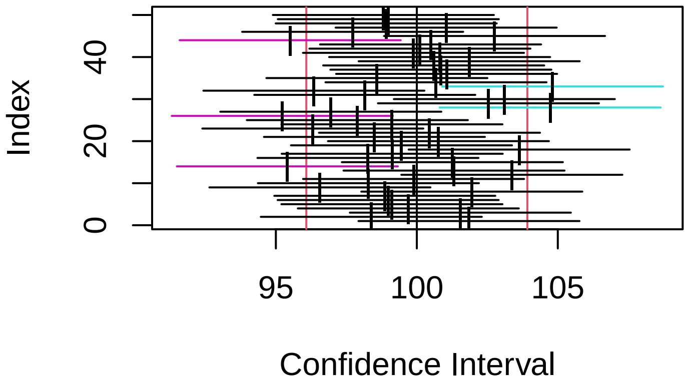


Importante El coeficiente de confianza es la probabilidad (confianza) de que un Intervalo contenga al parámetro estimado.

El coeficiente de confianza es un valor fijado por el investigador antes de comenzar la estimación. Así, si decide trabajar con una confianza del 95% para efectuar la estimación, el razonamiento será el siguiente:

- Sobre 100 muestras aleatorias de un cierto tamaño n de una población, si en cada una se calcula la medía muestral $\bar{x}$ y, a partir de ellas, se construyen sus correspondientes intervalos de confianza para el parámetro que se desea estimar, 95 contendrán al verdadero valor del parámetro poblacional, mientras que 5 no lo abarcarán.

Confidence intervals based on z distributi



Cada punto segmento pequeño vertical representa la media muestral y cada línea que pasa a través de estos estos segmentos es la amplitud del intervalo correspondiente. En este caso vemos que, de 50 intervalos obtenidos a partir de las 50 muestras aleatorias 5 de tales intervalos (10%) no contienen al verdadero valor del parámetro μ y 45 (90%) sí lo contienen.

Entonces para $\alpha = 0.05$ podemos escribir

$$P\left(-1.96 \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95$$

Ahora bien, como estamos tratando de estimar al parámetro μ lo razonable sería despejar convenientemente de modo que quede en el centro del intervalo el parámetros μ.

$$P\left(-1.96 \times \frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(-1.96 \times \frac{\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq 1.96 \times \frac{\sigma}{\sqrt{n}} - \bar{X}\right) = 0.95$$

Si multiplicamos por (-1) toda la desigualdad, cambia el sentido de la desigualdad:

$$P\left(1.96 \times \frac{\sigma}{\sqrt{n}} + \bar{X} \geq \mu \geq -1.96 \times \frac{\sigma}{\sqrt{n}} + \bar{X}\right) = 0.95$$

es decir

$$P\left(\bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

reacomodando convenientemente tenemos que

$$P\left(\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Este es el intervalo de confianza para el parámetro μ cuando trabajamos con una confianza del 95%.

- En situaciones reales de investigación, la población generalmente es grande y σ es un parámetro desconocido. Para solucionar este problema, σ también debe ser estimado. Su estimador lógico será s, la desviación estándar de la muestra.
- Si el tamaño de la muestra es suficientemente grande (muestras de tamaño mayor o igual a 30), no hay problemas en seguir utilizando la distribución de probabilidad Normal para medir la confianza de la estimación.
- En cambio, si la muestra es chica y no se puede aumentar su tamaño, por razones de costo o de tiempo u otras, para calcular la confianza de la estimación utilizaremos otra distribución de probabilidad: la distribución correspondiente a la variable *t de Student*, como veremos más adelante.

Supongamos, por ejemplo, que se quiere conocer el aumento promedio de peso de niños pertenecientes a un estrato social muy bajo cuando se los alimenta con una dieta fortificante.

La población objetivo fue especificada como los niños de dicho estrato social en el estado de Zacatecas.

Sería casi imposible alimentar con la dieta a todos los niños del estado. La única solución a este problema será recurrir a una muestra de $n = 40$ niños seleccionados aleatoriamente de la población en cuestión y alimentarlos con la dieta en estudio. Al cabo de un cierto tiempo se registrarán los aumentos de peso de cada niño y se calculará el promedio en la muestra.

Supongamos que resultó $\bar{x} = 4.300$ kg.

Utilizaremos esta estimación puntual para efectuar una estimación por intervalos del parámetro μ: aumento de peso promedio de niños pertenecientes al estrato social muy bajo del estado de Zacatecas.

Supongamos que, por estudios previos se conoce que el valor que toma el parámetro varianza poblacional es $\sigma^2 = 4\ kg^2$. Podemos utilizar la distribución de probabilidad normal para efectuar la estimación. Con una confianza de $1 - \alpha = 0.95$, el intervalo resultante será:

$$P\left(4.3 - 1.96 \times \frac{2}{\sqrt{40}} \leq \mu \leq 4.3 + 1.96 \times \frac{2}{\sqrt{40}}\right) = 0.95$$

$$P(3.69 \leq \mu \leq 4.69) = 0.95$$

Y nos da una amplitud del intervalo $4.96 - 3.68 = 1.24$

Ahora bien, si en lugar de tomar una muestra de 40 niños hubiésemos tomado una de $n = 100$ niños, manteniendo todos los demás valores constantes, el intervalo será:

$$P\left(4.3 - 1.96 \times \frac{2}{\sqrt{100}} \leq \mu \leq 4.3 + 1.96 \times \frac{2}{\sqrt{100}}\right) = 0.95$$

$$P(3.91 \leq \mu \leq 4.69) = 0.95$$

Y nos da una amplitud de $4.69 - 3.91 = 0.78$.

Vemos que al aumentar el tamaño de la muestra disminuyó la amplitud del intervalo de confianza, logrando así una mayor precisión en la estimación.

Pero, ¿Qué pasaría ahora si con los mismos datos la varianza poblacional σ^2 fuese 9 en lugar de 4?

En este caso estamos pensando que la variable aumento de peso en los niños tiene una variabilidad mayor de niño a niño.

$$P\left(4.3 - 1.96 \times \frac{3}{\sqrt{40}} \leq \mu \leq 4.3 + 1.96 \times \frac{3}{\sqrt{40}}\right) = 0.95$$

$$P(3.37 \leq \mu \leq 5.23) = 0.95$$

Lo que nos da una amplitud de intervalo de $5.23 - 3.37 = 1.86$.

Si comparamos este intervalo con el primero que hemos construido, vemos que ahora se han ampliado los límites provocando una merma en la precisión de la estimación.

Y ¿Qué podemos decir de la confianza de la estimación?

Cuando se efectúa una estimación por intervalos, a medida que se incrementa el nivel de confianza se incrementa también la amplitud del intervalo y, en consecuencia, disminuye la precisión del mismo.

Así, por ejemplo, si decidimos aumentar el **nivel de confianza** trabajando con $1 - 0.01 = \mathbf{0.99}$, el intervalo sería:

$$P\left(4.3 - 2.576 \times \frac{2}{\sqrt{40}} \leq \mu \leq 4.3 + 2.576 \times \frac{2}{\sqrt{40}}\right) = 0.99$$

$$P(3.49 \leq \mu \leq 5.11) = 0.99$$

Lo que nos da una amplitud de intervalo de $5.11 - 3.49 = 1.62$. Luego, construyendo el intervalo con un 99% de confianza hemos aumentado nuestra seguridad en la estimación, pero a costa de una precisión menor.

Pero, para evitar muchos cálculos y aprovechando la simetría de la distribución normal se expone la ecuación siguiente para el cálculo del intervalo de confianza de la media ($n \geq 30$):

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Algunos autores expresan el intervalo como $\bar{X} \pm Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$

Alfa	Coeficiente de Confianza	Z
0.10	0.90	1.645
0.05	0.95	1.960
0.01	0.99	2.576

Intervalo de la media con varianza desconocida

un caso muy común es cuando el investigador está obligado a trabajar con muestras chicas y pretende estimar la media poblacional con σ^2 desconocida.

En estos casos, dado que no se conoce la σ^2, para estimar por intervalos al parámetro μ debemos recurrir a un nuevo estadístico distribuido como una variable *"t" de Student*. Este estadístico surge de estandarizar la variable media muestral, pero tomando a S^2 como estimador puntual de la varianza poblacional.

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Esta variable ya no se distribuye normalmente cuando el tamaño de la muestra es chico, en la práctica $n < 30$. Entonces

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Donde $n - 1$, es el denominador con el que se calculó la varianza muestral S^2. Este valor corresponde a los grados de libertad de la variable "t" de Student.

Similarmente al caso de la normal obtenemos:

$$P\left(\bar{X} - t_{\frac{\alpha}{2},n-1} \times \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2},n-1} \times \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Algunos autores expresan el intervalo como $\bar{X} \pm t_{\frac{\alpha}{2},n-1} \times \frac{\sigma}{\sqrt{n}}$

Supongamos que un investigador desea estimar el rendimiento promedio de grasa que contiene la leche producida por vacas de cierta raza en un período de tiempo determinado.

Para ello se extrae una muestra de 10 vacas lecheras, obteniendo:

$n = 10$, $\bar{X} = 36.4grs$, $s^2 = 264.04grs^2$ y $s = 16.25grs$

Por lo tanto $s^2/n = 264.04/10 = 26.404$ y $s/\sqrt{n} = 5.138grs$

Los únicos datos que posee el investigador para llevar a cabo su investigación son la media y la varianza muestral.

Como la muestra es chica y no conocemos la varianza poblacional σ^2, el estadístico para confeccionar el correspondiente intervalo de confianza de μ, tendrá distribución *t de Student* con $n-1$ grados de libertad.

Para un nivel de confianza del 95% y buscando convenientemente en la tabla de probabilidades de la distribución *t de Student*, obtenemos:

$$t_{\frac{0.05}{2},10-1} = t_{0.025,9} = 2.62$$

El intervalo de confianza será

$$P\left(\bar{X} - t_{\frac{\alpha}{2},n-1} \times \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2},n-1} \times \frac{S}{\sqrt{n}}\right) = 0.95$$

$$P\left(36.4 - 2.62 \times 5.138 \leq \mu \leq 36.4 + 2.62 \times 5.138\right) = 0.95$$

$$P\left(24.78 \leq \mu \leq 48.02\right) = 0.95$$

El promedio de grasa en la leche de dichos animales se estima entre un 24.78 grs y 48.02 grs con un nivel de confianza del 95%.

Intervalos de confianza para proporciones

Lo visto en la sección anterior se puede usar con el fin de determinar intervalos de confianza para la media de cualquier población se la que se haya extraido una muestra grande. Cuando la población tiene una distribución de Bernoulli, esta expresión toma una forma especial.

Suponga que se tiene un proceso de fabricación y este tiene especificaciones. Se prueba una muestra de 144 productos y 120 satisfacen la especificación. Sea p la proporción de artículos en la población que satisfacen dicha especificación. Se desea encontrar un intervalo de confianza de 95% para p.

Se empieza construyendo un estimador de p. Sea X el número de artículos en la muestra que satisface la especificación. Entonces $X \sim Bin(n, p)$, donde $n = 144$ es el tamaño muestral. El estimador de p es $\hat{p} = X/n$. En este ejemplo, $X = 120$, por lo que $\hat{p} = 120/144 = 0.833$. La incertidumbre, o desviación estándar de $\hat{p}$, es $\sigma_{\hat{p}} = \sqrt{p(1-p)}$. Puesto que el tamaño muestral es grande, se tiene por el teorema del límite central que

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Con un razonamiento similar a los casos anteriores se puede obtener

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

A primera vista la expresión anterior parece un intervalo de confianza con un nivel $1-\alpha$ para p. Sin embargo los límites contienen una p desconocida, y por eso no se puede calcular. El punto de vista tradicional es sustituir esa p por $\hat{p}$. De tal manera que quedaría así:

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

Algunos autores expresan el intervalo como $\hat{p} \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Para el ejemplo obtenemos $.833 \pm 1.96 \times \sqrt{.833 \times (1-.833)/144}$ es decir 0.833 ± 0.0609 y por lo tanto el intervalo va de 0.7721 a 0.8939 con una confianza del 95%.

Intervalos de confianza para la varianza

Para utilizar a S^2 como estimador de σ^2 necesitamos conocer su distribución de probabilidad. De esta manera podremos establecer un cierto coeficiente de confianza de la estimación.

No existe una distribución conocida para S^2 pero sí para cierta transformación del mismo. Si la muestra proviene de una población en la cual la variable en estudio se distribuye normalmente, tenemos

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

donde χ^2_{n-1} es la distribución Ji-cuadrada co $n-1$ grados de libertad.

Una vez que contamos con esta información, podemos establecer un intervalo de confianza para estimar σ^2, de la siguiente manera:

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

Despejando tenemos

$$P\left(\frac{S^2 \times (n-1)}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{S^2 \times (n-1)}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) = 1-\alpha$$

Ejemplo

El Departamento de Control de Calidad de una empresa decide estudiar la homogeneidad de los productos comprados a un proveedor de materiales plásticos. Para ello toma 6 piezas de la producción de un envío y las somete a pruebas de resistencia. Las siguientes observaciones representan las cargas de rotura en unidades de 1,000 libras por pulgada cuadrada: 15.3, 18.7, 22.3, 17.6, 19.1 y 14.8

El Departamento de Calidad ha determinado que interesa conocer la variabilidad de las cargas de rotura, pues la uniformidad de este insumo es fundamental en el proceso de producción de la fábrica.

Como primera medida y habiendo comprobado que la variable cargas de rotura tiene distribución normal. Se procede a construir un intervalo de confianza del 90%:

$S^2 = 7.575$, $n-1=5$, $\alpha = 0.1$ por lo tanto $\chi_{.05,5} = 11.07$ y $\chi_{.95,5} = 1.15$, y el intervalo de confianza sería

$$P\left(\frac{7.575 \times (5)}{\chi^2_{\frac{0.10}{2},5}} \leq \sigma^2 \leq \frac{7.575 \times (5)}{\chi^2_{1-\frac{0.10}{2},5}}\right) = 0.90$$

$$P\left(3.42 \leq \sigma^2 \leq 32.93\right) = 0.90$$

Es decir el intevalo es de confianza de la varianza a un nivel de 90% es $(3.42, 32.93)$.

Obtención de Intervalos de confianza con R

Una muestra	Parámetro	Función	Ejemplo
Media	μ	t.test	t.test(x=altura, conf.level=0.90)$conf.int
Proporción	p	prop.test	prop.test(x=275, n=500, conf.level=0.90)$conf.int
Varianza	σ^2	stests::var.test()	var.test(x=altura, conf.level=0.98)$conf.int

Diferencia de	Parámetro	Función	Ejemplo
Medias de muestras independientes	$\mu_1 - \mu_2$	t.test	t.test(x=hombres$altura, y=mujeres$altura,,paired=FALSE, var.equal=FALSE,conf.level = 0.95)$conf.int
Medias de muestras pareadas	$\mu_1 - \mu_2$	t.test	t.test(x=Antes, y=Despues, paired=TRUE, conf.level=0.95)$conf.int
Proporciones	$p_1 - p_2$	prop.test	prop.test(x=c(75, 80), n=c(1500, 2000), conf.level=0.90)$conf.int
Varianza	σ^2	stests::var.test()	var.test(x=hombres$altura, y=mujeres$altura,,conf.level=0.95)$conf.int

Intervalos de confianza en R