Aula360
Aprendizaje para el siglo XXI

Apuntes Unidad 4

Estadísticos y Estimadores

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

CMM Centro de Modelamiento Matemático

Curso: Probabilidad y estadística descriptiva
Unidad 4 : Inferencia estadística
Tema: Intervalos de confianza.
Contenido: Estadísticos y Estimadores

ESTADÍSTICOS Y ESTIMADORES

Un **estadístico** es una función que toma valores de la muestra y que, por medio de operaciones matemáticas, entrega otro valor. Este valor suele ser distinto para cada muestra. Por ejemplo, el promedio muestral $\bar{X}$ es un estadístico que se calcula sumando los n datos de una muestra, y luego dividiendo por n.

Un **estimador** es cualquier estadístico que se utiliza para estimar un **parámetro** o alguna otra característica de una población de interés. Por ejemplo, el estadístico promedio muestral $\bar{X}$ puede usarse como un estimador para el parámetro media poblacional μ. Por último, llamaremos **estimación** a un valor particular del estimador calculado a partir de una muestra.

Para referirnos a un parámetro poblacional de forma general, utilizaremos la letra griega theta (θ), mientras que para referirnos a un estimador de este parámetro utilizaremos esta misma letra griega, pero en mayúscula y con un acento circunflejo o caret, es decir, $\hat{\Theta}$. Una estimación la denotaremos con la letra griega minúscula theta acompañada de un caret, es decir, $\hat{\theta}$.

Es importante notar dos aspectos acerca de la relación entre parámetros, estimadores y estimaciones. Estas son las siguientes:

1. Si se tienen varias muestras de una población, entonces un mismo estimador $\hat{\Theta}$ puede arrojar estimaciones diferentes para el mismo parámetro θ.

2. Pueden existir varios estimadores distintos para un mismo parámetro poblacional θ.

Dependiendo de lo que se quiere estudiar, las características de la población y las propiedades de cada estimador, puede haber un estimador que sea más apropiado que otro para estimar el parámetro de interés.

ESTIMACIÓN DE PARÁMETROS POBLACIONALES Y ERROR DE ESTIMACIÓN

Para realizar estimaciones de los parámetros de una población por medio de una muestra, es necesario que dicha muestra sea seleccionada de forma aleatoria.

Recordemos que una muestra es un conjunto de n individuos o elementos de la población. En términos sencillos, se tiene que una muestra es aleatoria si los individuos tienen la misma probabilidad de ser seleccionados. Formalmente, se tiene una variable aleatoria X que representa el resultado de la selección de una observación en la población. Supón que realizamos n veces el experimento de seleccionar una observación de esta población, con cada repetición llevada a cabo de forma independiente y bajo las mismas condiciones. Entonces, si denotamos como X_i a la i-ésima realización del experimento, la colección $X_1, X_2, \ldots, X_n$ representa una muestra aleatoria de la población.

Una vez que se haya seleccionado la muestra aleatoria $X_1, X_2, \ldots, X_n$ de la población de interés, utilizamos el estimador $\widehat{\Theta} = f(X_1, X_2, \ldots, X_n)$ para obtener una estimación $\widehat{\theta}$. Como la muestra es aleatoria, se asume que es representativa de la población. Por lo tanto, si la muestra es de gran tamaño, el parámetro θ de dicha población debería ser aproximadamente igual a la estimación $\widehat{\theta}$.

Es importante enfatizar que el valor obtenido por medio de la estimación puntual es, como su nombre lo indica, una estimación y no hay garantía de que sea muy cercano al valor real del parámetro de la población. De esta manera, definimos el **error de estimación** como el valor absoluto de la diferencia entre el valor real del parámetro y la estimación realizada, es decir:

$$\text{Error de estimación} = |\theta - \widehat{\theta}|$$

Como la definición de error de estimación depende del valor real del parámetro que estamos tratando de estimar, en la práctica no es posible calcular el error de estimación.

ESTIMAR UN PARÁMETRO POR MEDIO DE UNA ESTIMACIÓN PUNTUAL

Estimar un parámetro por medio de una estimación puntual tiene dos limitaciones importantes:

1. La estimación que se calcula a partir de un estimador varía de muestra en muestra, incluso si estas muestras son aleatorias y representativas de la población.
2. No es posible establecer algún nivel de confianza que se puede tener en una estimación específica. En otras palabras, una vez obtenida la estimación, no tenemos forma de saber que tan buena es.

Curso: Probabilidad y estadística descriptiva
Unidad 4 : Inferencia estadística
Tema: Intervalos de confianza.
Contenido: Estadísticos y Estimadores

Considera una variable aleatoria que sigue una distribución normal en una cierta población, con parámetros $E(X) = \mu$ y $DE(X) = \sigma$. Si se extrae una muestra aleatoria de tamaño n, la variable aleatoria $\bar{X}$, correspondiente al promedio muestral, sigue una distribución normal con parámetros $E(\bar{X}) = \mu$ y $DE(X) = \frac{\sigma}{\sqrt{n}}$:

$$\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Asimismo, una variable aleatoria X que sigue una distribución normal con parámetros μ y σ cumple la denominada regla empírica, es decir, se cumple que:

- La probabilidad $P(\mu - \sigma \leq X \leq \mu + \sigma)$ es igual a 68,2%.
- La probabilidad $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$ corresponde al 95,4%.
- La probabilidad de que $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma)$ corresponde al 99,6%.

Además, para cada variable aleatoria continua X que se distribuye como una normal con parámetros μy σ, es posible definir una nueva variable aleatoria:

$$Z = \frac{X - \mu}{\sigma}$$

Esta nueva variable sigue una **distribución normal estándar,** cuya media es $\mu_Z = 0$ y su desviación estándar es $\sigma_Z = 1$, lo que anotamos como $Z \sim N(0, 1)$. Esta transformación se conoce como **estandarización.**

Una forma de estimar un parámetro u otra característica de interés de la población es a través de intervalos. Esta estimación consiste en determinar un rango de valores donde es posible encontrar el valor real de un parámetro de la población. Este intervalo se calcula a partir de los datos de una muestra aleatoria extraída de la población.

ESTIMACIÓN POR INTERVALOS USANDO REGLA EMPÍRICA

Para aproximarnos a una estimación de la media poblacional a través de intervalos, podemos usar la regla empírica para la construcción de estos. Recordemos que la regla empírica está asociada a distribuciones que siguen una distribución aproximadamente normal, por lo que aplica en este caso.

Curso: Probabilidad y estadística descriptiva

Unidad 4 : Inferencia estadística

Tema: Intervalos de confianza.

Contenido: Estadísticos y Estimadores

Consideremos una variable aleatoria continua X que sigue una distribución normal con media μ y desviación estándar σ:

$$X \sim N(\mu, \sigma)$$

Si se extrae una muestra aleatoria de tamaño de n de la población, la variable aleatoria $\bar{X}$, correspondiente al promedio muestral, sigue una distribución normal con parámetros μ y $\frac{\sigma}{\sqrt{n}}$. De acuerdo con estos parámetros y la regla empírica, podemos representar la función de distribución de probabilidad de $\bar{X}$ de la siguiente manera:

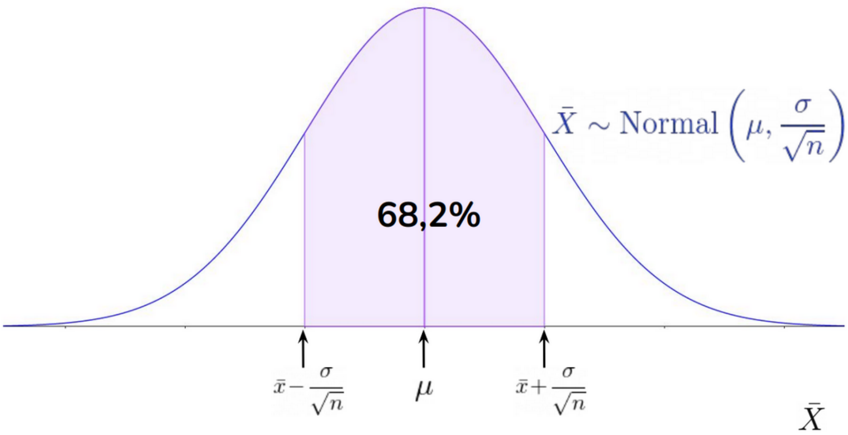
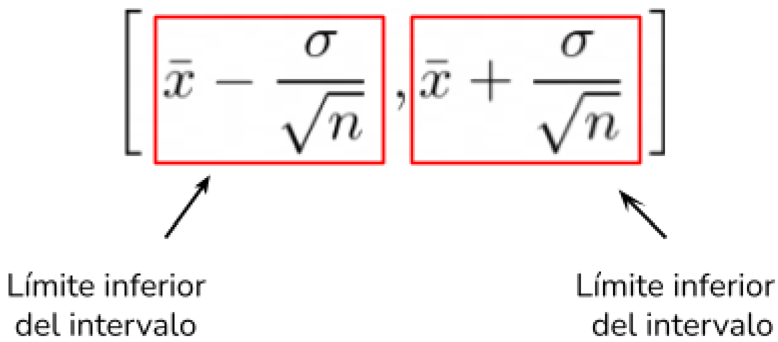


Figura 1: Gráfico de la función de distribución de $\bar{X}$.

Entonces, la probabilidad de que la media poblacional μ esté en este intervalo es aproximadamente 68,2%.

Los límites inferior y superior del intervalo corresponden a:



La interpretación es análoga para los otros dos intervalos asociados a la regla empírica:

- Si consideramos el intervalo de valores formado por los límites $\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}$ y $\bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}}$, aproximadamente el 95,4% de la distribución se encuentra a una distancia máxima de dos desviaciones estándar respecto a la media μ, es decir, la probabilidad de que la media poblacional μ esté a lo más, a dos desviaciones estándar del promedio muestral $\bar{x}$ es aproximadamente 95,4%.

- Si consideramos el intervalo de valores formado por los límites $\bar{x} - 3 \cdot \frac{\sigma}{\sqrt{n}}$ y $\bar{x} + 3 \cdot \frac{\sigma}{\sqrt{n}}$, el 99,6% de la distribución se encuentra a una distancia máxima de tres desviaciones estándar respecto a la media μ, es decir, la probabilidad de que la media poblacional μ esté a lo más, a tres desviaciones estándar del promedio muestral $\bar{x}$ es aproximadamente 99,6%.

FUNCIÓN INVERSA DE LA PROBABILIDAD ACOMULADA DE UNA DISTRIBUCIÓN NORMAL

Consideremos una variable aleatoria Z que sigue una distribución normal estándar, es decir, que su media es $\mu_Z = 0$ y su desviación estándar es $\sigma_Z = 1$.

En muchas ocasiones estaremos interesados en determinar el valor z que genera una probabilidad acumulada $P(Z \leq z) = \alpha$. Este cálculo corresponde a la función inversa de la probabilidad acumulada, es decir, ahora nos entregan la probabilidad y queremos determinar el valor que la generó.

La expresión z_α notará el valor sobre el eje horizontal para el cual el área bajo la curva queda a la izquierda de ese valor.

En caso de tener distribuciones normales que no sean la estándar, primero debemos realizar el proceso de estandarización y luego buscar el valor de z.

SÍNTESIS

- Un **estadístico** es una función que toma valores de una muestra y que, por medio de operaciones matemáticas, entrega otro valor. Un ejemplo es el promedio muestral $\bar{X}$, que se calcula sumando los n datos de una muestra, y luego dividiendo por n.

- Un **estimador** $\widehat{\Theta}$ es cualquier estadístico que se utiliza para estimar un **parámetro** θ o alguna otra característica de una población de interés. Un estimador para el parámetro media poblacional μ es el promedio muestral $\bar{X}$. Una **estimación** $\hat{\theta}$ es un valor particular del estimador calculado a partir de una muestra.

- Es importante notar dos aspectos acerca de la relación entre parámetros, estimadores y estimaciones:

  - Si se tienen varias muestras de una población, entonces un mismo estimador $\widehat{\Theta}$ puede arrojar estimaciones diferentes para el mismo parámetro θ.

  - Pueden existir varios estimadores distintos para un mismo parámetro poblacional θ.

- Dada una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una población de interés, se puede utilizar el estimador $\widehat{\Theta} = f(X_1, X_2, \ldots, X_n)$ para obtener una estimación $\hat{\theta}$, la que, si la muestra es grande, $\hat{\theta}$ debería ser un valor cercano al parámetro θ. Este proceso se conoce como **estimación puntual**, y está representado en el siguiente diagrama:

- Es importante enfatizar que el valor obtenido por medio de la estimación puntual es, como su nombre lo indica, una estimación y no hay garantía de que siempre sea muy cercano al valor real del parámetro de la población.

- La estimación por intervalos consiste en determinar un rango de valores donde es posible encontrar el valor real de un parámetro de la población.

- Consideremos una variable aleatoria continua X que sigue una distribución normal con media μ y desviación estándar σ. Si se extrae una muestra aleatoria de tamaño de n de la población, la variable aleatoria $\bar{X}$, correspondiente al promedio muestral, sigue una distribución normal con parámetros μ y $\frac{\sigma}{\sqrt{n}}$.

  De acuerdo a estos parámetros, podemos usar la regla empírica para construir intervalos de valores para estimar la media poblacional μ:

  - En el intervalo de valores formado por los límites $\bar{x} - \frac{\sigma}{\sqrt{n}}$ y $\bar{x} + \frac{\sigma}{\sqrt{n}}$, aproximadamente el 68,2% de la distribución se encuentra a una distancia máxima de una desviación estándar respecto a la media μ.

- Si consideramos el intervalo de valores formado por los límites $\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}$ y $\bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}}$, aproximadamente el 95,4% de la distribución se encuentra a una distancia máxima de dos desviaciones estándar respecto a la media μ.

- Si consideramos el intervalo de valores formado por los límites $\bar{x} - 3 \cdot \frac{\sigma}{\sqrt{n}}$ y $\bar{x} + 3 \cdot \frac{\sigma}{\sqrt{n}}$, el 99,6% de la distribución se encuentra a una distancia máxima de tres desviaciones estándar respecto a la media μ.