UNIVERSIDAD TECNOLÓGICA NACIONAL
Facultad Regional San Nicolás

PROBABILIDAD y ESTADÍSTICA II

UNIDAD Nº3

Licenciatura en Enseñanza de la Matemática

Año 2011

Mg. Lucía C. Sacco

UNIDAD Nº3 Estimaciones de parámetros.

Procedimientos de la Estadística inferencial: la estimación. Estimación de parámetros. Propiedades deseables de un estimador: insesgado, eficiente o con varianza mínima, coherencia y suficiencia. Interpretación gráfica de propiedades. Valor estimado de una parámetro.

Estimación puntual. Métodos de estimación puntual.

Estimación por intervalos. Intervalos de confianza. Intervalos de confianza para las medias con varianza conocida.

Propósitos:

Brindar oportunidades para la construcción de herramientas que permitan:

- Comprender los fundamentos teóricos y la lógica subyacente de la inferencia estadística en una de sus grandes ramas: la estimación de parámetros.
- Diferenciar las formas de estimación de parámetros poblacionales teniendo en cuenta las condiciones de los buenos estimadores y reconociendo las particularidades de cálculo en distintos casos -media, varianza y proporción poblacional-.
- Aplicar conceptos y procedimientos centrales de la inferencia estadística en la resolución de casos y problemas.

Bibliografía sugerida:

- Alfonso Lopes, Paulo (2000). *Probabilidad & Estadística. Conceptos, modelos y aplicaciones en Excel.* Prentice Hall.
- Lind, Douglas – Mason, Robert – Marchal, William (2000). *ESTADÍSTICA para administración y Economía.* Tercera edición. Irwin McGraw – Hill.
- Meyer, P. (1973). *Probabilidad y aplicaciones estadísticas.* Versión en español. Universidad Católica de Chile. Fondo Educativo Interamericano.
- Zylberberg, A. (2004). *Probabilidad y Estadística P(X).* Nueva Librería.

Páginas de Internet:

- http://recursostic.educacion.es/descartes/web/materiales_didacticos/inferencia_estadistica/estimac.htm
- http://es.wikipedia.org/wiki/Estimaci%C3%B3n_estad%C3%ADstica
- http://es.scribd.com/doc/31986198/Nociones-Basicas-de-Estadistica

1. Introducción

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 aquí que las decisiones que debe tomar el investigador ante la situación de incertidumbre que implica inferir de casos particulares a la generalidad, deben estar respaldadas por la objetividad que garantiza la aplicación del método científico.

De este modo, los resultados obtenidos en situaciones experimentales, serán idealizados de acuerdo a un modelo probabilístico conveniente, permitiendo al investigador medir en términos de probabilidad la incertidumbre que trae aparejada la generalización de sus resultados. En otras palabras, podrá medir y comunicar el 'error' que puede cometer o la confianza que deposita en sus decisiones.

Si la distribución de frecuencias de las observaciones puede asimilarse a la distribución de probabilidad teórica en la cual está basada la aplicación de la metodología inferencial elegida, entonces el investigador podrá estar seguro de que el error que informa en el proceso de prueba de hipótesis estadísticas planteadas o la confianza con que realiza sus estimaciones son correctos.

Aplicar cualquier metodología estadística inferencial sin estudiar a fondo el cumplimiento de los supuestos en los cuales ella está basada, lleva irremediablemente a conclusiones erróneas.

1.1 Ramas de la Estadística Inferencial:

Los dos principales procedimientos de la estadística inferencial son la estimación (puntual o por intervalos) y las pruebas de hipótesis. Comúnmente, los parámetros de una población son desconocidos, siendo necesario estimar el valor de éstos o, si no, efectuar indagaciones (pruebas de hipótesis) para comprobar si los valores a ellos atribuidos pueden ser considerados como verdaderos.

En esta Unidad veremos, entre otras cosas, cómo es posible obtener, para cada parámetro de interés, el **mejor estimador** de ese parámetro.

1.3 Población y muestra:

Una **población** se define como la totalidad de elementos sobre los cuales se desea estudiar un tema en particular. Por ejemplo, si se desea estudiar el ingreso promedio de las familias en la ciudad de San Nicolás, la población estará constituida por todas las familias que habitan en esta ciudad.

Es evidente que, por razones de costo y tiempo, sería casi imposible encuestadas a todas. Generalmente se encuesta a unas pocas seleccionadas de tal manera que representen al total de familias que componen la población en cuestión.

A ese conjunto de familias seleccionadas a partir de una cierta población, se lo denomina **muestra.**

El procedimiento que generalmente se sigue en cualquier investigación consiste en obtener resultados a partir de una muestra y luego generalizarlos a la población objetivo.

Una población cualquiera queda perfectamente especificada por ciertas medidas denominadas **parámetros poblacionales**.

Un parámetro poblacional es una medida que se calcula teniendo en cuenta todos los elementos que componen una cierta población.

Por ejemplo, si el ingreso promedio de las familias de la ciudad de San Nicolás se calcula considerando el ingreso de todas las familias que habitan en la ciudad, este ingreso promedio es un parámetro poblacional.

Es evidente que los parámetros poblacionales son generalmente imposibles de calcular. En la práctica, casi siempre se trabaja con muestras.

Las medidas calculadas a partir de las observaciones muestrales, se conocen con el nombre de **estadísticos muestrales**.

Un estadístico muestral es una medida que se calcula teniendo en cuenta solamente los elementos que integran una muestra determinada.

Así, si se toma una muestra de 100 familias de la ciudad de San Nicolás, se les realiza una entrevista en la que se pregunta el ingreso familiar y, en base a la información recogida se calcula un ingreso promedio, este promedio es un estadístico muestral.

Ahora bien, ¿para qué nos sirve calcular el ingreso promedio de las 100 familias que salieron seleccionadas en la muestra?

Nos sirve pues ella es la única información con la que contaremos para decir algo acerca de todas las familias de la ciudad de San Nicolás.

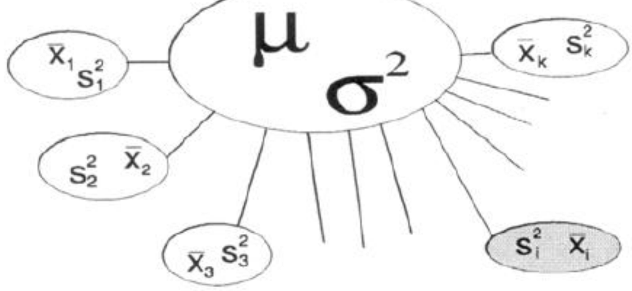
Debido a esto, la utilizaremos para estimar al parámetro poblacional (ingreso promedio de todas las familias de la ciudad de San Nicolás) que prácticamente nunca llegaremos a conocer.

El proceso por medio del cual se establecen relaciones entre los estadísticos muestrales y los parámetros poblacionales es el objeto de la **inferencia estadística**.

2. Estimación de parámetros

Como se dijo anteriormente, el objetivo de la estadística consiste en hacer inferencias acerca de los parámetros de una población teniendo en cuenta la información contenida en la muestra.

Ahora bien, como en general los parámetros poblacionales son desconocidos, existe una amplia gama de técnicas estadísticas que tienen como objetivo la estimación de estos parámetros a través de estadísticos muestrales adecuados a cada caso en particular.



La base teórica que sustenta la metodología que aplicamos a los resultados de una muestra, se fundamenta en la distribución de probabilidad del estimador calculado en cada una de las muestras posibles.

Existen dos tipos de estimaciones para parámetros, puntuales y por intervalo.

Una **estimación puntual** es un único valor estadístico y se usa para estimar un parámetro. El estadístico usado se denomina **estimador**.

Una **estimación por intervalo** es un rango, generalmente de ancho finito, que se espera que contenga el parámetro.

2.1 Estimadores

Un estimador es una función que permite calcular valores aproximados al del parámetro, además, se lo considera una variable aleatoria ya que puede tomar valores que pertenecen a un intervalo de números reales

Es posible definir muchos estadísticos para estimar un parámetro desconocido. Por ejemplo, puede elegirse la **media muestral** para estimar el valor de la **media poblacional**, o también la **mediana muestral**.

Por ejemplo, la media muestral se obtiene sumando todas las observaciones de la muestra y dividiendo esta suma por el tamaño de la muestra.

Cualquier persona podría definir otra combinación de las observaciones muestrales como estimador del parámetro μ y entonces cabría la pregunta:

¿Cuál es el "mejor" estimador de μ?

Un problema importante que debió resolver la teoría estadística, fue el de determinar el mejor estimador de cada parámetro en particular.

2.2 Propiedades de los estimadores

Cuando se analizan conceptos generales y métodos de inferencia es conveniente tener un símbolo genérico para el parámetro de interés. Se utilizará la letra griega θ para nombrar al parámetro y con $\hat{\theta}$ al estimador.

Entonces, ¿Cómo seleccionar un buen estimador de θ? , ¿Cuáles son los criterios para juzgar cuando un estimador es "bueno" o "malo"?

Si se piensa en términos de estimadores humanos como se encuentran en las grandes compañías, entonces, quizá un buen estimador es aquella persona cuyas estimaciones siempre se encuentran muy cercanas a la realidad.

De aquí surgen dos propiedades deseables de un estimador:

1. La distribución muestral de $\hat{\theta}$ debe tener una media igual al parámetro estimado θ, en este caso se dice que el **estimador es insesgado**.

   Si se usa la media muestral $\overline{X}$ para estimar la media poblacional μ, se sabe que la $\mu_{\overline{X}} = \mu$, por lo tanto la media es un estimador insesgado.

2. La varianza del estimador debe ser la menor posible, en este caso se dice que el **estimador es eficiente o con varianza mínima**.

Suponiendo que $\hat{\theta}_1$ y $\hat{\theta}_2$ son dos estimadores insesgados de θ. Entonces, aun cuando la distribución de cada estimador esté centrada en el valor verdadero de θ, las dispersiones de las distribuciones alrededor del valor verdadero pueden ser diferentes. Entre todos los estimadores $\hat{\theta}$ de θ que son insesgados, es necesario seleccionar al que tenga varianza mínima. En otras palabras, la eficiencia se refiere al tamaño de error estándar de la estadística.

Si comparamos dos estadísticas de una muestra del mismo tamaño y tratamos de decidir cuál de ellas es un estimador más eficiente, escogeríamos la estadística que tuviera el menor error estándar, o la menor desviación estándar de la distribución de muestreo.

Tiene sentido pensar que un estimador con un error estándar menor tendrá una mayor oportunidad de producir una estimación más cercana al parámetro de población que se está considerando. Como se puede observar las dos distribuciones tienen un mismo valor en el parámetro sólo que la distribución muestral de medias tiene una menor varianza, por lo que la media se convierte en un estimador eficiente e insesgado.

Es posible sumar a estas dos propiedades:

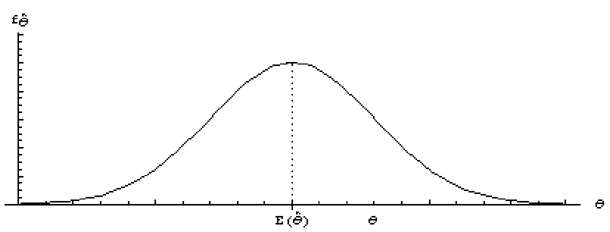
- **Coherencia:** Una estadística es un estimador coherente de un parámetro de población, si al aumentar el tamaño de la muestra se tiene casi la certeza de que el valor de la estadística se aproxima bastante al valor del parámetro de la población. Si un estimador es coherente se vuelve más confiable si tenemos tamaños de muestras más grandes.

- **Suficiencia:** Un estimador es suficiente si utiliza una cantidad de la información contenida de la muestra que ningún otro estimador podría extraer información adicional de la muestra sobre el parámetro de la población que se está estimando. Es decir se pretende que al extraer la muestra el estadístico calculado contenga toda la información de esa muestra. Por ejemplo, cuando se calcula la media de la muestra, se necesitan todos los datos. Cuando se calcula la mediana de una muestra sólo se utiliza a un dato o a dos. Esto es solo el dato o los datos del centro son los que van a representar la muestra. Con esto se deduce que si utilizamos a todos los datos de la muestra como es en el caso de la media, la varianza, desviación estándar, etc.; se tendrá un estimador suficiente.

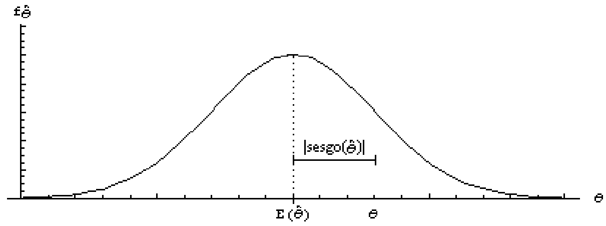
2.3 Interpretación gráfica de estas propiedades:

Supongamos que estamos estimando la media de una población normal. Es decir, la media de una población que sabemos que es normal, aunque no sepamos su media. Si como estimador de la media usamos, por ejemplo, alguna combinación lineal de los valores de una muestra tomada de esa población, entonces como el valor de cada valor de la muestra es una variable normal en si misma, y una combinación lineal de variables normales es una variable normal, nuestro estimador también es una variable aleatoria normal.

Si calculáramos como vimos antes el valor esperado del estimador y lo graficáramos, podríamos llegar a un gráfico como este:

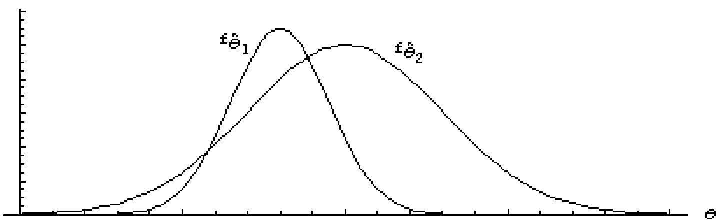


En este gráfico podemos apreciar que es deseable que el valor esperado del estimador coincida con el parámetro estimado. Denominamos sesgo a la diferencia $E(\hat{\theta})-\theta$ Por eso cuando el sesgo de un estimador es cero, se lo denomina "insesgado".



Como podemos observar, el estimador graficado no es insesgado. Por lo que dijimos antes, es deseable que el sesgo de un estimador sea pequeño.

Otra característica importante que analizamos fue la varianza. Es deseable que la varianza de un estimador sea pequeña, para que la variabilidad respecto de su valor esperado sea pequeña.



En el ejemplo graficado, la varianza de $\hat{\theta}_1$ es más pequeña que la de $\hat{\theta}_2$ Vemos que su variabilidad respecto de su valor esperado es menor.

Ejemplo 1

Consideremos una población compuesta por 5 escuelas rurales en las que se ha registrado el número de maestros obteniendo: 2, 3, 6, 8, 11.

Analizar cómo el estimador media muestral cumple con todas las propiedades de un buen estimador.

a. Insesgabilidad

En primer lugar calculamos la media aritmética y la varianza correspondiente a la variable X= cantidad de maestros por escuela. Se obtiene $\mu = 6$ maestros por escuela y $\sigma^2 = 10{,}8$ con una desviación típica de $\sigma = 3{,}29$

Estos resultados nos indican que el promedio de maestros en escuelas rurales en la población es de 6 maestros por escuela con una dispersión de 3,29 maestros por escuela.

Supongamos, ahora, que seleccionamos todas las muestras posibles de tamaño n = 2 por medio de un muestreo con reemplazo[1].

Como podemos observar en la Tabla 1, cada una de estas muestras es el resultado de un experimento aleatorio y todas tienen la misma probabilidad de ser seleccionada

[1] Se ha considerado incluir como primer ejemplo una situación sencilla utilizando un muestreo con reemplazo, más allá del contexto de la misma. La intención es estudiar en un principio el caso en que se incluyen todas las muestras posibles de tamaño n=2. En el caso de considerar el muestreo sin reemplazo, se cumple la insesgabilidad de la media aritmética. El valor de la varianza, en cambio, debe corregirse con el factor de corrección (N-n/N-1).

($1/N^n$ = 1/25). Luego, el muestreo es aleatorio y cada muestra es una muestra aleatoria simple.

En cada muestra podemos obtener la media aritmética, como vemos en la columna 4 de la Tabla 1.

Tabla 1: Muestras posibles y Promedios de cada muestra

Muestra	Observaciones en la muestra (x_i, x_i)	Probabilidad de cada muestra $p(x_i)$	Media muestra $\bar{x}_i$
1	(2,2)	1/25	2,0
2	(2,3)	1/25	2,5
3	(2,6)	1/25	4,0
4	(2,8)	1/25	5,0
5	(2,11)	1/25	6,5
6	(3,2)	1/25	2,5
7	(3,3)	1/25	3,0
8	(3,6)	1/25	4,5
9	(3,8)	1/25	5,5
10	(3,11)	1/25	7,0
11	(6,2)	1/25	4,0
12	(6,3)	1/25	4,5
13	(6,6)	1/25	6,0
14	(6,8)	1/25	7,0
15	(6,11)	1/25	8,5
16	(8,2)	1/25	5,0
17	(8,3)	1/25	5,5
18	(8,6)	1/25	7,0
19	(8,8)	1/25	8,0
20	(8,11)	1/25	9,5
21	(11,2)	1/25	6,5
22	(11,3)	1/25	7,0
23	(11,6)	1/25	8,5
24	(11,8)	1/25	9,5
25	(11,11)	1/25	11,0

Vemos que de esta forma, nos ha quedado definida una nueva variable aleatoria: la **variable aleatoria media muestral** (última columna de la tabla). El valor que ella toma, depende de la muestra a la que corresponda.

Como cada media aritmética está calculada con las observaciones muestrales, el valor obtenido en cada muestra será un estadístico muestral.

Ahora bien, por ser la media aritmética una variable aleatoria, podemos establecer su correspondiente distribución de probabilidad y calcular la esperanza matemática y varianza. Para eso, construiremos primero la tabla de frecuencias, computando los diferentes valores de xi y sus repeticiones.

Esto se presenta en las dos primeras columnas de la Tabla 2.

Las columnas restantes, resultan de asimilar las frecuencias relativas a las probabilidades (Teoría frecuencial de la probabilidad) y utilizarlas para obtener la Esperanza y la Varianza.

Tabla 2: Distribución de probabilidad de la variable

Aleatoria media muestral

$\bar{x}_i$	n_i	$p(\bar{x}_i)$	$x_i\ p(\bar{x}_i)$	$(\bar{x}_i - \bar{x})^2\ p(\bar{x}_i)$
2,0	1	1/25	2/25	16/25
2,5	2	2/25	5/25	24,5/25
3,0	1	1/25	3/25	9/25
3,5	0	0/25	0	0
4,0	2	2/25	8/25	8/25
4,5	2	2/25	9/25	4,5/25
5,0	2	2/25	10/25	P-125
5,5	2	2/25	11/25	0,5/25
6,0	1	1/25	6/25	0
6,5	2	2/25	13/25	0,5/25
7,0	4	4/25	28/25	4/25
7,5	0	0/25	0	0
8,0	1	1/25	8/25	4/25
8,5	2	2/25	17/25	12,5/25
9,0	0	0/25	0	0
9,5	2	2/25	19/25	24,5/25
10,0	0	0/25	0	0
10,5	0	0/25	0	0
11,0	1	1/25	11/25	25/25
	25	**1**	**150/25**	**135/25**

Se obtiene $E(\bar{x}) = \bar{x} = \sum_{i=1}^{N^n} \bar{x}_i \,.\, p(\bar{x}_i) = 6$ La primera conclusión importante que hemos obtenido es $E(\bar{x}) = \mu$ y esto equivale a decir que $\bar{x}$ **es un estimador insesgado de** μ

Sin embargo, es oportuno aclarar aquí que la propiedad de insesgamiento es un concepto teórico. Únicamente se da en términos de valores esperados, puesto que si nos fijamos en la Tabla 1 encontraremos que, de las 25 muestras posibles, sólo una media muestral coincide con el valor del parámetro μ.

Calculemos ahora la varianza de la variable aleatoria media muestral utilizando también la información proporcionada por la tabla.

Obtenemos $V(\bar{x}) = \sum_{i=1}^{N^n} (\bar{x}_i - \bar{x})^2 \,.\, p(\bar{x}_i) = 5{,}40$ Vemos que este valor es exactamente igual al de la varianza poblacional dividido por el tamaño de la muestra, es decir:

$$V(\bar{x}) = 5{,}40 = \frac{10{,}80}{2} = \frac{\sigma^2}{n}$$

La segunda conclusión importante a la que llegamos es que la varianza de la variable aleatoria media muestral es directamente proporcional a la varianza poblacional e inversamente proporcional al tamaño de la muestra. Esto quiere decir que a medida que se incremento el tamaño de la muestra menor es la variabilidad de la variable media muestral, mientras que, cuanto más variable es la característica en estudio en la población (expresada por una mayor varianza σ^2), mayor será también la varianza de la variable aleatoria media muestral.

b. Insesgabilidad de mínima varianza (eficiencia)

Veamos ahora si la media muestral es un estimador eficiente. Para verificar esta propiedad debemos comparar la varianza de esta variable aleatoria con la de algún otro estimador insesgado de μ.

Cuando la distribución de la característica en estudio en la población es perfectamente simétrica[2], la mediana también es un estimador insesgado de μ.

Ahora bien, la varianza del estimador mediana muestral es $V(Me) = \frac{\sigma^2}{n} . \frac{4}{\pi}$ (verificar), lo que nos permite verificar que siempre $V(Me) > V(\bar{x})$

Con ello verificamos que la media muestral es un estimador más eficiente que la mediana para estimar al parámetro poblacional μ. En símbolos:

$$\frac{V(Me)}{V(\bar{x})} = \frac{\frac{\sigma^2}{n} . \frac{4}{\pi}}{\frac{\sigma^2}{n}} = \frac{4}{\pi} > 1$$

En general, podemos verificar que la media muestral es un estimador que tiene varianza mínima cuando se lo compara con cualquier otro estimador del parámetro media poblacional.

c. Consistencia

En cuanto a la propiedad de consistencia, ella no necesita verificación. Es evidente que si la media muestral es un estimador insesgado también será consistente.

d. Distribución asintóticamente normal

Analizaremos ahora si el estimador media muestral cumple con la propiedad 4, es decir, tiene distribución asintóticamente normal.

Evidentemente hay dos situaciones posibles:

a) Cuando la muestra se selecciona aleatoriamente de una población **con distribución normal**.

b) Cuando la muestra se selecciona aleatoriamente de una población **sin distribución normal**.

En el caso a) la teoría estadística establece que la distribución de la variable aleatoria media muestral, calculada en base a una muestra seleccionada aleatoriamente de una población con distribución normal, responde a las características de la distribución normal de la población de origen.

[2] Comprobar que es así.

En el caso b) la teoría estadística establece que, aún cuando la distribución poblacional de la característica en estudio se aleja bastante de la forma normal, la distribución de la variable aleatoria media muestral se aproxima a la distribución normal a medida que se incremento el tamaño de la muestra.

Esta última afirmación está sustentada en un importante teorema de la teoría estadística conocido con el nombre de **teorema central del límite**.

Luego es fácil concluir que el estimador media muestral cumple con la propiedad 4.

Conclusiones:

Ahora estamos en condiciones de efectuar la siguiente síntesis sobre el ejemplo planteado: el estimador media muestral es un estimador insesgado, eficiente, consistente y asintóticamente normal por lo cual cumple con todas las propiedades de ser un buen estimador.

Actividad 1:

En base al Censo de Población y Vivienda, una empresa que brinda servicios de marketing ha listado las 50 viviendas de un radio censal, considerando la cantidad de personas que habitan en cada una de ellas.

Los resultados obtenidos fueron:

Vivienda N	Cant. de personas	Vivienda N	Cant. de personas
00	4	25	8
01	2	26	4
02	6	27	1
03	4	28	5
04	3	29	2
05	1	30	6
06	7	31	4
07	3	32	6
08	5	33	4
09	2	34	2
10	4	35	5
11	5	36	3
12	9	37	2
13	4	38	6
14	3	39	2
15	5	40	1
16	4	41	4
17	3	42	5
18	1	43	4
19	4	44	3
20	5	45	7
21	6	46	4
22	3	47	5
23	8	48	2
24	3	49	4

En relación a la situación planteada:

a) Obtenga 10 muestras con reemplazo de tamaño n = 3 y luego 10 muestras de tamaño n = 7. Utilice algún procedimiento aleatorio de selección, ya sea colocando 50 papelitos numerados (desde 00 hasta 49) en una caja y sacando luego hasta completar el tamaño de cada muestra o por algún otro procedimiento aleatorio como el uso de tablas de números al azar.
b) Calcule en cada muestra las correspondientes medias muéstrales. Construya una tabla de distribución de frecuencias para cada tamaño de muestra (una para muestras de n = 3 y otra para muestras de n = 7).
c) Compare la forma de las dos distribuciones.
d) Comente los resultados obtenidos.

3. Estimación puntual

La estimación puntual es un proceso mediante el cual se estima el parámetro en un punto, dando un valor específico como estimación.

Los dos métodos tradicionales de estimación puntual de parámetros son el de **mínimos cuadrados** y el de **máxima verosimilitud**.

El método de mínimos cuadrados fue trabajado en la Unidad Nº2 cuando hablamos del análisis de regresión.

El objetivo de la estimación puntual es seleccionar sólo un número, basados en datos de la muestra, que represente el valor más razonable de θ

Una **estimación puntual** de un parámetro θ es un sólo número que se puede considerar como el valor más razonable de θ. La estimación puntual se obtiene al seleccionar un estadístico apropiado y calcular su valor a partir de datos de la muestra dada. El estadístico seleccionado se llama **estimador puntual** de θ.

Los estimadores más frecuentes de los siguientes parámetros son:

Parámetro	Estimador más probable
μ	$\bar{x}$, la media muestral
σ^2 o σ	s^2 o s, la varianza muestral o desviación estándar muestral
ρ	$\hat{\rho}$ =X/n, la proporción muestral, donde x es el número de objetos en una muestra aleatoria de tamaño n que pertenece a la clase de interés
$\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2$, la diferencia de medias muestrales de dos muestras aleatoria independientes
$\rho_1-\rho_2$	$\hat{\rho}_1-\hat{\rho}_2$, la diferencia entre las proporciones de dos muestras aleatorias independientes

En el mejor de los casos, se encontrará un estimador $\hat{\theta}$ para el cual $\hat{\theta}=\theta$ siempre. Sin embargo, $\hat{\theta}$ es una función de los x_i muestrales, por lo que es en sí misma, una variable aleatoria. Entonces, $\hat{\theta}=\theta+error\,de\,estimacion$ lo cual nos permite deducir que el estimador preciso sería uno que produzca sólo pequeñas diferencias de estimación, de modo que los valores estimados se acerquen al valor verdadero.

La distancia entre una estimación y el parámetro estimado recibe el nombre de *error de estimación.*

Para cualquier estimador puntual con una distribución normal, la regla empírica dice que aproximadamente el 95% de todas las estimaciones puntuales estarán a no más de dos (exactamente 1,96) desviaciones estándar de la media de esa distribución.

Ejemplo 2

La longitud de los tornillos que produce una determinada máquina es una variable normal, pero no sabemos cuánto vale el parámetro (media poblacional) μ de esa distribución normal.

Podemos hacer el experimento de tomar 10 tornillos, calcular el promedio de sus longitudes, y usar ese promedio como estimación de μ. En este caso el estimador es:

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

El margen de error se estima como

$$\pm 1{,}96\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Ejemplo 3

En el futuro habrá cada vez más interés en desarrollar aleaciones de Mg de bajo costo, para varios procesos de fundición. En consecuencia, es importante contar con métodos prácticos para determinar varias propiedades mecánicas de esas aleaciones. Examine la siguiente muestra de mediciones del módulo de elasticidad obtenidas de un proceso de fundición a presión 44.2 43.9 44.7 44.2 44.0 43.8 44.6 43.1

Suponga que esas observaciones son el resultado de una muestra aleatoria. Se desea estimar la varianza poblacional σ^2

Un estimador natural es la varianza muestral[3]

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(44{,}2-44{,}0625)^2 + (43{,}9-44{,}0625)^2 + \ldots + (43{,}1-44{,}0625)^2}{8-1} = 0{,}251$$

Este valor nos permite encontrar el margen de error que se comete al estimar el valor de la media muestral.

[3] En "Probabilidad y Estadística I", Unidad Nº1, punto 4.3 Medidas de variabilidad, pág. 36. Se habló de este coeficiente de corrección utilizado.

4. Estimación por intervalos

Un estimado puntual, por ser un sólo número, no proporciona por sí mismo información alguna sobre la precisión y confiabilidad de la estimación.

Por ejemplo, si se tiene interés en estimar la resistencia promedio a la ruptura de cierto elemento estructural, es probable que un solo número no sea tan significativo como un intervalo, dentro del cual se espera encontrar el valor de este parámetro.

Una alternativa para reportar un solo valor del parámetro que se esté estimando es calcular e informar todo un intervalo de valores factibles, un *estimado de intervalo o intervalo de confianza (IC)*.

Una estimación por intervalos de un parámetro desconocido θ es un intervalo de la forma $I \leq \theta \leq u$, donde los puntos extremo l y u dependen del valor numérico de la estadística $\hat{\theta}$ para una muestra en particular y de la distribución de muestreo de $\hat{\theta}$.

De la distribución de muestreo de $\hat{\theta}$ es posible determinar los valores de l y u tales que la siguiente proposición sea verdadera:

$$P(I \leq \theta \leq u) = 1-\alpha \qquad ; \qquad 0 < \alpha < 1$$

Por tanto se tiene una probabilidad de $1-\alpha$ de seleccionar una muestra que produzca un intervalo que contiene el valor verdadero de θ El intervalo resultante $I \leq \theta \leq u$ se conoce como **Intervalo de Confianza del** $100(1-\alpha)$ **por ciento.**

Las cantidades l y u se denominan límites de confianza inferior y superior y $1-\alpha$ es el **coeficiente de confianza** (o ***nivel de confianza***, que es una medida del grado de fiabilidad en el intervalo).

De tal forma, cuando $\alpha = 0{,}05$, se tiene un intervalo de confianza del 95% y cuando $\alpha = 0{,}01$ se tiene uno del 99%. Entre mayor es el intervalo de confianza se tiene más seguridad de que el mismo contenga el parámetro desconocido.

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

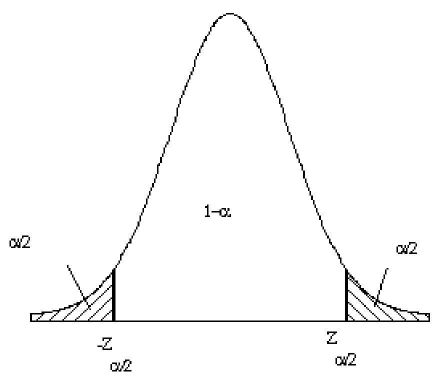
Un intervalo del tipo $I \leq \theta \leq u$ recibe el nombre más apropiado de **Intervalo de Confianza Bilateral**. También existen intervalos de confianza Unilaterales $\theta \geq I$ y $\theta \leq u$ donde los límites de confianza se eligen de modo que $P(\theta \geq I) = 1-\alpha$ y $P(\theta \leq u) = 1-\alpha$

4.1 Intervalos de confianza para las medias

Para estimar la media μ de una característica de la población, es necesario, primero saber si la varianza de la población es conocida o no lo es.

Para estimar la media μ de una característica de la población, cuando se considera conocida la varianza de esa población, se toma una muestra de tamaño "n" y se le calcula la media muestral $\overline{X}$

Por el Teorema del Límite Central, se sabe que de la distribución de la media muestral se obtiene que



$Z=\dfrac{\overline{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$ tenga una distribución como una normal estándar, con media $E(\overline{X})=\mu$ y varianza $\dfrac{\sigma^2}{n}$.

Teniendo en cuenta que la distribución de la media muestral sigue o tiende a una Distribución Normal, y considerando la varianza como conocida, el intervalo de confianza debe abarcar el área de $(1-\alpha)\%$ entre sus límites superior e inferior en dicha distribución. Cada límite es expresado en unidades de desviación típica y esas unidades esta expresadas por $Z_{a/2}$. En consecuencia, el intervalo de confianza bilateral del $100\,(1-\alpha)\%$ para μ dado esta dado por $I=\left[\overline{X}-Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}};\overline{X}-Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right]$

Tener en cuenta que esta expresión obedece al concepto:

Valor del parámetro=estimación puntual$\pm$una función de la confianza y la dispersión (directamente proporcionales) y del tamaño de la muestra (inversamente proporcionales).

Ejemplo 4

Supongamos que el director de investigaciones de mercado de una fábrica automotriz necesita hacer una estimación de la vida promedio de las baterías que su compañía produce. Selecciona aleatoriamente 200 usuarios y resulta tener una vida promedio de sus baterías de 36 años.

Encontrar el intervalo dentro del cual es probable que esté la media de población desconocida.

Para ello, es necesario encontrar el error estándar de la media $\sigma_{\overline{X}}=\dfrac{\sigma}{\sqrt{n}}$, conociendo que la desviación de la población es 10 meses, obtenemos $\sigma_{\overline{X}}=0{,}707$

Ahora podemos decir que la vida útil de las baterías esta dentro del intervalo $I=(\overline{X}-\sigma_{\overline{X}};\overline{X}+\sigma_{\overline{X}})$, es decir, $I=(35{,}293;36{,}707)$

Aunque estos datos son útiles no son suficientes, pues no tiene así un nivel de confianza significativo. Para esto debemos recordar que cuando trabajamos con la distribución normal de probabilidad, hemos visto que posiciones específicas del área bajo la curva normal están localizadas a una distancia de cierto número dado de desviaciones estándares por debajo y por arriba de la media. Esto se puede aplicar al error estándar de la media.

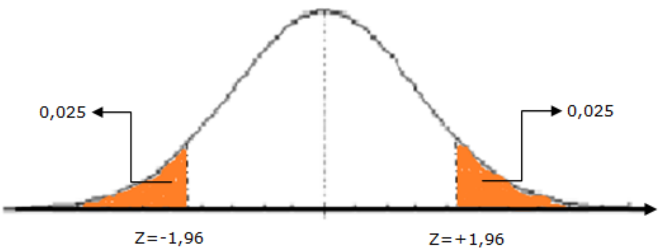
Así podemos decir que, por ejemplo, el 95,5% de las medias de muestra están dentro de $\pm 2\sigma$ de μ, y en consecuencia, μ está dentro de $\pm 2\sigma$ de la media de cada una de tales muestras. De manera parecida, también podemos decir que la probabilidad de que la media de la muestra esté dentro de $\pm\sigma$ de la media de la población μ es de 0,683.

Ejemplo 5

Un vendedor de partes del automotor, mayorista, necesita una estimación de la vida media que puede esperar de los limpiadores de parabrisas en condiciones normales de manejo. La administradora ya ha determinado que la desviación estándar de la vida útil de la población es de 6 meses. Se selecciona una sola muestra de 100 limpiadores y se obtienen que $\overline{x} = 21 meses$, encontrar el intervalo de confianza de un 95%

Como la muestra es mayor de 30, debemos calcular el error estándar de la media $\sigma_{\overline{X}} = 0{,}6$. Ahora, considerando un nivel de confianza del 95%, podemos obtener Z de la siguiente manera $\alpha = 100 - 95 = 5 \Rightarrow \frac{\alpha}{2} = 2{,}5 \Rightarrow \frac{2{,}5}{100} = 0{,}025$

Con este valor planteamos que $\begin{array}{l} F(Z) = 1 - F(-Z) \\ F(Z) = 1 - 0{,}025 = 0{,}975 \end{array}$ este valor lo buscamos en tabla y obtenemos Z=1,96.



Con este valor calculamos el intervalo de confianza en el que puede estar la media poblacional

$$I = \left[\overline{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} ; \overline{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = [19{,}824; 22{,}176]$$

Ejemplo 6

Se quiere estimar el ingreso medio anual de 700 familias que vive en una sección de 4 manzanas. Si se toma una muestra de 50 familias y se hallan los siguientes resultados $\overline{x} = \$11800$ y $s = \$950$ Encontrar un intervalo con un nivel de confianza del 90% en el que pueda encontrarse la media poblacional.

Calculo el error producido en la desviación estándar de la media $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 129{,}56$ [4]

Con el valor del 90% de nivel de confianza obtenemos que $\alpha = 100 - 90 = 10 \Rightarrow \frac{\alpha}{2} = 5 \Rightarrow \frac{5}{100} = 0{,}05$ De igual manera que lo hicimos antes, obtenemos Z=1,64.

Entonces $I = \left[\overline{X} - Z_{\alpha/2} \sigma_{\overline{X}} ; \overline{X} - Z_{\alpha/2} \sigma_{\overline{X}} \right] = [11587{,}52; 12012{,}48]$

---

[4] Como se trata de un muestreo sin reemplazo se ha utilizado para su cálculo el coeficiente de corrección.

Conclusiones:

Como síntesis, podemos decir que para reducir la amplitud de un intervalo de confianza y en consecuencia aumentar su precisión, debemos reducir el error estándar de la media muestral $\bar{x}$ *que es* σ/n*. Esto puede lograrse solamente disminuyendo la variabilidad de los datos ya sea homogeneizando el material experimental o, si esto no puede llevarse a cabo, aumentando el tamaño de la muestra.*

Es costumbre utilizar coeficientes de confianza del 90%, 95% y 99% Por este motivo es posible considerar la siguiente tabla que resume los valores de probabilidad de la distribución normal estandarizada para estos niveles de confianza.

Coeficiente de confianza	z
0,90	1,645
0,95	1,960
0,99	2,576

¿Por qué "intervalos de confianza" y no "de probabilidad"?

Si observamos las expresiones de los intervalos obtenidos para la media poblacional, de un 95% de confianza o de un 99% de confianza, se puede apreciar que en ellos no está implicada ninguna variable aleatoria, ya que en el centro del intervalo está un parámetro μ y en los extremos se tienen números obtenidos sumando y restando $Z\sigma/\sqrt{n}$ a la media obtenida en la muestra.

Al no existir variable aleatoria en la expresión, no sería correcto hablar de probabilidad. No obstante, se comprende que, cumpliéndose los supuestos (distribución normal de la población, o muestras suficientemente grandes), construyendo los intervalos con ésta metodología (sumar y restar $Z\sigma/\sqrt{n}$ a la media muestral), el 95% o 99% (según el valor de Z) de los posibles intervalos contendrán al verdadero valor de μ. De allí la expresión "existe entre un 95 (o 99) por ciento de confianza de que el intervalo contenga al parámetro".

Actividad 2:

Algunas publicaciones médicas aseguran que la vitamina C puede resultar de utilidad en la reducción del colesterol de las paredes internas de las arterias, disminuyendo así la posibilidad de ataques cardíacos. Para corroborar esta afirmación, un investigador médico observó el nivel de colesterol en 50 pacientes (todos ellos registraban niveles de colesterol mayores a los normales y, en consecuencia, conformaban un grupo de riesgo al infarto).

Durante un mes sometió a estos pacientes a una dosis diaria de 500 mg. de vitamina C.

Al finalizar el tratamiento, volvió a medir los niveles de colesterol de cada paciente registrando la disminución observada en cada uno de ellos. De estos registros obtuvo un promedio de $\bar{x}$ = 64,3 mg. por 100 ml. con una desviación estándar s = 18,9 mg. por 100 ml.

Con la información proporcionada por la muestra de pacientes seleccionados, estime por intervalos el verdadero valor de la disminución promedio de colesterol en la población objetivo considerada.

Actividad 3:

La Sección de Pruebas de una fábrica automotriz decide estudiar el consumo de nafta en autos de un nuevo modelo en experimentación.

Para ello realiza la experiencia de recorrer 100 Km. 30 veces ($n = 30$) registrando en cada prueba el consumo de combustible.

Una vez finalizado el experimento, calcula el promedio de consumo en las 30 pruebas obteniendo una media $\overline{x} = 10{,}8$ litros con una desviación estándar $s = 1{,}4$ litros.

Con estos resultados muéstrales, estime el verdadero consumo promedio de nafta en autos de este modelo en particular.