Inferencia Estadística para Economía y Negocios
Sesión I

Dr. Mauricio Huerta Aguiar

Facultad de Ciencias Sociales y Económicas
Departamento de Economía y Administración
Ingeniería Comercial

Outline

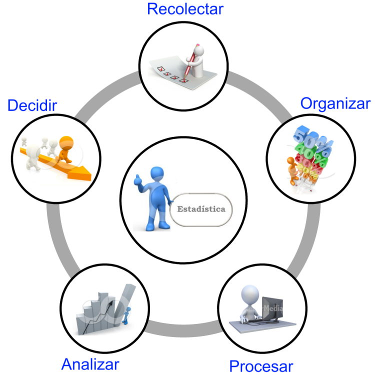
1. Nociones básicas
2. Inferencia estadística
3. Propiedades de los estimadores puntuales
4. Método de estimación de verosimilitud máxima
5. Distribuciones de muestreo
6. Distribuciones de muestreo para la media
7. Distribuciones de muestreo para la varianza
8. Intervalos de confianza
9. Método para obtener un IC
10. Intervalos de confianza para la media

Nociones Básicas

"Algún día
el pensamiento estadístico
será para un ciudadano eficiente
tan necesario como
habilidad para leer y escribir"
(H. G. Wells, 1904)

Introducción

La **estadística** es también conocida como la **ciencia de los datos** o **ciencia estadística** y comprende un conjunto de métodos que permiten recolectar, organizar y procesar datos, para después analizar los resultados obtenidos transformándolos en información, como un apoyo eficiente a la toma de decisiones.



- **Unidad de análisis (UA)**: entidad que proporciona un dato.
- **Variable estadística**: característica de interés a ser estudiada, la que puede ser medida u observada por el investigador en la UA.
- **Medición u observación:** es el proceso que asigna un valor de la variable a una UA.
- **Dato:** es el valor de la variable generado mediante su medición u observación en una UA.
- **Error de medición:** es la diferencia entre el valor exacto que posee la UA y el valor obtenido en la medición.

Conceptos estadísticos

- **Población**: es el universo o conjunto total de UAs. Es sobre este conjunto para quien se obtendrán las conclusiones finales.
- **Muestra**: es cualquier subconjunto de la población.
- **Muestra aleatoria**: es un subconjunto de la población que posee un grado de representatividad adecuado.

Conceptos estadísticos

- **Censo**: estudio que involucra a toda la población. No contiene margen de error (o error de precisión), pero es costoso.
- **Muestreo**: estudio que involucra a una muestra. Contiene error de muestreo, pero es menos costoso.
- **Sesgo**: componente de un estudio que hace que la muestra no sea representativa, pues distorsiona las conclusiones del estudio e impide que se lleve a cabo un proceso inferencial.
- **Parámetros**: información relacionada a la población y que es de interés para el investigador.
- **Estimadores**: aproximaciones de los parámetros basadas en una muestra aleatoria.

Clasificación de los métodos estadísticos

Los métodos estadísticos pueden clasificarse en dos:

1. Descriptivos.
2. Inferenciales.

Métodos descriptivos

Se preocupan de describir el conjuntos de datos. Generalmente están relacionados al análisis preliminar o exploratorio de los datos. En el caso de un censo son los únicos métodos que se pueden aplicar.

Métodos inferenciales

Son los encargados de llevar adecuadamente los resultados de una muestra aleatoria hacia la población objetivo.

Problemas a resolver con la estadística

- Evaluar el grado de satisfacción de los clientes.
- Identificar los factores que influyen en el endeudamiento.
- Comparar la lealtad de los clientes en una empresa.
- Evaluar el efecto de la publicidad en la conducta de compra.

Todos estos problemas son distintos, pero hay una línea general de razonamiento que es común a todos ellos llamada **metodología estadística**.

- Estudio mediante un **muestreo**

Encuesta CASEN 2024

(o Encuesta de **Ca**racterización **S**ocio-**E**conómica **N**acional)

Desarrollada entre Noviembre 2024 a Enero 2025.

- Estudio mediante un **censo**

Censo 2024

Finalizada el 31 de julio de 2024.

Encuesta CASEN

Es una encuesta a nivel nacional, regional y comunal, realizada por el Ministerio de Desarrollo Social y Familia desde 1985, con un periodicidad bienal y trienal. Los años en que se ha realizado esta encuesta son: 1985, 1987, 1990, 1992, 1994, 1996, 1998, 2000, 2003, 2006, 2009, 2010, 2011 (debido al terremoto), 2013, 2015, 2017, 2020 (debido a la pandemia) y 2024.

Objetivo general del estudio:
Medir las condiciones socioeconómicas de los hogares del país, en términos de acceso a la salud, la educación, el trabajo y la vivienda.

CASEN-2024 proporcionó información sobre temas de educación, trabajo, salud, vivienda, bienes ys ervicios, entre otros, para establecer políticas públicas. Ver más detalles en:

- `https://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-2024`
- `https://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2024/proceso_diseno_cuestionario_casen_2024.pdf`

CASEN

Casen: Pobreza según ingresos cae del 10,7% al 6,5% en primera medición post pandemia

Según los datos del sondeo presentado este jueves, el 2,0% de la población se sitúa en pobreza extrema y un 4,5% en pobreza no extrema.

27 de Julio de 2023 | 13:01 | Redactado por Daniela Toro, Emol.

Imagen referencial.

Aton / Archivo.

Censo de población en Chile

En Chile se realizan censos de población de manera más o menos sistemática desde mediados del siglo XIX, aunque anteriormente, en el periodo colonial, existieron otros intentos de empadronar a los habitantes del país. A partir de 1970, el Instituto Nacional de Estadísticas (INE) es el organismo encargado de realizar los censos a nivel nacional.

Objetivo general del estudio:
Determinar el número de habitantes que componen un país.

Un **censo** proporciona información principalmente acerca del **tamaño de la población chilena** para establecer políticas públicas. Ver más detalles en:

- `www.censo.cl`
- `https://es.wikipedia.org/wiki/Censo_chileno_de_2024`

Censo 2012

El XVIII Censo Nacional de Población y VII de Vivienda o Censo de Población y Vivienda 2012 se realizó en Chile entre el 9 de marzo y el 31 de julio de 2024.

Objetivo:
Actualizar los datos registrados en el censo realizado en abril de 2002.

El **CENSO-2012** estuvo a cargo del INE de Chile. El 2 de abril de 2013 se publicaron los resultados arrojando un total nacional de 16.634.603 habitantes. En septiembre de 2014 el INE publicó una nueva cifra en base a proyecciones del Censo de 2002, arrojando un total estimado de 17.819.054 habitantes. La información obtenida del censo 2012 se descartó debido a una serie de problemas en la obtención y tratamiento de los datos. Estos problemas de solucionaron en su versión del 2024.

LATERCERA Política | Nacional | Mundo | Negocios | Opinión | Santiago | Tendencias | Educación | Cultura | Entretención | Deportes | La Tercera TV

Nacional

Las principales fallas del Censo 2012 y las recomendaciones de la Comisión Externa del INE

- **La comisión criticó el escaso tiempo de preparación, el bajo presupuesto, el clima al interior del organismo, a la anterior administración, la imputación de no censadas y algunos errores en el cuestionario.**
- **Además de la realización de un censo abreviado, entre las recomendaciones destacan el no usar los datos del Censo 2012, crear una cifra provisoria para 2013 y fortalecer el INE.**

por La Tercera - 08/08/2013

- Estudio mediante un **muestreo**

Encuesta Adimark 2015

La serie mensual de encuestas Adimark de Evaluación de Gobierno se inicia en marzo de 2006, justo al comenzar el primer gobierno de Michelle Bachelet.

Objetivo general del estudio:
Evaluar la gestión del gobierno de Chile en relación al presidente, sus ministros, las coaliciones, el Congreso y la reformas, entre otros.

Adimark 2015

LATERCERA Política | Nacional | Mundo | Negocios | Opinión | Tendencias | Cultura | Entretención | Deportes | Finde | La Tercera TV

Política

Adimark: Aprobación de Bachelet cae a un 39% tras polémica por caso Caval

A pesar de la leve alza en el mes anterior, en el sondeo de febrero el apoyo a la Mandataria cayó cuatro puntos, cifrándose como la más baja desde que asumió el gobierno. Su desaprobación en tanto alcanzó un 52%. por Rosario Alvarez - 03/03/2015 - 09:02

Metodología

TIPO DE ESTUDIO:

Estudio cuantitativo con aplicación de encuestas a teléfonos de red fija y celulares mediante sistema CATI. La muestra es probabilística con selección aleatoria de hogares y de entrevistados, con utilización de doble marco muestral:1) Personas residentes en hogares conectados a red fija y 2) Personas con acceso a teléfonos celulares. Se utilizó un cuestionario estructurado.

GRUPO OBJETIVO:

Hombres y Mujeres, mayores de 18 años.

UNIVERSO DE LA MUESTRA:

El universo está compuesto por mayores de 18 años con acceso a teléfono de red fija o celulares en los principales centros urbanos de las 15 regiones del país.

El error se estima en aproximadamente +/- 3.0% con un 95% de confianza.

TAMAÑO DE LA MUESTRA: 1.043 casos

Teléfonos de red fija: 793 casos – Celulares: 250 casos

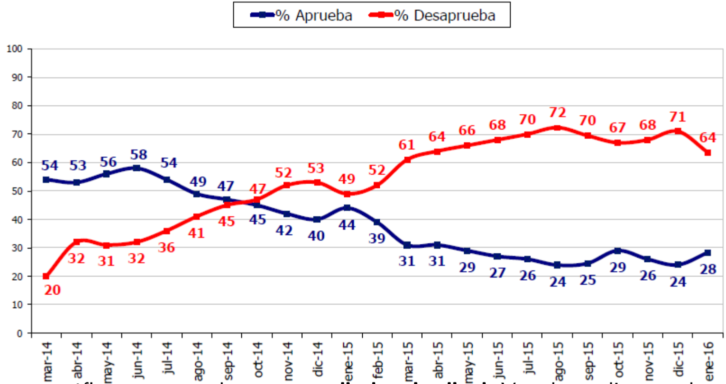
Santiago: 512 casos - Regiones: 531 casos

Los datos fueron ponderados por ciudad, sexo, edad y NSE, según datos del Censo 2002.

FECHA DE TERRENO: Entre el 05 y el 30 de agosto de 2014.

Adimark 2015: análisis de 1 pregunta (variable) de la encuesta

Independiente de su posición política, ¿Usted aprueba o desaprueba la forma como Michelle Bachelet está conduciendo su gobierno?



Note que este gráfico corresponde a un **estudio longitudinal**. Ver el estudio completo en:

`www.adimark.cl/es/estudios/documentos/6_eval%20gobierno%20ago_2014.pdf`

Outline

1 Nociones básicas

2 Inferencia estadística

3 Propiedades de los estimadores puntuales

4 Método de estimación de verosimilitud máxima

5 Distribuciones de muestreo

6 Distribuciones de muestreo para la media

7 Distribuciones de muestreo para la varianza

8 Intervalos de confianza

9 Método para obtener un IC

10 Intervalos de confianza para la media

Inferencia Estadística

Introducción

La estadística trata de responder tres preguntas básicas:

(i) ¿Cómo debo recolectar los datos?
(muestreo para generar datos)

(ii) ¿Cómo debo analizar y resumir los datos recopilados?
(métodos estadísticos para generar información)

(iii) ¿Qué tan precisa es la información obtenida desde los datos?
(inferencia estadística)

La **inferencia estadística** está formada por métodos utilizados para **apoyar la tomar decisiones** basados en **evidencia científica** o para obtener conclusiones sobre una población. Estos métodos utilizan la información contenida en una muestra representativa de la población para obtener tales conclusiones. La inferencia estadística puede dividirse en dos partes:

- Estimación de parámetros.
- Pruebas de hipótesis.

Inferencia estadística

Es el conjunto de métodos estadísticos que permiten deducir (inferir) cómo se distribuye una población bajo estudio o las relaciones estocásticas entre varias variables de interés a partir de la información proporcionada por una muestra.

- **Muestreo aleatorio simple:** es un procedimiento de selección en que todas las UAs de la población tienen la misma probabilidad de ser elegidas.

Una **muestra aleatoria simple** de tamaño n de una variable aleatoria (VA) X, con función de distribución F_X, es un conjunto de n VAs

$$X_1, \ldots, X_n$$

que son independientes e idénticamente distribuidas (IID) de acuerdo a F_X.

- **Parámetro**: éste es denotado por θ y corresponde a información relacionada a la población y que es de interés para el investigador. θ es cualquier característica medible de la distribución de la VA en estudio (media, varianza, etcétera). Así, un parámetro es una caracterización numérica de la distribución de la población, de manera que θ describe su función de distribución.
- **Espacio paramétrico**: conjunto de todos los posibles valores de un parámetro, denotado por Θ, esto es, $\theta \in \Theta$.
- **Estadístico:** es una función $T(X_1, \ldots, X_n)$ de la muestra que no contiene parámetros desconocidos. Por tanto, un estadístico es una VA, cuya distribución se llama **distribución de muestreo**.
- **Estimador**: es un estadístico que aproxima al parámetro y se denota por $\widehat{\theta} = T(X_1, \ldots, X_n)$.

Estimador puntual

Un **estimador** puntual $\widehat{\theta}$ de un parámetro θ es una función de la muestra aleatoria $X_1, \ldots, X_n$ que proporciona un valor único (o puntual). Un estimador puntual es en sí mismo una VA y, por consiguiente, tiene una distribución de muestreo. Un valor observado (realización) de un estimador se denomina **estimación** puntual del parámetro θ.

Error estándar

Sean $\widehat{\theta}$ un estimador puntual de θ y $\mathsf{Var}(\widehat{\theta})$ la varianza de este estimador. Una medida de la precisión de tal estimador es el error estándar (EE) de $\widehat{\theta}$ definido como

$$\mathsf{EE}(\widehat{\theta}) = \sqrt{\mathsf{Var}(\widehat{\theta})}.$$

Ejemplo (tiempo entre llegadas a la caja de un banco)

Considere una muestra aleatoria de tamaño $n = 2$ de una VA X con distribución exponencial de parámetro θ. Esto quiere decir que X_1 y X_2 son VAs independientes y con función de densidad común dada por

$$f_{X_i}(x) = \theta \exp(-\theta x) I_{]0,\infty]}(x), i = 1, 2.$$

Note que:

- El parámetro es θ y su espacio paramétrico es $\Theta =]0, \infty]$
- Un posible estadístico relacionado a θ es $T = T(X_1, X_2) = X_1 + X_2$ y un estimador de θ puede ser $\widehat{\theta} = \frac{X_1+X_2}{2}$.

Se pide:

- Hallar la distribución de muestreo de $\widehat{\theta}$.
- Encontrar el error estándar de $\widehat{\theta}$.

Ejemplo (continuación)

En relación al ejemplo anterior, considere que X es el tiempo entre llegadas (en minutos) de los clientes a la caja de un banco. Suponga que la población (para efectos pedagógicos) está constituida por los siguientes $N = 4$ tiempos entre llegadas: $\mathcal{P} \equiv \{0.4, 0.6, 0.8, 0.3\}$. Se pide:

- Hallar la distribución de muestreo empírica de $\widehat{\theta} = \frac{X_1+X_2}{2}$.
- Encontrar el error estándar empírico de $\widehat{\theta}$.
- Comparar los resultados teóricos con los empíricos. ¿Por qué difieren?

Outline

1. Nociones básicas
2. Inferencia estadística
3. Propiedades de los estimadores puntuales
4. Método de estimación de verosimilitud máxima
5. Distribuciones de muestreo
6. Distribuciones de muestreo para la media
7. Distribuciones de muestreo para la varianza
8. Intervalos de confianza
9. Método para obtener un IC
10. Intervalos de confianza para la media

Propiedades de los Estimadores Puntuales

Hasta ahora hemos definido el estimador de un parámetro. Entonces, la pregunta que surge es: ¿existen estimadores mejores que otros en algún sentido? A continuación definimos algunas propiedades que un estimador puede poseer o no, que nos ayudarán a decidir cuándo en estimador es mejor que otro.

Sesgo

Sea $\widehat{\theta}$ un estimador del parámetro θ. Así, el sesgo del estimador $\widehat{\theta}$ se define como

$$\mathsf{S}(\widehat{\theta}) = \mathsf{E}(\widehat{\theta}) - \theta.$$

Estimador insesgado

Se dice que el estimador $\widehat{\theta}$ es insesgado si $\mathsf{S}(\widehat{\theta}) = 0$, es decir,

$$\mathsf{E}(\widehat{\theta}) = \theta.$$

Error cuadrático medio

El error cuadrático medio de un estimador $\widehat{\theta}$ del parámetro θ está definido como

$$\mathsf{ECM}(\widehat{\theta}) = \mathsf{E}(\widehat{\theta} - \theta)^2.$$

Se puede demostrar (sumando y restando $\mathsf{E}(\widehat{\theta})$) que la expresión anterior es equivalente a

$$\mathsf{ECM}(\widehat{\theta}) = \mathsf{E}(\widehat{\theta} - \mathsf{E}(\widehat{\theta}))^2 + (\theta - \mathsf{E}(\widehat{\theta}))^2 = \mathsf{Var}(\widehat{\theta}) + \mathsf{S}(\widehat{\theta})^2.$$

Además, si el estimador $\widehat{\theta}$ es insesgado, entonces

$$\mathsf{ECM}(\widehat{\theta}) = \mathsf{Var}(\widehat{\theta}).$$

Estimador insesgado de varianza mínima

Si se consideran varios estimadores insesgados de θ, el que tiene la menor varianza recibe el nombre de estimador insesgado de varianza mínima.

Eficiencia relativa

Sean $\widehat{\theta}_1$ y $\widehat{\theta}_2$ dos estimadores de θ. Entonces, la eficiencia relativa de estos dos estimadorres se define como

$$\text{ER} = \frac{\text{ECM}(\widehat{\theta}_2)}{\text{ECM}(\widehat{\theta}_1)}.$$

Note que si además $\widehat{\theta}_1$ y $\widehat{\theta}_2$ son dos estimadores insesgados de θ, entonces la eficiencia relativa se reduce a

$$\text{ER} = \frac{\text{Var}(\widehat{\theta}_2)}{\text{Var}(\widehat{\theta}_1)}.$$

Por tanto, tenemos que:

- Si ER < 1, entonces el estimador $\widehat{\theta}_1$ es menos eficiente que $\widehat{\theta}_2$.
- Si ER $= 1$, entonces el estimador $\widehat{\theta}_1$ es igual de eficiente que $\widehat{\theta}_2$.
- Si ER > 1, entonces el estimador $\widehat{\theta}_1$ es más eficiente que $\widehat{\theta}_2$.

Ejemplos

Ejemplo 1 (Montgomery, ejercicio 6.2)

Sea $X_1, \ldots, X_7$ una muestra aleatoria de tamaño $n = 7$ de una población que tiene media μ y varianza σ^2. Considere los siguientes estimadores de μ:

- $\widehat{\theta}_1 = \frac{X_1 + X_2 + \cdots + X_7}{7}$
- $\widehat{\theta}_2 = \frac{2X_1 - X_6 + X_4}{2}$

- Indique por qué ambos son estimadores.
- ¿Alguno de estos estimadores es insesgado?
- ¿Cuál estimador es mejor? ¿En qué sentido es mejor?
- Calcule la eficiencia relativa y concluya cuál de ellos es más eficiente.

En Chile hay interés en la reforma tributaria

El Diario Financiero presentó un estudio de la caída en el respaldo a la reforma tributaria.

La variable de interés X es el "respaldo a la reforma tributaria". Esta variable toma el valor uno (1) si la reforma es respaldada y cero (0) si no.

Un ejemplo del mundo real

- La población está constituida por todos los chilenos mayores de 18 años.
- El parámetro θ es la proporción de chilenos que respalda la reforma.
- Una forma de calcular de manera exacta el valor de θ sería hacer un censo y preguntarle a todos los chilenos mayores de 18 años si respaldan o no la reforma. Sin embargo, realizar un censo es muy costoso, pues el tamaño de la población N es un número muy grande.
- La distribución de la variable X es Bernoulli con parámetro θ, es decir,

$$p_X(x) = \mathsf{P}(X = x) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1.$$

¿Cómo estimar el valor de θ si consideramos sólo una muestra aleatoría de tamaño n mucho menor que N?

Un ejemplo del mundo real

- Considere una muestra aleatoria de tamaño $n = 10$. En la práctica, la encuesta es aplicada a 10 chilenos mayores de 18 años.
- Lo anterior quiere decir que hay un conjunto de 10 variables aleatorias $X_1, \dots, X_{10}$ que son independientes (la opinión de un individuo no afecta la opinión de otro) e idénticamente distribuidas Bernoulli con función de probabilidad

$$p_{X_i}(x) = \mathsf{P}(X_i = x) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1, i = 1, \dots, 10.$$

- Como no es de interés la opinión de cada individuo, sino la opinión conjunta de todos ellos, se debe usar la distribución conjunta de las 10 variables aleatorias mediante

$$L(\theta) = p_{X_1, \dots, X_{10}}(x_1, \dots, x_{10}).$$

Un ejemplo del mundo real

- Como las 10 variables aleatorias son independientes, se tiene que

$$
\begin{aligned}
L(\theta) &= p_{X_1}(x_1) \times \cdots \times p_{X_{10}}(x_{10}) && (3.1)\\
&= \theta^{x_1}(1-\theta)^{1-x_1} \times \cdots \times \theta^{x_{10}}(1-\theta)^{1-x_{10}} && (3.2)\\
&= \theta^{\sum_{i=1}^{10} x_i}(1-\theta)^{10-\sum_{i=1}^{10} x_i}. && (3.3)
\end{aligned}
$$

- Supongamos que la encuesta se aplica a 10 individuos seleccionados de manera aleatoria. Si el individuo respalda la reforma se asigna un uno (1), en caso contrario se asigna un cero (0). Un ejemplo de una muestra contiene los datos:

0, 1, 1, 0, 0, 1, 0, 1, 0, 0

- Considerando los datos anteriores: $\sum_{i=1}^{10} x_i = 4$ y

$$L(\theta) = \theta^4(1-\theta)^6.$$

Un ejemplo del mundo real

¿Cuál podría ser una estimación de la proporción de individuos que respalda la reforma?

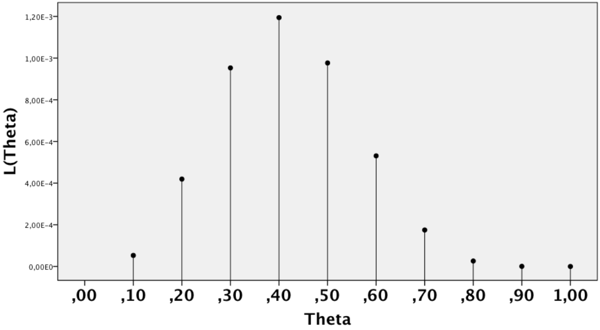
o

¿Cuál es el valor más probable de θ?

Esta pregunta se puede responder de tres formas:

- Usando la intuición. Como la estimación debe ser un valor del espacio paramétrico $\Theta = [0, 1]$, si no tenemos información desde una muestra aleatoria, podemos asumir este valor como $0,5$, pero si la tenemos y sabemos que 4 de 10 individuos respaldan la reforma, entonces $\hat{\theta} = 0,40$.

- Evaluando la función $L(\theta) = \theta^4(1-\theta)^6$ para distintos valores de θ y buscando el valor que maximiza $L(\theta)$. Un gráfico de $L(\theta) = \theta^4(1-\theta)^6$ es el siguiente.



Por lo tanto, $\hat{\theta} = 0,40$.

Un ejemplo del mundo real

- La tercera forma de responder la pregunta es derivando la función $L(\theta) = \theta^4(1-\theta)^6$ con respecto a θ e igualando esta derivada a cero como

$$L'(\theta) = 4\theta^3(1-\theta)^6 - 6(1-\theta)^5\theta^4 = 0.$$

Despejando θ de la ecuación anterior, se tiene $\hat{\theta} = 0,40$. En conclusión, se estima que el 40 % de los chilenos mayores de 18 años apoyará la reforma tributaria.

¿Esta estimación es exacta o posee error?

Error estándar

Sean $\widehat{\theta}$ un estimador puntual de θ y $\mathsf{Var}(\widehat{\theta})$ la varianza de este estimador. Una medida de la precisión de tal estimador es el EE de $\widehat{\theta}$ definido como

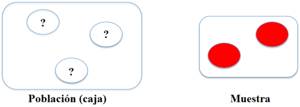
$$\mathsf{EE}(\widehat{\theta}) = \sqrt{\mathsf{Var}(\widehat{\theta})}.$$

Método de verosimilitud máxima

Un ejemplo didáctico (Mendenhall)

A continuación presentamos un ejemplo para ilustrar la lógica en la que está basado el método de verosimilitud máxima.

- Suponga que tenemos una caja que contiene tres bolitas. Sabemos que cada una de las bolitas puede ser roja o blanca, pero no sabemos el número total de cada uno de los dos colores.
- No obstante, podemos muestrear aleatoriamente dos de las bolitas sin reposición. Si la muestra aleatoria contiene dos bolitas rojas: **¿Cuál es una buena estimación del número total de bolitas rojas en la caja?**



Solución del ejemplo didáctico

Pongámonos en los dos siguientes escenarios. Obviamente, el número de bolitas rojas en la caja debe ser dos o tres (si hubiera cero o una bolita roja en la caja, sería imposible –probabilidad 0– sacar dos bolitas rojas cuando se muestrea sin reposición.

- Si hay dos bolitas rojas y una bolita blanca en la caja, la probabilidad de seleccionar aleatoriamente dos bolitas rojas es

$$\frac{\binom{2}{2}\binom{1}{0}}{\binom{3}{2}} = \frac{1}{3}.$$

- Si hay tres bolitas rojas en la caja, la probabilidad de seleccionar aleatoriamente dos bolitas rojas es

$$\frac{\binom{3}{2}\binom{0}{0}}{\binom{3}{2}} = 1.$$

Solución del ejemplo didáctico

- Parece razonable escoger el número tres como la estimación del número de bolitas rojas en la caja, porque esta estimación maximiza la probabilidad de obtener la muestra observada.

- Desde luego que es posible que la caja contenga sólo dos bolitas rojas, pero el resultado observado confiere más crédito a que haya tres bolitas rojas en la caja.

Outline

1. Nociones básicas
2. Inferencia estadística
3. Propiedades de los estimadores puntuales
4. Método de estimación de verosimilitud máxima
5. Distribuciones de muestreo
6. Distribuciones de muestreo para la media
7. Distribuciones de muestreo para la varianza
8. Intervalos de confianza
9. Método para obtener un IC
10. Intervalos de confianza para la media

Método de estimación de verosimilitud máxima

Método de estimación de verosimilitud máxima

Sea $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ una muestra aleatoria de tamaño n obtenida desde la variable aleatoria X con distribución de probabilidad $f_X(\cdot; \boldsymbol{\theta})$, donde $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$ es el vector de parámetros desconocidos a ser estimados. Sea $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ el vector de valores observados de la muestra aleatoria $\boldsymbol{X}$ de tamaño n. Entonces, la función de verosimilitud para $\boldsymbol{\theta}$ asociada con las observaciones $\boldsymbol{x}$ de $\boldsymbol{X}$ está dada por

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{i=1}^{n} f_X(x_i; \boldsymbol{\theta}).$$

De esta manera, el logaritmo de la función de verosimilitud, llamada función de log-verosimilitud, está dada por

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{i=1}^{n} \log(f_X(x_i; \boldsymbol{\theta})).$$

Note que la función de verosimilitud es una función de los parámetros desconocidos $\boldsymbol{\theta}$. El estimador de verosimilitud máxima de $\boldsymbol{\theta}$ es el valor de $\boldsymbol{\theta}$ que maximiza la función $L(\boldsymbol{\theta}; \boldsymbol{x})$ o equivalentemente el valor que maximiza la función $\ell(\boldsymbol{\theta}; \boldsymbol{x})$. Así, para encontrar el estimador de verosimilitud máxima de $\boldsymbol{\theta}$, debemos resolver el sistema de ecuaciones

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} \rightarrow \frac{\partial \ell}{\partial \theta_j} = 0, \quad j = 1, \ldots, k.$$

Este sistema es obtenido al derivar $\ell(\boldsymbol{\theta}; \boldsymbol{x})$ e igualar ésta a cero. De este modo, el estimador de verosimilitud máxima de $\boldsymbol{\theta}$ es

$$\hat{\boldsymbol{\theta}} = \arg \text{máx}\, \ell(\boldsymbol{\theta}, \boldsymbol{x}).$$

Sin pérdida de generalidad, considere que la muestra es obtenida desde una distribución de un parámetro (θ). Algunas propiedades del estimador verosimilitud máxima $\hat{\theta}$, bajo ciertas condiciones de regularidad, son

- **Insesgamiento asintótico:**

$$\lim_{n\to\infty} \mathsf{E}(\hat{\theta}) = \theta.$$

- **Eficiencia asintótica:** (varianza mínima)

$$\lim_{n\to\infty} \mathsf{Var}(\hat{\theta}) = -\frac{1}{\frac{\partial^2 L(\hat{\theta})}{\partial \theta^2}}.$$

- **Normalidad asintótica:**
  El estimador $\hat{\theta}$ sigue una distribución aproximadamente normal cuando n va a infinito.
- **Invarianza:**
  Si $\hat{\theta}$ es el estimador VM de θ y g es una función biyectiva y diferenciable, entonces $g(\hat{\theta})$ es el estimador VM de $g(\theta)$.
- **Unicidad:**
  Bajo condiciones generales, cuando un estimador VM existe, éste es único.

Ejemplos

Ejemplo compañía de seguros (Hildebrand)

Una compañía de seguros utiliza la distribución de Pareto, cuya FDP está dada por

$$f_X(x;\theta) = \theta(x+1)^{-(\theta+1)}, \quad x > 0,\, \theta > 0,$$

como la supuesta distribución de probabilidad de $X =$ valor (en miles de dólares) de la cantidad pagada por un seguro debido a la ocurrencia de un siniestro. Para un tipo de póliza, tres reclamos por pago proporcionan valores $x_1 = 0,82$, $x_2 = 0,63$ y $x_3 = 7,55$.

(i) Exprese la función de verosimilitud para θ.

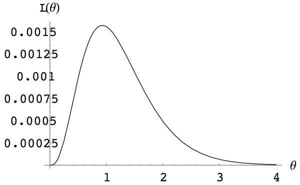
(ii) Evalúe la función de verosilimilitud para $\theta = 0, 1, 2, 3$. ¿Cuál de estos tres valores de θ le parece más cercano a la estimación VM?

(iii) Encuentre la estimación VM de θ para una muestra de tamaño n.

Solución

(i)

$$L(\theta) = \prod_{i=1}^{3} f_X(x_i;\theta) = \prod_{i=1}^{3} \theta(x_i+1)^{-(\theta+1)} \quad (4.1)$$
$$= \theta^3[(x_1+1)(x_2+1)(x_3+1)]^{-(\theta+1)} \quad (4.2)$$
$$= \theta^3 25,36^{-(\theta+1)}. \quad (4.3)$$



(ii) Evaluando $L(\theta)$ en $\theta = 0, 1, 2$ y 3, tenemos $L(0) = 0$, $L(1) = 0,0015544$, $L(2) = 0,0004902$ y $L(3) = 0,0000652$. El valor más cercano a la estimación VM es $\hat{\theta}_{\text{max}} = 1$. En la gráfica se nota que el máximo se alcanza cerca del valor uno.

(iii)

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = \prod_{i=1}^{n} \theta(x_i+1)^{-(\theta+1)} = \theta^n \prod_{i=1}^{n} (x_i+1)^{-(\theta+1)}.$$

Ahora, haciendo $\ell(\theta) = \log(L(\theta))$, tenemos

$$\ell(\theta) = n\log(\theta) - (\theta+1)\sum_{i=1}^{n} \log(x_i+1).$$

Derivando $\ell(\theta)$ con respecto a θ, tenemos

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \log(x_i+1).$$

Finalmente, igualando a cero la derivada anterior y despejando θ, evaluando θ en $\hat{\theta}$, se obtiene el estimador de verosimilitud máxima (que es una variable aleatoria) como

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n} \log(X_i + 1)}.$$

Para verificar los cálculos anteriores, note que, si un tipo de póliza proporciona tres reclamos por pago, digamos $x_1 = 0,82$, $x_2 = 0,63$ y $x_3 = 7,55$, entonces

$$\widehat{\theta} = \frac{3}{\sum_{i=1}^{3} \log(x_i + 1)} = \frac{3}{\log(1,82) + \log(1,63) + \log(8,55)} = 0,9278.$$

Outline

1. Nociones básicas
2. Inferencia estadística
3. Propiedades de los estimadores puntuales
4. Método de estimación de verosimilitud máxima
5. Distribuciones de muestreo
6. Distribuciones de muestreo para la media
7. Distribuciones de muestreo para la varianza
8. Intervalos de confianza
9. Método para obtener un IC
10. Intervalos de confianza para la media

Distribuciones de Muestreo

- Recuerde que la inferencia estadística tiene que ver con la toma de decisiones sobre una población, con base en la información contenida en una muestra aleatoria de ésta.
- Por ejemplo, suponga que se tiene interés en la proporción de personas que está a favor de la reforma tributaria en Chile. Un asesor de gobierno afirma que la proporción de personas (en todo Chile) que apoya la reforma es de $p = 0,46$.
- Una empresa consultora realiza un estudio por encuesta, la que aplica a 200 chilenos (seleccionados al azar) y calcula la proporción estimada de chilenos que está a favor de la reforma, estimación que resulta ser $\hat{p} = 0,50$.
- La empresa consultora puede repetir el estudio varias veces y entonces podría conocer la distribución de probabilidad empírica de $\hat{p} = \sum_{i=1}^{n} X_i/n$.

Teorema del límite central (TLC)

Sea $X_1, \ldots, X_n$ una muestra aleatoria de tamaño n tomada de una población (es decir, $X_1, \ldots, X_n$ es una sucesión de VAs IID, con distribución de probabilidad no especificada) de media μ y varianza σ^2. Entonces, si n es suficientemente grande, la VA

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

tiene aproximadamente una distribución normal con media $\mu_{\overline{X}_n} = \mu$ y varianza $\sigma^2_{\overline{X}_n} = \sigma^2/n$ (basado en ley de grandes números).

Observaciones:

• Formalmente decimos que la VA $Z_n = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}}$ converge a una VA con distribución normal estándar cuando $n \to \infty$.

• Independientemente del tipo de distribución (discreta, continua, simétrica o sesgada, unimodal o multimodal), la distribución muestral de $\overline{X}_n$ es normal (continua, simétrica y unimodal), si n es grande.

Ejemplo

Un asesor de gobierno afirma que el 46 % de la población de Chile está a favor de la reforma tributaria. Si extraemos una muestra aleatoria de tamaño 200, ¿ cuál es la probabilidad de que al menos 100 de ellos estén a favor de la reforma?

Solución:

Sabemos que $X_1, \ldots, X_{200}$ es una muestra aleatoria de tamaño $n = 200$ tomada de una población Bernoulli con media $\mu = 0,46$ y varianza $\sigma^2 = 0,46(1 - 0,46) = 0,2484$. Entonces, como $n = 200$ es suficientemente grande, la VA

$$\hat{p} = \overline{X}_n = \frac{1}{200} \sum_{i=1}^{200} X_i$$

tiene aproximadamente una distribución normal con media $\mu_{\overline{X}_n} = 0,46$ y varianza $\sigma^2_{\overline{X}_n} = 0,2484/200$.

TLC

- Sea $Y = \sum_{i=1}^{200} X_i$ la cantidad de chilenos que está a favor de la reforma tributaria, de un total de 200 chilenos.
- La probabilidad requerida es

$$\mathsf{P}(Y \geq 100) = \mathsf{P}(\hat{p} \geq 100/200) = \mathsf{P}(\hat{p} \geq 0,50).$$

- Utilizando el TLC, tenemos

$$\mathsf{P}(\hat{p} \geq 0,50) \approx \mathsf{P}\left(Z_n \geq \frac{0,50 - 0,46}{\sqrt{0,2484/200}}\right) = \mathsf{P}(Z_n \geq 1,135) = 0,128.$$

- Si no utilizamos el TLC, hay que considerar que Y se distribuye binomialmente con $n = 200$ y $p = 0,46$. Entonces,

$$\mathsf{P}(Y \geq 100) = \sum_{y=100}^{200} \binom{200}{y} 0,46^y (1 - 0,46)^{200-y} = 0,114.$$

Distribución de muestreo:

- Es la distribución de probabilidad de un estadístico.
- Depende de la distribución de la población y del tamaño de la muestra.

Error estándar de una estadística:

Es la desviación estándar de su distribución de muestreo.

Outline

1. Nociones básicas
2. Inferencia estadística
3. Propiedades de los estimadores puntuales
4. Método de estimación de verosimilitud máxima
5. Distribuciones de muestreo
6. Distribuciones de muestreo para la media
7. Distribuciones de muestreo para la varianza
8. Intervalos de confianza
9. Método para obtener un IC
10. Intervalos de confianza para la media

Distribuciones de muestreo para la media muestral

Sea $X_1, \ldots, X_n$ una muestra aleatoria de tamaño n con **distribución normal** de media μ y varianza σ^2. Entonces, la media muestral

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

tiene una distribución normal con media μ y varianza σ^2/n, es decir,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1) \quad \text{(distribución normal estándar).}$$

Distribución muestral de $\overline{X}$ (σ desconocida)

- Sean $Z \sim \mathsf{N}(0,1)$ y $U \sim \chi^2(\nu)$ (distribución chi-cuadrado con ν grados de libertad) dos VAs independientes. Entonces,

$$T = \frac{Z}{\sqrt{U/\nu}} \sim t(\nu),$$

es decir, la distribución t de Student con ν grados de libertad.

- Sea $X_1, \ldots, X_n$ una muestra aleatoria de tamaño n con distribución normal de media μ y varianza σ^2. Entonces,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

donde $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$.

Distribución muestral de $\overline{X} - \overline{Y}$ (σ_X^2, σ_Y^2 conocidas)

Suponga que el muestreo se lleva a cabo en dos poblaciones normales independientes, digamos, $X_1, \ldots, X_{n_X}$ e $Y_1, \ldots, Y_{n_Y}$, con varianzas conocidas, es decir, $X \sim \mathsf{N}(\mu_X, \sigma_X^2)$ e $Y \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$. Entonces,

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathsf{N}(0, 1).$$

Distribución muestral de $\overline{X} - \overline{Y}$ (σ_X^2, σ_Y^2 desconocidas)

Suponga que el muestreo se lleva a cabo en dos poblaciones normales independientes, digamos, $X_1, \ldots, X_{n_X}$ e $Y_1, \ldots, Y_{n_Y}$, con varianzas iguales, pero desconocidas, es decir, $X \sim \mathsf{N}(\mu_X, \sigma^2)$ e $Y \sim \mathsf{N}(\mu_Y, \sigma^2)$. Entonces,

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t(n_X + n_Y - 2),$$

donde

$$S_p = \sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}},$$

con S_X^2 y S_Y^2 siendo las varianzas muestrales de la población X e Y, respectivamente.

Outline

1. Nociones básicas
2. Inferencia estadística
3. Propiedades de los estimadores puntuales
4. Método de estimación de verosimilitud máxima
5. Distribuciones de muestreo
6. Distribuciones de muestreo para la media
7. Distribuciones de muestreo para la varianza
8. Intervalos de confianza
9. Método para obtener un IC
10. Intervalos de confianza para la media

Distribuciones de muestreo para la varianza

Distribución muestral de S^2 (μ conocida)

- Sean las VAs $Z_i \sim \mathsf{N}(0,1)$ independientes, para $i = 1, \ldots, n$. Entonces,

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n).$$

- Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución normal con media μ y varianza σ^2. Entonces,

$$\frac{nS_u^2}{\sigma^2} \sim \chi^2(n),$$

donde

$$S_u^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

y es necesario asumir que μ es conocida.

- En la realidad es muy difícil conocer el verdadero valor de μ. Entonces, la varianza muestral S_u^2 presentada anteriormente tiene poca utilidad práctica.
- De acuerdo a lo anterior, si se muestrea una población normal con media μ y varianza σ^2 (ambas desconocidas), la varianza muestral se define por

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Entonces,

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

Distribución muestral de S_X^2/S_Y^2

- Sean $U \sim \chi^2(\nu_1)$ e $V \sim \chi^2(\nu_2)$ dos VAs independientes. Entonces,

$$F = \frac{U/\nu_1}{V/\nu_2} \sim \mathcal{F}(\nu_1, \nu_2),$$

es decir, la distribución F de Fisher con ν_1 grados de libertad en el numerador y ν_2 grados de libertad en el denominador.

- Suponga que el muestreo se lleva a cabo en dos poblaciones normales independientes, digamos, $X_1, \ldots, X_{n_X}$ e $Y_1, \ldots, Y_{n_Y}$, es decir, $X \sim \mathsf{N}(\mu_X, \sigma_X^2)$ e $Y \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$. Entonces,

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim \mathcal{F}(n_X - 1, n_Y - 1).$$

Outline

1. Nociones básicas
2. Inferencia estadística
3. Propiedades de los estimadores puntuales
4. Método de estimación de verosimilitud máxima
5. Distribuciones de muestreo
6. Distribuciones de muestreo para la media
7. Distribuciones de muestreo para la varianza
8. Intervalos de confianza
9. Método para obtener un IC
10. Intervalos de confianza para la media

Intervalos de confianza

Introducción

- En muchas situaciones, una estimación puntual no proporciona información suficiente sobre un parámetro.
- Por ejemplo, si se tiene interés en estimar la proporción de personas que estará a favor de la reforma tributaria, sería mejor estimar ésta mediante un intervalo, dentro del cual se espera encontrar el valor de este parámetro (proporción).
- Este intervalo recibe el nombre de intervalo de confianza (IC).

Estimación por intervalo

Una estimación por intervalos o IC de un parámetro desconocido θ es un intervalo de la forma $[l; u]$, es decir, $l \leq \theta \leq u$, donde los límites l y u dependen del valor numérico del estimador $\hat{\theta}$, para una muestra particular de la distribución de la población.

Intervalo aleatorio (IA)

Es un intervalo cuyos límites son expresados en términos de estadísticos (correspondientes a VAs), es decir, es un estimador por intervalos.

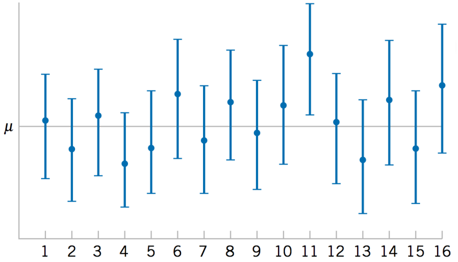
Intervalo de confianza

Es un IA evaluado en estimaciones puntuales por intervalo, es decir, es una estimación por intervalos.

Coeficiente de confianza

El coeficiente de confianza de un IC corresponde a la probabilidad con que el IA contiene al parámetro θ. Si el coeficiente de confianza es $1-\alpha$, con $0<\alpha<1$, entonces se dice que el IC es del $100(1-\alpha)\,\%$ de confianza.

La interpretación de un IC es que, si se toma un número infinito de muestras aleatorias y se calcula un IC del $100(1-\alpha)\,\%$ para θ, para cada una de estas muestras, entonces se espera que el $100(1-\alpha)\,\%$ de esos ICs contengan el verdadero valor del parámetro θ.



Por ejemplo, en la figura de arriba, se presentan 16 ICs para μ del 95 %. Empíricamente, se observa que 15 de ellos (93,75 %) contienen al verdadero valor del parámetro μ. El valor 93,75 % es lo que conoce como probabilidad de cobertura empírica del IC, que en teoría debería ser 95 %.

Outline

1. Nociones básicas
2. Inferencia estadística
3. Propiedades de los estimadores puntuales
4. Método de estimación de verosimilitud máxima
5. Distribuciones de muestreo
6. Distribuciones de muestreo para la media
7. Distribuciones de muestreo para la varianza
8. Intervalos de confianza
9. Método para obtener un IC
10. Intervalos de confianza para la media

Método para obtener un IC

Cantidad pivotal

Una cantidad pivotal $Q = Q(X_1, \ldots, X_n; \theta)$ es una función de la muestra aleatoria de tamaño n que depende del parámetro, pero su distribución no depende de éste.

La construcción de un IC para un parámetro θ consta en la selección de dos valores a y b tales que

$$\mathsf{P}\left[a < Q(X_1, \ldots, X_n; \theta) < b\right] = 1 - \alpha.$$

Mediante manipulación algebráica, se obtiene que

$$\mathsf{P}\left[T_1(X_1, \ldots, X_n; a) < \theta < T_2(X_1, \ldots, X_n; b)\right] = 1 - \alpha.$$

Ejemplo de la creación de un IC

Sea $X_1, \ldots, X_n$ una muestra aleatoria desde una distribución normal con media μ (desconocida) y varianza σ^2 (conocida). Construya un IC de $100(1-\alpha)\,\%$ para μ.

Solución:

Se puede demostrar que un estimador (por ejemplo de verosimilitud máxima) de μ es $\widehat{\mu} = \overline{X}$. Entonces, una cantidad pivotal es

$$Q(X_1, \ldots, X_n; \mu) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

porque, como σ es conocido, Q es una función de la muestra que depende del parámetro μ pero su distribución es N(0,1), que no depende de μ, es decir,

$$Q(X_1, \ldots, X_n; \mu) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

Ejemplo de la creación de un IC

Debemos seleccionar dos valores a y b tal que

$$\mathsf{P}\left[a < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < b\right] = \mathsf{P}\left[a < Z < b\right] = 1 - \alpha.$$

Sabemos que

$$\mathsf{P}\left[z_{\alpha/2} < Z < z_{1-\alpha/2}\right] = \underbrace{\mathsf{P}\left[Z < z_{1-\alpha/2}\right]}_{1-\alpha/2} - \underbrace{\mathsf{P}\left[Z < z_{\alpha/2}\right]}_{\alpha/2} = 1 - \alpha.$$

Además, $z_{\alpha/2} = -z_{1-\alpha/2}$. Entonces,

$$\mathsf{P}\left[-z_{1-\alpha/2} < Z < z_{1-\alpha/2}\right] = 1 - \alpha.$$

De esta forma, $a = -z_{1-\alpha/2}$, $b = z_{1-\alpha/2}$ y $Q = Z$.

Manipulando algebráicamente, tenemos que

$$\mathsf{P}\left(-z_{1-\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1-\alpha,$$

$$\mathsf{P}\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Así, podemos decir que el IA $\left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ contiene a μ con probabilidad $1-\alpha$. Por lo tanto,

$$\begin{aligned}\mathsf{IC}_{100(1-\alpha)\,\%}(\mu) &= \left[\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \\ &= \left[\overline{x} \mp z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].\end{aligned}$$

Ejemplo de la creación de un IC

Interpretación correcta

Un IA $\left[\overline{X} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n}; \overline{X} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n}\right]$ contiene al verdadero valor de la media μ con probabilidad $1-\alpha$.

Interpretación incorrecta: no se debe interpretar como que es μ el que cae en el IA $\left[\overline{X} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n}; \overline{X} + z_{1-\alpha/2} \cdot \sigma/\sqrt{n}\right]$ con probabilidad $1-\alpha$.

Error de estimación

La expresión $d = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ se denomina error de estimación (o error de precisión o margen de error) y $l = 2d$ es el largo del intervalo.

Outline

1 Nociones básicas
2 Inferencia estadística
3 Propiedades de los estimadores puntuales
4 Método de estimación de verosimilitud máxima
5 Distribuciones de muestreo
6 Distribuciones de muestreo para la media
7 Distribuciones de muestreo para la varianza
8 Intervalos de confianza
9 Método para obtener un IC
10 Intervalos de confianza para la media

Intervalos de confianza para la media

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una población normal con media μ y varianza σ^2 (conocida). Además, sea $\overline{x}$ la media muestral. Entonces, un IC del $100(1-\alpha)\,\%$ para μ es

$$\mathsf{IC}_{100(1-\alpha)\,\%}(\mu) = \left[\overline{x} \mp z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Tamaño de muestra

Para un error de estimación d y un nivel de confianza establecidos, además de contar con una estimación de la variabilidad de la población, el tamaño de muestra para estimar la media población está dado por

$$n = \frac{z_{1-\alpha/2}^2\,\sigma^2}{d^2}.$$

Ejemplo

Un gerente de una empresa afirma que el número de llamadas solicitando su servicio es a lo más de 15 por semana, en promedio. Para comprobar su afirmación, se revisaron los registros de solicitud de servicio para 36 semanas seleccionadas al azar, obteniendo $\overline{x} = 17$. Además, se sabe desde estudios anteriores que $\sigma^2 = 9$.

- ¿Contradice la evidencia de la muestra la afirmación del gerente? Considere que el muestreo se realizó sobre una población que tiene distribución normal y con una confianza del 95 %.
- Encuentre el tamaño de la muestra necesario dado un error de estimación no superior a $0,5$.

Solución del ejemplo

- El gerente afirma que el número medio de llamadas solicitando servicio es a lo más de 15 por semana ($\mu \leq 15$). Un IC del 95 % para la media poblacional está dado por $(17 - 1,96 \cdot \frac{3}{\sqrt{36}};\ 17 + 1,96 \cdot \frac{3}{\sqrt{36}}) = (16,02;\ 17,98)$, lo que contradice lo planteado por el gerente.
- Bajo estas condiciones, tenemos que

$$n = \frac{z_{1-\alpha/2}^2 \cdot \sigma^2}{d^2} = \frac{1{,}96^2 \cdot 9}{0,5^2} = 138,29,$$

Entonces, se debe considerar $n = 139$.

IC para μ (σ desconocida)

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una población normal con media μ y varianza σ^2 (desconocida). Además, considere que $\overline{x}$ y s^2 son la media y la varianza muestral, respectivamente. Entonces, un IC del $100(1-\alpha)\,\%$ para μ es

$$\text{IC}_{100(1-\alpha)\,\%}(\mu) = \left[\overline{x} \mp t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right].$$

Ejemplo

Se toma una muestra aleatoria de 30 vicepresidentes de una compañía y se determina el porcentaje del ingreso bruto que cada uno pagó como impuesto sobre la renta. Los datos son:

30,0	29,1	26,8	26,1	19,4	31,2
19,1	19,5	26,0	18,9	21,7	28,0
22,7	27,1	22,9	20,2	17,0	25,7
20,7	23,1	20,5	10,6	23,9	25,0
21,2	23,9	21,3	22,8	20,2	26,7

Estadísticos descriptivos

	N	Media	Varianza	Asimetría		Curtosis	
	Estadístico	Estadístico	Estadístico	Estadístico	Error típico	Estadístico	Error típico
Porcentaje del ingreso bruto	30	23,046	18,520	-,467	,427	1,125	,833
N válido (según lista)	30						

Ejemplo

Para los datos de los 30 vicepresidentes:

- Calcule un IC del 95 % para la media poblacional. Dé una interpretación verbal detallada de este intervalo.
- Determine el tamaño de muestra mínimo, suponiendo que el error de estimación no es superior a 0,2.
- ¿Es lógico suponer que la distribución muestral de la media es aproximadamente normal? ¿Qué teorema se está utilizando?

Fin

Gracias por asistir a esta clase