Universidad Nacional
Autónoma De México

Facultad de Estudios Superiores Acatlán

Actividad:

Estadística Inferencial

Técnicas Estadísticas y
Minería de Datos

Profesor:
Dr. Julio Cesar Galindo López

Módulo 1: Modelos estadísticos

Equipo 6

Integrantes:

Cariño Díaz David
Márquez Sánchez Moisés
Martínez Romualdo Valeria
Mondragón Miranda Néstor Yair
Reyes Cruz Alejandro
Torres Bustamante Dulce Jhoana

1. Investiga sobre el método de momentos y da dos ejemplos.

Solución:

El método de momentos, consiste en ver a la función de densidad de una variable aleatoria X de la forma $f(x, \theta)$, que depende de un parámetro θ fijo pero desconocido. Este método nos provee un mecanismo para estimar θ.

Este consiste en calcular los primeros k momentos poblacionales (Definimos al k-ésimo momento poblacional como $E[X^k]$), e igualarlos con los correspondientes k momentos muestrales (Definimos al k-ésimo momento muestral como $m_k = \frac{1}{n}\Sigma_{I=1}^{n} X_i^k$) de una muestra que suponemos que proviene de variables aleatorias independientes e idénticamente distribuidas, y con ello plantear y resolver un sistema de k ecuaciones para el parámetro o vector de parámetros que deseamos estimar.

Por ejemplo, supongamos que la muestra $x_1, x_2, x_3, \ldots x_n$ proviene de la sucesión de variables aleatorias $X_1, X_2, X_3, \ldots, X_n$ independientes e idénticamente distribuidas que siguen una distribución $Normal(\mu, \sigma^2)$. Para estimar ambos parámetros realizamos lo siguiente:

Conocemos cuales son las esperanzas de los primeros dos momentos de una variable aleatoria $X \sim Normal(\mu, \sigma^2)$

$$\mathbb{E}[X] = \mu \quad \text{y} \quad \mathbb{E}[X^2] = \sigma^2 + \mu^2$$
$$\Rightarrow$$
$$\mu = \mathbb{E}[X]$$
$$\wedge$$
$$\sigma^2 = \mathbb{E}[X^2] - \mu^2$$

Por lo que sustituyendo los momentos poblacionales por los momentos poblacionales obtenemos los siguientes estimadores.

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$\wedge$$
$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{[\sum_{i=1}^{n} x_i]^n}{n}$$

Ejemplo 2:

Busquemos ahora los estimadores por el método de momentos de una muestra aleatoria proveniente de una sucesión de variables aleatorias independientes e idénticamente distribuidas $X_i \sim Bin(n, p)$.

Nuevamente conocemos los primeros dos momentos poblacionales

$$\mathbb{E}[X] = np \quad \text{y} \quad \mathbb{E}[X^2] = np(1-p)$$

Igualamos los momentos poblacionales y los muestrales, y resolvemos el sistema de ecuaciones que dan como resultado.

$$m_1 = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{y} \quad m_2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

$$m_1 = np \quad \wedge m_2 = np(1-p) + (np)^2$$
$$\Rightarrow$$
$$m_2 = m_1(1-p) + (m_1)^2$$
$$\Rightarrow$$
$$\frac{m_2 - (m_1)^2}{m1} = 1 - p$$
$$\Rightarrow$$
$$p = 1 - \frac{m_2 - (m_1)^2}{m1} \quad \text{y} \quad n = \frac{m_1}{p}$$
$$\Rightarrow$$
$$\hat{p} = 1 - \frac{m_2 - (m_1)^2}{m1} \quad \text{y} \quad \hat{n} = \frac{m_1}{\hat{p}}$$

Los cuales son los estimadores por el método de momentos de las variables p y n

2. Calcula el MLE para la media μ y varianza σ^2 de una muestra normal.

Solución:

Sean $X_1, X_2, \ldots, X_n$ v.a.i.i.d tales que $X_i \sim Normal(\mu, \sigma^2)$

$$\Rightarrow$$
$$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$\Rightarrow$$
$$L(\mu,\sigma^2|x_1,..,x_n) = \prod_{i=1}^{n} f(x_i|\mu,\sigma^2)$$
$$\Rightarrow$$
$$ln(L(\mu,\sigma^2|x_1,..,x_n)) = ln\left(\prod_{i=1}^{n} f(x_i|\mu,\sigma^2)\right) = \sum_{i=1}^{n} ln(f(x_i|\mu,\sigma^2))$$
$$= \sum_{i=1}^{n} ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = -\frac{n}{2} ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Ahora derivamos con respecto a ambos parámetros para encontrar los valores de μ y de σ^2 tales que maximicen la log-similitud. Comenzamos con la primer variable:

$$\frac{d}{d\mu} ln(L(\mu,\sigma^2|x_1,..,x_n) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(xi-\mu)(-1) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$
$$\Leftrightarrow$$
$$\sum_{i=1}^{n}(x_i-\mu) = 0 = \sum_{i=1}^{n} x_i - n\mu$$
$$\Leftrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ahora repetimos el proceso, pero derivando con respecto a la segunda variable.

$$\frac{d}{d\sigma^2} ln(L(\mu,\sigma^2|x_1,..,x_n) = \frac{d}{d\sigma^2}\left(\left(-\frac{n}{2\sigma^2}\right)\left[ln(2\pi)+ln(\sigma^2)\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$=$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \Leftrightarrow \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{n}{2\sigma^2}$$

$$\Leftrightarrow$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = n \Leftrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Por lo tanto, los estimadores por el método de máxima verosimilitud son:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{Y} \quad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

La referencia a los siguientes ejercicios es
https://www.probabilitycourse.com

3. Resuelve el ejemplo Example 8.17 de la sección 8.3.2

Example 8.17 (Public Opinion Polling) We would like to estimate the portion of people who plan to vote for Candidate A in an upcoming election. It is assumed that the number of voters is large, and θ θ is the portion of voters who plan to vote for Candidate A. We define the random variable X X as follows. A voter is chosen uniformly at random among all voters and we ask her/him: "Do you plan to vote for Candidate A A?" If she/he says "yes," then X=1X=1, otherwise X=0 X=0. Then:

$$X \sim Bernoulli(\theta)$$

Let $X_1, X_2, X_3, \ldots X_n$ be a random sample from this distribution, which means that the X_is are i.i.d. and $Xi \sim Bernoulli(\theta)$. In other words, we randomly select n n voters (with replacement) and we ask each of them if they plan to vote for Candidate A. Find a (1−α)100% (1−α)100% confidence interval for θ based on $X_1, X_2, X_3, \ldots, X_n$

Solución:

Al ser cada X_i i.d Bernoulli, sabemos que $\mathbb{E}[X_i] = \theta$ y $Var(X_i) = \theta(1-\theta) = \sigma^2, \forall i \in \{1,2,...,n\}$. Ahora, queremos construir un intervalo de confianza de $(1-\alpha)\,\%100$ para el parámetro θ, el cual contruiremos con ayuda del TCL a tráves de lo siguiente:

$$IC_\alpha = \left[\bar{X} - z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right]$$

Para ello hace falta definir la σ, pero esta depende del mismo parámetro buscando θ, así que vamos a buscar una cota para esta. Emtonces definimos:

$$f(\theta) = \theta(1-\theta) \quad for \quad \theta \in (0,1)$$

Luego:

$$f^{(1)}(\theta) = \frac{d}{d\theta}\left[\theta - \theta^2\right] = 1 - 2\theta$$

$$f^{(2)}(\theta) = \frac{d}{d\theta}\left[1-2\theta\right] = -2$$

Teniendo que si hay máximo, igualamos al primera derivada en 0:

$$1 - 2\theta = 0, \quad \Rightarrow \quad \theta = \frac{1}{2}$$

Por lo tanto,

$$f(\theta) =\leq f(\frac{1}{2}) = \frac{1}{4} \quad \Rightarrow \quad \sigma_{máx} = \sqrt{\frac{1}{4}} = \frac{1}{2}$$

De aquí que concluimos que el intervalo de confianza ed $(1-\alpha)\,\%100$ esta dado por:

$$IC_\alpha = \left[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma_{máx}}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma_{máx}}{\sqrt{n}}\right]$$

$$= \left[\bar{X} - \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}, \bar{X} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}\right]$$

4. Resuelve los problemas Problem 2-Problem 6 (usa las applets o Python para calcular los cuantiles) de la sección 8.3.4.

Problem 6

A random sample $X_1, X_2, X_3, \ldots, X_{16}$ is given from a normal distribution with unknown mean μ=EXi and unknown variance $Var(X_i) = \sigma^2$. For the observed sample, the sample mean is $\bar{X} = 16.7$ and the sample variance is $S^2 = 7.5$

- Find a 95% confidence interval for μ
- Find a 95% confidence interval for σ^2

4.- Resuelve los problemas Problem 2-Problem 6 (usa las applets o Python para calcular los cuantiles) de la sección 8.3.4.

Problem 2

A random sample $X_1, X_2, X_3, \ldots, X_{100}$ is given from a distribution with known variance $\mathrm{Var}(X_i) = 16$. For the observed sample, the sample mean is $\overline{X} = 23.5$. Find an approximate 95% confidence interval for $\theta = EX_i$.

Problema 2

```python
import scipy.stats as stats

# Valores reportados
n = 100  # muestra
sample_mean = 23.5  # promedio de la muestra
variance = 16  # valor de la varianza
std_dev = variance ** 0.5  # desviación estandar

# Error cuadrático medio
standard_error = std_dev / (n ** 0.5)

# Invervalo de confinaza
# Para un 95% de confianza, usamos z-score 95%.
z_score = stats.norm.ppf(0.975)

# Margen de error
margin_of_error = z_score * standard_error

# Cálculo de intervalo de confianza
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)
confidence_interval
```

```
(22.71601440618398, 24.28398559381602)
```

Problem 3

To estimate the portion of voters who plan to vote for Candidate A in an election, a random sample of size n from the voters is chosen. The sampling is done with replacement. Let θ be the portion of voters who plan to vote for Candidate A among all voters. How large does n need to be so that we can obtain a 90% confidence interval with 3% margin of error? That is, how large n needs to be such that

$$P\left(\overline{X} - 0.03 \leq \theta \leq \overline{X} + 0.03\right) \geq 0.90,$$

where $\overline{X}$ is the portion of people in our random sample that say they plan to vote for Candidate A.

Problema 3

```python
# Z-score para 90%
z_score_90 = stats.norm.ppf(0.95)

# Margen de error
margin_of_error = 0.03

# Peor esenario
p_hat = 0.5

# Resolviendo para n usando la formula de margen de error
n_required = (z_score_90 ** 2 * p_hat * (1 - p_hat)) / (margin_of_error ** 2)
n_required
```

```
751.5398483598369
```

Problem 4

a. Let X be a random variable such that $R_X \subset [a, b]$, i.e., we always have $a \leq X \leq b$. Show that

$$\mathrm{Var}(X) \leq \frac{(b-a)^2}{4}.$$

b. Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from an unknown distribution with CDF $F_X(x)$ such that $R_X \subset [a, b]$. Specifically, EX and $\mathrm{Var}(X)$ are unknown. Find a $(1-\alpha)100\%$ confidence interval for $\theta = EX$. Assume that n is large.

Problema 4

```python
import numpy as np

# Definir valores arbitrarios para a, b y alfa
a = 0
b = 10
alpha = 0.05
n = 1000  # Tamaño de la muestra

# Generar una muestra aleatoria de tamaño n de una distribución uniforme entre a y b
# Esta es solo una elección arbitraria para demostrar; la distribución real es desconocida
sample = np.random.uniform(a, b, n)

# Calcular la media muestral
sample_mean = np.mean(sample)

# Desviación estándar de la muestra
sample_std = np.std(sample, ddof=1)

# Error estándar de la media
standard_error = sample_std / np.sqrt(n)

# Encontrar el valor crítico para un intervalo de confianza de (1 - alpha) * 100%
z_critical = stats.norm.ppf(1 - alpha / 2)

# Calcular el intervalo de confianza para la media
margin_of_error = z_critical * standard_error
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)

(sample_mean, standard_error, z_critical, confidence_interval)
```

```
(4.959171405311368,
 0.09087824552601774,
 1.959963984540054,
 (4.7810533171021845, 5.1372894935205515))
```

Problem 5

A random sample $X_1, X_2, X_3, \ldots, X_{144}$ is given from a distribution with unknown variance $\mathrm{Var}(X_i) = \sigma^2$. For the observed sample, the sample mean is $\overline{X} = 55.2$, and the sample variance is $S^2 = 34.5$. Find a 99% confidence interval for $\theta = EX_i$.

Probelma 5

```python
# Valores dados del problema
n = 144  # Tamaño de la muestra
sample_mean = 55.2  # Media muestral
sample_variance = 34.5  # Varianza muestral
alpha = 0.01  # Nivel de significancia para un 99% de intervalo de confianza

# Calcular la desviación estándar muestral
sample_std_dev = np.sqrt(sample_variance)

# Error estándar de la media
standard_error = sample_std_dev / np.sqrt(n)

# Distribución t de Student porque la varianza de la población es desconocida
# Grados de libertad para la t de Student
degrees_of_freedom = n - 1

# Valor crítico t para un intervalo de confianza del 99%
t_critical = stats.t.ppf(1 - alpha / 2, degrees_of_freedom)

# Intervalo de confianza para la media
margin_of_error = t_critical * standard_error
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)

confidence_interval
```

```
(53.92216008451527, 56.477839915484736)
```

Problem 6

A random sample $X_1, X_2, X_3, \ldots, X_{16}$ is given from a normal distribution with unknown mean $\mu = EX_i$ and unknown variance $\mathrm{Var}(X_i) = \sigma^2$. For the observed sample, the sample mean is $\overline{X} = 16.7$, and the sample variance is $S^2 = 7.5$.

a. Find a 95% confidence interval for μ.
b. Find a 95% confidence interval for σ^2.

Problema 6

```
[9] # Valores
    n = 16  # Tamaño de la muestra
    sample_mean = 16.7  # Media muestral
    sample_variance = 7.5  # Varianza muestral
    alpha = 0.05  # Nivel de significancia

    # Desviación estándar muestral
    sample_std_dev = np.sqrt(sample_variance)

    # Error estándar de la media
    standard_error = sample_std_dev / np.sqrt(n)

    # Valor crítico t para la media
    t_critical = stats.t.ppf(1 - alpha / 2, n - 1)

    # Intervalo de confianza para la media
    margin_of_error_mu = t_critical * standard_error
    confidence_interval_mu = (sample_mean - margin_of_error_mu, sample_mean + margin_of_error_mu)

    # Valores críticos chi-cuadrado para la varianza
    chi2_lower = stats.chi2.ppf(alpha / 2, n - 1)
    chi2_upper = stats.chi2.ppf(1 - alpha / 2, n - 1)

    # Intervalo de confianza para la varianza
    confidence_interval_var = (
        (n - 1) * sample_variance / chi2_upper,
        (n - 1) * sample_variance / chi2_lower
    )

    (confidence_interval_mu, confidence_interval_var)
```

```
((15.240696254641279, 18.15930374535872),
 (4.092636501481853, 17.965110906541934))
```