# Probabilidad y estadística

MAT 041, Primer semestre

Francisco Cuevas Pacheco

14 de noviembre de 2022

# Recuerdo: ¿Población o muestra?

**Definición (Población)**

*Una población es un conjunto que contiene la totalidad de individuos a ser estudiados por sus características. Una característica o variable estadística de interés se denota generalmente como $X$.*

**Definición (Muestra)**

*Una muestra es un subconjunto de una población seleccionada de acuerdo a algún método de muestreo. Generalmente, $n$ denotará el tamaño de la muestra y $X_1, X_2, \ldots, X_n$ las observaciones pertenecientes a la muestra.*

En el sentido estadístico, *población* se refiere al *todo* que es objeto de estudio y acerca del cuál se desea estudiar (cuyo tamaño se denota por $N$). En este mismo contexto, una *muestra* corresponde a un subconjunto, aleatorio o determinista, de esa población (cuyo tamaño se denota por $n$).

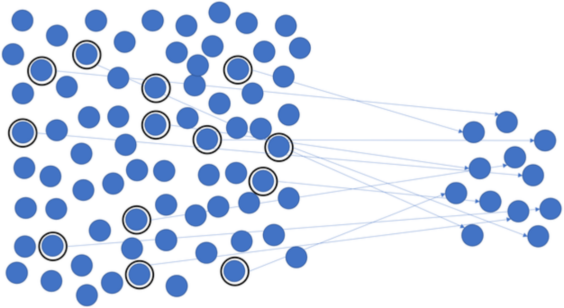

**Figura 1:** Población (Izquierda) y muestra (Derecha)

## El problema de la inferencia

La inferencia estadística es el proceso de usar el análisis de los datos para *inferir* propiedades de la distribución subyacente.

**Ejemplo**

*Con el propósito de promover una vida saludable se investigarán los hábitos deportivos de los* 5724 *estudiantes matriculados en un campus universitario. Por restricciones de tiempo y dinero se aplicará una encuesta sólo a* 120 *de ellos seleccionados al azar.*

Luego, la población serían los 5724 estudiantes, mientras que la muestra serían los 120 seleccionados

**Ejemplo (Cont.)**

*La encuesta registra la edad del estudiante, el sexo, la carrera en que está matriculado, si practica o no un deporte en forma regular, número de veces en la semana en que se practica deporte, horas semanales dedicadas a la práctica deportiva, etc.*

En este contexto, nos interesa la *población de datos*. Estos datos dependen de la característica aleatoria $X$ que se está midiendo en los individuos u objetos de la población.

**Ejemplo (Cont.)**

*La pregunta ¿usted practica o no un deporte de manera regular? tiene asociada una variable aleatoria $X$ con distribución Bernoulli de parámetro $p$, con $0 \leq p \leq 1$ y donde $p$ representa la probabilidad de que un estudiante cualquiera responda que el practica un deporte de forma regular.*

Notemos que $p$ puede ser interpretado como la proporción de estudiantes matriculados en ese campus que practican un deporte de forma regular. La muestra justamente tendrá como finalidad obtener información respecto al estado de actual de esa proporción $p$.

**Población y muestra**

Entonces, $X_1, X_2, ..., X_N$, representan la población de datos o respuestas que darían los $N = 5724$ estudiantes del campus y $X_1, X_2, ..., X_n$ es la muestra correspondiente a las respuestas que darían los $n = 120$ encuestados.

**Aleatoriedad de la respuesta**

Notemos además que no se sabe cuáles serán las respuestas y por tanto cada $X_i$; $i = 1, 2, ..., n$ es aleatoria. Después de responder las encuestas esas variables aleatorias se transforman en *datos* del tipo 0 ó 1 que se anotarán con minúsculas $x_1, x_2, ..., x_n$.

Por restricciones de tiempo, espacio y dinero a veces no es posible medir la caracteríatica $X$ en todos los individuos de una población y por tanto uno debe conformarse sólo con una muestra de mediciones.

Se busca que una muestra cumpla con al menos dos exigencias:

1. Debe se *representativa* de modo que los datos recolectados sean una imágen o espejo fiel de lo que sucede en la población
2. Debe ser *aleatoria* o al azar de modo que ningún individuo sea preferido en vez de otro. Todos deben tener la misma opción o posibilidad de ser seleccionados para la muestra.

**Ejemplo (Cont.)**

*Volviendo al ejemplo de la práctica deportiva de los estudiantes del campus universitario, la muestra es representativa si tiene la misma proporción de mujeres y hombres que en la población, si todas las edades y carreras están proporcionalmente representadas, etc.*

# Algunas hipótesis

Asumiremos que los elementos de la muestra satisfacen dos propiedades:

1. Independencia de las observaciones aleatorias $X_1, X_2, ..., X_n$
2. Que todas ellas tengan *igual modelo de probabilidad* $f(x;\theta)$ con los mismos parámetros $\theta$.

Esas dos exigencias facilitan la construcción del modelo de probabilidad conjunto para $X_1, X_2, ..., X_n$ y la derivación de los diversos métodos de inferencia estadística.

**Definición**

*Se dice que el conjunto de observaciones $X_1, X_2, ..., X_n$ es una **muestra aleatoria** si todas ellas son independientes entre si y todas tienen el mismo modelo de probabilidad.*

**Sinónimo**

Una muestra aleatoria también se suele estudiar como un conjunto de variables aleatorias *independientes e idénticamente distribuidas* (o iid)

# Inferencia: Introducción

Supongamos que estamos estudiando una población que sigue cierto modelo probabilístico dado por la función de densiad $f(x;\theta)$, donde $\theta$ puede ser un parámetro o un vectore de parámetros desconocido, entonces surge la pregunta:

¿Como podemos encontrar o estimar $\theta$ a partir de la información que podamos obtener de una muestra aleatoria?

Supongamos que estamos estudiando una población que sigue cierto modelo probabilístico dado por la función de densiad $f(x;\theta)$, donde $\theta$ puede ser un parámetro o un vectore de parámetros desconocido, entonces surge la pregunta:

¿Como podemos encontrar o estimar $\theta$ a partir de la información que podamos obtener de una muestra aleatoria?

Esto lo haremos mediante dos enfoques:

- Estimación puntual: Encontrar un único valor para $\theta$.
- Estimación intervalar: Entontrar un intervalo para $\theta$.

# Definiciones previas:

**Definición**

*Un **parámetro** poblacional, anotado* $\theta$*, es un número o un vector que representa alguna característica numérica de la población o del modelo* $f(x;\theta)$*.*

**Definición**

*El **espacio de parámetros** o **espacio paramétrico** , anotado* $\Theta \subset \mathbb{R}^d$*, es el conjunto de todos los valores posibles que puede asumir un parámetro o conjunto de parámetros.*

**Ejemplo**

*(Práctica deportiva) Si $X$ es la variable aleatoria asociada a la pregunta ¿Practicas algún deporte en forma regular?*

*entonces el modelo de probabilidad es Bernoulli($p$) con función de probabilidad (de cuantía o de masa) dada por*

$$f(x, p) = p^x(1 - p)^{1-x}; x \in 0, 1, 0 \leq p \leq 1$$

*Luego* $\theta = p$ *y* $\Theta = [0, 1]$

**Definición**

*Una* ***estadística muestral*** *$T$ es cualquier función de las variables que forman la muestra aleaoria. Se anota, $T = T(\underline{X}) = T(X_1, X_2, \ldots, X_n)$.*

Algunas estadísticas muestrales son:

| | |
|---|---|
| Total muestral | $t = \sum_{i=1}^{n} X_i$ |
| Media muestral | $\bar{X} = \sum_{i=1}^{n} X_i/n$ |
| Varianza muestral | $S_{n-1}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ |
| Máximo muestral | $X_{(n)} = \max\{X_1, X_2, ..., X_n\}$ |
| Mínimo muestral | $X_{(1)} = \min\{X_1, X_2, ..., X_n\}$ |

**Definición**

*Una **estadística muestral** $T$ es cualquier función de las variables que forman la muestra aleaoria. Se anota, $T = T(\underline{X}) = T(X_1, X_2, \ldots, X_n)$.*

Algunas estadísticas muestrales son:

| | |
|---|---|
| Total muestral | $t = \sum_{i=1}^{n} X_i$ |
| Media muestral | $\bar{X} = \sum_{i=1}^{n} X_i/n$ |
| Varianza muestral | $S_{n-1}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ |
| Máximo muestral | $X_{(n)} = \text{máx}\{X_1, X_2, ..., X_n\}$ |
| Mínimo muestral | $X_{(1)} = \text{mín}\{X_1, X_2, ..., X_n\}$ |

Observación: Al ser $T(\underline{X})$ una función de las variables $X_1, \ldots, X_n$ entonces $T$ es una variable aleatoria y por lo tanto tiene una densiad asociada. Cuando se reemplazan los datos $\underline{x} = (x_1, x_2, \ldots, x_n)$ en $T(\underline{X})$ se obtiene una **realización** de esa estadística que se anota $T = T(\underline{x})$ .

**Definición**

*Un* ***estimador*** *es cualquier estadística* $T = T(\underline{X})$ *empleada para estimar un parámetro* $\theta$*. Se anota* $\hat{\theta} = T(\underline{X})$ *y cuando se reemplazan los datos* $\underline{x} = (x_1, x_2, \ldots, x_n)$ *se obtiene una* ***estimación*** *de* $\theta$ *anotada* $\hat{\theta} = T(\underline{x})$*.*

Los métodos de estimación puntual que veremos en este curso son:

- Método de los Momentos
- Método de Máxima Verosimilitud

La idea natural de este métodos es estimar una media poblacional con una media muestral, estimar una varianza poblacional con la varianza de la muestra, etc. Más precisamente, el método estima los momentos poblacionales con los momentos muestrales.

**Definición**

*El* ***momento poblacional*** *de orden $r$ para la variable aleatoria $X$ es el número $\mu_r$ definido por*

$$\mu_r = E(X^r) = \begin{cases} \sum x^r p(x) & \textit{if } X \textit{es discreta} \\ \int_{-\infty}^{\infty} x^r f(x) dx & \textit{if } X \textit{es continua} \end{cases}$$

Recordemos que la media es $E(X) = \mu_1$ y la varianza es $Var(X) = \mu_2 - \mu_1^2$.

**Definición**

*Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria, entonces el **momento muestral** de orden $r$, anotado $m_r$, se dine como*

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

El **método de los momentos** igualar los momentos poblaciones con los momentos muestrales.

**Definición**

*Supongamos que el vector de parámetros* $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ *tiene* $k$ *componentes y que estos parámetros se pueden expresar en términos de los* $k$ *primeros momentos poblacionales*

$$\begin{aligned}
\theta_1 &= g_1(\mu_1, \mu_2, \ldots, \mu_k) \\
\theta_2 &= g_2(\mu_1, \mu_2, \ldots, \mu_k) \\
&\vdots \\
\theta_k &= g_k(\mu_1, \mu_2, \ldots, \mu_k)
\end{aligned}$$

*Entonces, sustituyendo los momentos poblacionales con los momentos muestrales, esto es haciendo* $\mu_1 = m_1, \mu_2 = m_2, \ldots, \mu_k = m_k$ *se obtienen los estimadores de momentos*

$$\begin{aligned}
\hat{\theta}_1 &= g_1(m_1, m_2, \ldots, m_k) \\
\hat{\theta}_2 &= g_2(m_1, m_2, \ldots, m_k) \\
&\vdots \\
\hat{\theta}_k &= g_k(m_1, m_2, \ldots, m_k)
\end{aligned}$$

**Ejemplo**

*Práctica deportiva (Cont) Si $X$ es la variable aleatoria asociada a la pregunta ¿Practicas algún deporte en forma regular?, encuentre el estimador de momentos, considerando que $X \sim Ber(p)$.*

**Ejemplo**

*Práctica deportiva (Cont) Si $X$ es la variable aleatoria asociada a la pregunta ¿Practicas algún deporte en forma regular?, encuentre el estimador de momentos, considerando que $X \sim Ber(p)$.*

**Solución:**

Dado que $X \sim Bernoulli(p)$, se tiene que la masa de probabilidad (de cuantía o de masa) está dada por

$$f(x, p) = p^x(1-p)^{1-x}; x \in 0, 1, 0 \leq p \leq 1$$

Se sabe entonces que:

$$\mathbb{E}[X] = p \text{ y que } m_1 = \bar{X}$$

luego, por el método de los momentos, se tiene que $\hat{p} = \bar{X}$.

**Ejemplo**

*Sea $X_1, \ldots, X_n$ una muestra aleatoria de tamaño $n$ con distribución Normal de media $\mu$ y varianza $\sigma^2$. Encuentre el estimador de momentos de $(\mu, \sigma^2)$.*

**Ejemplo**

*Sea $X_1, \ldots, X_n$ una muestra aleatoria de tamaño $n$ con distribución Normal de media $\mu$ y varianza $\sigma^2$. Encuentre el estimador de momentos de $(\mu, \sigma^2)$.*

**Solución:**

Si $X_i$ tiene una distribución Normal, entonces el estimador de momentos de $\mu$ y $\sigma^2$ se obtiene mediante resolver el sistema de ecuaciones $\mu_1 = m_1$ y $\mu_2 = m_2$. Luego se sabe que

$$\mu_1 = m_1 \implies \hat{\mu} = \bar{X}$$

$$\mu_2 = m_2 \implies \mu^2 + \sigma^2 = \frac{\sum_{i=1}^n X_i^2}{n}$$

reemplazando $\mu$ por $\hat{\mu}$ se obtiene que

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\end{aligned}$$

# Método de máxima verosimilitud

El método propone encontrar el valor del parámetro que hace más creible o verosímil la muestra que se está estudiando. Dicho de otra forma, buscamos maximizar la probabilidad de que los datos provengan de nuestro modelo.

Debemos notar que para una muestra aleatoria de tamaño $n$ con función de probabilidad $f(x;\theta)$, la función de probabilidad conjunta está dada por

$$f(x_1, x_2, ..., x_n, \theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

La función de verosimilitud se suele denotar por $L(\theta, \underline{x})$

Luego, definimos al estimador máximo verosímil, denotado por $\hat{\theta}_{ML}$, mediante

$$\hat{\theta}_{ML} = \mathrm{Argmax}_{\theta} L(\theta, \underline{x})$$

$$\hat{\theta}_{ML} = \mathrm{Argmax}_{\theta} \prod_{i=1}^{n} f(x_i, \theta)$$

## Estimador Máximo Verosímil

1. A veces es más fácil maximizar la función de **log-verosimilitud**: $l(\theta, \underline{x}) = lnL(\theta, \underline{x})$. Esto debido a que el logaritmo simplifica los cálculos transformando potencias en productos, productod en sumas y divisiones en restas. Además, la función $y = \ln x$ es continua y estrictamente creciente para $x > 0$. Por lo tanto, el valor de $\theta$ que maximiza $L(\theta; \underline{x})$ es el mismo que maximiza $l(\theta, \underline{x})$
2. Si el log-verosímil $l(\theta, \underline{x})$ es diferenciable, entonces el estimador máximo verosímil de $\theta$ se obtiene resolviendo la *ecuación de verosimilitud*:

$$\frac{\partial l(\theta, \underline{x})}{\partial \theta} = 0$$

**Ejemplo**

*(Práctica deportiva) Si $X$ es la variable aleatoria asociada a la pregunta ¿Practicas algún deporte en forma regular?*

*entonces el modelo de probabilidad es Bernoulli(p) con función de probabilidad (de cuantía o de masa) dada por*

$$f(x,p) = p^x(1-p)^{1-x}; x \in 0, 1, 0 \leq p \leq 1$$

*Encuentre la función de probabilidad conjunta de la muestra aleatoria $X_1, X_2, ..., X_n$. Luego, la función de probabilidad conjunta es,*

$$\begin{aligned} L(p, \underline{x}) &= f(x_1,p)f(x_2,p)\cdots f(x_n,p) \\ &= p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

$$l(p,\underline{x}) = \ln L(p,\underline{x}) = \sum_{i=1}^{n} x_i \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

Luego:

$$\frac{\partial l(p,\underline{x})}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}$$

Resolviendo $\frac{\partial l(p,\underline{x})}{\partial p} = 0$ se obtiene que

$$p^* = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$l(p,\underline{x}) = \ln L(p,\underline{x}) = \sum_{i=1}^{n} x_i \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

Luego:

$$\frac{\partial l(p,\underline{x})}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}$$

Resolviendo $\frac{\partial l(p,\underline{x})}{\partial p} = 0$ se obtiene que

$$p^* = \frac{\sum_{i=1}^{n} x_i}{n}$$

Verificamos que $p^*$ es máximo. En efecto:

$$\frac{\partial^2 l(p,\underline{x})}{\partial p^2} = -\frac{\sum_{i=1}^{n} x_i}{p^2} - \frac{n - \sum_{i=1}^{n} x_i}{(1-p)^2} < 0$$

**Ejemplo**

*Suponga ahora que $X$ es la variable aleatoria asociada a la pregunta "¿Cuántas horas a la semana dedicas a la práctica de algún deporte? entonces el modelo de probabilidad continuo podría ser el normal $N(\mu, \sigma^2)$ con función de probabilidad (de densidad) dada por*

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} - \infty < x < \infty, \sigma > 0$$

*Encuentre la función de probabilidad conjunta de la muestra aleatoria $X_1, X_2, ..., X_n$.*

**Solución:**

La función de probabilidad conjunta es,

$$\begin{aligned} L(\mu, \sigma^2, \underline{x}) &= f(x_1, \mu, \sigma^2) f(x_2, \mu, \sigma^2) \cdots f(x_n, \mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_1-\mu}{\sigma})^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_2-\mu}{\sigma})^2} \cdots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_n-\mu}{\sigma})^2} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i-\mu}{\sigma})^2} \end{aligned}$$

En consecuencia se tiene que

$$l(\mu, \sigma^2, \underline{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - \mu)}{\sigma^2}$$

Ahora, calculamos el gradiente $\nabla l(\mu, \sigma^2, \underline{x})$

$$\frac{\partial l(\mu, \sigma^2, \underline{x})}{\partial \mu} = \sum_{i=1}^{n}\frac{(x_i - \mu)}{\sigma^2}$$

$$\frac{\partial l(\mu, \sigma^2, \underline{x})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

Haciendo $\nabla l(\mu, \sigma^2, \underline{x}) = 0$ obtenemos que

$$\hat{\mu}_{ML} = \bar{X}$$

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