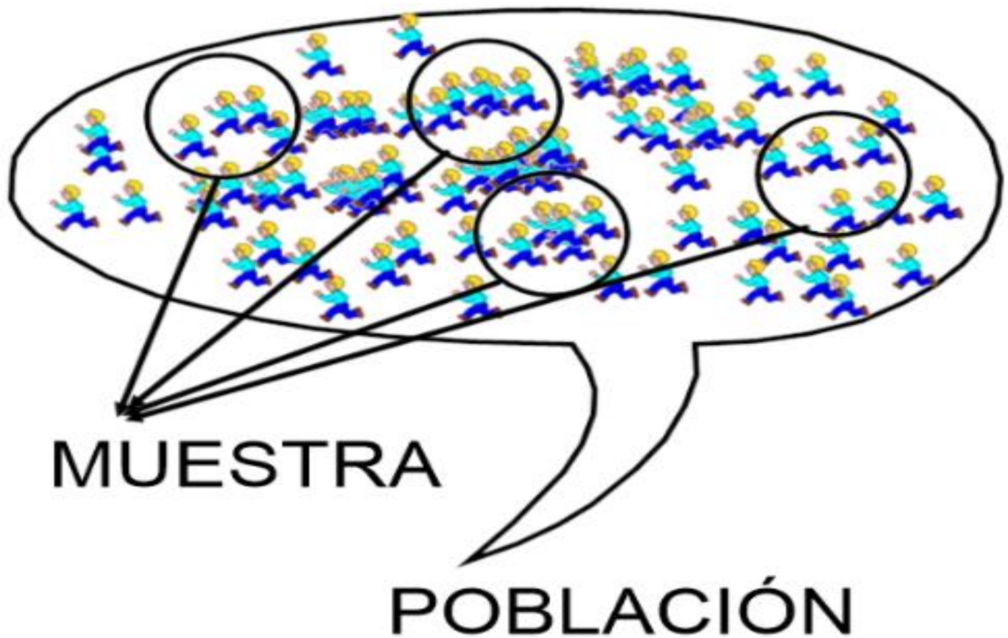
Unidad III

Muestras Aleatorias
Distribución de Muestreo
Parte I



El problema de interés

1) La **media poblacional** es un parámetro de gran interés en muchas situaciones prácticas

Por ejemplo, queremos conocer :

- el tiempo promedio de compilación de un programa .
- calificación promedio de alumnos que aprueban el segundo parcial de Modelos.
- el precio promedio mensual de una laptop.

2) El **desvío poblacional** es también un parámetro de gran interés en muchas aplicaciones

Por ejemplo, queremos conocer :

- el desvío del tiempo de ejecución de un algoritmo medido en una computadora de cierta marca.
- el desvío del tiempo que la CPU ocupa en ejecutar el código de aplicación.

Y una estimación del **error** en la aproximación

Muestreo

En muchas oportunidades uno desea conocer determinadas características de la población y en muchos casos resulta imposible o poco práctico estudiarla en su totalidad. Por diversos motivos:

- **Económicos de la experimentación.**
- **De tiempo.**
- **Por desconocer su tamaño.**
- **El hecho de que muchos de los métodos de medida son destructivos.**
- **Por no disponer de tecnología suficiente para acceder a ciertas poblaciones.**

Por ello, es necesario utilizar una muestra de observaciones tomadas de la población de interés con objeto de obtener conclusiones sobre ella.

Muestreo

Ejemplo

- Si se intenta determinar la duración promedio de una cierta marca de baterías para laptop sería imposible probarlas todas si no se quiere dejar de vender ninguna y además, requeriría mucho tiempo.

- Si un biólogo desea evaluar algunas características de una determinada especie de pez del mar Argentino resultaría imposible medir a toda la población, no sólo por una cuestión económica, sino también, por que difícilmente se conozca el tamaño de la población, se posean los medios suficientes para lograr acceder a toda la población, etc.

Inferencia estadística

La **Inferencia Estadística** es aquella rama de la Estadística mediante la cual se trata de sacar conclusiones de una población en estudio, a partir de la información que proporciona una muestra representativa de la misma.

Con frecuencia se está tentado a elegir una **muestra** seleccionando a los miembros más convenientes de la población. Tal procedimiento puede conducir a **inferencias erróneas** respecto de la misma, es decir, inferencias que **sobreestiman o subestiman** alguna característica de la población, en este caso el procedimiento se dice que esta **sesgado.**

Para eliminar cualquier posibilidad de sesgo la muestra a seleccionar debe ser:

- **Aleatoria:** el azar determina las observaciones que se seleccionan ,
- Las observaciones **independientes** entre si**.**
- **Representativa de la población:** en el sentido de que debe tener una composición similar en cuanto la proporción de distintas características de la población. Nunca se podrá estar totalmente seguro de que el resultado sea una muestra representativa, pero sí se puede actuar de manera que esta condición se alcance con una probabilidad alta. Esta garantizada la representatividad con la elección correcta del método de muestreo.

Por ejemplo, una muestra para un estudio de estaturas no incluirá solamente individuos bajos o solamente altos, sino individuos de ambas clases en proporciones similares a las de la población.

Tipo de muestreo

Muestreo aleatorio simple:

En una muestra aleatoria simple de tamaño n, cada elemento de la población tiene la misma probabilidad de ser elegido para formar parte de la misma.

El procedimiento empleado es el siguiente:

1) se asigna un número a cada individuo de la población.

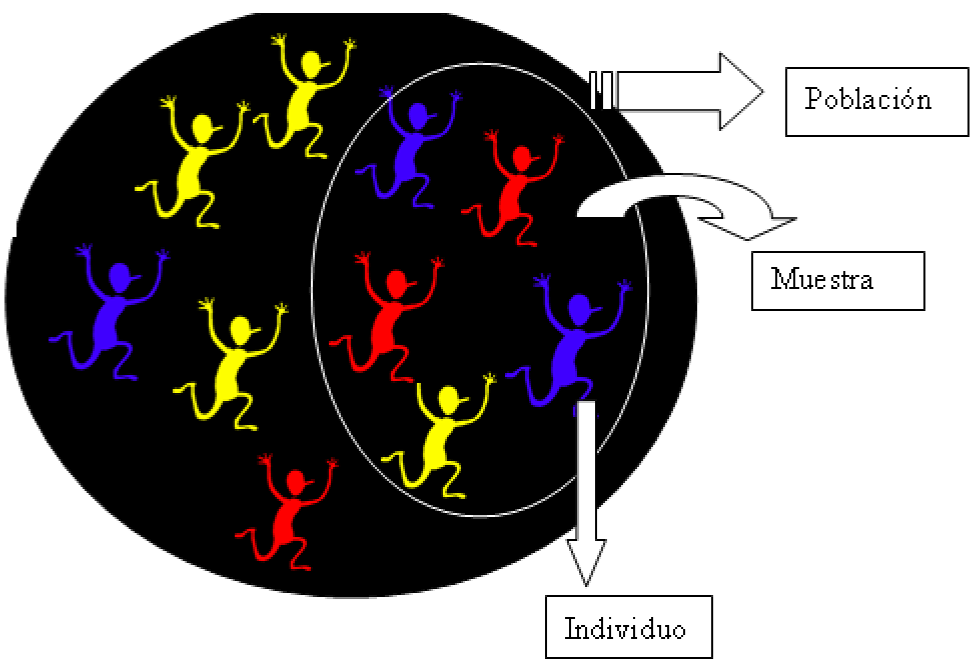
2) a través de algún medio mecánico (*bolillero, tablas de números aleatorios, números aleatorios generados con una calculadora o una computadora, etc*.) se eligen tantos sujetos como sea necesario para completar el tamaño de muestra requerido.

3) Los elementos de la muestra se eligen en forma independiente unos de otros.

Este procedimiento, atractivo por su ***simpleza***, por la ***representatividad*** *de la muestra obtenida* y ***produce estimadores*** de los parámetros desconocidos ***próximos a los valores reales*** de los mismos. Sin embargo, no tiene utilidad práctica cuando la población que estamos manejando es *muy grande*.

Ejemplo

Supongamos que tenemos una población de **50 000** individuos, y que disponemos de un listado con sus nombres. Si queremos elegir **100** personas, es necesario una ***computadora*** o un ***bolillero*** o una ***tabla de números al azar*** para elegir a **100** individuos de los **50 000** disponibles.



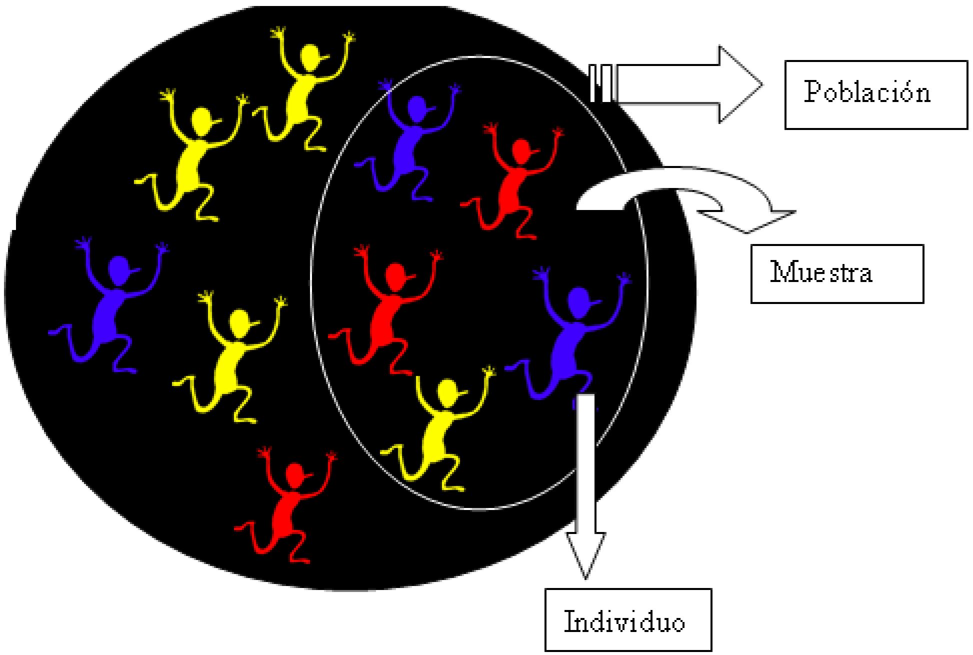
Ejemplo

Numeraremos a los individuos de la población de **1** a **50000**.

Individuo 1 = 00001

Individuo 2 = 00002

Individuo 3 = 00003



.

.

.

Individuo 50000 = 500000

Tabla de Números Aleatorios

Comenzamos en **fila 1** y **columna 2:**

94699605300659384844309201995073688448048419900606989883876735114039395681214 25

77911871138564403555252065841181918542313340633396153447124597459208947297368412

0022943510984165069238518680855438661686648568774340590536486570684260862998711

1811548918339486630565309291983094510520900386680423119312206916780715822056911

6503836967674354492782501506725781021430984732088859208823338144458466089028979

Elemento de la muestra aleatoria:

1 = Individuo 46996
2 = Individuo 05300

Ejemplo

Los individuos elegidos al azar:

Comenzamos en **fila 1** y **columna 2:**

1 = Individuo 46996
2 = Individuo 05300
3 = Individuo 48443
4 = Individuo 09201

.

.

.

Se continúa hasta lograr tener las 100 observaciones

Muestreo aleatorio simple

Poblaciones finitas

El muestreo **aleatorio simple** se realiza “con reemplazo”, es decir: Se selecciona un elemento de la población al azar, se observa el valor de la variable aleatoria **X**, *se devuelve a* la población y se vuelve a seleccionar otro elemento. Así hasta obtener los ***n*** *elementos. Este procedimiento* garantiza la **independencia** de las observaciones.

Poblaciones infinitas

El muestreo **aleatorio simple** se realiza “sin reemplazo”, es decir: Se selecciona un elemento de la población al azar, se observa el valor de la variable aleatoria **X,** no se devuelve a la población y se vuelve a seleccionar otro elemento. Así hasta obtener los ***n*** elementos. En este caso, la **independencia** de las observaciones está garantizada por la dimensión de la población.

Trabajaremos:

- con poblaciones infinitas, es decir, tamaños desconocidos.

- con muestreo aleatorio simple, que garantiza una muestra **representativa** de la población y la obtención de observaciones **independientes**.

Las observaciones de una muestra se obtienen al observar la característica medible X de manera independiente bajo las mismas condiciones, n veces. La muestra se denota:

$$X_1, X_2, \ldots, X_n.$$

donde

X_i la v.a. que representa la i-ésima observación de la v.a. X,

Los valores numéricos obtenidos son $x_1, x_2, \ldots, x_n$ de una muestra particular .

Si se considera la v.a X= "***área de un lote elegido al azar (en unidades cuadradas)***"

Muestra Aleatoria Simple
de tamaño n = 5 de la población (v.a. X)

⇨

$X_1, X_2, \ldots, X_5.$

Muestra de tamaño n = 5 de Camila:

$x_1 = 1u^2$ $x_2 = 8u^2$ $x_3 = 9u^2$ $x_4 = 8u^2$ $x_5 = 12u^2$

Muestra de tamaño n = 5 de Nicolás:

$x_1 = 9u^2$ $x_2 = 6u^2$ $x_3 = 8u^2$ $x_4 = 16u^2$ $x_5 = 3u^2$

Las observaciones de una muestra se obtienen al observar la característica medible X de manera independiente bajo las mismas condiciones, n veces. La muestra se denota:

$$X_1, X_2, \ldots, X_n.$$

donde

X_i la v.a. que representa la i-ésima observación de la v.a. X,

Definición de Muestra Aleatoria Simple

Una muestra aleatoria de tamaño *n* de la v.a. X, la constituyen un conjunto de *n* v.a. $X_1, X_2, \ldots, X_n$, que satifacen las siguientes propiedades:

- **las X_i son v.a. independientes entre sí.**
- **todas las X_i tienen la misma distribución de probabilidad que la v.a. X.**

El propósito principal de una muestra aleatoria es obtener información sobre los parámetros no conocidos de la población.

Ejemplos

1. Supongamos que deseamos conocer el gasto promedio semanal en fotocopias de un alumno de la UNS. Resulta poco práctico interrogar a cada uno de los alumnos de la universidad para determinar el valor de μ. Para hacer una inferencia con respecto al verdadero **gasto promedio** semanal por alumno, μ, un procedimiento más razonable consiste en seleccionar una muestra aleatoria de un tamaño apropiado y utilizar la media de las observaciones obtenidas, es decir,

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

La media muestral es una función de los valores observados en la muestra aleatoria. Por lo tanto es una v.a.

X = "gasto semanal en fotocopias de un alumno de la UNS (en pesos)",

$\overline{X}$ = **"gasto promedio semanal en fotocopias de n alumnos de la UNS elegidos al azar (en pesos)"**

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

v.a. media muestral

Si seleccionamos 5 alumnos al azar: $\overline{X} = \frac{X_1 + X_2 + \cdots + X_5}{5}$

$\overline{X}$ **= "gasto promedio semanal en fotocopias de 5 alumnos de la UNS elegidos al azar (en pesos)"**

Muestra 1: x_1= 5000\$ x_2= 5600\$ x_3= 2301 x_4 = 800\$ y x_5 = 3302,4\$

$$\overline{x}_1 = 4840{,}68\$$$

Si seleccionamos otros 5 alumnos al azar:

Muestra 2: x_1= 3400\$ x_2= 2904\$ x_3= 3007\$ x_4 = 2808\$ y x_5 = 3101\$

$$\overline{x}_2 = 3044\$$$

Se muestran dos de los valores posibles de la v.a. gasto promedio de 5 alumnos seleccionados al azar

2. Si deseamos conocer la **variabilidad del gasto** semanal en fotocopias de un alumno de la UNS. Resulta poco práctico interrogar a cada uno de los alumnos de la universidad para determinar el valor de σ^2**.** Para hacer una inferencia con respecto a la verdadera varianza del gasto semanal por alumno, σ^2, un procedimiento más razonable consiste en seleccionar una muestra aleatoria de un tamaño apropiado y utilizar la varianza de las observaciones obtenidas, es decir,

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

La varianza muestral es una función de los valores observados en la muestra aleatoria. Por lo tanto es una v.a.

X = "gasto semanal en fotocopias de un alumno de la UNS (en pesos)"

S^2 = "varianza del gasto semanal en fotocopias de n alumnos seleccionados al azar de la UNS (en pesos^2)"

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

v.a. varianza muestral

S^2 = "***varianza del gasto semanal en fotocopias de 5 alumnos seleccionados al azar de la UNS (en pesos^2)***"

Si seleccionamos 5 alumnos al azar:

Muestra 1: x_1= 5000\$ x_2= 5600\$ x_3= 2301 x_4 = 8000\$ y x_5 = 3302,4\$

La varianza muestral del gasto semanal seria : $s^2 = 2202{,}24^2\ \$^2$

Si seleccionamos otros 5 alumnos al azar :

Muestra 2: x_1= 3400\$ x_2= 2904\$ x_3= 3007\$ x_4 = 2808 \$ y x_5 = 3101\$

La varianza muestral del gasto semanal es: $s^2 = 277{,}29^2\ \$^2$

Se muestran dos de los valores posibles de la v.a. varianza muestral del gasto de 5 alumnos seleccionados al azar.

Estadístico

Definición:
Cualquier función de las variables aleatorias que componen una muestra aleatoria se llama estadístico.

$$Estadístico = f(X_1, \cdots, X_n)$$

donde , X_1, X_2, . . ., X_n constituyen una muestra aleatoria de tamaño *n*

Observación Importante!!!!

Un **estadístico** es una **v.a.** por ser **función de variables aleatorias**. Es decir, los valores de un estadístico varían de muestra a muestra.

Ejemplos de estadísticos Importantes:

1. Si X_1, X_2, ..., X_n representa una muestra aleatoria de tamaño *n* de una v.a. X se define **media muestral**, y se denota: $\overline{X}$

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

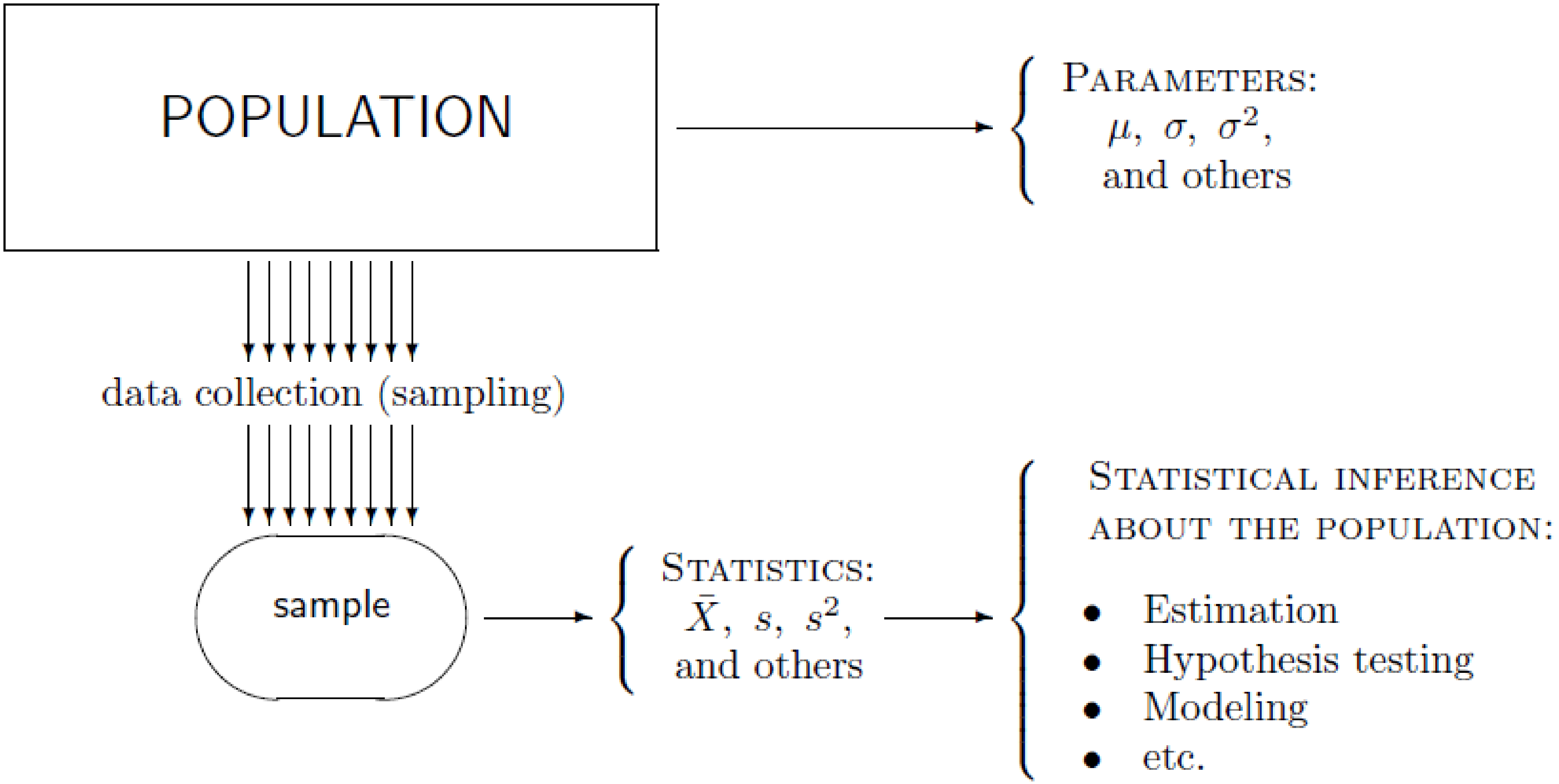
2. Si X_1, X_2, ..., X_n representa una muestra aleatoria de tamaño *n* de una v.a. X, se define **varianza muestral** y se denota: $\mathbf{S^2}$, al estadístico:

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

3. Si X_1, X_2, ..., X_n representa una muestra aleatoria de tamaño *n* de una v.a. X, se define el **desvío estándar muestral** y se denota: **S**, al estadístico:

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

Población, Parámetros, Muestra y estadísticos



Distribución de muestreo

Puesto que un estadístico es una v.a. tiene una distribución de probabilidad, que se conoce como distribución de muestreo.

La distribución de muestreo de un estadístico depende:

- de la distribución de probabilidad de la v.a. **X** (de la población),
- del tamaño de la muestra ,y
- del Tipo de Muestreo utilizado para seleccionar la muestra.

En la mayor parte de los casos supondremos que nuestra población, X, tiene distribución Normal. El muestreo siempre será considerado aleatorio simple.

Estadístico:

Media muestral

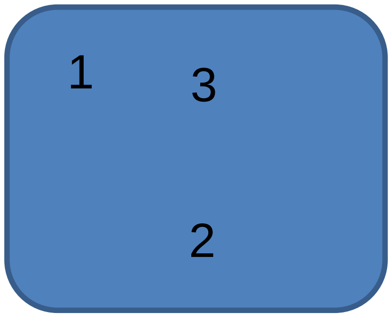
Distribución de muestreo

Construcción de una v.a. Media muestral

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Ejemplo: Tenemos una población con los siguientes $N = 3$ elementos:

N = 3 estudiantes



X= «N° de computadoras que posee un estudiante en su casa»

R_X $= \{1, 2, 3\}$

X	1	2	3
$P(X = x)$	$1/3$	$1/3$	$1/3$

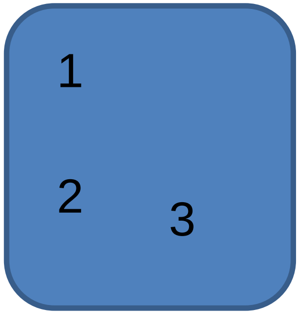
donde $\mu = 2$ y $\sigma^2 = 0,67$

En efecto,

E(X) = **μ** =

Distribución de muestreo

N = 3 estudiantes



X= «*N° de computadoras que posee un estudiante en su casa*»

X	1	2	3
$P(X = x)$	$1/3$	$1/3$	$1/3$

donde $\mu = 2$ y $\sigma^2 = 0,67$

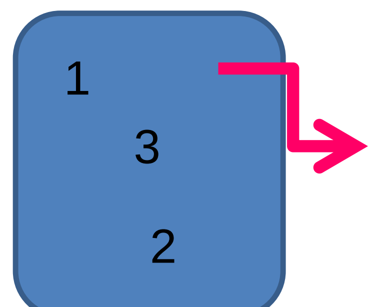
En efecto,

E(X) =

V(X) = σ^2 = E(X^2) - [E(X)]^2= ( ) - ( )^2

=

N = 3 estudiantes



n =2
Con
repos.

Se extraen muestras de $n = 2$ elementos:
Con reposición tenemos 9 posibles muestras:

$(1,1)\ (1,2)\ (1,3)\ (2,1)\ (2,2)\ (2,3)\ (3,1)\ (3,2)\ (3,3)$

v.a. Media muestral

Se extraen muestras de $n = 2$ elementos:
Con reposición tenemos 9 posibles muestras:

$$\overline{X} = \frac{X_1 + X_2}{2}$$

$(1,1)\ (1,2)\ (1,3)\ (2,1)\ (2,2)\ (2,3)\ (3,1)\ (3,2)\ (3,3)$

En cada una de las muestras se puede calcular la media muestral:

- Las medias muestrales ($\bar{x}$) serían: 1; 1,5; 2; 1,5; 2; 2,5; 2; 2,5 y 3.

$\bar{x} = \frac{1+1}{2} = 1$ $\bar{x} = \frac{1+2}{2} = 1.5$ $\bar{x} = \frac{1+3}{2} = 2$ Y así siguiendo

	(1, 1)	(1, 2)	(1, 3)	(2, 1)	(2, 2)	(2, 3)	(3, 1)	(3, 2)	(3, 3)
$\bar{X}$	1	1,5	2	1,5	2	2,5	2	2,5	3

$$R_{\bar{X}} = \{1, 1.5, 2, 2.5, 3\}$$

Por lo tanto la media es una variable aleatoria que puede adoptar diferentes valores

$\overline{X}$ = *«N° promedio de computadoras que poseen en su casa 2 estudiantes elegidos al azar»* **Es una v.a.**

Distribución de la v.a. Media Muestral

$\overline{X}$ = *«N° promedio de computadoras que poseen en su casa 2 estudiantes elegidos al azar»*

	$(1,1)$	$(1,2)$	$(1,3)$	$(2,1)$	$(2,2)$	$(2,3)$	$(3,1)$	$(3,2)$	$(3,3)$
$\bar{X}$	1	$1,5$	2	$1,5$	2	$2,5$	2	$2,5$	3

$$\mathbf{P(\overline{X} = 1)}$$

$$\mathbf{P(\overline{X} = 1.5)}$$

$$\mathbf{P(\overline{X} = 2)}$$

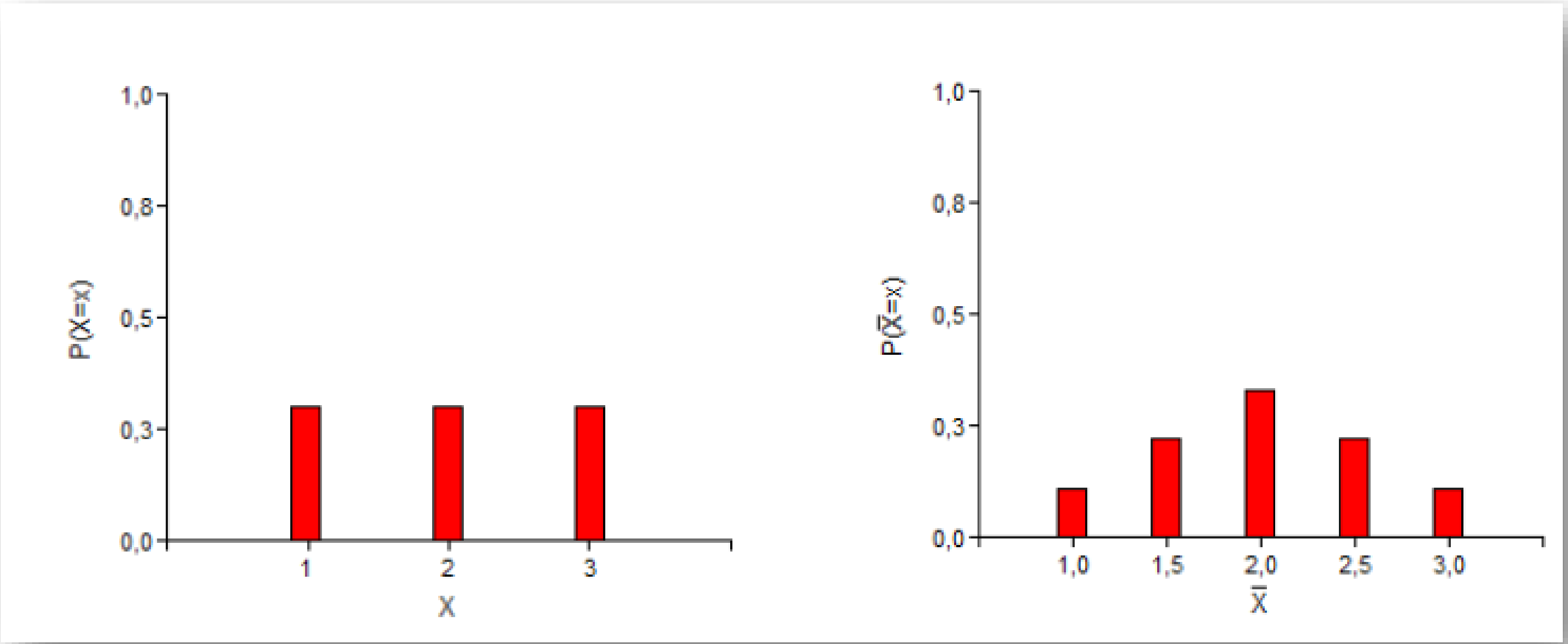
$$\mathbf{P(\overline{X} = 2.5)}$$

$$\mathbf{P(\overline{X} = 3)}$$

Funciones de distribución de las v.a X y la v.a media muestral $\bar{X}$

X	1	2	3
$P(X=x)$	$1/3$	$1/3$	$1/3$

$\bar{X}$	1	1,5	2	2,5	3
$P(\bar{X}=x)$	$1/9$	$2/9$	$3/9$	$2/9$	$1/9$



Calcular $E(\bar{X})$ y $V(\bar{X})$

Función de distribución de probabilidad de la v.a. media muestral

$\bar{X}$	1	1,5	2	2,5	3
$P(\bar{X} = x)$	1/9	2/9	3/9	2/9	1/9

$$\mathbf{E(\overline{X})} = \sum_{\overline{x} \in R_{\overline{X}}} \vec{x} P(\overline{X} = \overline{x}) =$$

$$\mathbf{V(\overline{X})} = \sum_{\overline{x} \in R_{\overline{X}}} \vec{x}^2 P(\overline{X} = \overline{x}) - \left[E(\overline{X})\right]^2 = \quad - \quad {}^2 =$$

Luego tenemos que :

E($\overline{X}$ **) =**

V($\overline{X}$ **) =**

Estadístico: Media muestral

Sea $\mathbf{X_1, X_2, ..., X}_n$ una muestra aleatoria de una v.a. **X** con media finita μ y varianza finita σ^2. Por lo tanto, las ***n*** variables aleatorias son **independientes** y poseen la **misma distribución** que la v.a. **X,** es decir, $\mathbf{E(X_i) = \mu}$ **y** $\mathbf{V(X_i) = \sigma^2}$.

El **valor esperado** y la **varianza** del estadístico **media muestral**, $\overline{X}$ son:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

v.a. media muestral

$$E(\overline{X}) = E\left(\frac{X_1 + ... + X_n}{n}\right) = E\left(\frac{1}{n}(X_1 + ... + X_n)\right) = \frac{1}{n}\left(E(X_1) + ... + E(X_n)\right) = \frac{1}{n} n\mu = \mu$$

y

$$V(\overline{X}) = V\left(\frac{X_1 + ... + X_n}{n}\right) = V(\frac{1}{n}(X_1 + ... + X_n)) = \frac{1}{n^2}\left(V(X_1) + ... + V(X_n)\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Distribución de muestreo

$$E(\bar{X}) = \mu \quad y \quad V(\bar{X}) = \frac{\sigma^2}{n}$$

Importante!!!! Este resultado es válido sin importar la distribución de probabilidad de la v.a. de interés, X, siempre que la varianza tenga un valor finito.

De la definición de V $(\bar{X})$ se deduce que el desvío estándar es

$$\sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Recibe el nombre de desvío estándar o error estándar de la media muestral.
Conforme el tamaño de la muestra crece, el desvío estándar, y por lo tanto la varianza, de $\bar{X}$ decrece.

Función de distribución de probabilidad de la v.a. media muestral es:

$\bar{X}$	1	1,5	2	2,5	3
$P(\bar{X} = x)$	1/9	2/9	3/9	2/9	1/9

Se tiene que:

$$\boxed{\mu_{\bar{X}} = E(\bar{X}) = \qquad\qquad = E(X) = \mu_X}$$

$$\boxed{V(\bar{X}) \cong \qquad\qquad = \frac{V(X)}{n} = \frac{\sigma_X^2}{n} \cong \underline{\qquad\qquad}}$$