Inferencia Estadística

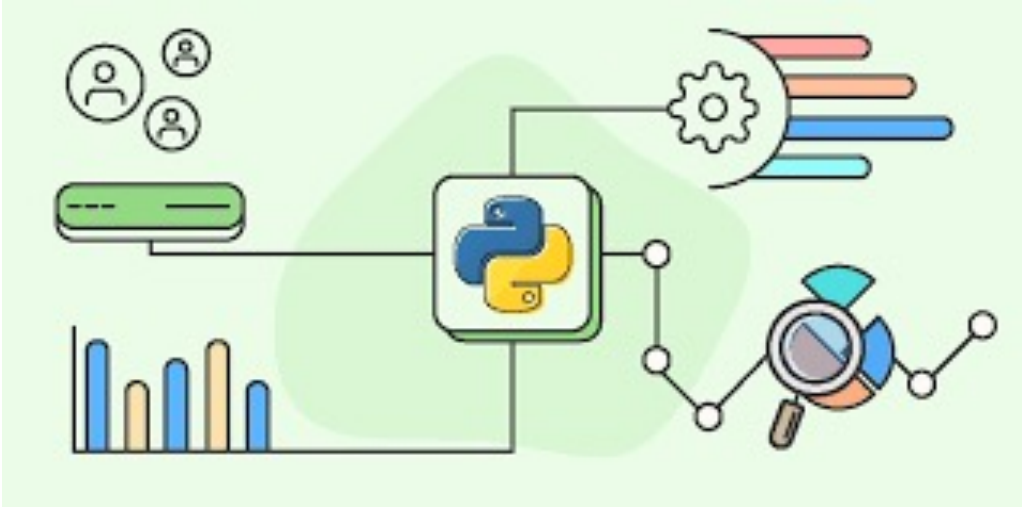
Muestras y Muestreo

Especialización en Ciencia de Datos

2021

Objetivos

- Utiliza los conceptos básicos de estadística Inferencial
- Aprender sobre técnicas de muestreo
- Realizar cálculos de probabilidad utilizando la distribución muestral para resolver un problema.

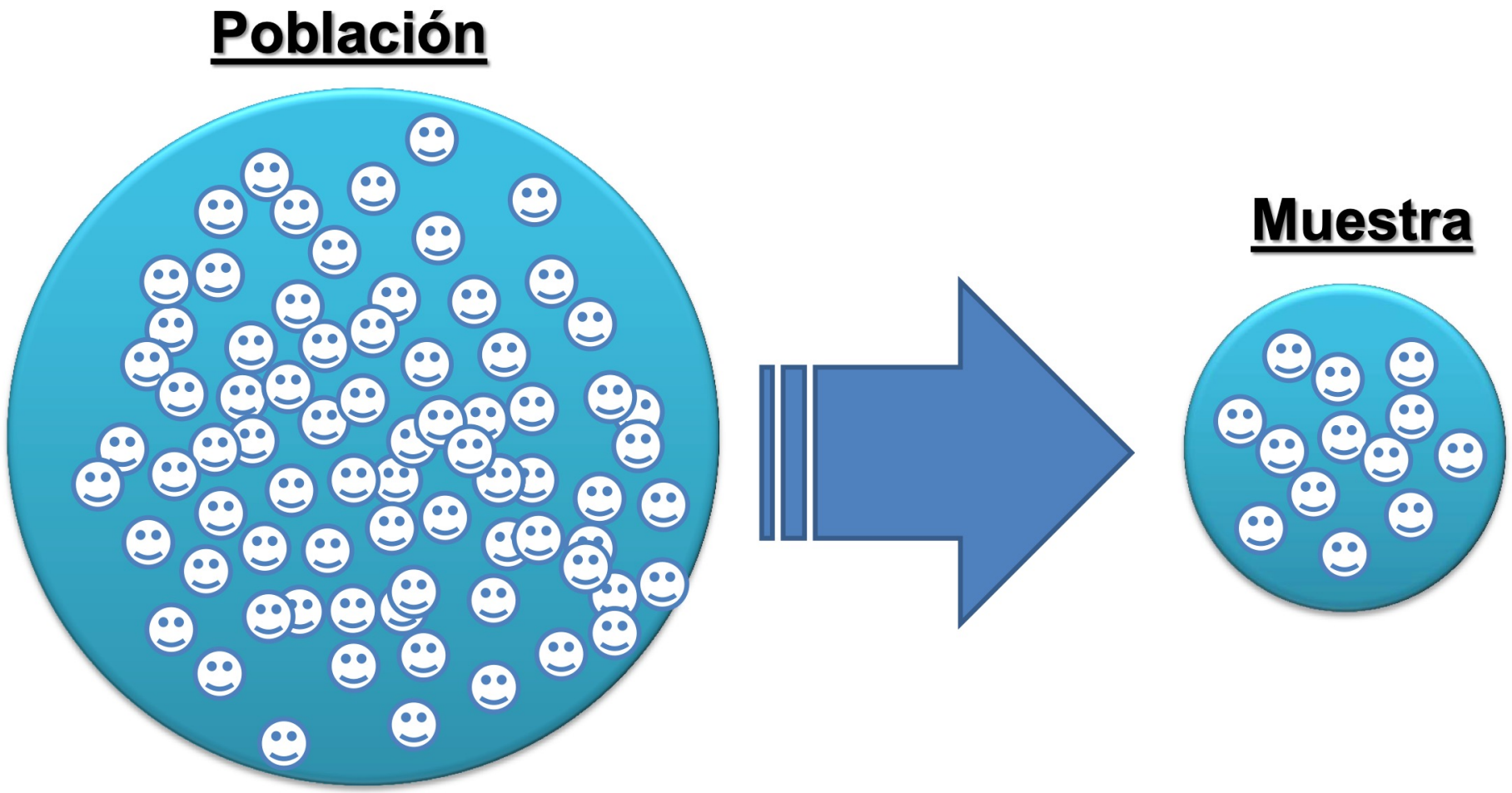


Contenido:

1. Muestras representativas
2. Muestreo aleatorio
3. Muestreo sistémico
4. Muestreo estratificado
5. Muestreo por conglomerados
6. Teorema del Límite Central
7. Dsitribución muestral

1. Muestras

Muestra poblacional



Población

Los términos ***población*** (o *universo*) y ***muestra*** son términos **relativos**. La ***población*** es **el conjunto de elementos (sujetos, objetos o indicadores) que presentan determinada característica o propiedad en común**, que el investigador **quiere analizar** al realizar la investigación, y que satisfacen un conjunto predeterminado de **criterios** establecidos (definidos) por el analista. Es decir, son **los "casos" investigados**, que pueden ser personas, animales, registros de cualquier tipo, muestras de laboratorios, etc., pero que son siempre **elementos que comparten una determinada característica predefinida por el investigador**, en base a la cual se agrupan en una determinada población.

Muestra Poblacional

El analista debe **definir** precisamente los **criterios** que permitan decidir , ante cada caso o elemento, **si pertenece o no a la población investigada**, es decir, debe determinar estrictamente el ***marco muestral*** o los límites de la población.

Cuando el tamaño de la población es **muy grande**, la investigación no se realiza en toda su extensión , sino en un ***subconjunto*** o parte de ella , denominada ***muestra***, y después **se *generalizan* los hallazgos** obtenidos **a toda la población**.

La muestra debe ser **representativa** de la población.

Muestra Poblacional

La muestra es el **subconjunto** de la población donde se efectúa o lleva a cabo la investigación con la finalidad de **generalizar** posteriormente los resultados a toda la población.

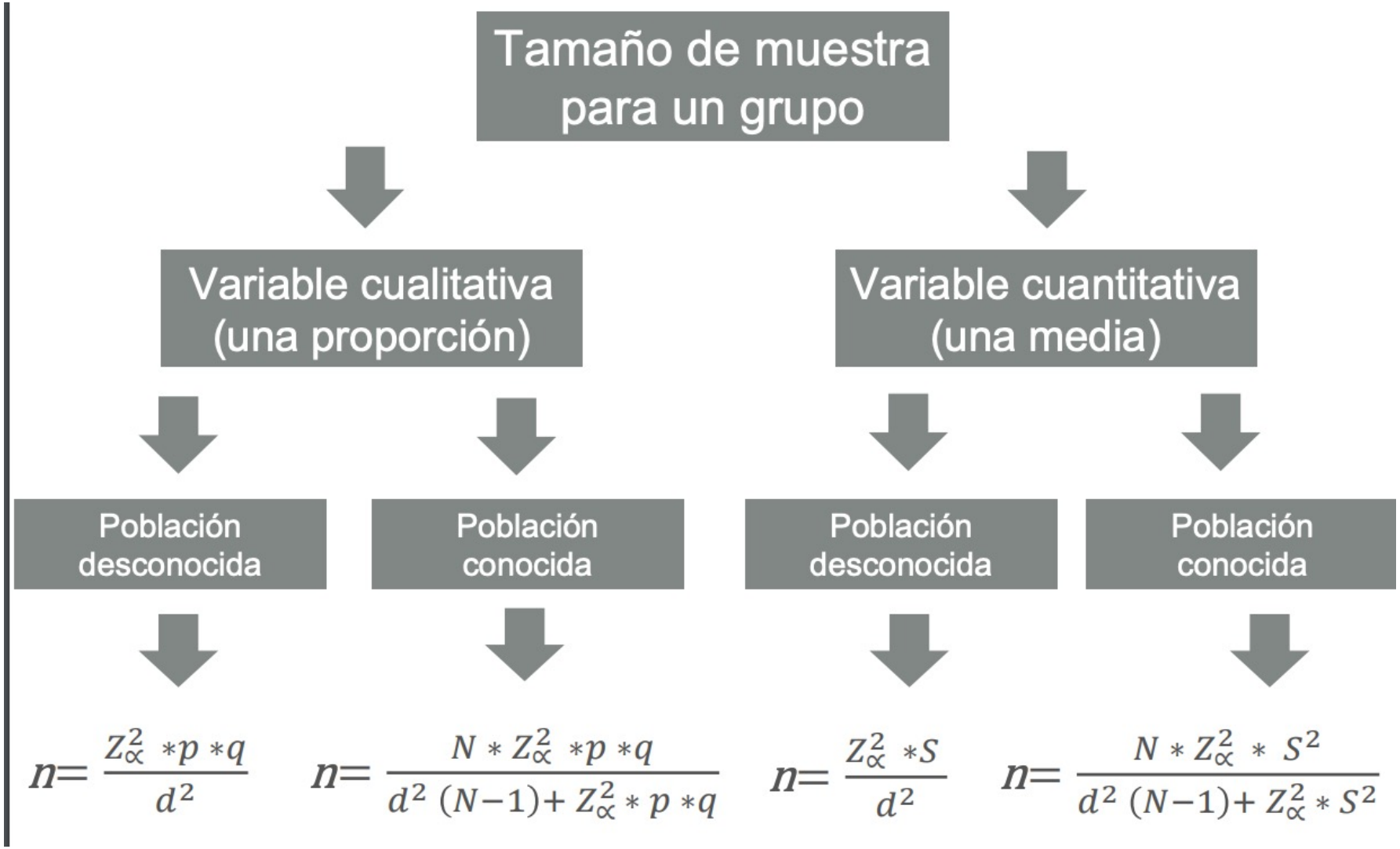
Para que dicha generalización sea lícita, la muestra debe poseer ***las mismas (o muy similares) características básicas (relevantes) de la población investigada,*** es decir, debe ser ***representativa*** de la población.

Muestra Poblacional

Si **N** es el tamaño de la población y **n** el tamaño de la muestra,

siendo **N** suficientemente **grande** , pueden extraerse un cierto número de muestras ***distintas*** de tamaño **n**. Si en cambio **N** es un número **pequeño** (por ejemplo, 30 o 40 casos) , convendrá **analizar directamente a toda la población**, es decir, **no extraer una muestra** o subconjunto.

Tamaño de la muestra



Cómo calcular el tamaño de muestra para una población finita

$$n = \frac{N * Z_{\alpha}^{2} * p * q}{e^{2} * (N - 1) + Z_{\alpha}^{2} * p * q}$$

n = Tamaño de muestra buscado

N = Tamaño de la Población o Universo

z = Parámetro estadístico que depende el Nivel de Confianza (NC)

e = Error de estimación máximo aceptado

p = Probabilidad de que ocurra el evento estudiado (éxito)

q = (1 - p) = Probabilidad de que no ocurra el evento estudiado

Tamaño de la muestra

Tu nivel de confianza corresponde a una puntuación Z. Este es un valor constante necesario para esta ecuación. Aquí están las puntuaciones Z para los niveles de confianza más comunes:

90% - Puntuación Z = 1,645

95% - Puntuación Z = 1.96

99% - Puntuación Z = 2.576

Supongamos que nos piden calcular el tamaño para una población de 543.098 consumidores de una marca de bebidas energéticas, donde el investigador asigna un nivel de confianza de 95% y un margen de error de 3%. Donde se desconoce la probabilidad “p” del evento.

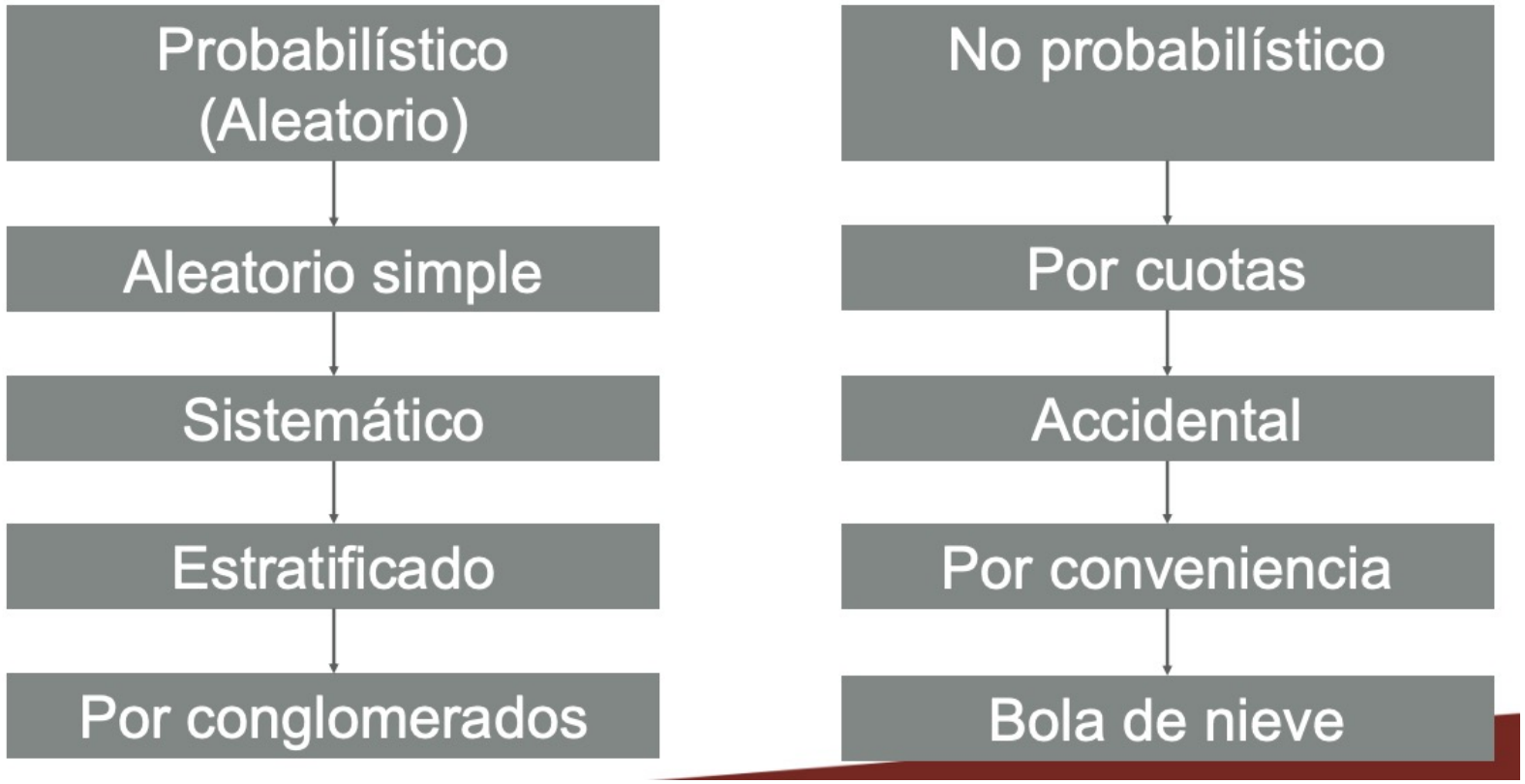
Basándonos en este ejemplo, y en nuestra fórmula, el "N" será 543.098, nuestro Z será 1.96 (recuerda que el investigador asignó un nivel de confianza de 95%) y “e” será de 3%. Y como nuestro ejemplo dice que se desconoce la probabilidad de que ocurra el evento, se asigna un 50% a "p" y un 50% a "q".

El resultado de nuestro tamaño de muestra sería: 1065.2, y tendría que ser redondeado pues estamos hablando de personas.

Tipos de Muestreo

La **representatividad** de la muestra tiene que ver, entonces, con que ésta **posea aproximadamente las mismas características básicas** que posee la población. Y esto, a su vez, tiene que ver con ***la manera de seleccionar u obtener la muestra*** (es decir, con los ***procedimientos de extracción*** de la muestra) y con el ***tamaño*** *de la muestra.* Distintos **procedimientos de obtención de muestras** definen distintos ***tipos de muestreo***. En general, los muestreos se califican en ***probabilísticos*** y ***no probabilísticos***.

Tipos de Muestreo



todos y cada uno de los elementos que integran la población tienen la misma probabilidad conocida de ser seleccionados

cuando **no todos** tienen la misma posibilidad de ser elegidos, o esta probabilidad **no se conoce**

Tipos de Muestreo

Pese a que nunca hay garantías absolutas de representatividad, en los muestreos probabilísticos el **error de muestreo**, es decir, **el margen de error o riesgo de equivocarse al generalizar los resultados** obtenidos en la muestra a toda la población, **puede calcularse.** Este margen de error ("desviación estándar") **se define de antemano**. En los muestreos **no probabilísticos** el margen de error **se desconoce**, y por ende **no puede calcularse**.

Ventajas del Muestreo

¿Por qué calcular el tamaño de la muestra?

- Una muestra puede estudiarse con mayor rapidez que una población.
- El estudio de una muestra es menos costoso.
- Toma menos tiempo el estudio a realizar.
- Los resultados son mas precisos.

¿Cuándo calcular el tamaño de la muestra?

- Cuando no se puede estudiar toda la población
- Cuando se quieren estudiar dos o mas grupos y establecer diferencias.
- Cuando se quieren estimar parámetros, prevalencia, promedio, porcentajes y tasas.

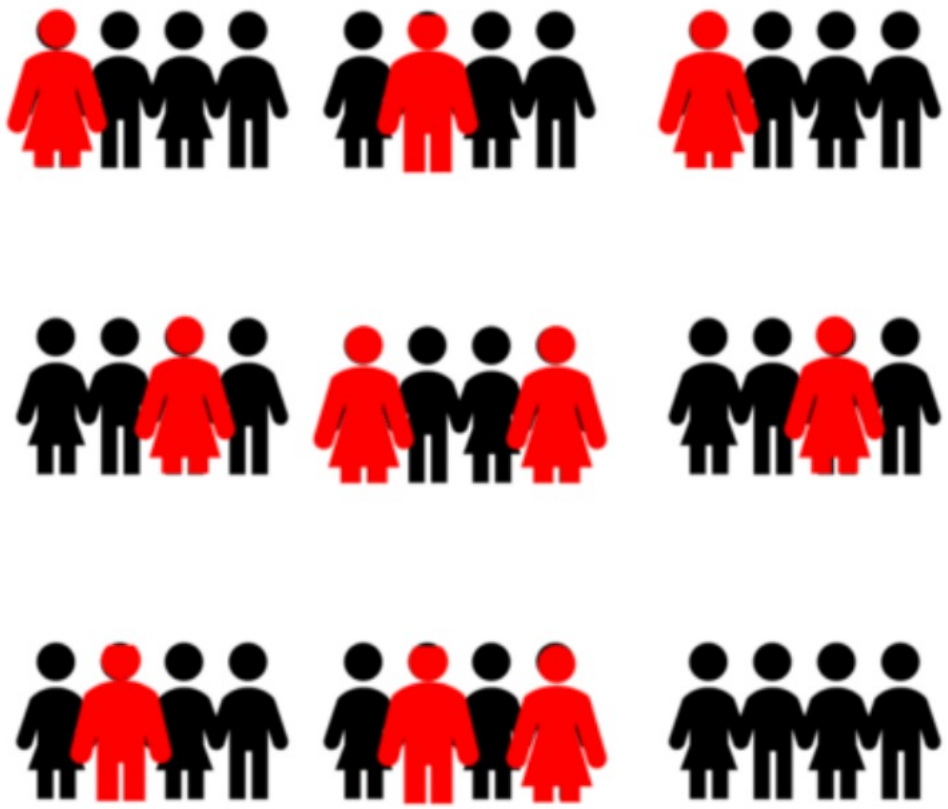
permite ***profundizar más el análisis de las variables*** involucradas en el fenómeno investigado ; permite ***mayor control*** **de dichas variables.**

Unidad de observación y de muestreo

La ***unidad de observación*** es cada uno de los **elementos** (sujetos, objetos o indicadores) **que integran la población**, y en los que **se analizarán las variables investigadas**. La ***unidad de muestreo*** *o de análisis* es **el elemento utilizado para seleccionar la muestra**, es decir, **cada uno de los elementos que integran la muestra**. Por lo general, la **unidad de observación** (poblacional) y **la unidad de análisis** (muestral) son **la misma**, pero hay casos en que **no** : si se desea investigar **el maltrato familiar de los menores** , y **no** hay modo de seleccionar directamente las **unidades de observación** (los **menores maltratados**) , se seleccionan **las unidades de análisis** (los **hogares o casas donde habitan los menores maltratados**) para poder llegar a ellos.

2. Muestreo aleatorio

Muestreo aleatorio simple



Cada uno de los elementos o unidades de la población tiene aquí **la misma probabilidad conocida de ser seleccionado**, y esto se logra mediante la **selección al azar** de dichos elementos.

Se confecciona primero **un listado numerando correlativamente *todas* las unidades de la población** (denominada "**marco muestral**"), para lo cual es necesario, previamente, haber **definido correctamente la población** (es decir, haberla **delimitado** de un modo estricto y concreto).

Muestreo aleatorio simple

Se determina luego, usando las **fórmulas** usuales , el **tamaño** de la muestra, lo que implica también definir el **margen de error** o **desviación estándar**, y el **nivel de confianza** deseados para generalizar los resultados de la investigación de la muestra a toda la población.

Posteriormente, se procede a **seleccionar el número calculado de unidades** mediante cualquier **procedimiento aleatorio**, por ejemplo, por sorteo o rifa, o Tabla de Números aleatorios.

Muestreo aleatorio simple

Ventajas:

- la simpleza de su procedimiento y el bajo costo,
- no es necesario dividir la población en subgrupos ni tomar ningún otro paso adicional antes de seleccionar miembros de la población al azar.

Desventajas:

- No puede usarse cuando la población es demasiada grande, o potencialmente infinita
- no es posible confeccionar el listado numerado de todas las unidades (es decir, el *marco o estructura muestral)*.

Además, dependiendo del tamaño de la población, puede tornarse un método muy lento .

3. Muestreo sistemático

Muestreo sistemático

El muestreo ***sistemático*** es también un método **probabilístico aleatorio**,
que exige la confección del **marco muestral**
(es decir, el **listado** numerado de elementos) .
Presenta la ventaja , respecto del muestreo aleatorio simple,
de que **evita el riesgo de que la muestra pierda representatividad**
o quede **sesgada** por la existencia de algún tipo de **regularidad**
o periodicidad en la población que el investigador **no conoce.**

Muestreo sistemático

Ventajas: puede eliminar la selección agrupada y es fácil de ejecutar.

Contras: necesitamos predeterminar el tamaño estimado de la población. No funciona bien si la población tiene un tipo de patrón estandarizado.

Caso de uso: se usa cuando los datos relevantes no exhiben patrones.

Muestreo sistemático

Este tipo de muestreo consiste en obtener el *número de selección sistemático* (***K***) dividiendo el tamaño de la población (***N***) por el tamaño calculado de la muestra (***n***)

Así : $$K = \frac{N}{n}$$

determina el **número de selección periódica** que se utiliza sistemáticamente para escoger cada unidad de observación, hasta completar la muestra.

Luego se decide por sorteo por cuál número natural (menor o igual que ***K***) se **inicia** la selección , y así hasta completar la muestra. Supongamos que la población (**N**) es de 100 unidades y la muestra **(n)** de 25 unidades. Por ende, ***K*** = 100/25 = 4.

Colocamos cuatro papelitos con los números 1, 2, 3, 4 en un recipiente, y elegimos uno al azar (por caso, el número **3**), el **primer elemento** o unidad muestral será el que esté numerado en **tercer lugar** (es decir, **el elementos número 3 del listado**) ; el siguiente, será el elemento cuyo número resulta de sumar el número del elemento anterior (**3**) con el número de selección sistemática (**4**) , es decir, **7**. El siguiente, será **7 + 4 = 11** , y así sucesivamente hasta **completar la muestra**, es decir, hasta elegir **25** unidades.

Con este procedimiento, el ***último*** **elemento** elegido será un número ***menor al tamaño de la población.*** En nuestro ejemplo, la sucesión sería, para A = 3 ,

K = 4, N = 100 y n = 25, la siguiente :

3, 7, 11, 15, 19, 23, 27, 31, 35, 39, 43, 47, 51, 55, 59, 63, 67, 71, 75, 79, 83, 87, 91, 95, **99**. Este último número, 99, es **menor que 100**, el tamaño de la población (**N**).

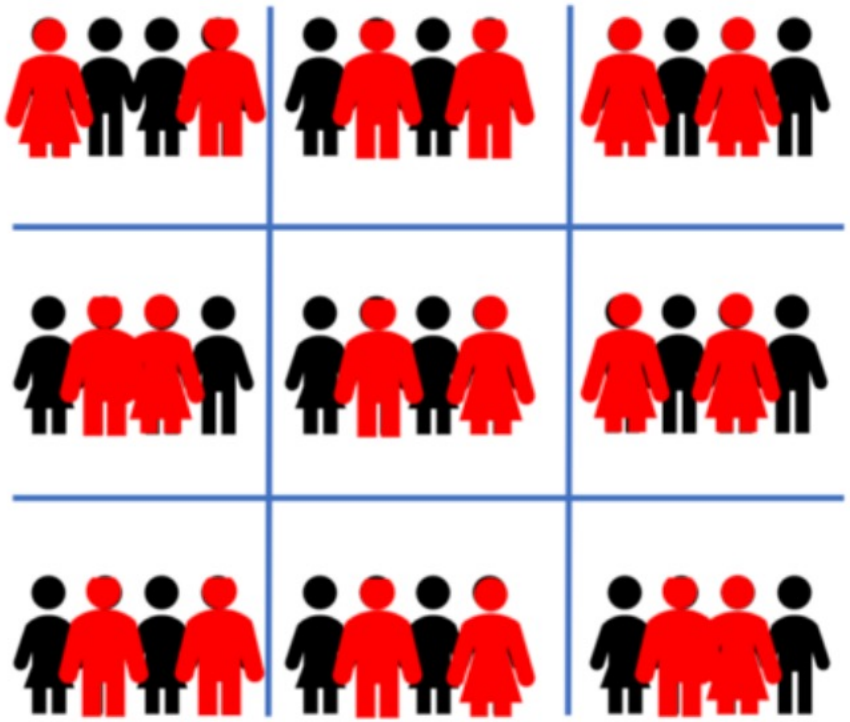
En general, si llamamos ***A*** al primer número elegido (es decir, 3) , y ***K*** al número de selección sistemática (es decir, 4) , el segundo número será ***A*** + ***K*** (es decir, 7) el tercero ***A*** + ***2 K*** (es decir, 11), el cuarto ***A*** + ***3 K*** (es decir, 15) y así sucesiva -mente hasta el último número elegido :

$$A + ((n - 1) K)$$

es decir: 3 + ((25 – 1) x 4) = **99**.

4. Muestreo estratificado

Muestreo Estratificado

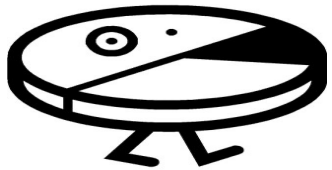


Se emplea cuando se conoce la ***distribución*** y la ***variabilidad*** de la variable en toda la población.

Esto significa que el investigador sospecha, o sabe, que la **variable principal** que está analizando ***se comporta de manera diferente*** (es decir, **varía**) en cada ***estrato o subgrupo*** de la población. Y también, que el investigador conoce ***cómo se distribuye porcentualmente la variable***.

Este procedimiento se aplica cuando la población ***no es homogenea***, pues aunque cada *estrato* o subgrupo sea *internamente homogéneo* , **los estratos son heterogéneos (diferentes) entre sí**. Cada estrato se configura o conforma en base a alguna ***variable*** que el investigador **supone *relacionada* con el *fenómeno investigado***, pues sospecha que este fenómeno **se comportará de modo diferente en cada estrato** así constituido.

Por ejemplo, un investigador está analizando la relación entre las variables ***hábitos de estudio*** y ***nivel de aprendizaje logrado*****.** Si el investigador sospecha que la variable ***hábitos de estudio*** se comporta de manera ***diferente*** respecto de la variable ***nivel socioeconómico***, porque supone, acertadamente o no, que los alumnos de **bajo** nivel socioeconómico tienen ***menos*** hábitos de estudio que los alumnos de elevado nivel socioeconómico , podría formar tres estratos o subgrupos en la población total : alumnos de **bajo** nivel socioeconómico, alumnos de nivel **medio**, y alumnos de **alto** nivel.



- Una vez dividida la población en estratos o subgrupos, ***tomaría una muestra de cada estrato*** (denominada ***submuestra***) para estudiar la **relación** entre las variables ***hábitos de estudio*** y ***nivel de aprendizaje logrado*** en **cada estrato**, para finalmente establecer una ***comparación*** entre los **distintos subgrupos** y arribar a una conclusión global.
- Nótese que **la estratificación es artificial**, en el sentido de que tal agrupación no está dada de un modo natural en la realidad, sino que el agrupamiento en estos estratos es una ***decisión teórica*** del investigador (podría haber hecho **otra** estratificación)

- Cuando se habla de "**variabilidad de la variable principal**" se refiere al comportamiento **diferencial** de la variable **por estrato.**

- Nótese que si el universo o población fuese **homogéneo** (es decir, que la variable ***hábitos de estudio*** se comporta ***de la misma manera*** *en* ***toda la población***) la estratificación **carece de sentido**. Además, no es conveniente estratificar la población en base a **demasiadas variables**, pues se generarían **muchos estratos** y esto complicaría innecesariamente el análisis ("**cruces de variables**").

Una vez que el investigador dividió la población en estratos , debe **obtener una muestra de cada estrato** utilizando cualquiera de los procedimientos aleatorios indicados (**azar simple** o **sistemático**) . Es decir, que una vez divido el universo en subgrupos, cada estrato se considera como si fuera una población particular en sí misma, de tamaño más reducida, de la cual se extrae una muestra.

La ***distribución*** de la variable se refiere al hecho de que el investigador **conoce cómo se distribuye la variable** en estudio (por ejemplo, el hábito de fumar más de 20 cigarrillos diarios) **en la población total** (por ejemplo, porque sabe, mediante encuestas, que en una región determinada **el 70 % de los pobladores fuma esa cantidad** de cigarrillos y el 30 % no) .

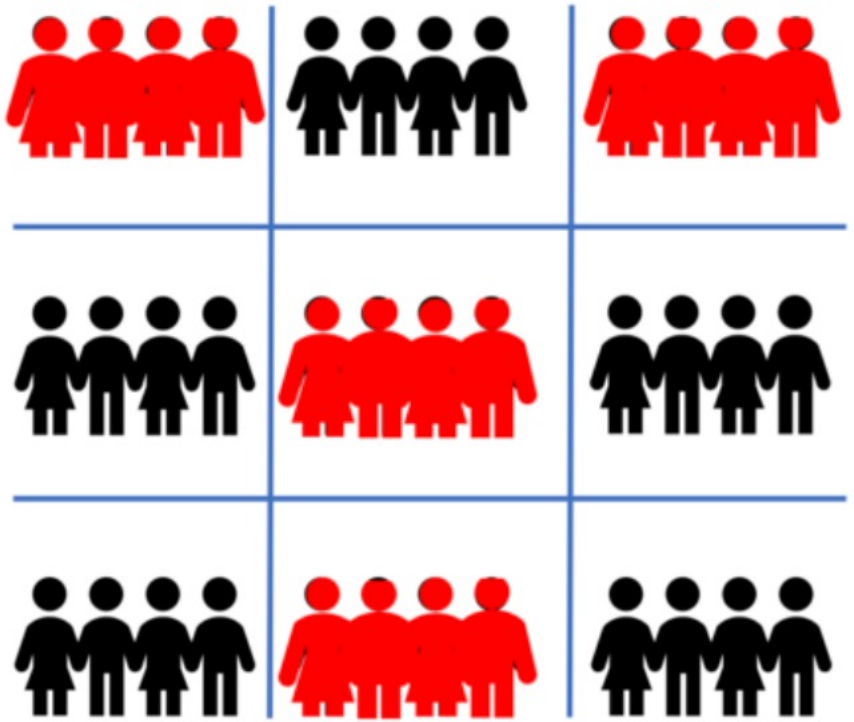
Una vez definida la población, y determinados **los estratos** según la **distribución** y la **variabilidad** de la variable principal investigada, el investigador debe obtener **una lista numerada** de las unidades de análisis o elementos que componen cada estrato. A continuación, debe decidir si desea aplicar un muestreo ***proporcional*** o ***no proporcional***

Ventajas: captura las características clave de la población, por lo que la muestra es más representativa de la población.

Contras: es ineficaz si no se pueden formar subgrupos.

5. Muestreo por conglomerados

Muestreo por conglomerado (clustering)



Se utiliza para poblaciones grandes y dispersas, en lugar de individuos se seleccionan conglomerados que están agrupados de forma natural por ejemplo; casas, cuadras, manzanas, colonias, etc. Se selecciona en primer lugar el conglomerado mas alto y a partir de este se selecciona un subgrupo. Y así sucesivamente hasta legar a las unidades de análisis. También se denomina muestreo por etapas múltiples

Muestreo por conglomerado (clustering)

Ventajas: reduce la variabilidad y es fácil de realizar.

Contras: es posible introducir sesgos durante el muestreo.

Caso de uso: se utiliza cuando todos los individuos de cada grupo pueden ser representativos de las poblaciones.

6. Teorema del Límite Central

Teorema del Límite Central

Teorema (del límite central): Sea $X_1, X_2, ..., X_n$ un conjunto de variables aleatorias, independientes e idénticamente distribuidas de una distribución con media $\boldsymbol{\mu}$ y varianza $\boldsymbol{\sigma^2 \neq 0}$. Entonces, si **n** es suficientemente grande, la variable aleatoria

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

tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$.

Independiente de la distribución original, la distribución de los promedios de todas las posibles muestras de tamaño n será siempre de tipo normal

Tomando todas las posibles muestras de una población, la distribución de frecuencia de sus promedios será una distribución Normal

Teorema del Límite Central

Ejemplo

- Suponga una población compuesta de solamente de cinco unidades
- A,B,C, D y E
- Si tomo muestras de tamaño tres tengo las siguientes posibilidades
- A,B,C A,B,D A,B,E B,C,D, B,C,E B,D,E C,D,E

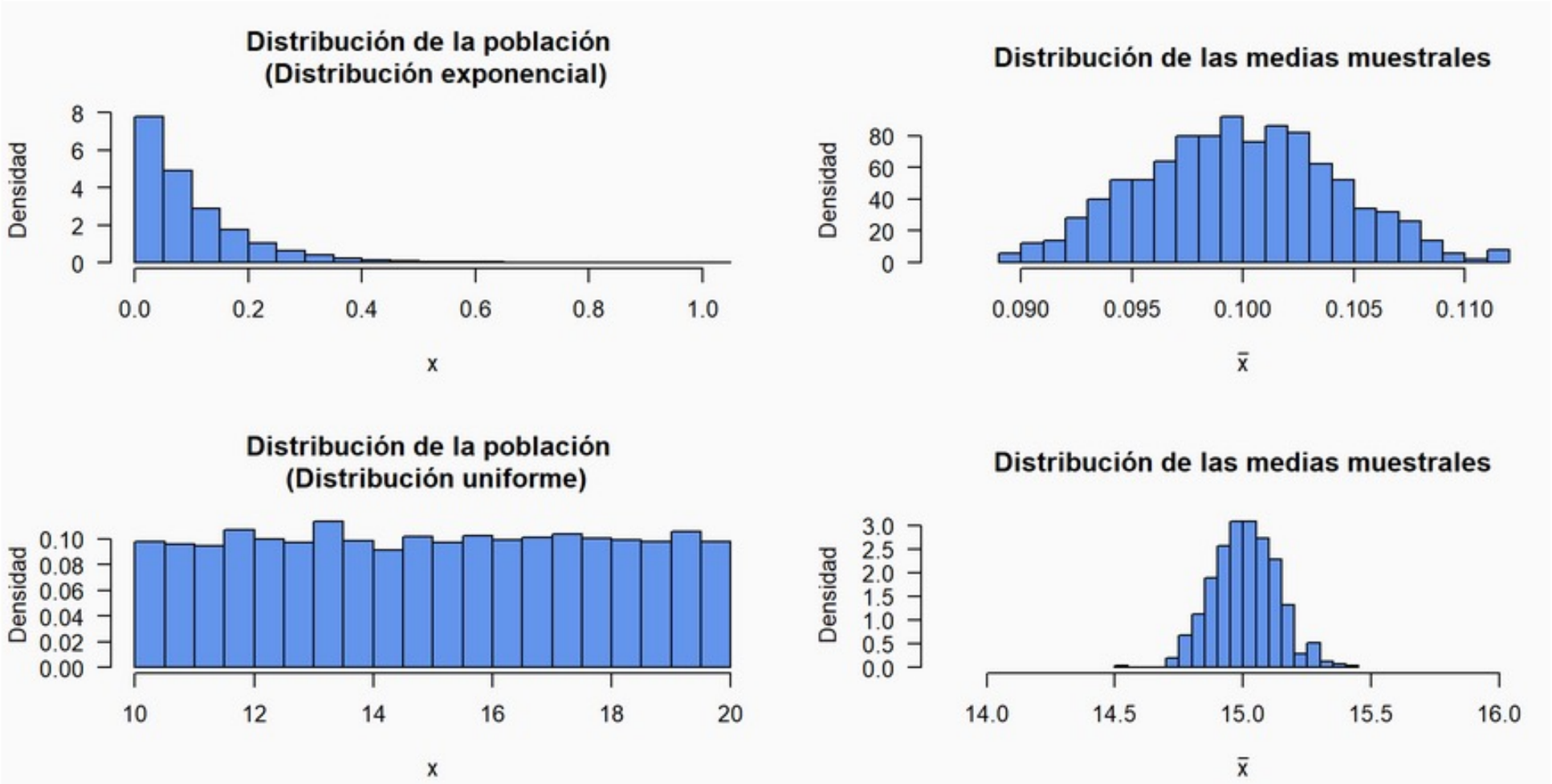
Teorema del Límite Central

- Imagine que A=1, B=2, C=3, D= 4, E= 5

Ahora con Números

			Total	Promedio
1	2	3	6	2
1	2	4	7	2.33
1	2	5	8	2.66
1	3	4	8	2.66
1	3	5	9	3
2	3	4	9	3
2	3	5	10	3.33
3	4	5	12	4

Teorema del Límite Central



7. Distribución muestral

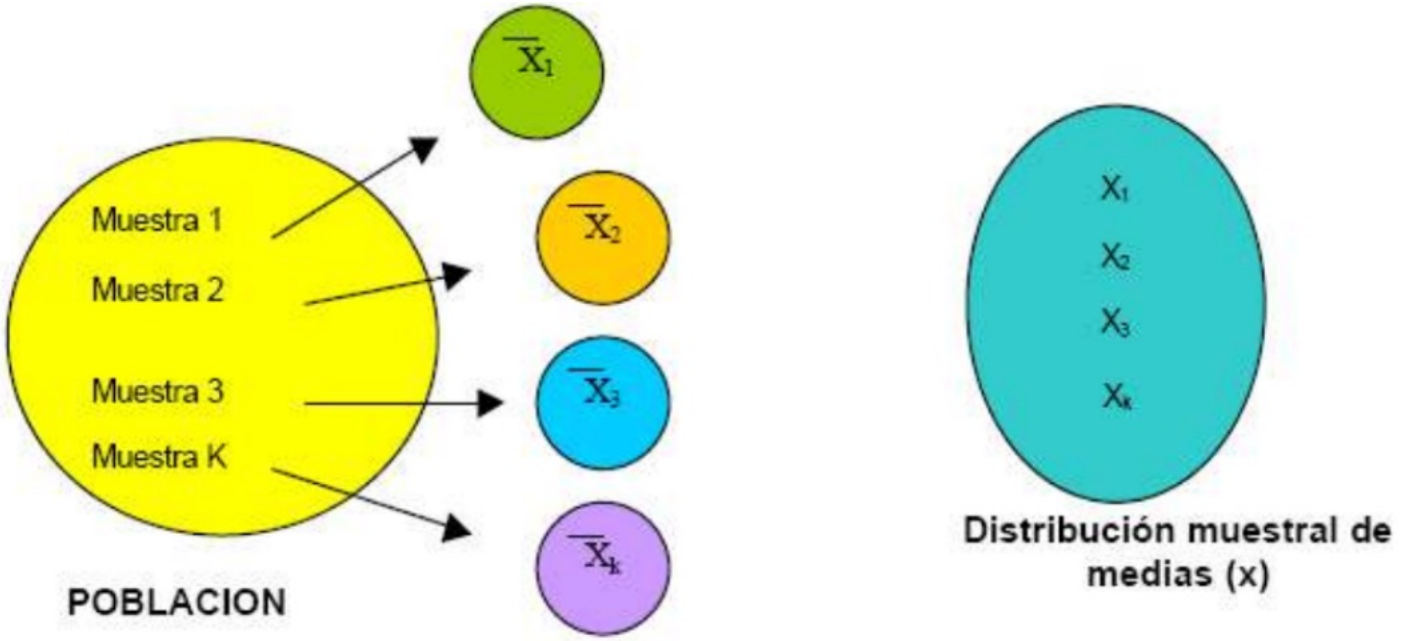
Distribución muestral

Las muestras aleatorias obtenidas de una población son, por naturaleza propia, impredecibles. No se esperaría que dos muestras aleatorias del mismo tamaño y tomadas de la misma población tenga la misma media muestral o que sean completamente parecidas; puede esperarse que cualquier estadístico, como la media muestral, calculado a partir de las medias en una muestra aleatoria, cambie su valor de una muestra a otra, por ello, se quiere estudiar la distribución de todos los valores posibles de un estadístico

Como los valores de un estadístico, tal como la media, varían de una muestra aleatoria a otra, se le puede considerar como una variable aleatoria con su correspondiente distribución de frecuencias.
La distribución de frecuencia de un estadístico muestral se denomina distribución muestral.

Distribución muestral de medias

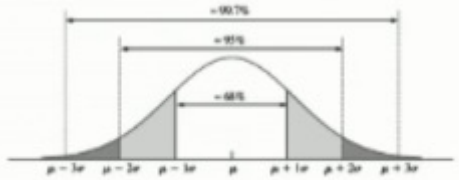
Suponga que se han seleccionado muestras aleatorias de tamaño 20 en una población grande. Se calcula la media muestral x para cada muestra; la colección de todas estas medias muestrales recibe el nombre de distribución muestral de medias, lo que se puede ilustrar en la siguiente figura:

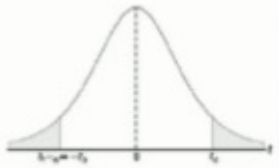


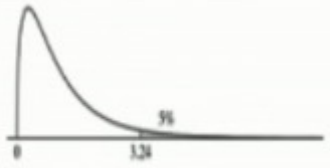
Distribuciones muestrales

Parametro poblacional a estimar

	Media, Proporción o Lambda		Varianza	
Mejor Estimador	Media Muestral		Varianza Muestral	
Caso	Distribución con Varianza conocida	Distribución con Varianza Desconocida	Distribución de una sola Varianza	Distribución del Cociente de Varianza X sobre Varianza Y
Estadístico	$\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}$	$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$\frac{(n-1)S^2}{\sigma^2{}_0}$	$\frac{S_X^2}{S_Y^2}$
Distribución del Estadístico (Parámetros)	Normal Estandar (**0, 1**)	t-Student (**N-1**)	Chi-cuadrada (**N-1**)	F Fisher-Snedecor (**Nx-1, Ny-1**)







A Practicar!!!!

KIBERNUM

Creando Juntos Excelencia

Software Engineering | IT Staffing | IT Academy | IT Consulting

Proyecto apoyado por CORFO