ALGUNOS ELEMENTOS DE MUESTREO
(Guía de clase)

FICHA Nº 23

A. La **selección de casos** constituye un problema metodológico que depende tanto del problema de investigación, como de los recursos disponibles. En términos generales, se abren dos alternativas:i) los métodos aleatorios y ii) los métodos no-aleatorios de selección.

✎ El muestreo aleatorio es aquel en el que todas las unidades de la población tienen una probabilidad conocida de ser seleccionados; probabilidad que es establecida a priori de la misma selección.

✎ El muestreo no aleatorio se caracteriza por la selección intencional de los elementos. Existen varias formas de hacer esta selección:

- "bola de nieve": primero se selecciona un elemento por ejemplo un informante calificado, y a partir de el se confecciona una lista de otros individuos que serían relevantes de ser entrevistados
- "por valores de la variable de interés". Por ejemplo, en un estudio cuyo objeto es el proceso político que lleva a la adopción de políticas sociales, se seleccionan distintas políticas sectoriales focalizadas.
- por valores de una o varias "variables hipotéticamente explicativas". En el ejemplo anterior, se supone que los gobiernos encabezados por partidos de derecha instrumentarán políticas focalizadas; por lo tanto se seleccionan distintos tipos de gobiernos atendiendo a una definición ideológica (derecha, centro, izquierda) y se analiza qué tipo de política social fue adoptado.
- "la selección de sujetos en el marco de diseños experimentales"

ALGUNOS ELEMENTOS DE MUESTREO
(Guía de clase)

✎ La diferencia más importante entre los distintos métodos de selección tienen que ver con las implicancias que acarrean al momento de la inferencia:

- la posibilidad de realizar afirmaciones sobre los parámetros poblaciones conociendo cuantitativamente el error que puedan tener esas afirmaciones. **Sólo el muestreo aleatorio permite conocer el error que se comete al estimar los parámetros de población.**

- La posibilidad o imposibilidad de realizar afirmaciones causales, es decir de generalizar a la población las características de las distribuciones conjuntas halladas en una muestra y a las cuales se le asocia una teoría causal. **Las muestras aleatorias en sí mismas no resuelven este problema.**

✎ El **caso más simple en el muestreo aleatorio**, del cual se derivarán luego los distintos tipos de muestras, tiene dos propiedades:

- Un elemento *i* de la población tiene la misma probabilidad de ser seleccionado que el resto de los elementos. Formalmente:

$$P_i = \frac{1}{N}$$

- La población tiene un tamaño tal que puede ser considerada infinita. Poblaciones mayores a 100 mil casos son tratadas como infinitas por lo general.

- En términos aplicados, para la selección de una muestra se requiere contar y trabajar sobre un **marco muestral** o listado identificatorio actualizado y completo de todos los elementos que componen la población. El siguiente cuadro 1 proviene de la Encuesta Nacional de Ingresos y Gastos de los Hogares levantada en Honduras recientemente. Se muestra el total de viviendas que según proyecciones con base censal existían al momento de la Encuesta. Para realizar la selección se supone que se disponía de un listado de todas las viviendas de Honduras con su respectiva ubicación geográfica (departamento, localidad) del cual extraer los elementos de la muestra y terminar visitando los hogares. En este contexto, por lo general se utiliza la base de datos relevada por el último censo disponible.

Cuadro 1
Marco muestral para la Encuesta de Ingresos y Gastos de Honduras

Cantidad de viviendas en el marco muestral, cantidad de Unidades Primarias de Muestreo elegidas y hogares efectivamente entrevistados, por estrato

Estrato	Cantidad de Viviendas en el país	Hogares Efectivamente entrevistados
Distrito Central	136.195	983
San Pedro Sula	67.374	880
Otras Ciudades	187.622	1219
Zona Rural	510.861	664
Total	**902.052**	**3746**

Fuente: www.ine-hn.org/enigh/tipodemuestreo.html

B. Fijemos estas ideas generales con un ejemplo que se irá desarrollando conforme se exponen los temas siguientes.

- Supóngase que se desea realizar una caracterización de la estructura de la organización y de la tecnología educativa disponible en las escuelas que imparten educación primaria en México.
- Según las últimas cifras disponibles para el ciclo escolar 2003-2004, el tamaño de la población estaba integrada por 101.631 escuelas, por lo que la población se puede tomar como infinita. La probabilidad de que una escuela sea seleccionada es:

$$P_i = \frac{1}{101631}$$
$$P_i = 0.00000983$$

- En el universo se registran las siguientes proporciones (h_i) para cada tipo de escuela. Utilizando la definición de que las probabilidades son las **proporciones en el límite** (e.g. en una población inifinita), podríamos establecer que la proporción es la probabilidad asociada con cada tipo de escuela de ser incluida en la muestra.

Cuadro 2

Universo de las escuelas primarias de México

Sector	Número	h_i
Escuela de CONAFE	15189	14,95%
Escuela del sector Indígena	9041	8,90%
Escuela Pública (rural y urbana común)	71141	70,00%
Escuela Privada	6260	6,16%
Total	101631	100,00%

- Tales probabilidades (h_i) definidas **en el límite** como probabilidades permitirían establecer el "desideratum" de que la proporción de escuelas en la muestra de cada uno de los sectores institucionales sea bastante similar al observado en la población. La promesa del muestreo aleatorio consistirá en que tales proporciones se mantendrán bastante razonablemente próximas (entre algunos muestristas esto se denomina "muestra auto-ponderada"). Sin embargo, si esto no llegara a ocurrir, se diría que la muestra tiene algunos sesgos y según la gravedad del problema detectado, debería instrumentarse alguna corrección mediante **ponderadores**. Estos se definen como una razón entre las proporciones que teóricamente debería pesar la *j-ésima* categoría y lo empíricamente observado. Otra forma de definirlo es diciendo que se trata de una **razón** entre la **fracción de muestreo** teórica y la fracción de muestro observada.

$$W_J = \frac{p_j}{h_j} = \frac{\frac{n_T}{N}}{\frac{n_o}{N}}$$

- En la anterior fórmula:

  p_J es la probabilidad que tiene la categoría *j* de ser incluida en la población

  h_i es la proporción empíricamente observada de la categoría *j* en la muestra obtenida

- Si por ejemplo, sucediera que tuvimos una muestra donde las proporciones observadas indican una sobrerepresentación del sector indígena y del CONAFE deberíamos ajustar los ponderadores de la muestra. Repliquemos el cuadro 1, incluyendo la descripción de la muestra.

Cuadro 3

Universo y muestra para las Escuelas Primarias de México

	UNIVERSO		**MUESTRA**		**W_J**
Sector	Número	p_j	Número	h_J	
Escuela de CONAFE	15189	14.95%	340	10.56%	1.41584
Escuela del sector Indígena	9041	8.90%	682	21.17%	0.42014
Escuela Pública (rural y urbana común)	71141	70.00%	2052	63.71%	1.09877
Escuela Privada	6260	6.16%	147	4.56%	1.34965
Total	101631	100.00%	3221	100.00%	1.00000

Fuente: elaborado con base en el universo de escuelas primarias de México difundido por la Secretaría de Educación Pública de México (junio 2003) y la base de datos del 4º levantamiento de Estándares Nacionales de Matemática y Español (2001).

- De la anterior tabla se deprenden los ponderadores que deberán ser activados en el paquete estadístico para todos los análisis (última columna del cuadro 3, titulada W_J.

- **Sin embargo,** en el tratamiento siguiente supondremos que la muestra aún no fue tomada, que no existe, y que por lo tanto no sabemos nada de ella. Se verá que las cuestiones teóricas postuladas detrás de la selección aleatoria de una muestra tienen que ver con un **juego de probabilidades**, donde se hacen supuestos y se utiliza la imaginación.

C. La decisión de seleccionar los casos a estudiar mediante una muestra aleatoria, permite que la estadística ponga a disposición del investigación un conjunto de conocimientos teóricos entre los que ocupa un lugar crucial el **teorema del límite central** fundado en la noción de **distribución muestral** de estadísticos**.** Todos estos conocimientos son provistos por la teoría de las probabilidades y por lo tanto, son parte de un **juego de azar** donde se trata de saber que tan bien podría llegar a resultar una muestra.

✎ Los **elementos de una población** se pueden seleccionar ya sea *reponiendo* los elementos una vez seleccionados, o *sin reponerlos.*

- En el primer caso, se habla de muestreo con reemplazamiento y en el segundo, sin reemplazamiento.

- La diferencia entre uno y otro es que al extraer el segundo elemento para la muestra, su probabilidad ya no será la misma que para el primero, sino que vendrá dado por:

$$P_i = \frac{1}{N-1}$$

- Si se combinan las dos distinciones anteriores hechas, da lugar al cuadro 3. En el contexto de las investigaciones empíricas, las muestras será con re-emplazamiento y en varias ocasiones las poblaciones podrán ser tratadas como infinitas porque son mayores a 100 mil elementos.

Cuadro 4

	Población infinita	Población finita
Sin re-emplazamiento		
Con re-emplazamiento		

ALGUNOS ELEMENTOS DE MUESTREO
(Guía de clase)

✎ Utilizando el análisis combinatorio, se puede calcular el total de todas las distintas *"n"* **muestras sin reemplazamiento** cada una con un número exacto "*r*" de elementos que se pueden tomar de una población ("*N*") viene definido por:

$$_{N}C_{r} = \frac{N!}{r!(N-r)!}$$

- Si esto se aplicara al ejemplo de las escuelas primarias, y si supusiéramos que nuestra muestra será de 3200 escuelas, el número total de muestras que se podrían obtener de un universo de 101631 es:

$$_{101631}C_{3200} = \frac{101631!}{3200! * 98431!}$$

✎ En la población N existen algunas propiedades (X_1, X_2, X_3, X_4, ...X_k) que son de interés para la investigación y que han sido definidos teóricamente.

- La muestra se extrae porque los **parámetros (Θ)** que caracterizan la **distribución poblacional** de cada una esas propiedades son desconocidos.

- Para cada una de estas propiedades se dispondrá de al menos una definición operativa para su medición (es decir, una variable). Cada una de esas variables que serán registradas en las unidades de la muestra, se podrán aplicar los **estadísticos** uni, bi y multivariados que se han definido con anterioridad.

- Para cada variable, se podrá obtener una **distribución observada** *en el caso más simple* del estadístico de tendencia central "media aritmética" (si es una variable por lo menos interval) o una distribución observada de proporciones en el caso de variables nominales u ordinales. A los efectos del siguiente desarrollo, supondremos que hay un único parámetro de interés para una única variable y que se trata de la media aritmética poblacional.

$$\Theta = \mu$$

- El objetivo de la investigación será (al menos) caracterizar la **μ** de la distribución poblacional de cada uno de los indicadores utilizando el **estadístico $\bar{x}$ como estimador**. Qué tan buena sea esta operación es un **problema de estimación** que será revisado en la siguiente ficha de actividad.

- En el ejemplo desarrollado las escuelas interesan ser descritas en dos dimensiones principales: en su estructura y en su tecnología. La estructura de la organización tiene distintas dimensiones, pero una de las más cruciales es si la escuela tiene un maestro por grado enseñado o por el contrario, un mismo maestro enseña a alumnos de distintos grados en forma paralela. Al primer tipo de estructura se le llama comúnmente “escuela completa” en tanto que la segunda es una “escuela incompleta”. La “tecnología” de la enseñanza que más interesa es el número de computadores que tienen las escuelas. Sean por lo tanto:

$P_{(E=1)}$ la proporción poblacional de escuelas que tienen una estructura incompleta (E=1). Este será el primer parámetro de interés.

μ el promedio poblacional del número de computadores (C) que tienen las escuelas. Este será el segundo parámetro de interés.

✎ **Ahora viene un importante ejercicio de IMAGINACIÓN. Supóngase** que se está en condiciones de extraer sin re-emplazamiento la **totalidad** de las muestras de tamaño “**n=r**” de una población infinita de elementos:

- Para cada una de estas muestras se calcula la media aritmética ($\bar{x}$**)** de la variable de interés X, de tal forma que se cuenta con una nueva matriz de datos en la cual los renglones constituyen “resúmenes” hechos para cada una de todas las m-ésimas muestras posibles y cuya única variable es el valor de la media aritmética en cada muestra.

Cuadro 5

Número de muestra	Media aritmética de X en la muestra
m = 1	$\bar{x}_{\mathbf{1}}$
2	$\bar{x}_{\mathbf{2}}$
3	$\bar{x}_{\mathbf{3}}$
..	**..**
M - ésima muestra (de un total de $_N C_r$)	$\bar{x}_{\mathbf{m}}$

- A esta nueva matriz teórica o imaginaria de todas las posibles muestras, se pueden aplicar nuevamente estadísticos para caracterizar la **distribución muestral** de la variable (que ahora repetimos, es la media aritmética de cada muestra).

- Dado que ahora se trata de caracterizar una distribución **teórica** se utilizará el operador de valor esperado para definir la medida de tendencia central “media” para las medias aritméticas; el cual viene dado por:

$$E(\bar{x}) = \sum_{m=1}^{m} \overline{x_m}$$

$$Var(\bar{x}) = \left\{ \sum_{m=1}^{M} [\bar{x}_m - E(\bar{x})]^2 \right\} / m$$

- También se puede estimar una medida de dispersión, por ejemplo, la varianza:

✎ El **teorema del límite central** señala que la distribución muestral de las medias calculadas cada una de todas las posibles muestras aleatorias

$$[1] E(\bar{x}) = \mu$$

$$[2] Var(\bar{x}) = \frac{\sigma^2}{n}$$

$$[3] \overline{x}_{n \to \infty} \approx N(\mu; \sigma^2 / n)$$

de tamaño **n** una población, se caracterizará por:

- Como se puede apreciar en las expresiones [1] y [2], el teorema está ligando los estadísticos que caracterizan una distribución muestral con los estadísticos de la distribución poblacional.

  - Conceptualmente se dice que la media aritmética de todas las medias aritméticas posibles de calcular en muestras aleatorias es igual a la media poblacional. Con lo cual, existe una relación entre la media realmente observada y la media ignorada.

- Se está diciendo además, que la dispersión de esta distribución de medias muestrales es igual a la varianza de la población pero dividida entre el tamaño **n** que tiene cada muestra (recuérdese que todas las muestras son del mismo tamaño **n=r**).

➭ En la expresión [3] se está afirmando además que la distribución muestral de medias se aproxima a una **distribución normal** conforme el tamaño **n** de cada muestra tiende a infinito (es decir, es lo suficientemente grande).

- Se afirma que la distribución **teórica** lograda por todas las medias aritméticas calculadas en cada muestra será una **DISTRIBUCIÓN NORMAL** por lo tanto, simétrica y perfectamente caracterizable por la expresión:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2}\frac{\left(x-\bar{x}\right)^2}{\sigma^2}\right]}$$

- Esta distribución será normal sin importar cuál sea la distribución que X tenga en la población o que X tenga en cada una de las muestras.

➭ Si suponemos que las medias aritméticas han sido estandarizadas, es decir transformadas desviaciones de la media o puntajes típicos, entonces podemos escribir el teorema de la siguiente forma:

$$[1] E(\bar{x}) = 0$$

$$[2] Var(\bar{x}) = 1$$

$$[3] \overline{\underset{n \to \infty}{x}} \approx N(0;1)$$

- Por lo tanto estamos en condiciones de utilizar la tabla general de Z que permite calcular las probabilidades entre dos puntos (z_1 y z_2) que son según lo venimos suponiendo, valores de medias aritméticas obtenidos en distintas muestras. Es decir, podemos decir con qué probabilidad se podría obtener un determinado rango de valores de medias aritméticas obtenidas con base en

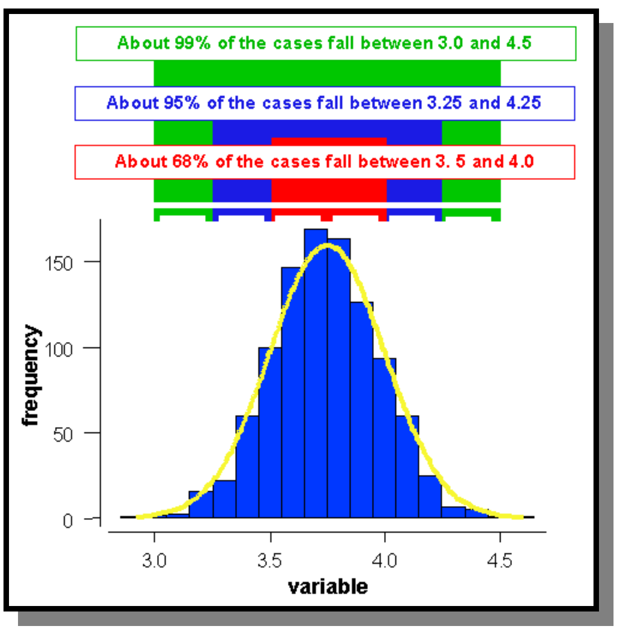

Figura 1
Áreas bajo la curva normal suponiendo una media poblacional de 3.75

el muestreo aleatorio.

✎ La distribución normal es una propiedad de **la distribución teórica** de las medias muestrales que permite que la inferencia estadística quede liberada de cualquier otro supuesto que no sea la existencia de una muestra aleatoria grande tomada de una población infinita. **También se aplica este teorema a la distribución de otros estadísticos, por ejemplo, las** proporciones**.**

D. De este teorema se deriva directamente la determinación del **tamaño de la muestra**.

✎ El **tamaño muestral** en el caso más simple de todos y que se denomina muestreo aleatorio simple (M.A.S.) viene para una variable continua dado por[1] y cuando se trata de una variable discreta por [2]:

$$[1] n = \frac{Z^2 \sigma^2}{d^2}$$

$$[2] n = \frac{Z^2 p * q}{d^2}$$

- En la anterior fórmula se ocupan 4 valores:

n	representa el tamaño de la muestra, una incógnita a despejar
Z^2	representa el nivel de confianza con que se harán las estimaciones a partir de los datos muestrales.

σ^2 o p*q	indica la varianza poblacional de la variable de interés X
d^2	representa la discrepancia máxima que se está dispuesto a aceptar entre el estadístico y el parámetro.

- Debe observarse que en la anterior fórmula no ingresa el tamaño de la población (o N) que va a ser estudiada. Este es el caso más simple de todos y se supone que se trata de una población lo suficientemente grande.

- El tamaño de la muestra se fija en relación a una variable de interés X para la cual se utilizará su varianza poblacional. Si éstas son varias, entonces deberán calcularse varios tamaños de muestra y luego seleccionar entre ellos.

✎ La fórmula del tamaño de la muestra incluye por tanto, un parámetro poblacional (la varinza de X) y otros dos valores que son el resultado de **adoptar decisiones fundamentales** con base en primer lugar a sus objetivos de investigación y en segundo lugar, en relación a los recursos y tiempo que dispondrá para realizar la investigación. Tanto el nivel de confianza como la discrepancia máxima admisible son decisiones, no datos.

- La primera decisión es respecto del **nivel de confianza** ("Z") que se desea que tengan las estimaciones que se realizarán:

    - Este nivel de confianza se establece en términos de porcentajes. Por lo general suele adoptarse un 95% de confianza pero cuando el objetivo de la investigación es altamente exigente, se suele adoptar un 99%.
    - La confianza refiere a probabilidades bajo el área de una curva normal. Un 95% refiere a una probabilidad P= 0.95.
    - El nivel de confianza del 95% es la respuesta que el investigador da a

la siguiente pregunta: *¿En cuántas muestras se podría encontrar la media observada entre los valores estandarizados - 1.96 y + 1.96?.* La respuesta sería en 95 de cada 100 muestras posibles a tomar del total de todas las posibles.

- Es por esta razón que el nivel de confianza se transforma y se denota por la letra Z (de puntajes típicos) dentro de la fórmula. Está expresado en probabilidades bajo el área de una curva normal estandarizada.

➭ La **discrepancia** (d) se define por la diferencia máxima que se está dispuesto a aceptar entre el estadístico a calcular y el parámetro poblacional que se quiere estimar.

- Se define:

$$d^2 = (\overline{x} - \mu)^2$$

- Se establece la medida previamente a la investigación, **sin conocer** cuál será la media aritmética observada (aún no se ha extraído la muestra) ni conocer realmente cuál es la media poblacional.
- La discrepancia es ya una cuestión sustantiva y específica de cada investigación. Por ejemplo, si el objetivo es conocer el resultado final en un ballotage presidencial, la diferencia máxima aceptable para el candidato A entre la proporción de votantes observado y el parámetro no puede ser mayor por ejemplo que 0.05 (es decir 5%)
- En sí misma es interpretable como una **medida de la precisión** que se desea tener al realizar la inferencia.

➭ La **varianza poblacional** σ^2 es el único dato empírico (no decidido) que se requiere para computar la muestra.

- Dado que el objetivo de la investigación es conocer la distribución poblacional de X, es razonable suponer que no se tenga información actual sobre dicha varianza.

- En tales casos, se utiliza como información el valor de la varianza para la variable que haya sido determinado en otros estudios antecedentes, bajo el supuesto de que ésta no ha variado.
- En el caso de que no se dispongan antecedentes sobre la varianza, o que no se puede aceptar el supuesto de que la varianza se ha mantenido constante a través del tiempo, se debe suponer la peor situación. Ésta es la **máxima varianza** teórica para la variable X y viene dada por la varianza de una variable dicotómica con **p=q =0.50.**

E. Regresando al ejemplo iniciado sobre el cálculo del tamaño de una muestra aleatoria simple para las escuelas primarias de México, apliquemos la fórmula con base en las siguientes decisiones:

- Dos son las variables de interés, la estructura y la tecnología. De estudios anteriores se sabe que la proporción de escuelas incompletas en el universo es de 0.35 .Sin embargo, no se tiene información sobre la distribución poblacional del número de computadoras, por lo que será necesario contar con un estimador de su varianza.

- Esto supone que se deberán estimar dos tamaños de muestras: uno para cada variable. Dado que se conoce la proporción de escuelas incompletas en el universo, convendrá comenzar por ahí. Entonces:

$$\sigma^2 = p * q$$
$$= 0.35 * 0.65$$

- La segunda decisión es respecto de cuál es el nivel de confianza que se deseará adoptar. Siguiendo la convención más frecuente, se tomará el 95%

$$P_{(-1.96 \leq Z \leq +1.96)} = 0.955 \Rightarrow z = 2$$

- La tercera decisión adoptada es que la discrepancia máxima aceptable entre la proporción que sea observada y la proporción poblacional para la estructura es de 0.05.

- Aplicando estas decisiones en la fórmula se tiene que el tamaño de la muestra debiera ser de:

$$n = \frac{2^2 0.35 * 0.65}{0.05^2}$$
$$n = 364$$

- La muestra requerida será de 364 escuelas primarias si se desea estimar cuál es la proporción de escuelas de estructura incompleta en la población. Ahora es necesario estimar el tamaño de la muestra para la variable tecnológica. Se recordará que no se tiene información sobre su varianza. En consecuencia se recomienda adoptar la varianza máxima que viene definida por p=q=0.5 y aplicar la fórmula del tamaño muestral con base en proporciones. Si se mantienen las mismas decisiones, el tamaño esta vez será de:

$$n = \frac{2^2 0.5 * 0.5}{0.05^2}$$
$$n = 400$$

- Esta muestra será de 400 casos. Dado que se desean estimar *ambas* variables el tamaño final de la muestra aleatoria simple será de 400 y no de 364.

- Como se observa, la ignorancia respecto a un valor razonable de la varianza ha costado 36 escuelas más en la muestra. Supóngase que encuestar cada escuela cuesta aproximadamente 250 dólares exclusivamente en términos de trabajo de campo: el aumento de la muestra ha significado 9 mil dólares adicionales.

- En las siguientes dos tablas se hace el ejercicio se modificar primero las varianzas y luego la discrepancia máxima simultáneamente con el nivel de confianza para observar cómo se modifica el tamaño de la muestra. Este ejercicio sirve para observar cómo impactan pequeños cambios en la precisión sobre el tamaño muestra.

- Obsérvese que al incrementar las exigencias, una muestra en la peor situación (varianza máxima) pasa de 400 casos a 8100.

Cuadro 6

Ejercicios de cálculo del tamaño de muestra variando decisiones

MODIFICACIÓN DEL TAMAÑO DE LA VARIANZA				
Discrepancia	0.05		Nivel de confianza, Z=	2.00
Ejemplo	P	Q	p*q	n
1	0.10	0.90	0.09	144
2	0.20	0.80	0.16	256
3	0.30	0.70	0.21	336
4	0.40	0.60	0.24	384
5	0.50	0.50	0.25	400
6	0.60	0.40	0.24	384
7	0.70	0.30	0.21	336
8	0.80	0.20	0.16	256
9	0.90	0.10	0.09	144
MODIFICACIÓN DE LA DISCREPANCIA MÁXIMA ACEPTABLE Y DEL NIVEL DE CONFIANZA				
Discrepancia	0.01		Nivel de confianza, Z=	3.00
Ejemplo	P	Q	p*q	n
1	0.10	0.90	0.09	8100
2	0.20	0.80	0.16	14400
3	0.30	0.70	0.21	18900
4	0.40	0.60	0.24	21600
5	0.50	0.50	0.25	22500

F. La teoría elemental del muestreo constituye el núcleo duro del cual se parte para diseñar muestras concretas para objetivos concretos. Éste es el campo del **muestro aplicado** en el cual han desarrollado tres tipos distintos de muestreo:

- El **muestreo aleatorio simple** (o M.A.S) se caracteriza por que todos los elementos tienen la misma probabilidad (equiprobabilidad) de ser seleccionados para integrar una muestra. Es el tipo más simple y sobre el cual se ha desarrollado originalmente la teoría elemental del muestreo.

- En el **muestreo aleatorio estratificado** (M.A.E) la población es segmentada o partida en diferentes grupos definidos teóricamente por el investigador y que se denominan estratos. La hipótesis que se hace es que la distribución de la variable de interés X es más homogénea dentro de cada uno de los estratos.

- El muestreo **aleatorio por conglomerados** (M.A.C) se define cuando los elementos del universo ya están agrupados según cuestiones de proximidad geográfica o física. No se hace ningún supuesto sobre si los conglomerados son o no internamente homogéneos.

- El muestreo estratificado y el de conglomerados son dos variaciones aparentemente similares pero que acarrean consecuencias muy diferentes para las estimaciones como para el trabajo de campo. El siguiente cuadro 5 compara ambos aspectos:

Cuadro 7

Comparación entre el muestreo estratificado y el muestreo por conglomerados

MUESTREO ESTRATIFICADO	MUESTREO POR CONGLOMERADOS
La Unidad Primaria de Muestreo (PSU en inglés) es la misma unidad de análisis definida en el estudio. Las unidades de análisis seleccionadas están definidas por alguno de los tres criterios siguientes: a) cupos iguales, b) afijanción proporcional y c) afijación no-proporcional. Supone una disminución del error muestral debido a que con la estratificación se ha eliminado la principal fuente de varianza hipotetizada.	La unidad primaria de muestreo (PSU) es un cluster o conglomerado pre-existente de unidades de análisis. En la versión *simple,* todos los integrantes del conglomerado aleatoriamente seleccionado son incluidos en la muestra. En la versión compleja, se toman *muestras* en cada conglomerado El error muestral generalmente se incrementa, a no ser que el conglomerado sea altamente heterogeneo en las variables de interés. La cantidad de conglomerados entre los cuales seleccionar la muestra de PSUs dependerá de la heterogeneidad. El tamaño de la muestra de PSUs depende de una función de producción en el que se incluyen la varianza intraconglomerados, los costos del trabajo de campo y el tamaño promedio de los conglomerados.

✎ La elección del tipo de muestreo generalmente obecede a cuestiones de costos del trabajo de campo y del tiempo disponible para realizar el relevamiento.

ALGUNOS ELEMENTOS DE MUESTREO
(Guía de clase)

- Imagínese obtener una muestra de 400 escuelas bajo el tipo MAS para las primarias de México. Este tipo de muestreo supone que cada escuela puede car en cualquier parte del país: desde Tijuana hasta Tapachula. Los costos para cubrir dicha muestra serían gigantescos. Esta es una de las razones que pueden conducir a elegir un muestreo por conglomerados por ejemplo.

✎ **Pero es importante recordar que cualquiera sea la estrategia** particular que será empleada para seleccionar aleatoriamente los casos en una investigación, la determinación del tamaño muestral se realiza partiendo de una misma fórmula, aquella desarrollada para el caso más simple.

Bibliografía de referencia

CEA D'ANCONA

1996 *Metodología cuantitativa*. Editorial Síntesis. Madrid

COCHRAM,

1963 *Sampling Technics* . Ed. Joh Willey. New York. [hay versión en español, editada en México]

KISH, Leslie

1995 *Diseño Estadístico para la Investigación*. Editorial Centro de Investigaciones Sociológicas. Madrid.

SUDAM, Seymour

1976 *Applied Sampling*. John Willey. New York.