

Tamaños de muestra

Johanna Trochez

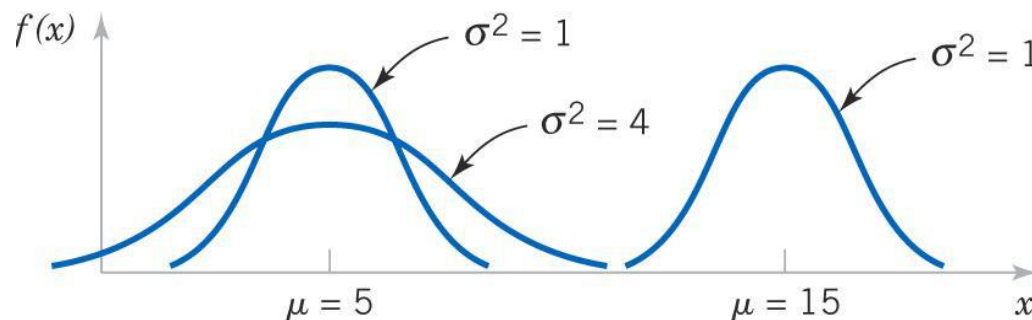
DISTRIBUCIÓN NORMAL

Una distribución ampliamente usada es la distribución normal o gaussiana



Parámetros de la distribución

Esta distribución depende de los parámetros de localización y escala, determinados por la media (μ) y la desviación estándar (σ).



Función de distribución de probabilidad normal

Una variable aleatoria x con pdf

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

es una variable aleatoria con parámetros $\mu \in \mathbb{R}$ y $\sigma > 0$.

La variable aleatoria se denota de la forma

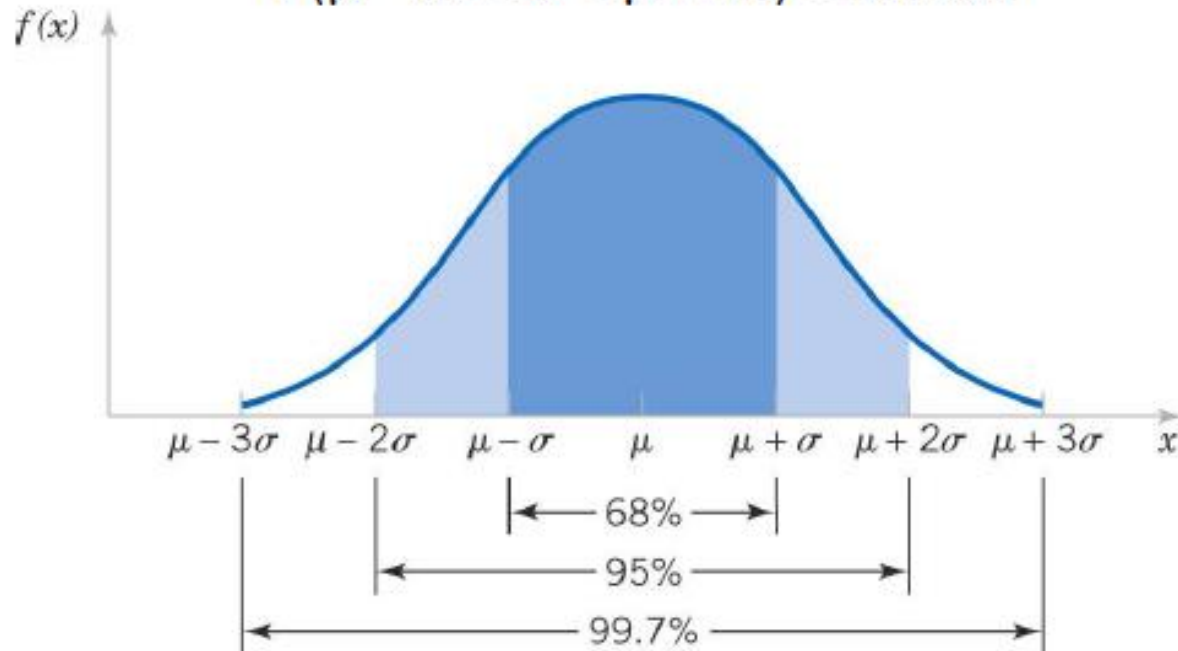
$$x \sim N(\mu, \sigma^2)$$

REGLA EMPIRICA

$$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$$

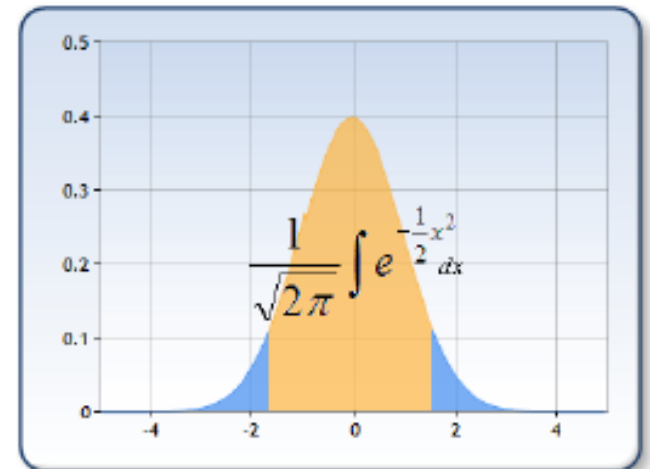
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$$



DISTRIBUCIÓN NORMAL ESTÁNDAR

Una variable aleatoria normal con $\mu = 0$ y $\sigma = 1$, es llamada una variable aleatoria normal estándar y se denota como z

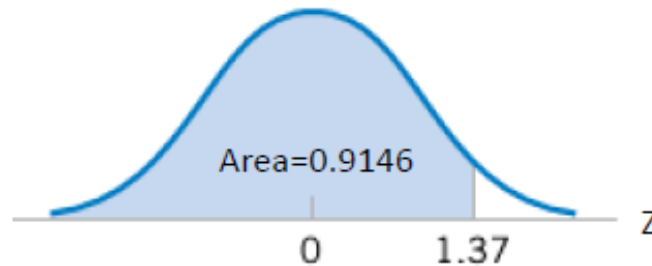
$$z \sim N(0,1)$$




FUNCIÓN DE DISTRIBUCIÓN ACUMULADA

La función de distribución acumulada se denota como

$$\Phi(x) = P(Z < z) = F(z)$$



Estructura de la tabla normal

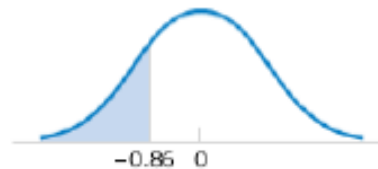


Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.5	0.0044	0.0044	0.0043	0.0043	0.0042	0.0042	0.0041	0.0041	0.0040	0.0040
-1.4	0.0044	0.0044	0.0043	0.0043	0.0042	0.0042	0.0041	0.0041	0.0040	0.0040
-1.3	0.0044	0.0044	0.0043	0.0043	0.0042	0.0042	0.0041	0.0041	0.0040	0.0040
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.01	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.02	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.03	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.04	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.05	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.06	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.07	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.08	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.09	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359

EJEMPLOS

Assume Z is a standard normal random variable.

Find $P(Z \leq -0.86)$.

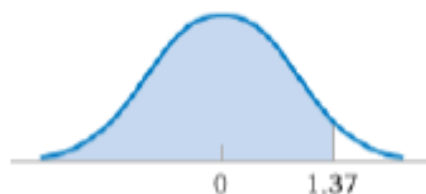


On table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
\vdots										
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148

Assume Z is a standard normal random variable.

Find $P(Z \leq 1.37)$.

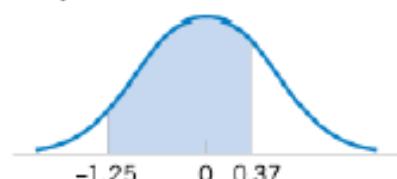


On table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
•										
•										
•										
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

Assume Z is a standard normal random variable.

Find $P(-1.25 \leq Z \leq 0.37)$.

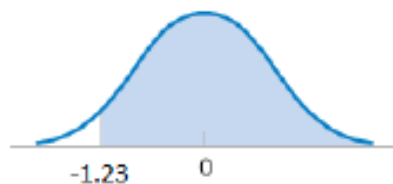


On table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
•										
•										
•										
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
•										
•										
•										
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517

Assume Z is a standard normal random variable.

Find $P(Z > -1.23)$.



On table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
\vdots										
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985

Estadística

Dos conceptos fundamentales: muestra y población

Población objetivo: conjunto de elementos sobre los que queremos hacer afirmaciones

Muestra: subconjunto de la población que se extrae para ser estudiado

Population
(N)



Sample
(n)

¿Porqué una muestra?

Imposibilidad o costo excesivo de realizar un **censo** en que se mide toda la población.

El muestreo se realiza para obtener información acerca de los parámetros desconocidos de la población, por medio de un experimento que permite observar o medir las características de la población, de las cuáles se tiene incertidumbre

Herramientas con dos Objetivos Básicos

Describir la muestra: **Estadística Descriptiva**

Obtener conclusiones de la población a partir de la muestra: **Inferencia Estadística.**

Parámetro: número derivado del estudio de una variable estadística de toda una población.

Estadístico: Medida de resumen numérica que se calcula a partir de la muestra

Definiciones

- **Elemento y Unidad de observación:** Objeto sobre el cual se realiza una medición. En poblaciones humanas las unidades de observación son los humanos.
- **Unidad de muestreo:** Unidad donde realizamos el muestreo
- **Marco de muestreo:** Lista de las unidades de muestreo
- Podríamos estudiar personas pero no tenemos la lista de todos los individuos que pertenecen a la población objetivo.
- Las familias sirven como unidades de muestreo y las unidades de observación son los individuos que viven en una familia

¿Cómo elegir el tamaño de la Muestra (n) ?

- ¿Qué se va a medir?
- ¿Qué se quiere determinar?
- Nivel máximo de error admisible
- Nivel de confianza con qué se quiere obtener la estimación del tamaño muestral
- Variabilidad de las características a medir

Se dice que las variables aleatorias X_1, X_2, \dots, X_n forman una **muestra aleatoria** simple de tamaño n si:

1. Las X_i son variables aleatorias independientes.
2. Cada X_i tiene la misma distribución de probabilidad, en la mayoría de veces se asume normal
3. Cualquier función de las variables aleatorias que forman una muestra se llaman estadístico .

Es decir las X_i son *independientes e idénticamente distribuidas* (iid) $X_i \sim \text{iid } N(0, \sigma^2)$

Teorema del Límite central

Si \bar{X} es la media de una muestra aleatoria de tamaño n , tomada de una población de tamaño N , con media μ y varianza finita σ^2 , entonces la forma límite de la distribución es;

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Conforme $n \rightarrow \infty$, es la distribución de la normal estándar $n(Z;0,1)$

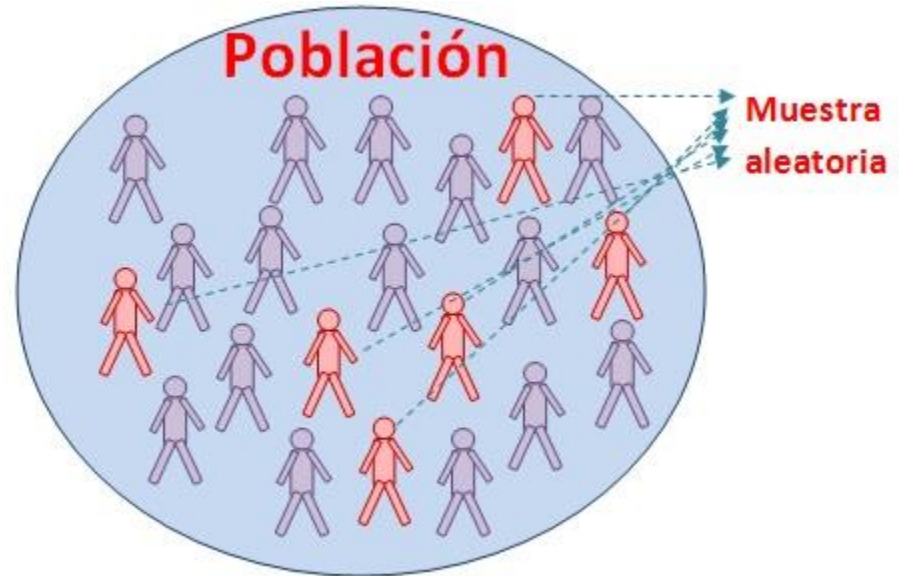
Tamaño de muestra para μ

Población Finita

$$n = \frac{NZ^2\sigma^2}{(N-1)e^2 + Z^2\sigma^2}$$

Población infinita:

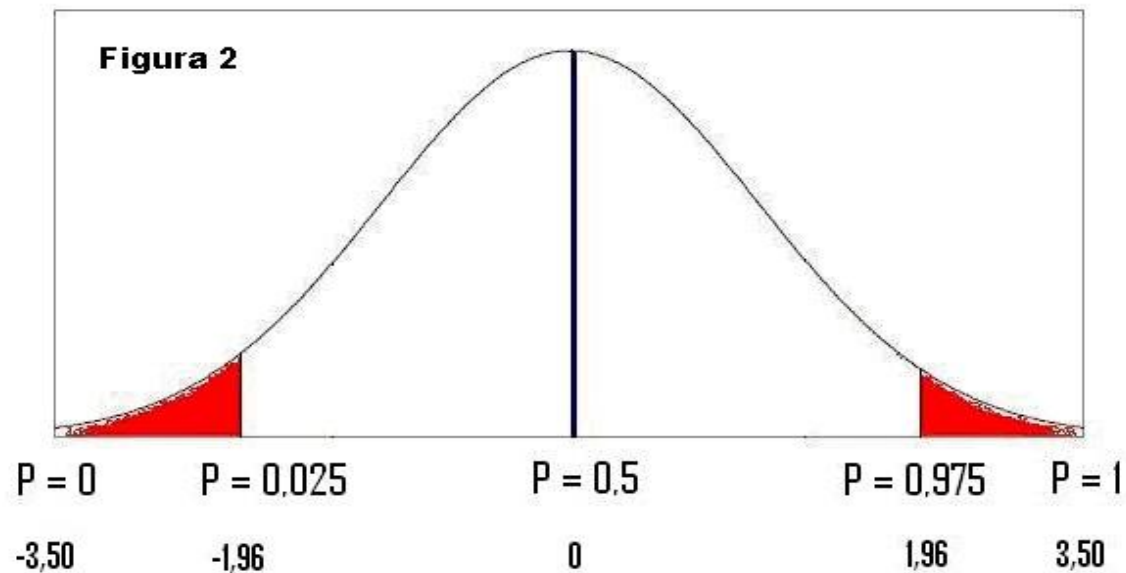
$$n = \frac{Z^2\sigma^2}{e^2}$$



Video: <https://www.youtube.com/watch?v=JX5m7o6rOAAQ>

¿Qué se necesita?

- NC=Nivel de confianza
- σ =Desviación estándar
- e = error máximo admisible



Se desea estimar el contenido medio de un refresco con un nivel de confianza del 93%, con un error máximo de estimación de 5ml. Muestras previas indican que la desviación del contenido es 12 ml. Calcular el tamaño de muestra.

NC: 0.93

$\alpha=0.07$

$\alpha/2=0.035$

$Z_{\alpha/2} = 1.81$

$\sigma=12$ ml

$e= 5$ ml

$$n = \frac{Z^2 \sigma^2}{e^2}$$

$$n = \frac{1.81^2 * 12^2}{5^2}$$

$$n=19$$

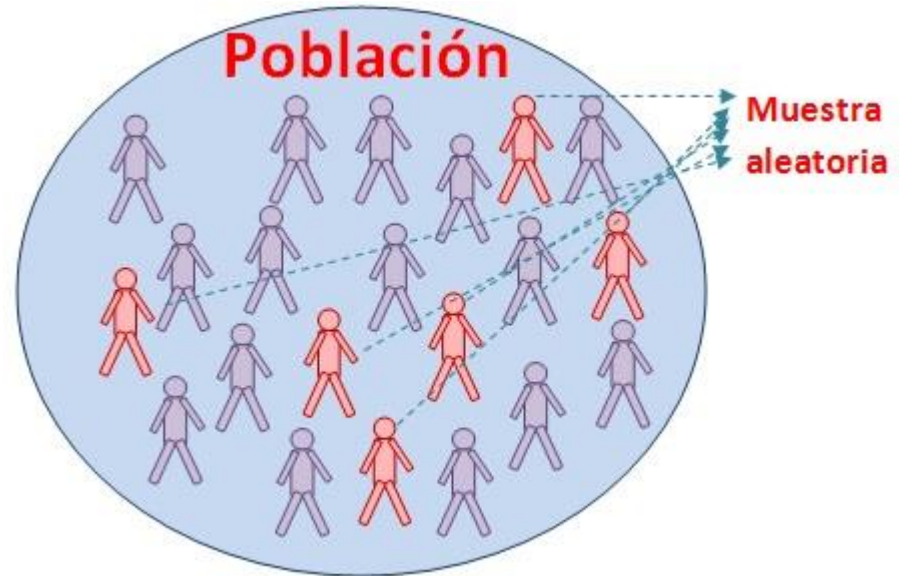
Tamaño de muestra para estimar p

Población Finita

$$n = \frac{NZ^2pq}{(N-1)e^2 + Z^2pq}$$

Población infinita:

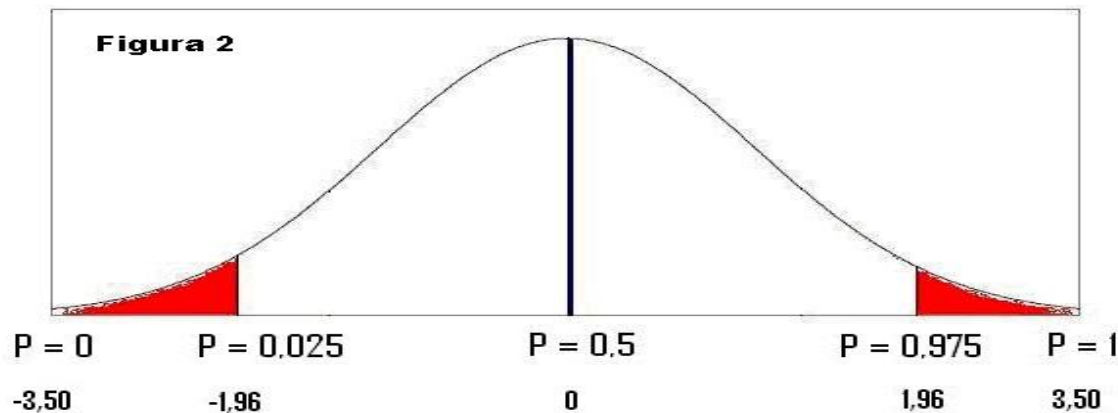
$$n = \frac{Z^2pq}{e^2}$$



Video: <https://www.youtube.com/watch?v=4G2kKHx5O8U>

¿Qué se necesita?

- NC=Nivel de confianza
- $Z_{\alpha/2}$ =valor del cuantil
- p=proporción estimada
- e= error máximo admisible



Se desea estimar con un nivel de confianza del 97% el porcentaje de clientes que compraría un nuevo producto. para esto se toma una muestra previa de 80 clientes de los cuales 65 manifestarían que comprarían el nuevo producto.

Si se desea un error máximo de estimación de 6% calcule el tamaño de muestra

$$NC=97\%$$

$$\alpha=0.03$$

$$\alpha/2=0.015$$

$$Z_{\alpha/2}=2.17$$

$$p = \frac{65}{80} \approx 0.8$$

$$q = 1 - 0.8 = 0.2$$

$$e=0.06$$

$$n = \frac{Z^2 pq}{e^2}$$

$$n = \frac{2.17^2 * 0.8 * 0.2}{0.06^2}$$

$$n \approx 210$$

Sesgo de medición

- Ocurre cuando el instrumento con el que se mide tiene una tendencia a diferir del valor verdadero en alguna dirección.

Técnicas de Muestreo

- Muestreo No-Aleatorizado (o No-Probabilista)
 - i. Se basa en el juicio personal del investigador.
 - ii. Puede generar buenas muestras pero no permite una evaluación estadística de confianza.
 - iii. Frecuentemente usado como primera aproximación
- Muestreo Aleatorizado (o Probabilista)
 - i. Se controla la probabilidad de seleccionar un determinado individuo
 - ii. Permite estudiar objetivamente la confianza de las generalizaciones hacia la población objetivo.

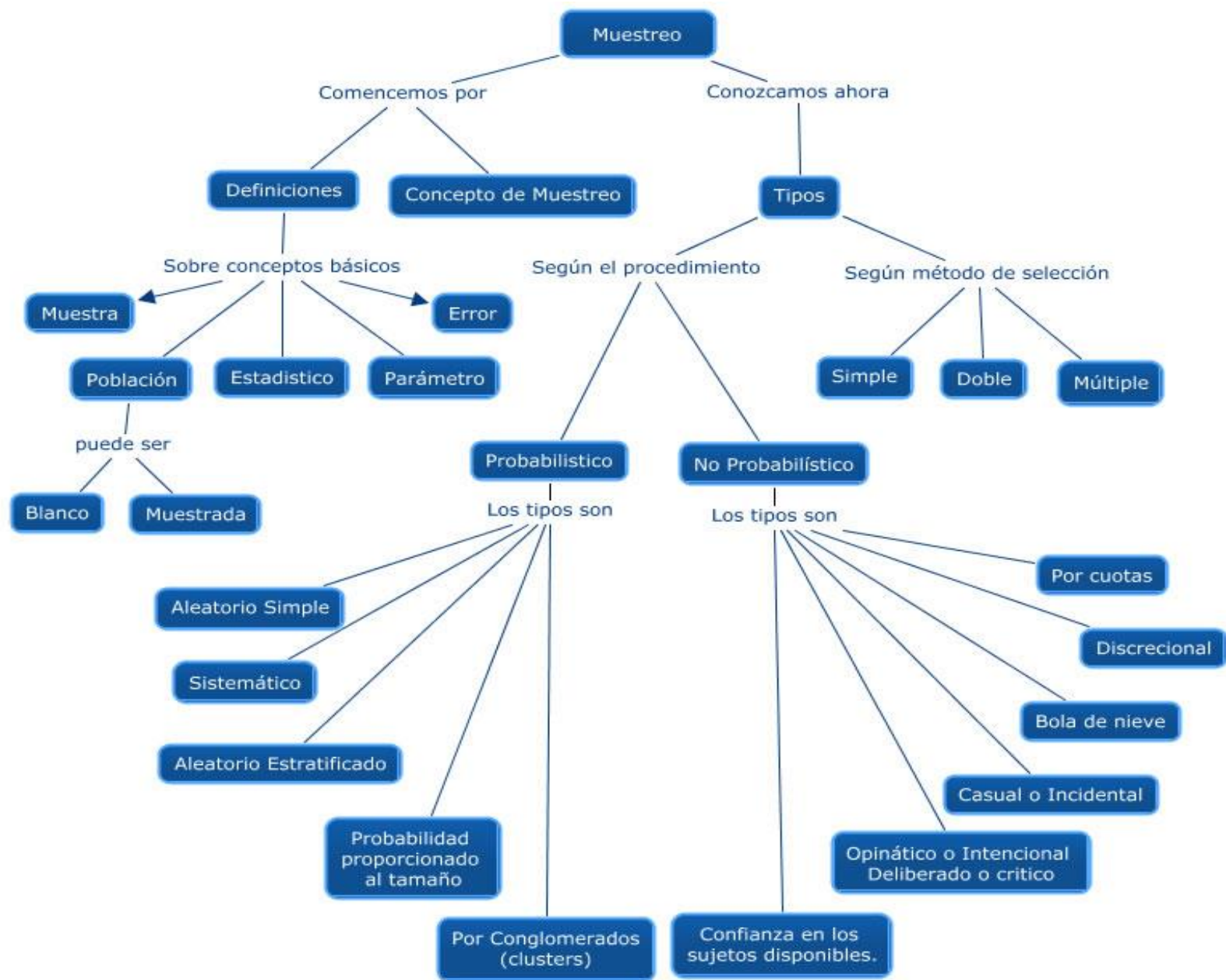
Técnicas de Muestreo

Muestreo no-Aleatorizado o no-Probabilístico

- Muestreo por convenciencia
- Muestreo por juicio
- Muestreo por cuota
- Muestreo tipo “bola de nieve” (snowball)

Muestreo Aleatorizado o Probabilístico:

- Muestreo aleatorio simple
- Muestreo sistemático
- Muestreo estratificado
- Muestreo por grupos



Muestreo no probabilista

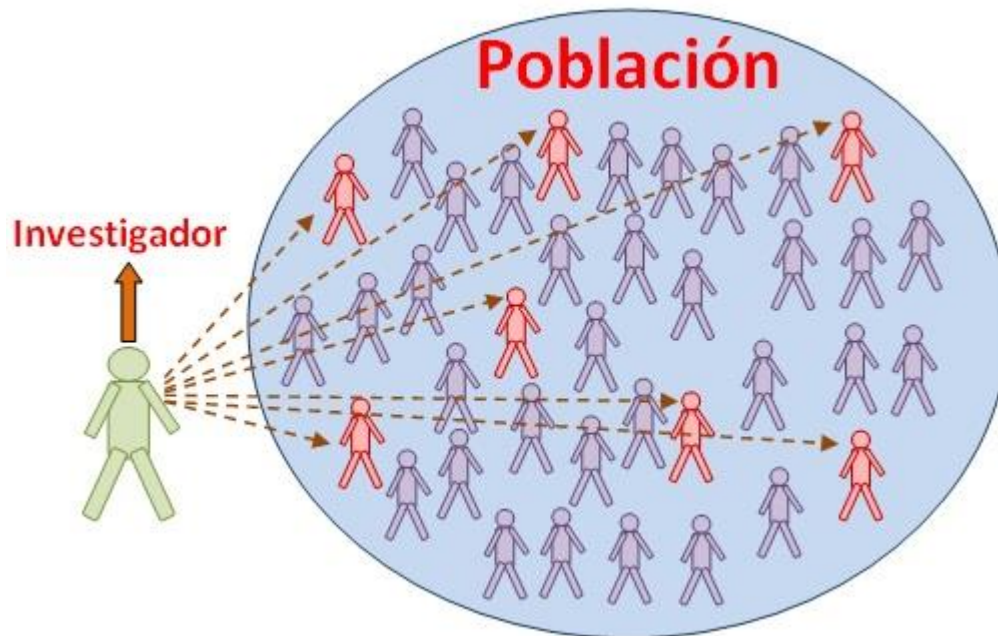
Muestreo por Conveniencia

- Los elementos de la muestra se eligen por estar en el lugar o en el momento adecuado para la investigación.
- El criterio de selección (lugar, tiempo y demás) es completamente dependiente del investigador, sin reglas predeterminadas.
- **Ejemplos:**
 - encuestas en la calle
 - encuestas a estudiantes
 - encuestas web



Muestreo por Juicio

- Se selecciona de acuerdo a alguna característica específica del encuestado juzgada por el encuestador
- Muestreo por conveniencia
- Clientes / Consumidores de un cierto tipo
- Expertos en un tema o aspecto de la organización
- Personajes “líderes de opinión”



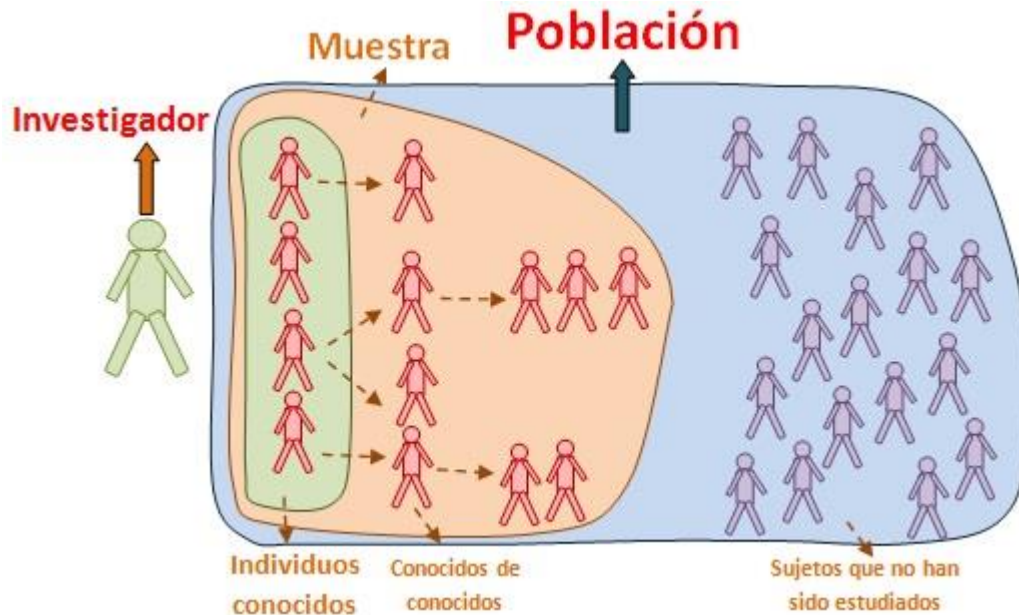
Muestreo por Cuota

- Separa la población de acuerdo a variables de control: edad, sexo, raza, nivel socio-económico
- A cada subgrupo se le asigna una proporción de muestreo, típicamente un % de la población
- La elección de la unidad en la muestra esta basada en el criterio del investigador de modo que se elige una muestra de conveniencia dentro de cada subpoblación



Muestreo tipo bola de nieve

- Se selecciona un grupo inicial
- Los nuevos encuestados se seleccionan en base a las referencias de los encuestados anteriores, explotando sus “redes sociales” .
- Muy utilizado en ciencias sociales, cuando la característica a estudiar es rara o escasa y cuando es difícil conseguir encuestados.



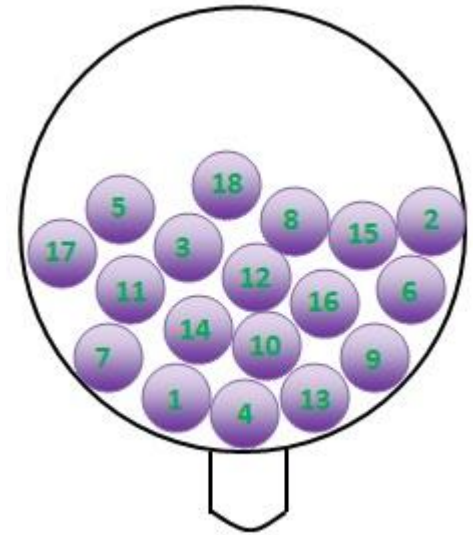
Un investigador quiere hacer un estudio sobre el comportamiento de los individuos de una secta secreta.

Empieza estudiando a tres integrantes de la misma secta que conoce y ellos le van presentando a otros sujetos para incluirlos en su estudio.

Muestreo probabilístico

Muestreo Aleatorio Simple

- Cada elemento del marco muestral tiene la misma probabilidad de ser seleccionado y cada elemento se selecciona de manera independiente de los otros.
- Con reemplazo: se pueden repetir elementos
- Sin reemplazo: no se pueden repetir elementos
- Se indexa a la población y luego se elige un índice de manera aleatoria hasta completar el tamaño deseado de la muestra.



Muestreo Aleatorio Estratificado

Antes de seleccionar los elementos, se agrupa la población muestral en estratos de acuerdo a una variable importante: edad, género, ocupación.

Objetivo: reducir la variabilidad que se puede observar dentro de cada estrato

Dentro de cada estrato se puede proceder con muestreo simple



Muestreo por conglomerados

El método de muestreo por conglomerados se utiliza cuando la población está agrupada naturalmente.

Si se supone que los conglomerados son muestra significativa de la variable que se está estudiando, se puede seleccionar algunos grupos al azar (todos los conglomerados deben tener las mismas probabilidades de ser seleccionados) y utilizarlos en representación de la población.

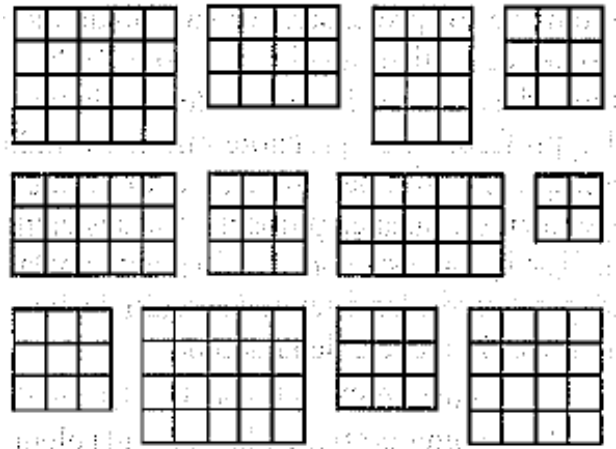
En la práctica, el conglomerado más utilizado es el geográfico. Si queremos hacer un estudio en un país, podemos dividir el país en conglomerados como las comunidades, provincias, ciudades.



Diferencia entre

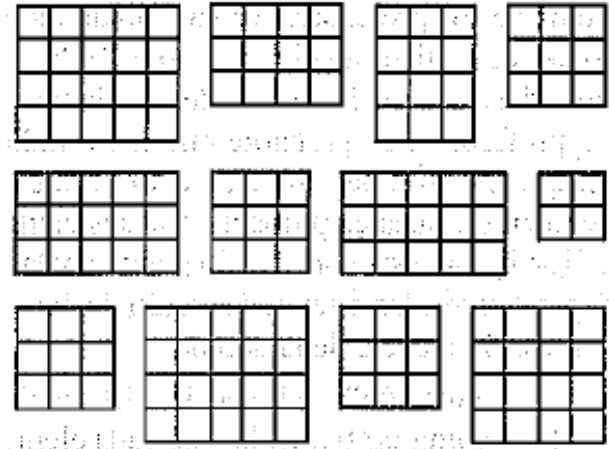
Muestreo por estratos

Población de H estratos, el estrato h tiene n elementos



Muestreo por conglomerados

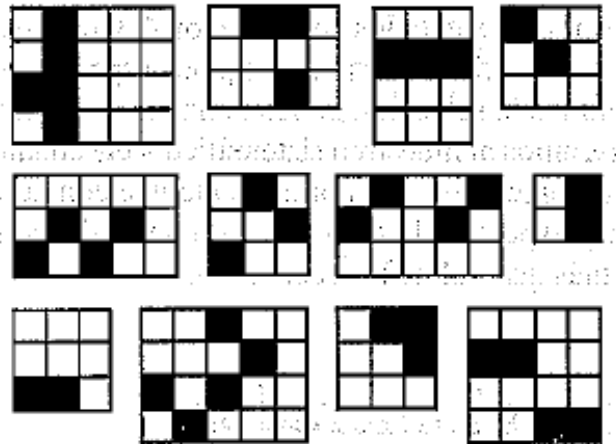
Población de N conglomerados



Diferencia entre

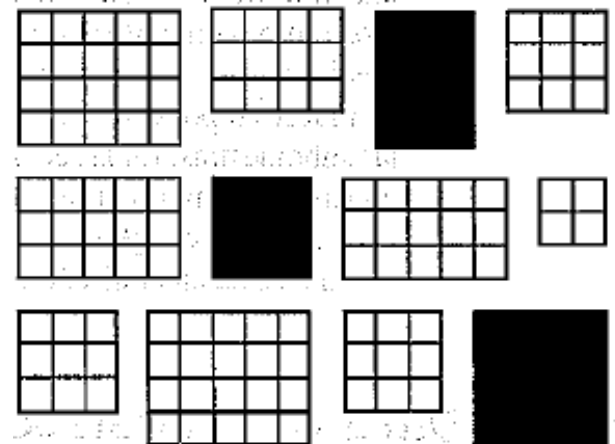
Muestreo por estratos

Se extrae una muestra aleatoria simple de cada estrato



Muestreo por conglomerados

Se extrae una muestra aleatoria simple de conglomerados, observe que todos los elementos están en la muestra



Muestreo sistemático

- Se utiliza en muestras ordenadas del 1 al N .
- Supongamos que tenemos una población de N individuos ordenados del 1 al N . Queremos seleccionar una muestra de tamaño n .
- Sea k el entero más próximo a N/n .
- Escogemos al azar un número i entre 1 y k (utilizando los números aleatorios, sacar una bola de un bombo, etc.).
- La muestra será el elemento i y los elementos $i+k, i+2k$, etc. Es decir, el elemento k y los elementos a intervalos fijos k hasta conseguir los n sujetos:

$$M = (i, i+k, i+2k, \dots, i+(n-1)k)$$

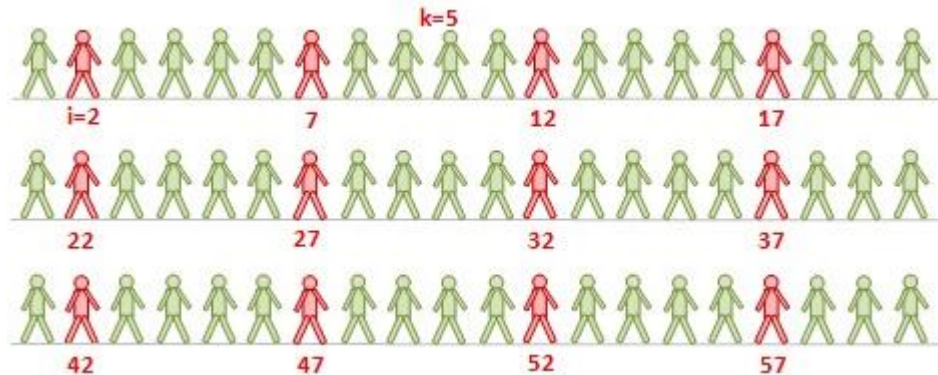


EJEMPLO

- Se quiere saber la opinión sobre un profesor de una clase de 60 personas. Dichas personas están ordenadas por orden alfabético en la lista de alumnos de clase. Para realizar la encuesta, seleccionamos a 12 personas. Por lo tanto, $N=60$ y $n=12$. El intervalo fijo entre sujetos es:

$$k = N/n = 60/12 = 5$$

Ahora elegimos al azar un número entre 1 y $k=5$. Suponemos que nos sale $i=2$. La muestra resultado mediante el muestreo sistemático será:



EJEMPLO DE UNA MUESTRA

Encuestas acerca de la preferencia de un candidato presidencial, las cuales se hacen sobre una muestra.

Indecisos, inconformes y el voto en blanco priman

El presidente candidato lidera el sondeo con el 13,6, seguido de lejos por el candidato del Centro Democrático que llegó al 6,6% de las preferencias.



Juan Manuel Santos
13,6%



Óscar Iván Zuluaga
6,6%



Enrique Peñalosa
4,9%



Antonio Navarro
4,3%



Clara López
2,0%



Martha Lucía Ramírez
0,7%

Ficha técnica

Persona natural o jurídica que la realizó:	JPG Investigación de Mercados con el apoyo de: Mediciones y Servicios de Marketing, Yamil Cure Ruiz
Persona natural o jurídica que la encomendó:	JPG Investigación de Mercados
Fuente de financiación:	Recursos propios de: JPG Investigación de Mercados, Mediciones y Servicios de Marketing, Yamil Cure Ruiz
Referencia:	Conocimiento, Imagen e Intención de Voto Primera Vuelta; Proyección candidatos finalistas segunda vuelta próximas elecciones a la Presidencia de la Republica de Colombia.
Tipo de estudio y tipo de investigación:	Exploratorio / Cuantitativa
Universo:	Área urbana de las ciudades de Bogotá, Medellín, Cali, Barranquilla y Bucaramanga
Tamaño del Universo:	El potencial de Sufragantes de estas cinco ciudades es de 9.583.254, distribuidos así: Bogotá: 5.188.174, Cali: 1.545.205, Medellín: 1.407.617, Barranquilla: 967.425 y Bucaramanga: 474.833 Fuente: Registraduría Nacional del Estado Civil 10/03/2014
Marco muestral:	Censo Electoral Colombiano discriminado por municipios.
Tipo de muestreo:	Aleatorio estratificado, modelo polietápico afijación proporcional al NSE de cada una de las ciudades objeto del estudio, en hogares, con reposición
Técnica de muestreo:	Entrevista personal cara a cara en hogares, asistido por tarjetón
Elemento muestral:	Hombres y Mujeres de 18 o más años de edad, de niveles socioeconómicos 1, 2, 3, 4, 5 y 6 residentes en las ciudades de Bogotá, Medellín, Cali, Barranquilla y Bucaramanga que <u>manifestaron intención de voto</u> para las Próximas Elecciones a la Presidencia de la Republica de Colombia
Tamaño y distribución de la muestra:	6.000 Entrevistas distribuidas así: 1.200 Entrevistas para cada una de las ciudades objeto del estudio: Bogotá, Medellín, Cali, Barranquilla y Bucaramanga
Margen de error:	Nacional: +/- 1,2267% y para cada una de las ciudades objeto del estudio: +/- 2,8132%
Nivel de confianza :	95%
Fecha del trabajo de campo:	Del 05 al 09 de Abril de 2014
Preguntas que se formularon:	Favor Referirse al Cuestionario, (5 Preguntas)
Candidatos presidenciales (personas) por los que se indago ordenados de acuerdo a las posiciones que tienen en el tarjetón electoral:	Clara López Obregón, Marta Lucía Ramírez, Juan Manuel Santos, Enrique Peñalosa, Óscar Iván Zuluaga y el Voto en Blanco