Cea D'Ancona, "La selección de las unidades de observación: el diseño de la muestra"

La selección de las unidades de observación constituye un estadio primordial en todo proceso investigador. De cómo se materialice dicha elección dependerá, en buena medida, la calidad de la información que se recoja. Razón por la cual se recomienda al investigador que no escatime, ni tiempo ni esfuerzo, en la planificación y ejecución del diseño de la muestra. Pero, ¿en qué consiste el diseño muestral?.

Una de las primeras decisiones a tomar en cualquier investigación es la especificación y acotación de la población a analizar. Por población (o universo de estudio) comúnmente se entiende "un conjunto de unidades, para las que se desea obtener cierta información". En la definición y acotación de la población han de mencionarse características esenciales que la ubiquen en un **espacio y tiempo concreto.**

Ej: "Población de 65 y más años que residen en el municipio de Madrid"

Ej: Artículos de prensa publicados en los periódicos El País, ASC y El Mundo entre 1990 y 1995"

Una vez definida la población, se procede al diseño de la muestra: **la selección de unas unidades concretas de dicha población**. Aunque el universo fuese de pequeña dimensión, por razones de economía (en tiempo y dinero), rara vez se observa a cada una de las unidades que lo forman. Por el contrario, se decide la extracción de una **muestra** de entre los integrantes del universo. Si bien, en este hacer también existen divergencias. Depende, fundamentalmente, de la estrategia de investigación que se haya escogido para la consecución de los objetivos del estudio.

El tamaño de la muestra

El número de unidades a incluir en la muestra constituye una de las decisiones preliminares en cualquier diseño muestral. En esta decisión participan diferentes factores. Éstos pueden resumirse en los seis siguientes:

a) El tiempo y los recursos disponibles para llevar a cabo la investigación.

b) La modalidad de muestreo seleccionada en orden a alcanzar los objetivos esenciales de la investigación.

c) La diversidad de los análisis de datos prevista.

d) La varianza o heterogeneidad poblacional.

e) El margen de error máximo admisible para la estimación de los parámetros poblacionales.

f) El nivel de confianza de la estimación muestral .

La varianza o heterogeneidad poblacional.

Cuando se desconoce el valor de la varianza poblacional (situación muy habitual en la práctica de la investigación social), se recurre al supuesto más desfavorable: se toma el producto de las probabilidades **"P"** (de aparición de un suceso) y **"Q"** (que indica la no ocurrencia del suceso o evento; siendo su valor igual a **"1 - P"**) como equivalente a la varianza poblacional; ambas probabilidades presentarían el valor de 0,50.

CUADRO 5.1. Tamaño muestral para poblaciones infinitas a un nivel de confianza del 95,5% (2 sigma).

Límites de error (%) para ±2σ	*Valores presupuestos de P y Q (%)*					
	1/99	10/90	20/80	30/70	40/60	50/50
0,1	39.600	360.000	640.000	840.000	960.000	1.000.000
0,5	1.584	14.400	25.600	33.600	38.400	40.000
1,0	396	3.600	6.400	8.400	9.600	10.000
1,5	176	1.600	2.844	3.733	4.267	4.444
2,0	99	900	1.600	2.100	2.400	2.500
2,5	63	576	1.024	1.344	1.536	1.600
3,0	44	400	711	933	1.067	1.111
3,5	32	294	522	686	784	816
4,0	25	225	400	525	600	625
5,0	16	144	256	336	384	400

A medida que aumenta el volumen del tamaño de la muestra, se produce un decrecimiento en el valor del error muestral. Si el tamaño de la muestra se amplía, por ejemplo, de 816 a 2.500 unidades, el error muestral disminuye en un 1,5% (pasando de representar un 3,5% a sólo un 2%).

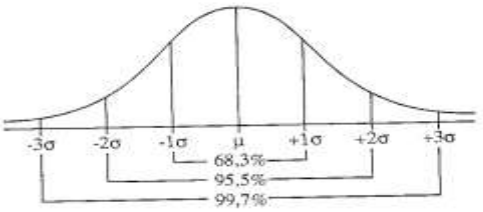
Por lo que, el investigador deberá encontrar un **punto intermedio entre el tamaño y el error muestral**, ante la tendencia observada en muestras grandes a proporcionar mínimos incrementos en adecuación en la estimación de parámetros.

El error muestral interviene en la decisión sobre el tamaño de la muestra siempre que el diseño muestral sea probabilístico. En este caso, el investigador fija el error a priori, sopesando la

precisión que desea para sus estimaciones, con los costes que supondría la reducción del error muestral. Los errores comprendidos entre el 2,5% y el 2% son los más frecuentes en la investigación social.

La distribución normal se aplica en la estadística inferencial para la estimación de la probabilidad de que un determinado evento acontezca. Representa una curva perfectamente simétrica, en forma de campana, y que admite infinitos valores (unidades "Z": unidades de desviación típica). El área total bajo la curva normal es 1 (dado que la probabilidad siempre es un valor comprendido entre 0 y 1). **En función de cuál sea el valor de "Z", variará la probabilidad concedida al evento en cuestión** (véase la tabla del área bajo la curva normal en el Anexo).

Valores de la distribución normal		*Probabilidad comprendida en el intervalo*
Entre	±1σ	0,683
	±2σ	0,955
	±3σ	0,997



CÁLCULOS

Todos los aspectos referidos participan en el cálculo del tamaño de una muestra probabilística. La fórmula genérica para una muestra aleatoria (simple o sistemática) sería la siguiente, cuando el universo o población estuviese compuesto por más de 100.000 unidades:

$$n = \frac{Z^2 \hat{S}^2}{E^2} \quad \text{o} \quad n = \frac{Z^2 \hat{P}^2 (1 - \hat{P})}{E^2}$$

donde: "Z" representa las *unidades de desviación típica* correspondientes al *nivel de confianza* elegido (2σ o 3σ, fundamentalmente).
"$\hat{S}^2$ constituye el valor de la *varianza poblacional*. Este equivale al producto de las proporciones $\hat{P}$ y $\hat{Q}$ siendo $\hat{Q} = 1 - \hat{P}$.
"E" denota el *error* máximo permitido que el investigador establece *a priori*.

Si el universo estuviese integrado por 100.000 unidades o menos, se trataría de una población finita. En este caso, habría que introducir un factor de corrección, quedando las fórmulas transformadas de la siguiente manera:

$$n = \frac{Z^2 \hat{S}^2 N}{E^2 (N-1) + Z^2 \hat{S}^2} \quad \text{o} \quad n = \frac{Z^2 \hat{P} \hat{Q} N}{E^2 (N-1) + Z^2 \hat{P} \hat{Q}}$$

donde: "N" es el tamaño de la población.
En *muestras aleatorias estratificadas* y *por conglomerados* se introducen otras variaciones que se comentarán cuando se describan ambas modalidades de muestreo (apartado 5.4).

EJEMPLOS

_____Se desea conocer el tamaño de la muestra necesario para una encuesta a la población mayor de 18 años, con objeto de medir el voto en las próximas elecciones municipales. El error máximo permitido será ±2,5%.

a) Si la encuesta se efectúa en Madrid capital, sin información previa sobre el porcentaje de voto. Para un nivel de confianza de 95,5% (2σ), el tamaño muestral necesario sería:

$$n = \frac{Z^2\hat{P}\hat{Q}}{E^2} = \frac{2^2 \cdot 50 \cdot 50}{2,5^2} = 1.600 \text{ unidades}$$

En cambio, para un nivel de confianza del 99,7% (3σ), el tamaño muestral sería más del doble:

$$n = \frac{Z^2\hat{P}\hat{Q}}{E^2} = \frac{3^2 \cdot 50 \cdot 50}{2,5^2} = 3.600 \text{ unidades}$$

b) Si se parte de la predicción (a partir de sondeos realizados con anterioridad) de que el PSOE puede alcanzar el 24% del total de votos:

$$n = \frac{Z^2\hat{P}\hat{Q}}{E^2} = \frac{2^2 \cdot 24 \cdot 76}{2,5^2} = 1.167 \text{ unidades para } 2\sigma$$

$$n = \frac{Z^2\hat{P}\hat{Q}}{E^2} = \frac{3^2 \cdot 24 \cdot 76}{2,5^2} = 2.627 \text{ unidades para } 3\sigma$$

c) Si la encuesta se llevase a cabo en el municipio de Torrelodones, con una población de hecho (censada en 1991) de 7.113 individuos.

$$n = \frac{Z^2\hat{P}\hat{Q}N}{E^2(N-1) + Z^2\hat{P}\hat{Q}} = \frac{4 \cdot 50 \cdot 50 \cdot 7.113}{(2,5^2 \cdot 7.112) + (4 \cdot 50 \cdot 50)} = 1.306 \text{ para } 2\sigma$$

$$n = \frac{Z^2PQN}{E^2(N-1) + Z^2\hat{P}\hat{Q}} = \frac{9 \cdot 50 \cdot 50 \cdot 7.113}{(2,5^2 \cdot 7.112) + (9 \cdot 50 \cdot 50)} = 2.390 \text{ para } 3\sigma$$

EL ERROR MUESTRAL

Pero, por muy perfecta que sea la muestra, como únicamente se analiza una parte de la población (y esa muestra sólo representa una de todas las posibles muestras que pueden extraerse de una misma población), **siempre habrá alguna divergencia entre los valores obtenidos de la**

muestra (estimaciones) y los valores correspondientes en la población (parámetros). Esa disparidad se denomina **error muestral: el grado de inadecuación existente entre las *estimaciones muestrales* y los *parámetros poblacionales.***

Para el cálculo del error muestral (en muestras probabilísticas, exclusivamente), se acude al estadístico llamado **"error típico"**. En la teoría de la *probabilidad* **este estadístico mide la extensión a la que las estimaciones muestrales se distribuyen alrededor del parámetro poblacional.** Concretamente, se especifica que aproximadamente el 68% de las estimaciones muestrales se hallarán comprendidas entre ±1 vez el error típico del parámetro poblacional; el 95,5%, entre ±2 veces el error típico; y, finalmente, el 99,7%, entre ±3 veces el error típico.

En suma, el **nivel de confianza en la estimación aumenta conforme se amplía el margen de error.** El nivel de confianza más utilizado es -como ya se indicó en el apartado anterior- 20' (sigma). Este nivel de confianza expresa que hay un 95,5% de probabilidad de que la estimación muestral se halle comprendida en el intervalo definido por dos veces el error típico del parámetro poblacional.

CÁLCULOS

Dadas estas variaciones en el error muestral debidas al diseño realizado, los cálculos del error típico han de adecuarse a las fórmulas apropiadas a cada diseño muestral.

Para una **muestra aleatoria simple** o **sistemática**, las fórmulas correspondientes al error típico (E) serían las siguientes:

	Universo infinito	*Universo finito (≤100.000 unidades)*
Error típico de la media	$E_{(\bar{x})} = \sqrt{\frac{\hat{S}^2}{n}}$	$E_{(\bar{x})} = \sqrt{\frac{\hat{S}^2}{n}\frac{N-n}{N-1}}$
Error típico de una proporción	$E_{(p)}\sqrt{\frac{\hat{P}\hat{Q}}{n}}$	$E_{(p)}\sqrt{\frac{\hat{P}\hat{Q}}{n}\frac{N-n}{N-1}}$

Las proporciones "P" y "Q" pueden expresarse tanto en porcentajes como en tantos por uno, obteniéndose los mismos resultados. En todos los casos, los resultados obtenidos se multiplicarán por el nivel de confianza adoptado. Si es 95,5%, por 2σ; y si fuera 99,7%, por σ.

A partir de los valores obtenidos, se calculan los intervalos de confianza para el nivel de probabilidad fijado. Ello posibilita conocer cuánto se aproxima la estimación muestral al parámetro poblacional.

EJEMPLOS

_____a) un equipo de investigadores desea conocer los hábitos de consumo de la población juvenil española. Para ello entrevistan (mediante cuestionario) a 2.000 jóvenes distribuidos por toda España. Del estudio se concluye (entre otros aspectos) que la media de gastos mensuales es de 30.000 pts al mes, con una desviación típica de 5.600. **Calcular el error de la estimación muestral. A partir de él, inferir cuál será el parámetro poblacional correspondiente a un nivel de probabilidad de 2σ.**

$$E_{(\bar{x})} = \sqrt{\frac{\hat{S}^2}{n}} = \sqrt{\frac{5.600^2}{2.000}} = 125,21$$

Si el error se multiplica por **2σ**, se realiza una estimación del intervalo de confianza que comprende el parámetro poblacional, sumando y restando a la estimación muestral el producto del error por el nivel de confianza elegido: **30.000 ± (2)*(125,21)**. En consecuencia, **hay un 95,5% de probabilidad de que la media de gastos mensuales de la población juvenil española se halle comprendida entre 29.750 Y 30.250 pesetas al mes.**

b) De los 2.950 menores internados en un centro de reforma se ha extraído una muestra de 780 menores de distintas edades. De ellos, el 49% fueron acusados de hurto. ¿Cuál será la proporción de detenidos por hurto en la población total recluida en dicho centro a un nivel de confianza de **2σ**?

$$E_{(p)} = \sqrt{\frac{\hat{P}\hat{Q}}{n}\,\frac{N-n}{N-1}} = \sqrt{\frac{(49)(51)}{780}\,\frac{2950-780}{2950-1}} = 1,54\%$$

Los límites de confianza serían, respectivamente: **49 - (2)*(1,5) Y 49 + (2)*(1,5).** Por tanto **habría un 95,5% de probabilidad de que el porcentaje de menores detenidos por hurto (en todo el centro) estuviese comprendido entre el 46 y el 52% de la población total.**

TIPOS DE MUESTREO

Las modalidades de muestreo son variadas, aunque cabe agruparlas en dos amplias categorías (**muestreo probabilístico** y **no probabilístico**), dependiendo de si el **azar** interviene en todas las fases de la selección.

• Muestreo probabilístico o aleatorio

Se fundamenta en la **aleatorización como criterio esencial de selección** muestral. Ello favorece que:

i) Cada unidad de la población tenga una probabilidad igual (y conocida a priori) de participar en la muestra; **ii)** La elección de cada unidad muestral sea independiente de las demás; **iii)** El cálculo de la adecuación de la estimación muestral (error muestral) a los parámetros poblacionales pueda hacerse dentro de unos márgenes de probabilidad específicos.

El muestreo probabilístico se adecúa más a propósitos de: **a) Estimación de parámetros. b) Comprobación de hipótesis** (test de significatividad).

• Muestreo no probabilístico

A diferencia del anterior, la extracción de la muestra **se efectúa siguiendo criterios diferentes de la aleatorización (como la conveniencia u otros criterios subjetivos). Ello da cabida a cualquier discreción por parte del equipo investigador.** Además, repercute en:

i) La desigual probabilidad de las unidades de la población para formar parte de la muestra; **ii)** La dificultad de calcular el error muestral; **iii)** La introducción de sesgos en el proceso de elección muestral. Esto redunda en riesgos superiores de invalidez de los hallazgos de la investigación.

No obstante, el muestreo no probabilístico presenta dos ventajas notorias que le hacen atractivo en la práctica investigadora: **i)** No precisa de la existencia de un marco de muestreo; **ii)** Su materialización resulta más sencilla y económica que los muestreos probabilísticos.

El muestreo no probabilístico se muestra más apropiado para: **a) La indagación exploratoria** (estudios piloto). **b) Estudios cualitativos**, más interesadas en profundizar en la información aportada que en su representatividad estadística. **c) Investigaciones sobre población "marginal"** (como prostitutas, delincuentes, drogadictos, homosexuales, inmigrantes ilegales, e incluso miembros menos "marginales" como parejas cohabitantes, por ejemplo), de

difícil registro y localización. Lo que complica la aplicación de diseños muestrales probabilísticos.

CUADRO 5.4. Modalidades de muestreo.

Probabilísticos	*No probabilísticos*
Simple Sistemático Estratificado Por conglomerados De rutas aleatorias	Por cuotas Estratégico "Bola de nieve"

MUESTREO ALEATORIO SIMPLE

Constituye el prototipo de muestreo, en referencia al cual se estiman las fórmulas básicas para el cálculo del tamaño y del error muestral. Como muestreo probabilístico, su realización exige la existencia de un marco muestral. Una vez localizado, se asigna a cada unidad de la población (en el listado) un número de identificación (si éste no figura anexo al marco muestral, siguiendo un orden consecutivo) para, posteriormente, proceder a la extracción aleatoria de los integrantes de la muestra.

La selección muestral puede hacerse siguiendo una tabla de números aleatorios, mediante un programa de ordenador u otro procedimiento.

Cuando el marco muestral se halla en soporte magnético (como el último censo de población -de 1991-, por ejemplo), se facilita la selección muestral mediante ordenador. En este caso, el programa de ordenador configurado ejecuta todas las tareas correspondientes: a) Enumerar a cada uno de los integrantes del marco muestral (o listado de las unidades de la población). b) Genera su propia serie de números aleatorios. c) Selecciona e imprime la lista de las unidades muestrales finalmente elegidas.

CUADRO 5.5. Ventajas e inconvenintes del muestreo aleatorio simple.

Ventajas	*Inconvenientes*
Facilidad de los cálculos estadísticos.	Requiere listar y enumerar a las unidades de la población.
Elevada probabilidad de lograr la equiparabilidad entre las características de la muestra y las correspondientes a la población.	Resulta monótono y arduo en muestras y poblaciones elevadas (sobre todo cuando se emplean procedimientos manuales).
	La dispersión alcanzada en la muestra repercute negativamente en los costes de la investigación.

MUESTREO ALEATORIO SISTEMÁTICO

Esta variedad de muestreo probabilístico es más practicada que la anterior. Exige igualmente la existencia de un listado de la población, pero difiere del muestreo aleatorio simple en dos aspectos fundamentales:

a) Sólo la primera unidad de la muestra se elige al azar (mediante una tabla de números aleatorios, por sorteo u otro procedimiento), con la condición de que el número seleccionado sea inferior al coeficiente de elevación. El coeficiente de elevación se calcula dividiendo el tamaño del universo por el tamaño de la muestra (N/n). Expresa el número de veces que la muestra se halla contenida en el universo.

b) Los restantes elementos de la muestra se obtienen sumando, sucesivamente, el coeficiente de elevación (a partir del primer número elegido aleatoriamente), hasta completar, al menos, el tamaño muestral.

CUADRO 5.6. Ventajas e inconvenientes del muestreo aleatorio sistemático.

Ventajas	*Inconvenientes*
No precisa del uso continuo de una tabla de números aleatorios (u otro procedimiento de asignación aleatoria). No exige que el marco muestral sea un listado. Éste, en cambio, puede adoptar varias formas (fichas, papeletas,... hasta la propia presencia física de las unidades de la población).	Necesita del recuento constante de las unidades de la población. Antes de la selección muestral, hay que desordenar el marco muestral (si éste se encuentra ordenado de de acuerdo con algún criterio que favorezca la mayor representación de determinados segmentos de la población en la muestra).

MUESTREO ALEATORIO ESTRATIFICADO

El muestreo aleatorio estratificado constituye una de las modalidades de muestreo más practicadas en la investigación social, cuando se dispone de información sobre características de la población de interés. Supone la clasificación de las unidades de población (contenidas en el marco de muestreo escogido), en un número reducido de grupos (estratos), en razón de su similaridad, dictada por las características observadas en el marco muestral. Con ello se persigue que cada estrato tenga representación en la muestra final.

Al igual que en el muestreo por cuotas, en el estratificado la muestra se distribuye en diferentes grupos de población, en función de los valores que presente en las variables elegidas

para la estratificación. No obstante, ambas modalidades de muestreo difieren en la forma de elección de las unidades muestrales. Mientras que en el muestreo por cuotas intervienen criterios subjetivos, en el estratificado sólo el azar. En él, la extracción de las unidades finales de la muestra (en cada estrato) se hace siguiendo exclusivamente procedimientos aleatorios de selección muestral (muestreo simple, sistemático, de rutas aleatorias).

En suma, si con la estratificación se persigue el logro de una mayor precisión en la estimación muestral, ésta se alcanzará cuando se cumplan dos condiciones esenciales:

a) Sean máximas las diferencias entre los estratos y mínimas dentro de cada estrato.

b) Las variables de estratificación se hallen relacionadas con los objetivos de la investigación (con las variables independientes y/o dependientes). De no ser así, se obtendría una precisión similar a la alcanzada sin la estratificación.

Las variables de estratificación más empleadas son las variables sexo y edad, debido a que ambas se hallan recogidas en la mayoría de los marcos muestrales. A ellas pueden añadirse otras variables, como la clase social, la ocupación, el nivel de instrucción; depende del objetivo fundamental de la investigación. En estudios a nivel nacional (e internacional), suele estratificarse por ubicación geográfica: ámbito territorial (país, comunidad autónoma, provincia, municipio), tipo de hábitat (urbano, semiurbano, rural), o por tamaño de hábitat (número de habitantes). Tras la clasificación de la población en estratos, se procede a fijar la muestra en cada estrato. Por afijación se entiende la distribución del tamaño muestral global entre los estratos diferenciados.

Estratificación no proporcional: La representación de los estratos en la muestra final no es proporcional a su peso en el conjunto de la población, al haberse dado una probabilidad desigual de selección en cada estrato (mediante la **afijación simple** o la **óptima**). Esta "no proporcionalidad" puede deberse al deseo de analizar, con mayor detalle, unos estratos concretos, a los cuales les correspondería un tamaño muestral inferior, si se hubiese optado por la estratificación proporcional; o, simplemente, para propiciar la representatividad de las estimaciones muestrales en todos los estratos.

Uno de los inconvenientes fundamentales de la estratificación no proporcional es la **necesidad de ponderar la muestra**, si se desea la obtención de estimaciones muestrales para el

conjunto de la población. Por el contrario, no se precisa de la ponderación, cuando sólo se realizan análisis individuales y/o comparativos entre los estratos.

Por **ponderación** se entiende el proceso de asignación de "pesos" a cada estrato, de manera que logre compensar la desigual probabilidad de selección dada a cada unidad de población que compone el estrato. A tal fin, se comparan los datos muestrales con características de la población de interés publicadas en el último censo de población, padrón de habitantes u otro sondeo a cuyos datos se conceda una significativa validez. Por lo que, antes de ponderar, el investigador deberá comprobar la adecuación de los datos que toma como referente de las características poblacionales.

Ventajas	*Inconvenientes*
Supone un menor error muestral y, por tanto, una mayor precisión de la estimación muestral. Asegura la representación de las variables de estratificación (y de las variables relacionadas con ellas). Pueden emplearse procedimientos muestrales variados en los distintos estratos. Facilita la organización del trabajo de campo.	Precisa más información del marco muestral que el muestreo aleatorio simple (para identificar a la población de cada estrato). Ello puede resultar costoso. Lleva consigo cálculos estadísticos complejos.

MUESTREO ALEATORIO POR CONGLOMERADOS

El muestreo por conglomerados también representa un procedimiento de selección aleatoria de un conjunto de individuos (referidos ahora como conglomerados).

Con el muestreo aleatorio estratificado comparte la característica básica de seccionar la población total en grupos, como fase previa a la extracción muestral. Si bien difiere de él en varios aspectos importantes:

a) En el muestreo estratificado se busca la homogeneidad dentro del estrato y la heterogeneidad entre los estratos. En el muestreo por conglomerados es a la inversa: el error muestral disminuye conforme aumenta la heterogeneidad dentro del grupo (conglomerado). Ello se debe a la necesidad de que cada conglomerado

constituya una representación, lo más ajustada posible, de la variedad de componentes del universo.

b) En el muestreo estratificado se selecciona aleatoriamente una muestra para cada estrato. En el muestreo por conglomerados lo que se extrae es una muestra aleatoria de conglomerados. Sus integrantes formarán la muestra.

c) En el muestreo estratificado, la unidad de muestreo es el individuo. En cambio, en el muestreo por conglomerados es el conglomerado (o conjunto de individuos).

Los **conglomerados** pueden ser las **áreas geográficas** que dividen a la población que se analiza (país, comunidad autónoma, municipio, distrito, áreas censales, viviendas); pero, también, **organizaciones o instituciones** (colegios, hospitales, tribunales, centros penitenciarios).

a) Los conglomerados han de estar bien **definidos y delimitados.** Cada unidad de la población sólo puede pertenecer a un único conglomerado.

b) El **número de elementos** que componen el conglomerado ha de ser conocido previamente (aunque sea de manera aproximada).

c) Los conglomerados elegidos **han de ser pocos**, si realmente quieren reducirse los costes de la investigación.

d) Los conglomerados deberían escogerse de manera que se consiguiera **disminuir el aumento en el error muestral, generado por la agrupación** (o aconglomeración).

e) Los conglomerados **no tienen por qué hallarse idénticamente definidos en todos los lugares.**

Cuando se muestrearon individuos u hogares en áreas urbanas, los conglomerados suelen ser bloques o conjuntos de bloques. En cambio, en las áreas rurales, los conglomerados serán segmentos geográficos limitados por carreteras y fronteras naturales (como ríos y lagos).

Si, a partir de una muestra por conglomerados, se extrae una **nueva muestra**, con referencia a cada uno de los conglomerados previamente elegidos, y así, sucesivamente, se está ante un diseño muestral muy habitual en la investigación social: el **muestreo polietápico por conglomerados.**

El **muestreo polietápico (o polifásico) por conglomerados** representa una extensión del muestreo por conglomerados. En él la unidad de muestreo final no son los conglomerados, sino subdivisiones de estos. Por lo que no se toman cada uno de los integrantes de los conglomerados elegidos aleatoriamente, sino sólo a una parte de ellos, escogidos también de forma aleatoria. Ello supone muestrear -como su nombre indica- en distintos niveles, implicando varios procedimientos de selección muestral (como la estratificación).

La modalidad de muestreo polietápico por conglomerados más sencilla implica la extracción muestral en dos fases : a) En la **primera fase**, se seleccionan las agrupaciones de los miembros de la población de estudio (conocidas como las unidades de muestreo primarias), que son análogas a los conglomerados. b) En la **segunda fase**, se eligen aleatoriamente los miembros de la población a observar, de las unidades de muestreo primarias previamente seleccionadas.

En el caso más general (en muestras nacionales), se recurre a **muestreos polietápicos, estratificados, por conglomerados:** primero, se afija la muestra por estratos; después, se extraen (de forma aleatoria proporcional) los municipios, las secciones estadísticas; y, por último, los hogares en los que se realizarán las entrevistas.

En general, se recomienda aumentar el número de conglomerados (en este caso de facultades) con preferencia a elevar el número de individuos a observar en cada uno de ellos (los profesores). La razón está en la probable **homogeneidad de los conglomerados elegidos**.

Cuando éstos son bastante homogéneos, no se precisa añadir más elementos del conglomerado a la muestra, ya que se obtendría una información redundante, al ser similares las características de las unidades que forman el conglomerado. En este caso (**cuando los conglomerados son homogéneos**), se aconseja ampliar el número de conglomerados para, de esta forma, abarcar una mayor variedad de la población de interés. Por el contrario, **si los conglomerados fuesen heterogéneos**, la mejor opción (entendida como reducción del error muestral) sería la opuesta: reducir el número de conglomerados, aumentando las unidades a observar en cada uno de ellos.

• En suma, el muestreo aleatorio por conglomerados se muestra de especial interés cuando: a) Resulte difícil compilar una lista exhaustiva de todos los componentes de la población. Lo que imposibilitaría la práctica de otra variedad de muestreo probabilístico . b) Se quiera reducir la duración y los costes económicos del trabajo de campo en la investigación. c) Se

realicen **estudios de ámbito nacional** o internacional, que supongan una considerable dispersión de la muestra.

Ventajas	*Inconvenientes*
No exige un listado de toda la población de interés; sólo de las unidades del conglomerado. Al concentrar el trabajo de campo en un número limitado de puntos de muestreo, disminuyen los costes de la investigación.	Mayor error muestral (y, por tanto menor precisión de las estimaciones muestrales) porque los conglomera do suelen ser muy homogéneos *. Requiere cálculos estadísticos complejos en la estimación del error muestral, principalmente.

* El error muestral puede reducirse aumentando el número de *conglomerados*.

MUESTREO POR CUOTAS

Una de las modalidades de muestreo más populares es el **muestreo por cuotas**. Esta constituye una variedad de muestreo no probabilístico que parte, igualmente, de la segmentación de la población de interés en grupos, a partir de variables sociodemográficas relacionadas con los objetivos de la investigación. Por lo que, su puesta en práctica conlleva, también, la elaboración de una matriz con las características básicas de la población que se analiza (proporciones de población diferenciadas por sexo y edad, nivel de instrucción, clase social). Esta información suele obtenerse del último censo de población, padrón de habitantes u otra fuente estadística similar.

El propósito es seleccionar una muestra que se ajuste a la distribución de las características fundamentales de la población. Además de los objetivos del estudio, en la elección de las variables intervienen otros factores: la precisión que el investigador desee para su indagación, junto a la accesibilidad de las variables elegidas. Ésta dependerá de la facilidad de su obtención en el marco de muestreo elegido pero, también, de su practicabilidad (si el entrevistador puede acceder fácilmente) a los grupos de población definidos por las variables escogidas.

Las cuotas más habituales son las determinadas por la conjunción de las variables sexo y edad, en consonancia con su **mayor accesibilidad.** La mayoría de los marcos muestrales contienen ambas variables. A esto se suma la relativa **facilidad (para el entrevistador) de localizar visualmente a personas que pertenezcan a los distintos grupos de sexo y edad.**

Después de la delimitación de las cuotas, se proporciona a cada entrevistador su asignación correspondiente: **el perfil y el número de personas a entrevistar en cada cuota.**

Lo que distingue al muestreo por cuotas, respecto del estratificado, es la libertad que se da al entrevistador para la elección de las unidades finales de la población a entrevistar. Aunque el azar intervenga en las fases iniciales del diseño muestral (en la elección de áreas o zonas geográficas, por ejemplo), la selección de los elementos concretos de la población es totalmente arbitraria. Es **el propio entrevistador quien elige al entrevistado,** en cualquier momento y lugar (en una calle comercial, a la salida del metro, en la parada del autobús, en un mercado, en un parque). La única condición que se le impone es que la persona se ajuste a las cuotas fijadas por el equipo investigador.

Ventajas	*Inconvenientes*
Resulta más económico (en tiempo y dinero) que los muestreos probabilísticos.	Supone un mayor error muestral que los diseños probabilísticos.
Fácil de administrar.	Inexistencia de algún método válido para calcular el error típico (al no ser un muestreo probabilístico).
No precisa de un listado de la población.	Limites en la representatividad de la muestra para las características no especificadas en los controles de cuotas.
	Dificultad para controlar el trabajo de campo.

Hernández Sampieri, R. (2010) Metodología de la Investigación "¿Cómo seleccionar una muestra?

¿EN UNA INVESTIGACIÓN SIEMPRE TENEMOS UNA MUESTRA?

No siempre, pero si en la mayoría de los casos realizamos el estudio de una muestra.

¿SOBRE QUÉ O QUIENES SE RECOLECTAN LOS DATOS?

Aquí el interés se centra sobre "*qué* o *quiénes*", es decir, en los *sujetos, objetos, sucesos, eventos* o *contextos de estudio*. Eso depende del planteamiento inicial de la investigación.

Al definir la **unidad de análisis**, el sobre que o quienes se van a recolectar datos depende de: i) el enfoque elegido (cuantitativo, cualitativo o mixto, ii) el planteamiento del problema a investigar y iii) de los alcances del estudio. Estas acciones nos llevarán al siguiente paso que consiste en **delimitar una población.**

•**Muestra (enfoque cuantitativo)**: subgrupo de la población del cual se recolectan los datos y debe ser representativo de dicha población.

•**Muestra (enfoque cualitativo)**: unidad de análisis o conjunto de personas, contextos, eventos o sucesos sobre el (la) cual se recolectan los datos sin que necesariamente sea representativo (a) del universo.

¿CÓMO SE DELIMITA UNA POBLACIÓN?

Lo primero es definir si nos interesa o no delimitar una población y si pretendemos que esto sea antes de recolectar los datos o durante el proceso. En los estudios cualitativos por lo común la población o el universo no se limita a priori. En los cuantitativos casi siempre si.

Para el enfoque cuantitativo las unidades deben situarse claramente en torno a sus características de contenido, de lugar y en el tiempo.

¿CÓMO SE SELECCIONA UNA MUESTRA BAJO EL ENFOQUE CUANTITATIVO?

La muestra es en esencia un subgrupo de la población. Digamos que es un subconjunto de elementos que pertenecen a ese subconjunto definido en sus características al que llamamos población.

Ritchey, F. (2001), Estadística para las ciencias sociales. El potencial de la imaginación estadística

DISPERSIÓN

Dispersión refiere a cómo se extienden las puntuaciones de una variable de intervalo/razón de la menor a la mayor y la forma de distribución entre estas. Existe un número infinito de posibles formas de distribución para una variable con una media dada. Los

estadísticos de dispersión permiten descripciones precisas de la frecuencia de casos en cualquier punto de una distribución.

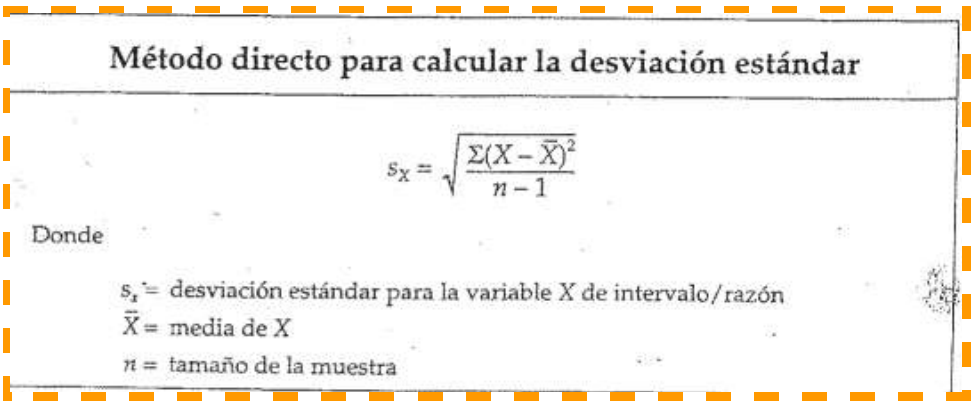
EL RANGO

El **rango** es una expresión de cómo las puntuaciones de una variable de intervalo/razón se distribuyen de la menor a la mayor. (rango= puntuacion maxima - puntuacion minima)

LA DESVIACIÓN ESTÁNDAR

El **desvío estándar** describe cómo las puntuaciones de una variable de intervalo/razón u ordinal de tipo intervalo se extienden a lo largo de una distribución, en relación a la puntuación media.

La desviación estándar se calcula determinando que tan lejos está cada punto de la media (que tan lejos se desvía de la media). En este sentido, la desviación estándar es un derivado de la media y las dos medidas siempre se informan juntas. La desviación estándar nos dice con qué amplitud se agrupan las puntuaciones alrededor de la media.



Boado, M. (2013) ***Revisión de análisis de tablas e introducción a modelos loglineares.***

LAS CHANCES Y CHANCES RELATIVAS (odds y odds ratio).

Una forma alternativa y equivalente a la fundada en las probabilidades estimadas y esperadas es la basada en las chances y chances relativas. La chance es la ventaja de ser **i** dado **j** frente a ser **j** dado **j**. Es una razón que se estima entre dos valores. Da una idea de competencia u oportunidad. Obviamente se deriva de una tradición matemática vinculada a los juegos de azar. Y a los efectos de los cálculos, como se verá, permite muchas más opciones que los procedimientos anteriores.

Cuando se aplica a los **valores marginales** se denomina **chance marginal**. Cuando se aplica a los **valores internos** de la tabla se denomina **chance condicional.**

Las chances condicionales son muy importantes, y sirven para estimar las ventajas de un resultado frente a otro. **Chance (odd) y probabilidad** implican conceptos diferentes, pero relacionados entre sí: **Chance:** Uno, trata una ventaja de ocurrencia (o mejor dicho, de lo ocurrido); **Probabilidad:** y el otro, la proporción de los casos de una combinación en el total de casos, o en el total de casos por fila o columna.

odd = probabilidad/(1 - probabilidad) y **probabilidad = odd/(1 + odd)**

Es usual que entre los diversos resultados que aporta una tabla, se relacionan las chances que se observan.

La nueva forma de comparación que surge es una **'razón de chances',** o sea una razón de las razones previamente observadas, y se la denomina **'chance relativa'**, o 'ventaja relativa', u **'odds ratio'**, o 'razón de momios'. La **'chance relativa es idéntica' al 'producto cruzado' de una tétrada de celdas,** por razones aritméticas claras.

¿Qué comparo cuando leo una razón de chances, o chance relativa?

La ventaja de ser Bi antes que Bj dado que se es Ai, frente a ser Bi antes que ser Bj dado que se es Aj.

La razón de chances, u odds ratio, estima y **mide una ventaja que nos interesa** en relación a una 'base de comparación'.

Var fila Voto	Var col: generaciones		
	1=Joven	2=Viejo	Total
1=Pcol	50	73	123
2=Pnal	43	21	64
3=FA	80	19	99
Total	173	113	286

•La chance de 1,1 frente a 1,2 , es decir de ser 1 antes que 2 dado que son ambos 1, o en categorías de ser joven y votar colorado respecto a ser viejo y votar colorado, es (50/73) = 0,68. La chance de ser joven y votar colorado es casi un tercio menor para los jóvenes respecto de los viejos.

•La chance de 3,1 frente a 3,2 , es decir de ser 1 antes que 2 dado que son ambos 3, o en categorías de ser joven y votar FA respecto a ser viejo y votar FA, es (80/19)=4,21. La chance de ser joven y votar FA es 4 veces mayor para los jóvenes que para los viejos.

• La chance relativa u odds ratio es la razón de ambas razones. Así: (50/73) / (80/19), que equivale a (50 x 19) / (80 x 73), da por resultado 0,16.

Entonces la chance de votar colorado en los jóvenes se reduce a 1/6, y por su parte la chance de votar FA en los viejos se reduce a 1/6. O, puesto de otra manera más contundente: la chance relativa de votar FA antes que colorado es casi 6 veces superior en los jóvenes que en los viejos, y viceversa la chance relativa de votar colorado antes que FA en los viejos es 6 veces mayor que en los jóvenes.

PROPIEDADES DE LAS CHANCES RELATIVAS U ODDS RATIO

1. Son siempre positivas.

2. Son invariantes.

3. Cuando el resultado de la chance relativa adquiere valor igual a 1 es sinónimo de independencia (NO asociación) en la dicotomía o tétrada de celdas que se considera.

4. Cuando adquiere valor mayor o menor que 1 es sinónimo de asociación.

5. Si bien su distribución se sesga hacia la derecha, porque la chance relativa u odds ratio varía entre 0 y + ∞, ello se 'corrige' en los paquetes estadísticos convencionales estimando el log odds ratio, o logaritmo natural de la razón de chances, que varía entre - ∞ y + ∞, con el valor 0 como indicativo de independencia.

Cea D'Ancona, M. A. (1996), Metodología cuantitativa. Estrategias y técnicas de investigación social

EL ANÁLISIS MULTIVARIABLE

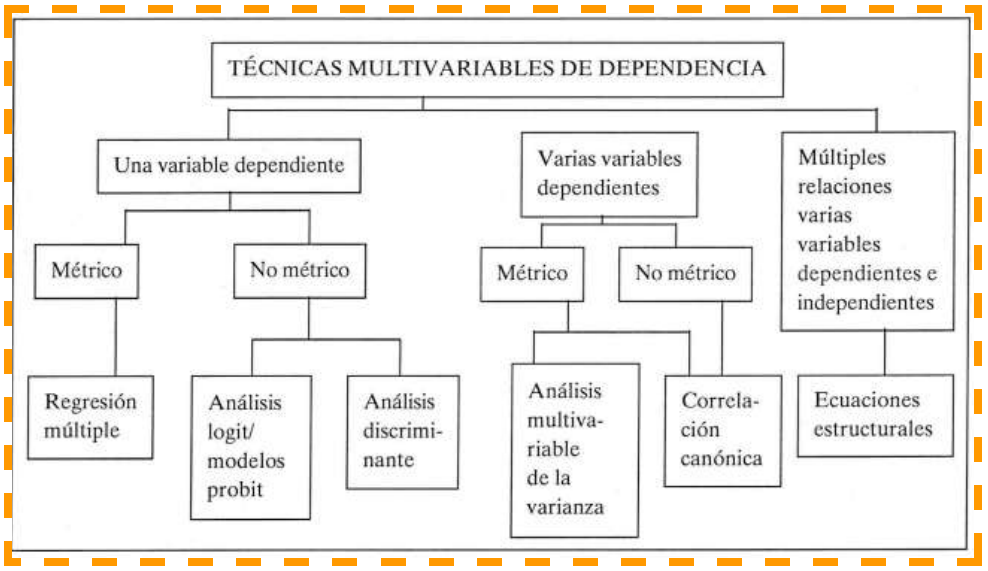
Los análisis univariables y bivariables con frecuencia se muestran insuficientes para cubrir los objetivos de la investigación. El proporcionar una visión conjunta e integrada, que describa y/o explique la realidad que se analiza, demanda la realización de **análisis multivariables** (de más de dos variables al mismo tiempo).La peculiaridad del análisis multivariable reside en operar con un número elevado de variables, y de manera simultánea, basándose en el cálculo matricial.

Para el análisis multivariable existe un amplio abanico de técnicas. Estas técnicas pueden agruparse en dos grandes categorías (**técnicas de dependencia y técnicas de interdependencia**), en función de si se diferencia, o no, entre variables dependientes e independientes.

En la elección de la técnica concreta a aplicar intervienen, básicamente, el objetivo de la investigación, y las características de las variables que se analicen (su número y nivel de medición).

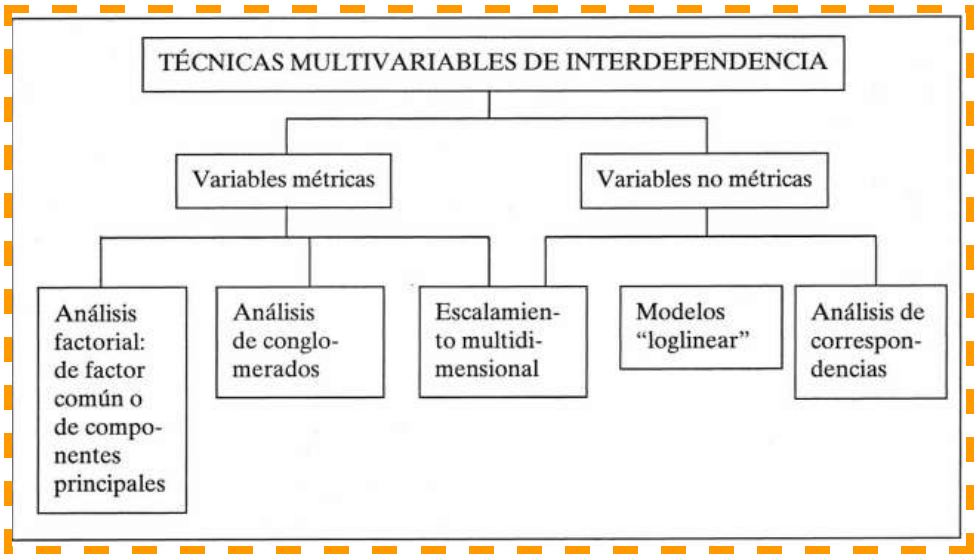
TÉCNICAS MULTIVARIADAS DE DEPENDENCIA

Un conjunto de técnicas analíticas unidas por un mismo propósito: medir la existencia de relaciones causales entre un conjunto de variables, el grado y significatividad de la misma. Sin embargo, difieren en el número de variables dependientes que incluyen, y en el nivel de medición exigido (métrico o no métrico).



TÉCNICAS MULTIVARIADAS DE INTERDEPENDENCIA

A diferencia de las técnicas analíticas anteriores, las de interdependencia presentan un menor poder predictivo. Mediante ellas se analiza la existencia de asociación o relación mutua entre varias variables, sin diferenciar entre dependientes e independientes. Ahora la diferencia básica entre las técnicas se establece en función del nivel de medición mínimo exigido en las variables para su cumplimentación: métrico o no métrico.



Sánchez Vizcaíno, G. (2009). "Regresión logística". Técnicas de análisis de datos en investigación de mercados.

Puntaje Z clase

¿Cómo calcular porcentajes de puntaje Z?

X= Valor a buscar en la tabla Z

μ= Media aritmética

σ= Desvío típico

$$Z = \frac{x - \mu}{\sigma}$$

1) Buscar el puntaje z en la tabla.
2) Sumarle el otro 50% para ver el total de casos del otro lado de la curva.

¿Cómo calcular valores porcentuales a partir del puntaje Z?

***Z*= Valor a buscar en la tabla Z**

(Recordar que el puntaje Z va DEL CENTRO hacia los costados)

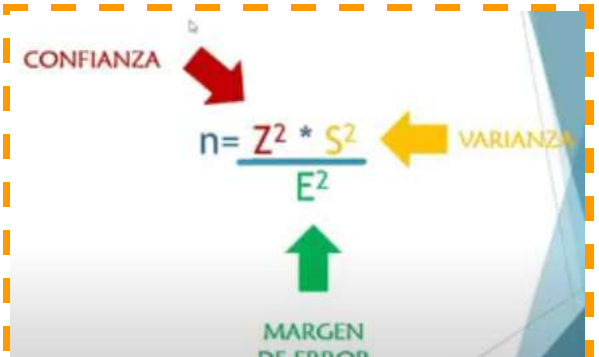
μ= Media aritmética

σ= Desvío típico

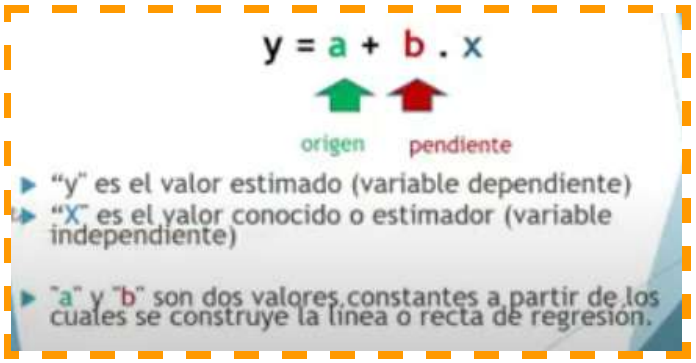
$$x = \mu \pm z * \sigma$$

1) Buscar el puntaje Z porcentual que buscó resolver
2) Resolver calculando a partir de que si se busca un número por debajo del -50% restando y si se busca un número por sobre el +50% sumando.
3) Se resuelve la ecuación.

¿Cómo calcular el tamaño de una muestra?

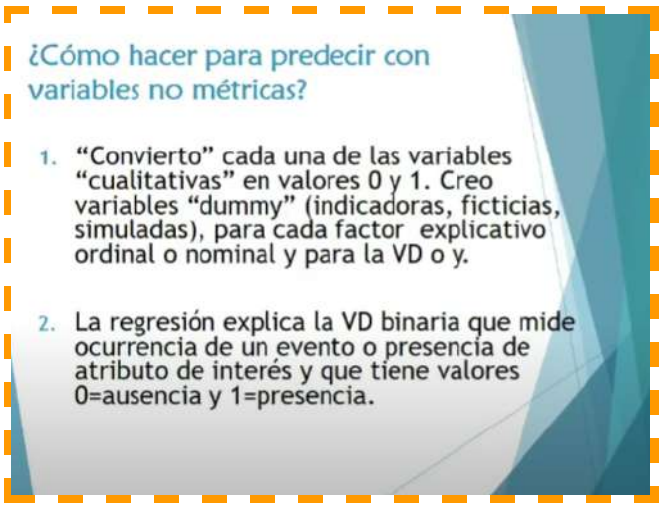


Regresión Lineal



Regresión logística Binomial

Se usa para variables cuantitativas dicotómicas. ¿Va a suceder tal fenómeno? Son de variables con 2 posibilidades únicamente. Es para medir qué tan verosímil es un conjunto de VI para medir esa VD. Pero debe transformar las variables cualitativas en variables numéricas.



Confrontas 2 variables cualitativas entre sí (escuelas técnicas vs todas las demás).

Análisis de Clusters