Estadística fundamental Introducción al muestreo estadístico Máster en Data Science y Big Data (Finanzas)

Pilar Barrios

Diciembre 2021



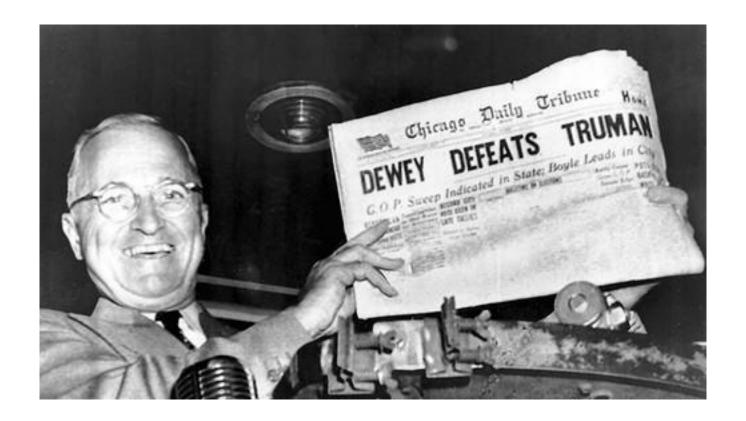


Índice

- 1. Introducción
- 2. Tipos de muestreo
- 3. Conceptos básicos
 - a. Precisión y error de muestreo
 - b. Confianza
 - c. Errores no atribuibles al muestreo
 - d. Tamaño de la muestra
 - e. Fases de la investigación por muestreo







https://www.history.com/news/dewey-defeats-trumanelection-headline-gaffe



En la práctica profesional, nos enfrentamos con frecuencia a la necesidad de **conocer realidades respecto al universo de datos de estudio, N,** que desea verificar.

Sin embargo, muchas veces, el **grupo en estudio es tan grande** que hace difícil, casi imposible el examen de todas sus partidas, por lo que se procede comúnmente a **examinar un determinado número de ellas**, observar el comportamiento de las variables en estudio en este conjunto limitado para luego extender los resultados al conjunto total, con cierto grado de confianza. Es entonces, importante conocer que se puede estudiar un subconjunto de la población para luego **inferir los resultados al conjunto total de elementos** con un cierto grado de precisión, el cual dependerá de la selección de los elementos del subconjunto a estudiar y de los métodos de inferencia para la obtención de premisas generales.

Esto es la "teoría del muestreo", la selección adecuada en cantidad y calidad de las partidas que constituirán el grupo de observación. Para ello es necesario tener conceptos muy claros respecto a población o universo y muestra.



El muestreo es una técnica muy habitual en muchos ámbitos: encuestas del INE, cálculo del IPC, estudios demográficos, de control de calidad, etc. Por ejemplo en las entidades financieras es muy usado en auditoría y consiste en aplicar pruebas de auditoría a menos de la totalidad de un conjunto de transacciones o saldos contables, a fin de sacar conclusiones sobre todo el conjunto.

Al aplicar las pruebas a cada transacción de la muestra, se podrá comprobar si esta cumple con ciertos requisitos preestablecidos. En caso de que no cumpla estaremos ante un **error o desviación**, y tratándose de una prueba de controles se estaría ante la evidencia de que el control bajo análisis no funcionó correctamente en ese caso particular.

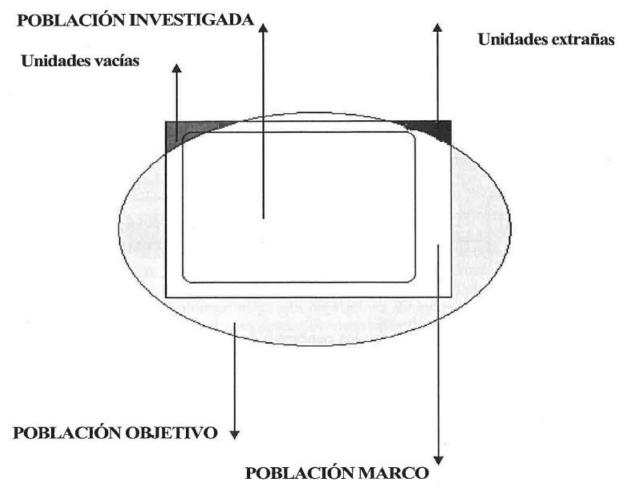
Dado que las desviaciones se proyectan a la totalidad de la población de acuerdo a los resultados de las pruebas sobre las transacciones de la muestra, es deseable que una muestra extraída sea **representativa** de la población total. Las muestras representativas permitirán hacer estimaciones sobre la población más aproximadas a la realidad.



Cada aplicación de muestreo debe ser establecida teniendo en cuenta las siguientes **preguntas**:

- ¿Qué queremos conseguir? (**Objetivos** de la prueba)
- ¿Qué es lo que abarcará la muestra? (Definición del universo y de la unidad de muestreo)
- ¿Cómo se realizará el muestreo? (El **método de selección**, muestreo estadístico o no estadístico)
- ¿Cuánto se incluirá en el muestreo? (Tamaño de la muestra, n)
- ¿Qué significan los resultados? (**Evaluación** e interpretación)





Fuente: Pérez, C. Técnicas de muestreo estadístico



Población objetivo → Población que se desea investigar

Problemas: inaccesibilidad a algunas unidades

negativas a colaborar

ausencias

errores en los instrumentos de medida

Población investigada → Población que realmente es objeto de estudio (teniendo en cuenta los problemas citados

Población marco → Lista de unidades de muestreo de entre las que se selecciona la muestra. Es la contrapartida en el mundo real de la población objetivo

Unidad extraña → Unidad que no es realmente del marco y que no tiene relación de ningún tipo con la población objetivo

Unidad vacía → Unidad erróneamente incluida en el marco y que no pertenece a la población objetivo aunque esté directamente relacionada con ella



Duplicaciones → Unidades repetidas en el marco

Omisiones → Unidades omitidas en el marco que son del marco realmente

Marcos imperfectos → Marcos con unidades vacías, extrañas, duplicaciones y omisiones

Marco en sentido amplio → incluye información complementaria (variables auxiliares, encuestas piloto)

Muestra → conjunto de unidades de muestreo seleccionadas de un marco o de varios

Unidad elemental de muestreo → elemento más simple de la población

Unidad de muestreo compuesta o primaria → se compone de varias unidades elementales



Los objetivos deberían:

- estar claros y
- explicitarse antes de avanzar sobre el programa de muestreo,

ya que estos inciden sobre el plan de muestreo que habrá de elegirse.

A medida que se va puliendo el programa de muestreo, será necesario iterar con los objetivos, con el alcance, y con la precisión del muestreo, ya que puede resultar necesario considerar la suficiencia de los recursos con los que se cuenta.

Al finalizar el proceso, el alcance de la auditoría debe ser consistente:

- con los recursos disponibles y
- con el programa de muestreo asociado.



Los objetivos del muestreo son, por tanto, inferir de una muestra (o subgrupo de elementos extraídos del universo a analizar) conclusiones sobre la población. Es decir:

Nos interesa conocer un dato poblacional \rightarrow parámetro θ

Para ello estimamos valores de una muestra \rightarrow estimador $\hat{\theta}$

Se define la esperanza matemática del estimador $\hat{\theta}$ del parámetro poblacional θ

$$E(\hat{\theta}) = \sum_{S} \hat{\theta}(s_i) P(s_i) = \sum_{R} t P^{T} (\hat{\theta} = t)$$

Siendo P(s_i) la probabilidad de que a muestra i salga elegida y $\hat{\theta}(s_i)$ el valor del estimador en la muestra i.

 $P^{T}(\hat{\theta} = t)$ es la probabilidad de que el estimador tome los distintos valores posibles tal que $(X_1 \cdots X_n) \to \hat{\theta}(X_1 \cdots X_n) = t$, donde X_i representan los valores que toma el parámetro de la población.



Al ser $\hat{\theta}$ un valor estimado, no tiene por qué coincidir con el valor poblacional, y nos interesará conocer cuánto nos estamos desviando del valor poblacional.

Queremos saber cuánto nos desviamos \rightarrow varianza $V(\hat{\theta})$

Pero esta información provendría de la población por lo que también es desconocida y necesitaremos estimarlo.

Hallamos la estimación de la desviación \rightarrow varianza estimada $\hat{V}(\hat{\theta})$

Cuanto menor sea la desviación, mayor será la precisión del análisis.

Así, la varianza del estimador se define como una medida que cuantifica la concentración de las estimaciones alrededor de su valor medio.

$$V(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2 = \sum_{R} (t - E(\hat{\theta}))^2 P^T(\hat{\theta} = t) = E(\hat{\theta}^2) - (E(\hat{\theta}))^2$$





El muestreo se puede dividir en "probabilístico" y "no probabilístico", también denominados "estadístico" y "no estadístico", o "al azar" y "no al azar", respectivamente.

El hecho de extraer una muestra **no significa azar o aleatoriedad**. En ello es determinante la fase mecánica de selección de las unidades de muestreo que conformarán la muestra.

La diferencia fundamental entre estos tipos de muestreo radica en que en el muestreo **probabilístico** se puede **medir el riesgo que se asume al muestrear**, mientras que en el muestreo no probabilístico ello no es posible. En el estadístico, la fundamentación es teórica y las **conclusiones pueden ser cuantitativas y precisas**.

En el muestreo probabilístico, se le asigna a cada elemento de la población una probabilidad de aparecer en la muestra.



Mientras que en el **no estadístico**, la fundamentación es práctica y las **conclusiones pueden ser cualitativas** y, en el mejor de los casos, razonables, además de perder ante determinados casos la posibilidad de ser extrapolables.

Ante estas diferencias, ¿por qué recurrir al muestreo no probabilístico?

En algunas circunstancias sucede que no es posible por falta de tiempo, por escasez de recursos, por limitaciones para acceder a la población u otras dificultades operativas llevar a cabo un muestreo probabilístico.

Será preferible, entonces, acudir a determinadas pruebas específicas, para las que habrán de tenerse en cuenta ciertos cuidados mínimos y necesarios para sostener fundadamente las opiniones vertidas en un informe.



Muestreo probabilístico

- Basado en el cálculo de probabilidades
- Cada unidad elemental tiene una probabilidad conocida de pertenecer a la muestra y su valor se utiliza en el cálculo de los estimadores

```
\{U_1, U_2, ..., U_N\} Se toma una muestra de tamaño n
```

Ejemplo: de una urna llena de bolas de colores (población), se extraen n bolas (muestra).

El muestreo puede ser además:

- Sin reemplazamiento (SR)
- Con reemplazamiento (CR)



Por otro lado los métodos de selección también pueden ser clasificados según otras características.

- Hasta ahora solo nos fijábamos en si la unidad se reponía o no (si volvíamos a depositar la bola en la urna o no).
- En estos casos, todos los elementos de la población tenían la misma probabilidad de aparecer en la muestra (salir elegidos) inicialmente. Este método de selección se denomina de **probabilidades iguales**.
 - → todas las bolas tenían el mismo tamaño
- No obstante, puede haber casos en los que no nos interese que la probabilidad de todos los elementos sea la misma, sino que dependa del tamaño o de ciertas características. Por ejemplo, auditando las operaciones de cierta entidad, nos interesará que tengan más probabilidades de aparecer en la muestra aquellas operaciones más grandes, por lo que se podrían escoger probabilidades distintas para cada elemento, proporcionales al tamaño (en unidades monetarias) de la operación (MUS por sus siglas en inglés o MUM). Este método será el de probabilidades desiguales.
 - → las bolas tienen distinto tamaño y las más grandes se escogen más



Las conclusiones derivadas de **muestreos no probabilísticos** deben ser cuidadosas: deberá evitarse informar sobre ratios ya que puede conducir a otros a generalizar o alcanzar conclusiones erróneas.

En cambio, se podrá informar sobre la **cantidad de casos que cumplieron o no con el criterio**. Una vez informada la cantidad de casos, deberá ensayarse una conclusión que permita interpretarlos, es decir, transformar los datos en información.

Sin embargo, **los hallazgos no podrán considerarse representativos**, ¿dan cuenta de la existencia de algún problema serio? ¿Permiten conocer la naturaleza del problema?

A los efectos de fortalecer los resultados del muestreo no probabilístico, es recomendable, en la medida de lo posible, hacer uso de información de otras fuentes (**triangulación**).



Además dentro del muestreo probabilístico existen también diferentes tipos:

- Muestreo aleatorio con y sin reemplazamiento
- Muestreo estratificado
- Muestreo por conglomerados
- Muestreo sistemático
- Otros

https://www.youtube.com/watch?v=f_Hx0pOJEuY



2. Tipos de muestreo Factor de elevación y factor de muestreo

Ejemplo: Supongamos, por ejemplo, que hemos seleccionado una muestra de 28 personas de una población de 560 estudiantes de una escuela para preguntarles si tienen conexión a internet en su casa. Nos planteamos las siguientes preguntas.

- 1. ¿Qué significa que hemos seleccionado 28 de 560?
- 2. ¿Qué proporción de la población se esta seleccionando?
- 3. Como queremos extraer conclusiones sobre la población, ¿a cuántos estudiantes de la población representa cada estudiante de la muestra?

En realidad estamos haciendo una encuesta a 28/560=5% de la población y cada estudiante de la muestra representa a 560/28=20 estudiantes de la población.



2. Tipos de muestreo Factor de elevación y factor de muestreo

En el ejemplo anterior, ¿qué tiene más sentido un muestreo con o sin reemplazamiento?

Aunque estos dos tipos de muestreo son diferentes, cuando el tamaño de la población es infinito o muy grande, ambos muestreos proporcionan resultados similares. En general si n/N es mayor que 10% podría dar lugar a resultados diferentes.

Factor de elevación: N/n. Representa el número de elementos existentes en la población por cada elemento de la muestra.

Factor de muestreo: n/N. Porcentaje de la población representada en la muestra.





a. Precisión y error de muestreo

En la medida en que una muestra es representativa, habrá similitud entre el valor del estimador (de la muestra) y el valor del parámetro (de la población). Es el grado de similitud entre un estimador y un parámetro, siendo el parámetro el valor asociado a un determinado atributo de la población, y el estimador el valor asociado al mismo atributo, pero de una muestra, lo que determina la **precisión** del análisis.

Ejemplo:

Supóngase una población de la cartera de préstamos particulares de una entidad de crédito, a partir de los cuales se podría calcular la probabilidad de incumplimiento, valor que desconocemos y que alcanza 10% (el parámetro).

De extraerse una muestra, podrá calcularse la probabilidad de incumplimiento promedio de las unidades muestreadas, que conformará el estimador, cuyo valor muy probablemente oscilará en torno al valor del parámetro, sin necesariamente coincidir.



a. Precisión y error de muestreo

No obstante, en los casos en que se tome una muestra, **no se contará con la información poblacional** para comparar cuánto se desvía la estimación del parámetro muestral, por lo que será necesario **estimar el error** que se comete al realizar el muestreo.

El concepto inverso al de precisión es, por tanto, el "error de muestreo". Se denomina error de muestreo a cualquier diferencia que exista entre el estimador y el parámetro, debido al azar.

Se pueden tomar mediciones del error absoluto (sujeto a las unidades de medida del indicador que se esté estimando):

$$e = \sigma(\widehat{\theta})$$

Y el error relativo o coeficiente de variación, que es independiente de las unidades.

$$e_r = CV(\hat{\theta}) = \frac{\sigma(\hat{\theta})}{E(\hat{\theta})}$$



a. Precisión y error de muestreo

Los objetivos del muestreo son, por tanto, inferir de una muestra (o subgrupo de elementos extraídos del universo a analizar) conclusiones sobre la población. Es decir:

Nos interesa conocer un dato poblacional \rightarrow parámetro θ

Para ello estimamos valores de una muestra \rightarrow estimador $\hat{\theta}$

Al ser un valor estimado, no tiene por qué coincidir con el valor poblacional, y nos interesará conocer cuánto nos estamos desviando del valor poblacional.

Queremos saber cuánto nos desviamos \rightarrow varianza $V(\hat{\theta})$

Pero esta información provendría de la población por lo que también es desconocida y necesitaremos estimarlo.

Hallamos la estimación de la desviación \rightarrow varianza estimada $\hat{V}(\hat{\theta})$

Cuanto menor sea la desviación, mayor será la precisión del análisis.



a. Precisión y error de muestreo

El **tamaño de la muestra** incide sobre la precisión de los estimadores. Si se desea mayor precisión en los estimadores, más grande deberá ser la muestra.

El tamaño de la muestra incide sobre el error de muestreo de modo inverso a como sucedía con la precisión. Si se desea menor error de muestreo, más grande deberá ser la muestra. El riesgo de una muestra pequeña es que no sea representativa de la población por el hecho de extraer unidades de muestreo particularmente extremas en valor.

Por ejemplo, supóngase que para estimar la altura promedio de la población mencionada precedentemente se extrae una muestra pequeña, pero en virtud del azar, algunas extracciones corresponden a jugadores de baloncesto, sin que se hayan extraído (también por azar) personas con valores de atributo en el otro extremo del continuo y que permitan compensar a los jugadores de baloncesto.



a. Precisión y error de muestreo

El error de muestreo se mide como margen de error admisible.

$$e = \sigma(\hat{\theta}) = \sqrt{V(\hat{\theta})}$$

Ejemplo: se espera que el valor del parámetro esté comprendido en +/- 5% del valor del estimador que se ha calculado o, en otras palabras, que la probabilidad de incumplimiento promedio esté comprendida entre 9.31% y 10.29% (cuando la probabilidad de incumplimiento promedio de la muestra dio por resultado 9.8%). El error de muestreo se suele representar bajo la letra "E".

Sabiendo además que la varianza es dependiente de n, una vez conocida la forma de la varianza se podrá estimar n en función del error prefijado.



b. Confianza

El **nivel de confianza** es una medida de la probabilidad de que el valor del estimador no difiera del valor del parámetro más allá del margen de error.

La confianza reside sobre el grado de representatividad de la muestra sobre la población. En términos coloquiales, podría decirse que es una manera de estipular **como de parecidas son las muestras que podrían extraerse**, una y otra vez, de la misma población. Así, dado un nivel de confianza del 95%, estamos reflejando la idea de que de cada 100 veces que escojamos una muestra de la población, 95 de esas muestras serán representativas.

Si se desea mayor grado de confianza sobre los estimadores, más grande deberá ser la muestra. De este modo, cuánto más grande sea la muestra, existe menor probabilidad de diferir significativamente entre ellas.

En el extremo, si la muestra fuera tan grande como la población, entonces no habría diferencias entre muestras extraídas una y otra vez, por lo que el nivel de confianza en el muestreo sería máximo.



b. Confianza

Unido al concepto de nivel de confianza está el **nivel de significación** (α) de tal forma que el nivel de confianza se define como 1- α . El nivel de significación es la probabilidad de escoger una muestra que no sea representativa de la población. Por tanto, este concepto es la diferencia entre la certeza (1) y el nivel de confianza (1 – α).

Además, es necesario conocer el concepto de "valor crítico" designado generalmente como $z_{\alpha/2}$, asociado a las distribuciones y al nivel de confianza. Sería el valor que deja a la derecha de la distribución un área igual a $\alpha/2$ (o lo que es lo mismo, acumula a su izquierda 1- $\alpha/2$).

$$P\left(Z > z\alpha_{/2}\right) = \alpha_{/2}$$



b. Confianza

A través de estos conceptos podemos dar forma al intervalo de confianza¹, que se define como los valores entre los que se encuentra el verdadero valor poblacional con un grado de confianza igual a 1- α .

Para el caso bajo una distribución normal, por ejemplo, tendríamos:

$$[\hat{\theta} - z\alpha_{/2}\sigma(\hat{\theta}), \hat{\theta} + z\alpha_{/2}\sigma(\hat{\theta})]$$

Normalmente se trabaja con niveles de confianza comprendidos entre el 90% y el 95%. Un nivel de confianza del 95% significa, por ejemplo, que se tiene un 95% de confianza en que el valor del estimador esté dentro del intervalo de confianza definido.

Siguiendo con el ejemplo anterior, podría decirse que con un 95% de confianza, la probabilidad de incumplimiento promedio está comprendida entre 9.31% y 10.29% (cuando la probabilidad de incumplimiento promedio de la muestra dio por resultado 9.8%).

¹ Los intervalos de confianza se explican en el apartado correspondiente a distribuciones de probabilidad.



c. Error no atribuible al muestreo

Hay otras fuentes de error, además del error de muestreo, que pueden derivar en estimadores imprecisos.

Cualquier diferencia que exista entre el estimador y el parámetro, y que no se deba al azar ni al tamaño de muestra, si no a fallos en el programa de muestreo, o a su aplicación en campo, se denomina "error no atribuible al muestreo".

A diferencia del error de muestreo, este tipo de error **no es mensurable** y hasta puede pasar desapercibido, conduciendo irremediablemente a conclusiones incorrectas.

El origen del error no atribuible al muestreo puede clasificarse como: i) problemas de diseño del programa, ii) equívocos, iii) falta de respuesta. En el caso de encuestas, por ejemplo, cabe mencionar los siguientes: preguntas ambiguas, anidadas o correlativas, respuestas por reputación, por conveniencia o falta de respuesta, procedimientos ambiguos, ponderación incorrecta, cobertura inapropiada y falsificación o equívocos.



d. Tamaño de la muestra

El tamaño de una muestra esta determinado en gran medida por tres factores:

- Prevalencia estimada de la variable considerada
- Nivel de fiabilidad
- Margen de error aceptable
- 1). Cuando la varianza (S²) es conocida:
 - a) Para poblaciones infinitas o tamaños de población desconocida

$$n = \frac{Z^2 S^2}{d^2}$$

b) Para poblaciones finitas o conocidas

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



d. Tamaño de la muestra

- 2). Cuando la varianza (S²) es desconocida:
 - a) Tamaño de la población N es desconocida

a) Tamaño de la población N es conocida

$$n = \frac{Z^2 PQ}{d^2}$$

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



d. Tamaño de la muestra

Donde:

n: Es el tamaño de la muestra.

N: Tamaño de la población.

Z: Factor de confiabilidad. Es 1.96 cuando es un 95% de confianza y es 2.57 cuando se establece un 99% de confianza (valor de distribución normal estandarizada correspondiente al nivel de confianza escogida).

$$P = 0.5$$

$$Q = 1-P = 0.5$$

d: Es el margen de error permisible. Establecido por el investigador.



d. Tamaño de la muestra

Ejemplo: Estimar el tamaño de la muestra de estudio para una población de 120 préstamos.

$$n = \frac{120(1.96)^2(0.5)(0.5)}{(0.1)^2(120-1) + (1.96)^2(0.5)(0.5)} \longrightarrow n = \frac{115.248}{2.1504} = 53.59$$

$$n \approx 53$$



d. Tamaño de la muestra. Factores que influyen

Las limitaciones presupuestarias o temporales hacen que en ciertas oportunidades resulte necesario revisar el alcance del trabajo, o iterar con los niveles de confianza y error de muestreo hasta adecuar el tamaño de la muestra a las posibilidades reales.

En el extremo, será necesario abandonar las pretensiones de llevar a cabo un muestreo probabilístico, y volcar los esfuerzos a la elección del mejor método de muestreo no probabilístico. Pero además del *error de muestreo* y de la *confianza*, en rigor, las únicas dos variables endógenas, existen otros factores que influyen sobre el tamaño de la muestra:

1) La dispersión. Cuanta mayor variabilidad en los valores de un atributo, menos preciso será el estimador, por lo que se requiere de tamaños de muestra más grandes frente a un mismo error de muestreo asumido.



d. Tamaño de la muestra. Factores que influyen

- 2) La proporción de un atributo. Incide en virtud de su frecuencia relativa y de tal forma que cuanto más cercana al 50%, más grande deberá ser el tamaño de la muestra. Si bien la proporción puede ser estimada a grandes rasgos a los efectos de introducir el dato en la fórmula, en caso de imposibilidad para hacer este cálculo preliminar, debe recurrirse, justamente, a calcular el tamaño de muestra bajo un criterio conservador, es decir, asumiendo que la proporción es del 50%, de modo de obtener el máximo tamaño de muestra posible. Este concepto es equivalente al de "error esperado", que se encuentra en algunas fórmulas para la estimación del tamaño de muestra.
- **3)** El **tamaño de la población.** A medida que la población aumenta de tamaño, en especial cuando N>1000, su tamaño incide cada vez menos sobre el tamaño de la muestra, obteniéndose grandes economías. A grandes trazos, el análisis de una parte cada vez menor de una población cada vez más grande, es suficiente para concluir sobre la población.



e. Fases de la investigación por muestreo

- Delimitar la población objetivo y la población investigada. Establecer un marco.
- Diseñar la muestra.
- Extraerla: trabajos de campo (qué medimos, a quién y cómo desarrollo de los programas informáticos).
- Procesamiento de los datos.
- Evaluación y conclusiones.
- Difusión.



© 2021 Afi Escuela de Finanzas. Todos los derechos reservados.