ANALISIS DE DATOS

Dr. Edgar Acuna

http://academic.uprm.edu/eacuna

DEPARTAMENTO DE CIENCIAS MATEMATICAS
UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGUEZ

Agosto, 2019

Porque estudiar Estadistica?

1. Hay datos por todas partes.
2. No importa cual es tu area de especialidad, frecuentente tienes que tomar decisiones que envuelven datos. Por lo tanto, aprender metodos estadisticos para analizar datos nos ayuda a una toma de decisiones mas efectiva.
3. Trabajar con una muestra pequena puede ser mas eficiente que trabajar con todos los datos (Big data).

INTRODUCCIÓN

En este capítulo, primero se introducirán algunos conceptos estadísticos básicos, luego se dará una definición y división de la estadística. Finalmente se hará una clasificación de los distintos tipos de datos que aparecen en un estudio estadístico y de que forma pueden ser recolectados.

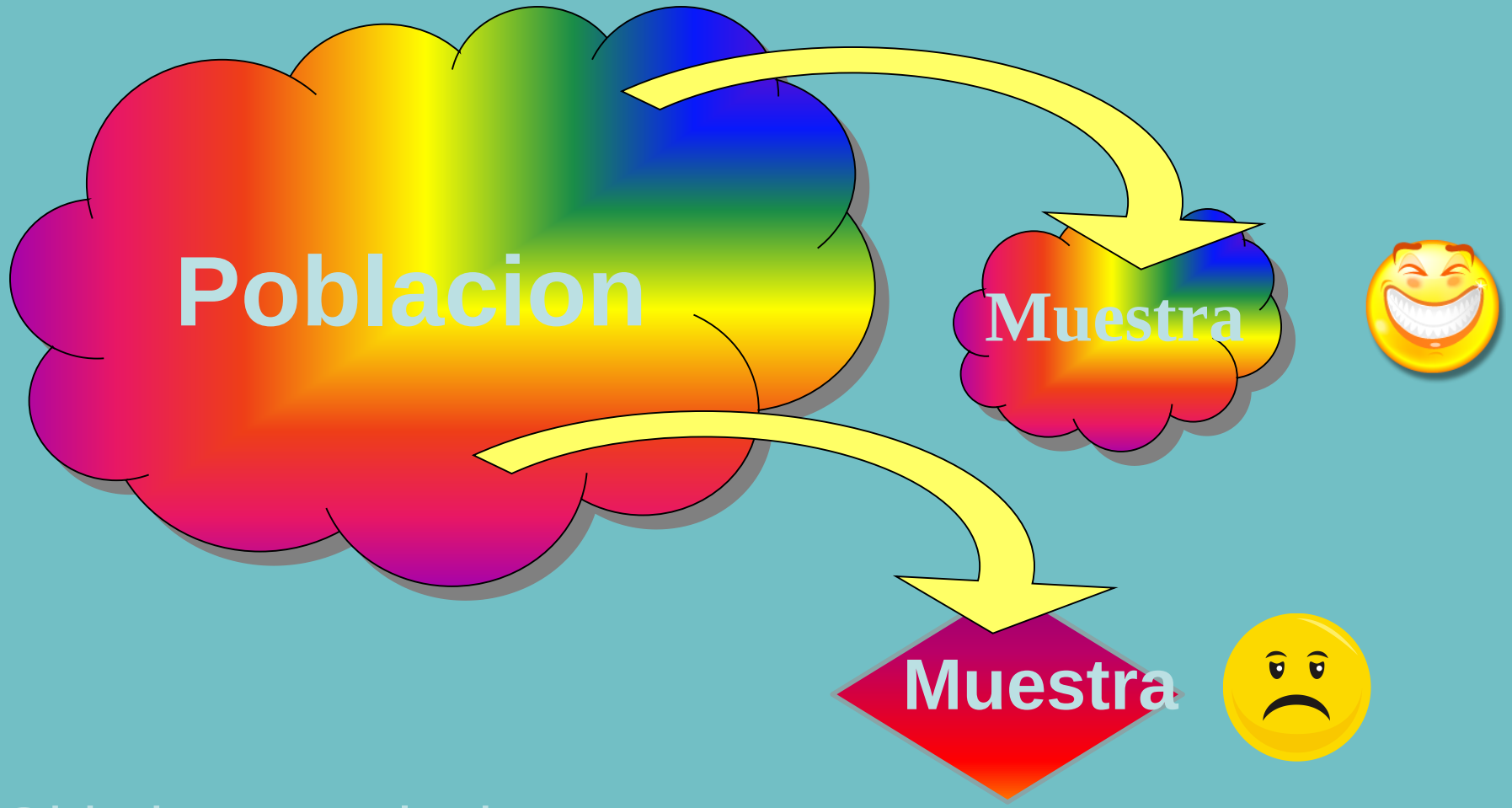
1.1 Conceptos Estadísticos Básicos

a) **Población:** una población es un conjunto de individuos u objetos que poseen la característica que se desea estudiar. En un sentido más estadístico, una población es el conjunto de mediciones de una cierta característica en todos los individuos u objetos que poseen dicha característica

b) **Muestra:** Es el conjunto de mediciones que han sido realmente recolectados. La extracción de la muestra es un paso bien importante porque es a partir de ella que se sacan conclusiones acerca de la población. Si el diseño es sencillo la muestra tiene que ser relativamente grande, alrededor de un 10% del tamaño de la población.

c) **Muestra Aleatoria:** Es una muestra bien representativa de la población. Se considera que cada elemento de la población ha tenido la misma oportunidad de formar parte de la muestra. Las conclusiones basadas en una muestra aleatoria son confiables.

Una muestra sin/con sesgo



Objetivo es selecionar una muestra que sea representativa de la poblacion

El sesgo del muestreo

Sesgo de muestreo ocurre cuando el metodo de seleccionar una muestra causa que la muestra no sea bien representativa de la poblacion

- Si hay sesgo de muestreo, entonces las conclusiones extraidas acerca de la poblacion basadas en la muestra tomada no son confiables.

El Poder del Muestreo Aleatorio

- Antes de la eleccion del 2008, la encuestadora Gallup tomo una ***muestra aleatoria*** de 2,847 votantes (total de votantes es ~150 millones). 52% de los entrevistados apoyaban a Obama.
- En la eleccion verdadera, 53% votaron por Obama.
- En la eleccion del 2016,Gallup daba una ventaja del 5% de Clinton sobre Trump.

Caso Trump:Porque Fallaron las encuestas ?

18 de 20 encuestas vaticinaban la victoria de Hillary Clinton sobre Trump. Algunas le daban un margen de victoria del 15% al 30%. Solo la encuesta de USC/LA Times y IBD/TIPP hicieron la prediccion correcta.

1-Los modelos pierden efectividad conforme pasa el tiempo.
2-No hay que seguir siempre lo que piensa la mayoria.
3-Hay que medir el comportamiento irracional de los entrevistados (las encuestas politicas son mayormente por telefono).
4-Buscar informacion adicional mas alla de las encuestas.
5-Tomar muestras mas grandes
6-El factor de las redes sociales (Analisis de sentimientos en Twitter)

1.1 Conceptos Estadísticos Básicos

d) Variable: Es la característica que se desea estudiar.

e) Dato: Es un valor particular de la variable.

f) Parámetro: Es un valor que caracteriza a una población. El valor del parámetro es constante y por lo general es desconocido.

g) Estadístico: Es un valor que se calcula en base a los datos que se toman en la muestra y el cual es usado para estimar el valor del parámetro. El valor del estadístico es conocido y varía con la muestra tomada.

1.1 Conceptos Estadísticos Básicos

h) Censo: Es un listado de una o más características de todos los elementos de una población. Los censos poblacionales se hacen cada 10 años a nivel mundial.

i) Encuesta: Es un listado de una o más características de todos los elementos de una muestra.

1.2 Definición de la Estadística

La Estadística es la ciencia donde se aprende acerca de la población a partir de la información recolectada de una muestra extraída de ella. La Estadística comprende los métodos usados para recolectar la muestra, la organización y presentación de los datos recolectados y la extracción de conclusiones mediante la aplicación de técnicas adecuadas a los datos de la muestra.

1.3 División de la Estadística

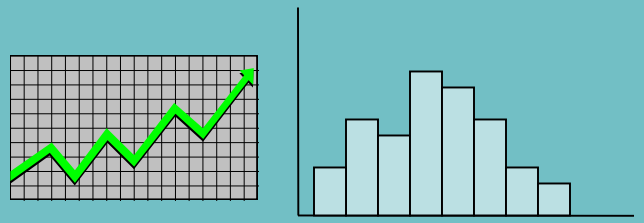
- **Estadistica Descriptiva:** Conjunto de técnicas y métodos que son usados para recolectar, organizar, y presentar en forma de tablas y gráficas información numérica. También se incluyen aquí el cálculo de medidas estadísticas de centralidad y de variabilidad.

- **Estadistica Inferencial:** Conjunto de técnicas y métodos que son usados para sacar conclusiones generales acerca de una población usando datos de una muestra tomada de ella.

Estadistica Descriptiva

- Recolectar datos

  - Ej. Encuestas
- Presentar datos



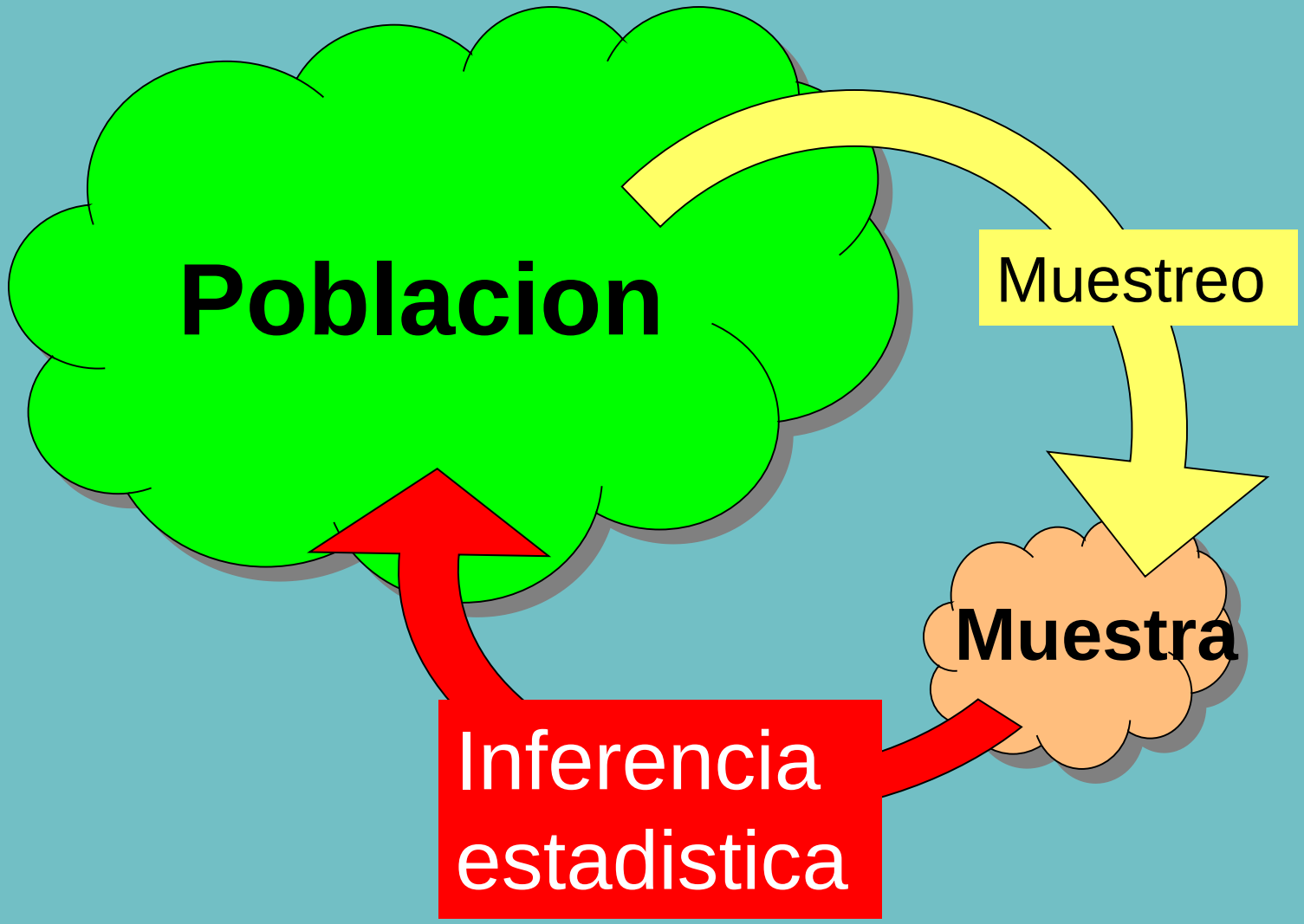
  - Ej.Tablas y graficas
- Resumir datos
  - Ej. Media muestral $= \frac{\sum X_i}{n}$

Estadistica Inferencial

- Estimacion
  - Ej. Estimar el salario promedio poblacional using el salario promedio de una muestra.
- Prueba de hipotesis
  - Ej. Probar la afirmacion de que el salario promedio poblacional es mas de 20 mil dolares.

Inferencia es el proceso of extraer conclusiones o tomar decisiones acerca de una poblacion basados en los resultado de una muestra.

Como funciona la estadistica



1.4 Tipos de Datos

A. Datos Cuantitativos. Son aquellos que resultan de hacer mediciones o conteos. Se clasifican a su vez en dos subtipos:

A1. Datos Discretos. Son los que resultan de hacer conteos y por lo general son números enteros.

A2. Datos Continuos. Son los que resultan de hacer mediciones y pueden asumir cualquier valor de la recta real.

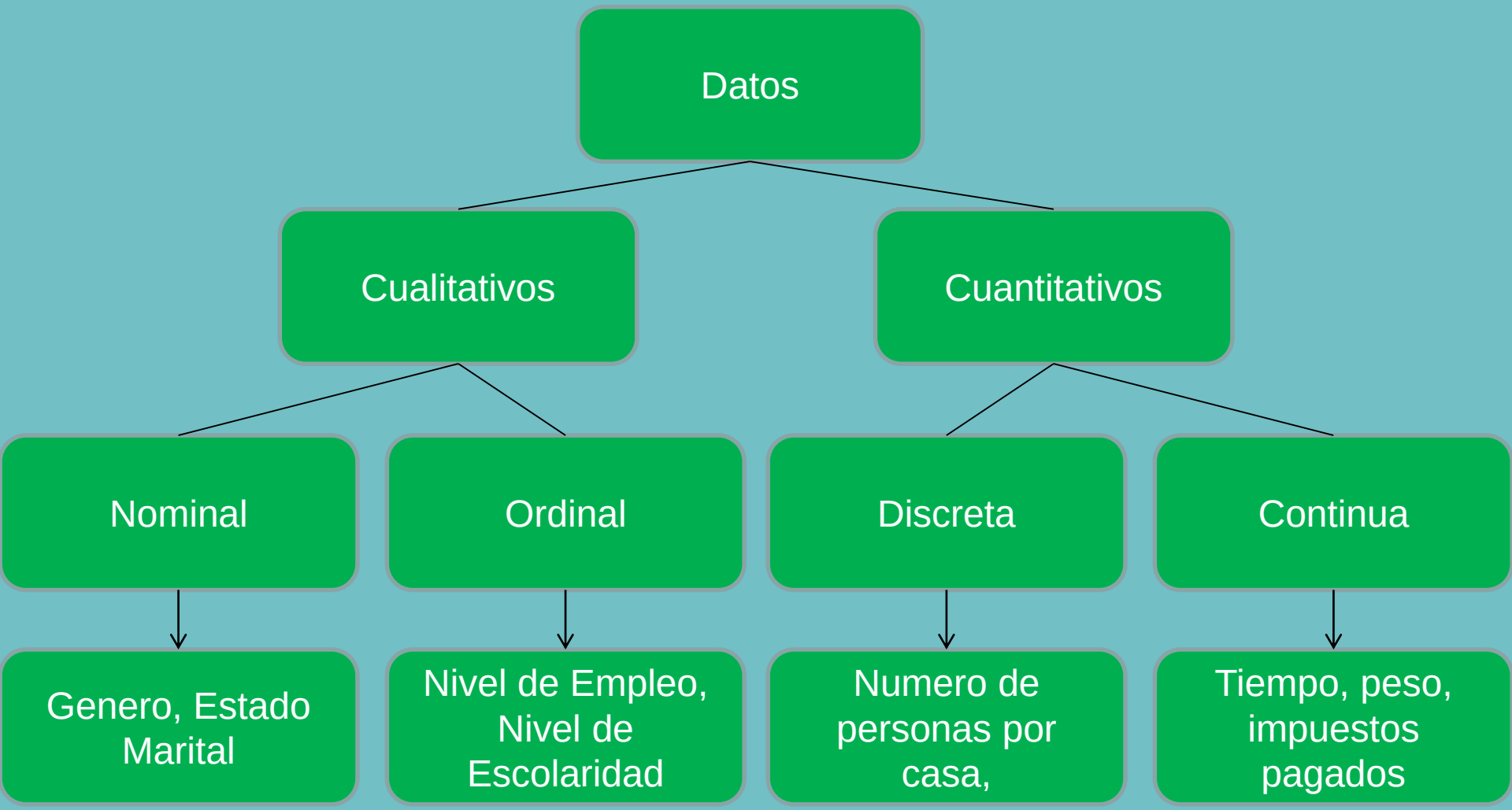
1.4 Tipos de Datos

B. Datos Cualitativos o Categóricos. Son aquellos que expresan atributos o categorías. Para facilitar el análisis estadístico de este tipo de datos frecuentemente se codifican a números, esta codificación da lugar a dos subtipos de datos categóricos:

B1. Datos Nominales. Son aquellos datos categóricos que pueden ser codificados numéricamente pero donde hay una relación arbitraria entre los números asignados y el valor de la variable.

B2. Datos Ordinales. Son aquellos que al ser codificados numéricamente deben guardar una correspondencia entre los números asignados y el verdadero valor de la variable.

Tipos de datos



1.5 Técnicas de Muestreo

a) **Muestreo Aleatorio**. Se usa cuando a cada elemento de la población se le quiere dar la misma oportunidad de ser elegido en la muestra.

b) **Muestreo Estratificado**. Se usa cuando se conoce de antemano que la población está dividida en estratos, que son equivalentes a categorías y los cuales por lo general no son de igual tamaño. Luego, de cada estrato se saca una muestra aleatoria, usualmente proporcional al tamaño del estrato.

c) **Muestreo por conglomerados ("Clusters").** En este caso la población se divide en grupos llamados conglomerados. Luego se elige al azar un cierto número de ellos y todos los elementos de los conglomerados elegidos forman la muestra.

c) **Muestreo Sistemático.** Se usa cuando los datos de la población están ordenados en forma numérica. La primera observación es elegida al azar de entre los primeros elementos de la población y las siguientes observaciones son elegidas guardando la misma distancia entre si.

1.6 Maneras de Recolectar Datos

a) Haciendo entrevistas personales. Puede ser el método más efectivo en muchas ocasiones pero es costoso y requiere bastante tiempo para ser ejecutado.

b) Haciendo entrevistas por teléfono. Tiene la desventaja de que el entrevistado puede no ser sincero en sus contestaciones.

c) Mediante cuestionarios emitidos por correo. Es costoso y por lo general no más del 30% de los entrevistados retornan el cuestionario.

d) Por observación directa.

e) A través de la Internet.

f) Usando simulación por computadoras.