UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

FACULTAD DE ECONOMÍA

MAESTRÍA EN ESTUDIOS SUSTENTABLES REGIONALES Y METROPOLITANOS

APUNTES DE LA

UNIDAD DE APRENDIZAJE: ESTADÍSTICA APLICADA I

Elaborado por: Ricardo Rodríguez Marcial

Septiembre de 2018

CONTENIDO

INTRODUCCIÓN

El objetivo del presente documento es presentar las técnicas de la estadística, tanto descriptiva como inferencial; su aplicación al ámbito regional y metropolitano, con la finalidad de extraer las señales de la información y poder construir mejores descripciones de los fenómenos bajo estudio.

Para los ejercicios que se presentan se utilizan los software Excel, Eviews y QSIG con lo cual se cubre la totalidad de los temas tratados.

Debe señalarse que el contenido del documento va dirigido a los estudiantes que cursan la Unidad de Aprendizaje Estadística Aplicad I, de la Maestría en Estudios Sustentables Regionales y Metropolitanos, que actualmente ofrecen la Facultad de Economía, Arquitectura y el Instituto de Investigaciones sobre la Universidad de la U.A.E.M.

En la Maestría en Estudios Sustentables Regionales y Metropolitanos que se ofrece en el espacio académico; la estadística aplicada, sirve para extraer conclusiones acerca del fenómeno que se estudia y, con ello, estar en posibilidades de comprenderlo y predecirlo, cobra relevancia, toda vez que esta maestría está enfocada a resolver problemas concretos que aquejan en el ámbito regional y metropolitano. Permitiendo, diseñar políticas encaminadas a elevar el bienestar de la población.

PRSENTACIÓN

El Programa de Estudios denominado Estadística Aplicada I es parte trascendente del Programa de Maestría en Estudios Sustentables Regionales y Metropolitanos que se imparte en la Universidad Autónoma del Estado de México.

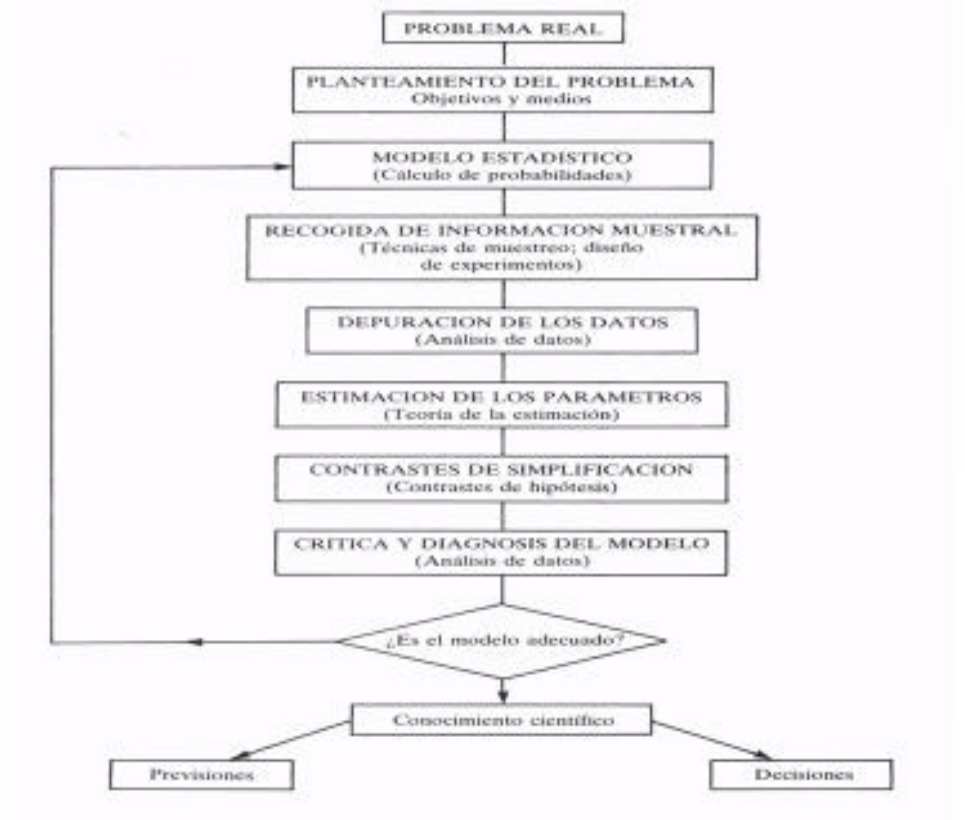
La importancia del estudio de la estadística descriptiva e inferencial, radica en que los estudiantes tienen la misión de detectar una problemática donde se vea involucrada la sustentabilidad en los entornos regional y metropolitano, conocerla con la mayor precisión posible, cuantificándola para determinar los elementos que la ocasionan y estar en posibilidad de recomendar las medidas de política acordes a los objetivos propuestos. Sin embargo, los datos tratados directamente, pueden ser de difícil interpretación e inducir con ello a decisiones equivocadas. Se hace necesario algún tipo de tratamiento o análisis de los mismos que nos ayuden a extraer conclusiones más fiables.

Se debe señalar que estos apuntes están estructurados de acuerdo al contenido de la Unidad de Aprendizaje Estadística Aplicada I, de ahí que en el índice se encuentre todos los elementos y en el mismo orden que se presentan en dicha Unidad.

MÓDULO I. NIVEL CONCEPTUAL Y TEÓRICO DE LA ESTADÍSTICA DESCRIPTIVA E INFERENCIAL

La estadística, es una rama de las matemáticas que sirve como herramienta para estudiar y analizar datos numerosos en diversos campos de la ciencia; en el ámbito profesional, permite el tratamiento de la información y la obtención de conclusiones para una correcta toma de decisiones. El objetivo del primer módulo es proporcionar los conceptos básicos suficientes para que el estudiante, vaya *entendiendo el idioma* de la estadística, y con ello, pueda hacer uso correcto de esta poderosa herramienta para darle la aplicación que más desee en su estudio.

El método de la estadística, ocupa los dos tipos de razonamiento que distinguen al método científico: el razonamiento deductivo que consiste en partir de lo general e ir a lo particular del fenómeno de estudio y, el razonamiento inductivo, que recorre el sentido inverso, de lo particular a lo general. En la siguiente ilustración se aprecia de manera nítida lo señalado anteriormente.



Uno de los conceptos más elementales del lenguaje estadístico, es el **muestreo** (Mendenhall, et al., 2010), puesto que, en casi todos los casos estadísticos, solamente se trabaja con un número limitado de datos, esto es, una muestra, mismo que se toma a partir de un número mayor de datos, la población. Entenderemos por muestreo a la operación de

seleccionar o elegir qué elementos de la población van a constituir la muestra en la que van a estudiarse uno o varios caracteres.

Ilustración 1: Muestra y Población.

Fuente (Mendenhall, et al., 2010)

Ahora bien, ¿qué es más importante estudiar para la estadística? Dados los conceptos de población y muestra, pareciese que a la estadística le importara conocer el conjunto total de los datos, es decir, de la población, sin embargo, en la práctica resulta difícil e inclusive imposible ser capaces de recolectar toda la información de todos los elementos de la población, por lo que en estadística **tratamos de describir o pronosticar el comportamiento de la población con base en información obtenida de una muestra representativa de esa población** (Mendenhall, et al., 2010)**.**

Distinguiremos dos tipos de muestreo, a saber: Probabilístico (aleatorio) y No Probabilístico.

El primero consiste en determinar por azar a cada uno de los individuos de la muestra, existiendo diversos métodos para ello: 1) muestreo aleatorio simple garantiza, en términos teóricos, que cada uno de los individuos de la población tenga la misma oportunidad de aparecer en la muestra, selección al azar; para realizar esto se trabaja con tablas de números aleatorios, computadoras o calculadoras, el requisito para usar este método, es contar con una lista de cada uno delos individuos de la población a investigar; 2) muestreo aleatorio estratificado, es semejante al caso anterior, con la diferencia de que la población se divide en estratos; por ejemplo, si se tiene la población estudiantil de una Facultad, puede dividirse entre el estrato hombres y el estrato mujeres; 3) muestreo sistemático, una vez que se determinó el número de elementos que integran la muestra, la selección de los individuos que formarán parte de la misma, se seleccionan sistemáticamente; por ejemplo,

se van a elegir estudiantes en un aula que tiene 40 alumnos y, decidimos que formarán parte de la muestra uno de cada cinco estudiantes contados por filas.

Respecto al muestreo no probabilístico, distinguiremos dos maneras de hacerlo: por cuotas y por accidente.

Otro elemento importante dentro del tema del muestreo es la determinación del tamaño de la muestra, para ello existen distintas fórmulas, cuyo su uso depende de la información con que contemos; sin embargo, de manera general, el cálculo del tamaño de la muestra está explicado por: a) el porcentaje de confianza con que se desea generalizar los datos a la población, cuanto mayor sea el porcentaje de confianza que se desea, mayor será la cantidad de sujetos necesarios para la muestra; b) el porcentaje de error, se refiere a la elección de la probabilidad de aceptar una hipótesis siendo falsa o a la inversa; y, c) nivel de variabilidad para comprobar la hipótesis, se refiere a especificar la probabilidad de que el evento suceda o no.

Si contamos con las proporciones de acierto y no acierto, la expresión es:

Proporciones

$$n = \frac{Z^2.p.q}{e^2}$$

Donde:

n: tamaño de la muestra

Z: nivel de confianza

p: variabilidad positiva

q: variabilidad negativa

e: error

Si conocemos el tamaño de la población:

$$n = \frac{Z^2 p q N}{NE^2 + Z^2 p q}$$

Donde:

N: tamaño de la población

n: tamaño de la muestra

Z: nivel de confianza

p: variabilidad positiva

q: variabilidad negativa

E: error

Si conocemos la desviación estándar de la población:

$$n = \frac{Z^2{}_{\alpha/2} S^2}{\varepsilon^2}$$

Donde:

n: tamaño de la muestra

Z: nivel de confianza

S: desviación estándar de la población

ε: error

Si conocemos el tamaño y la desviación estándar de la población:

$$n = \frac{S^2}{\frac{\varepsilon^2}{Z^2} + \frac{S^2}{N}}$$

Donde:

N: tamaño de la población

n: tamaño de la muestra

Z: nivel de confianza

S: desviación estándar de la población

ε: error

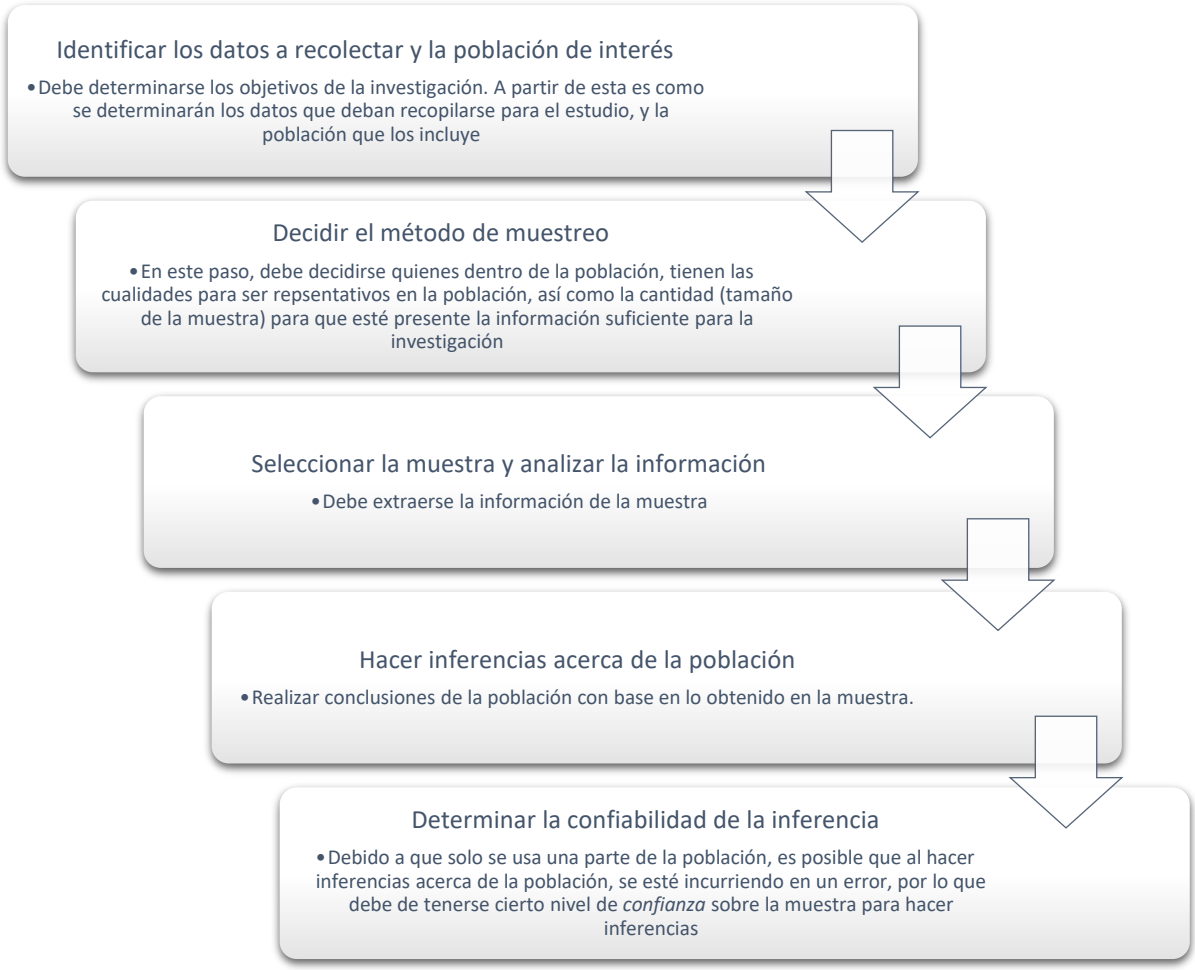
Estadística descriptiva e inferencial

Dentro de la estadística existe una serie de *subdivisiones*. Al analizar una serie de datos (ya sea que provengan de una muestra o de una población), deben existir técnicas para organizarlo o resumirlo, a dicho conjunto de técnicas se denomina **estadística descriptiva**. La estadística descriptiva se materializa en gráficas de barras, de dispersión, pastel, histogramas, entre otras. Por otra parte, como se mencionó anteriormente, conseguir todos los datos de la población puede resultar imposible, ya sea por tiempo, costo, distribución, etc., es por ello, que nace la **estadística inferencial**, que siendo una rama de la estadística busca explicar a la población basados únicamente en datos obtenidos de una muestra. En

otras palabras, se busca hacer *inferencias*[1] acerca de las características poblacionales con base en las características de una muestra.

El procedimiento a seguir para hacer uso de la estadística inferencial, puede resumirse en los siguientes pasos:

Ilustración 2: Procedimiento Estadística Inferencial.



Fuente: elaboración propia con datos de (Mendenhall, et al., 2010)

[1] Hacer inferencias significa tener la capacidad de predecir, obtener conclusiones y tomar decisiones (Mendenhall, et al., 2010).

Variables y Datos

Ya se analizó anteriormente, a grandes rasgos, las utilidades tanto de la estadística descriptiva como de la inferencial, sin embargo, se ha hablado mucho respecto a los *datos* sin llegar a definirlos, este concepto sin duda, forma parte fundamental del lenguaje estadístico. Algunos de los conceptos importantes para el análisis estadístico son los siguientes:

- Variable: característica que cambia con el tiempo o con diferentes situaciones;
- Unidad experimental: individuo que provee las mediciones de las variables, al conjunto total de unidades experimentales se le conoce como población, mientras que a un subconjunto de dichas unidades experimentales se le llama muestra.
- Datos univariados: se mide una sola variable por cada unidad experimental;
- Datos multivariados: se miden dos o más variables por cada unidad experimental.

Ejemplo: Si se desease analizar el bienestar económico en Toluca, Estado de México, una de las variables sería el ingreso mensual, debido a que este cambia dependiendo de la persona a la que se le pregunte y en el tiempo en el que se le haga la pregunta; la unidad experimental será cada una de las personas que conforman la fuerza laboral toluqueña, mientras que la población es el **total** de los trabajadores de Toluca. En este caso, el resultado serán datos univariados, pues únicamente se investiga respecto al ingreso económico mensual; sin embargo, si se estuviese interesado en el número de servicios públicos con que cuenta, los bienes que posee, entre otras, estaríamos presente a un estudio de datos multivariados.

Adentrándose más en los datos, debe diferenciarse entre los siguientes tipos:

Ilustración 3: Tipos de datos.

Fuente (Mendenhall, et al., 2010)

Existen dos tipos de datos, esto es, que nuestras variables pueden clasificarse entre cualitativas o cuantitativas. La primera, como su nombre lo indica, mide características de las unidades experimentales, mientras que la segunda se limita a medir numéricamente las unidades experimentales.

Cuadro 1.- Ejemplos de variables cualitativas y cuantitativas

Cualitativas	Cuantitativas
• Grado de marginación: Bajo, Medio, Alto	• Promedio del semestre anterior; • Percepción económica mensual;

• Nivel del IDH: Bajo, Medio, Alto • Materia favorita: Matemáticas, español, Ciencias naturales, etc. • Música favorita: Rock, Pop, Rap, etc.	• Tasa de interés del mercado; • Número de habitantes en pobreza extrema en un país; • Tipo de cambio actual;

Nótese que, dentro de las variables cuantitativas, aún existe una mayor separación, esto es, las variables cuantitativas discretas y continuas. Las variables discretas se caracterizan por solo integrar números finitos y contables, es decir $x = 0,1,2,3 \ldots . n$, por ejemplo, en el caso de *número de habitantes en pobreza extrema de un país* únicamente se puede describir con un número entero, puesto que es erróneo decir que existen 12.4 habitantes en pobreza. En cambio, una variable cuantitativa continua, si puede tomar todos los valores posibles, tal es el caso de la percepción económica mensual, o el tipo de cambio, donde se pueden incluir decimales (los que sean necesarios) para explicar la variable (es decir, comentar que el tipo de cambio se encuentra en 17.69 si es correcto).

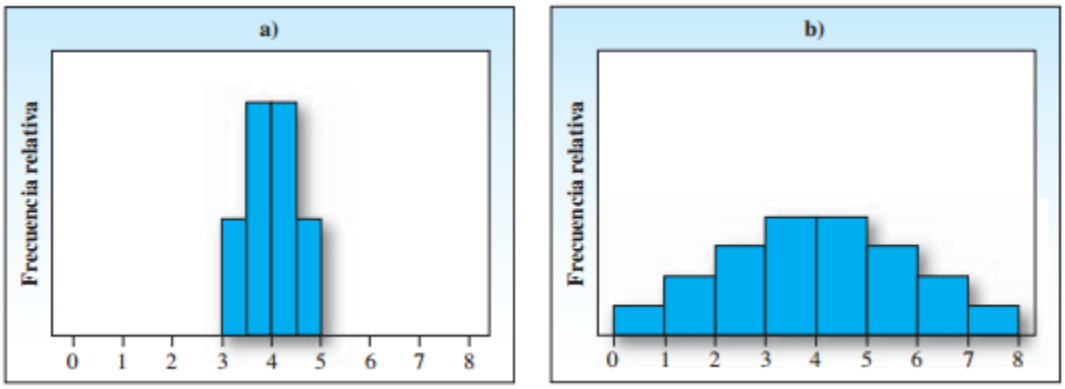
Descripción de la estadística

En un principio, alguien pudiese pensar que la única aplicación de la estadística es la elaboración de mapas y gráficos que expliquen la realidad, si bien es cierto que los gráficos son un apoyo para el investigador, también es cierto que las medidas numéricas (de centro, posición relativa y variabilidad) permiten estructurar y describir el comportamiento de alguna variable. Cabe recordar que, las mediciones asociadas a la población se denominan parámetros, mientras que las asociadas con la muestra llevan el nombre de estadísticas.

- Medidas de centro: son aquellas que ayudan al investigador a obtener información respecto a la distribución de los datos, en estos se encuentra la media, mediana y moda:
    - Media aritmética: representada con $\bar{x} = \frac{\sum x_i}{n}$ (para el caso muestral) y con $\mu = \frac{\sum x_i}{N}$ (para el caso poblacional). El resultado, implica el punto de equilibrio de la distribución, además, la media muestral $\bar{x}$ es el mejor aproximado para la media poblacional μ;
    - Mediana: al ser ordenadas las mediciones de menor a mayor, el dato que quede justo en medio de la recta, será el denominado mediana m;
    - Moda: se refiere al dato que ha tenido mayor coincidencia en el estudio, esto es, el dato más repetido de la muestra o población (según sea el caso).
- Medidas de variabilidad: si bien la media aritmética representa el punto de equilibrio de la distribución, es cierto que no todos los datos caen en la misma media, por lo que los datos (sea de la muestra o de la población) cuentan con un cierto grado de variabilidad o dispersión, por lo que esta medida debe ser contenida en el estudio de estadística:

- Rango: dado un conjunto de mediciones, es la diferencia entre el dato más grande y el más pequeño;
- Varianza: la varianza se estructura de la siguiente forma: $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$ (para el caso de la población) y $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$ (para el caso de la muestra). Así, la varianza se define como el promedio de los cuadrados de las desviaciones alrededor de su media. En otras palabras, mientras mayor sea la varianza (en función del tamaño de los datos) mayor es la dispersión de los datos respecto a la media. Si se observa la ilustración 4, se puede apreciar que en el gráfico a, los datos están muy cercanos a la media, mientras que, en el b, existe más datos que se encuentran alejados de la media. En este sentido, es de esperar que la varianza del gráfico b sea superior al de a.
- Desviación estándar: corresponde a la raíz cuadrada de la varianza, y se define como la variabilidad de las medidas originales de medición respecto a su media.

Ilustración 4: La varianza entre dos distribuciones.

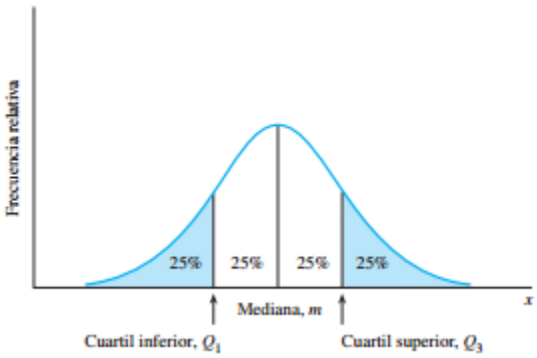


Fuente: (Mendenhall, et al., 2010).

- Mediciones de posición relativa: A veces es necesario encontrar en los datos, incentivos para realizar comparaciones entre muestras o poblaciones con un diverso número de datos. Pare ello se presentan estas medidas de posición relativa:

- Puntaje z muestral: es una medida de posición relativa medida de la siguiente manera. $puntaje\ z = \frac{x-\bar{x}}{s}$, el resultado se lee como unidades de desviación estándar, esto es, que si el resultado fuese 1.7, el dato en cuestión es superior 1.7 desviaciones estándar a la media. ¿De qué sirve lo anterior? Este indicador permite corroborar la naturaleza de los datos, puesto que *a manera de recomendación* si existiese un dato con un $puntaje\ z$ igual o mayor a 3, habrá que prestarle atención pues, si bien el dato puede ser sólo un caso atípico, pudiera también que el investigador esté incurriendo en un error de captura o de muestreo;
- Percentil: siendo un conjunto de n mediciones de la variable x, mismos que están acomodados de manera ascendente. El p-ésimo percentil es el valor donde x es mayor que el porcentaje p del resto de las mediciones, por lo que es menor que el restante $(1-p)$. ¿Cómo interpretar eso? Esto indica que, de ubicarse en el 75avo percentil (esto es, el tercer cuartil o cuartil superior), el 75% de las muestras son inferiores al valor ubicado, mientras que lógicamente, el 25% restante es superior. El caso de la mediana, como se estudió anteriormente, supone que el 50% es inferior a ella y el 50% es superior, por lo que la mediana queda exactamente en el centro. Para encontrar el valor de los cuartiles, se ubica la siguiente formula: $Q_1 = 0.25(n+1), Q_3 = 0.75(n+1)$ para el cuartil inferior y superior respectivamente.

Ilustración 5: Distribución y percentiles.



Fuente: (Mendenhall, et al., 2010)

Análisis y descripción de datos bivariados

Como se mencionó anteriormente, existen los datos univariados y los multivariados (donde se analiza una variable y dos o más variables respectivamente). Dentro de los datos multivariados, se encuentran aquellos donde se relacionan dos variables. Un ejemplo podría ser el siguiente: al analizar el desarrollo económico de la ciudad de Toluca, se pueden recopilar los datos de las variables: i) años de instrucción de las personas de la población de Toluca, Estado de México, y ii) ingreso percibido mensual. A partir de la recolección de ambas variables, puede realizarse una relación entre ambas para contestar las siguientes preguntas: ¿Los años de instrucción de las personas toluqueñas explican el ingreso percibido? ¿Existe una relación entre estas dos variables? ¿Es positiva/negativa la relación?

Se puede comenzar con la realización de una comparación entre las variables cuantitativas, en nuestro ejemplo, permitirá establecer en qué grado de estudios (primaria, secundaria, medio superior, superior) se obtienen los mejores ingresos. Para realizarlo, es posible el uso de gráficas que expliquen de manera sencilla la relación que existe entre el grado de estudios y los ingresos obtenidos, (una gráfica de barras o de pastel son buenos instrumentos para describir este tipo de eventos). Sin embargo, puede presentarse la situación donde las gráficas sean muy similares o sencillamente sea difícil vislumbrar la relación; ante esta situación, se pueden utilizar otros instrumentos de la estadística descriptiva como el uso de las proporciones.

$$Proporción = \frac{A}{B}, \qquad A \in B$$

Aplicando este instrumento, se puede afirmar que el 71% de las personas que ganan 10,000 o más pesos cuentan con un grado de licenciatura, o que el 44% de las personas que ganan entre 0 y 2,999 es porque cuentan sólo con una educación primaria.

Sueldo	**Primaria**	**Secundaria**	**Preparatoria**	**Licenciatura**	**Total**

0-2,999	$\frac{50}{114} = 0.44$	$\frac{34}{114} = 0.30$	$\frac{20}{114} = 0.18$	$\frac{10}{114} = 0.09$	$\frac{114}{114} = 1$
3,000-7,999	$\frac{30}{106} = 0.28$	$\frac{30}{106} = 0.30$	$\frac{23}{106} = 0.22$	$\frac{23}{106} = 0.22$	$\frac{106}{106} = 1$
8,000-9,999	$\frac{14}{123} = 0.11$	$\frac{22}{123} = 0.18$	$\frac{42}{123} = 0.34$	$\frac{45}{123} = 0.27$	$\frac{123}{123} = 1$
10,000 o más	$\frac{2}{85} = 0.02$	$\frac{8}{85} = 0.09$	$\frac{15}{85} = 0.18$	$\frac{60}{85} = 0.71$	$\frac{85}{85} = 1$

Cuadro 1: Proporciones en datos bivariados. Fuente: elaboración propia

Regresando a la segunda interrogante que nos hicimos: ¿existe una relación entre estas dos variables? Para construir la respuesta se utiliza una forma gráfica que suele ocuparse para entender si existe o no una (fuerte) relación entre estas dos variables.

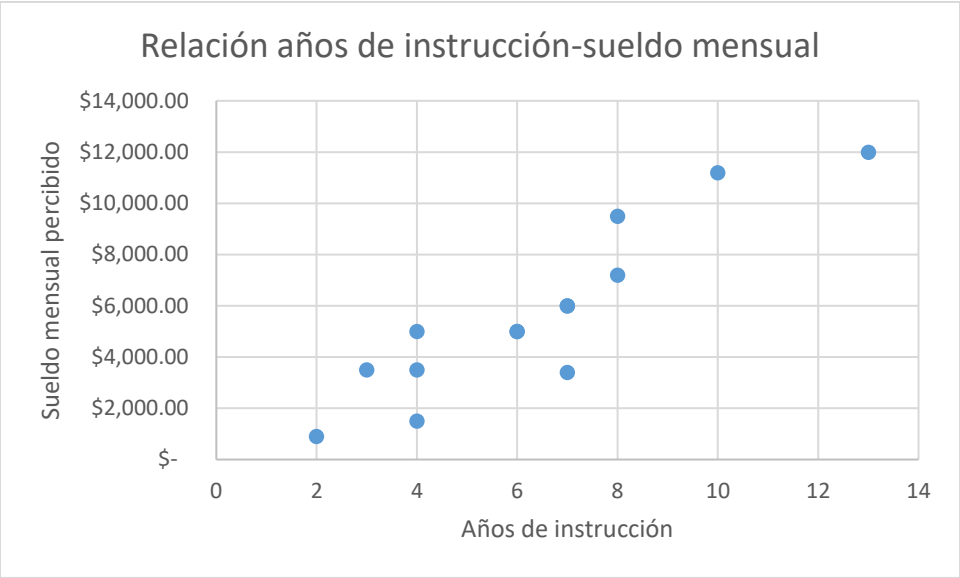
Suponga los datos del cuadro de la derecha.

Años de instrucción	Sueldo actual
6	$ 5,000.00
8	$ 7,200.00
3	$ 3,500.00
4	$ 3,500.00
7	$ 6,000.00
13	$ 12,000.00
4	$ 5,000.00
10	$ 11,200.00
7	$ 6,000.00
8	$ 9,500.00
7	$ 3,400.00
6	$ 5,000.00
4	$ 1,500.00
2	$ 900.00

A partir de dos variables cuantitativas, es posible establecer una relación entre estas dos (si existiese). Ahora bien, para poder apreciar si existe (o no) una relación, se hará uso de los gráficos, uno de los más utilizados son los gráficos de dispersión. Lo que hace un gráfico de dispersión es establecer una relación entre una variable cuantitativa (eje de las abscisas) y su homólogo en el eje de las ordenadas. Esto significa, que las primera tres observaciones de nuestro ejemplo, quedaría con coordenadas (6,5000); (8,7200) y (3,3500).

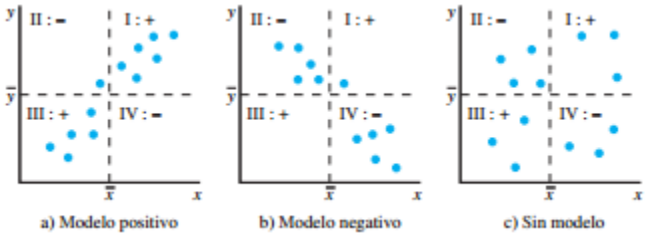
El gráfico 1 muestra el diagrama de dispersión de nuestro ejemplo. En él puede observarse que existe una fuerte relación entre los años de instrucción y el sueldo actual, nos conduce a pensar que, estadísticamente los años de instrucción explican el sueldo percibido, y a mayor nivel de instrucción, mayores ingresos mensuales.

Gráfico 1: Gráfico de dispersión



Es necesario estacar que no siempre la relación entre dos variables es positiva (recta hacia arriba), dependiendo del fenómeno bajo análisis puede darse el caso donde la recta tienda hacia abajo (tendencia negativa) o donde exista una gran dispersión de modo que no pueda ser posible (gráficamente) identificar una tendencia; en este último caso nos conduciría a pensar que estadísticamente no existe una relación entre ambas variables.

Ilustración 6: Tipos de modelos bivariados.



Fuente: (Mendenhall, et al., 2010)

La inferencia estadística

Como ya se mencionó anteriormente, la estadística no solo tiene como función describir una serie de datos, sino que también busca hacer inferencias sobre una población con los datos recolectados a partir de una muestra representativa. En esta sección revisaremos algunos métodos utilizados para la extracción y análisis de datos de una muestra que permitan realizar observaciones y generar conclusiones sobre una población.

Conceptos básicos:

- Parámetro poblacional: son las características de una variable de una población (media, varianza, moda poblacional);
- Estimador: es un estadístico que a partir de datos de una muestra permite estimar el valor del parámetro.

Ejemplo, si se desea conocer el parámetro μ, que corresponde a la media poblacional normal, su estimador será el valor $\bar{X}$, que tomará los valores particulares de $\bar{x}$, esto es, las medias de las diversas muestras que pudiesen generarse. Entonces, ¿cómo debe de ser el estimador $\bar{X}$ para que este explique de manera correcta al parámetro μ?

Se dice que para que un estimador (A) sea óptimo para un parámetro poblacional (a), este debe cumplir las propiedades de ser insesgado, eficiente y consistente. La primera consiste en que el valor del estimador coincide en su media (o esperanza matemática) poblacional, es decir, que:

$$E(A) = \mu = a.$$

La propiedad de eficiencia hace referencia a que si se tuvieran dos estimadores A_1, A_2 de un parámetro, quien tenga menor varianza será considerado como el más eficiente (recordar que la varianza se traduce como la dispersión de los datos alrededor de la media). En este caso, si $\sigma^2_{A_1} < \sigma^2_{A_2}$ implica que el estimador A_1 es más eficiente que el estimador A_2.

La tercera propiedad indica que un estimador se considerará consistente si, conforme se incrementa el tamaño de la muestra, el valor del estimador se aproxima al verdadero valor poblacional y, además, la varianza se anula. Esto es: $\lim_{n\to\infty} A = a$; $\lim_{n\to\infty} \sigma^2 = 0$

En la práctica no siempre se cumplen estas propiedades; se presentan casos donde los estimadores pueden estar sesgados y ser poco eficientes, sin embargo, un requisito mínimo debe ser, que el estimador sea consistente (Gorgas, et al., 2011). En cuanto a la estimación de un estimador, se tienen dos maneras de determinar su valor: la primera es a través de la expresión de un único valor, al cual se le denomina estimación puntual; la segunda, se realiza a través de la construcción de un intervalo donde se tienen probabilidades que el parámetro pueda estar dentro, a este se le llama estimación por intervalos de confianza.

- Principales estimadores puntuales

Un estimador puntual, como se expresó anteriormente, es un indicador que permite estimar el valor de un parámetro poblacional. Visto de forma práctica, los estimadores puntuales de los parámetros $\mu \; y \; \sigma^2$ (media y varianza poblacionales, respectivamente) serán $\bar{X} \; y \; S^2$ (esto es, media y varianzas de una muestra, respectivamente) en el caso de una población normal. En resumen:

$$E(\bar{X}) = \mu \; ; E(S^2) = \sigma^2$$

Esto ocurre si los estimadores cuentan con sus características de ser insesgados, eficientes y consistentes.

- Estimación por intervalos de confianza

Un estimador puntual puede incurrir en un error al no explicar correctamente el valor del parámetro poblacional, es por ello que en algunos casos suele recomendarse hacer uso de intervalos de confianza. Un intervalo de confianza permite no capturar un único estimador,

sino un rango que vaya desde el límite inferior L_1 al límite superior L_2. De esta manera, se explica que existe la probabilidad de que el parámetro poblacional esté dentro del intervalo $[L_1, L_2]$. En términos de probabilidad se tiene que:

$$P(L_1 < \beta < L_2) = 1 - a$$

Lo anterior se explica de la siguiente manera: existe una probabilidad de $1 - a$ de seleccionar una muestra que conduzca a un intervalo que contenga el parámetro poblacional, es decir, que a indica el porcentaje de la posibilidad de seleccionar una muestra que **no** contenga al parámetro poblacional.

A continuación, se presentan las siguientes fórmulas para el cálculo de intervalos de confianza en poblaciones normales:

- Media con varianza poblacional conocida y con población infinita o muestro con remplazo: $P\left(\bar{X} - z_{\frac{a}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\frac{a}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - a$;
- Media con varianza poblacional conocida y con población finita o muestro sin remplazo: $P\left(\bar{X} - z_{\frac{a}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + z_{\frac{a}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}\right) = 1 - a$;
- Media con varianza poblacional desconocida y $n > 30$:
  $P\left(\bar{X} - z_{\frac{a}{2}}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\frac{a}{2}}\left(\frac{S}{\sqrt{n}}\right)\right) = 1 - a$;
- Media con varianza poblacional desconocida y $n < 30$:

  $P\left(-t_{\frac{a}{2,n-1}} < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < t_{\frac{a}{2,n-1}}\right) = 1 - a$ o; $P\left(\bar{X} - t_{\frac{a}{2,n-1}}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{a}{2,n-1}}\frac{S}{\sqrt{n}}\right) = 1 - a$

- Varianza: $P\left(\frac{(n-1)S^2}{\chi^2_{a,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-a,n-1}}\right) = 1 - a$

MÓDULO II. APLICACIÓN A LA ECONOMÍA DEL DESARROLLO REGIONAL Y METROPOLITANO SUSTENTABLE DE LA ESTADÍSTICA BÁSICA

En el módulo anterior, se ofreció un somero repaso de los conceptos básicos de la estadística como herramienta para el análisis de datos y funciones en diversas ciencias, incluidas la ciencia económica. Ahora bien, en este módulo se hará un énfasis en la aplicación de la estadística dentro del desarrollo regional. El alumno debe estar consciente que el uso de la estadística dentro de este campo es prácticamente ilimitado, por lo que en estas notas únicamente se muestra un ejemplo del uso de dicho instrumento.

Supóngase el caso donde se desee conocer la estructura de las aglomeraciones productivas en el Estado de México. En primera instancia, deben de recolectarse los datos que se usarán durante el estudio, una fuente de suma importancia son los Censos Económicos que ofrece el Instituto Nacional de Estadística y Geografía (INEGI). A pesar de ser levantados cada cinco años, no existe otra fuente que tenga una cobertura similar a esta (García & Carranco, 2008).

Para el caso de estudio, García y Carranco (2008) proponen recoger la información del sector manufacturero puesto que permite identificar el núcleo de las actividades relacionadas, pues permite ubicar fácilmente la cadena productiva de un bien. Para realizar el ejercicio es necesario recoger los datos de empleo en el sector manufacturero de los 125 municipios que integran el Estado de México.

Como existe la totalidad de los datos de los 125 municipios que es el universo poblacional, no hay necesidad de realizar alguna inferencia estadística; en caso contrario, si no se existiera la información de todos los municipios y se desea hablar de la entidad federativa como tal, sería necesario hacer uso de la inferencia estadística (contando con la opción de realizar el análisis con un intervalo de confianza o un estimador puntual, según la conveniencia del investigador).

Considerando la población, el modelo requiere de los siguientes indicadores (García & Carranco, 2008):

$E_{ij} = Empleo\ del\ sector\ i\ en\ el\ municipio\ j$;

$E_j = Empleo\ manufacturero\ del\ municipio\ j$;

$E_{iM} = Empleo\ del\ sector\ i\ en\ el\ Estado\ de\ México$;

$E_M = Empleo\ manufacturero\ en\ el\ Estado\ de\ México$;

$X_k = Proporción\ acumulada\ de\ municipios\ contabilizados$;

$Y_k = Proporción\ acumulada\ del\ empleo\ aportado\ en\ la\ actividad\ i\ por\ los\ municipios\ contabilizados$;

$QL_{ij} = Coeficiente\ de\ Localización\ de\ la\ actividad\ i\ en\ el\ municipio\ j$;

$PR = \text{Índice de participación Relativa del municipio } j \text{ y la actividad } i;$

$HH = Coeficiente\ Hirschman - Herfindahl\ de\ la\ actividad\ i\ y\ el\ municipio\ j;$

El Coeficiente de Localización permite comparar la estructura sectorial entre dos espacios distintos. Indica que, de existir $QL > 1$ el sector en el municipio tiene mayor peso que en la región. Este coeficiente se calcula de la siguiente manera:

$$QL_{ij} = \frac{E_{ij}/E_j}{E_{iM}/E_M}$$

El coeficiente de participación relativa, mide la aportación del municipio al Estado. Este toma los valores entre 0 y 1, mientras mayor sea el número, mayor participación dentro del Estado tiene.

$$PR = \frac{E_{ij}}{E_{iM}}$$

El índice $Hirschman - Herfindahl$ (HH) modificado muestra el peso de un sector en la estructura productiva local. De tomar un valor positivo se indica que la actividad económica tiene fuerte importancia en el municipio.

$$HH = \frac{E_{ij}}{E_{iM}} - \frac{E_j}{E_M}$$

El cuadro 2 muestra el caso del municipio de Toluca, Estado de México, en él se rescatan los datos obtenidos del Censo Económico 2014 de INEGI. Puede notarse que, para el caso de la actividad económica *Fabricación de equipos de transporte,* el municipio aporta el 34% del empleo de toda la entidad (PR). Asimismo, a través del Coeficiente de Localización (3.15) se puede observar que esta actividad toma mayor importancia para el municipio de Toluca que para el caso estatal.

Cuadro 2: Empleo, participaciones relativas e indicadores del Municipio de Toluca.

Concepto	***Valor***
Empleo total Estatal	2,023,837
Empleo total Toluca	221,323
Empleo total de la actividad económica "Fabricación de equipos de transporte" estatal	46,005
Empleo total de la actividad económica "Fabricación de equipos de transporte" Toluca	15,870
Coeficiente de Localización	3.154429
Coeficiente de Participación Relativa	0.344962
Índice HH	0.235604

Participación de actividad	0.071705

Fuente: elaboración propia con datos del Censo Económico 2014 (INEGI)

Al observar el índice $Hirschman - Herfindahl$ (HH), cuyo valor es superior a 0, expresa que la actividad económica sí resulta ser importante para el municipio. Además, la *participación de actividad*, revela que la actividad económica aporta el 7% el empleo en el municipio.

Ahora obsérvese el caso de Nezahualcóyotl (cuadro 3), los resultados obtenidos son sumamente diversos a los obtenidos en el caso de la capital mexiquense. En ellos se observa que el Coeficiente de Localización no supera la unidad, esto indica que en Nezahualcóyotl no tiene el mismo peso la actividad económica que en la entidad.

El índice $Hirschman - Herfindahl$ obtenido es negativo, esto responde a que en el municipio, esta actividad económica no tiene gran peso en el municipio. Esto resulta significativo al analizar la *participación de actividad* misma que indica que en el empleo total municipal, la fabricación de equipos de transporte aporta un 0.8%, cifra muy inferior al 7% del municipio capitalino.

Cuadro 3: Empleo, participaciones relativas e indicadores del Municipio de Nezahualcóyotl.

Concepto	**Valor**
Empleo total Estatal	2,023,837
Empleo total Nezahualcóyotl	12,607
Empleo total de la actividad económica "Fabricación de equipos de transporte" estatal	46,005
Empleo total de la actividad económica "Fabricación de equipos de transporte" Nezahualcóyotl	106
Coeficiente de Localización	0.3698832
Coeficiente de Participación Relativa	0.0023041
Índice HH	-0.0039251
Participación de actividad	0.0084080

Fuente: elaboración propia con datos del Censo Económico 2014 (INEGI)

Con el ejemplo realizado, se puede observar el uso de la estadística como una herramienta para el análisis de distintas regiones y de sectores económicos particulares. Si bien, la estadística provee de mucha información que el investigador requiere para realizar su trabajo, éste deberá elaborar sus propias hipótesis y conclusiones; por consecuencia, interpretar los resultados que la estadística le ofrece.

El INEGI a través de sus censos económicos ofrece información precisa posible[2] respecto a la población de estudio. Sin embargo, si suponemos que solo se cuenta con poca información, nos conduciría a contestar la pregunta: ¿cómo se hace uso de la inferencia estadística?

En primer lugar, se considera que la población sigue una distribución de probabilidad normal, si bien puede suceder el caso en que no se cuente con esta información (puesto que precisamente se desconocen los datos de la población) es posible que deba de suponerse esta distribución para hacer uso de la estadística inferencial[3]. En el gráfico 2, y a manera de demostración únicamente, se presentan los datos de la población de los municipios del Estado de México. La variable de estudio es la relación Industrias Alimenticias-empleo de industrias manufactureras total. Esto es:

$$Relación = \frac{Empleo_{Industrias\ Alimenticias,j}}{Empleo_{Industrias\ manufactureras,j}}$$

Los datos obtenidos con el programa E-views en su versión 9, indica que, en promedio, el 39.5% de la población que trabaja en la industria manufactura, labora en la industria de alimentos.

La serie de datos, estadísticamente seguirán una distribución normal si la asimetría[4] esigual a 0 y la curtosis[5] igual a 3. Para el caso que nos ocupa se aproximan mucho a estos valores teóricos: asimetría igual a 0.33 y, curtosis igual a 2.17.

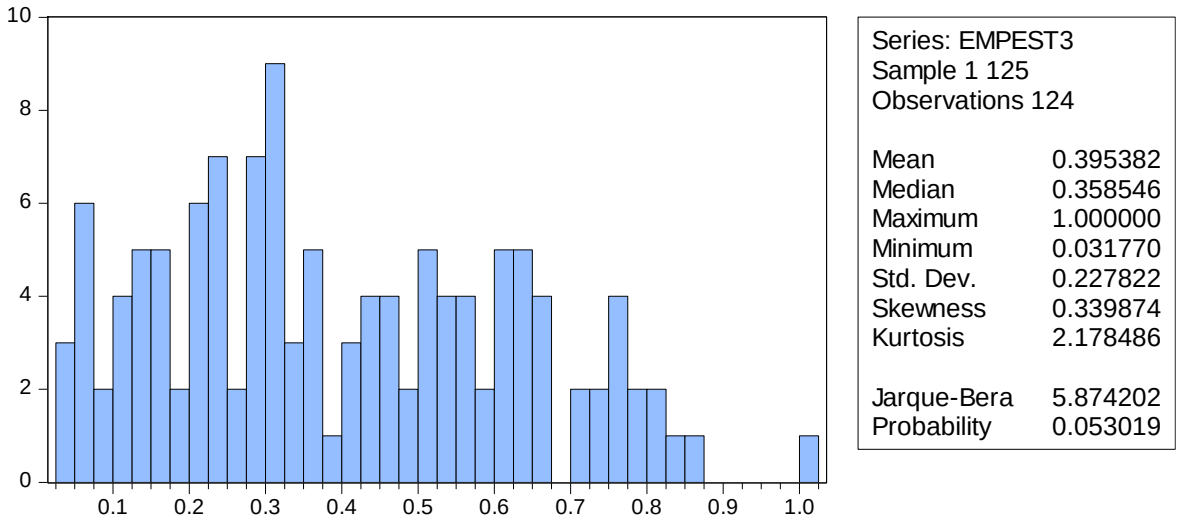
Gráfico 2: Estadística básica proporción empleo en Industria Alimenticia sobre empleo en manufactura total municipal en el Estado de México.

[2] Cabe destacar que los Censos pueden incurrir en un breve sesgo debido a las condiciones de confidencialidad del INEGI, una de ellas es que la información no es difundida si el número de unidades económicas en una entidad es inferior a 3 (García & Carranco, 2008).

[3] Esto puede realizarse debido a la existencia del teorema del límite central que señala que independientemente de la función de distribución que dio origen a la serie de datos, conforme crezca el tamaño de la muestra se comportará como un normal.

[4] Asimetría ($Skewness$) igual a 0, indica que la distribución se encuentra en equilibrio y, por ende, la media no presenta algún sesgo a ningún lado.

[5] Curtosis ($Kurtosis$) igual 3, indica que tan *picuda* se encuentra la distribución de probabilidad (Vela, 2010). En su conjunto, $asimetría = 0\ y\ curtosis = 3$ se puede concluir que la distribución sigue una forma normal. $D = N(0, \sigma^2)$



Fuente: elaboración propia con datos de INEGI.

La pregunta que sigue es: ¿qué pasaría si no se cuentan con todos los datos de la población?. Supóngase que sólo contamos con datos de 25 municipios[6]. De la forma anterior, se obtiene la proporción $\frac{Empleo_{Industrias\ Alimenticias,j}}{Empleo_{Industrias\ manufactureras,j}}$ y se calcula una media cuyo valor es de 0.4543. Esta media indica que, de los 25 municipios aleatoriamente seleccionados, el 45.43% del empleo en manufacturas es generado gracias a industrias alimenticias. Para obtener un intervalo de confianza que permita establecer límites entre los cuales se encuentre la media poblacional y, por las características de la muestra, esta tiene varianza poblacional desconocida y un número de datos menor a 30, se usará la expresión:

$$P\left(\bar{X} - t_{\frac{a}{2},n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{a}{2},n-1}\frac{S}{\sqrt{n}}\right) = 1 - a$$

Sustituyendo la expresión con los datos obtenidos se tiene que:

$$P\left(0.4543 - 2.064\left(\frac{0.2498}{\sqrt{25}}\right) < \mu < 0.4543 + 2.064\left(\frac{0.2498}{\sqrt{25}}\right)\right) = 1 - a; a = 0.05$$

$$I = [0.3511, 0.5574]$$

El resultado anterior indica que, con un nivel de confianza $1 - a$ igual a 0.95 o 95%, la muestra elegida denota un intervalo donde se encuentre la media poblacional. Se puede suponer que la media poblacional se encuentra en el intervalo generado que comprende entre de 0.3511 a 0.5574. En efecto, retomando el ejercicio anterior se observa que la media poblacional fue de 0.3953, misma que se encuentra dentro del intervalo de confianza, por lo que efectivamente, la estadística inferencial fue una herramienta inequívoca al describir la media.

[6] Véase el Anexo 1.1, los municipios seleccionados fueron extraído de forma aleatoria.

Ahora bien, durante al capítulo anterior, se explicó que un estimador puntual (como lo es, la media) cuenta con tres características, que es ser insesgado, debe tener máxima eficiencia y además debe ser consistente. Para lo anterior, se presentará una segunda muestra donde se incluyen los mismos 25 municipios aleatorio más 10 extras, dando una total de 35 municipios (Véase Anexo 1.2).

Ahora se cuentan con dos estimadores, uno perteneciente a la muestra de 25 municipios, y otra que incluye 10 municipios más a los cuales denotaremos por $\bar{\bar{X}}_{25}, \bar{X}_{35}$ respectivamente.

Ambas medias de la muestra son superiores a la media poblacional, por lo que se dice que los estimadores están sesgados. Al corroborar la eficiencia, se puede observar que la desviación estándar de la muestra con 35 municipios es menor que la muestra que recoge únicamente a 25 municipios, por lo cual, asumiremos que el estimador $\bar{X}_{35}$ es más eficiente que el estimador $\bar{\bar{X}}_{25}$.

Por último, ¿el estimador será consistente? De acuerdo al concepto, mientras más se amplíe el tamaño de la muestra, el estimador tenderá a igualarse con el parámetro poblacional. Efectivamente, mientras $\bar{\bar{X}}_{25} = 0.4543$, si se incrementan 10 municipios, entonces $\bar{X}_{35} = 0.4016$, el cual es un número inferior pero más cercano al valor $\mu = 0.3953$. Así, se dice que nuestros estimadores presentan sesgo, $\bar{X}_{35}$ es más eficiente que $\bar{\bar{X}}_{25}$, y que además son consistentes, esto último es de esperarse puesto que nuestra población sigue una distribución normal y conforme se obtenga más información de la población, los resultados se parecerán más al comportamiento poblacional.

Se recomienda que el alumno defina las metas de su investigación, recopile los datos que puede utilizar para el cumplimiento de sus objetivos y, por medio de diversas técnicas estadísticas puede tanto explicar los datos observados, como realizar una inferencia e inclusive proyecciones de los mismos.

Distribución Normal o gaussiana

Características de la curva normal.

La distribución gaussiana, recibe también el nombre de *distribución normal*, ya que una gran mayoría de las variables aleatorias (v. a.) continuas de la naturaleza siguen esta distribución.

- Es unimodal ya que, la moda, que es el punto sobre el eje horizontal donde la curva tiene su máximo, ocurre donde $x = \mu$ y con lo cual en μ coinciden la media, la mediana y la moda.
- La curva es simétrica alrededor de su eje vertical donde se tiene la media μ por tanto $P[X \leq \mu] = P[X \geq \mu] = ½$,
- Está determinada por 2 parámetros: la media y la desviación estándar.
- El área comprendida bajo la curva entre la media y cualquier valor Xi de la variable aleatoria se expresa en función del número de desviaciones estándar que dicho valor Xi diste de la media aritmética.
- La curva tiene sus puntos de inflexión en $\mu \pm \sigma$. Es cóncava hacia abajo si $\mu - \sigma < X < \mu + \sigma$, y es cóncava hacia arriba en cualquier otro punto.
- La curva normal se acerca al eje horizontal en forma asintótica en cualquiera de las dos direcciones, alejándose de la media.
- El área total bajo la curva y arriba del eje horizontal es igual a 1.
- El número de valores Xi que toma la variable aleatoria X es infinita.

Se dice que una v.a. *X* sigue una **distribución normal** de parámetros μ y σ^2, lo que representamos del modo $X \rightarrow N(\mu,\sigma^2)$ si su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty \prec x \prec \infty, \ \forall x \in \Re$$

Donde:

μ = Promedio

σ = desviación estándar

π = 3.14159

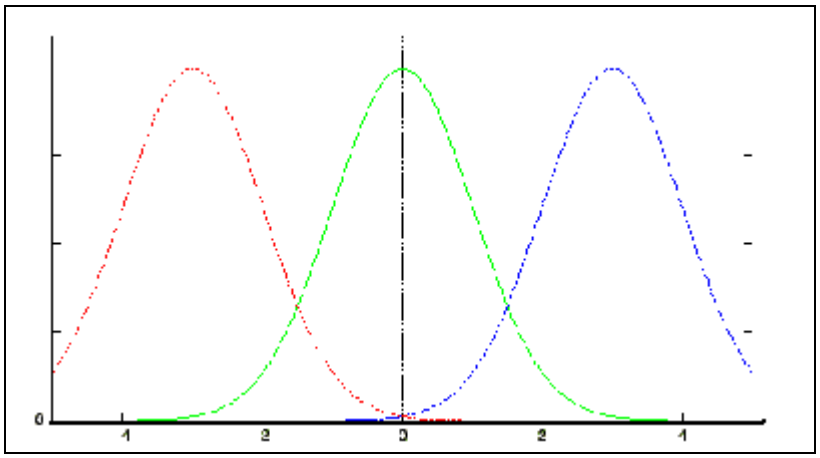
e = 2.711828

Estos dos parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\sigma}^2$ coinciden además con la media (esperanza) y la varianza respectivamente de la distribución:

$$E[X] = \mu$$
$$VAR[X] = \sigma^2$$

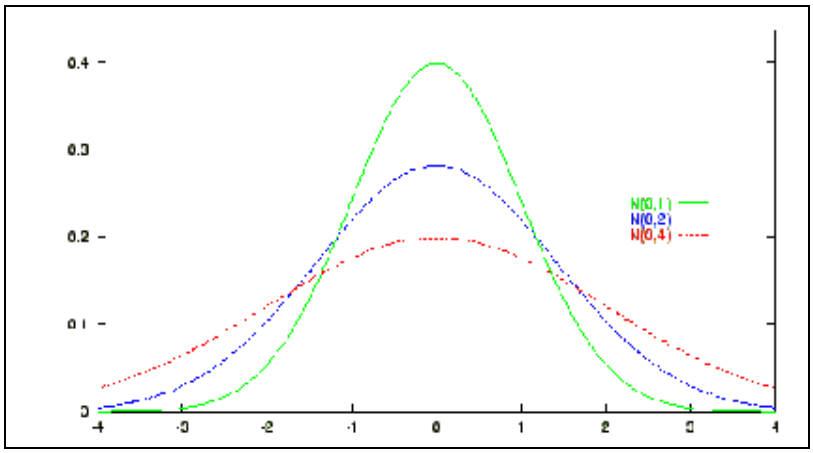
La forma de la función de densidad es la llamada *campana de Gauss*. La forma de la campana de Gauss depende de los parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\sigma}$: $\boldsymbol{\mu}$ indica la posición de la campana *(parámetro de centralización)*;

Distribuciones gaussianas con diferentes medias e igual dispersión.



($\boldsymbol{\sigma}^2$ o $\boldsymbol{\sigma}$) será el parámetro de dispersión. Cuanto menor sea, mayor cantidad de masa de probabilidad habrá concentrada alrededor de la media y cuanto mayor sea “más aplastado" será.

Distribuciones gaussianas con igual media pero varianza diferente.



Como se ha mencionado, la ley de probabilidad gaussiana la encontramos en la mayoría de los fenómenos que observamos en la naturaleza. Sin embargo, a pesar de su utilidad, hay que apuntar un hecho *negativo* para esta ley de probabilidad:

La función e^{-x^2} no posee primitiva conocida.

Las consecuencias desde el punto de vista práctico son importantes, ya que eso impide el que podamos escribir de modo sencillo la función de distribución de la normal, y nos tenemos que limitar a decir que:

$$F(x) = P[X \leq x] = \int_{-\infty}^{\infty} f(t)dt = \frac{1}{\sigma\sqrt{2x}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dt$$

sin poder hacer uso de ninguna expresión que la simplifique. Afortunadamente esto no impide que para un valor de *x* fijo, *F*(*x*) pueda ser calculado. Para la utilización en problemas prácticos de la función de distribución *F*, existen ciertas tablas donde se ofrecen (con varios decimales de precisión) los valores *F*(*x*) para una serie limitada de valores x_i dados. Normalmente *F* se encuentra tabulada para una distribución *Z*, normal de media 0 y varianza 1 que se denomina *distribución normal estandarizada*:

$$Z \to N(0,1) \Leftrightarrow fz(z) = \frac{1}{\sqrt{2x}} e^{-\frac{n2}{2}} \forall z \in R$$

En el caso de que tengamos una distribución diferente $X \sim N(\mu,\sigma^2)$, se obtiene *Z* haciendo el siguiente cambio:

$$X \to N(\mu, \sigma^2) \Rightarrow z = \frac{X - \mu}{\sigma} \to N(0,1)$$

De manera general se tiene:

Proposición (Cambio de origen y escala)

Sean $a, b \in R$ Entonces

$$X \to N(\mu, \sigma^2) \Rightarrow Y = a + b \cdot X \to N(a + b\mu, (b\sigma)^2)$$

Este resultado puede ser utilizado del siguiente modo: Si $X \sim N(\mu, \sigma^2)$, y nos interesa calcular $Fx(X) = P[X \leq x]$,

1. Hacemos el cambio $Z = \frac{X - \mu}{\sigma} \to N(0,1)$ y calculamos $Z = \frac{x - \mu}{\sigma}$;

2. Usamos la *tabla* relativa a la distribución N(0,1) para obtener (de modo aproximado)

$Fz(z) = P[Z \leq z]$

3. Como $P[z \leq z] = P\left[\frac{x - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right] = P[X \leq x] = Fx(x)$

tenemos que el valor obtenido en la tabla, $F_Z(z)$ es la probabilidad buscada.

La productividad media que alcanza un trabajador en la maquila es de 2.75 televisores diarios con una desviación estándar de 0.63.

a) ¿Cuál es la probabilidad o proporción de trabajadores con una productividad mayor a 3.50 televisores?
b) si seleccionamos unos trabajadores al azar, cuál es la probabilidad que produzcan entre 2 y 3 televisores.

$$P(x > 3.5) = 0.500 - 0.3830 = 0.117$$

$$P(2 \leq x \leq 3) = 0.3830 + 0.1517 = 0.5347$$

$$Z = \frac{2 - 2.75}{0.63} = \frac{-0.75}{0.63} = -1.1904$$

$$Z = \frac{3 - 2.75}{0.63} = \frac{0.25}{0.63} = 0.3968$$

c) Cuál es la probabilidad o proporción de una productividad mayor de 1.75 televisores.

$$P(x > 1.75) = 0.500 + 0.4429 = 0.9429$$

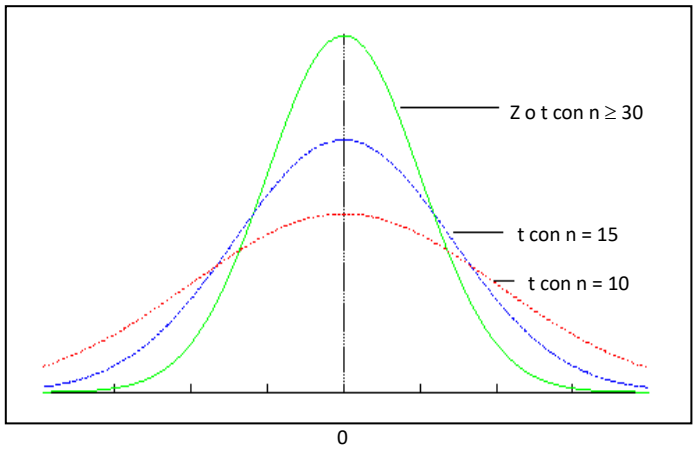
DISTRIBUCIÓN "T" DE STUDENT

Cuando se toma una muestra pequeña, la distribución normal puede no aplicarse. El teorema del límite central asegura normalidad en el proceso de muestreo sólo si la muestra es grande. Cuando se utiliza una muestra pequeña, puede ser necesaria una distribución alternativa, la distribución t Student. Específicamente la distribución t se utiliza solo cuando se cumplen las tres condiciones.

1. La muestra es pequeña
2. σ es desconocida
3. La población es normal o casi normal

Si σ es conocida, la distribución Z se usa inclusive si la muestra es pequeña.

Al igual que la distribución Z, la distribución t tiene una media de cero, es simétrica con respecto a la media y oscila entre -∞ y +∞. Sin embargo, mientras que .la distribución Z tiene una varianza de $s^2 = 1$, la varianza de la distribución t es mayor que 1. Por tanto, es más plana y más dispersa que la distribución Z. La varianza para la distribución t es:



$$S^2 = \frac{n-1}{n-3}$$

Si el tamaño de la muestra es pequeño, los valores de S^2 fluctúan considerablemente de muestra a muestra.

En realidad la distribución t es una familia de distribuciones cada una con su propia varianza. La varianza decrece hasta 1 cuando los grados de libertad aumentan. La varianza depende de los grados de libertad (g.l.) definidos como el número de observaciones que se pueden escoger libremente.

Grados de libertad: Es el número de observaciones que se pueden escoger menos el número de restricciones impuestas sobre tales observaciones, en donde una restricción es algún valor que tales observaciones deben poseer.

Para un número alto de grados de libertad se puede aproximar la distribución de Student por la normal, es decir,

$$t_n \xrightarrow{n\to\infty} N(0,1)$$

El estadístico t se calcula en gran parte como el estadístico Z

$$t = \frac{\bar{x} - \mu}{S_{\bar{x}}}$$

La prueba t de Student, es un método de análisis estadístico, que compara las medias de dos categorías dentro de una variable dependiente, o las medias de dos grupos diferentes. Es una prueba paramétrica, o sea que solo sirve para comparar variables numéricas de distribución normal. En caso de tener que analizar variables numéricas de distribución anormal, se debe utilizar otro tipo de pruebas no paramétricas.

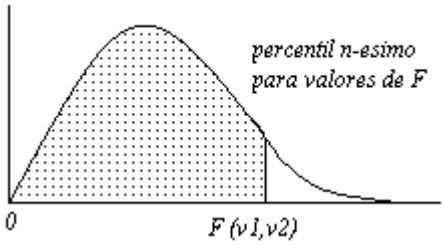
La prueba t Student, arroja el valor del estadístico t. Según sea el valor de t, corresponderá un valor de significación estadística determinado.

La prueba t para muestras independientes se utiliza para comparar la media de dos grupos o dos categorías dentro de una misma variable dependiente.

LA DISTRIBUCIÓN F

La distribución existe sólo para los valores NO negativos de **F**, presenta asimetría positiva y tiene dos parámetros:

v_1 ($n_1 - 1$) grados de libertad del numerados y v_2 ($n_2 - 1$) grados de libertad del denominador



Se usa como estadístico de prueba para saber si dos muestras provienen de poblaciones que poseen varianzas iguales y también se aplica cuando se trata de comparar simultáneamente varias medias poblacionales.

Es importante prestar atención a que la tabla se refiera al porcentaje adecuado de F y no confundir las entradas (v_1, v_2)

Características de la distribución *F*

- Existe una "familia" de distribuciones *F*.
- Cada miembro de la familia está determinado por dos parámetros: los grados de libertad (*gl*) en el numerador y los grados de libertad en el denominador. La forma de la curva varia a medida que lo hacen los grados de libertad.
- El valor de *F* no puede ser negativo y es una distribución continua.
- La distribución *F* tiene sesgo positivo.
- Sus valores varían de 0 a ∞ . Con forme $F \to \infty$ la curva se aproxima al eje *X*.

Contraste de hipótesis

Además de la estadística básica, a partir de estimadores como se trabajó en la sección anterior, es necesario conocer que, para la aplicación de la estadística se debe tener conocimientos acerca del *contraste de hipótesis*.

Gorgas, et al. (2011), menciona que la hipótesis la apreciamos dentro del método científico, como la explicación de lo que sucede a nuestro entorno. Posteriormente (dentro del mismo método), se llega a un proceso de experimentación y evaluación para verificar si la hipótesis (idea/afirmación/teoría) planteada en el inicio se rechaza o no. La hipótesis, se contrasta con los datos experimentales, si coincide dentro de un margen de error, la hipótesis se mantiene, de otra manera, debe rechazarse y elaborar un nuevo modelo.

La hipótesis, es una afirmación de alguna característica estadística, en esencia busca explicar algún fenómeno de una población; sin embargo, como se ha demostrado en el texto, estudiar completamente a una población raramente llega a ser factible por cuestión geográfica, de tiempo, costos, entre otras. Es por ello, que se debe realizar el estudio con alguna muestra de la población y verificar si nuestra afirmación puede ser aplicable dentro

de la población. Si se desea evaluar que el nivel de estudios de la población mexiquense es igual o inferior que el nivel de estudios de la población europea, deberá plantearse la hipótesis que el nivel de estudios de la población mexiquense es superior.

A dicha hipótesis se le conoce como hipótesis nula H_0, por el contrario, se encuentra la hipótesis alternativa H_1, que es aquella que se acepta si H_0 resulta ser rechazada. En el ejemplo: H_0 es que la población mexiquense cuenta con un nivel de estudios superior a la europea, y H_1 será que la población mexiquense cuenta con un nivel de estudios igual o inferior a la europea.

Sin embargo, en este planteamiento se puede incurrir en dos tipos de error, por lo cual habrá de tener mucho cuidado al estructurar nuestras hipótesis:

- Rechazar H_O cuando en realidad es verdadera (error tipo I);
- No rechazar H_1 cuando en realidad es falsa (error tipo II).

El error tipo I, suele denotarse con a, denominado nivel de significancia de la prueba y, permite establecer las regiones de aceptación y de rechazo. ¿Qué es esto? Bien, al momento de establecer una hipótesis, pueden ocurrir dos cosas, o que se acepte H_O o que se rechace.

Suponiendo, por ejemplo, que se desea demostrar que la población toluqueña cuenta con más de 12 años de instrucción como promedio, entonces $H_1: \mu > 12; H_0: \mu \leq 12$. De esta manera se buscará demostrar que debe rechazarse H_0. Con un $a = 0.01$, una muestra recogida a partir de 100 cuestionarios arroja que el promedio de años de instrucción es de 13.5 con una desviación estándar igual a 2.9.

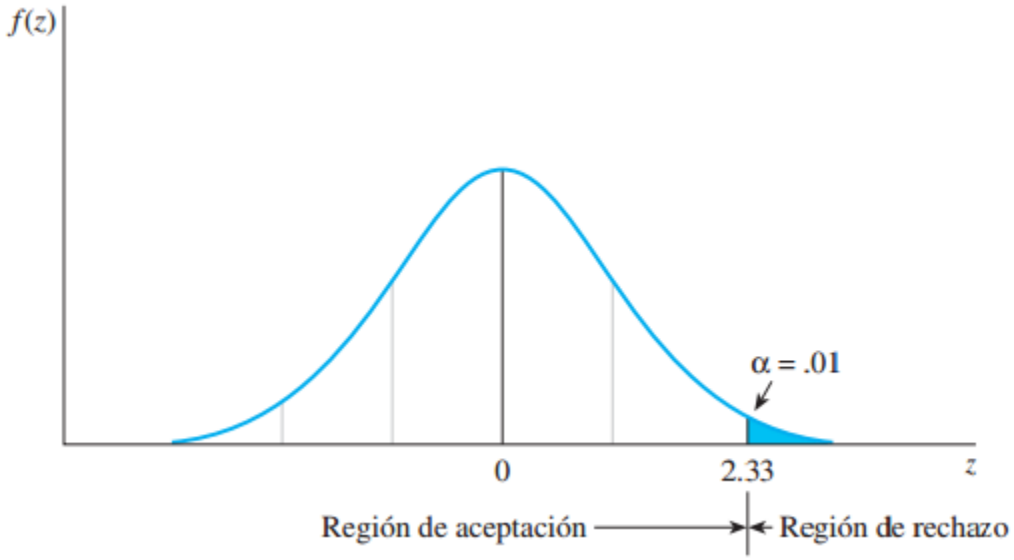


Ilustración 7: Regiones de aceptación y rechazo de Ho. Fuente: (Mendenhall, et al., 2010)

Puede observarse en la ilustración 7, que, al distribuirse la población de una forma normal, puede hacerse uso del valor Z.

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}; Considere\ que\ SE = \frac{s}{\sqrt{n}}$$ [7]

Sustituyendo los datos del ejercicio se tiene que:

$$z = \frac{13.5 - 12}{2.9/\sqrt{100}} = 5.1724$$

Como el valor z cayó dentro del área de rechazo, se puede decir que con un nivel de confianza del 99%, H_0 se rechaza por lo que efectivamente, puede aceptarse que el nivel de instrucción de los ciudadanos de Toluca supera los 12 años.

Ahora supongamos que deseamos demostrar que la población toluqueña cuenta con 13 años de instrucción, por ende, $H_0: \bar{x} = 13$ mientras que $H_1: \bar{x} \neq 13$. Como en este ejemplo no se busca que sea menor o mayor, sino que sea igual, la distribución normal ahora estará acotada por dos colas como se observa en la ilustración 8. En ella se observa que ahora buscamos que los datos se encuentren dentro de $z = 0$.

Retomando la fórmula anterior, además suponiendo una nueva muestra de 50 personas cuyo promedio de años de instrucción es de 13.7 y desviación estándar de 3 y sustituyendo se obtiene que:

$$z = \frac{13.7 - 13}{3/\sqrt{50}} = 1.6499$$

Lo anterior posiciona al valor $z = 1.6499$ dentro de la región de no rechazo, por lo cual se comprueba que con un nivel de confianza del 99% se puede considerar que en promedio, la población toluqueña cuenta con 13 años de instrucción.

[7] De esta manera, el valor z_α indica el número de errores estándar que tendrá como límite para rechazar H_0.

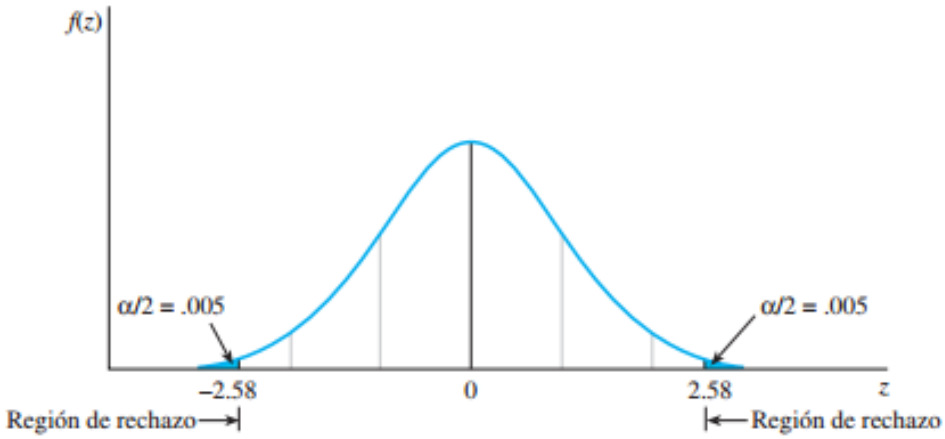


Ilustración 8: Regiones de rechazo y no rechazo Ho. Fuente: (Mendenhall, et al., 2010)

Para muestras pequeñas (inferiores a 30), el valor crítico considerado a partir del valor z no será el más eficiente, por lo que se recomienda considerar la distribución t-Student:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}; n < 30$$

Nota: Es importante que el estudiante considere la resolución de ejercicios para facilitar su comprensión matemática, mismos ejemplos que puede encontrar en libros que se encuentran en la presente bibliografía o en textos de su preferencia.

Diagrama de dispersión y regresión lineal

Como se comentó anteriormente, es posible realizar la estimación de modelos da la forma $y = f(x)$, es decir, una variable y está explicada por una variable x. Algunos ejemplos son:

- El PIB de una región esta explicada por su nivel de empleo: $PIB = f(Empleo)$;
- El rendimiento de un estudiante esta explicada por sus calificaciones y el número de materias aprobadas: $Rend_{est} = f(Calf_{est}, MatAprob_{est})$;
- La inflación está explicada por el nivel de desempleo: $\pi = f(u)$.

De esta manera, se construye una hipótesis respecto si existe una relación positiva o negativa entre dos tipos de variables. Retomando el gráfico 1, puede suponerse que existe una relación positiva entre los años de instrucción y el sueldo mensual adquirido. Por esta razón se proveerá a estructura un modelo lineal simple de la forma:

$$y = a + bx$$

Donde y es la variable dependiente, x es la variable independiente, a es el cruce con el eje de las ordenadas ($a = y_o$), y por último b es la pendiente de la recta. En este sentido, y con base en los resultados que se observan en el gráfico 3, indica que la regresión de la forma $y = a + bx$ toma los valores $y = -860.24 + 1{,}019.6x$, donde y es el sueldo mensual percibido y x son los años de instrucción.

¿Cómo puede interpretarse lo anterior? Bien, primero habrá que observar el signo que toma el coeficiente b, en este caso, al ser positivo rectifica nuestra hipótesis de que existía una relación positiva entre estas variables, es decir, a mayores años de instrucción, mayor sueldo percibido. Ahora el alumno se preguntará, ¿en cuánto? Aquí es donde se analiza el coeficiente b, como toma el valor de $1{,}019.6$, se puede interpretar que, en promedio, por cada año adicional de años de instrucción, el sueldo medio percibido se incrementará en $\$1{,}019.6$.

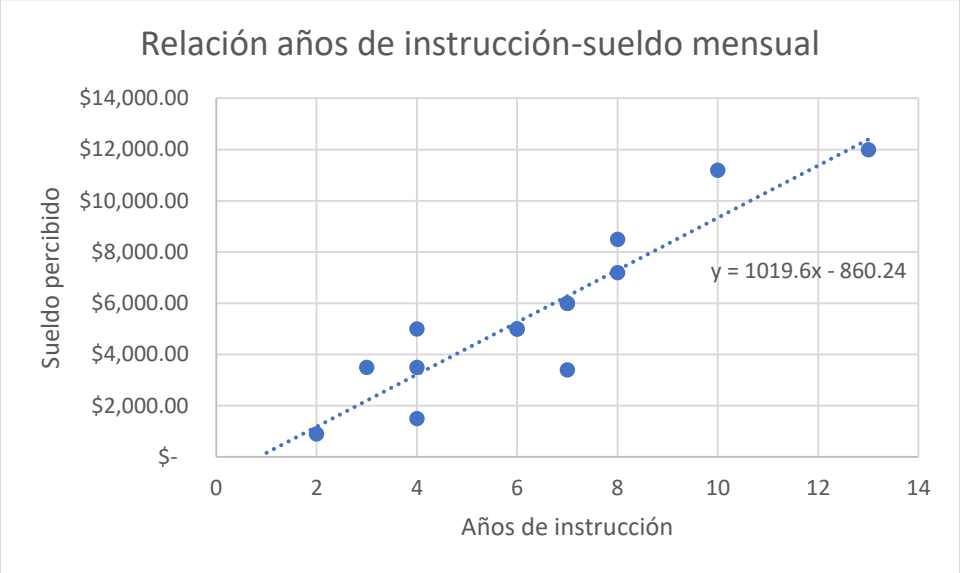


Gráfico 3: Regresión años de instrucción-sueldo

Por último, el valor de $a = -860.14$, indica el nivel de ingresos de una persona sin años de instrucción, este dato, aunque es físicamente incorrecto (una persona no puede "ganar negativos"), si que ayuda si queremos estimar el promedio de una persona con x años de instrucción. Con ello, se propone deducir, ¿cuánto puede ganar en promedio una persona que estudió 10 años en su vida? Para resolver esa sencilla pregunta, únicamente debe sustituirse el valor deseado en x, esto es:

$$y = -860.24 + 1{,}019.6(10) = 9{,}335.76$$

Esto quiere decir, que una persona que ha estudiado 10 años en su vida, puede esperar ganar la cantidad de $\$9{,}335.76$, según los datos del modelo estimado.

Ahora bien, ¿cómo saber qué recta se acerca más a la realidad? Está claro que pueden definirse infinidad de rectas en el modelo, entonces definir la recta más pertinente que permita al modelo simular de la mejor manera lo observado en la realidad será el proceso a seguir a continuación.

Mínimos Cuadrados Ordinarios

El método llamado Mínimos Cuadrados Ordinarios (MCO o MLS por sus siglas en inglés) es aquel que permite, encontrar la recta de mejor ajuste. A grandes rasgos, el MCO busca

minimizar la distancia entre la recta de mejor ajuste y cada uno de los puntos observados. De esta manera, el error $(y - \hat{y})$[8]será mínimo y el modelo estará lo mejor ajustado a la realidad.

El procedimiento, será minimizar la suma de los cuadrados del error SSE:

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$$

Donde:

$b = \frac{S_{xy}}{S_{xx}}; a = \bar{y} - b\bar{x};$

$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y_i}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n};$

$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}.$

En este sentido, y retomando el ejercicio de ejemplo, se tiene que:

Observación (i)	**Años** (x)	**Sueldo** (y)	x^2	$x * y$	y^2
1	6	5,000	36	30,000	25,000,000
2	8	7,200	64	57,600	51,840,000
3	3	3,500	9	10,500	12,250,000
4	4	3,500	16	14,000	12,250,000
5	7	6,000	49	42,000	36,000,000
6	13	12,000	169	156,000	144,000,000
7	4	5,000	16	20,000	25,000,000
8	10	11,200	100	112,000	125,440,000
9	7	6,000	49	42,000	36,000,000
10	8	8,500	64	68,000	72,250,000
11	7	3,400	49	23,800	11,560,000
12	6	5,000	36	30,000	25,000,000
13	4	1,500	16	6,000	2,250,000
14	2	900	4	1,800	810,000
Suma	89	78,700	677	613,700	579,650,000

De esta manera, se puede calcular los siguientes coeficientes:

$$S_{xy} = 613{,}700 - \frac{(89)(78{,}700)}{14} = 113{,}392.85$$

$$S_{xx} = 677 - \frac{7{,}921}{14} = 111.21$$

[8] A continuación, el error se definirá como la diferencia entre la serie real y y la estimada $\hat{y}$.

$$b = \frac{113{,}392.85}{111.21} = 1{,}019.59$$

$$a = \bar{y} - b\bar{x} = 5{,}621 - 1{,}019.6 * 6 = -860.24$$

De esta manera, se logra llegar a la ecuación $y = f(x) \rightarrow y = -860.24 + 1{,}019.5x$ que es aquella donde la recta se ajusta de mejor manera a los datos minimizando el error al cuadrado entre ellos.

Ahora puede salir una nueva interrogante, ¿la variable x sirve para explicar el comportamiento de la variable y?Si la variable x toma diferentes valores y no hace que la variable y de inmute, estamos en presencia que la relación entre estas dos variables no es lineal, esto es que en un modelo poblacional $y = \alpha - \beta x + \epsilon$, β es igual a 0. Para verificar que la variable x si explica a y se realizará una prueba de hipótesis de la forma siguiente:

$$H_0: \beta = \beta_0$$
$$H_1: \beta \neq \beta_0$$

El estadístico t para validar la prueba será dado por:

$$t = \frac{b - \beta_0}{\sqrt{\frac{MSE}{S_{xx}}}} \; ; MSE = s^2; s^2 = varianza\ del\ error$$

Sustituyendo con los valores del ejemplo:

$$t = \frac{1{,}019.6 - 0}{\sqrt{\frac{1{,}663{,}805}{111.12}}} = \ 8.3325$$

Los valores críticos de $t_{\frac{\alpha}{2},12}$ es de ± 3.055, y como el valor 8.3325 en términos absolutos es superior al valor crítico de t, se rechaza la hipótesis nula, por la cual se acepta que $\beta \neq \beta_0$, y con ello, se puede concluir que la variable x explica a la variable y.

Ejercicio práctico

En el anexo 2, se encuentran los datos de población empleada y producción bruta total de los 125 municipios que integran el Estado de México. Los datos fueron recogidos del Censo Económico 2014 que realizó INEGI. Ahora supongamos que se desea establecer una relación entre estas dos variables, y demostrar que dicha relación es positiva, es decir, que, a mayor número de personas empleadas en un municipio, mayor producción bruta existirá.

En primera instancia, se plasmarán los datos en un gráfico de dispersión para observar de forma práctica, si es visible una recta entre ambas variables. El gráfico 4 que se muestra a continuación denota que pudiese existir una recta con tendencia positiva, por lo que nuestra hipótesis pudiera ser verdadera. Para verificar lo anterior, habrá que construir el modelo de la forma $y = a + bx$, para ello se tiene que:

$$Y = a + b * EMP$$

Donde Y es igual a la producción bruta total, y EMP el número de empleados en el municipio, ahora siguiendo los MCO:

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 142{,}234{,}890{,}084 - \frac{(2{,}023{,}837)(1{,}116{,}235)}{125} = 124{,}162{,}262{,}074$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 212{,}653{,}877{,}829 - \frac{4{,}095{,}916{,}202{,}569}{125} = 179{,}886{,}548{,}208.45$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{124{,}162{,}262{,}074}{179{,}886{,}548{,}208.45} = 0.6902$$

$$a = \bar{y} - b\bar{x} = 8{,}930 - 0.6902(16{,}191) = -2{,}245.35$$

Por lo tanto, el modelo final está dado por: $Y = -2{,}245.35 + 0.6902EMP$

La interpretación será la siguiente:

- La relación es positiva, puesto que el valor b es mayor que 0, esto indica que, a mayor número de empleados, mayor producción;
- Por cada persona que es empleada adicionalmente, en promedio la producción se incrementa en 0.6902 millones de pesos (690.2 mil pesos);
- En caso que hipotéticamente no existiese una sola persona empleada en el municipio, este incurriría en producción negativa de $2{,}245.35$. (Debido a que no es físicamente posible, se limitará a decir 0).

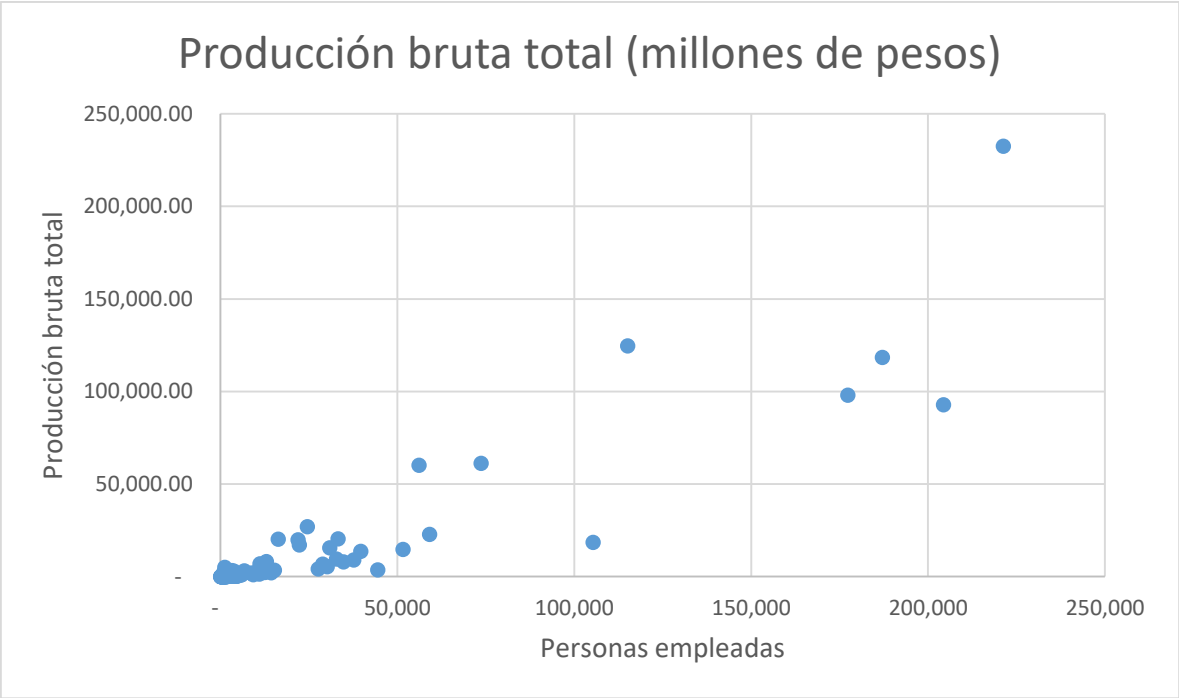


Gráfico 4: Relación empleo total y producción bruta total de los municipios del Estado de México. Fuente: elaboración propia con datos de INEGI

Ahora que ya se tiene el modelo, se procede a realizar la prueba de hipótesis para validar que efectivamente, el número de personas empleadas explica al valor de la producción en los municipios del Estado de México. Para ello, se buscará rechazar la hipótesis nula.

$$H_0: \beta = \beta_0$$
$$H_1: \beta \neq \beta_0$$

Cuyo valor estadístico $t_{\frac{a}{2}}$ estará dado por:

$$t = \frac{b - \beta_0}{\sqrt{\frac{MSE}{S_{xx}}}}$$

Sustituyendo:

$$t = \frac{0.690225 - 0}{\sqrt{\frac{144{,}870{,}216}{179{,}886{,}548{,}208.45}}} = 24.3221$$

Al contrastar con el valor crítico $t_{\frac{a}{2},\infty} = 2.576$[9], se observa que el t estimado es superior en términos absolutos al crítico, por ende, se ubica en la zona de rechazo, por lo cual se acepta

[9] Al ser 123, los grados de libertad superiores a los 30 recomendados para el uso de la tabla t. Es por ello que con muestras muy grandes, los valores de la distribución t tienden a la z, considerando ∞ grados de libertad.

la hipótesis alternativa y se concluye que la variable x y es explicativa de la variable y. En el ejemplo, el número de personas empleadas si explican a la producción bruta total.

De forma general, puede expresarse una regresión de la siguiente manera (Tusell, 2011):

$$y_1 = \beta_0 x_{1,0} + \beta_1 x_{1,1} + \cdots + \beta_{p-1} x_{1,p-1} + \epsilon_1$$

$$y_2 = \beta_0 x_{2,0} + \beta_1 x_{2,1} + \cdots + \beta_{p-1} x_{2,p-1} + \epsilon_1$$

$$\vdots$$

$$y_N = \beta_0 x_{N,0} + \beta_1 x_{N,1} + \cdots + \beta_{p-1}\, x_{N,p-1} + \epsilon_1$$

Y en su forma matricial:

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}$$

$\vec{y}$: $Vector\ N * 1\ de\ observaciones\ de\ la\ variable\ Y$;
X: $Matriz\ N * p\ de\ los\ valores\ de\ la\ variable\ X\ de\ elementos\ x_{i,j}.j - \acute{e}sima\ variable\ y\ i - observación$;
$\vec{\beta}$: $Vector\ de\ parámetros\ \beta_0, \beta_1, \dots, \beta_{p-1}$;
$\vec{\epsilon}$: $Vector\ de\ perturbaciones\ aleatoria\ N * 1$.

Es importante recalcar que, para llevar a cabo una regresión lineal, deben de cumplirse los siguientes supuestos (Gujarati & Porter, 2010):

- El modelo debe ser lineal;
- Los valores de X son independientes el término de error ϵ, esto es: $cov(\epsilon_i, x_i) = 0$;
- El valor medio del vector de error es cero: $E(\epsilon_i) = 0$;
- Existe homoscedasticidad (varianza constante): $var(\epsilon_i) = \sigma^2$;
- No autocorrelación en los errores: $cov(\epsilon_i, \epsilon_j) = 0$;
- El número de observaciones es superior al de parámetros;
- Deben variar los valores de x;
- No hay multicolinealidad, esto es, relación entre las variables x;
- No hay sesgo de especificación, el modelo está especificado correctamente.

De cumplirse dichos supuestos, se podrá considerar una buena regresión y, por ende, los resultados del modelo serán aceptados para la interpretación del problema de investigación.

MÓDULO IV: EJERCICIOS APLICADOS A LOS SISTEMAS DE INFORMACIÓN GEOGRÁFICA

El objetivo de este último módulo, es la aplicación de la herramienta estadística para el uso eficiente de los Sistemas de Información Geográfica. Un Sistema de Información Geográfica (SIG), se define como el conjunto de herramientas diseñadas para obtener, almacenar, recuperar y desplegar datos espaciales del mundo real (Instituto Nacional de Estadística y Geografía, 2014).

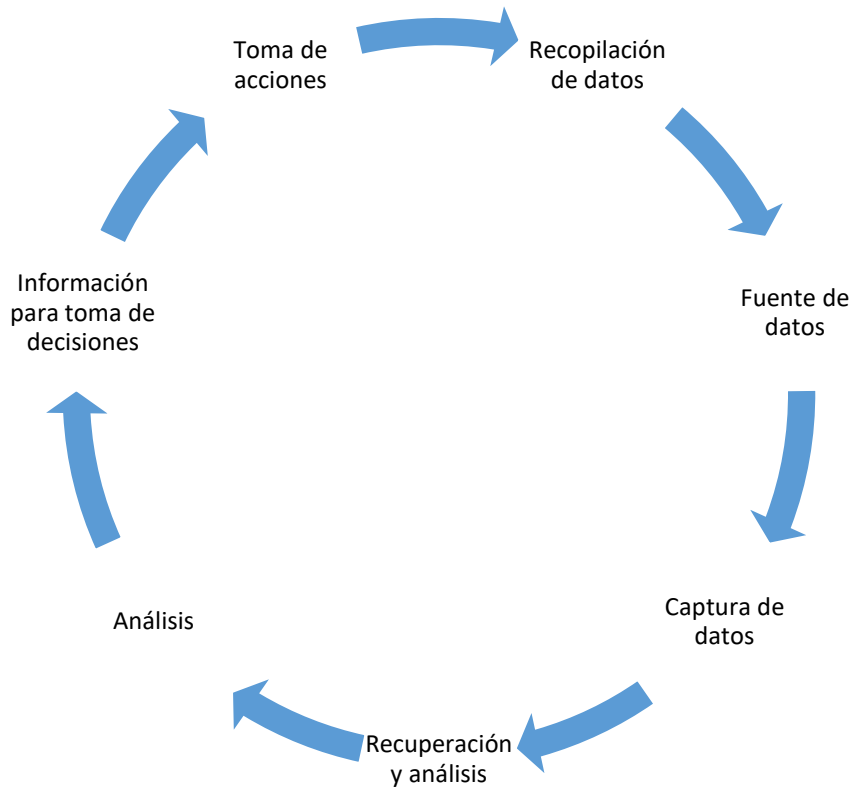


Ilustración 9: Ciclo de un SIG: Fuente: INEGI (2014)

Como se aprecia en la ilustración 9, el SIG forma parte de un ciclo donde, a partir de un problema real, se realiza la obtención de datos para la correcta toma de decisiones.

Es a través de mapas, imágenes, datos tabulares y textuales como se construye una base de datos para uso del SIG.

El SIG debe de ser capaz de resolver:

¿Qué hay? ¿Dónde se produce alguna circunstancia? ¿Qué cambios se han producido? ¿Qué modelo existe? ¿Qué pasaría sí? (Instituto Nacional de Estadística y Geografía, 2014). Uno de los principales usos de los SIG es el estudio de interrelaciones que mejoran el entendimiento sobre actividades y/o recursos de distribución. De esta manera, se puede combinar conjuntos aparentemente heterogéneos de datos.

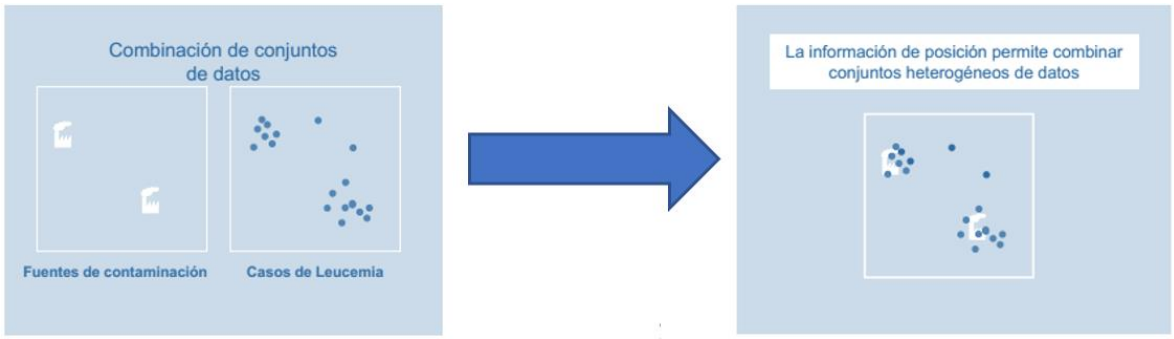


Ilustración 10: Reconocimiento de una estructura espacial. Fuente: INEGI (2014)

Sin embargo, uno de los principales problemas al momento de codificar los datos espaciales es la utilización de un código informático (código binario), es por ello que existen dos maneras de *simplificar* la realidad geográfica (Peña, s.f.). Para lo anterior, existen dos formatos de datos: vectorial y ráster:

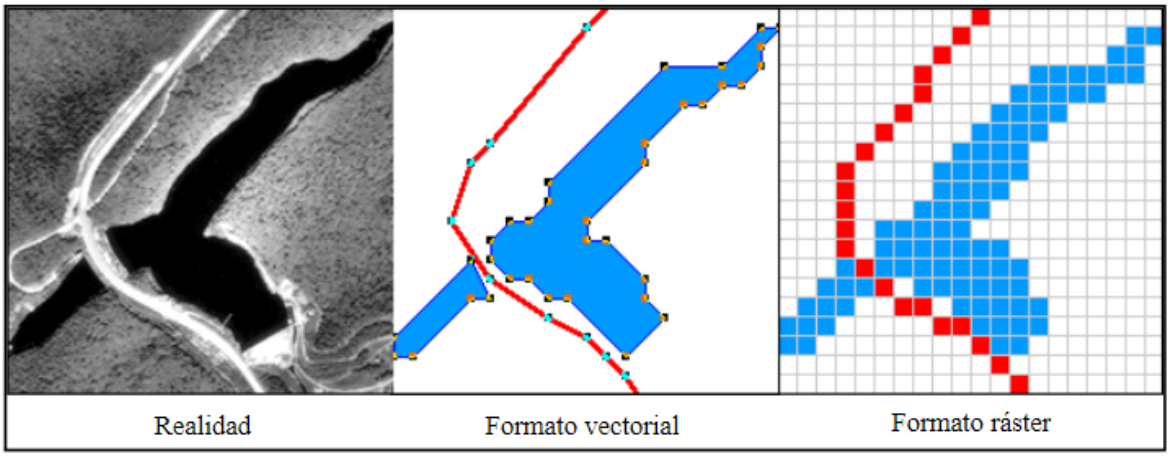


Ilustración 11: Formatos digitales de representación de datos geográficos. Fuente: (Peña, s.f.)

Como puede apreciarse en la ilustración 11, la estructura vectorial se caracteriza por el uso de puntos, líneas y polígonos que hacen que se replique la forma de una manera muy precisa. Este proceso es útil para la descripción de carreteras, terrenos, cuencas hidrológicas, limites administrativos, etc. (Peña, s.f.).

El método de estructuración se basa en el diseño de coordenadas (X, Y), mismas que forman *nodos* y que al unirse da la creación de polígonos.

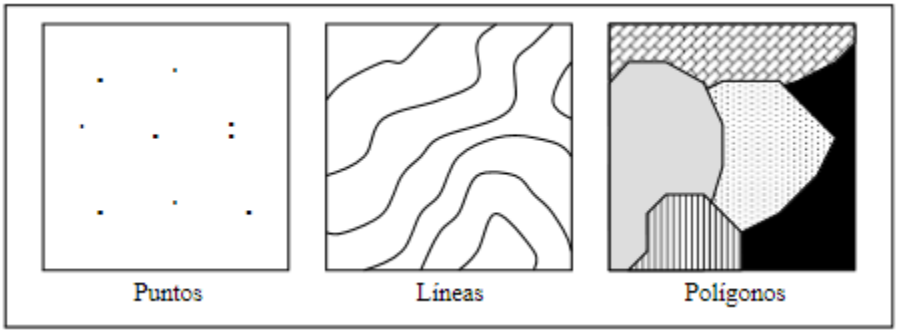


Ilustración 12: Proceso de creación de Polígonos. Fuente: (Peña, s.f.)

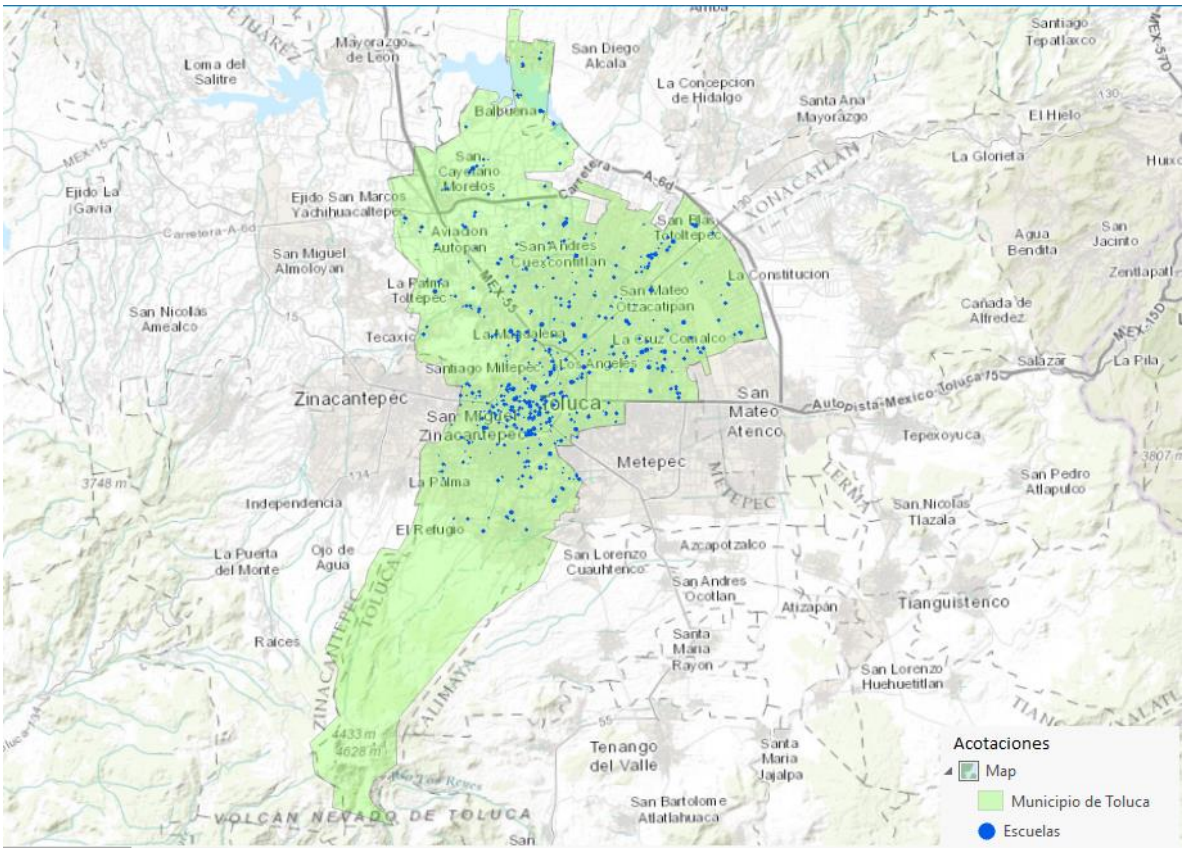
Por su parte, el proceso de modelización ráster, se realiza a través de celdas o píxeles, lo que ayuda a describir el espacio de una región mediante el número de rejillas (píxeles) que explican cierta característica de un terreno, tal como lo es la actitud, temperatura o precipitación, entre otros (Peña, s.f.).

A continuación se presenta un pequeño ejercicio del uso del SIG en un modelo real, se hace uso de los datos de vectores ya que, por su sencillez y facilidad permite una mejor interpretación de los datos a comparación del modelo ráster.

Supóngase que se desea conocer la distribución de las diversas escuelas que se ubican en el municipio de Toluca, Estado de México. Para ello, INEGI da a conocer la **Cartografía Geoestadística Urbana y Rural Amanzanada**, en este documento se trabajó con la versión a junio del año 2016. De igual manera se hizo uso del software ArcGIS Pro en su versión 1.4.0 así como del software libre QSIG en su versión 2.18.10.[10]

En el mapa 1, se puede observar la distribución de estas en el municipio de Toluca, a partir de esto, el alumno puede comenzar a elaborar sus teorías y/o hipótesis para su investigación al comenzar a relacionar distintas variables para su estudio.

[10] En el presente documento se trabajó exclusivamente con los archivos SHP, siendo que existen más variantes de archivos, se escogió este por su versatilidad y facilidad de uso.



Mapa 1: Distribución de las escuelas en el municipio de Toluca. Fuente: Elaboración propia con datos de INEGI.

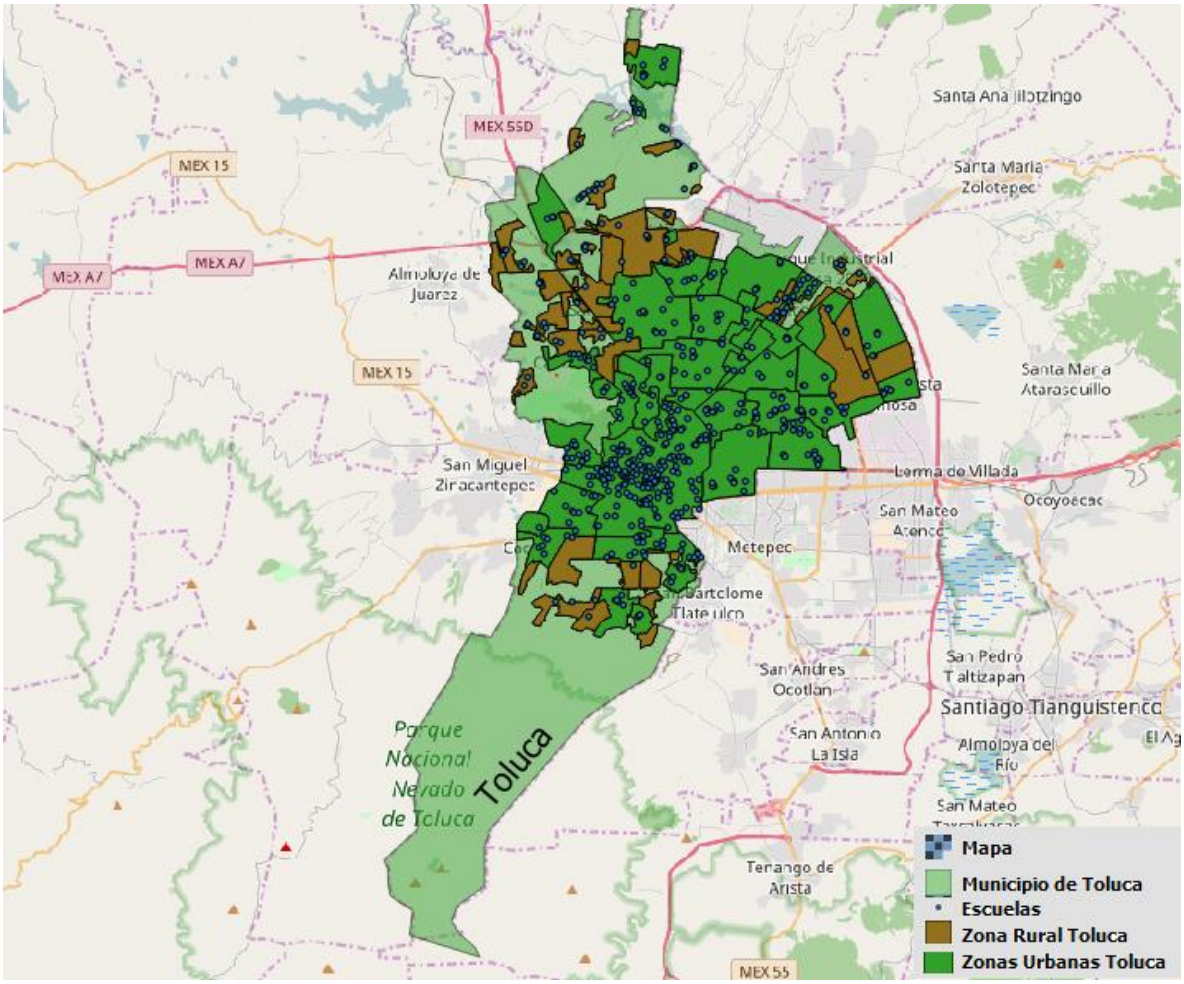
Nuevamente supongamos que, se desea encontrar la relación entre el número de escuelas y las zonas rurales que existen en la capital mexiquense. Un Sistema de Información Geográfica permitirá al investigador obtener una respuesta gráfica a su interrogativa. Así, se estructura el mapa 2, donde se distinguen las escuelas del municipio y las zonas rurales que existen en el mismo.

De aquí, pueden distinguirse varias situaciones:

- El INEGI, y el software utilizado, manejan datos vectoriales, que por su mejor nivel de detalle favorecen la interpretación de la distribución en un SIG;
- Aparentemente, pudiera existir una relación inversa entre zonas rurales y número de escuelas, puesto que la mayoría de estas se ubica dentro de zonas no rurales;
- Un SIG ofrece una descripción gráfica y sencilla de la situación actual de la realidad;
- Será cuestión del investigador, indagar las causas y consecuencias del fenómeno geográfico estudiado, siempre apoyado de las herramientas que el SIG puede ofrecer.

Ahora habrá que demostrar estadísticamente que, existe un menor número de escuelas en las zonas rurales del municipio de Toluca. El sistema arroja un total de 564 escuelas dentro

del municipio de Toluca, si se considera que la superficie total del municipio de Toluca es de $456.17\ km^2$ (COESPO, 2010), se puede decir que existen 1.23 escuelas por cada kilómetro cuadrado que conforma a la entidad.



Mapa 2: Distribución de las escuelas y zonas rurales en el municipio de Toluca. Fuente: Elaboración propia con datos de INEGI (2017)

De igual manera, a través del SIG es posible calcular el tamaño que ocupa el territorio rural y urbano de la entidad, este asciende a $239.84\ km^2$, por lo cual, el número de escuelas por kilómetro cuadrado de territorio ocupado asciende a 2.35.

Continuando con el análisis, se rescata que INEGI divide al municipio de Toluca en 96 zonas, 58 de ellas son consideradas zonas rurales y 38 zonas urbanas. El espacio ocupado por las zonas rurales asciende a la cifra de $59.97\ km^2$, mientras que las zonas urbanas cuentan con $179.88\ km^2$. Además, el SIG arroja la siguiente distribución de las escuelas en el municipio de Toluca: 72 escuelas ubicadas en los espacios rurales y 492 escuelas en los urbanos. Esto indica que, en el caso de las zonas rurales, existe 1.2 escuelas por cada kilómetro cuadrado, cifra muy inferior a las 2.733 escuelas por kilómetro cuadrado que existen en los espacios urbanos.

De esta manera se demuestra que los espacios urbanos si tienen una mayor concentración de escuelas que los espacios rurales, y mismo resultado debe ser atendido a la hora del diseño, elaboración y ejecución de la política económica.

De la misma manera en que se realizó el estudio entre la razón escuelas por kilómetro cuadrado de las zonas rurales y urbanas, el estudiante puede investigar acerca de rutas de movilidad, hospitales, zonas verdes, ríos, establecimientos gubernamentales, etc., y compararlos entre espacios, manzanas, zonas, municipios, estados e inclusive, a través del tiempo. En este sentido, el único límite de un SIG será el deseo del investigador.

V.- COMENTARIOS FINALES

Como se pudo demostrar en el ejercicio anterior, es posible realizar la ubicación, descripción y uso de la herramienta estadística dentro de un Sistema de Información Geográfica. La información que recopila el INEGI a través de los Censos Económicos, de Población y Vivienda, Agropecuarios, entre otros; permite al usuario realizar el análisis y la estructuración del SIG.

La idea del presente documento es mostrar al estudiante, que el uso de la estadística no es en lo absoluto ajeno al desarrollo regional, metropolitano y sustentable, sino más bien, debe verse a ésta como una herramienta útil y prácticamente ilimitada al momento de realizar un análisis territorial y económico.

El estudiante debe de hacer uso de estas herramientas para el propósito de su investigación, puesto que debe de tener el conocimiento para, a través de la estadística, poder sustentar sus proyectos, diseños de política, estructuras, investigaciones, propuestas de desarrollo, etc. El diseño de la política pública no debe ser algo trivial, esta, para su óptimo funcionamiento, debe estar siempre respaldada tras datos verídicos, investigaciones previas y análisis estadísticos que ciertamente, permitan dan fundamente al desarrollo de las políticas.

Con este documento, no se busca que el lector busque replicar el ejercicio o que únicamente lo contemple para su lectura, sino que se le invita que recopile todos los recursos necesarios (bases de datos, software, teoría económica, teoría de movilidad, teoría de sustentabilidad y sostenibilidad, entre otras) y busque nuevas alternativas de la alta gama de posibilidades que existen en los softwares y la amplia base de datos que el INEGI ofrece periódicamente.

REFERENCIAS

Andrade, M. I. & Klimsza, C., s.f. *Aplicación de un Sistema de Información Geográfica para una dirección de estadística municipal,* s.l.: s.n.

Creative Commons Attribution-ShareAlike 3.0, 2014. *QSIG2.2.* [En línea] Available at: http://docs.qgis.org/2.2/es/docs/training_manual/foreword/index.html [Último acceso: 1 Julio 2017].

García, M. B. & Carranco, Z., 2008. Concentración regional en Veracruz. Un enfoque de identificación de aglomeraciones productivas locales. *Análisis Económico,* pp. 291-310.

Gorgas, J., Cardiel, N. & Zamorano, J., 2011. *Estadística Básica para estudiantes de ciencias.* Madrid: Universidad Complutense de Madrid.

Gujarati, D. & Porter, D., 2010. *Econometría.* Quinta ed. México, D.F.: Mc Graw Hill.

Instituto Nacional de Estadística y Geografía, 2014. *Sistemas de Información Geográfica,* s.l.: s.n.

Mendenhall, W., Beaver, R. & Beaver, B., 2010. *Introducción a la probabilidad y estadística.* México, D.F.: Cengage Learning Editores.

Peña, J., s.f. *Sistemas de Información Geográfica aplicados a la gestión del territorio,* s.l.: s.n.

Tusell, F., 2011. *Análisis de Regresión. Introducción Teórica y Práctica basada en R,* Bilbao: s.n.

Vela, F., 2010. *Normalidad en los errores,* México, D.F.: Universidad Autónoma Metropolitana.

ANEXOS

Anexo 1: Muestra de 25 municipios

Clave y nombre del municipio	*Proporción* $\frac{Empleo_{Industrias\ Alimenticias,j}}{Empleo_{Industrias\ manufactureras,j}}$
004 Almoloya de Alquisiras	0.8046
008 Amatepec	0.6312
011 Atenco	0.2377
014 Atlacomulco	0.0789
017 Ayapango	0.3636
022 Cocotitlán	0.6000
024 Cuautitlán	0.1649
028 Chiautla	0.2911
032 Donato Guerra	0.8378
038 Isidro Fabela	0.7674
047 Jiquipilco	0.7533
049 Joquicingo	0.6645
055 Mexicaltzingo	0.6552
057 Naucalpan de Juárez	0.0733
063 Ocuilan	0.5789
068 Ozumba	0.4235
071 Polotitlán	0.2416
077 San Simón de Guerrero	0.4118
081 Tecámac	0.2700
088 Tenancingo	0.6009
093 Tepetlaoxtoc	0.1553
096 Tequixquiac	0.4522
101 Tianguistenco	0.0479
107 Tonatico	0.5274
117 Zacualpan	0.7237
Media	0.4543
Desviación estándar	0.2498

Anexo 1.1 Muestra de 35 municipios

Clave y nombre del municipio	*Proporción Empleo Ind. Alim./empleo Ind. Man.*
004 Almoloya de Alquisiras	0.8046
005 Almoloya de Juárez	0.2772
008 Amatepec	0.6312
010 Apaxco	0.1775
011 Atenco	0.2377
014 Atlacomulco	0.0789
017 Ayapango	0.3636
020 Coacalco de Berriozábal	0.3263
022 Cocotitlán	0.6000
024 Cuautitlán	0.1649
026 Chapa de Mota	0.2134
028 Chiautla	0.2911
032 Donato Guerra	0.8378
035 Huehuetoca	0.0703
038 Isidro Fabela	0.7674
040 Ixtapan de la Sal	0.4885
042 Ixtlahuaca	0.1795
045 Jilotepec	0.0318
047 Jiquipilco	0.7533
049 Joquicingo	0.6645
055 Mexicaltzingo	0.6552
057 Naucalpan de Juárez	0.0733
063 Ocuilan	0.5789
068 Ozumba	0.4235
071 Polotitlán	0.2416
076 San Mateo Atenco	0.2246
077 San Simón de Guerrero	0.4118
081 Tecámac	0.2700
086 Temascaltepec	0.7091
088 Tenancingo	0.6009
093 Tepetlaoxtoc	0.1553
096 Tequixquiac	0.4522
101 Tianguistenco	0.0479
107 Tonatico	0.5274
117 Zacualpan	0.7237
Media	0.4016
Desviación estándar	0.2487

Anexo 2.- Datos regresión municipios Estado de México

Clave y municipio	Población empleada	Producción bruta total (millones de pesos)	Clave y municipio	Población empleada	Producción bruta total (millones de pesos)	Clave y municipio	Población empleada	Producción bruta total (millones de pesos)	Clave y municipio	Población empleada	Producción bruta total (millones de pesos)
001 Acambay	2462	232.439	**032 Donato Guerra**	600	32.415	**063 Ocuilan**	1662	133.711	**094 Tepetlixpa**	1656	225.618
002 Acolman	12024	6844.931	**033 Ecatepec de Morelos**	204423	92921.498	**064 El Oro**	1468	173.043	**095 Tepotzotlán**	22025	19909.646
003 Aculco	2166	543.418	**034 Ecatzingo**	458	18.500	**065 Otumba**	2877	297.141	**096 Tequixquiac**	3650	506.740
004 Almoloya de Alquisiras	699	67.314	**035 Huehuetoca**	11294	5634.770	**066 Otzoloapan**	181	12.765	**097 Texcaltitlán**	707	107.055
005 Almoloya de Juárez	4928	1411.807	**036 Hueypoxtla**	2747	176.392	**067 Otzolotepec**	3751	246.382	**098 Texcalyacac**	499	35.089
006 Almoloya del Río	1597	79.244	**037 Huixquilucan**	30920	15615.166	**068 Ozumba**	3518	346.187	**099 Texcoco**	32835	9642.821
007 Amanalco	677	36.749	**038 Isidro Fabela**	570	42.924	**069 Papalotla**	478	123.714	**100 Tezoyuca**	3895	899.874

008 Amatepec	1146	92.480		**039 Ixtapaluca**	39698	13688.543		**070 La Paz**	33287	20481.773		**101 Tianguistenco**	16331	20373.695
009 Amecameca	4951	617.347		**040 Ixtapan de la Sal**	4821	760.682		**071 Polotitlán**	1248	5157.299		**102 Timilpan**	761	70.317
010 Apaxco	3907	3082.708		**041 Ixtapan del Oro**	150	12.002		**072 Rayón**	1592	474.911		**103 Tlalmanalco**	3145	380.691
011 Atenco	3806	639.360		**042 Ixtlahuaca**	15314	3565.958		**073 San Antonio la Isla**	3268	3289.703		**104 Tlalnepantla de Baz**	187106	118560.624
012 Atizapán	2093	396.757		**043 Xalatlaco**	1769	208.010		**074 San Felipe del Progreso**	4675	290.964		**105 Tlatlaya**	465	25.895
013 Atizapán de Zaragoza	59186	22824.340		**044 Jaltenco**	1642	150.725		**075 San Martín de las Pirámides**	2385	537.092		**106 Toluca**	221323	232585.008
014 Atlacomulco	22305	17110.258		**045 Jilotepec**	12241	3383.086		**076 San Mateo Atenco**	11077	1416.581		**107 Tonatico**	1414	107.816
015 Atlautla	1930	119.746		**046 Jilotzingo**	702	85.819		**077 San Simón de Guerrero**	133	19.024		**108 Tultepec**	10529	2828.171
016 Axapusco	1586	169.020		**047 Jiquipilco**	1071	65.049		**078 Santo Tomás**	199	18.600		**109 Tultitlán**	73709	61189.149
017 Ayapango	198	14.131		**048 Jocotitlán**	12995	8170.926		**079 Soyaniquilpan de Juárez**	1126	228.345		**110 Valle de Bravo**	8893	1831.107
018 Calimaya	3101	277.047		**049 Joquicingo**	831	48.146		**080 Sultepec**	530	36.385		**111 Villa de Allende**	900	182.061

019 Capulhuac	4946	1344.218		**050 Juchitepec**	1554	180.810		**081 Tecámac**	37720	9010.933		**112 Villa del Carbón**	2138	236.848
020 Coacalco de Berriozábal	28946	6845.425		**051 Lerma**	56138	60226.160		**082 Tejupilco**	5812	929.198		**113 Villa Guerrero**	2859	398.210
021 Coatepec Harinas	1860	252.153		**052 Malinalco**	3236	253.318		**083 Temamatla**	671	54.787		**114 Villa Victoria**	2703	322.741
022 Cocotitlán	755	51.871		**053 Melchor Ocampo**	3861	552.636		**084 Temascalapa**	2389	167.355		**115 Xonacatlán**	4535	609.080
023 Coyotepec	3152	318.451		**054 Metepec**	51627	14847.027		**085 Temascalcingo**	4422	789.908		**116 Zacazonapan**	890	2118.294
024 Cuautitlán	24625	26961.021		**055 Mexicaltzingo**	1151	140.112		**086 Temascaltepec**	836	346.366		**117 Zacualpan**	844	228.857
025 Chalco	34797	8068.459		**056 Morelos**	2470	452.861		**087 Temoaya**	4204	379.136		**118 Zinacantepec**	13481	2819.152
026 Chapa de Mota	1037	68.007		**057 Naucalpan de Juárez**	177388	98018.100		**088 Tenancingo**	9333	1081.865		**119 Zumpahuacán**	412	39.704
027 Chapultepec	901	98.244		**058 Nezahualcóyotl**	105392	18612.896		**089 Tenango del Aire**	728	56.173		**120 Zumpango**	14413	2089.195
028 Chiautla	2688	343.529		**059 Nextlalpan**	1948	187.741		**090 Tenango del Valle**	8643	2052.660		**121 Cuautitlán Izcalli**	115175	124755.464

029 Chicoloapan	12543	2307.593		**060 Nicolás Romero**	27704	4295.229		**091 Teoloyucan**	6055	1146.149		**122 Valle de Chalco Solidaridad**	30241	5423.741
030 Chiconcuac	4484	352.942		**061 Nopaltepec**	770	115.322		**092 Teotihuacán**	6785	3240.402		**123 Luvianos**	1468	108.019
031 Chimalhuacán	44508	3786.748		**062 Ocoyoacac**	11214	6920.040		**093 Tepetlaoxtoc**	1360	236.523		**124 San José del Rincón**	1119	67.077
												125 Tonanitla	600	41.945