Ministerio de Hacienda
Dirección General del Presupuesto

PLAN DE CAPACITACIÓN Y ASISTENCIA TÉCNICA PARA LA REFORMA PRESUPUESTARIA

Módulo III
ACTUALIZACIÓN DE CONOCIMIENTOS GENERALES

Curso 3

Conceptos y Métodos Básicos de Estadística

MANUAL DEL PARTICIPANTE

[Edición 9 de abril 2013]

"Más y mejores servicios públicos con equilibrio y sostenibilidad fiscal"

Elaborado por : Rebeca Elizabeth Ramos
Revisado por : Dr. José Nerys Funes Torres

Integrantes del Equipo de Instrucción:
DR. JOSÉ NERYS FUNES TORRES
INGRID CAROLINA MARTÍNEZ
JOSÉ DAVID ESCOBAR MUÑOZ
HÉCTOR DOUGLAS MOLINA GUEVARA
DANIEL ALEJANDRO RIVAS RIVAS
REBECA ELIZABETH RAMOS

Dirección y Supervisión del Plan de Capacitación y Asistencia Técnica: Roger DÍAZ, Consultor GIZ.

Contenido

I. ASPECTOS GENERALES

El curso desarrolla los métodos básicos de la estadística descriptiva y la estadística inferencial para calcular medidas de tendencia central y dispersión e interpretarlas comparativamente, y establecer criterios para realizar una prueba de hipótesis de diferentes parámetros. El curso también orienta sobre las principales fuentes de datos oficiales, tales como censos, encuestas, bases de datos especializadas, así como sobre las principales herramientas de búsqueda de información en la web.

1.1 Objetivo general

Proveer a los participantes de las herramientas básicas de análisis tanto al momento de elaborar diagnósticos de la problemática como punto de partida para diseñar programas presupuestarios, así como durante la ejecución, seguimiento y evaluación de programas.

1.2 Requisitos

- Manual del participante (el alumno debe haber recibido y leído con anticipación el contenido del manual)
- Presentación en Power Point a cargo del instructor
- Manual del instructor el curso
- Proyector Multimedia, Pizarra acrílica, Plumones

II. DESARROLLO

LECCION 1: Consideraciones Generales de la Estadística

Resumen de la lección: Este curso desarrolla los conceptos generales de la estadística como: la estadística descriptiva e inferencial, criterios generales que se deben tener en cuenta para realizar un análisis estadístico, así como la utilización de los diversos tipos de variables a utilizar.
Tiempo total: 30 minutos

1.1 Objetivos de la Lección

Al finalizar esta lección los participantes serán capaces de:

a. Diferenciar el alcance de la estadística Descriptiva e Inferencial
b. Conocer los criterios básicos para realizar análisis estadístico
c. Diferenciar los diversos tipos de variables que se utilizan para el análisis de datos

1.2 Desarrollo del contenido de la Lección

Definiciones clave

- Concepto de Estadística
- Clasificación de la Estadística
- Tipos de variables estadísticas

Punto 1: Clasificación de la Estadística

La estadística es una ciencia que estudia la recolección, análisis e interpretación de datos, ya sea para ayudar en la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado. Sin embargo, la estadística es más que eso, en otras palabras es el vehículo que permite llevar a cabo el proceso relacionado con la investigación científica.

La Estadística tiene por objeto recolectar, organizar, resumir, presentar y analizar datos relativos a un conjunto de objetos, personas, procesos, etc. A través de la cuantificación y el ordenamiento de los datos intenta explicar los fenómenos observados, por lo que resulta una herramienta de suma utilidad para la toma de decisiones.

Bajo este contexto, la Estadística se divide en dos áreas**: Estadística Descriptiva y Estadística Inferencial.**

Estadística Descriptiva: permite organizar y presentar un conjunto de datos de manera que describan en forma precisa las variables analizadas facilitando su lectura e interpretación. Obviamente, la materia prima de la Estadística Descriptiva la constituyen los datos.

Ejemplo: Se realiza una encuesta en el departamento de San Salvador para para deducir las preferencias de votos para la próximas elecciones presidenciales, y se concluye que el 30% votara por el partido XXF, el 45% por el partido XXK y el resto equivalente al 25% por el XXL.

Sin embargo, cuando existen limitantes (económicas, tiempo, recurso humano, entre otras) en las cuales es imposible obtener información de toda la población, se toma una parte de ella a la que se le llama **"Muestra"** el cual es definido como un subconjunto representativo de una población.

La Estadística Inferencial: permite generalizar los resultados (datos estadísticos) de una muestra a la población total, es decir se realizan conclusiones o inferencias, basándose en los datos simplificados y analizados en una muestra.

Ejemplo: En línea con el ejemplo anterior la intención de voto de los habitantes del departamento de San Salvador se pudiese obtener a partir de una muestra representativa tomada de los habitantes de los diferentes municipios del depto. y a partir de dichos resultado se infieren las intenciones de voto de la población, en este caso los habitantes del dpto. de San Salvador.

Punto 2: Etapas o pasos mínimos necesarios para realizar un análisis Estadístico

La estadística suministra valores que ayudan a descubrir interrelaciones entre múltiples parámetros, así como también es una herramienta útil para realizar predicciones de valores futuros.

Un análisis estadístico es un proceso sistemático en el cual se requiere realizar una serie de etapas con el mayor grado de acuciosidad posible, estas se describen a continuación:

1. Definición del problema de estudio y objetivos del mismo.
2. Selección de la información necesaria para realizar el estudio.
3. Recogida de la información que va a depender del presupuesto con el que contemos y de la calidad de los datos exigida.
4. Ordenación y clasificación de la información en tablas y gráficos.
5. Resumen de los datos mediante medidas de posición, dispersión, entre otras que se consideren convenientes.
6. Análisis estadístico formal obteniendo hipótesis y predicciones.
7. Interpretación de resultados y extracción de conclusiones.

Punto 3: Tipos de Variables

Las variables estadísticas son características que poseen todas las unidades de la población o muestra a estudiar.

Las variables que se observan y analizan pueden ser de dos tipos:

a) **Variables cualitativas o atributos:** no se pueden medir numéricamente, representan características o atributos de las variables (por ejemplo: nacionalidad, sexo, religión).

   a.1 Nominal: Característica o cualidad cuyas categorías no tienen un orden preestablecido. Ejemplos: Sexo, Deporte Favorito, etc.

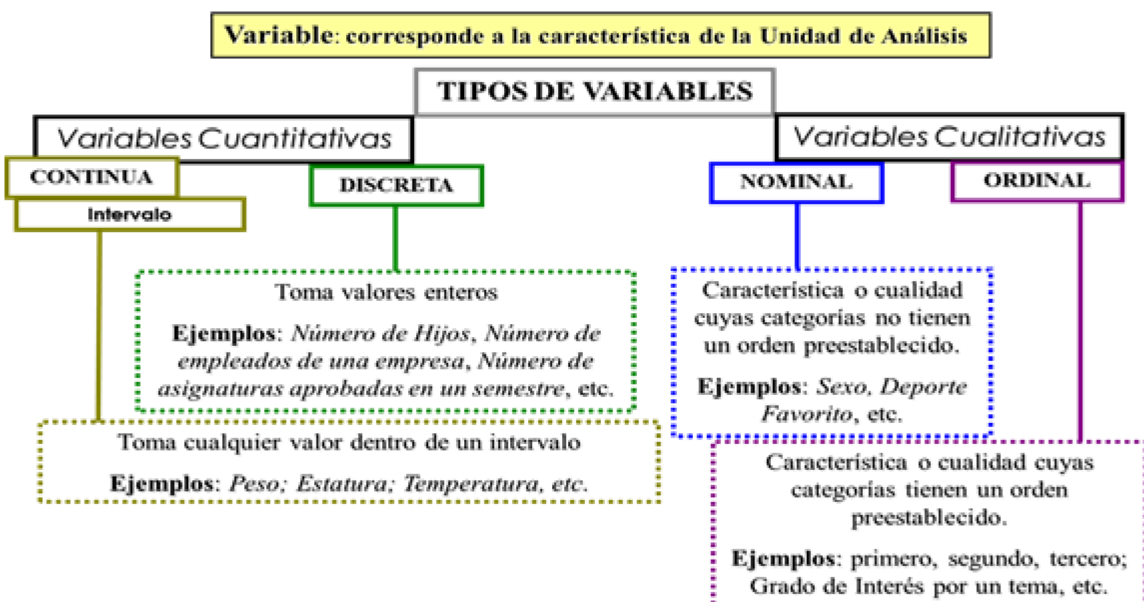
   a.2 Ordinal: Característica o cualidad cuyas categorías tienen un orden preestablecido.
   Ejemplos: primero, segundo, tercero; Grado de Interés por un tema, etc.

b) **Variables cuantitativas:** toman valores numéricos (edad, altura, precio de un producto, ingresos anual es).

Por su parte, las variables cuantitativas se pueden clasificar atendiendo a los valores que pueden tomar en discretas y continuas:

b.1) Discretas: sólo pueden tomar valores enteros (1, 2, 8, -4, etc.). Por ejemplo: número de hermanos (puede ser 1, 2, 3...., etc., pero, por ejemplo, nunca podrá ser 3.45).

b.2) Continuas: Son las variables que pueden adquirir cualquier valor dentro de un intervalo especificado de valores. Por ejemplo la masa (2.3 kg, 2.4 kg, 2.5 kg,...) o la altura (1.64 m, 1.65 m, 1.66 m,...), o el salario. Solamente se está limitado por la precisión del aparato medidor, en teoría permiten que siempre exista un valor entre dos valores numéricos.



Discusión de los contenidos: Lección 1.

1. Contestar verdadero o falso y comentar su respuestas según sea el caso:
   a) La Estadística es una ciencia que estudia y describe las características de un conjunto de casos. V___ F___
   b) La estadística inferencial generaliza los resultados de una muestra a los de la población total. V___ F___
   c) Durante los últimos dos días se ha informado de un total de cinco homicidios diarios en San Salvador, este es un ejemplo de estadística inferencial.

V___ F___

2. Establecer las diferencias entre variables cualitativas y cuantitativas.
3. Establecer las diferencias entre variables discretas y continuas.
4. Definir al menos dos ejemplos de variables discretas y de variables continuas.
5. ¿De qué tipo de escala son las siguientes variables, respectivamente?
   a. Niños, Jóvenes, Adultos.
   b. Ingeniería Eléctrica, Mecánica, Química, de Sistemas.

De las cuatro alternativas siguientes, subraye la respuesta correcta:

A. Cualitativa y ordinal.
B. Ordinal y nominal.
C. Ambas cualitativas.
D. Ambas ordinales.

6. De las siguientes **variables** indica cuáles son **discretas, continuas o cualitativas.**

a) Tiempo de duración de un automóvil.
b) Número de hijos de 50 familias.
c) Número de empleados en la procuraduría.
d) Número de goles marcados por tu equipo favorito en la última temporada.
e) Número de médicos por ley de salario en el Hospital Rosales.

1.4 Conclusiones/Ideas fuerza a tener presente

1. La estadística es la ciencia que facilita, el análisis, presentación e interpretación de datos.
2. La diferencia entre la estadística descriptiva y la inferencial, es más que nada el universo de acción, la estadística descriptiva trabaja con datos poblacionales la inferencia con una muestra de esa población y a través de esa información infiere los datos poblacionales.
3. Las diferencias entre las variables cualitativas y cuantitativas es que las primeras se refieren a atributos y las segundas reflejan un valor numérico.
4. Las variables cuantitativas a su vez pueden ser de dos tipos: discretas, cuando toman un número entero y continuas cuando su valor puede oscilar entre dos números.

1.5 Bibliografía

1. Gildaberto Bonilla, Estadística I, Elementos de Estadística Descriptiva y Probabilidad.
2. Anderson Sweeney Williams, Estadistica para Administración y Economia
3. http://www.eumed.net/cursecon/libreria/drm/ped-drm-est.htm
4. http://www.vitutor.com/estadistica/descriptiva/a_1.html

LECCION 2: Tabulación y Gráficos

Resumen de la lección: Entre los principales elementos que contempla esta lección se encuentra la adopción de herramientas que contribuyan a presentar los datos de una manera más ordenada y que facilite la comprensión del lector ya sea a través de tablas de frecuencias o de forma gráfica.

Tiempo total: 1 horas

2.1 Objetivos de la Lección 2

Al finalizar esta lección los participantes serán capaces de:

a. Obtener criterios para presentar de una forma más ordena la información utilizando tablas de frecuencias.
b. Conocer la aplicabilidad de las tablas de frecuencias simples y tablas de frecuencias para datos agrupados.
c. Adquirir elementos para utilizar la representación gráfica de la manera más apropiada considerando entre otros aspectos, el tipo de variable, escala de medición, entre otras.

2.2 Desarrollo del contenido de la Lección 2

Definiciones clave

Tabla de frecuencia simple, frecuencia absoluta, relativa y acumulada, rango, tabla de frecuencia con datos agrupados en clases, intervalo de clases, ancho de clase, marca de clase o punto medio, principales tipos de gráficos.

Punto 1: Tabulación de la Información

La distribución de frecuencias o tabla de frecuencias es una ordenación en forma de tabla de los datos estadísticos, asignando a cada dato su frecuencia correspondiente. Para entender cómo funcionan las tablas de frecuencia, analicemos el siguiente ejemplo:

Ejemplo 1.: Suponga que se ha preguntado a 37 familias sobre el número de hijos. La forma de simplificar los datos, equivale a contar cuantas familias tienen el mismo número de hijos. A esta operación la conoceremos como "frecuencia absoluta".

N. de Hijos	N. de Familias
0	4
1	7
2	8
3	5
4	10
5 o más	3
Total	**37**

Se observa que 7 familias tienen solamente un hijo. También, se puede tener interés en conocer el porcentaje de familias que tienen 3 ó menos, para ello se requiere sumar el número de familias que tienen: 0, 1, 2 o 3 hijos y dividirlo entre el total de familias, esto es, ((4+7+8+5)/37)*100 = 64.86%, este valor se interpreta que aproximadamente el 65% de las familias tienen 3 ó menos hijos.

Frecuencia Absoluta (fi): Número de veces que se repite un valor dentro de un conjunto de datos.

Podemos identificar dos tipos de tablas de frecuencias, las cuales denotaremos como tablas de frecuencia simple y tablas de frecuencias con datos agrupados.

1.1 Tabla de frecuencia simple

Se caracterizan por manejar un conjunto pequeño de posibles resultados de una variable dentro de la muestra o población. Por lo general, su uso tiende al manejo de datos cualitativos o variables cuantitativas discretas.

Ejemplo 2. El gobierno está interesado en medir el grado de aceptación que tendría si se construyera una carretera cerca de cierta comunidad, para ello, se selecciona una muestra de 10 personas del sector en que habitan. Para tal fin, se les pide que valoren dicho proyecto empleando una escala del 1 al 5, su opinión sobre dicho proyecto **(1 = Muy Malo, 2= Malo, 3 = Regular, 4 = Bueno y 5 = Excelente)**. Las respuestas tabuladas de las 10 personas son:

Persona	Respuesta (Grado de aceptación)
1	2
2	5
3	4
4	5
5	4
6	3
7	4
8	5
9	3
10	5

SOLUCIÓN

Como se puede observar, el número de resultados que puede alcanzar la variable "grado de aceptación" son relativamente pocos (solo cinco posibilidades), lo cual sirve como parámetro para catalogar la tabla de frecuencia resultante como "tabla de frecuencia simple".

Otra forma de catalogar los datos es conociendo la distancia o variación que hay entre el valor menor (Xmin) y el valor mayor (Xmax), diferencia que se conoce como “Rango”.

Rango (R): Diferencia existente entre el valor Máximo (X_{max}) y el valor Mínimo (X_{min}) de un conjunto de datos. La fórmula empleada es: $R= X_{max} - X_{min}$

En el ejemplo R sería igual a 4 puesto que:
R= 5 -1 = 4
Si el rango manejado es pequeño, bastara representar los datos con una tabla de frecuencia simple. Para crear esta tabla deberemos seguir los siguientes pasos:
PASO 1: Contar las veces que se repite cada valor dentro de la muestra.
PASO 2: Ubicar estas frecuencias en una tabla ordenada.

Grado de Aceptación	Frecuencia (fi)
1	0
2	1
3	2
4	3
5	4
Total	**10**

Ninguna de las personas valoró el proyecto de construcción de carretera como muy malo (grado de aceptación igual a 1), mientras que la mayoría de las respuestas se centraron en Excelente y Bueno (grado de aceptación iguales a 5 y 4 respectivamente).

Observando los datos resumidos, podemos concluir que la mayoría de las personas encuestadas tienen una visión favorable del proyecto de construcción de la carretera. En este ejemplo se visualiza, como la tabla de frecuencia agiliza el análisis de los datos.

La estadística considera otros tipos de frecuencias auxiliares que complementan el análisis de las tablas de frecuencia.

Frecuencia Absoluta Acumulada (Fa): Presenta un saldo acumulado de las frecuencias de los intervalos. Esta frecuencia se calcula sumando el acumulado de las frecuencias de los intervalos anteriores más la frecuencia absoluta del intervalo actual. La última frecuencia absoluta acumulada deberá ser igual a *N*.

$$\mathbf{FA= Fa_{-1} + f_i}$$

La Frecuencia Absoluta Acumulada del último intervalo es igual al tamaño de la muestra (o población). Siguiendo con el ejemplo, tenemos:

- Grado de aceptación 1: $Fa_1 = 0$
- Grado de aceptación 2: $Fa_2 = 0 + 1 = 1$
- Grado de aceptación 3: $Fa_3 = 1 + 2 = 3$
- Grado de aceptación 4: $Fa_4 = 3 + 3 = 6$
- Grado de aceptación 5: $Fa_5 = 6 + 4 = 10$

Esta frecuencia nos proporciona de inmediato el número de casos que queda por debajo de cada clase. La Fa_4, por ejemplo, nos dice que seis personas opinaron que la ejecución del proyecto de construcción de carretera se encontraba entre muy malo y bueno.

Frecuencia Relativa (h): Equivale a la razón de las frecuencias de cada intervalo sobre la totalidad de los datos.
Matemáticamente se expresa: $h_i = f_i / n$

Para el ejemplo, las frecuencias relativas son:

Grado de aceptación 1: $h_1 = \frac{0}{10} = 0$

Grado de aceptación 2: $h_2 = \frac{1}{10} = 0.1 \; ó \; 10\%$

Grado de aceptación 3: $h_3 = \frac{2}{10} = 0.2 \; ó \; 20\%$

Grado de aceptación 4: $h_4 = \frac{3}{10} = 0.3 \; ó \; 30\%$

Grado de aceptación 5: $h_5 = \frac{4}{10} = 0.4 \; ó \; 40\%$

La sumatoria de las frecuencias relativas debe ser igual a 1 (si se trabaja estos valores como porcentaje, equivaldría al 100% de los datos). El 40% de las personas encuestadas (h_5), opinaron que al realizar el proyecto de construcción sería excelente.

Frecuencia Relativa Acumulada (H_i): Presenta un saldo acumulado de las frecuencias relativas de cada intervalo de clase. Su cálculo resulta de la suma del acumulado de las frecuencias relativas de los intervalos anteriores más la frecuencia relativa del intervalo actual. $\mathbf{H_i = H_{i-1} + h_i}$

La última de las Frecuencias Relativas Acumuladas equivale a 1. Las tablas de frecuencias suelen mostrar tanto las frecuencias absolutas, como relativas.

Grado de Aceptación (Clase)	fi	Fa	h_i	H_i
1	0	0	0,0	0,0
2	1	1	0,1	0,1
3	2	3	0,2	0,3
4	3	6	0,3	0,6
5	4	10	0,4	1,0
TOTAL	**10**		**1,0**	

Características de las Tabla de frecuencia simple

- ✓ El número de posibles valores que toma la variable debe ser reducido (Rango pequeño).
- ✓ Suele ser utilizada en la cuantificación de las variables cualitativas y cuantitativas discretas.
- ✓ Su construcción es sencilla y equivale a especificar la frecuencia de cada resultado.

Ejercicio Propuesto

Por ejemplo, se quiere saber la cantidad de Población Económicamente Activa (PEA) en 50 familias de la zona rural, con el propósito de terminar que familias serán las beneficiarias de un programa de ayuda social, se ha estimado que cada persona económicamente activa aporta a la familia un promedio de $125.

Como analista del programa será necesario saber:
* ¿Cuál es el rango de PEA en los 50 hogares?
* ¿Cuál es la frecuencia de PEA en las 50 familias?
* ¿Cuál es la frecuencia relativa de la PEA en los 50 familias?
* ¿Cuál es el porcentaje de familias con 3 o más personas económicamente activas?
* Realizar el análisis correspondiente de acuerdo a las respuestas de las interrogantes antes señaladas.

Número de PEA en 50 hogares

2	1	2	2	1	2	4	2	1	1
2	3	2	1	1	1	3	4	2	2
2	2	1	2	1	1	1	3	2	2
3	2	3	1	2	4	2	1	4	1
1	3	4	3	2	2	2	1	3	3

1.2 Tabla de Frecuencia con Datos Agrupados en Clases

Cuando el tamaño de la población y/o muestra y el rango de la variable son grandes, será necesario agrupar dichos valores en "**Intervalos de Clases".**

Por ejemplo, en el caso de contar con una valoración del 1 al 100 (un rango equivalente a 99), la tabla de frecuencia simple se encargaría de buscar cuantas veces se repite cada uno de los 99 posibles resultados en un conjunto de datos, teniendo una función contraria a la de resumir los datos.

Agrupar los valores de la variable en intervalos podría simplificar estas fuentes de datos. Por ejemplo, podríamos hablar de las frecuencias para los valores comprendidos entre 0-20, 20-40, 40-60, 60-80 y 80-100.

En el intervalo 0-20 (que de ahora en adelante le llamaremos **intervalo de clase**), se sumaran las frecuencias de los datos cuyos resultados estén entre 0 y 20.

Intervalo de clase: Intervalos empleados en las Tablas de Frecuencias Estadísticas, capaz de contener diversas medidas de una variable. Consta de un límite inferior (Lm) y un límite superior (Ls).

Otro punto importante que el estadista debe definir, es la cantidad de intervalos de clase que empleará en la tabla de frecuencia. Esta cantidad de intervalos no debería ser grande, debido a que no se cumpliría el objetivo de resumir la información, y no tan pocos intervalos, ya que se perdería mucha información.

No existe una fórmula, ni unos principios únicos para establecer el número de intervalos. Para nuestro propósito, optaremos por manejar un número de intervalos convenientes entre 5 y 15.

Algunos autores han propuestos fórmulas que permiten ayudar en la tarea de conseguir el número ideal de intervalos.

Numero de intervalos (Nc): Cantidad de intervalos con los cuales se compone una tabla de frecuencia.

La primera, la más conocida, establece el número de intervalos al obtener la raíz cuadrada del total de elementos considerados en el estudio.

$$\mathbf{Nc} = \sqrt{\mathbf{n}}$$

Cuando se trabajan con muestras mayores a 225, con la fórmula se obtiene un Nc superior a 15, por tanto, recomendaremos para estos casos la siguiente fórmula: **1 + 3,22 log n**

Si en ambas fórmulas obtenemos un Nc mayor a 15, simplemente tomaremos 15 intervalos. El estadista podrá omitir los resultados de las fórmulas y conseguirá seleccionar el número de intervalos que crea que son los más adecuados, de acuerdo al objeto de estudio o las características que desea mostrar de la variable.

Cada intervalo posee un número máximo de resultados que puede agrupar. A este valor lo conoceremos como el “Ancho del Intervalo de Clase (A)”.

Ancho del intervalo de Clase (A): Equivale a la diferencia entre el Límite superior (Ls) y el Límite inferior (Lm) de cada intervalo. Matemáticamente se expresa:

A = Ls − Lm

Su cálculo resulta de la división del Rango (R) entre el Número de Intervalos (Nc)

A = R / Nc

Ejemplo con Datos de ingresos de 24 familias.

Variable: Ingresos semanales en US$ por familia, n = 24 datos.

1,450	1,443	1,536	1,394	1,623	1,650
1,480	1,355	1,350	1,430	1,520	1,550
1,425	1,360	1,430	1,450	1,680	1,540
1,304	1,260	1,328	1,304	1,360	1,600

Antes de elaborar la tabla de frecuencia, debemos definir cuál de los dos tipos propuestos es el que mejor se adapta (tabla de frecuencia simple o tabla de frecuencia con datos agrupados en clases)

Si resumimos los datos en una tabla de frecuencia simple, tendríamos una tabla muy extensa, en la cual algunas frecuencias de los ingresos de las familias serian 0. Esto se debe a que el rango manejado es muy amplio (R =$1,680 – $1,260 = 420).

Ingresos	Frecuencia (fi)
1,260	1
1,304	2
1,328	1
1,350	1
1,355	1
1,360	2
1,394	1
1,425	1
1,430	2
,,,,,	……
1,680	1
TOTAL	**24**

En el caso que queramos agrupar aún más estos datos, trabajaríamos con el concepto de intervalos de clase (tabla de frecuencia con datos agrupados en clases).

Paso 1: Determinar el número de intervalos (Nc)

Optaremos por utilizar la primera fórmula expuesta: $Nc = \sqrt{n}$

Nc = √24 = 4.898 ≈ 5

Se debe siempre aproximar el número de intervalos al entero más próximo, recordando que este valor no será menor a 5, ni un valor mayor a 15. Nuestra tabla estará constituida por cinco intervalos.

Paso 2: Determinar el ancho de cada intervalo

Antes de hallar el ancho de los intervalos de clase, se debe calcular el rango (R) como primera medida. En nuestro ejemplo el rango fue calculado anteriormente cuyo resultado fue igual a $420.
Con el Rango y el número de intervalos, podremos hallar el ancho:

$$A = R / Nc$$

A = 420/5 = 84

Paso 3: Determinar los intervalos de clases

Con el valor mínimo de la serie como punto de partida y el ancho de clase, se procede a construir los intervalos de clase. El primer intervalo parte del valor mínimo equivalente al límite inferior del primer intervalo, al cual le agregamos el ancho del intervalo de clase con lo cual obtenemos el límite superior; el segundo intervalo se construye a partir del valor siguiente al límite superior calculado en el primer intervalo, constituyéndose de esta forma en el límite inferior, a dicho valor se le suma el ancho de clase con lo cual se obtiene el límite superior, y así sucesivamente se van construyendo el resto de intervalos.

Continuando con nuestro ejemplo en el cual se determinó que el número de intervalos (Nc) es igual a 5, y que el ancho de clase igual a 84, se procede a construir los intervalos correspondientes:

INTERVALO	LIMITE INFERIOR	LIMITE SUPERIOR
1	1,260	1,344
2	1,345	1,429
3	1,430	1,514
4	1,515	1,599
5	1,600	1,684

Como se puede observar en los intervalos construidos, el último límite superior ($1,684) cubre al valor máximo ($1,680) de la serie de datos de ingresos de las 24 familias, por consiguiente se determina la valides del límite superior del intervalo construido, puesto que contempla el valor máximo de la serie de datos.

Paso 4: Determinar las frecuencias absolutas, frecuencias relativas y marcas de clases

Un valor representativo de los intervalos en las tablas de frecuencia son las “Marcas de Clase”.

Marcas de Clase (Mc): Se define como el punto medio de un intervalo de clase y se obtiene sumando los límites inferior y superior de la clase y dividiendo por 2.

$$Mc = \frac{Ls + Lm}{2}$$

Continuando con el ejemplo anterior se construye la tabla siguiente:

INTERVALO	LIMITE INFERIOR	LIMITE SUPERIOR	fi	Fa	h_i	H_i	Mc
1	1,260	1,344	4	4	0.167	0.167	1,302
2	1,345	1,429	6	10	0.250	0.417	1,387
3	1,430	1,514	6	16	0.250	0.667	1,472
4	1,515	1,599	4	20	0.167	0.833	1,557
5	1,600	1,684	4	24	0.166	1.000	1,642
Total			**24**		**1.000**		

Punto 2: Gráficos

Los gráficos son considerados como el método de presentación de la información más simple para el lector porque puede captar el panorama general o la tendencia de los datos en una sola mirada. Es mucho más fácil de comprender que una tabla o un texto. La sencillez de líneas, una atractiva manera de presentación, la posibilidad de usar las tres dimensiones (3D), junto con colores, hacen de los gráficos una de las herramientas más poderosas para transmitir ideas en forma rápida y simple al lector. Su desventaja más notoria es la pérdida de precisión y exactitud, si se le compara con una tabla.

Componentes de un gráfico:

Un gráfico para que sea de fácil comprensión para el lector deberá estar compuesto de las partes siguientes:
a.- Título del gráfico, hace referencia que situaciones se quiere representar.
b.- Cuerpo del gráfico o gráfico propiamente dicho (incluye leyenda en el eje de las “x” y eje de las “y” de ser necesarias).
c.- Fuente de información (se coloca como pie del gráfico la fuente de información de la que se recopilo la información para hacer el gráfico).

A continuación se describirán de forma breve los gráficos más utilizados:

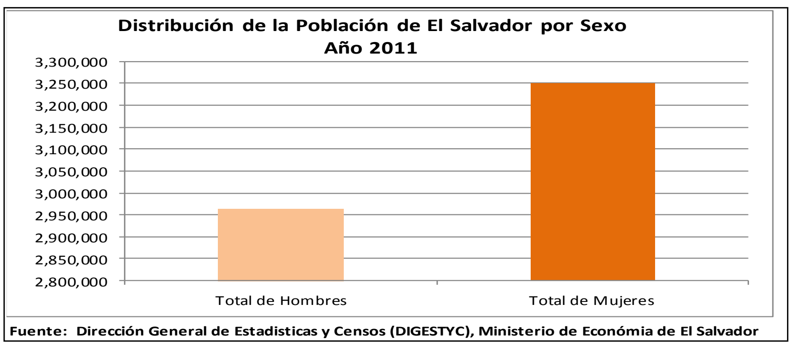
Diagramas de barras:

Se llama así porque las frecuencias de cada categoría de la distribución se hacen figurar por trazos o columnas de longitud proporcional (verticales u horizontales), separados unos de otros. Se usa fundamentalmente para representar distribuciones de frecuencias de una **variable cualitativa o cuantitativa discreta**, y ocasionalmente en la representación de series cronológicas o históricas. Uno de los ejes sirve para inscribir las frecuencias, ya sean absolutas o relativas (%), y el otro para la escala de clasificación utilizada.

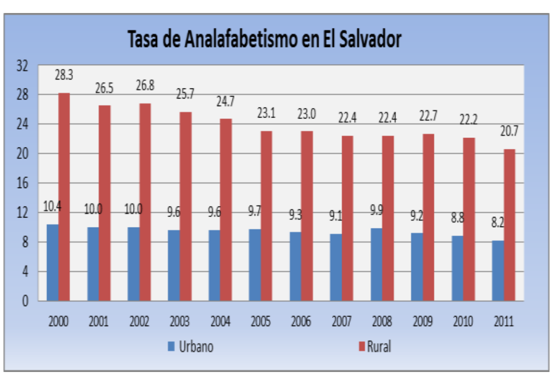
Existen tres principales clases de gráficos de barras:

Barra simple: se emplean para graficar hechos únicos.

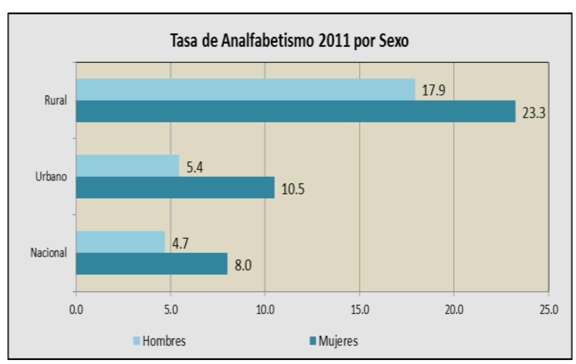
Ejemplo:



Barras múltiples: es muy recomendable para comparar una serie estadística con otra, para ello emplea barras simples de distinto color o tramado en un mismo plano cartesiano, una al lado de la otra.
Ejemplos:



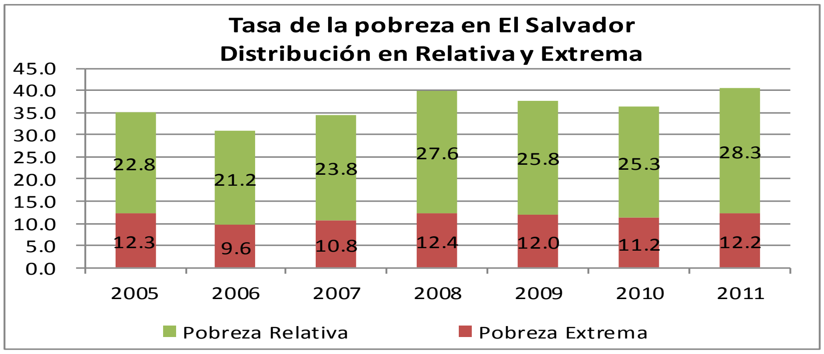
Fuente: Dirección General de Estadisticas y Censos (DIGESTYC), Ministerio de Económia de El Salvador



Fuente: Dirección General de Estadisticas y Censos (DIGESTYC), Ministerio de Económia de El Salvador

Barras compuestas: en este método de graficación las barras de la segunda serie se colocan encima de las barras de la primera serie en forma respectiva.

Ejemplo:



Fuente: Dirección General de Estadisticas y Censos (DIGESTYC), Ministerio de Económia de El Salvador

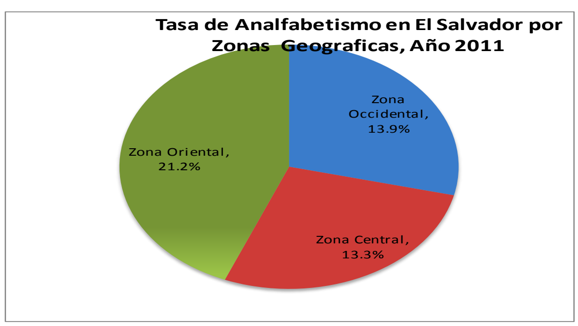
Diagramas de sectores (también llamados gráficos circulares)

Se divide un círculo en tantas porciones como clases existan, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa (figura 3). Se muestra el diagrama en dos y tres dimensiones; para una mejor ilustración se le pueden agregar colores.

Características de los gráficos de sectores

- No muestran frecuencias acumuladas.
- Se prefiere para el tratamiento de datos cualitativos o cuasicuantitativas.
- La mayor área (o porción de la figura) representa la mayor frecuencia.
- Son muy fáciles de elaborar.
- La figura completa equivale al 100% de los datos (360º).

Ejemplo de Gráficos de Sectores:



Fuente: Ministerio de Economia de El Salvador, Dirección General de Estadisticas y Censos, Encuesta de Hogares y Propositos Multiples 2011.

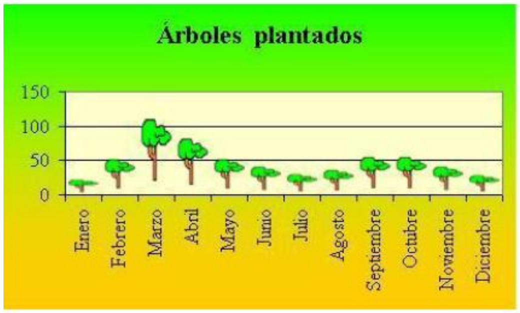
Pictogramas

Es un gráfico con dibujos alusivos al carácter que se está estudiando y cuyo tamaño es proporcional a la frecuencia que representan, dicha frecuencia se suele indicar.

Algunas características de este tipo de gráfico son:

- Su formato es libre.
- Emplean una secuencia de símbolos para representar frecuencias.
- Se emplean para el tratamiento de datos tanto cualitativos como cuantitativos.

Ejemplos de Pictograma: Plantación de árboles por mes y nacimientos por departamentos.



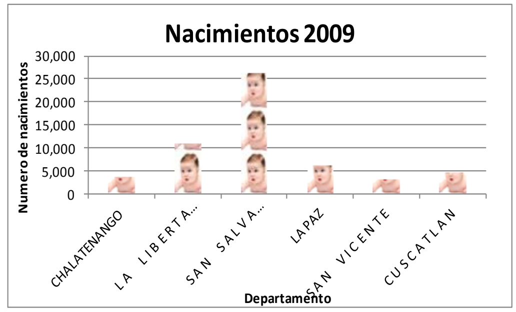
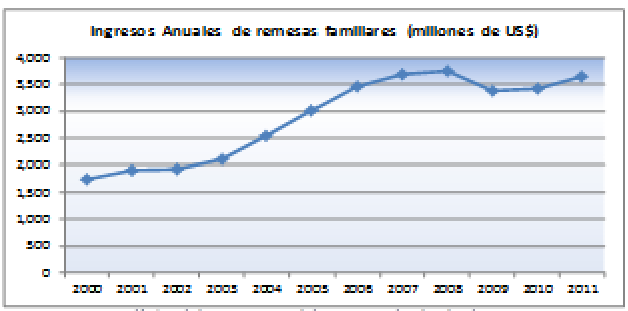


Grafico Lineal: Consiste en un conjunto de líneas o segmentos de recta que muestran los cambios que experimenta una determinada variable cuantitativa, generalmente, en función del tiempo. En el eje horizontal se describe el tiempo y en el eje vertical la frecuencia con que aparece la unidad de tiempo.
Ejemplo: Ingresos Anuales de remesas familiares.



Fuente: Base Estadística del Banco Central de Reserva de El Salvador

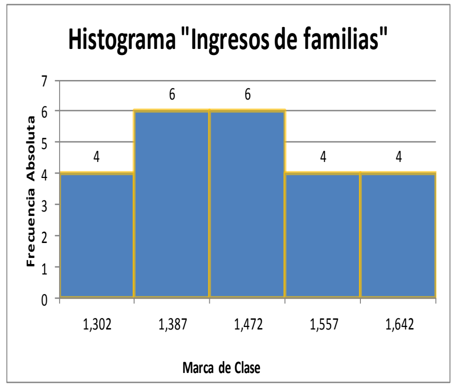
Histograma: Se puede considerar como un gráfico de columnas especial. Se realiza sobre el primer cuadrante del plano cartesiano. La diferencia radica en que el histograma se utiliza más a menudo para representar Tabla de Frecuencia con Datos Agrupados en Clases, donde el ancho de la columna equivale al ancho del intervalo de clase.

Las frecuencias absolutas se colocan en el eje vertical y también pueden emplearse las frecuencias relativas y en el eje horizontal las marcas de clases. Otra diferencia importante es que no existe espacio entre las barras.

Se estima que por el tipo de información brindada y por la manera en que ésta es dispuesta, los histogramas son de especial utilidad y eficacia para las ciencias sociales ya que permiten comparar datos sociales como los resultados de un censo, la cantidad de mujeres y/o hombres en una comunidad, el nivel de analfabetismo o mortandad infantil, etc.

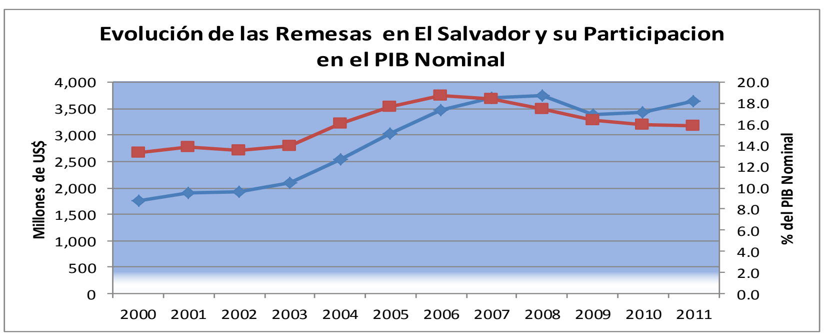
Ejemplo de histograma con la información siguiente: Ingresos semanales en US$ de 24 familias.

INTERVALO	LIMITE INFERIOR	LIMITE SUPERIOR	fi	Mc
1	1,260	1,344	4	1,302
2	1,345	1,429	6	1,387
3	1,430	1,514	6	1,472
4	1,515	1,599	4	1,557
5	1,600	1,684	4	1,642
Total			24	



Gráficos que Representan Dos tipos de Escalas (utilización de eje principal y eje secundario): este tipo de gráficos es muy utilizado cuando se quiere presentar información que utiliza dos tipos de escala de medición diferentes por ejemplo cantidades porcentajes y números enteros.

Ejemplo: Remesas en El Salvador y PIB nominal.



Fuente: Base Estadística del Banco Central de Reserva de El Salvador

2.3 Conclusiones/Ideas fuerza a tener presente

1. La tabla de frecuencia simple se caracteriza por manejar un conjunto pequeño de posibles resultados de una variable dentro de la muestra o población. Por lo general, su uso tiende al manejo de datos cualitativos o variables cuantitativas discretas, sin embargo cuando el tamaño de la población y/o muestra y el rango de la variable es grande, será necesario agrupar dichos valores en "**Intervalos de Clases"** por lo que se hace necesario utilizar una tabla de frecuencia para datos agrupados.

2. Los gráficos son considerados como el método de presentación de la información más simple para el lector porque pueden captar el panorama general o la tendencia de los datos en una sola mirada ya que son mucho más fácil de comprender que una tabla o un texto, sin embargo su principal desventaja radica en la pérdida de precisión y exactitud que podrían tener si se compara con una tabla.

3. Dentro de los gráficos más utilizados para representar variables cualitativas se encuentran los gráficos de barras, de sectores y pictogramas, sin embargo los gráficos de barras también son utilizados para la representación de variables cuantitativas, así como los gráficos de líneas que se utilizan principalmente para representación de series cronológicas.

Es importante mencionar que si se desea representar dos tipos de escalas en un gráfico, puede ser útil la utilización de dos ejes uno principal y uno secundario, en el que en cada eje se especifique la escala de medición que se utilice, lo cual facilitará la comprensión del lector.

2.4 Bibliografía

1. Gildaberto Bonilla, Estadística I, Elementos de Estadística Descriptiva y Probabilidad.
2. Jhonson, R y Kuby, P. (1999). Estadística Elemental, lo Esencial. México: Thomson.
3. Martínez Bencardino, Ciro. (2006). Estadística Básica Aplicada. Colombia: ECOE EDICIONES, 3° ED.
4. Montgomery, Douglas C. y Runger, George C. (1996). Probabilidad y Estadística. McGrawHill.
5. Juan Carlos Vergara Schmalbach y Víctor Manuel Quesada Ibarguen, Estadística Básica con Aplicaciones en Excel.

LECCION 3: Medidas de Posición Centrales y no Centrales

Resumen de la lección: Esta lección pretende brindar conocimientos generales, sobre el cálculo y utilidad de las medidas de tendencia central, dentro de las cuales se consideran para efectos de estudio la media aritmética y geométrica, mediana y moda.

Tiempo total: 1:30 horas

3.1 Objetivos de la Lección 3

Proporcionar conocimientos sobre la utilidad práctica que tienen las medidas de posición central y no centrales, en el diseño de programas puesto, que estos pueden simplificar un conjunto de datos por medio de un solo número e indicar donde se concentran los valores de estudio e identificar de una mejor manera la población objetivo beneficiaria de una intervención.

3.2 Desarrollo del contenido de la Lección 3

Definiciones clave

Media aritmética y geométrica, mediana, moda, cuartil, quintil, decil, percentil.

Punto 1: Medidas de Posición Central

Los promedios o medidas de posición proporcionan valores típicos o representativos de la variable en estudio. Dentro de las medidas de posición centrales más conocidas se encuentran: la media (aritmética y geométrica), la mediana y la moda. Las cual a continuación se describe un breve resumen del mismo.

Media aritmética

Es la medida más conocida, la más fácil de calcular y con la que siempre estamos más familiarizados, ya que siempre hemos calculado el promedio de calificaciones obtenidas en cada periodo escolar (Ciclo, año, etc.) A veces se le denomina simplemente *media o promedio*, y es utilizada con tanta frecuencia, que en algunas ocasiones nos conduce a resultados que no revelan lo que se pretende presentar, ya que la distribución de los datos puede requerir de la aplicación de un promedio diferente a la media, ya sea, media geométrica o media armónica.

Media Aritmética simple: se define como el cociente que se obtiene al dividir la suma de los valores de la variable por el número total de observaciones. Su fórmula está dada por:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La media aritmética es un promedio estándar que a menudo se denomina "promedio".

Ejemplo. Supongamos que un almacén tiene empleados a 12 vendedores, y sus ingresos mensuales son: $ 585, $ 521, $ 656, $ 465, $ 536, $ 487, $ 564, $ 490, $ 563, $ 1234, $ 469 y $ 547. Se pide determinar la media de los ingresos de los 12 vendedores.

X =1/12 (7,117) = **$593.08 “Promedio de Ingreso de los vendedores”**

La media aritmética es muy sensible a los valores extremos de la variable, ya que todas las observaciones intervienen en el cálculo de la media, la aparición de una observación extrema, hará que la media se desplace en esa dirección. En consecuencia, no es recomendable usar la media como medida central en las distribuciones muy asimétricas ya que podría no ser una medida muy representativa; **ejemplos:** el número de hijos promedio de las familias de cierto país puede no ser un valor muy representativo, ya que si sale elevado puede ser ocasionado porque en algunas área del país la población tiene un número elevado de hijos y dicho dato no se puede inferir a la población total ya que no sería representativo, así también el PIB per cápita, considerado como el promedio de ingresos al año que reciben las personas de determinado país puede no ser una variable muy representativa, por lo que ciertos analistas utilizan para determinar el promedio de ingresos por persona el llamado salario modal es decir el salario que más se repite.

Media aritmética ponderada

Cuando el número de observaciones es grande, las operaciones para calcular la media se simplifican si agrupamos los datos en una tabla de frecuencias. La fórmula matemática está dada por:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n}$$

Si los datos están agrupados en clase, no se conoce el valor de x, por lo tanto se toma el punto medio de cada clase en vez de x (marca de clase).

Por otra parte, la media aritmética de una muestra dividida en submuestras, es igual, a la media ponderada de las submuestras, tomando como ponderación los tamaños de las submuestras. Esto es,

$$\bar{x} = \frac{\sum_{i=1}^{m} \bar{x}_i n_i}{n} \quad donde \;\; n = n_1 + n_2 + \ldots + n_m$$

Ejercicios propuestos de media aritmética

Media Simple

1. El jefe de una delegación policial inició un estudio acerca de las horas de tiempo extra de los policías. Se seleccionaron al azar 15 de ellos y durante el mes de junio se anotaron las siguientes horas extras laboradas. 13 , 13, 12, 15, 7, 15, 5, 12, 6, 7, 12, 10, 9, 13, 12. Calcule la media aritmética.

2. El banco B analiza el número de veces que se utiliza por día un cajero automático ubicado en el Superselectos. A continuación se indican las veces que dicho aparato se utilizó en cada uno de los últimos 30 días: 83, 63, 96, 64, 80, 36, 84, 84, 78, 76, 73, 61, 84, 68, 59, 54, 52, 84, 75, 65, 95, 59, 90, 47, 70, 52, 87, 61, 77, 60,

Determine la media del número de veces que la máquina fue utilizada por día.

Media ponderada

3. La siguiente tabla muestra el porcentaje de desempleo de la PEA y el número de personas de la PEA de algunos municipios seleccionados de cierta región.

Municipio	% de desempleo	PEA
X	4.5	74520
Y	7.6	126310
Z	8.3	98760

Cuál es el promedio de desempleo para la región.

4. Suponga que en junio un inversionista compró 300 acciones del Banco Agrícola a un precio de $ 20 por acción, en agosto compró 400 acciones más a $ 25 cada una, y en noviembre 400 a $ 23 por acción. Cuál es el precio medio ponderado por acción.

Solución

$$\overline{X} = \frac{300(20) + 400(25) + 400(23)}{300 + 400 + 400} = 22.9$$

5. En un supermercado trabajan 35 mujeres, con un salario promedio mensual de $650 dólares y 15 hombres, en promedio ganan un 12% más que las mujeres ¿Cuál es el salario promedio de los empleados del supermercado?

La Media Geométrica (MG)

La media geométrica (Mg), de un conjunto de "n" números positivos se define como la raíz n - ésima de la multiplicación de los "*n*" valores de la variable. Por tanto, la fórmula para la media geométrica es dada por:

$$MG = \sqrt[n]{X_1 \cdot X_2 \cdot .. \cdot X_n}$$

Existen dos usos principales de la media geométrica:

1. Para promediar porcentajes, índices y cifras relativas y
2. Para determinar el incremento porcentual promedio en ventas, producción u otras actividades o series económicas de un periodo a otro.

Ejemplo

Supóngase que las utilidades obtenidas por una compañía constructora en cuatro proyectos fueron de 3, 2, 4 y 6%, respectivamente. ¿Cuál es la media geométrica de las ganancias?

En este ejemplo la media geométrica es determinada por:

$$MG = \sqrt[4]{(3)(2)(4)(6)}$$
$$= 3.464101615$$

Y así la media geométrica de las utilidades es el 3.46%. La media aritmética de los valores anteriores es 3.75%. Aunque el valor 6% no es muy grande, hace que la media aritmética se incline hacia valores elevados. La media geométrica no se ve tan afectada por valores extremos.

Ejercicio 1. El 1 de mayo de 2006 se ahorraron $ 50,000 en un banco al 7.6% de interés anual, capitalizados semestralmente. Obtener la cantidad media depositada en la cuenta, entre el 1 de mayo de 2002 y el 31 de octubre de 2012, suponga que no se hicieron retiros durante el periodo.

La Moda:

Es una medida de posición, menos importante que los promedios y su uso es bastante limitado. Se utiliza en distribuciones cuando la variable o el atributo presentan una frecuencia demasiado grande con respecto a las demás.

"La moda se define como aquel valor de la variable o del atributo que presenta la mayor densidad, es decir, la mayor frecuencia."

Si se tiene un atributo o una variable con máxima frecuencia, la distribución es ***unimodal***. Si hay dos valores en la variable con la misma frecuencia máxima, la distribución es ***bimodal***. Si hay más de dos, la distribución es ***multimodal***. Cuando ninguno de los valores que toma la variable se repite, no existe moda.

La Mediana:

La mediana de una distribución de frecuencia corresponde al valor, supuesto los datos ordenados de menor a mayor, que deja a ambos lados el mismo número de observaciones. Cuando calculamos la mediana en datos no agrupados, ordenamos las observaciones de menor a mayor o viceversa. En su cálculo se presentan dos casos:

a) **Cuando el número de datos es impar:** En este caso la mediana coincide con el dato central.

Ejemplo: Consideremos los salarios en dólares para 11 vendedores: 243, 320, 311, 254, 234, 261, 239, 310, 218, 267, 287.

Calcular la mediana.

Solución:

Primero ordenar los datos de menor a mayor: 218, 234, 239, 243, 254, 261, 267, 287, 310, 311, 320. La posición donde se encuentra la mediana: **(11+1)/2=6**, la mediana se encuentra en la sexta posición y corresponde al valor de: **Md=261.**

b) Cuando los datos son pares: La mediana será el término medio de los dos valores centrales.

Ejemplo: Consideremos los salarios en dólares para 12 vendedores; los cuales se han presentado ordenados anteriormente 218, 234, 239, 243, 254, 261, 267, 287, 310, 311, 320 y 322: Calcular la mediana.

Solución:

Para obtener la posición central se aplica la siguiente fórmula: (N+1)/2 $\Longrightarrow$ (12+1)/2 = 6.5, entonces la mediana corresponde al promedio de los dos valores sombreados, esto es: **Md=(261+267)/2=264.**

Punto 2. Medidas de Dispersión

Las medidas de dispersión estudian la separación existente entre los diversos valores que toma la variable. Se dividen en medidas de dispersión absoluta y relativa. Las absolutas suelen hacer referencia a un promedio, y permiten estudiar su representatividad. En este tipo de medidas depende de las unidades, lo que es un inconveniente para realizar comparaciones entre poblaciones. En este sentido, las medidas de dispersión relativas no dependen de las unidades y permiten comparar variabilidad entre poblaciones.

VARIANZA

La varianza es una medida de dispersión que sirve para estudiar la representatividad de la media. Viene definida como la media de las diferencias cuadráticas de las puntuaciones respecto a su media aritmética. Normalmente a partir de la varianza se obtiene la desviación típica o estándar y se define como la raíz cuadrada positiva de la varianza, a continuación se presentan dichas fórmulas:

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \qquad\qquad S = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}$$

Una varianza grande es indicativa de que la media no es representativa, mientras que una varianza pequeña indica que la media es un buen representante de los datos.

COEFICIENTE DE VARIACIÓN

En ocasiones puede interesar comparar la dispersión de dos muestras y la desviación típica no es válida, si las dos muestras tienen unidades diferentes. Para evitar este inconveniente se define el coeficiente de variación CV como:

$$CV = \frac{S}{\bar{x}}$$

Utilidad del coeficiente de variación

VALOR DEL C.V.	**GRADO EN QUE LA MEDIA REPRESENTA AL CONJUNTO DE DATOS**
0-<10%	Media altamente representativa
10% - < 20%	Media bastante representativa
20% - < 30%	Media tiene representatividad
30%- < 40%	Media con representatividad dudosa
40% o más	Media carente de representatividad

Ejemplo. Una compañía requiere los servicios de un técnico especializado. De los expedientes presentados, se han seleccionado 2 candidatos: A y B, los cuales reúnen los requisitos mínimos requeridos. Para decidir cuál de los 2 se va a contratar, los miembros del Jurado deciden tomar 7 pruebas a cada uno de ellos. Los resultados se dan a continuación:

Pruebas	1	2	3	4	5	6	7
Puntaje obtenido por A	57	55	54	52	62	55	59
Puntaje obtenido por B	80	40	62	72	46	80	35

Estadísticamente ¿Cuál de los candidatos debe ser contratado? Fundamente su respuesta.

Solución.

Podemos notar que B ha tenido mayores notas que A en algunas pruebas, pero también ha tenido menores notas que A. El rendimiento de A de alguna manera fue constante. Analicemos la media y la desviación para cada uno.

Candidato A:

$$\bar{x}_A = \frac{57+55+54+52+62+55+59}{7} = \frac{394}{7} = 56.29$$

$$S_A^2 = \frac{(57-56.29)^2+(55-56.29)^2+(54-56.29)^2+(52-56.29)^2+(62-56.29)^2+(55-56.29)^2+(59-56.)^2}{7} = 11.23$$

$$S_A = 3.35$$

Candidato B:

$$\bar{x}_B = \frac{80+40+62+72+46+80+35}{7} = \frac{415}{7} = 59.28$$

$$S_B^2 = \frac{(80-59.28)^2+(40-59.28)^2+(62-59.28)^2+(72-59.28)^2+(46-59.28)^2+(80-59.28)^2+(35-59.28)^2}{7} = 360.90$$

$$S_B = 18.99$$

Podemos también analizar el coeficiente de variación (C.V).

$C.V._A = \frac{3.35}{56.29}$ =0.05=5%   la media de A es altamente representativa.

$C.V._B = \frac{18.99}{59.28}$ =0.32=32%   la media de B tiene representatividad dudosa.

Aunque B tuvo mejor promedio que A, B presenta mayor variabilidad en sus notas, por lo que sería mejor contratar al candidato A.

Ejercicio propuesto: Durante un cierto mes el precio de los frijoles por libra tuvo $\bar{x} = \$1.55$ y $S_x = \$0.25$. Mientras que en ese mismo mes el precio del barril de petróleo tuvo una media de $\bar{y} = \$98\ S_y = \$5$. ¿Dónde hubo mayor variabilidad en los precios?

Punto 3: Medidas de Posición no Centrales: Cuartiles, Quintiles, Deciles y Percentiles

Cuartiles

Son medidas de posición que dividen en cuatro partes porcentuales iguales a una distribución ordenada de datos.

Cuando la distribución de datos contiene un número determinado de datos y se requiere obtener un porcentaje o una parte de la distribución de datos, se puede dividir la distribución en cuatro partes iguales, cada parte tiene la misma cantidad de datos y cada una de las partes representa un 25% de la totalidad de datos. Es decir:

Cuartil 1 **25 %**	**Cuartil 2** **50 %**	**Cuartil 3** **75 %**	**Cuartil 4** **100%**

Fórmula General:

Para calcular el valor de uno de los cuatro Cuartiles, se utiliza la fórmula:

$$Q_k = k\ (N/4)$$

Donde:
Q_k = Cuartil número 1, 2, 3 ó 4
N = total de datos de la distribución.

Para cada cuartil, su ecuación se establece así:

$Q_1 = 1 (N / 4)$ $Q_2 = 2 (N / 4)$ $Q_3 = 3 (N / 4)$ $Q_4 = 4 (N / 4)$

Cada cuartil tiene un significado estadístico particular o representa de la distribución de datos un porcentaje establecido; por ejemplo:

a) $Q_1 = 1 (N/4)$
El valor obtenido al realizar el cálculo en una serie de datos nos proporciona el valor que representa el 25 % de esa serie de datos. También, nos indica que el 25% de la serie de datos está bajo él y sobre él, se encuentra el 75% de los datos de la serie.

b) $Q_2 = 2 (N/4)$
Para el cuartil 2, se tiene como caso especial, primero porque su valor representa la mitad de la serie de datos, igual que la mediana. Segundo, bajo esté valor se encuentra el 50% de la serie de datos y tercero, sobre ese valor calculado se encuentra el otro 50% de la serie de datos.

c) $Q_3 = 3 (N/4)$
El cuartil 3, nos indica que el valor obtenido representa bajo sí el 75 % de la distribución de los datos y sobre sí, se encuentra el 25 % de la distribución de datos.

d) $Q_4 = 4 (N/4)$
El cuartil 4, nos indica que el valor obtenido tiene bajo sí el 100% de la distribución de datos. Por lo general no se calcula, ya que es un hecho que el último valor de la distribución él lo representa.

<u>Quintiles</u>
Se representan con la letra K. Su fórmula aproximada es i*n/5.

- El primer quintil. Separa a la muestra dejando al 20 % de los datos a su izquierda.
- El segundo quintil. Es el valor que indica que el 40 % de los datos son menores.
- El tercer quintil. Indica que el 60 % de los datos son menores que él.
- El cuarto quintil. Separa al 80 % de los datos inferiores del otro 20 %.

<u>Deciles</u>

- Se representan con la letra D. Son 9 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en diez tramos iguales, en los que cada uno de ellos concentra el 10% de los resultados. Su fórmula aproximada es i*n/10.

- Es el decil i-ésimo, donde la i toma valores del 1 al 9. El (i*10) % de la muestra son valores menores que él y el 100-(i*10) % restante son mayores.

Ejemplo de medidas de posición:
Una empresa prestigiosa asegura que sus empleados tienen salarios superiores al salario mínimo. Se toma una muestre de 15 personas y sus salarios fueron: 300, 275, 180, 325, 200, 250, 350, 260, 280, 310, 400, 380, 260, 290, 370. Calcular Q_1, Q_2 y Q_3, del salario en dólares, deciles 1 y 7 y percentil 50 y 70.

Solución:
Recordemos que: **Q_1 = 1 (N / 4)** **Q_2 = 2 (N / 4)** **Q_3 = 3 (N / 4)**

Para encontrar los cuartiles necesitamos tener los datos ordenados de menor a mayor, esto es:

180, 200, 250, 260, 260, 275, 280, 290, 300, 310, 325, 350, 370, 380, 400.

Luego hacemos los respectivos cálculos:

$Q_1 = \frac{1(15)}{4} = 3.75$, éste valor es el que se encuentra en la posición 4, luego: $Q_1 = 260$

$Q_2 = \frac{2(15)}{4} = 7.5$, éste valor es el que se encuentra en la posición 8, luego $Q_2 = 290$

$Q_3 = \frac{3(15)}{4} = 11.25$, éste valor es el que se encuentra en la posición 12, luego $Q_3 = 350$

Se tiene que el 75% de los trabajadores gana menos de $350.00

Deciles:

Calcular D_1 y D_7 del salario de los 15 trabajadores anteriores.
D_i = i (N /10)

Así:

$D_1 = \frac{1(15)}{10} = 1.5$, éste valor es el que se encuentra en la posición 2, entonces: $D_1 = 200$

$D_7 = \frac{7(15)}{10} = 10.5$, éste valor es el que se encuentra en la posición 11, entonces: $D_7 = 325$

Podemos decir que el 10% de los empleados gana menos de $200 y el 25% ganan más de $325.

Percentiles:
Encontrar los percentiles P_{50} y P_{70} del salario de los 15 trabajadores anteriores.

P_i = i (N /100)

$P_{50} = \frac{50(15)}{100} = 7.5$, éste valor es el que se encuentra en la posición 8, así: $P_{50} = 290$

Que coincide con el valor de Q_2.

$P_{70} = \frac{70(15)}{100} = 10.5$, éste valor es el que se encuentra en la posición 11, así: $P_{70} = 325$

Que coincide con el valor de D_7.

De aquí se puede observar que el 50% de los trabajadores gana menos de $290.

3.3. Conclusiones/Ideas fuerza a tener presente

1. Las medidas de tendencia central simplifican la información de un conjunto de datos y que brindan información sobre donde se encuentran más concentrados.

2. Cuando existe mucha disparidad en los datos la media puede no ser una medida muy representativa, ya que sesgaría la información; para estos casos pueden considerarse otras medidas de tendencia central como la moda o la mediana.

3. Las medidas de posición no centrales son muy útiles principalmente en el análisis de información económica, puesto que ayudan a determinar por ejemplo en que segmentos de la población se concentran más los ingresos.

3.4 Bibliografía

1. Introducción a la Probabilidad y Estadística Inferencial, Dr. Nerys Funes
2. http://www.eumed.net/cursecon/libreria/drm/ped-drm-est.htm
3. http://www.vitutor.com/estadistica/descriptiva/a_1.html

LECCION 4: Análisis de Regresión

Resumen de la lección: Esta lección tiene como propósito conocer la utilidad de la regresión lineal en la gestión de programas presupuestarios, puesto que ayudaría a identificar si existe o no relación entre variables, también se exponen parámetros estadísticos que ayudarían a identificar ese grado de relación.

Tiempo total: 1 horas

4.1 Objetivos de la Lección 4

- comprender la utilidad de la regresión lineal
- identificar los diferentes tipos de correlación lineal
- interpretar la confiablidad del modelo a través de parámetros estadísticos.

4.2 Desarrollo del contenido de la Lección 4

Definiciones clave

– Regresión lineal
– Diagrama de dispersión
– Coeficiente de determinación,
– Coeficiente de correlación.

Punto 1: Regresión Lineal Simple

El análisis de regresión: es una técnica estadística para investigar la relación funcional entre dos o más variables, ajustando algún modelo matemático. La regresión lineal tiene una finalidad muy importante que es la de estimar los valores de una variable con base a los conocidos de la otra.

Cuando es utilizado este modelo para predicciones, es importante que solo se considere lo relevante de la variable independiente. Los valores de la variable dependiente se predicen cuando es conocida la variable independiente.

Diagrama de Dispersión:
El diagrama de dispersión es la gráfica donde se encuentran todos los puntos de las observaciones, tanto de la variable dependiente (Y), como de la variable independiente (X). El diagrama de dispersión puede revelarnos dos tipos de información:

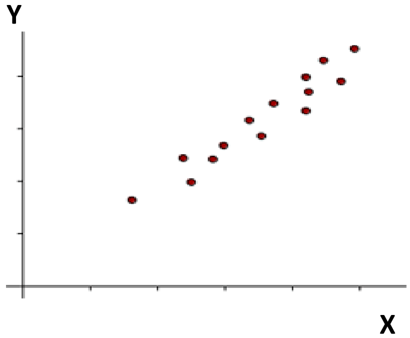
1) Relación de las variables
2) Tipo de línea o ecuación de estimación

El análisis de un diagrama de dispersión puede mostrar varios tipos de correlaciones entre las variables con un intervalo de confianza determinado. La correlación puede ser positiva (Y aumenta con X), negativa (Y disminuye con X), o nula (las variables no están correlacionadas).

El diagrama de dispersión es una de las herramientas básicas de gestión de la calidad, muy útil al analizar las causas de un problema y para Identificar oportunidades de mejora continua.

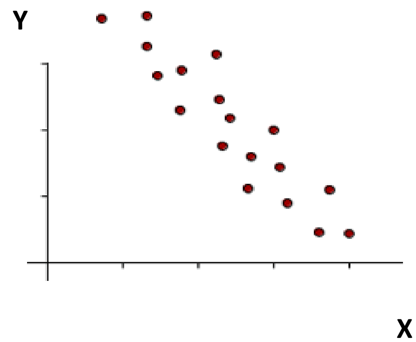
Correlación directa

La recta correspondiente a la nube de puntos de la distribución es una recta creciente.



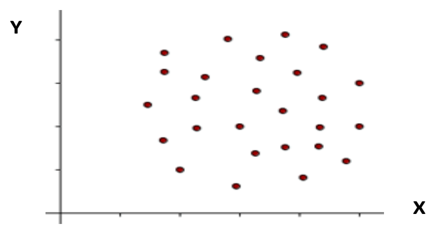
Correlación inversa

La recta correspondiente a la nube de puntos de la distribución es una recta decreciente.



Correlación nula

En este caso se dice que las variables son *no correlacionadas* y la nube de puntos tiene una forma redondeada.



Cálculo de la Recta de Regresión

Uno de los métodos más utilizados para calcular la recta de regresión es el de Mínimos Cuadrados es una técnica de análisis numérico en la que, dados un conjunto de pares ordenados: variable independiente (X), variable dependiente (Y), y una familia de funciones intenta encontrar la ecuación que mejor se aproxime a los datos (un "mejor ajuste"), de acuerdo con el criterio de *mínimo error cuadrático*.

Y = a + bx.
Y = Estimaciones de “Y” a partir de “X”.
b = Aumento de “Y” por unidad de aumento de “X”.
a = valor de “Y” cuando “X” vale cero (ordenada al origen).

El coeficiente de correlación lineal (r): frecuentemente denominado correlación. Una medida estadística ampliamente utilizada que mide el grado de intensidad de la relación lineal entre dos variables aleatorias.

Los valores que puede tomar el coeficiente de correlación "r" son: -**1 < r < 1**

Si **r > 0**, la correlación lineal es positiva (si sube el valor de una variable sube el de la otra). La correlación es tanto más fuerte cuanto más se aproxime a 1.

Ejemplo: ingesta de nutrientes en niños y rendimiento escolar: los niños que ingieren **más** nutrientes tendrán **mayor** rendimiento escolar.

Si **r < 0**, la correlación lineal es negativa (si sube el valor de una variable disminuye el de la otra). La correlación negativa es tanto más fuerte cuanto más se aproxime a -1.

Ejemplo: dispositivos de seguridad en la carretera y número de accidentes en carreteras: mientras **más** dispositivos de seguridad se implementen en carretera ocurrirán **menos** accidentes de tránsito.

Si **r = 0**, no existe correlación lineal entre las variables. Aunque podría existir otro tipo de correlación (parabólica, exponencial, etc.)

Ejemplo: número de cigarrillos diarios y estatura de fumadores.

Coeficiente de determinación (r^2): una vez ajustada la recta de regresión a la nube de observaciones es importante disponer de una medida que indique la bondad del ajuste (discrepancia entre los valores observados y los valores esperados en el modelo de estudio, mientras menor sea esa diferencia mejor es la bondad de ajuste del modelo) realizado y que permita decidir si el ajuste lineal es suficiente o se deben buscar modelos alternativos.

El coeficiente de determinación se encuentra entre los valores 0 y 1, pero sin olvidar que un r^2 cercano a 1 indica una fuerte correlación entre "X" y "Y", mientras que un r^2 cercano a 0 significa que las variables tienen poca correlación. El coeficiente de determinación resulta al elevar al cuadrado el coeficiente de correlación (r).

Parámetros para determinar la significancia del coeficiente de determinación

Valor	Significado
0	Correlación nula
0.25 - 0.49	Correlación débil
0.5 - 0.74	Correlación moderada
0.75 - 0.99	Correlación intensa
1	Correlación perfecta

Cálculo de los coeficientes de la recta de regresión.

Los coeficientes a y b de la recta de regresión $y = ax + b$, se calculan con las siguientes fórmulas:

$$a = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$b = \bar{Y} - a\bar{X}$, donde $\bar{X} = \frac{\sum x}{n}$ y $\bar{Y} = \frac{\sum Y}{n}$

Se observa que es necesario calcular cinco cantidades para determinar a y b:
$n, \Sigma X, \Sigma Y, \Sigma X^2 \; y \; \Sigma XY$

Para calcular el coeficiente de determinación se utiliza la siguiente fórmula:

$$r^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Ejemplo: Obtener la recta de regresión para las puntuaciones de una prueba de aprovechamiento en matemáticas (X) y las calificaciones finales (Y) para estudiantes universitarios de primer año. Calcular también, su coeficiente de determinación.

Solución

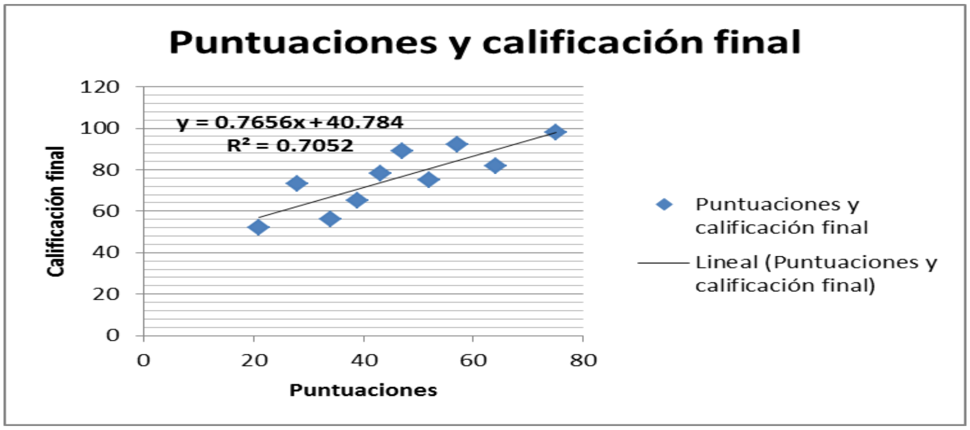
Estudiante	Puntuación del examen de aprovechamiento	Calificación final	X^2	XY	Y^2
1	39	65	1521	2535	4225
2	43	78	1849	3354	6084
3	21	52	441	1092	2704
4	64	82	4096	5248	6724
5	57	92	3249	5244	8464
6	47	89	2209	4183	7921
7	28	73	784	2044	5329
8	75	98	5625	7350	9604
9	34	56	1156	1904	3136
10	52	75	2704	3900	5625
Total (Σ)	**460**	**760**	**23634**	**36854**	**59816**

En este caso $n = 10$, y los coeficientes de la recta de regresión son:

$$a = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{10(36854) - 460(760)}{10(23634) - (460)^2} = 0.76556$$

$$b = \bar{Y} - b\bar{X} = \frac{760}{10} - (0.76556)\left(\frac{460}{10}\right) = 40.78424$$

Por lo tanto, la recta de regresión es: $Y = 0.76556X + 40.78424$



El coeficiente de determinación será:

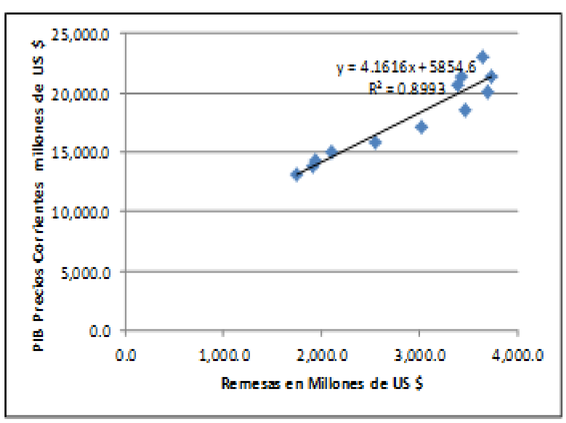
$$r^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$Sustituyendo\ valores\ se\ tiene\ \frac{(1894)^2}{(2474)(2056)} = 0.7052$$

<u>Ejemplo práctico de Regresión Lineal: PIB a precios corrientes y Flujo anual de remesas familiares</u>

Año	PIB Precios Corrientes millones de US$ (Y)	Remesas millones de US$ (X)
2000	13,134.1	1,750.7
2001	13,812.7	1,910.5
2002	14,306.7	1,935.2
2003	15,046.7	2,105.3
2004	15,798.3	2,547.6
2005	17,093.8	3,017.2
2006	18,550.7	3,470.9
2007	20,104.9	3,695.3
2008	21,431.0	3,742.1
2009	20,661.0	3,387.2
2010	21,427.9	3,431.0
2011	23,054.1	3,648.8

Fuente: Base Estadística del Banco Central de Reserva de El Salvador



La importancia que han tenido las remesas en la economía salvadoreña es muy relevantes estas representan en el periodo de 2000-2011 aproximadamente un 16% del PIB a precios corrientes.

Con la información proporcionada, se puede generar la recta de regresión utilizando una hoja de cálculo de Excel lo que facilitaría la obtención de información, dicho modelo estadístico podemos decir lo siguiente:

1. El diagrama de regresión nos dice que existe una correlación directa entre las variables (x) y (y) es decir a mayor flujo de remesas mayor PIB a precios corrientes.
2. El coeficiente de determinación es 0.8993 es decir que existe una correlación intensa entre las variables.
3. Según la recta de mínimos cuadrados, al incrementarse en un millón de dólares las remesas familiares el PIB corriente aumentaría en 4.1616 millones, y cuando no se perciba ningún flujo de remesas el PIB sería positivo pero disminuirá drásticamente.

Ejercicios propuestos

1. Para los hogares salvadoreños, disponemos del promedio mensual redondeados sobre los gastos en productos alimenticios ($Y) e ingresos promedio del hogar ($X), tomados de una muestra de hogares, para el período 2005- 2012.

Año	2005	2006	2007	2008	2009	2010	2011	2012
Yt	258	273	289	308	331	355	377	400
Xt	381	402	426	454	486	520	553	590

Considerando que los gastos se puede expresar como función lineal de los ingresos ($Y_t=a+b\cdot X_t$), determine:

a) Los estimadores de los parámetros a y b de la recta de regresión.
b) El coeficiente de determinación de dicha regresión.
d) La predicción del valor que tomará el gasto para un hogar que tiene ingresos de $650.

2. Se supone que se puede establecer cierta relación lineal entre las exportaciones de un país y la producción interna de dicho país. Por ejemplo, tenemos los datos anuales (expresados en miles de quintales de maíz) para tales variables correspondientes al quinquenio 2008-2012 en la siguiente tabla:

Año	Producción	Exportaciones
2008	52654	10420
2009	53972	11841
2010	57383	14443
2011	61829	16732
2012	65381	18760

A partir de tal información, y considerando como válida dicha relación lineal, se pide:

a) Los estimadores de los parámetros a y b de la recta de regresión.
b) Si la producción de maíz para el año 2013 fuera de 70,000 quintales, ¿cuál sería la predicción de las exportaciones para dicho año?

3. En un estudio se registran los salarios iniciales Y (en miles de dólares) y los años de estudio X de 10 empleados:

Salario inicial (Y)	Años de estudio (X)
35	12
46	16
48	16
50	15
40	13
65	19
28	10
37	12
49	17
55	14

a) Dibujar el diagrama de dispersión correspondiente a estos datos.
b) Encontrar la recta de regresión.
c) Determinar el coeficiente de determinación.
d) ¿Existe un buen grado de correlación entre las variables?
e) Si un empleado tuviera 20 años de estudio, ¿cuánto se espera que gane?

4.3 Conclusiones/Ideas fuerza a tener presente

1. El análisis de regresión es una técnica estadística para determinar si existe relación entre dos o más variables. Aplicaciones de regresión son numerosas y ocurren en casi todos los campos, incluyendo ingeniería, la física, ciencias económicas, ciencias biológicas y de la salud, como también ciencias sociales.
2. El análisis de regresión nos arroja información importante como por ejemplo:

- **Grado de asociación de las variables de estudio:** dicho análisis puede determinarse a través del coeficiente de correlación, sin embargo en la práctica se utiliza más el coeficiente de determinación puesto que es una medida de bondad de ajuste, es decir puede ayudarnos a decidir si el ajuste lineal es suficiente o se deben buscar modelos alternativos.
- **Determinar la tendencia:** los gráficos de dispersión nos indican la tendencia que tienen las variables en estudio y si existe gráficamente algún tipo de relación entre ellas.
- **Para predicción y estimación.**
  Algunos casos de esta utilidad del análisis de regresión son:
  a) La respuesta de un cultivo al variar la cantidad de los fertilizantes; el objetivo puede ser establecer la forma de la relación, o predecir la combinación optima de fertilizantes.
  b) La relación entre varias medidas meteorológicas y la producción del cultivo; el más obvio objetivo podría ser tratar de entender los efectos meteorológicos sobre el crecimiento del cultivo.

4.4 Bibliografía

1. Gildaberto Bonilla, Estadística I, Elementos de Estadística Descriptiva y Probabilidad.
2. http://www.eumed.net/cursecon/libreria/drm/ped-drm-est.htm
3. http://www.vitutor.com/estadistica/descriptiva/a_1.html
4. Domar Guajarati, Econometría.

APENDICE 1: FUENTES DE BUSQUEDA DE INFORMACIÓN

Tiempo Total: 30 minutos (Material de consulta)

Fuentes y recolección de datos

La base para el análisis estadístico y formación de indicadores son los datos, los cuales se pueden obtener de manera indirecta, por ejemplo al utilizar la información de los censos nacionales o de algún tipo de organización o institución, o de manera directa, llenando fichas, cédulas de observación o aplicando cuestionarios, entre otros.

En muchas ocasiones en la investigación científica es necesario recurrir a la obtención directa de datos, ya sea al emplear fichas o cédulas de observación en las que se registran características de interés, o bien al diseñar un cuestionario cuyas respuestas nos permitan conocer las características de quien responde, su opinión, las condiciones reales en las cuales se encuentra en relación con algún aspecto específico, como podría ser trabajo, educación y capacitación, salud, relaciones interpersonales, empleo del tiempo libre, etcétera.

Cuando los datos se obtengan de manera directa, será necesario apoyarse en el muestreo, diseñar el instrumento de recolección, coordinar la recopilación de datos y su procesamiento, para finalmente hacer en análisis de la información y efectuar el informe o informes finales en que se basará la toma de decisiones.

Tipos de Fuentes de Recolección de datos

Las fuentes de recolección de datos de una investigación estadística son:

1. **Fuentes Primarias:** Es cuando la persona o institución ha recolectado los datos.

2. **Fuentes Secundarias:** Si la persona o institución que ha publicado los datos, no es la que ha efectuado la investigación. Se utilizan cuando la oficina que las pública tenga suficiente solvencia técnica.

3. **Fuentes Oficiales:** Es cuando los datos son provistos por cualquier dependencia gubernamental.

4. **Fuentes Privadas:** Es cuando son provistos por agencias, personas, organizaciones, etc., no gubernamentales.

A continuación se presentan un conjunto de **Fuentes Oficiales** que brindan información de utilidad para estudios estadísticos y realización de indicadores de desempeño.

Nombre de la Institución	Información Proporcionada	Dirección de electrónica
Banco Central de Reserva de El Salvador	- Estadísticas económicas (PIB nominal, PIB constante, PIB percápita, exportaciones, importaciones, remesas, indicadores económicos, entre otras). - Estadísticas Monetarias y Financieras.	**Página Oficial:** http://www.bcr.gob.sv **Base de datos económica:** http://www.bcr.gob.sv/bcrsite/?cat=1000&lang=es
Ministerio de Hacienda	- Estadísticas Fiscales (ingresos y gasto público, déficit fiscal, carga tributaria, balance primario, entre otras estadísticas de carácter fiscal)	**Página Oficial:** http://www.mh.gob.sv **Base de datos:** http://www.transparenciafiscal.gob.sv/portal/page/portal/PTF/estadisticas

Dirección General de Estadísticas y Censos de El Salvador	- Estadísticas Sociales (pobreza, empleo, vivienda, poblacionales, educativas, salud y acceso a servicios básicos). -Estadísticas Económicas (Índice de Precios al consumidor, canasta básica, actividad económica por área geográfica)	**Página Oficial:** http://www.digestyc.gob.sv **Base de datos:** **información de la EHPM:** http://www.digestyc.gob.sv/index.php/temas/des/ehpm/resultados-encuesta.html **Estadísticas económicas:** http://www.digestyc.gob.sv/index.php/temas/ee.html **Censo de Población y vivienda 2007 (redatan):** http://www.digestyc.gob.sv/servers/redatam/htdocs/CPV2007P/index.html
Ministerio de Agricultura y Ganadería	Estadísticas Agropecuarias	**Página Oficial:** http://www.mag.gob.sv **Base de datos:** http://www.mag.gob.sv/index.php?option=com_phocadownload&view=section&id=9:estadisticas-de-produccion-agropecuaria&Itemid=221
Fiscalía General de la República Policía Nacional Civil	Estadísticas de Seguridad Pública	**Página Oficial:** http://www.fiscalia.gob.sv **Base de datos:** http://www.fiscalia.gob.sv/index.php/estadisticas-2/ **Página Oficial:** http://www.seguridad.gob.sv **Base de datos:** http://www.pnc.gob.sv/core/index.php/oir/estadisticas
Ministerio de Salud	Estadísticas de Salud	**Página Oficial:** http://www.salud.gob.sv **Base de datos:** http://www.salud.gob.sv/index.php/oir/estadisticas
Instituto Salvadoreño del Seguro Social	Estadísticas de Seguridad Social	**Página Oficial:** www.isss.gob.sv **Base de datos:** http://www.isss.gob.sv/index.php?option=com_docestandar&view=docestandar&categoria=44&showview=0&Itemid=234
Ministerio de Turismo	Estadísticas de Turismo	**Página Oficial:** http://www.mitur.gob.sv **Base de datos:** http://www.mitur.gob.sv/uaip/estadisticas.html

APENDICE 2: Método de Predicción

Método de Regresión Lineal: La regresión lineal es una alternativa bastante sencilla, que permite realizar estimaciones, a partir de una línea recta o ecuación matemática, que describe la relación entre dos variables. En este método, se procede a analizar una serie cronológica, es decir un conjunto de datos u observaciones, ordenados en términos de tiempo o en otra variable que influya en la variable que se requiere analizar. En nuestro caso, puede tratarse de una serie de datos sobre gastos o ingresos presupuestarios.

REGRESIÓN EN FUNCIÓN DEL TIEMPO

FORMULA: $Y = a + bX$

donde:

Con los datos del ejemplo:

$Y = 665.35 + 88.41\ X$

$Y = 665.35 + 88.41\ (14) = 1{,}903.09$
$Y = 665.35 + 88.41\ (15) = 1{,}991.5$
$Y = 665.35 + 88.41\ (16) = 2{,}079.91$
$Y = 665.35 + 88.41\ (17) = 2{,}168.32$

X= DATOS DEL PERIODO EN ESTUDIO (VARIABLE INDEPENDIENTE)

Y= DATOS DE ESTUDIO (VARIABLE DEPENDIENTE)

n= NUMERO DE PERIODOS

$$a = \frac{\sum Y - b\sum X}{n} \qquad b = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - (\sum x)^2}$$

TABLA DE DATOS:

AÑO	X (PERIODO)	Y (DATOS)	XY	X^2
2000	1	799.8	799.8	1.0
2001	2	866.3	1,732.6	4.0
2002	3	903.9	2,711.7	9.0
2003	4	960.3	3,841.2	16.0
2004	5	1,026.2	5,131.0	25.0
2005	6	1,169.9	7,019.4	36.0
2006	7	1,362.5	9,537.5	49.0
2007	8	1,506.8	12,054.4	64.0
2008	9	1,615.2	14,536.8	81.0
2009	10	1,256.2	12,562.0	100.0
2010	11	1,565.6	17,221.1	121.0
2011	12	1,801.5	21,618.0	144.0
2012	13	1,860.7	24,189.1	169.0
$\sum$	91	16,694.9	132,954.6	819.0
2013	14	1,903.1		
2014	15	1,991.5		
2015	16	2,079.9		
2016	17	2,168.3		

LOS DATOS DE ESTUDIO QUE SE REQUIEREN OBTENER SE OBTIENEN SUSTITU
LOS COEFICIENTES "a" Y "b", y "X=14,15,16 Y 17" , EN ECUACION Y=A+BX

$$b = \frac{209{,}177.80}{2366}$$

$$b = 88.41$$

$$a = \frac{\sum Y - b\sum X}{n}$$

$$a = \frac{8649.55}{13}$$

$$a = 665.35$$

REGRESIÓN LINEAL CONSIDERANDO AL PIB COMO VARIABLE INDEPENDIENTE

FORMULA: $Y = a + bX$ → Con los datos del ejemplo:
$Y = -455.05 + 0.095\,X$

donde:

$Y = -455.05 + 0.095\,(26{,}163.4) = 2{,}030.5$
$Y = -455.05 + 0.095\,(28{,}016.3) = 2{,}206.5$
$Y = -455.05 + 0.095\,(29{,}942.4) = 2{,}389.5$
$Y = -455.05 + 0.095\,(32{,}068.5) = 2{,}591.5$

X= DATOS DEL PIB A PRECIOS CORRIENTE (VARIABLE INDEPENDIENTE)

Y= DATOS DE ESTUDIO (VARIABLE DEPENDIENTE)

n= NUMERO DE PERIODOS

$$a = \frac{\sum Y - b\sum X}{n} \qquad b = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - (\sum x)^2}$$

TABLA DE DATOS:

AÑO	X (PIB a Precios Corrientes)	Y (DATOS)	XY	X^2
2000	13,134.1	799.8	10,504,653.2	172,504,582.8
2001	13,812.7	866.3	11,965,942.0	190,790,681.3
2002	14,306.7	903.9	12,931,826.1	204,681,664.9
2003	15,046.7	960.3	14,449,346.0	226,403,180.9
2004	15,798.3	1,026.2	16,212,215.5	249,586,282.9
2005	17,093.8	1,169.9	19,998,036.6	292,197,998.4
2006	18,550.7	1,362.5	25,275,328.8	344,128,470.5
2007	20,104.9	1,506.8	30,294,063.3	404,207,004.0
2008	21,431.0	1,615.2	34,615,351.2	459,287,761.0
2009	20,661.0	1,256.2	25,954,348.2	426,876,921.0
2010	21,427.9	1,565.6	33,546,448.8	459,154,898.4
2011	23,054.1	1,801.5	41,531,961.2	531,491,526.8
2012	24,304.0	1,860.7	45,222,452.8	590,684,416.0
∑	238725.9	16,694.9	322,501,973.7	4,551,995,388.9
2013	26,163.4	2,030.5		
2014	28,016.3	2,206.5		
2015	29,942.4	2,389.5		
2016	32,068.5	2,591.5		

← LOS DATOS DE ESTUDIO QUE SE REQUIEREN OBTENER SE OBTIENEN SUSTITUYENDO LOS VALORES OBTENIDOS PARA LOS COEFICIENTES "a" Y "b", y "X=14,15,16 Y 17" , EN ECUACION Y=A+BX

$$b = \frac{207{,}032{,}566.16}{2185884725}$$

$$b = 0.095$$

$$a = \frac{\sum Y - b\sum X}{n}$$

$$a = \frac{-5915.690764}{13}$$

$$a = -455.05$$

Método de Tendencias: Este método se fundamenta en gran medida en la tendencia histórica de los datos, con énfasis en el primero y el último dato de la serie histórica, sobre el cálculo de una tasa de variación (de carácter compuesto), llamada Tasa Promedio de Crecimiento (Tpc). Una vez obtenida esa Tasa, es posible estimar el dato para el siguiente periodo. El dato obtenido en esta forma, es producto de la tendencia de la serie histórica, por lo cual será alcista si la serie ha sido creciente, y decrementalista si la serie es decreciente. Como en todos los métodos, sus resultados se deberán tomar con cautela, y compararse con los obtenidos por otros métodos, con las condiciones macroeconómicas previstas por el BCR, y con variaciones previstas los precios, cobros por servicios, tarifas, aranceles, etc.

$Y_{(n+1)} = Yt + Tpc\,(Yt)$

Tpc: Tasa Promedio de Crecimiento

Formula para obtener la Tpc:

$$Tpc = [\,(Y_t \,/\, Y_o)^{1/n} - 1.0\,] \times 100$$

donde:

$Y_{(n+1)} =$ = Dato estimado por Tendencia

Y_t = Dato real del ultimo periodo

Y_o = Dato real del primer periodo

n = Numero de periodos en analisis

$$Tpc = [\,(1{,}860.7 \,/\, 799.8)^{1/13} - 1.0\,] \times 100$$
$$Tpc = [\,(2.32)^{1/13} - 1.0\,] \times 100$$
$$Tpc = [\,1.067105 - 1.0\,] \times 100$$
$$Tpc = [\,0.067105\,] \times 100$$
$$Tpc = 6.71053\ \%$$

Tabla de Datos:

AÑO	n	DATOS REALES Y
2000	1	799.8
2001	2	866.3
2002	3	903.9
2003	4	960.3
2004	5	1,026.2
2005	6	1,169.9
2006	7	1,362.5
2007	8	1,506.8
2008	9	1,615.2
2009	10	1,256.2
2010	11	1,565.6
2011	12	1,801.5
2012	13	1,860.7
2013	14	1,985.6
2014	15	2,118.8
2015	16	2,260.9
2016	17	2,412.6

año 2013 = $Y_{(n+1)} = Yt + Tpc\,(Yt)$

$Y_{(n+1)} = 1{,}860.7 + 0.067105*(1860.7)$

$Y_{(n+1)} = 1{,}985.6$

año 2014 = $Y_{(n+1)} = 1{,}985.6 + 0.0671053*(1{,}985.6)$

$Y_{(n+1)} = 2{,}118.8$

año 2015 = $Y_{(n+1)} = 2{,}118.8 + 0.067105*(2{,}118.8)$

$Y_{(n+1)} = 2{,}260.9$

año 2016 = $Y_{(n+1)} = 2{,}260.9 + 0.067105*(2{,}260.9)$

$Y_{(n+1)} = 2{,}412.6$

Suavizado Exponencial: Este método contiene un mecanismo de autocorrección que ajusta los pronósticos en dirección opuesta a los errores pasados. Es un caso particular de promedios móviles ponderados de los valores actuales y anteriores en el cual las ponderaciones disminuyen exponencialmente. Se emplea tanto para suavizar como para realizar pronósticos.

$$Y_{t+1} = \alpha \cdot X_t + (1-\alpha) \cdot Y_t$$

Donde:

Y_{t+1} = pronóstico para cualquier período futuro.
α = constante de suavización, a la cual se le da un valor entre 0 y 1.
X_t = valor real para el período de tiempo.
Y_t = pronóstico hecho previamente para el período de tiempo

En este caso se tomara un α =0.4

AÑO	X (PERIODO)	Y (DATOS)	Cálculo	Pronosticos
2000	1	799.8	año base	
2001	2	866.3	2001=2000	799.8
2002	3	903.9	0.4*(866.3)+(1-0.4)*(799.8)	826.4
2003	4	960.3	0.4*(903.9)+(1-0.4)*(826.4)	857.4
2004	5	1,026.2	0.4*(960.3)+(1-0.4)*(857.4)	898.6
2005	6	1,169.9	0.4*(1026.2)+(1-0.4)*(898.6)	1,037.7
2006	7	1,362.5	0.4*(1169.9)+(1-0.4)*(1037.7)	1,090.6
2007	8	1,506.8	0.4*(1362.5)+(1-0.4)*(1090.6)	1,199.4
2008	9	1,615.2	0.4*(1506.8)+(1-0.4)*(1199.4)	1,322.4
2009	10	1,256.2	0.4*(1615.2)+(1-0.4)*(1322.4)	1,439.5
2010	11	1,565.6	0.4*(1256.2)+(1-0.4)*(1439.5)	1,366.2
2011	12	1,801.5	0.4*(1565.6)+(1-0.4)*(1366.2)	1,446.0
2012	13	1,860.7	0.4*(1801.5)+(1-0.4)*(1446)	1,588.2
Valores Estimados				
2013	14	1,985.6	0.4*(1860.7)+(1-0.4)*(1588.2)	1,697.2
2014	15	2,118.8	0.4*(1985.6)+(1-0.4)*(1697.2)	1,812.6
2015	16	2,260.9	0.4*(2118.8)+(1-0.4)*(1812.6)	1,935.1
2016	17		0.4*(2,260.9)+(1-0.4)*(1935.1)	2,065.4

Pronósticos realizados por suavizado exponencial

Para estimar los pronosticos por este metodo se tomaron de base para los años 2013-2015 los realizados por el metodo de tendencia