Ministerio de Hacienda

Dirección General de Presupuesto

Módulo III

ACTUALIZACION DE CONOCIMIENTOS GENERALES

Curso 4

Conceptos y Métodos Básicos de Estadística

Tiempo total

4 horas

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

Contenido

- **LECCION 1: Consideraciones Generales de la Estadística.**
- **LECCION 2: Tabulación y Gráficos**
- **LECCION 3: Medidas de Posición Centrales y no Centrales**
- **LECCION 4: Análisis de Regresión**
- **APENDICE 1: Fuentes de Búsqueda de Información**
- **APENDICE 2. Métodos de Predicción**

Objetivos del Plan de Reforma del Presupuesto Público

- **Objetivo estratégico 1:** Aumentar la eficiencia en el uso de los recursos públicos, financiando más y mejores servicios bajo condiciones de equilibrio y sostenibilidad fiscal.

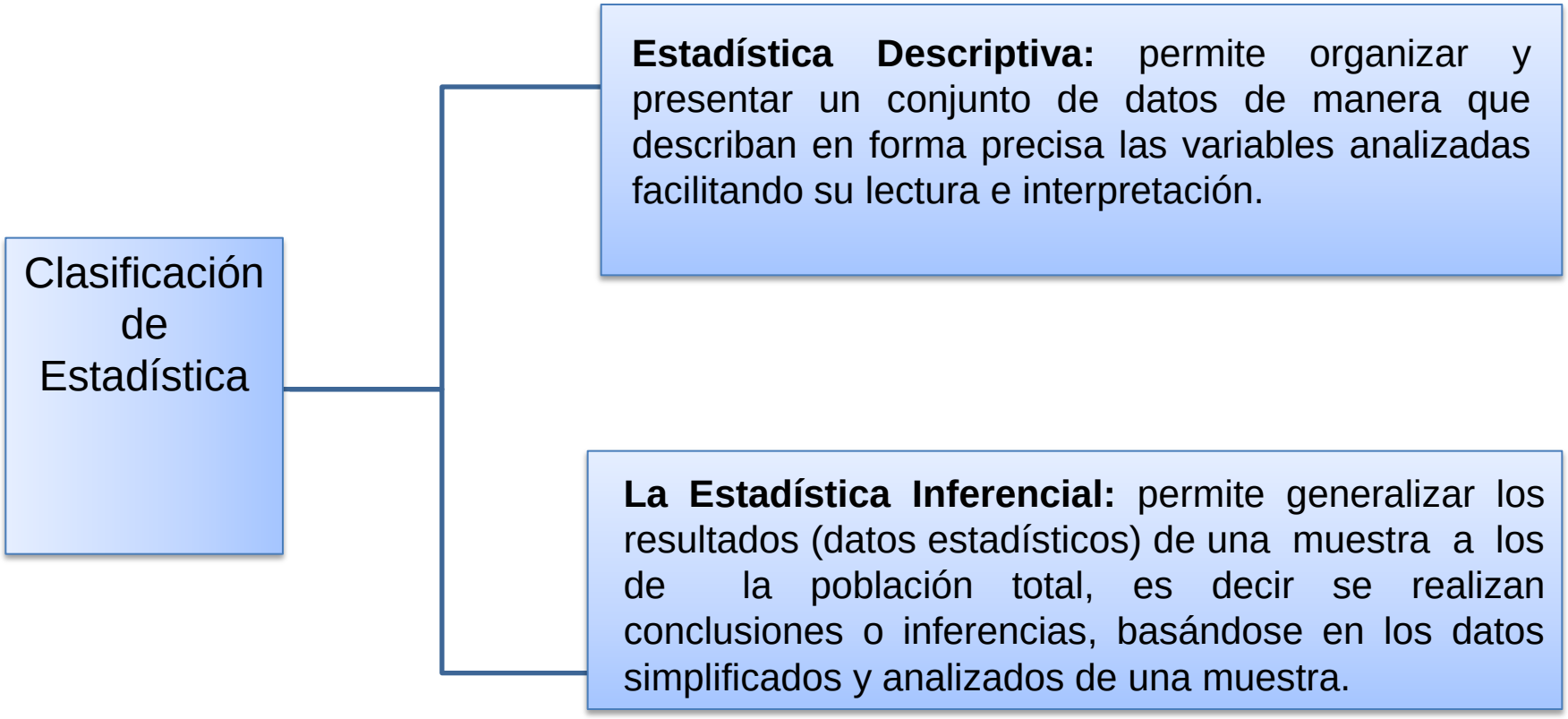
- **Objetivo estratégico 2:** Mejorar la asignación de los recursos presupuestarios en función de las prioridades y metas de un desarrollo sostenido del país.

- **Objetivo estratégico 3:** Transformar el presupuesto público en un instrumento de gerencia, transparencia y rendición de cuentas.

- **Objetivo estratégico 4:** Crear la capacidad fiscal en el país para afrontar situaciones de emergencia derivadas de crisis económica y/o desastres naturales.

Lección 1: Consideraciones Generales de la Estadística.

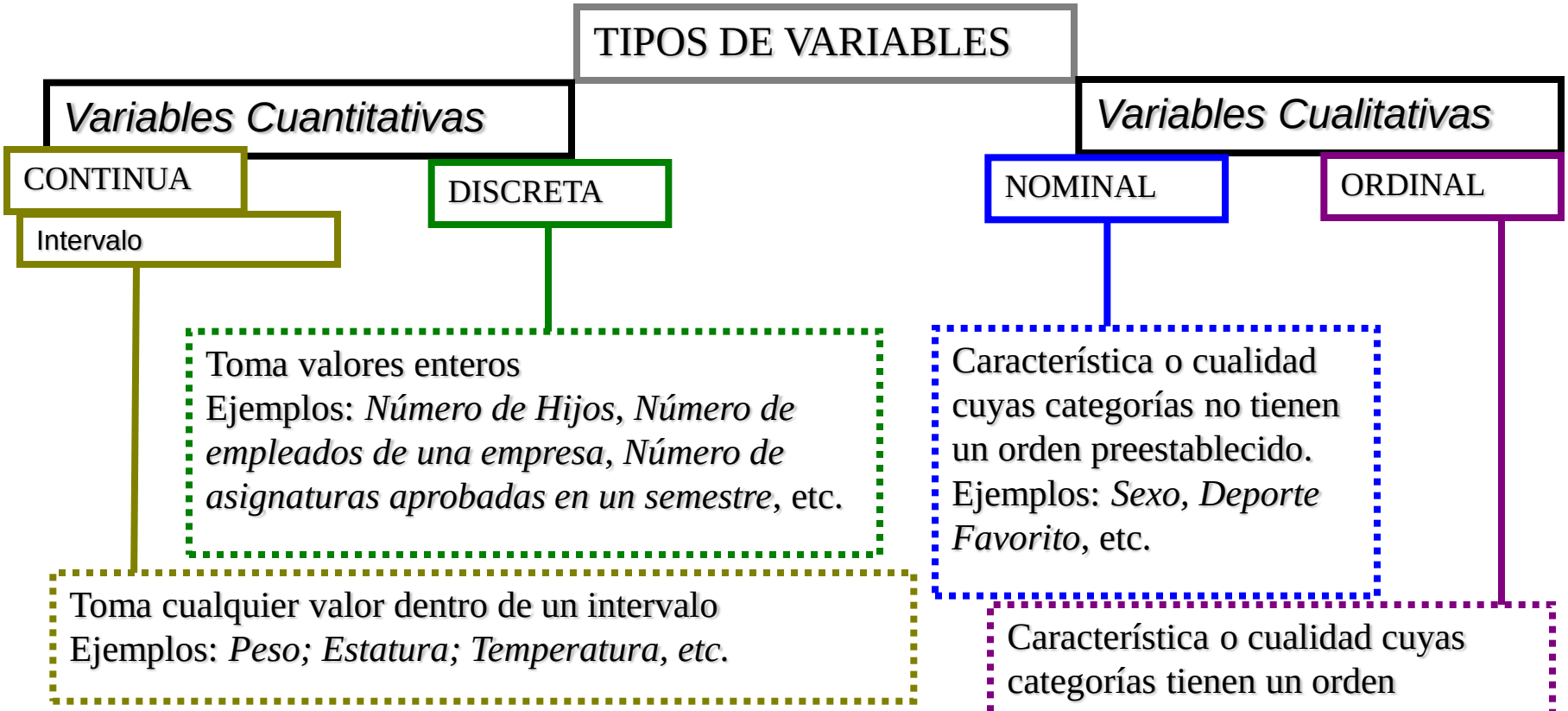
Punto 1: Clasificación de la Estadística



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

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

Variable: corresponde a la característica de la Unidad de Análisis



Unidad de Medida: Gramos o Kilos para la variable Peso; Grados C o F para Temperatura

Lección 1: Consideraciones Generales de la Estadística. Discusión

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.

Proponer más ejemplos de variables ordinales y nominales

LECCION 2: Tabulación y 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.

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

Ejemplo: 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".

Se observa que 7 familias tienen solamente un hijo. También, ((4+7+8+5)/37)*100 = 64.86%, aproximadamente el 65% de las familias tienen 3 ó menos hijos.

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

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.

- Por ejemplo el gobierno realiza una muestra a 10 personas con el propósito de medir el grado de aceptación que tendría si se construyera una carretera cercana al 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: Tabla de Frecuencias

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), la mayoría de las respuestas se centraron en Excelente y Bueno (grado de aceptación iguales a 5 y 4 respectivamente), por lo que se puede concluir que la mayoría de las personas encuestadas tienen una **visión favorable del proyecto de construcción de la carretera.**

Tabla de Frecuencias

- **Frecuencia Absoluta Acumulada (Fa):** Esta frecuencia se calcula sumando el acumulado de las frecuencias de los intervalos anteriores más la frecuencia absoluta del intervalo actual.

$$Fa = Fa_{-1} + f_i$$

- **Frecuencia Relativa** (h): Equivale a la razón de las frecuencias de cada intervalo sobre la totalidad de los datos.

$$h_i = f_i / n$$

- **Frecuencia Relativa Acumulada (Hi): S**u cálculo resulta de la suma del acumulado de las frecuencias relativas de los intervalos anteriores más la frecuencia relativa del intervalo actual.

$$Hi = Hi\text{-}1 + hi$$

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**	

Tabla de frecuencia agrupadas

- **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).

- **Número 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.

$$Nc = \sqrt{n}$$

- Cuando se trabajan con muestras mayores a 225, la fórmula obtiene un Nc superior a 15, por tanto, recomendaremos para estos casos la siguiente fórmula:

$$1 + 3{,}22 \log n$$

- **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)

$$A = R / Nc$$

Tabla de frecuencia agrupadas

- **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

- **PASO 1: Determinar el número de intervalos (Nc).**

Optaremos por utilizar la primera fórmula expuesta:

$$\mathbf{Nc = \sqrt{24} = 4.898 \approx 5}$$

- **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:

$$A = 420/5 = 84$$

- **Paso 3: Determinar los intervalos de clases**

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

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

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

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

- De acuerdo a los pasos anteriores 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		

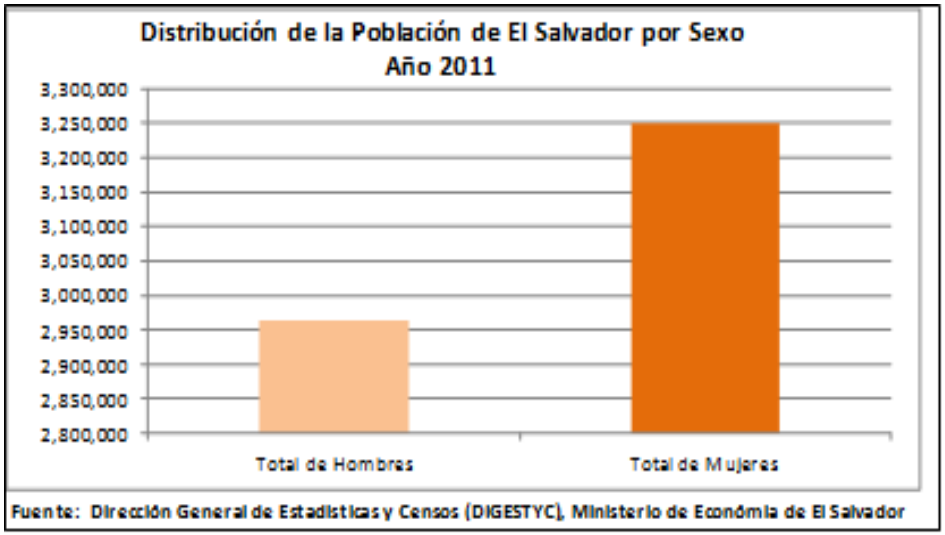
Punto 2: Gráficos

- **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.

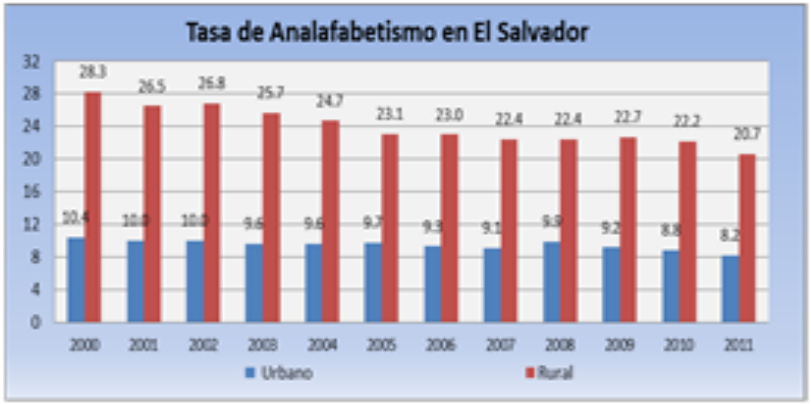
Existen tres principales clases de gráficos de barras:

Barra simple: se emplean para graficar hechos únicos. Ejemplo:



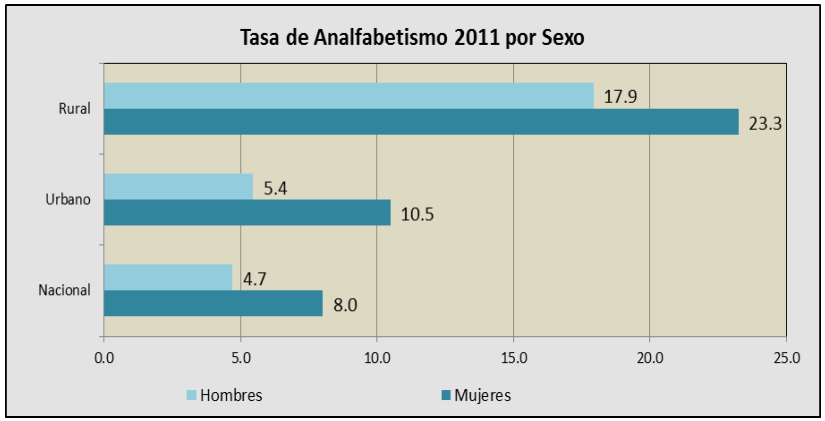
Barras múltiples: es muy recomendable para comparar una serie estadística con otra, para ello emplea barras simples de distinto color o tramado en un mismo plano cartesiano, una al lado de la otra.

•Ejemplo N° 1:



Fuente: Dirección General de Estadísticas y Censos (DIGESTYC), Ministerio de Economía de El Salvador

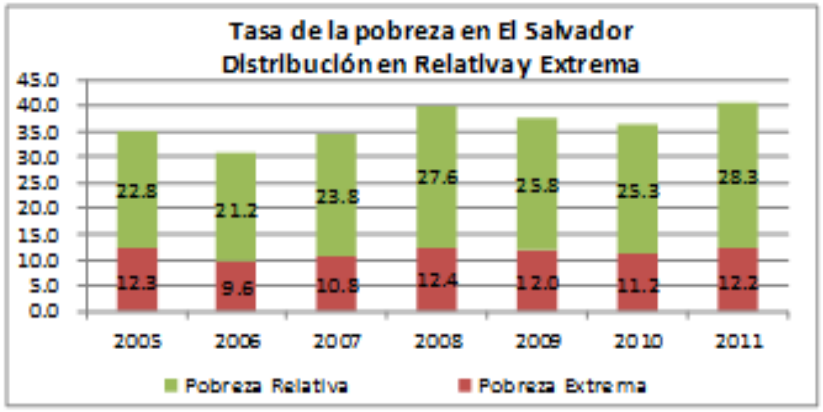
Ejemplo N° 2:



Fuente: Dirección General de Estadísticas y Censos (DIGESTYC), Ministerio de Economía 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 Estadísticas y Censos (DIGESTYC), Ministerio de Economía de El Salvador

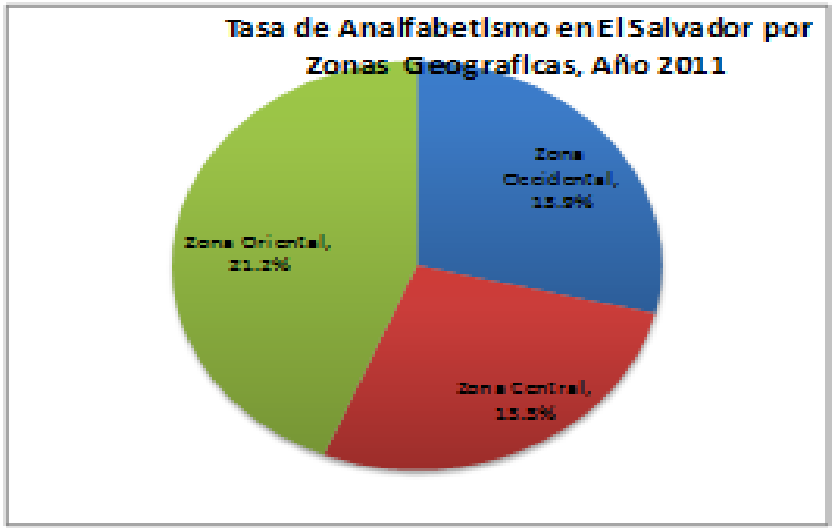
Diagramas de sectores (también llamados gráfico circular)

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 . 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 cuasicualitativos.
- 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:

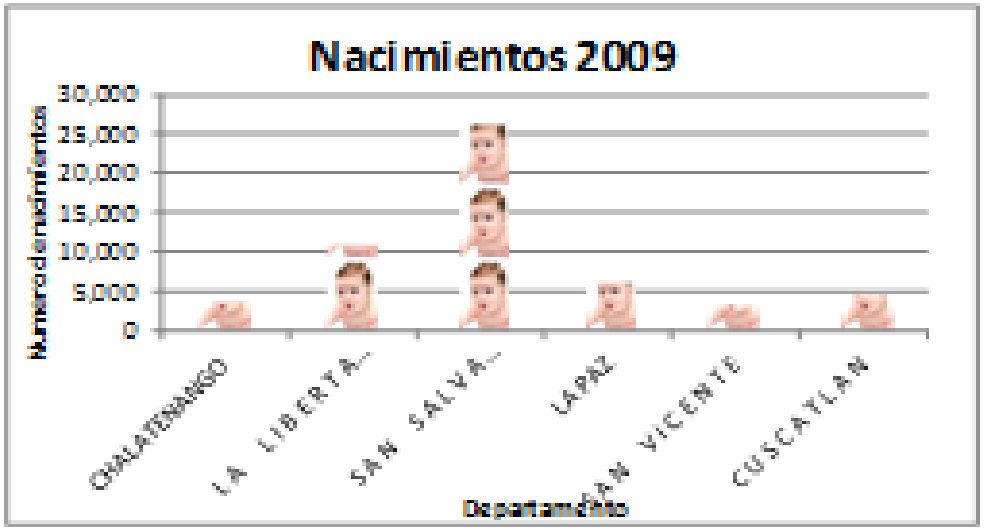


Fuente: Ministerio de Economía de El Salvador, Dirección General de Estadísticas y Censos, Encuesta de Hogares y Propósitos Múltiples 2011.

Pictograma

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.

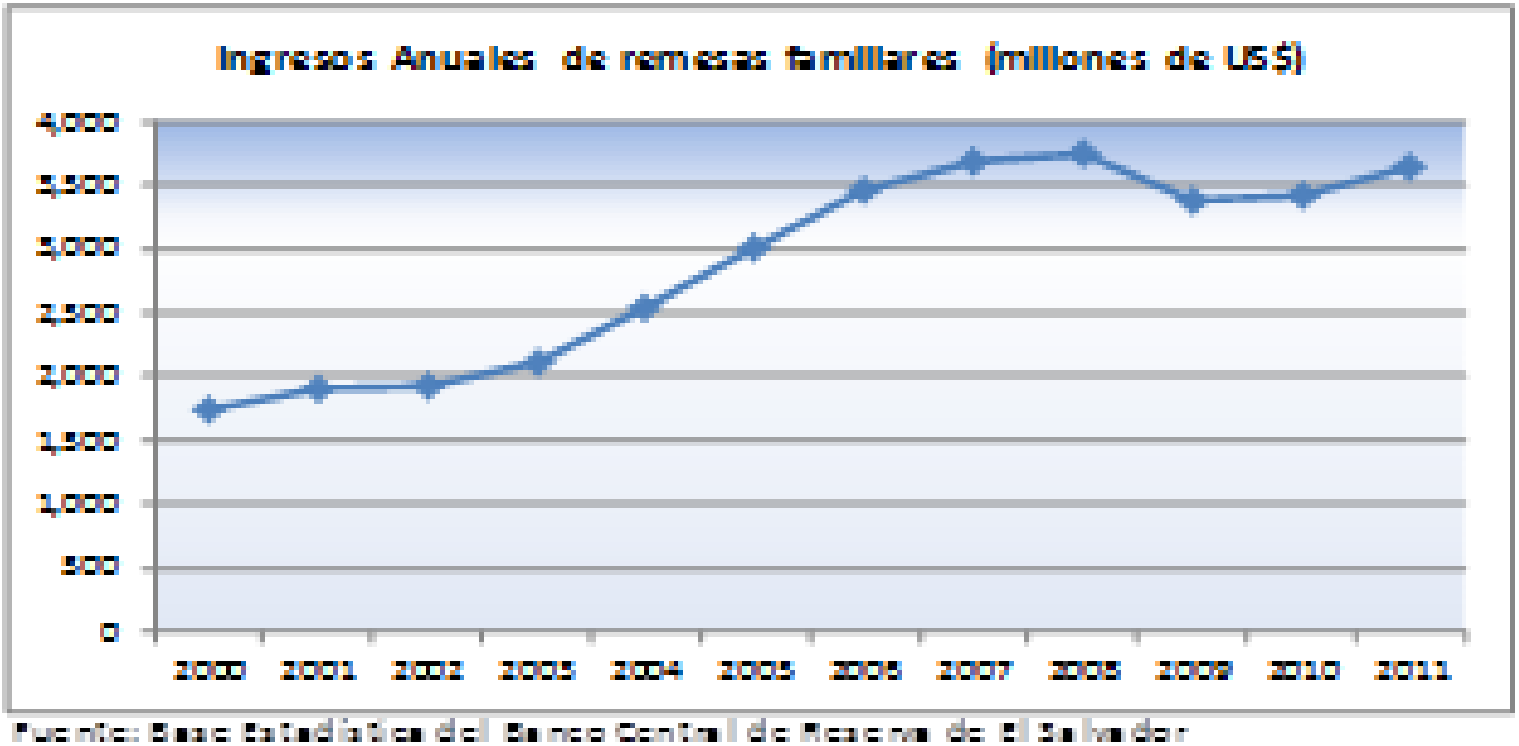
- Ejemplo de Pictograma: Plantación de árboles y nacimientos 2009.



Breve descripción de los gráficos más utilizados:

Gráfico 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 de remesas**

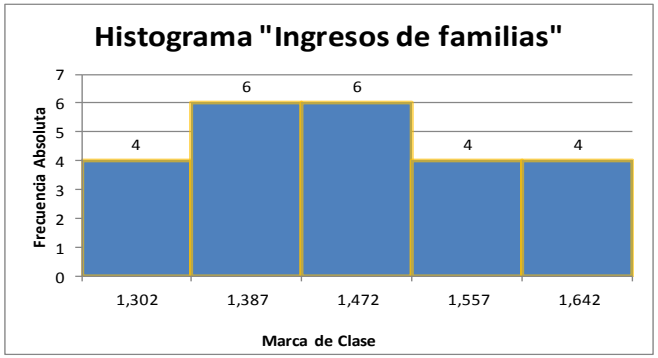


Fuente: Base Estadística del Banco Central de Reserva de El Salvador

Histograma: Se utiliza 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.

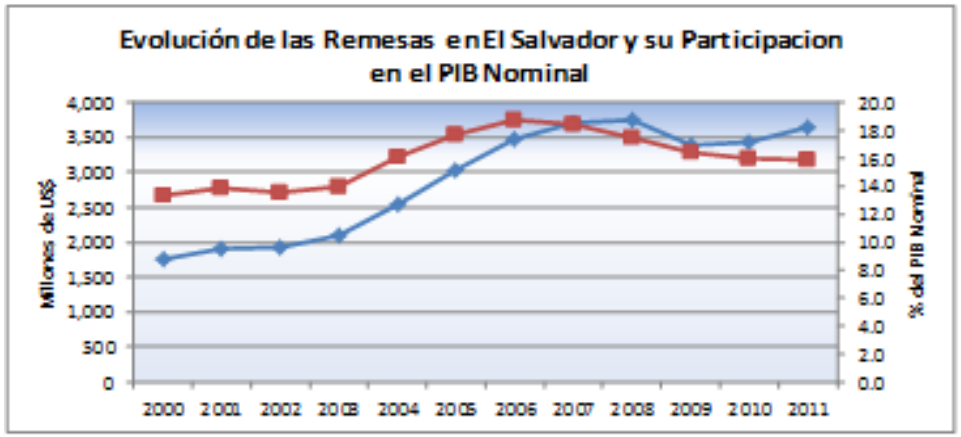
Ejemplo de histograma con la información siguiente: Ingresos semanales en US$ de 24 familias.

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	



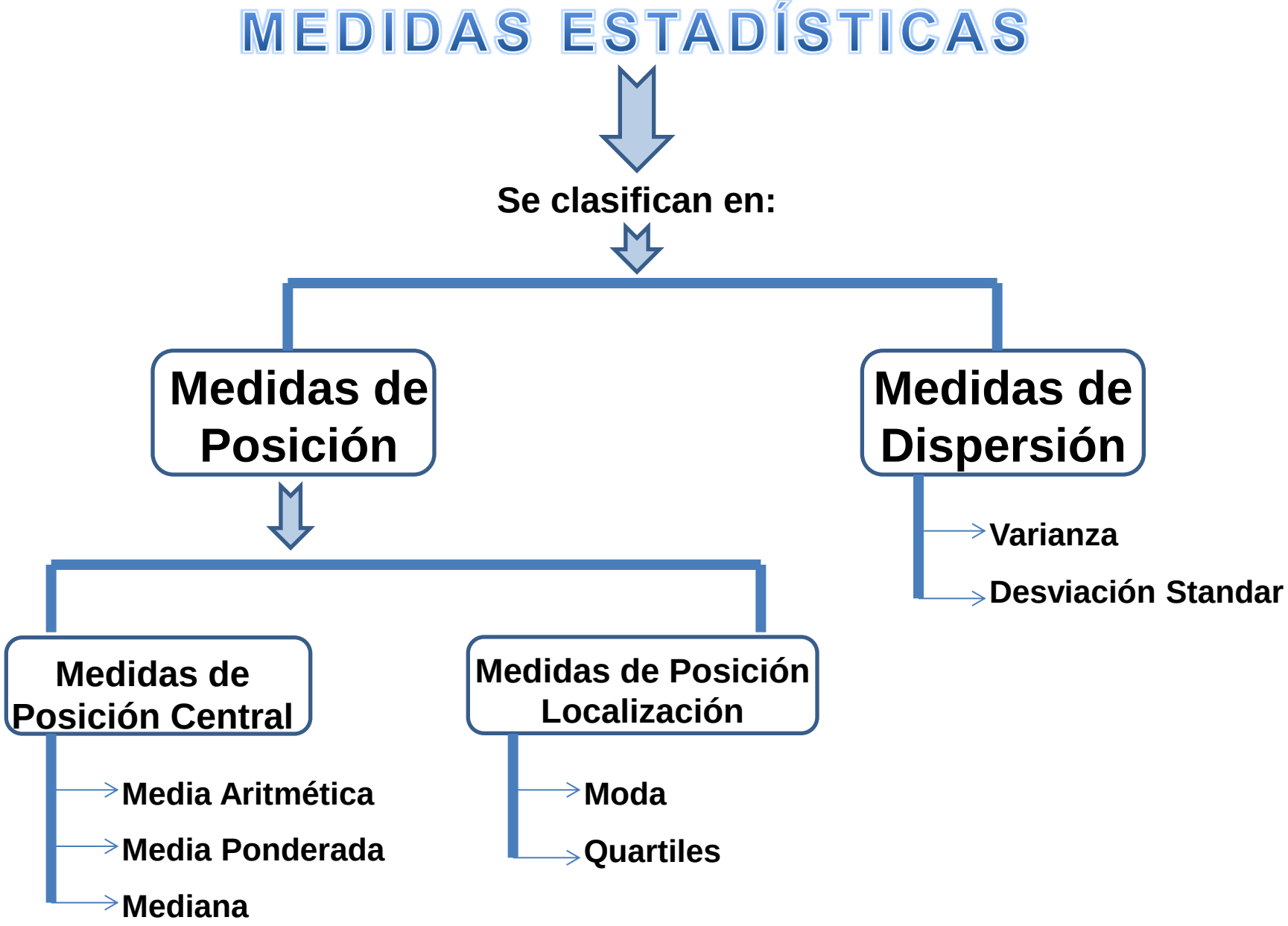
- **Gráficos que Representan Dos tipos de Escalas (utilización de eje principal y eje secundario):** este tipo de gráficos es muy utilizado cuando se quiere presentar información que utiliza dos tipos de escala de medición diferentes por ejemplo cantidades porcentajes y números enteros.

Ejemplo:



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

LECCION 3: Medidas de Posición Centrales y no Centrales



LECCION 3: Medidas de Posición Centrales y no Centrales

Punto 1: Medidas de Posición Central

Los promedios o medidas de posición proporcionan valores típicos o representativos de la variable en estudio. a continuación se describe un breve resumen del mismo.

Media aritmética

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

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

- **Ejemplo.** Supongamos que en un almacén tienen 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"**

LECCION 3: Medidas de Posición Centrales y no Centrales

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:

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

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).

Ejemplo:

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

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

La Media Geométrica (MG)

$$MG = \sqrt[n]{(X_1)(X_2)\cdots(X_n)}$$

- Existen dos usos principales de la media geométrica:
  - Para promediar porcentajes, índices y cifras relativas y
  - 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.

La Moda:

"La moda se define como aquel valor de la variable o del atributo que presenta 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 el número de datos es impar: 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.**

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 (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

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.

Matemáticamente la varianza y desviación estándar están dadas por:

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

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

Punto 2. Medidas de dispersión

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

Punto 2. Medidas de dispersión

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?

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.

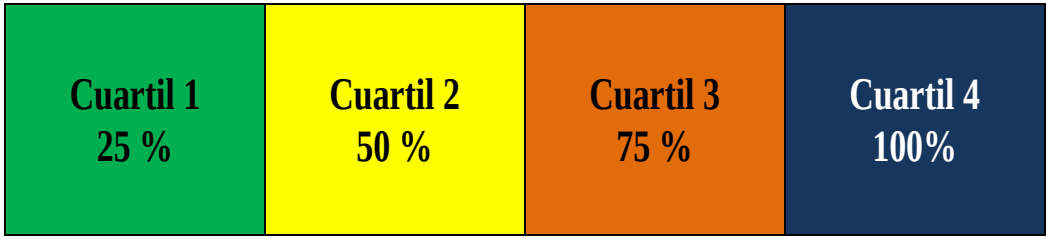
$$C.V_B = \frac{18.99}{59.28} = 0.32 = 32\%$$ la media de B tiene representatividad dudosa.

Se recomienda el candidato A

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

Cuartiles

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:



- **Fórmula General:**

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

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

En 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:

- **$Q_1 = 1\,(N / 4)$** **$Q_2 = 2\,(N / 4)$** **$Q_3 = 3\,(N / 4)$** **$Q_4 = 4\,(N / 4)$**

Cuartiles:

- **Ejemplo:**
- Calcular Q_1, Q_2 y Q_3 del salario en dólares de 15 personas: 300, 275, 180, 325, 200, 250, 350, 260, 280, 310, 400, 380, 260, 290, 370.
- Recordemos que: **$Q_1 = 1\ (N / 4)$, $Q_2 = 2\ (N / 4)$, $Q_3 = 3\ (N / 4)$, $Q_4 = 4\ (N / 4)$**
- datos ordenados de menor a mayor.
- 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

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

- $D_1 = \frac{1(15)}{10} = 1.5$, 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

Percentiles:

- Se representan con la letra P. Son 99 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/100.
- **Ejemplo**
- 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: $P_{50} = 290$

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

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

LECCION 4: Análisis de Regresión

- ***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.

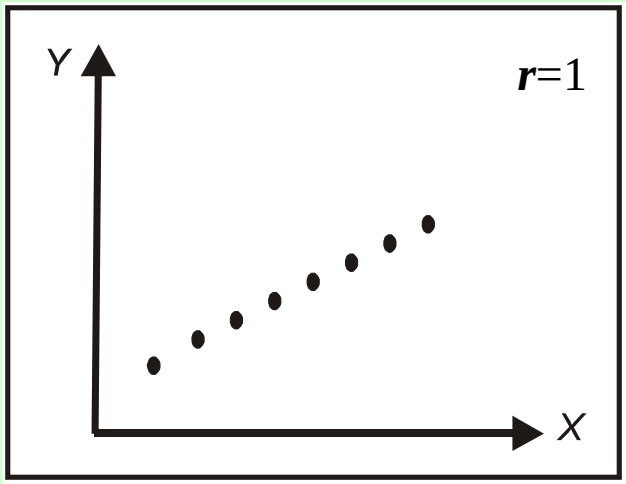
Diagrama de Dispersión:

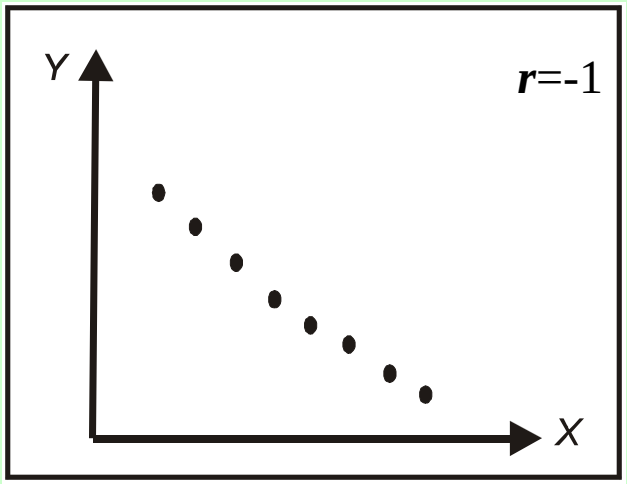
- El diagrama de dispersión es la gráfica donde se encuentran todos los puntos de las observaciones, tanto de la variable dependiente (Y), como de la variable independiente (X).
- El diagrama de dispersión puede revelarnos dos tipos de información:

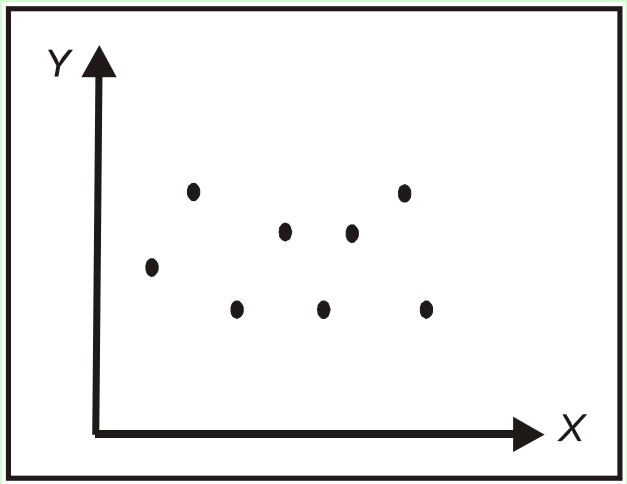
  a) Relación de las variables

  b) Tipo de línea o ecuación de estimación

Ejemplos de Diagramas de Dispersión







Utilidad de los Análisis de Regresión y Correlación

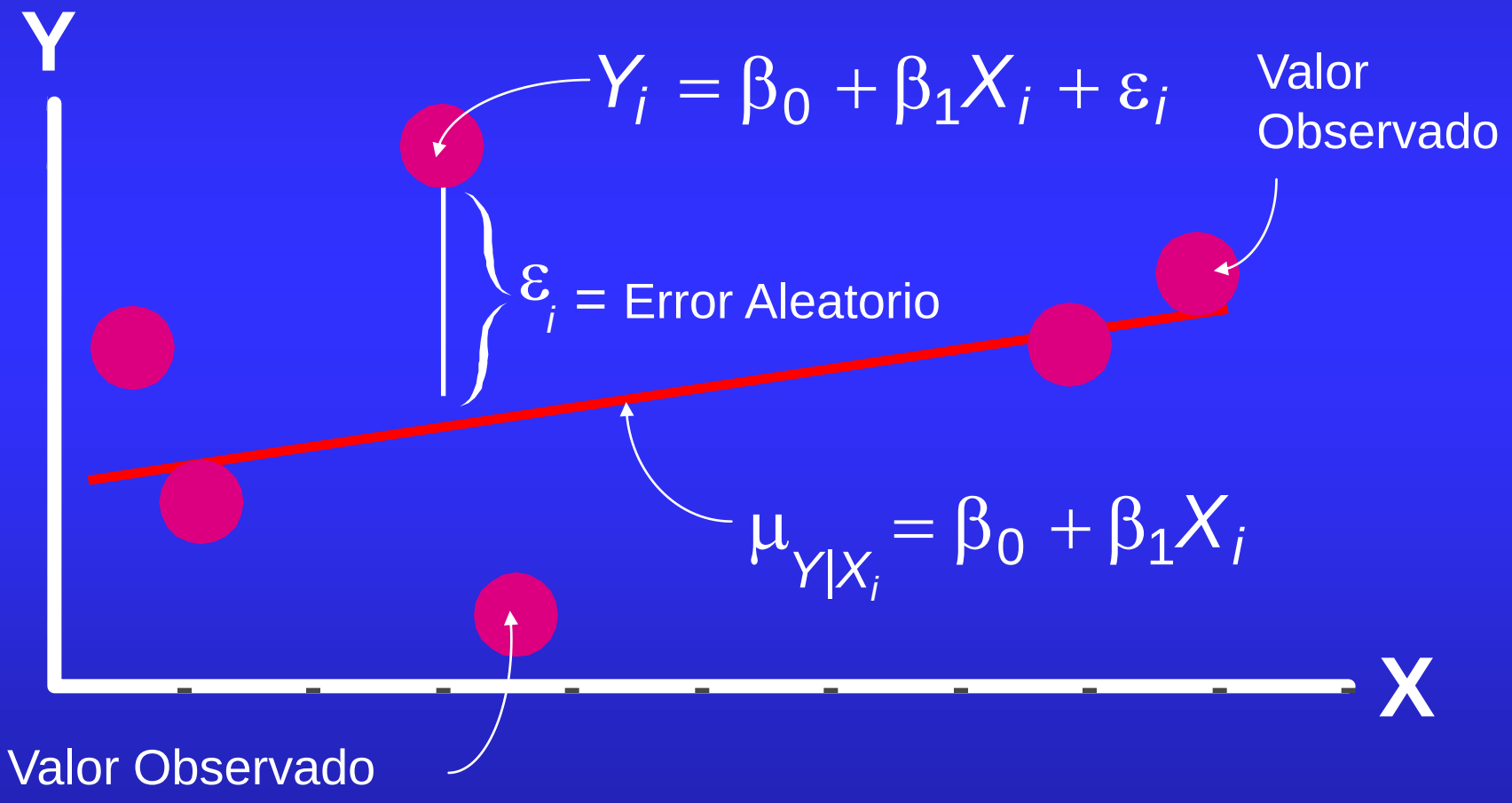
- El **Análisis de Regresión** es utilizado para la

  - **Predicción** : El modelo estadístico se usa para predecir (por interpolación o extrapolación) los valores de una variable dependiente o respuesta basados en los valores de, al menos, una variable independiente o explicativa y

  - **Estimación** de los efectos de las variables explicativas de interés en la variable respuesta.

- El **Análisis de Correlación** es utilizado para medir y evaluar estadísticamente la magnitud de la asociación entre variables numéricas.

Modelo de Regresión Lineal Poblacional



Estimación de coeficientes

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

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

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

Se observa que es necesario calcular cinco cantidades para determinar a y b:

$$n, \Sigma X, \Sigma Y, \Sigma X^2 \text{ y } \Sigma XY$$

Ejemplo (Gastos e ingresos)

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 ($Yt=a+b\cdot Xt$), determine:

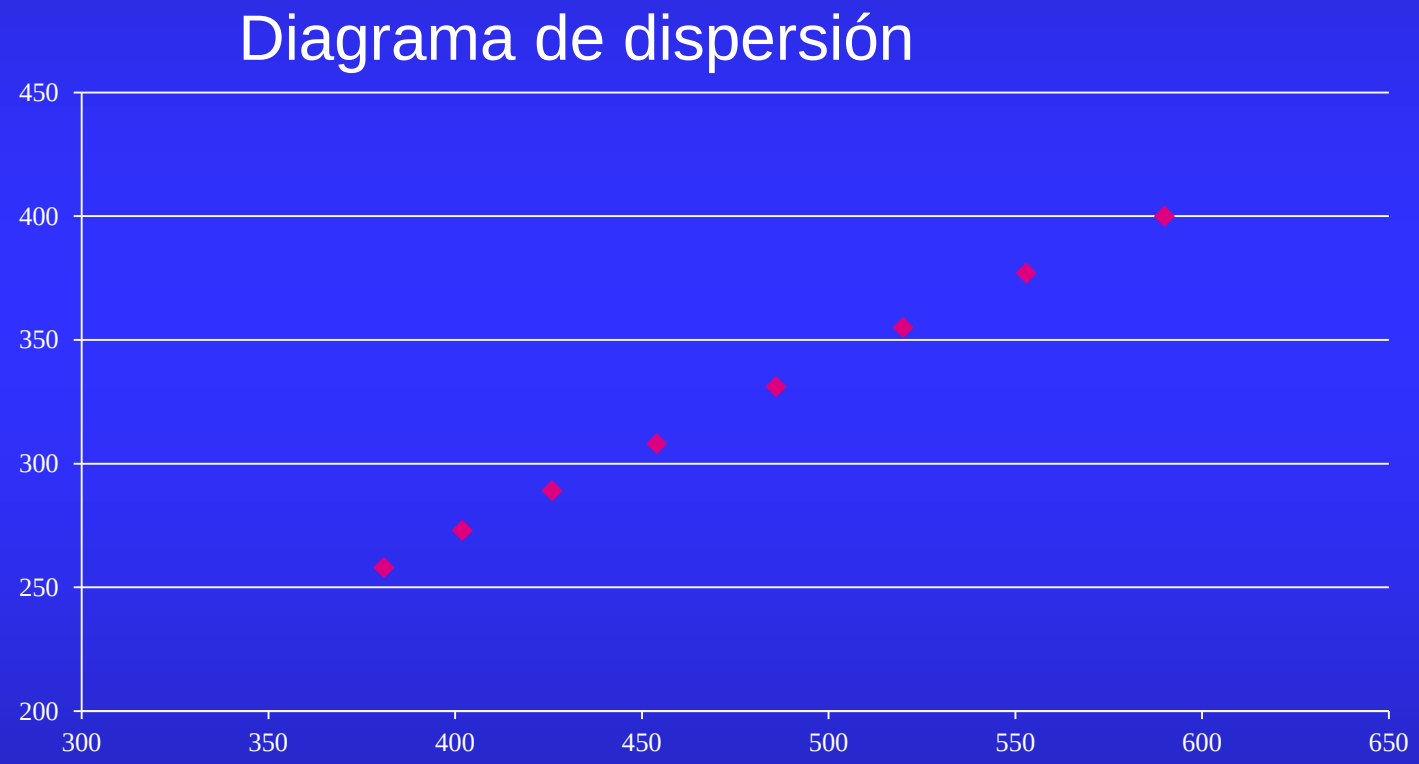
a) Los estimadores de los parámetros a y b de la recta de regresión.

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.

Solución Ejemplo (Gastos e ingresos)

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



Ajuste del modelo: y = 0.6848x - 2.4204 R^2 = 0.9997

Coeficiente de determinación (r^2): una vez ajustada la recta de regresión a la nube de observaciones es importante disponer de una medida que mida la bondad del ajuste y que permita decidir si el ajuste lineal es suficiente o se deben buscar modelos alternativos.

Parámetros para determinar la significancia del coeficiente de determinación

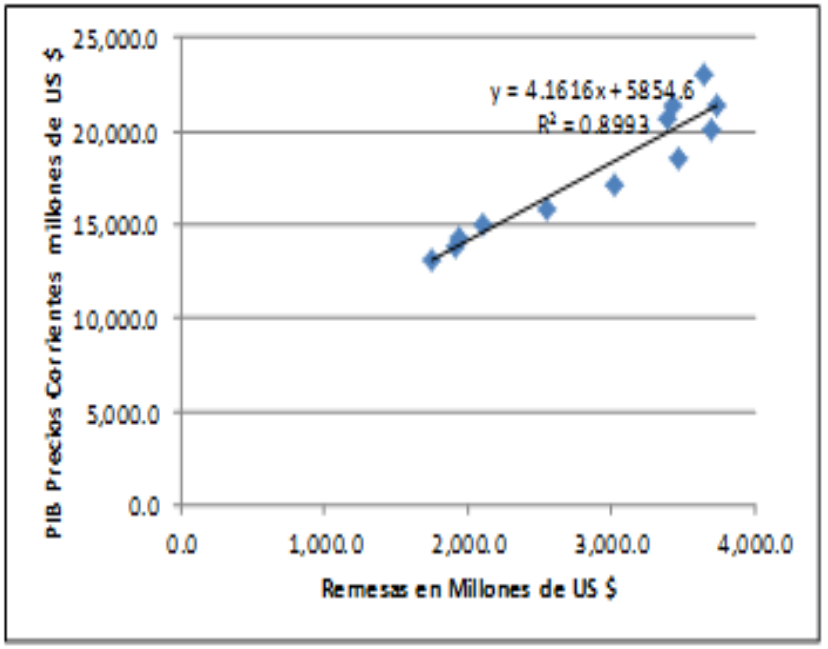
Valor	Significado
0	Correlación nula
0.25 - 0.49	Correlación débil
0.5 - 0.74	Correlación Moderada
0.75 - 0.99	Correlación Intensa
1	Correlación Perfecta

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

- **Ejemplo práctico de Regresión Lineal: PIB a precios corrientes y Flujo anual de remesas familiares**

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



- ¿Qué tipo de diagrama de dispersión se presenta?
- ¿Qué tipo de coeficiente r presenta la gráfica?
- ¿Qué tipo de significancia según el coeficiente **r^2** ? ¿baja, moderada, nula, intensa?
- Interprete la ecuación obtenida

Módulo III
ACTUALIZACION DE CONOCIMIENTOS GENERALES

Curso 4

Conceptos y Métodos Básicos de Estadística

Tiempo total
4 horas

GRACIAS!