MANUAL

ESTADISTICA

1. Conceptos generales
1.1 Antecedentes históricos
1.2 Representación e interpretación de graficas
1.3 Elementos fundamentales
1.4 Distribución de frecuencia
1.5 Organización de la información

2. Medidas de tendencia central
2.1 Media aritmética
2.2 Mediana
2.3 Moda
2.4 Media geométrica
2.5 Media armónica

3. Medidas de dispersión
3.1 Rango
3.2 Desviación media
3.3 Desviación típica
3.4 Varianza

4. Medidas de forma
4.1 Sesgo
4.2 Apuntamiento
4.3 Momentos

5. Medidas de correlación
5.1 Coeficiente de correlación
5.2 Recta de regresión
5.3 Error estándar de estimación

Unidad I Conceptos generales

ESTADISTICA

Propósito general.

Desarrollar la capacidad del razonamiento formal por medio de las herramientas básicas de la estadística descriptiva para muestrear, procesar y comunicar información social y científica, para la toma de decisiones en la vida cotidiana en un clima de colaboración.

HISTORIA DE LA ESTADÍSTICA

Como dijera Huntsberger: "La palabra estadística a menudo nos trae a la mente imágenes de números apilados en grandes arreglos y tablas, de volúmenes de cifras relativas a nacimientos, muertes, impuestos, poblaciones, ingresos, deudas, créditos y así sucesivamente. Huntsberger tiene razón pues al instante de escuchar esta palabra estas son las imágenes que llegan a nuestra cabeza.

La Estadística es mucho más que sólo números apilados y gráficas bonitas. Es una ciencia con tanta antigüedad como la escritura, y es por sí misma auxiliar de todas las demás ciencias. Los mercados, la medicina, la ingeniería, los gobiernos, etc. Se nombran entre los más destacados clientes de ésta.

La ausencia de ésta conllevaría a un caos generalizado, dejando a los administradores y ejecutivos sin información vital a la hora de tomar decisiones en tiempos de incertidumbre.

La Estadística que conocemos hoy en día debe gran parte de su realización a los trabajos matemáticos de aquellos hombres que desarrollaron la teoría de las probabilidades, con la cual se adhirió a la Estadística a las ciencias formales.

En este breve material se expone los conceptos, la historia, la división así como algunos errores básicos cometidos al momento de analizar datos Estadísticos.

Definición de Estadística

La Estadística es la ciencia cuyo objetivo es reunir una información cuantitativa concerniente a individuos, grupos, series de hechos, etc. y deducir de ello gracias al análisis de estos datos unos significados precisos o unas previsiones para el futuro.

La estadística, en general, es la ciencia que trata de la recopilación, organización presentación, análisis e interpretación de datos numéricos con e fin de realizar una toma de decisión más efectiva.

Otros autores tienen definiciones de la Estadística semejantes a las anteriores, y algunos otros no tan semejantes. Para Chacón esta se define como “la ciencia que tiene por objeto el estudio cuantitativo de los colectivos”; otros la definen como la expresión cuantitativa del conocimiento dispuesta en forma adecuada para el escrutinio y análisis. La más aceptada, sin embargo, es la de Minguez, que define la Estadística como “La ciencia que tiene por objeto aplicar las leyes de la cantidad a los hechos sociales para medir su intensidad, deducir las leyes que los rigen y hacer su predicción próxima”.

Los estudiantes confunden comúnmente los demás términos asociados con las Estadísticas, una confusión que es conveniente aclarar debido a que esta palabra tiene tres significados: la palabra estadística, en primer término se usa para referirse a la información estadística; también se utiliza para referirse al conjunto de técnicas y métodos que se utilizan para analizar la información estadística; y el término estadístico, en singular y en masculino, se refiere a una medida derivada de una muestra.

Historia de la Estadística

Los comienzos de la estadística pueden ser hallados en el antiguo Egipto, cuyos faraones lograron recopilar, hacia el año 3050 antes de Cristo, prolijos datos relativos a la población y la riqueza del país. De acuerdo al historiador griego Heródoto, dicho registro de riqueza y población se hizo con el objetivo de preparar la construcción de las pirámides. En el mismo Egipto, Ramsés II hizo un censo de las tierras con el objeto de verificar un nuevo reparto.

En el antiguo Israel la Biblia da referencias, en el libro de los Números, de los datos estadísticos obtenidos en dos recuentos de la población hebrea. El rey David por otra parte, ordenó a Joab, general del ejército hacer un censo de Israel con la finalidad de conocer el número de la población.

También los chinos efectuaron censos hace más de cuarenta siglos. Los griegos efectuaron censos periódicamente con fines tributarios, sociales (división de tierras) y militares (cálculo de recursos y hombres disponibles). La investigación histórica revela que se realizaron 69 censos para calcular los impuestos, determinar los derechos de voto y ponderar la potencia guerrera.

Pero fueron los romanos, maestros de la organización política, quienes mejor supieron emplear los recursos de la estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos tenían la obligación de anotar nacimientos, defunciones y matrimonios, sin olvidar los recuentos periódicos del ganado y de las riquezas contenidas en las tierras conquistadas. Para el nacimiento de Cristo sucedía uno de estos empadronamientos de la población bajo la autoridad del imperio.

Durante los mil años siguientes a la caída del imperio Romano se realizaron muy pocas operaciones Estadísticas, con la notable excepción de las relaciones de tierras pertenecientes a la

Iglesia, compiladas por Pipino el Breve en el 758 y por Carlomagno en el 762 DC. Durante el siglo IX se realizaron en Francia algunos censos parciales de siervos. En Inglaterra, Guillermo el Conquistador recopiló el Domesday Book o libro del Gran Catastro para el año 1086, un documento de la propiedad, extensión y valor de las tierras de Inglaterra. Esa obra fue el primer compendio estadístico de Inglaterra.

Aunque Carlomagno, en Francia; y Guillermo el Conquistador, en Inglaterra, trataron de revivir la técnica romana, los métodos estadísticos permanecieron casi olvidados durante la Edad Media.

Durante los siglos XV, XVI, y XVII, hombres como Leonardo de Vinci, Nicolás Copérnico, Galileo, Neper, William Harvey, Sir Francis Bacon y René Descartes, hicieron grandes operaciones al método científico, de tal forma que cuando se crearon los Estados Nacionales y surgió como fuerza el comercio internacional existía ya un método capaz de aplicarse a los datos económicos.

Durante el siglo XVII y principios del XVIII, matemáticos como Bernoulli, Francis Maseres, Lagrange y Laplace desarrollaron la teoría de probabilidades. No obstante durante cierto tiempo, la teoría de las probabilidades limitó su aplicación a los juegos de azar y hasta el siglo XVIII no comenzó a aplicarse a los grandes problemas científicos.

Godofredo Achenwall, profesor de la Universidad de Gotinga, acuñó en 1760 la palabra estadística, que extrajo del término italiano statista (estadista). Creía, y con sobrada razón, que los datos de la nueva ciencia serían el aliado más eficaz del gobernante consciente. La raíz remota de la palabra se halla, por otra parte, en el término latino status, que significa estado o situación; Esta etimología aumenta el valor intrínseco de la palabra, por cuanto la estadística revela el sentido cuantitativo de las más variadas situaciones.

Jacques Quételect es quien aplica las Estadísticas a las ciencias sociales. Este interpretó la teoría de la probabilidad para su uso en las ciencias sociales y resolver la aplicación del principio de promedios y de la variabilidad a los fenómenos sociales. Quételect fue el primero en realizar la aplicación práctica de todo el método Estadístico, entonces conocido, a las diversas ramas de la ciencia.

Entretanto, en el período del 1800 al 1820 se desarrollaron dos conceptos matemáticos fundamentales para la teoría Estadística; la teoría de los errores de observación, aportada por Laplace y Gauss; y la teoría de los mínimos cuadrados desarrollada por Laplace, Gauss y Legendre. A finales del siglo XIX, Sir Francis Gaston ideó el método conocido por Correlación, que tenía por objeto medir la influencia relativa de los factores sobre las variables. De aquí partió el desarrollo del coeficiente de correlación creado por Karl Pearson y otros cultivadores de la ciencia biométrica como J. Pease Norton, R. H. Hooker y G. Udny Yule, que efectuaron amplios estudios sobre la medida de las relaciones.

Los progresos más recientes en el campo de la Estadística se refieren al ulterior desarrollo del cálculo de probabilidades, particularmente en la rama denominada indeterminismo o relatividad, se ha demostrado que el determinismo fue reconocido en la Física como resultado de las investigaciones atómicas y que este principio se juzga aplicable tanto a las ciencias sociales como a las físicas.

División de la Estadística

La Estadística para su mejor estudio se ha dividido en dos grandes ramas: la Estadística Descriptiva y la Inferencial.

Estadística Descriptiva: consiste sobre todo en la presentación de datos en forma de tablas y gráficas. Esta comprende cualquier actividad relacionada con los datos y está diseñada para resumir o describir los mismos sin factores pertinentes adicionales; esto es, sin intentar inferir nada que vaya más allá de los datos, como tales.

Estadística Inferencial: se deriva de muestras, de observaciones hechas sólo acerca de una parte de un conjunto numeroso de elementos y esto implica que su análisis requiere de generalizaciones que van más allá de los datos. Como consecuencia, la característica más importante del reciente crecimiento de la estadística ha sido un cambio en el énfasis de los métodos que describen a métodos que sirven para hacer generalizaciones. La Estadística Inferencial investiga o analiza una población partiendo de una muestra tomada.

Método Estadístico

El conjunto de los métodos que se utilizan para medir las características de la información, para resumir los valores individuales, y para analizar los datos a fin de extraerles el máximo de información, es lo que se llama métodos estadísticos. Los métodos de análisis para la información cuantitativa se pueden dividir en los siguientes seis pasos:

1. Definición del problema.
2. Recopilación de la información existente.
3. Obtención de información original.
4. Clasificación.
5. Presentación.
6. Análisis.

ESTADÍSTICA DESCRIPTIVA

Definición de Estadística

La **Estadística** trata del recuento, ordenación y clasificación de los datos obtenidos por las observaciones, para poder hacer comparaciones y sacar conclusiones.
Un **estudio estadístico** consta de las siguientes fases:
Recogida de datos.
Organización y representación de datos.
Análisis de datos.
Obtención de conclusiones.

Conceptos de Estadística

Población
Una **población** es el conjunto de todos los elementos a los que se somete a un estudio estadístico.
Individuo
Un **individuo** o **unidad estadística** es cada uno de los elementos que componen la población.
Muestra
Una **muestra** es un conjunto representativo de la población de referencia, el número de individuos de una muestra es menor que el de la población.
Muestreo
El **muestreo** es la reunión de datos que se desea estudiar, obtenidos de una proporción reducida y representativa de la población.
Valor
Un **valor** es cada uno de los distintos resultados que se pueden obtener en un estudio estadístico. Si lanzamos una moneda al aire 5 veces obtenemos dos valores: cara y cruz.
Dato
Un **dato** es cada uno de los valores que se ha obtenido al realizar un estudio estadístico. Si lanzamos una moneda al aire 5 veces obtenemos 5 datos: cara, cara, cruz, cara, cruz.

Definición de variable

Una **variable estadística** es cada una de las **características o cualidades** que poseen los i**ndividuos de una población.**

Tipos de variable estadísticas

Variable cualitativa

Las **variables cualitativas** se refieren a **características o cualidades** que **no** pueden ser medidas con **números.** Podemos distinguir dos tipos:

Variable cualitativa nominal

Una **variable cualitativa nominal** presenta **modalidades no numéricas** que **no** admiten un **criterio de orden**.
Por ejemplo:
El estado civil, con las siguientes modalidades: soltero, casado, separado, divorciado y viudo.
Variable cualitativa ordinal o variable cuasicuantitativa

Una **variable cualitativa ordinal** presenta **modalidades no númericas**, en las que existe un **orden.** Por ejemplo:
La nota en un examen: suspenso, aprobado, notable, sobresaliente.
Puesto conseguido en una prueba deportiva: 1º, 2º, 3º, ...
Medallas de una prueba deportiva: oro, plata, bronce.

Variable cuantitativa

Una **variable cuantitativa** es la que se expresa mediante un **número**, por tanto se pueden realizar o**peraciones aritméticas** con ella. Podemos distinguir dos tipos:

Variable discreta

Una **variable discreta** es aquella que toma **valores aislados**, es decir **no** admite **valores intermedios** entre dos valores específicos. Por ejemplo:
El número de hermanos de 5 amigos: 2, 1, 0, 1, 3.

Variable continua

Una **variable continua** es aquella que puede tomar **valores comprendidos entre dos números**. Por ejemplo:
La altura de los 5 amigos: 1.73, 1.82, 1.77, 1.69, 1.75.
En la práctica medimos la altura con dos decimales, pero también se podría dar con tres decimales.

Distribución de frecuencias

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**.

Tipos de frecuencias

Frecuencia absoluta

La **frecuencia absoluta** es el **número de veces** que aparece un determinado **valor** en un estudio estadístico.

Se representa por f_i.
La **suma de las frecuencias absolutas** es igual al número total de datos, que se representa por **N**.

$$f_1 + f_2 + f_3 + \ldots + f_n = N$$

Para indicar resumidamente estas sumas se utiliza la letra griega **Σ** (sigma mayúscula) que se lee suma o sumatoria.

$$\sum_{i=1}^{i=n} f_i = N$$

Frecuencia relativa

La **frecuencia relativa** es el **cociente** entre la f**recuencia absoluta** de un determinado valor y el **número total de datos**.
Se puede expresar en tantos por ciento y se representa por n_i.

$$n_i = \frac{f_i}{N}$$

La suma de las frecuencias relativas es igual a 1.

Frecuencia acumulada

La **frecuencia acumulada** es la **suma de las frecuencias absolutas** de todos los **valores inferiores o iguales** al **valor** considerado.
Se representa por F_i.

Frecuencia relativa acumulada

La **frecuencia relativa acumulada** es el **cociente** entre la **frecuencia acumulada** de un determinado **valor** y el **número total de datos**. Se puede expresar en tantos por ciento.

Ejemplo

Durante el mes de julio, en una ciudad se han registrado las siguientes temperaturas máximas:
32, 31, 28, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 30, 32, 31, 31, 30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33, 29, 29.
En la primera columna de la tabla colocamos la variable ordenada de menor a mayor, en la segunda hacemos el recuento y en la tercera anotamos la frecuencia absoluta.

x_i	Recuento	f_i	F_i	n_i	N_i
27	I	1	1	0.032	0.032
28	II	2	3	0.065	0.097
29	~~IIII~~ I	6	9	0.194	0.290
30	~~IIII~~ II	7	16	0.226	0.516
31	~~IIII~~ III	8	24	0.258	0.774
32	III	3	27	0.097	0.871
33	III	3	30	0.097	0.968
34	I	1	31	0.032	1
		31		1	

Este tipo de **tablas de frecuencias** se utiliza con **variables discretas**.

Distribución de frecuencias agrupadas

La **distribución de frecuencias agrupadas** o **tabla con datos agrupados** se emplea si las **variables** toman un **número grande de valores** o la **variable es continua**.

Se **agrupan** los **valores** en **intervalos** que tengan la **misma amplitud** denominados **clases**. A cada **clase** se le asigna su **frecuencia correspondiente**.

Límites de la clase

Cada **clase** está **delimitada** por el **límite inferior de la clase** y el **límite superior de la clase**.

Amplitud de la clase

La **amplitud de la clase** es la **diferencia** entre el **límite superior e inferior** de la **clase**.

Marca de clase

La **marca de clase** es el **punto medio** de cada **intervalo** y es el **valor** que representa a todo el **intervalo** para el **cálculo** de algunos **parámetros**.

Construcción de una tabla de datos agrupados

3, 15, 24, 28, 33, 35, 38, 42, 43, 38, 36, 34, 29, 25, 17, 7, 34, 36, 39, 44, 31, 26, 20, 11, 13, 22, 27, 47, 39, 37, 34, 32, 35, 28, 38, 41, 48, 15, 32, 13.

1º Se localizan los valores menor y mayor de la distribución. En este caso son 3 y 48.

2º Se restan y se busca un número entero un poco mayor que la diferencia y que sea divisible por el número de intervalos queramos establecer.

Es conveniente que el número de intervalos oscile entre 6 y 15.

En este caso, 48 - 3 = 45, incrementamos el número hasta 50 : 5 = 10 intervalos.

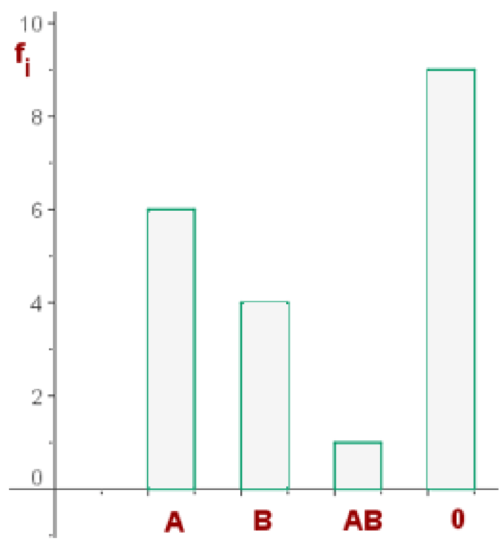
Se forman los intervalos teniendo presente que el límite inferior de una clase pertenece al intervalo, pero el límite superior no pertenece intervalo, se cuenta en el siguiente intervalo.

	c_i	f_i	F_i	n_i	N_i
[0, 5)	2.5	1	1	0.025	0.025
[5, 10)	7.5	1	2	0.025	0.050
[10, 15)	12.5	3	5	0.075	0.125
[15, 20)	17.5	3	8	0.075	0.200
[20, 25)	22.5	3	11	0.075	0.2775
[25, 30)	27.5	6	17	0.150	0.425
[30, 35)	32.5	7	24	0.175	0.600
[35, 40)	37.5	10	34	0.250	0.850
[40, 45)	42.5	4	38	0.100	0.950
[45, 50)	47.5	2	40	0.050	1
		40		1	

Diagrama de barras y polígonos de frecuencias

Un **diagrama de barras** se utiliza para de presentar **datos cualitativos** o **datos cuantitativos de tipo discreto.**
Se representan sobre unos ejes de coordenadas, en el **eje de abscisas** se colocan los **valores de la variable**, y sobre el **eje de ordenadas** las **frecuencias absolutas o relativas o acumuladas.**
Los **datos** se representan mediante **barras** de una **altura proporcional** a la **frecuencia.**
Ejemplo
Un estudio hecho al conjunto de los 20 alumnos de una clase para determinar su grupo sanguíneo ha dado el siguiente resultado:

Grupo sanguíneo	f_i
A	6
B	4
AB	1
0	9
	20

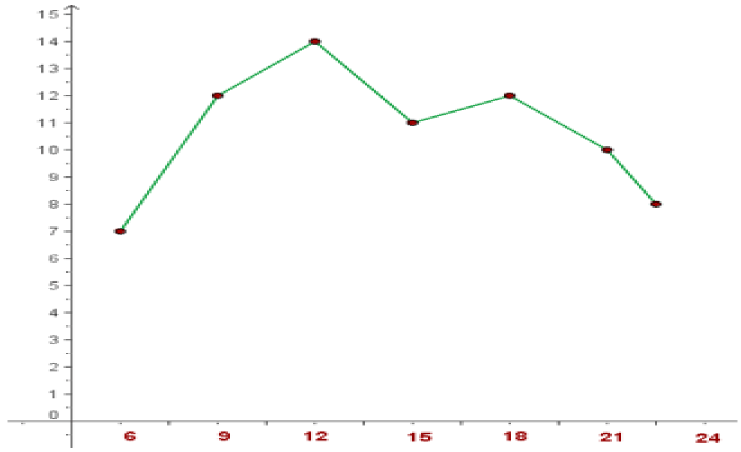


Polígonos de frecuencia

Un **polígono de frecuencias** se forma uniendo los **extremos** de las **barras** mediante **segmentos.**
También se puede realizar trazando los **puntos** que representan las **frecuencias** y uniéndolos mediante **segmentos.**

Ejemplo
Las temperaturas en un día de otoño de una ciudad han sufrido las siguientes variaciones:

Hora	Temperatura
6	7º
9	12°
12	14°
15	11°
18	12°
21	10°
24	8°



Un **diagrama de sectores** se puede utilizar para todo tipo de *variables*, pero se usa frecuentemente para las **variables cualitativas**.
Los **datos** se representan en un **círculo**, de modo que el **ángulo** de cada **sector** es **proporcional** a la **frecuencia absoluta** correspondiente.

$$\alpha = \frac{360°}{N} \cdot f_i$$

El diagrama circular se construye con la ayuda de un transportador de ángulos.
Ejemplo
En una clase de 30 alumnos, 12 juegan a baloncesto, 3 practican la natación, 4 juegan al fútbol y el resto no practica ningún deporte.

$$\alpha_1 = \frac{360°}{30} \cdot 12 = 144° \qquad \alpha_2 = \frac{360°}{30} \cdot 3 = 36°$$

$$\alpha_3 = \frac{360°}{30} \cdot 9 = 108° \qquad \alpha_4 = \frac{360°}{30} \cdot 6 = 72°$$

	Alumnos	Ángulo
Baloncesto	12	144°
Natación	3	36°
Fútbol	9	108°
Sin deporte	6	72°
Total	30	360°

Un **histograma** es una **representación gráfica** de una **variable** en forma de **barras**.
Se utilizan para **variables continuas** o para **variables discretas**, con un gran número de datos, y que se han agrupado en **clases**.
En el **eje abscisas** se construyen unos **rectángulos** que tienen por **base la amplitud del intervalo**, y por **altura**, la **frecuencia absoluta** de cada **intervalo**.
La **superficie** de cada **barra** es **proporcional** a la **frecuencia** de los **valores** representados.

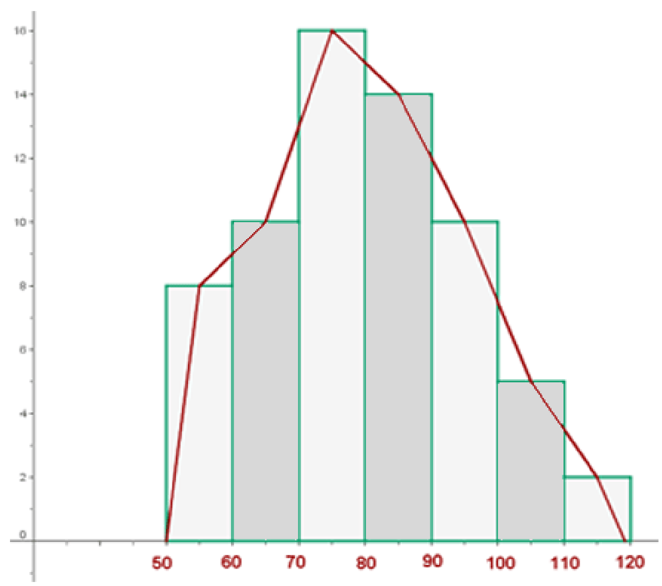
Polígono de frecuencia

Para construir el **polígono de frecuencia** se toma la **marca de clase** que coincide con el **punto medio** de cada **rectángulo**.

Ejemplo

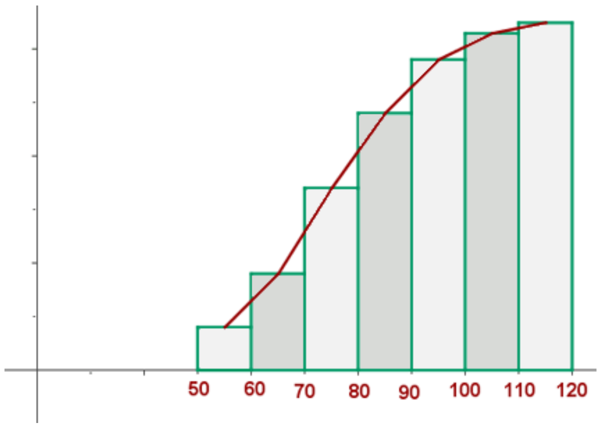
El peso de 65 personas adultas viene dado por la siguiente tabla:

	c_i	f_i	F_i
[50, 60)	55	8	8
[60, 70)	65	10	18
[70, 80)	75	16	34
[80, 90)	85	14	48
[90, 100)	95	10	58
[100, 110)	110	5	63
[110, 120)	115	2	65
		65	



Histograma y polígono de frecuencias acumuladas

Si se representan las **frecuencias acumuladas** de una **tabla de datos agrupados** se obtiene el **histograma de frecuencias acumuladas** o su correspondiente **polígono**.



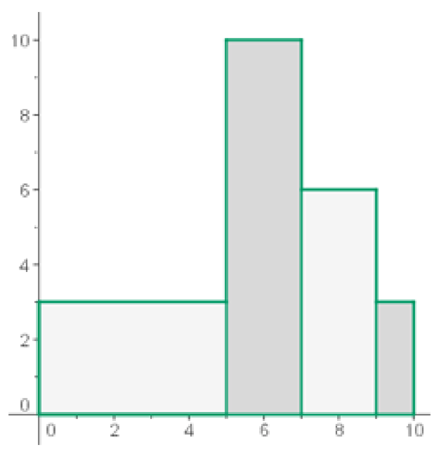
Histogramas con intervalos de amplitud diferente

Para **construir** un **histogramas** con **intervalo de amplitud diferente** tenemos que **calcular** las **alturas** de los **rectángulos** del **histograma**.

$$h_i = \frac{f_i}{a_i}$$

h_i es la altura del intervalo.
f_i es la frecuencia del intervalo.
a_i es la amplitud del intervalo.
Ejemplo
En la siguiente tabla se muestra las calificaciones (suspenso, aprobado, notable y sobresaliente) obtenidas por un grupo de 50 alumnos.

	f_i	h_i
[0, 5)	15	3
[5, 7)	20	10
[7, 9)	12	6
[9, 10)	3	3
	50	



Definición de parámetro estadístico

Un **parámetro estadístico** es un **número** que se obtiene a partir de los **datos** de una **distribución estadística**.
Los **parámetros estadísticos** sirven para sintetizar la información dada por una tabla o por una gráfica.

Tipos de parámetros estadísticos
Hay **tres tipos parámetros estadísticos**:

Medidas de centralización
Nos indican en torno a qué valor (centro) se distribuyen los datos.
La **medidas de centralización** son:
Media aritmética
La **media** es el valor **promedio** de la distribución.
Mediana
La **mediana** es la **puntación** de la escala que **separa la mitad superior** de la distribución y **la inferior**, es decir divide la serie de datos en **dos partes iguales**.
Moda
La **moda** es el **valor** que **más se repite** en una distribución.
Medidas de posición

Las **medidas de posición** dividen un conjunto de datos en grupos con el mismo número de individuos.
Para calcular las **medidas de posición** es necesario que los **datos** estén ordenados de **menor a mayor**.
La **medidas de posición** son:
Cuartiles
Los **cuartiles dividen** la serie de datos en **cuatro partes iguales**.
Deciles
Los **deciles** dividen la serie de datos en **diez partes iguales**.
Percentiles
Los **percentiles** dividen la serie de datos en **cien partes iguales**.

Medidas de dispersión
Las **medidas de dispersión** nos informan sobre cuanto se alejan del centro los valores de la distribución.
Las **medidas de dispersión** son:
Rango o recorrido
El **rango** es la **diferencia** entre el **mayor** y el **menor** de los **datos** de una distribución estadística.
Desviación media
La **desviación media** es la **media aritmética** de los **valores absolutos** de las **desviaciones** respecto a la **media**.
Varianza
La **varianza** es la **media aritmética** del **cuadrado de las desviaciones** respecto a la **media**.
Desviación típica
La **desviación típica** es la **raíz cuadrada** de la **varianza**.

Unidad II Medidas de tendencia central

Media aritmética.

La **media aritmética** es el **valor** obtenido al **sumar** todos los **datos** y **dividir** el resultado entre el **número** total de **datos**.

$\bar{x}$ es el símbolo de la **media aritmética**.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{N}$$

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

Ejemplo:
Los pesos de seis amigos son: 84, 91, 72, 68, 87 y 78 kg. Hallar el peso medio.

$$\bar{x} = \frac{84 + 91 + 72 + 68 + 87 + 78}{6} = 80\ Kg$$

Media aritmética para datos agrupados
Si los **datos** vienen **agrupados** en una tabla de frecuencias, la expresión de la **media** es:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \dots + x_n f_n}{N}$$

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

Ejercicio de media aritmética
En un test realizado a un grupo de 42 personas se han obtenido las puntuaciones que muestra la tabla. **Calcula la puntuación media**.

	x_i	f_i	$x_i \cdot f_i$
[10, 20)	15	1	15
[20, 30)	25	8	200
[30,40)	35	10	350
[40, 50)	45	9	405
[50, 60	55	8	440
[60,70)	65	4	260
[70, 80)	75	2	150
		42	1 820

$$\bar{x} = \frac{1820}{42} = 43.33$$

Mediana.

Es el **valor** que ocupa el **lugar central** de todos los **datos** cuando éstos están **ordenados de menor a mayor**.
La **mediana** se representa por **M_e**.
La **mediana** se puede **hallar** sólo para **variables cuantitativas**.
Cálculo de la mediana

1 Ordenamos los **datos** de **menor a mayor**.
2 Si la serie tiene un **número impar de medidas** la **mediana** es la **puntuación central** de la misma.
2, 3, 4, 4, 5, 5, 5, 6, 6 **Me= 5**
3 Si la serie tiene un **número par** de puntuaciones la **mediana** es la **media** entre las dos **puntuaciones centrales**.
7, 8, 9, 10, 11, 12 **Me= 9.5**
Cálculo de la mediana para datos agrupados

La **mediana** se encuentra en el **intervalo** donde la **frecuencia acumulada** llega hasta la **mitad de la suma de las frecuencias absolutas**.

Es decir tenemos que buscar el intervalo en el que se encuentre. $\frac{N}{2}$

$$Me = L_i + \frac{\frac{N}{2} - F_{i-1}}{f_i} \cdot a_i$$

L_i es el límite inferior de la clase donde se encuentra la mediana.
$\frac{N}{2}$ es la semisuma de las frecuencias absolutas.
F_{i-1} es la **frecuencia acumulada** anterior a la clase mediana.
a_i es la amplitud de la clase.
La **mediana** es **independiente** de las **amplitudes** de los **intervalos**.

Ejemplo:
Calcular la **mediana** de una distribución estadística que viene dada por la siguiente tabla:

	f_i	**F_i**
[60, 63)	**5**	**5**
[63, 66)	**18**	**23**
[66, 69)	**42**	**65**
[69, 72)	**27**	**92**
[72, 75)	**8**	**100**
	100	

100/2 = 50
Clase de la mediana: [66, 69)

$$Me = 66 + \frac{50 - 23}{42} \cdot 3 = 67.93$$

Moda.

La **moda** es el **valor** que tiene **mayor frecuencia absoluta**.
Se representa por M_o.
Se puede hallar la **moda** para **variables cualitativas** y **cuantitativas**.
Hallar la **moda** de la distribución:
2, 3, 3, 4, 4, 4, 5, 5 **M_o= 4**
Si en un grupo hay **dos o varias puntuaciones** con la **misma frecuencia** y esa frecuencia es la máxima, la **distribución** es **bimodal** o **multimodal**, es decir, tiene **varias modas**.
1, 1, 1, 4, 4, 5, 5, 5, 7, 8, 9, 9, 9 **M_o= 1, 5, 9**
Cuando todas las **puntuaciones** de un grupo tienen la **misma frecuencia**, **no** hay **moda**.
2, 2, 3, 3, 6, 6, 9, 9
Si **dos puntuaciones adyacentes** tienen la **frecuencia máxima**, la **moda** es el **promedio** de las dos puntuaciones adyacentes.
0, 1, 3, 3, 5, 5, 7, 8 **Mo = 4**

Cálculo de la moda para datos agrupados
1º Todos los intervalos tienen la misma amplitud.

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i$$

L_i es el límite inferior de la clase modal.
f_i es la frecuencia absoluta de la clase modal.
f_{i-1} es la frecuencia absoluta inmediatamente inferior a la clase modal.
f_{i+1} es la frecuencia absoluta inmediatamente posterior a la clase modal.
a_i es la amplitud de la clase.
También se utiliza otra **fórmula** de la **moda** que da un **valor aproximado** de ésta:

$$Mo = L_i + \frac{f_{i+1}}{f_{i-1} + f_{i+1}} \cdot a_i$$

Ejemplo
Calcular la **moda** de una distribución estadística que viene dada por la siguiente tabla:

	f_i
[60, 63)	**5**
[63, 66)	**18**
[66, 69)	**42**
[69, 72)	**27**
[72, 75)	**8**
	100

$$Mo = 66 + \frac{(42-18)}{(42-18)+(42-27)} \cdot 3 = 67.846$$

$$Mo = 66 + \frac{27}{18+27} \cdot 3 = 67.8$$

2º Los intervalos tienen amplitudes distintas.
En primer lugar tenemos que hallar las alturas.

$$h_i = \frac{f_i}{a_i}$$

La clase modal es la que tiene mayor altura.

$$Mo = L_i + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} \cdot a_i$$

La **fórmula** de la **moda aproximada** cuando existen distintas amplitudes es:

$$Mo = L_i + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot a_i$$

Ejemplo
En la siguiente tabla se muestra las calificaciones (suspenso, aprobado, notable y sobresaliente) obtenidas por un grupo de 50 alumnos. **Calcular la moda.**

	f_i	h_i
[0, 5)	15	3
[5, 7)	20	10
[7, 9)	12	6
[9, 10)	3	3
	50	

$$Mo = 5 + \frac{10 - 3}{(10 - 3) + (10 - 6)} \cdot 2 = 6.27$$

$$Mo = 5 + \frac{6}{3 + 6} \cdot 2 = 6.33$$

La **media geométrica** de una cantidad arbitraria de números (digamos *n* números) es la raíz n-ésima del producto de todos los números.

$$\bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Por ejemplo, la media geométrica de 2 y 18 es

$$\sqrt[2]{2 \cdot 18} = \sqrt[2]{36} = 6$$

Otro ejemplo, la media de 1, 3 y 9 sería

$$\sqrt[3]{1 \cdot 3 \cdot 9} = \sqrt[3]{27} = 3$$

La **media armónica** , denominada **H**, de una cantidad finita de números es igual al recíproco, o inverso, de la media aritmética de los recíprocos de dichos valores

Así, dados los números $a_1, a_2, \ldots, a_n$, la media armónica será igual a:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}} = \frac{n}{\left(\frac{1}{a_1} + \cdots + \frac{1}{a_n}\right)}$$

La media armónica resulta poco influida por la existencia de determinados valores mucho más grandes que el conjunto de los otros, siendo en cambio sensible a valores mucho más pequeños que el conjunto.

La media armónica no está definida en el caso de la existencia en el conjunto de valores nulos.

Medidas de posición

Los **cuartiles** son los **tres valores** de la variable que **dividen** a un **conjunto** de **datos ordenados** en **cuatro partes iguales**.
Q_1, Q_2 y Q_3 determinan los valores correspondientes al **25%, al 50% y al 75%** de los **datos**.
Q_2 coincide con la **mediana**.

Cálculo de los cuartiles
1 Ordenamos los **datos** de **menor a mayor**.

2 Buscamos el lugar que ocupa cada **cuartil** mediante la expresión $\frac{k \cdot N}{4},\ k = 1, 2, 3$.
Número impar de datos
2, 5, 3, 6, 7, 4, 9

2, 3, 4, 5, 6, 7, 9
↓ ↓ ↓
Q_1 Q_2 Q_3

Número par de datos
2, 5, 3, 4, 6, 7, 1, 9

1, 2, 3, 4, 5, 6, 7, 9
2.5 4.5 6.5
↓ ↓ ↓
Q_1 Q_2 Q_3

Cálculo de los cuartiles para datos agrupados

En primer lugar buscamos la **clase** donde se encuentra $\frac{k \cdot N}{4},\ k = 1, 2, 3$, en la **tabla de las frecuencias acumuladas**.

$$Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i \qquad k = 1, 2, 3$$

L_i es el límite inferior de la clase donde se encuentra el cuartil.
N es la suma de las frecuencias absolutas.
F_{i-1} es la **frecuencia acumulada** anterior a la clase del cuartil.
a_i es la amplitud de la clase.

Ejercicio de cuartiles

Calcular los cuartiles de la distribución de la tabla:

	f_i	F_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	65	

Cálculo del primer cuartil

$$\frac{65 \cdot 1}{4} = 16.25$$

$$Q_1 = 60 + \frac{16.25 - 8}{10} \cdot 10 = 68.25$$

Cálculo del segundo cuartil

$$\frac{65 \cdot 2}{4} = 32.5$$

$$Q_2 = 70 + \frac{32.5 - 18}{16} \cdot 10 = 79.0625$$

Cálculo del tercer cuartil

$$\frac{65 \cdot 3}{4} = 48.75$$

$$Q_3 = 90 + \frac{48.75 - 48}{10} \cdot 10 = 90.75$$

Los **deciles** son los **nueve valores** que **dividen** la serie de **datos** en **diez partes iguales**.

Los **deciles** dan los valores correspondientes al 10%, al 20%... y al 90% de los datos.

D_5 coincide con la **mediana**.

Cálculo de los deciles

En primer lugar buscamos la clase donde se encuentra $\frac{k \cdot N}{10}, \quad k = 1, 2, ..9$, en la tabla de las frecuencias acumuladas.

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad\qquad k = 1, 2, ..9$$

L_i es el límite inferior de la clase donde se encuentra el decil.

N es la suma de las frecuencias absolutas.

F_{i-1} es la **frecuencia acumulada** anterior a la clase el decil..

a_i es la amplitud de la clase.

Ejercicio de deciles
Calcular los deciles de la distribución de la tabla:

	f_i	F_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	65	

Cálculo del primer decil

$$\frac{65 \cdot 1}{10} = 6.5$$

$$D_1 = 50 + \frac{6.5 - 0}{8} \cdot 10 = 58.12$$

Cálculo del segundo decil

$$\frac{65 \cdot 2}{10} = 13$$

$$D_2 = 60 + \frac{13 - 8}{10} \cdot 10 = 65$$

Cálculo del tercer decil

$$\frac{65 \cdot 3}{10} = 19.5$$

$$D_3 = 70 + \frac{19.5 - 18}{16} \cdot 10 = 70.94$$

Cálculo del cuarto decil

$$\frac{65 \cdot 4}{10} = 26$$

$$D_4 = 70 + \frac{26 - 18}{16} \cdot 10 = 75$$

Cálculo del quinto decil

$$\frac{65 \cdot 5}{10} = 32.5$$

$$D_5 = 70 + \frac{32.5 - 18}{16} \cdot 10 = 79.06$$

Cálculo del sexto decil

$$\frac{65 \cdot 6}{10} = 39$$

$$D_6 = 80 + \frac{39 - 34}{14} \cdot 10 = 83.57$$

Cálculo del séptimo decil

$$\frac{65 \cdot 7}{10} = 45.5$$

$$D_7 = 80 + \frac{45.5 - 34}{14} \cdot 10 = 88.21$$

Cálculo del octavo decil

$$\frac{65 \cdot 8}{10} = 52$$

$$D_8 = 90 + \frac{52 - 48}{10} \cdot 10 = 94$$

Cálculo del noveno decil

$$\frac{65 \cdot 9}{10} = 58.5$$

$$D_9 = 100 + \frac{58.5 - 58}{5} \cdot 10 = 101$$

Los **percentiles** son los **99 valores** que **dividen** la serie de **datos** en **100 partes iguales**.
Los **percentiles** dan los valores correspondientes al 1%, al 2%... y al 99% de los datos.
P_{50} coincide con la **mediana**.
Cálculo de los percentiles

En primer lugar buscamos la clase donde se encuentra $\frac{k \cdot N}{100}, \quad k = 1, 2, ..99$, en la tabla de las frecuencias acumuladas.

$$P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i \qquad k = 1, 2, ..99$$

L_i es el límite inferior de la clase donde se encuentra el percentil.
N es la suma de las frecuencias absolutas.
F_{i-1} es la **frecuencia acumulada** anterior a la clase del percentil.
a_i es la amplitud de la clase.

Ejercicio de percentiles
Calcular el percentil 35 y 60 de la distribución de la tabla:

	f_i	F_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	65	

Percentil 35

$$\frac{65 \cdot 35}{100} = 22.75$$

$$P_{35} = 70 + \frac{22.75 - 18}{16} \cdot 10 = 72.97$$

Percentil 60

$$\frac{65 \cdot 60}{100} = 39$$

$$P_{60} = 80 + \frac{39 - 34}{14} \cdot 10 = 83.57$$

Unidad III Medidas de dispersión

Las **medidas de dispersión** nos informan sobre cuánto se alejan del centro los valores de la distribución.

Las **medidas de dispersión** son:

Rango o recorrido

El **rango** es la **diferencia** entre el **mayor** y el **menor** de los **datos** de una distribución estadística.

Desviación media

La **desviación respecto a la media** es la **diferencia** entre cada **valor** de la variable estadística y la **media aritmética.**

D_i = x - x

La **desviación media** es la **media aritmética** de los **valores absolutos de las desviaciones respecto a la media**.

La **desviación media** se representa por $D_{\bar{x}}$

$$D_{\bar{x}} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|}{N}$$

$$D_{\bar{x}} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{N}$$

Ejemplo

Calcular la **desviación media** de la distribución:

9, 3, 8, 8, 9, 8, 9, 18

$$\bar{x} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$D_{\bar{x}} = \frac{|9-9| + |3-9| + |8-9| + |8-9| + |9-9| + |8-9| + |9-9| + |18-9|}{8} = 2.25$$

Desviación media para datos agrupados

Si los datos vienen agrupados en una **tabla de frecuencias**, la expresión de la **desviación media** es:

$$D_{\bar{x}} = \frac{|x_1 - \bar{x}| f_1 + |x_2 - \bar{x}| f_2 + \ldots + |x_n - \bar{x}| f_n}{N}$$

$$D_{\bar{x}} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}| f_i}{N}$$

Ejemplo
Calcular la **desviación media** de la distribución:

	x_i	f_i	$x_i \cdot f_i$	|x - x|	|x - x| · f_i
[10, 15)	12.5	3	37.5	9.286	27.858
[15, 20)	17.5	5	87.5	4.286	21.43
[20, 25)	22.5	7	157.5	0.714	4.998
[25, 30)	27.5	4	110	5.714	22.856
[30, 35)	32.5	2	65	10.174	21.428
		21	457.5		98.57

$$\bar{x} = \frac{457.5}{21} = 21.786$$

$$D_{\bar{x}} = \frac{98.57}{21} = 4.69$$

Varianza

La **varianza** es la **media aritmética del cuadrado de las desviaciones respecto a la media** de una distribución estadística.

La varianza se representa por σ^2.

$$\sigma^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{N}$$

Varianza para datos agrupados

$$\sigma^2 = \frac{(x_1-\bar{x})^2 f_1 + (x_2-\bar{x})^2 f_2 + \ldots + (x_n-\bar{x})^2 f_n}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 f_i}{N}$$

Para simplificar el **cálculo de la varianza** vamos o utilizar las siguientes expresiones que son equivalentes a las anteriores.

$$\sigma^2 = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{N} - \bar{x}^2 \qquad \sigma^2 = \sum_{i=1}^{n}\frac{x_i^2}{N} - \bar{x}^2$$

Varianza para datos agrupados

$$\sigma^2 = \frac{x_1^2 f_1 + x_2^2 f_2 + \ldots + x_n^2 f_n}{N} - \bar{x}^2 \qquad \sigma^2 = \sum_{i=1}^{n}\frac{x_i^2 f_i}{N} - \bar{x}^2$$

Ejercicios de varianza
Calcular la varianza de la distribución:
9, 3, 8, 8, 9, 8, 9, 18

$$\bar{x} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$\sigma^2 = \frac{(9-9)^2 + (3-9)^2 + (8-9)^2 + (8-9)^2 + (9-9)^2 + (8-9)^2 + (9-9)^2 + (18-9)^2}{8} = 15$$

Calcular la varianza de la distribución de la tabla:

	x_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
[10, 20)	15	1	15	225
[20, 30)	25	8	200	5000
[30,40)	35	10	350	12 250
[40, 50)	45	9	405	18 225
[50, 60	55	8	440	24 200
[60,70)	65	4	260	16 900
[70, 80)	75	2	150	11 250
		42	1 820	88 050

$$\bar{x} = \frac{1820}{42} = 43.33$$

$$\sigma^2 = \frac{88050}{42} - 43.33^2 = 218.94$$

Propiedades de la varianza

1 La **varianza** será siempre un **valor positivo o cero**, en el caso de que las puntuaciones sean iguales.

2 Si a todos los **valores** de la variable se les **suma** un **número** la **varianza no varía**.

3 Si todos los **valores** de la variable se **multiplican** por un **número** la **varianza** queda **multiplicada** por el **cuadrado** de dicho **número**.

4 Si tenemos varias distribuciones con la misma **media** y conocemos sus respectivas **varianzas** se puede calcular la **varianza total**.

Si todas las muestras tienen el mismo tamaño:

$$\sigma^2 = \frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}{n}$$

Si las muestras tienen distinto tamaño:

$$\sigma^2 = \frac{k_1 \cdot \sigma_1^2 + k_2 \cdot \sigma_2^2 + \ldots + k_n \cdot \sigma_n^2}{k_1 + k_2 + \ldots + k_n}$$

Observaciones sobre la varianza

1 La **varianza**, al igual que la media, es un índice muy sensible a las puntuaciones extremas.

2 En los casos que no se pueda hallar la media tampoco será posible hallar la **varianza**.

3 La **varianza** no viene expresada en las mismas unidades que los datos, ya que las desviaciones están elevadas al cuadrado.

Desviación típica

La **desviación típica** es la **raíz cuadrada de la varianza**.
Es decir, la raíz cuadrada de la media de los cuadrados de las puntuaciones de desviación.
La **desviación típica** se representa por **σ**.

$$\sigma = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{N}} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{N}}$$

Desviación típica para datos agrupados

$$\sigma = \sqrt{\frac{(x_1-\bar{x})^2 f_1+(x_2-\bar{x})^2 f_2+\ldots+(x_n-\bar{x})^2 f_n}{N}} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 f_i}{N}}$$

Para simplificar el cálculo vamos o utilizar las siguientes expresiones que son equivalentes a las anteriores.

$$\sigma = \sqrt{\frac{x_1^2+x_2^2+\ldots+x_n^2}{N}-\bar{x}^2} \qquad \sigma = \sqrt{\sum_{i=1}^{n}\frac{x_i^2 f_i}{N}-\bar{x}^2}$$

Desviación típica para datos agrupados

$$\sigma = \sqrt{\frac{x_1^2 f+x_2^2 f_2+\ldots+x_n^2 f_n}{N}-\bar{x}^2} \qquad \sigma = \sqrt{\sum_{i=1}^{n}\frac{x_i^2}{N}-\bar{x}^2}$$

Ejercicios de desviación típica
Calcular la **desviación típica** de la distribución:
9, 3, 8, 8, 9, 8, 9, 18

$$\bar{x} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$\sigma^2 = \sqrt{\frac{(9-9)^2+(3-9)^2+(8-9)^2+(8-9)^2+(9-9)^2+(8-9)^2+(9-9)^2+(18-9)^2}{8}} = 3.87$$

Calcular la desviación típica de la distribución de la tabla:

	x_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
[10, 20)	15	1	15	225
[20, 30)	25	8	200	5000
[30,40)	35	10	350	12 250
[40, 50)	45	9	405	18 225
[50, 60)	55	8	440	24 200
[60,70)	65	4	260	16 900
[70, 80)	75	2	150	11 250
		42	1 820	88 050

$$\bar{x} = \frac{1820}{42} = 43.33$$

$$\sigma = \sqrt{\frac{88050}{42} - 43.33^2} = 14.797$$

Propiedades de la desviación típica

1 La **desviación típica** será siempre un **valor positivo o cero**, en el caso de que las puntuaciones sean iguales.

2 Si a todos los **valores** de la variable se les **suma** un **número** la **desviación típica no varía**.

3 Si todos los **valores** de la variable se **multiplican** por un **número** la **desviación típica** queda **multiplicada** por dicho **número**.

4 Si tenemos varias distribuciones con la misma **media** y conocemos sus respectivas **desviaciones típicas** se puede calcular la **desviación típica total**.

Si todas las muestras tienen el mismo tamaño:

$$\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}{n}}$$

Si las muestras tienen distinto tamaño:

$$\sigma = \sqrt{\frac{k_1 \cdot \sigma_1^2 + k_2 \cdot \sigma_2^2 + \ldots + k_n \cdot \sigma_n^2}{k_1 + k_2 + \ldots + k_n}}$$

Observaciones sobre la desviación típica

1 La **desviación típica**, al igual que la media y la varianza, es un índice muy sensible a las puntuaciones extremas.

2 En los casos que no se pueda hallar la media tampoco será posible hallar la **desviación típica**.

3 Cuanta más pequeña sea la **desviación típica** mayor será la **concentración de datos** alrededor de la **media**.

Unidad IV Medidas de forma

Medidas de forma: Son indicadores estadísticos que permiten identificar si una distribución de frecuencia presenta uniformidad

Sesgo

En estadística se llama **sesgo** de un estimador a la diferencia entre su esperanza matemática y el valor del parámetro que estima. Un estimador cuyo sesgo es nulo se llama *insesgado* o *centrado*.En notación matemática, dada una muestra $x_1, \ldots, x_n$ y un estimador $T(x_1, \ldots, x_n)$ del parámetro muestral θ, el sesgo es

$$E(T) - \theta.$$

El no tener sesgo es una propiedad deseable de los estimadores. Una propiedad relacionada con ésta es la de la consistencia: un estimador puede tener un sesgo pero el tamaño de éste converger a cero conforme crece el tamaño muestral.

Dada la importancia de la falta de sesgo, en ocasiones, en lugar de estimadores *naturales* se utilizan otros corregidos para eliminar el sesgo. Así ocurre, por ejemplo, con la varianza muestral.

Coeficiente de apuntamiento

La otra medida de forma que vamos a considerar es el apuntamiento, al igual que con la simetría hemos de tomar una referencia para ver si la distribución de los datos es apuntada o no.
Esa referencia será la distribución normal, distinguiremos tres casos que la distribución sea más picuda que la normal, igual a ella o más aplastada. Para poder comparar las distribuciones con la normal podemos tomar el estadístico

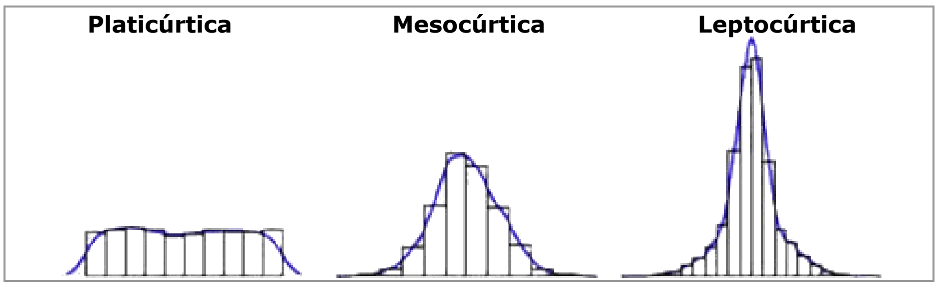
$$a_4 = \frac{1}{\sigma^4} \frac{\sum_{i=0}^{k} \left(x_i - \bar{x}\right)^4 n_i}{n}$$

La distribución normal toma para a_4 el valor 3, por tanto podemos hacer dos cosas tomar este estadístico y clasificar el apuntamiento en función de que su valor sea mayor, igual o menor que 3, o bien hacer una corrección para que el centro de referencia esté en cero. Con esta premisa se define el **coeficiente de aplastamiento de Fisher** (curtosis) como

$$\gamma_2 = \frac{1}{\sigma^4} \frac{\sum_{i=0}^{k} \left(x_i - \bar{x}\right)^4 n_i}{n} - 3 = a_4 - 3$$

Teniendo en cuenta el coeficiente de aplastamiento de Fisher podemos decir que:

- Si $\gamma_2 > 0$ la distribución se llama **Leptocúrtica,** las frecuencias son más apuntadas que la normal.
- Si $\gamma_2 = 0$ la distribución se llama **Mesocúrtica,** la distribución tiene el mismo apuntamiento que la normal.
- Si $\gamma_2 < 0$ se denomina **Platicúrtica** , es menos apuntada que la normal.



Momentos de una variable

Dada una variable aleatoria X con función de probabilidad o densidad f(x) podemos definir una función de X que sea igual a la variable elevada a un exponente entero no negativo.

$$z(x) = x^k \quad \text{siendo} \quad k \in Z \quad k \geq 0$$

El valor esperado de z(x) es el k-ésimo momento de la variable X respecto a su origen y se llama

$$\mu'_k = E[x^k] = \begin{cases} \sum_x x^k f(x) & \text{Si X es discreta} \\ \int_{-\infty}^{\infty} x^k f(x)dx & \text{Si X es continua} \end{cases}$$

- k = 0 $\mu'_0 = E[x^0] = 1$
- k = 1 $\mu'_1 = E[x^1] = E[x] = \mu_X = \mu$

A este primer momento respecto al origen que es igual al valor esperado se le llama también **media aritmética de la variable** y se le denomina μ_X, simplemente **μ**.En la mayoría de los casos, la media μ expresa la tendencia central de la variable o el orden de magnitud de sus valores.

El resto de los momentos respecto al origen tienen escaso interés en la mayoría de los casos.

Unidad V Medidas de correlación

La **correlación** es una medida sobre el grado de relación entre dos variables, sin importar cual es la causa y cual es el efecto. La dependencia de la que se habla en este sentido es la dependencia entre la varianza de las variables.

Como hemos visto el manejo de unidades adimensionales nos permiten tener un coeficiente sobre el que de forma cómoda se pueda trabajar, por lo que podemos dividir entre el producto de las desviaciones de las variables, es decir:

$$r = \frac{S_{xy}}{n\left(S_x S_y\right)}$$

los valores para este coeficiente están comprendidos entre -1 y 1.

Se tiene los siguientes criterios para r

$$r = \begin{cases} r = 1 & \text{la correlación lineal es perfecta, directa o correlación lineal positiva} \\ r = 0 & \text{no existe correlación lineal o correlación lineal nula} \\ r = -1 & \text{la correlación lineal es perfecta, inversa o correlación lineal negativa} \end{cases}$$

entre mas se aproxima a los valores 1 y -1 la aproximación a una correlación se considera buena. Cuando mas se aleja de 1 o de -1 y se acerca a cero se tiene menos confianza en la dependencia lineal por lo que una aproximación lineal será lo menos apropiado, sin embargo no significa que no existe dependencia, lo único que podemos decir es que la dependencia no es lineal. Un valor positivo para *r* indica que a medida que una variable crece la otra también lo hace, por el contrario si su valor es negativo, lo que podemos decir es que a medida que una variable crece la otra decrece.

Datos influyentes

Ejemplos de correlación

Una vez que se determina que existe dependencia lineal un aspecto sumamente relevante es el investigar las características del modelo matemático que relaciona una variable con otra, así de esta forma podemos decir, una variable puede clasificarse como determinístico y probabilistico. El modelo determinístico, que no será abordado en este curso, esta ligado a la ecuación que regula de forma determinante el comportamiento de un fenómeno, así por ejemplo podemos determinar a partir de la obtención de una ecuación sobre el potencial de frenado en un material, que ante cambios de la longitud de onda la relación es lineal no permitirá predecir cuales serán sus valores. Ecuaciones que permiten ver como es la oposición a la corriente eléctrica, o resistencia eléctrica, al aumentar la temperatura de un metal, entre otros, es un claro indicio de una ecuación que es determinística, en ella se podrá describir como cambiara la resistencia eléctrica del material en cuestión ante el aumento de una temperatura en el material. Por otro lado, los fenómenos probabilísticos están sujetos a la modelos que aunque puedan ser descritos por una ecuación no implica que todos los valores que intervienen en el estudio puedan ser localizados en el gráfico que los representan, y por supuesto un dato mas no es garantía que sea localizado en la ecuación.

Recta de regresión.

La **recta de regresión** es la que mejor se ajusta a la **nube de puntos**.

La **recta de regresión** pasa por el punto $(\bar{x}, \bar{y})$ llamado **centro de gravedad**.
Recta de regresión de Y sobre X
La recta de regresión de Y sobre X se utiliza para estimar los valores de la Y a partir de los de la X.
La **pendiente** de la recta es el cociente entre la covarianza y la varianza de la variable X.

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

Recta de regresión de X sobre Y
La **recta de regresión** de X sobre Y se utiliza para estimar los valores de la X a partir de los de la Y.
La **pendiente** de la recta es el cociente entre la covarianza y la varianza de la variable Y.

$$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

Si la correlación es nula, r = 0, las rectas de regresión son perpendiculares entre sí, y sus eucaciones son:

y = $\bar{y}$

x = $\bar{x}$

Ejemplo

Las notas de 12 alumnos de una clase en Matemáticas y Física son las siguientes:

Matemáticas	2	3	4	4	5	6	6	7	7	8	10	10
Física	1	3	2	4	4	4	6	4	6	7	9	10

Hallar las **rectas de regresión** y representarlas.

x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
2	1	2	4	1
3	3	9	9	9
4	2	8	16	4
4	4	16	16	16
5	4	20	25	16
6	4	24	36	16
6	6	36	36	36
7	4	28	49	16
7	6	42	49	36
8	7	56	64	49
10	9	90	100	81
10	10	100	100	100
72	60	431	504	380

1º Hallamos las medias ariméticas.

$$\bar{x} = \frac{72}{12} = 6 \qquad \bar{y} = \frac{60}{12} = 5$$

2º Calculamos la covarianza.

$$\sigma_{xy} = \frac{431}{12} - 6 \cdot 5 = 5.92$$

3º Calculamos las varianzas.

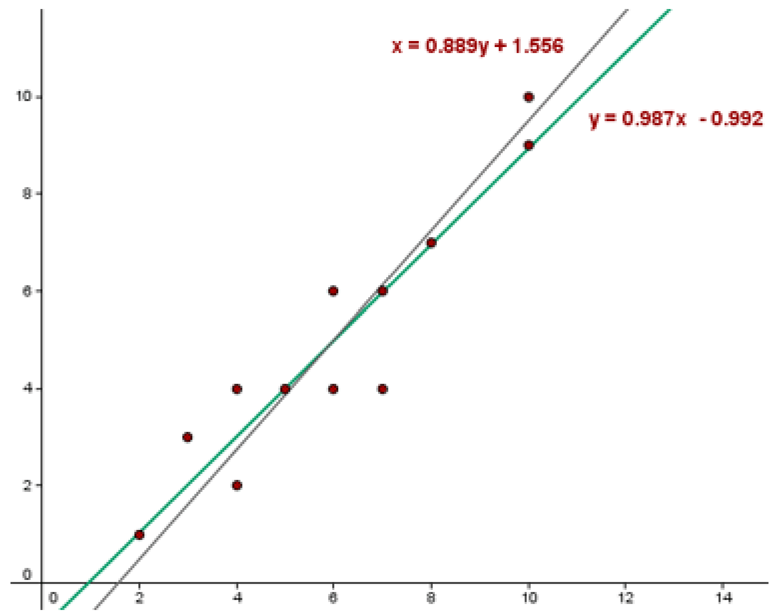
$$\sigma_x^2 = \frac{504}{12} - 6^2 = 6 \qquad \sigma_y^2 = \frac{380}{12} - 25 = 6.66$$

4º Recta de regresión de Y sobre X.

$$y - 5 = \frac{5.92}{6}(x - 6) \qquad y = 0.987x - 0.922$$

4º Recta de regresión de X sobre Y.

$$x - 6 = \frac{5.92}{6.66}(y - 5) \qquad x = 0.889y + 1.556$$



EL error estándar de estimación

La determinante primaria de la exactitud es el grado de dispersión de la población: cuanto mas dispersa este, menor será la exactitud de la estimación. El grado de dispersión en la población se puede estimar a partir del grado de dispersión en las observaciones de la muestra con respecto a la línea de regresión calculada, utilizando la formula.

Se = "———————————————— (yi -yc)

n-2

en la cual:

yi = cada valor de y

yc = valor de línea de regresión correspondiente a partir de la ecuación de regresión.

n = números de observaciones.

La formula anterior no se utiliza por lo general para cálculos reales, es mas fácil trabajar con la formula simplificada

Se "————————————————y2 - a ————————————————y - b ————————————————xy

n - 2

Ejercicios

Resolver lo siguiente:

1. Indica que **variables** son **cualitativas** y cuales **cuantitativas**:
1 Comida Favorita.
2 Profesión que te gusta.
3 Número de goles marcados por tu equipo favorito en la última temporada.
4 Número de alumnos de tu Instituto.
5 El color de los ojos de tus compañeros de clase.
6 Coeficiente intelectual de tus compañeros de clase.

2. De las siguientes **variables** indica cuáles son **discretas** y cuales **continuas**.
1 Número de acciones vendidas cada día en la Bolsa.
2 Temperaturas registradas cada hora en un observatorio.
3 Período de duración de un automóvil.
4 El diámetro de las ruedas de varios coches.
5 Número de hijos de 50 familias.
6 Censo anual de los españoles.

3. Clasificar las siguientes **variables** en **cualitativas** y **cuantitativas discretas** o **continuas**.

1 La nacionalidad de una persona.
2 Número de litros de agua contenidos en un depósito.
3 Número de libros en un estante de librería.
4 Suma de puntos tenidos en el lanzamiento de un par de dados.
5 La profesión de una persona.
6 El área de las distintas baldosas de un edificio.

4. Las puntuaciones obtenidas por un grupo en una prueba han sido:
15, 20, 15, 18, 22, 13, 13, 16, 15, 19, 18, 15, 16, 20, 16, 15, 18, 16, 14, 13.

Construir la **tabla de distribución de frecuencias** y dibuja el **polígono de frecuencias**.

5. El número de estrellas de los hoteles de una ciudad viene dado por la siguiente serie:

3, 3, 4, 3, 4, 3, 1, 3, 4, 3, 3, 3, 2, 1, 3, 3, 3, 2, 3, 2, 2, 3, 3, 3, 2, 2, 2, 2, 2, 3, 2, 1, 1, 1, 2, 2, 4, 1.

Construir la tabla de distribución de frecuencias y dibuja el diagrama de barras.

6. Las calificaciones de 50 alumnos en Matemáticas han sido las siguientes:5, 2, 4, 9, 7, 4, 5, 6, 5, 7, 7, 5, 5, 2, 10, 5, 6, 5, 4, 5, 8, 8, 4, 0, 8, 4, 8, 6, 6, 3, 6, 7, 6, 6, 7, 6, 7, 3, 5, 6, 9, 6, 1, 4, 6, 3, 5, 5, 6, 7.

Construir la **tabla de distribución de frecuencias** y dibuja el **diagrama de barras**.

7. Los pesos de los 65 empleados de una fábrica vienen dados por la siguiente tabla:

Peso	**[50, 60)**	**[60, 70)**	**[70, 80)**	**[80,90)**	**[90, 100)**	**[100, 110)**	**[110, 120)**
f_i	**8**	**10**	**16**	**14**	**10**	**5**	**2**

1 Construir la **tabla de frecuencias.**
2 Representar el **histograma** y el **polígono de frecuencias.**

8. Los 40 alumnos de una clase han obtenido las siguientes puntuaciones, sobre 50, en un examen de Física.

3, 15, 24, 28, 33, 35, 38, 42, 23, 38, 36, 34, 29, 25, 17, 7, 34, 36, 39, 44, 31, 26, 20, 11, 13, 22, 27, 47, 39, 37, 34, 32, 35, 28, 38, 41, 48, 15, 32, 13.

1 Construir la **tabla de frecuencias**.
2 Dibujar el **histograma** y el **polígono de frecuencias.**

9. Sea una distribución estadística que viene dada por la siguiente tabla:

x_i	61	64	67	70	73
f_i	5	18	42	27	8

Calcular:
1 La **moda, mediana y media.**
2 El **rango, desviación media, varianza y desviación típica.**

10. Calcular la **media**, la **mediana** y la **moda** de la siguiente serie de números: 5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5, 4.

11 Hallar la **varianza y la desviación típica** de la siguiente serie de datos:12, 6, 7, 3, 15, 10, 18, 5.

12 Hallar la **media, mediana y moda** de la siguiente serie de números:

3, 5, 2, 6, 5, 9, 5, 2, 8, 6.

13. Hallar la **desviación media, la varianza y la desviación típica** de la series de números siguientes:

2, 3, 6, 8, 11.12, 6, 7, 3, 15, 10, 18, 5.

14 Se ha aplicado un test a los empleados de una fábrica, obteniéndose la siguiente tabla:

	f_i
[38, 44)	7
[44, 50)	8
[50, 56)	15
[56, 62)	25
[62, 68)	18
[68, 74)	9
[74, 80)	6

Dibujar el **histograma** y el **polígono de frecuencias acumuladas**.

15. Dadas las series estadísticas:
3, 5, 2, 7, 6, 4, 9.
3, 5, 2, 7, 6, 4, 9, 1.
Calcular:
La **moda**, la **mediana** y la **media**.
La **desviación media, la varianza** y la **desviación típica**.
Los **cuartiles** 1º y 3º.
Los **deciles** 2º y 7º.
Los **percentiles** 32 y 85.

16. Una distribución estadística viene dada por la siguiente tabla:

	[10, 15)	[15, 20)	[20, 25)	[25, 30)	[30, 35)
f_i	3	5	7	4	2

Hallar:
La **moda, mediana** y **media**.
El **rango**, **desviación media** y **varianza.**
Los **cuartiles** 1º y 3º.
Los **deciles** 3º y 6º.
Los **percentiles** 30 y 70.

17. Dada la distribución estadística:

	[0, 5)	[5, 10)	[10, 15)	[15, 20)	[20, 25)	[25, ∞)
f_i	3	5	7	8	2	6

Calcular:
La **mediana y moda**.
Cuartil 2º y 3º.
Media.