ANTOLOGIA

Materia

Estadística descriptiva

Licenciatura:

Administración y estrategia de negocios

Cuatrimestre:

Tercero

Marco Estratégico de Referencia

Antecedentes históricos

Nuestra Universidad tiene sus antecedentes de formación en el año de 1979 con el inicio de actividades de la normal de educadoras "Edgar Robledo Santiago", que en su momento marcó un nuevo rumbo para la educación de Comitán y del estado de Chiapas. Nuestra escuela fue fundada por el Profesor Manuel Albores Salazar con la idea de traer educación a Comitán, ya que esto representaba una forma de apoyar a muchas familias de la región para que siguieran estudiando.

En el año 1984 inicia actividades el CBTiS Moctezuma Ilhuicamina, que fue el primer bachillerato tecnológico particular del estado de Chiapas, manteniendo con esto la visión en grande de traer educación a nuestro municipio, esta institución fue creada para que la gente que trabajaba por la mañana tuviera la opción de estudiar por las tardes.

La Maestra Martha Ruth Alcázar Mellanes es la madre de los tres integrantes de la familia Albores Alcázar que se fueron integrando poco a poco a la escuela formada por su padre, el Profesor Manuel Albores Salazar; Víctor Manuel Albores Alcázar en julio de 1996 como chofer de transporte escolar, Karla Fabiola Albores Alcázar se integró en la docencia en 1998, Martha Patricia Albores Alcázar en el departamento de cobranza en 1999.

En el año 2002, Víctor Manuel Albores Alcázar formó el Grupo Educativo Albores Alcázar S.C. para darle un nuevo rumbo y sentido empresarial al negocio familiar y en el año 2004 funda la Universidad Del Sureste.

La formación de nuestra Universidad se da principalmente porque en Comitán y en toda la región no existía una verdadera oferta Educativa, por lo que se veía urgente la creación de una institución de Educación superior, pero que estuviera a la altura de las exigencias de los jóvenes

que tenían intención de seguir estudiando o de los profesionistas para seguir preparándose a través de estudios de posgrado.

Nuestra Universidad inició sus actividades el 18 de agosto del 2004 en las instalaciones de la 4ª avenida oriente sur no. 24, con la licenciatura en Puericultura, contando con dos grupos de cuarenta alumnos cada uno. En el año 2005 nos trasladamos a nuestras propias instalaciones en la carretera Comitán – Tzimol km. 57 donde actualmente se encuentra el campus Comitán y el corporativo UDS, este último, es el encargado de estandarizar y controlar todos los procesos operativos y educativos de los diferentes campus, así como de crear los diferentes planes estratégicos de expansión de la marca.

Misión

Satisfacer la necesidad de Educación que promueva el espíritu emprendedor, aplicando altos estándares de calidad académica, que propicien el desarrollo de nuestros alumnos, Profesores, colaboradores y la sociedad, a través de la incorporación de tecnologías en el proceso de enseñanza-aprendizaje.

Visión

Ser la mejor oferta académica en cada región de influencia, y a través de nuestra plataforma virtual tener una cobertura global, con un crecimiento sostenible y las ofertas académicas innovadoras con pertinencia para la sociedad.

Valores

- Disciplina
- Honestidad
- Equidad
- Libertad

Escudo

El escudo del Grupo Educativo Albores Alcázar S.C. está constituido por tres líneas curvas que nacen de izquierda a derecha formando los escalones al éxito. En la parte superior está situado un cuadro motivo de la abstracción de la forma de un libro abierto.

Eslogan

"Mi Universidad"

ALBORES

Es nuestra mascota, un Jaguar. Su piel es negra y se distingue por ser líder, trabaja en equipo y obtiene lo que desea. El ímpetu, extremo valor y fortaleza son los rasgos que distinguen.

ESTADISTICA DESCRIPTIVA

Objetivo de la materia:

La asignatura de Estadística Descriptiva pretende introducir al estudiante en el conocimiento de las técnicas estadísticas básicas, con el objetivo de que lo ayuden en su futuro profesional a la hora de tomar decisiones en cualquier entorno laboral y de relaciones humanas.

CONTENIDO

INDICE

UNIDAD I

INTRODUCCION A LOS DATOS AGRUPADOS

1.- La estadística descriptiva

La estadística descriptiva es, junto con la inferencia estadística o estadística inferencial, una de las dos grandes ramas de la estadística. Su propio nombre lo indica, trata de describir algo. Pero no describirlo de cualquiera forma, sino de manera cuantitativa. Pensemos en el peso de una caja de verduras, en la altura de una persona o en la cantidad de dinero que gana una empresa. De estas variables podríamos decir muchas cosas. Por ejemplo, podríamos indicar que esta o aquella caja de tomates pesan mucho o pesan menos que otras. Siguiendo con otro ejemplo, podríamos decir que el ingreso de una empresa varía mucho a lo largo del tiempo o que una persona tiene una altura promedio.

Para dictar las afirmaciones anteriores, sobre mucho, poco, alto, bajo, muy variable o poco variable necesitamos variables de medidas. Esto es, necesitamos cuantificarlas, ofrecer un número. Con esto en mente, podríamos utilizar los gramos o los kilogramos como unidad de medida para saber el peso de tantas cajas de tomates como consideremos. Una vez pesemos treinta cajas, sabremos cuales pesan más, cuales pesan menos, que cuantía es la que más se repite o si existe mucha disparidad entre los pesos de las diferentes cajas.

Con esta idea nace la estadística descriptiva, con la de recoger datos, almacenarlos, realizar tablas o incluso gráficos que nos ofrezcan información sobre un determinado asunto. Adicionalmente, nos ofrecen medidas que resumen la información de una gran cantidad de datos.

1.1.- Tipos de variables

Dentro de la estadística descriptiva, podemos describir los datos de manera cualitativa o cuantitativa.

Variable cualitativa: Hace referencia a una cualidad. Ejemplos: el color de ojos de una persona o el color de pelo.

Variable cuantitativa: Hace referencia a una medida cuantitativa. Ejemplos: la altura de una persona en centímetros o el peso de una persona en kilogramos.

Así pues, sobre estas variables se pueden calcular ciertos parámetros. Especialmente sobre las variables cuantitativas. Ya que, por ejemplo, ¿cuál es el valor promedio del color de ojos? Si hay cinco personas con color de ojos azul y cinco con color de ojos verde, el promedio no será que tienen un color de ojos promedio de azul-verde. Por tanto, en ese caso no sería posible calcular algunos de los parámetros que veremos a continuación.

Variable estadística

Parámetros estadísticos básicos

Con el objetivo de resumir la información, se idearon diversas fórmulas que ofrecían medidas de un determinado tipo. Así, están aquellas que nos ofrecen información sobre el centro, otras sobre la dispersión o variabilidad y otras sobre la posición de un valor.

Medidas de tendencia central: Denominadas así porque ofrecen información sobre el centro de conjunto de datos. Por ejemplo, la media es una medida de tendencia o posición central ya que el promedio nos ofrece un valor centrado del conjunto de datos. ¿Dónde podríamos decir que se encuentra el punto medio? En el centro, en la mitad aproximadamente. Otro ejemplo de medida de tendencia central es la mediana.

Medidas de dispersión: También son conocidas como medidas de variabilidad. Por ejemplo, la desviación típica es una medida de variabilidad ya que nos dice si los valores de un conjunto de datos son muy dispares o no. Dos ejemplos más sobre medidas de dispersión podrían ser la varianza y el rango estadístico.

Medidas de posición: No son las más conocidas, pero se utilizan frecuentemente. Un ejemplo de ello, se encuentra en los percentiles o los deciles. Cuando un dato en concreto se encuentra

en el percentil 90, quiere decir que por debajo de ese dato se encuentran el 90% de datos. Existen otras medidas de posición como los cuartiles o algunas variantes como el primer cuartil.

1.2.- Conceptos básicos estadística

Universo:

En estadística es el nombre especifico que recibe particularmente en la investigación social la operación dentro de la delimitación del campo de investigación que tienen por objeto la determinación del conjunto de unidades de observaciones del conjunto de unidades de observación que van a ser investigadas. Para muchos investigadores él termino universo y población son sinónima. En general, el universo es la totalidad de elementos o características que conforman el ámbito de un estudio o investigación.

Población:

En estadística el concepto de población va más allá de lo que comúnmente se conoce como tal. En términos estadísticos, población es un conjunto finito o infinito de personas, animales o cosas que presentan características comunes, sobre los cuales se quiere efectuar un estudio determinado. En otras palabras, la población se define como la totalidad de los valores posibles (mediciones o conteos) de una característica particular de un grupo especificado de personas, animales o cosas que se desean estudiar en un momento determinado. Así, se puede hablar de la población de habitantes de un país, de la población de estudiantes universitarios de la zona sur del Estado Anzoátegui, de la población de casas de la Urbanización Los Ríos de la ciudad de El Tigre, el rendimiento académico de los estudiantes del IUTJAA, el número de carros marca Corola de la ciudad de El Tigre, la estatura de un grupo alumnos del IUTJAA, la talla, etc.

Muestra:

La muestra es un subconjunto de la población, seleccionado de tal forma, que sea representativo de la población en estudio, obteniéndose con el fin de investigar alguna o algunas de las propiedades de la población de la cual procede. En otras palabras es una parte de la

población que sirve para representarla. Según el DRAE, es una parte o porción extraída de un conjunto por métodos que permiten considerarla como representativa del mismo.

Entonces, una muestra no es más que una parte de la población que sirve para representarla. La muestra debe obtenerse de la población que se desea estudiar; una muestra debe ser definida sobre la base de la población determinada, y las conclusiones que se obtengan de dicha muestra sólo podrán referirse a la población en referencia.

Muestreo:

Es el procedimiento mediante el cual se obtiene una o más muestras de una población determinada. Existen dos tipos de muestreos a saber:

Los Parámetros:

Son cualquiera característica que se pueda medir y cuya medición se lleve a cabo sobre todos los elementos que integran una población determinada, los mismos suelen representarse con letras griegas. El valor de un parámetro poblacional es un valor fijo en un momento dado. Ejemplo: La media Aritmética = m (miu), La desviación Típica = s, (Sigma) etcétera.

Dato estadístico:

Es un conjunto de valores numéricos que tienen relación significativa entre sí. Los mismos pueden ser comparados, analizados e interpretados en una investigación cualquiera. Se puede afirmar que son las expresiones numéricas obtenidas como consecuencia de observar un individuo de la población; por lo tanto, son las características que se han tomado en cuenta de cualquiera población para una investigación determinada.

Frecuencia:

La frecuencia es el número de veces que se repite (aparece) el mismo dato estadístico en un conjunto de observaciones de una investigación determinada, las frecuencias se les designan con las letras fi, y por lo general se les llaman frecuencias absolutas.

Distribución de Frecuencia:

En estadística existe una relación con cantidades, números agrupados o no, los cuales poseen entre sí características similares. Existen investigaciones relacionadas con los precios de los productos de la dieta diaria, la estatura y el peso de un grupo de individuos, los salarios de los empleados, los grados de temperatura del medio ambiente, las calificaciones de los estudiantes, etc., que pueden adquirir diferentes valores gracias a una unidad apropiada, que recibe el nombre de variable. La representación numérica de las variables se denomina dato estadístico. La distribución de frecuencia es una disposición tabular de datos estadísticos, ordenados ascendente o descendentemente, con la frecuencia (fi) de cada dato. Las distribuciones de frecuencias pueden ser para datos no agrupados y para datos agrupados o de intervalos de clase.

Distribución de frecuencia para datos no Agrupados:

Es aquella distribución que indica las frecuencias con que aparecen los datos estadísticos, desde el menor de ellos hasta el mayor de ese conjunto sin que se haya hecho ninguna modificación al tamaño de las unidades originales. En estas distribuciones cada dato mantiene su propia identidad después que la distribución de frecuencia se ha elaborado. En estas distribuciones los valores de cada variable han sido solamente reagrupados, siguiendo un orden lógico con sus respectivas frecuencias.

Distribución de frecuencia de clase o de datos Agrupados:

Es aquella distribución en la que las disposiciones tabulares de los datos estadísticos se encuentran ordenados en clases y con la frecuencia de cada clase; es decir, los datos originales de varios valores adyacentes del conjunto se combinan para formar un intervalo de clase. No existen normas establecidas para determinar cuándo es apropiado utilizar datos agrupados o datos no agrupados; sin embargo, se sugiere que cuando el número total de datos (N) es igual o superior 50 y además el rango o recorrido de la serie de datos es mayor de 20, entonces, se

utilizará la distribución de frecuencia para datos agrupados, también se utilizará este tipo de distribución cuando se requiera elaborar gráficos lineales como el histograma, el polígono de frecuencia o la ojiva.

La razón fundamental para utilizar la distribución de frecuencia de clases es proporcionar mejor comunicación acerca del patrón establecido en los datos y facilitar la manipulación de los mismos. Los datos se agrupan en clases con el fin de sintetizar, resumir, condensar o hacer que la información obtenida de una investigación sea manejable con mayor facilidad.

Componentes de una distribución de frecuencia de clase

1.- Rango o Amplitud total (recorrido).- Es el límite dentro del cual están comprendidos todos los valores de la serie de datos, en otras palabras, es el número de diferentes valores que toma la variable en un estudio o investigación dada. Es la diferencia entre el valor máximo de una variable y el valor mínimo que ésta toma en una investigación cualquiera. El rango es el tamaño del intervalo en el cual se ubican todos los valores que pueden tomar los diferentes datos de la serie de valores, desde el menor de ellos hasta el valor mayor estando incluidos ambos extremos. El rango de una distribución de frecuencia se designa con la letra R.

2.- Clase o Intervalo de clase.- Son divisiones o categorías en las cuales se agrupan un conjunto de datos ordenados con características comunes. En otras palabras, son fraccionamientos del rango o recorrido de la serie de valores para reunir los datos que presentan valores comprendidos entre dos límites. Para organizar los valores de la serie de datos hay que determinar un número de clases que sea conveniente. En otras palabras, que ese número de intervalos no origine un número pequeño de clases ni muy grande. Un número de clases pequeño puede ocultar la naturaleza natural de los valores y un número muy alto puede provocar demasiados detalles como para observar alguna información de gran utilidad en la investigación.

Tamaño de los Intervalos de Clase

Los intervalos de clase pueden ser de tres tipos, según el tamaño que estos presenten en una distribución de frecuencia:

a) Clases de igual tamaño, b) clases desiguales de tamaño y c) clases abiertas.

3.- Amplitud de Clase, Longitud o Ancho de una Clase

La amplitud o longitud de una clase es el número de valores o variables que concurren a una clase determinada. La amplitud de clase se designa con las letras Ic. Existen diversos criterios para determinar la amplitud de clases, ante esa diversidad de criterios, se ha considerado que lo más importante es dar un ancho o longitud de clase a todos los intervalos de tal manera que respondan a la naturaleza de los datos y al objetivo que se persigue y esto se logra con la práctica.

4.-Punto medio o Marca de clase

El centro de la clase, es el volar de los datos que se ubica en la posición central de la clase y representa todos los demás valores de esa clase. Este valor se utiliza para el cálculo de la media aritmética.

5.-Frecuencia de clase

La frecuencia de clase se le denomina frecuencia absoluta y se le designa con las letras fi. Es el número total de valores de las variables que se encuentran presente en una clase determinada, de una distribución de frecuencia de clase.

6.- Frecuencia Relativa

La frecuencia relativa es aquella que resulta de dividir cada uno de los fi de las clases de una distribución de frecuencia de clase entre el número total de datos(N) de la serie de valores. Estas frecuencias se designan con las letras fr; si cada fr se multiplica por 100 se obtiene la frecuencia relativa porcentual (fr %).

7.-Frecuencias acumuladas

Las frecuencias acumuladas de una distribución de frecuencias son aquellas que se obtienen de las sumas sucesivas de las fi que integran cada una de las clases de una distribución de frecuencia de clase, esto se logra cuando la acumulación de las frecuencias se realiza tomando en cuenta la primera clase hasta alcanzar la ultima. Las frecuencias acumuladas se designan con las letras fa. Las frecuencias acumuladas pueden ser menor que (fa< que) y frecuencias acumuladas mayor que (fa>que).

8.- Frecuencia acumulada relativa

La frecuencia acumulada relativa es aquella que resulta de dividir cada una de las fa de las diferentes clases que integran una distribución de frecuencia de clase entre el número total de datos (N) de la serie de valores, estas frecuencias se designan con las letras far. Si las frace multiplican por 100 se obtienen las frecuencias acumuladas relativas porcentuales y las mismas se designan así: far %.

La mediana

La mediana (Md) es una medida de posición que divide a la serie de valores en dos partes iguales, un cincuenta por ciento que es mayor o igual a esta y otro cincuenta por ciento que es menor o igual que ella. Es por lo tanto, un parámetro que esta en el medio del ordenamiento o arreglo de los datos organizados, entonces, la mediana divide la distribución en una forma tal que a cada lado de la misma queda un número igual de datos.

Para encontrar la mediana en una serie de datos no agrupados, lo primero que se hace es ordenar los datos en una forma creciente o decreciente y luego se ubica la posición que esta ocupa en esa serie de datos; para ello hay que determinar si la serie de datos es par o impar, luego el número que se obtiene indica el lugar o posición que ocupa la mediana en la serie de valores, luego la mediana será el número que ocupe el lugar de lo posición encontrada.

La moda

La moda es la medida de posición que indica la magnitud del valor que se presenta con más frecuencia en una serie de datos; es pues, el valor de la variable que más se repite en un conjunto de datos. De las medias de posición la moda es la que se determina con mayor facilidad, ya que se puede obtener por una simple observación de los datos en estudio, puesto que la moda es el dato que se observa con mayor frecuencia. La moda se designa con las letras Mo.

Desviación típica o estándar

Es la medida de dispersión más utilizada en las investigaciones por ser la más estable de todas, ya que para su cálculo se utilizan todos los desvíos con respecto a la media aritmética de las

observaciones, y además, se toman en cuenta los signos de esos desvíos. Se le designa con la letra castellana S cuando se trabaja con una muestra y con la letra griega minúscula s (Sigma) cuando se trabaja con una población. Es importante destacar que cuando se hace referencia a la población él número de datos se expresa con N y cuando se refiere a la muestra él número de datos se expresa con n. La desviación típica se define como:

Interpretación de la desviación estándar

La desviación típica como medida absoluta de dispersión, es la que mejor nos proporciona la variación de los datos con respecto a la media aritmética, su valor se encuentra en relación directa con la dispersión de los datos, a mayor dispersión de ellos, mayor desviación típica, y a menor dispersión, menor desviación típica.

Varianza

Es otra de las variaciones absolutas y la misma se define como el cuadrado de la desviación típica; viene expresada con las mismas letras de la desviación típica pero elevada al cuadrado, así S2 y s2. Las fórmulas para calcular la varianza son las mismas utilizadas por la desviación típica, exceptuando las respectivas raíces, las cuales desaparecen al estar elevados el primer miembro al cuadrado

La Estadística dentro del Método Científico

La estadística no se puede utilizar como una caja mágica para extraer certezas, donde se introducen datos y se extraen leyes. La estadística, en el contexto de probabilidades y técnicas de inferencia, es incapaz por sí misma de suplantar al Método Científico, sólo es un gran apoyo.

1.3.- Tabla de datos agrupados

¿Cómo se elabora una Tabla de Distribución de Frecuencias para Datos Agrupados?

Por lo general una tabla de frecuencias con datos agrupados se realiza cuando la cantidad de datos es grande y/o la variable es continua.

Básicamente consiste en agrupar los datos en intervalos de una misma amplitud, denominados clases. A cada clase se le asignan valores de cada tipo de frecuencias.

Vamos directo al punto con un ejemplo:

Consultamos a 50 personas sobre cuál era su edad y obtuvimos los siguientes resultados:

38 – 15 – 10 – 12 – 62 – 46 – 25 – 56 – 27 – 24 – 23 – 21 – 20 – 25 – 38 – 27 – 48 – 35 – 50 – 65 – 59 – 58 – 47 – 42 – 37 – 35 – 32 – 40 – 28 – 14 – 12 – 24 – 66 – 73 – 72 – 70 – 68 – 65 – 54 – 48 – 34 – 33 – 21 – 19 – 61 – 59 – 47 – 46 – 30 – 30

Paso 1: Identificar el valor máximo y mínimo

38 – 15 – *10* – 12 – 62 – 46 – 25 – 56 – 27 – 24 – 23 – 21 – 20 – 25 – 38 – 27 – 48 – 35 – 50 – 65 – 59 – 58 – 47 – 42 – 37 – 35 – 32 – 40 – 28 – 14 – 12 – 24 – 66 – *73* – 72 – 70 – 68 – 65 – 54 – 48 – 34 – 33 – 21 – 19 – 61 – 59 – 47 – 46 – 30 – 30

Valor máximo: 73 años
Valor mínimo: 10 años

Paso 2: Calcular el Rango

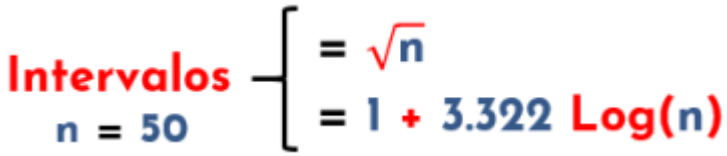
Obtener el rango de edades en que se encuentran los encuestados, sólo basta con determinar la diferencia que hay entre el más joven y el más adulto:

Rango = Valor máximo - Valor mínimo
Rango = 73 - 10
Rango = 63 años

Paso 3: Calcular la cantidad de Intervalos

A los intervalos también se les conoce como clases. Simplemente son las «*categorías*» en las cuales vamos a encasillar a nuestros encuestados.

Hay varias formas de calcular cuántos intervalos debemos utilizar. Vamos a analizar un par:



Para ambas formas de calcular la cantidad de intervalos a utilizar, el valor de n corresponde a la cantidad de datos que tenemos para analizar. En este caso son 50 datos.

Con la primera forma tendríamos que redondear el resultado, ya que los intervalos corresponde a cantidades enteras (no puedes tener un intervalo y medio… o un intervalo y alguito… debes aproximar como NORMALMENTE lo harías).

La segunda forma se conoce como Regla de Sturges, y el resultado obtenido lo debes aproximar por ARRIBA, es decir, al entero siguiente (por ejemplo si te da 5.1 lo debes aproximar a 6 y no a 5). Para nuestro ejemplo:

Por ambas formas obtuvimos que debemos utilizar 7 intervalos.

Paso 4: Calcular la Amplitud de los Intervalos

Ya sabemos el Rango de edad en la que se mueven nuestros encuestados…. y sabemos entre cuántos intervalos hay que REPARTIR las categorías… Así se calcula la amplitud:

Paso 5: Construcción de los intervalos

El primer intervalo viene con límite inferior igual al valor mínimo de los datos, en este caso 10 años. Súmale el valor de la amplitud, es decir, 9 años, y obtendrás el límite superior de 19 años. Eso nos daría el primer intervalo:

$$[10 - 19)$$

Ojo! Fíjate bien, se utiliza corchete para el dato que SE INCLUYE… y se utiliza paréntesis para el dato que NO SE INCLUYE. Eso significa que los datos de 10 años se cuentan pero los de 19 NO.

El 19 se cuenta en el siguiente intervalo y allí vendría siendo el límite inferior. Súmale el valor de la amplitud, es decir, 9 años, y obtendrás el límite superior de 28 años. Eso nos daría el segundo intervalo:

$$[19 - 28)$$

El uso del corchete implica que SÍ vamos a contar acá el 19 pero el paréntesis indica que NO vamos a incluir a los de 28 años. Ese se incluye en el siguiente.

Veamos los 7 intervalos construidos:

Edad (x)
[10 - 19)
[19 - 28)
[28 - 37)
[37 - 46)
[46 - 55)
[55 - 64)
[64 - 73]

Si te fijas bien, el último intervalo debe finalizar en el valor máximo, es decir, 73 años. Lógicamente ese último intervalo debe concluir con corchetes para no dejar por fuera el dato de 73 años.

Paso 6: Cálculo de la Marca de Clase de cada intervalo

La marca de clase simplemente es el punto medio que hay en cada intervalo.

Lo que debes hacer es sumar límite inferior y superior de cada intervalo y dividir el resultado entre 2. Así:

Edad (x)	Marca de Clase (X_i)
[10 - 19)	14.5
[19 - 28)	23.5
[28 - 37)	32.5
[37 - 46)	41.5
[46 - 55)	50.5
[55 - 64)	59.5
[64 - 73]	68.5

$$\frac{10+19}{2} = 14.5 \qquad \frac{19+28}{2} = 23.5 \qquad \frac{28+37}{2} = 32.5 \qquad \frac{37+46}{2} = 41.5$$

$$\frac{46+55}{2} = 50.5 \qquad \frac{55+64}{2} = 59.5 \qquad \frac{64+73}{2} = 68.5$$

Paso 7: Determinar la Frecuencia Absoluta de cada intervalo La frecuencia absoluta sólo consiste en CONTAR la cantidad de datos que caen en cada intervalo. Se representa con la f minúscula y un subíndice (número chiquito abajo) que indica el intervalo en el cual está ubicada la frecuencia absoluta (f_i).

Veamos cuántos datos caen en el primer intervalo de [10 – 19)

Edades de 50 personas: 38 - 15 - 10 - 12 - 62 - 46 - 25 - 56 - 27 - 24 - 23 - 21 - 20 - 25 - 38 - 27 - 48 - 35 - 50 - 65 - 59 - 58 - 47 - 42 - 37 - 35 - 32 - 40 - 28 - 14 - 12 - 24 - 66 - 73 - 72 - 70 - 68 - 65 - 54 - 48 - 34 - 33 - 21 - 19 - 61 - 59 - 47 - 46 - 30 - 30

Si te fijas bien, NO estamos contando los datos de 19 años… esos se cuentan en el siguiente intervalo. Para el primer intervalo tenemos 5 datos, esa será su frecuencia absoluta, su CONTEO.

Veamos cuántos datos caen en el segundo intervalo de [19 – 28)

Edades de 50 personas: 38 - 15 - 10 - 12 - 62 - 46 - 25 - 56 - 27 - 24 - 23 - 21 - 20 - 25 - 38 - 27 - 48 - 35 - 50 - 65 - 59 - 58 - 47 - 42 - 37 - 35 - 32 - 40 - 28 - 14 - 12 - 24 - 66 - 73 - 72 - 70 - 68 - 65 - 54 - 48 - 34 - 33 - 21 - 19 - 61 - 59 - 47 - 46 - 30 - 30

Si te fijas bien, NO estamos contando los datos de 28 años… esos se cuentan en el siguiente intervalo. Para el segundo intervalo tenemos 11 datos, esa será su frecuencia absoluta, su CONTEO.Veamos cuántos datos caen en el tercer intervalo de [28 – 37)

Edades de 50 personas: 38 - 15 - 10 - 12 - 62 - 46 - 25 - 56 - 27 - 24 - 23 - 21 - 20 - 25 - 38 - 27 - 48 - 35 - 50 - 65 - 59 - 58 - 47 - 42 - 37 - 35 - 32 - 40 - 28 - 14 - 12 - 24 - 66 - 73 - 72 - 70 - 68 - 65 - 54 - 48 - 34 - 33 - 21 - 19 - 61 - 59 - 47 - 46 - 30 - 30

Si te fijas bien, NO estamos contando los datos de 37 años… esos se cuentan en el siguiente intervalo. Para el tercer intervalo tenemos 8 datos, esa será su frecuencia absoluta, su CONTEO. Estas son las frecuencias absolutas de los 7 intervalos:

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)
[10 - 19)	14.5	5
[19 - 28)	23.5	11
[28 - 37)	32.5	8
[37 - 46)	41.5	5
[46 - 55)	50.5	8
[55 - 64)	59.5	6
[64 - 73]	68.5	7
	Total	50

Evidentemente la sumatoria de todas las frecuencias absolutas debe arrojar el número de datos que tenemos, en este caso 50.

Paso 8: Determinar la Frecuencia Absoluta Acumulada de cada intervalo

No te compliques, ACUMULAR es SUMAR todo lo que llevo hasta el momento.

La Frecuencia Absoluta Acumulada (F_i) de cada intervalo consiste en sumar todas las frecuencias absolutas de los intervalos anteriores y el actual. Para diferenciar su símbolo de la frecuencia absoluta, simplemente utiliza la F mayúscula.

La primer frecuencia absoluta acumulada es la misma primer frecuencia absoluta porque recién estamos empezando… no hay nada que acumular todavía.

La segunda frecuencia absoluta acumulada vale 16 porque debemos sumar 5+11 porque son las frecuencias absolutas que llevamos hasta ahora para ACUMULAR.

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)
[10 - 19)	14.5	5	5
[19 - 28)	23.5	11	16
[28 - 37)	32.5	8	
[37 - 46)	41.5	5	
[46 - 55)	50.5	8	
[55 - 64)	59.5	6	
[64 - 73]	68.5	7	
	Total	50	

La tercera frecuencia absoluta acumulada vale 24 porque debemos sumar 5+11+8 porque son las frecuencias absolutas que llevamos hasta ahora para ACUMULAR.

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)
[10 - 19)	14.5	5	5
[19 - 28)	23.5	11	16
[28 - 37)	32.5	8	24
[37 - 46)	41.5	5	
[46 - 55)	50.5	8	
[55 - 64)	59.5	6	
[64 - 73]	68.5	7	
	Total	50	

La cuarta frecuencia absoluta acumulada vale 29 porque debemos sumar 5+11+8+5 porque son las frecuencias absolutas que llevamos hasta ahora para ACUMULAR.

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)
[10 - 19)	14.5	5	5
[19 - 28)	23.5	11	16
[28 - 37)	32.5	8	24
[37 - 46)	41.5	5	29
[46 - 55)	50.5	8	
[55 - 64)	59.5	6	
[64 - 73]	68.5	7	
	Total	50	

Cuando llegues al último intervalo, deberás obtener un ACUMULADO igual al TOTAL de datos, en este caso 50:

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)
[10 - 19)	14.5	5	5
[19 - 28)	23.5	11	16
[28 - 37)	32.5	8	24
[37 - 46)	41.5	5	29
[46 - 55)	50.5	8	37
[55 - 64)	59.5	6	43
[64 - 73]	68.5	7	50
	Total	50	

Paso 9: Determinar la Frecuencia Relativa de cada intervalo

La palabra RELATIVA nos indica que vamos a RELACIONAR cada Frecuencia Absoluta con su Total… y en matemáticas cuando te dicen relacionar algo con algo… es DIVIDIR ese algo con ese algo.Un pequeño ejemplo con dinero (eso hace más llamativas las cosas… ¿no?)Todos en mi familia aportan plata para el mercado mensual… entre todos aportamos un TOTAL de 200 dólares. De esos 200, yo sólo aporto 20 dólares.Vamos a obtener la RELACIÓN de MI APORTE respecto al TOTAL.Fácil, 20 ÷ 200 = 0.1

Si lo convierto a porcentaje… 0.1 x 100% = 10%

Entonces MI APORTE RELATIVO es del 10% del TOTAL.

Espero que hayas entendido a qué se refiere la palabra RELATIVO.

La Frecuencia Relativa (f_r) de cada intervalo consiste en dividir la Frecuencia Absoluta de es mismo intervalo entre el Total de datos.

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		
[10 - 19)	14.5	5	5	0.1	10%	5 ÷ 50
[19 - 28)	23.5	11	16	0.22	22%	11 ÷ 50
[28 - 37)	32.5	8	24	0.16	16%	8 ÷ 50
[37 - 46)	41.5	5	29	0.1	10%	5 ÷ 50
[46 - 55)	50.5	8	37	0.16	16%	8 ÷ 50
[55 - 64)	59.5	6	43	0.12	12%	6 ÷ 50
[64 - 73]	68.5	7	50	0.14	14%	7 ÷ 50
	Total	50	**Total**	1	100%	

De la tabla construida hasta ahora, podemos observar que la frecuencia relativa se puede expresar en decimal o en porcentaje, y que la suma de todas las frecuencias relativas debe dar el 100%.

Paso 10: Determinar la Frecuencia Relativa Acumulada de cada intervalo

Vuelve y juega lo acumulado… no te compliques, ACUMULAR es SUMAR todo lo que llevo hasta el momento.

La Frecuencia Relativa Acumulada (F_r) de cada intervalo consiste en sumar todas las frecuencias relativas de los intervalos anteriores y el actual. Para diferenciar su símbolo de la frecuencia relativa, simplemente utiliza la F mayúscula.

La primera frecuencia relativa acumulada es la misma primera frecuencia relativa porque recién estamos empezando… no hay nada que acumular todavía.

La segunda frecuencia relativa acumulada vale 0.32 porque debemos sumar 0.1+0.22 porque son las frecuencias relativas que llevamos hasta ahora para ACUMULAR.

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		Frecuencia relativa acumulada (F_r)	
[10 - 19)	14.5	5	5	0.1	10%	0.1	10%
[19 - 28)	23.5	11	16	0.22	22%	0.32	32%
[28 - 37)	32.5	8	24	0.16	16%		
[37 - 46)	41.5	5	29	0.1	10%		
[46 - 55)	50.5	8	37	0.16	16%		
[55 - 64)	59.5	6	43	0.12	12%		
[64 - 73]	68.5	7	50	0.14	14%		
	Total	50	**Total**	1	100%		

La tercer frecuencia relativa acumulada vale 0.48 porque debemos sumar 0.1+0.22+0.16 porque son las frecuencias relativas que llevamos hasta ahora para ACUMULAR.

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		Frecuencia relativa acumulada (F_r)	
[10 - 19)	14.5	5	5	0.1	10%	0.1	10%
[19 - 28)	23.5	11	16	0.22	22%	0.32	32%
[28 - 37)	32.5	8	24	0.16	16%	0.48	48%
[37 - 46)	41.5	5	29	0.1	10%		
[46 - 55)	50.5	8	37	0.16	16%		
[55 - 64)	59.5	6	43	0.12	12%		
[64 - 73]	68.5	7	50	0.14	14%		
	Total	50	**Total**	1	100%		

Seguro ya entendiste la dinámica… veamos de una vez todas las Frecuencias Relativas Acumuladas de nuestro ejemplo:

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		Frecuencia relativa acumulada (F_r)	
[10 - 19)	14.5	5	5	0.1	10%	0.1	10%
[19 - 28)	23.5	11	16	0.22	22%	0.32	32%
[28 - 37)	32.5	8	24	0.16	16%	0.48	48%
[37 - 46)	41.5	5	29	0.1	10%	0.58	58%
[46 - 55)	50.5	8	37	0.16	16%	0.74	74%
[55 - 64)	59.5	6	43	0.12	12%	0.86	86%
[64 - 73]	68.5	7	50	0.14	14%	1	100%
	Total	50	**Total**	1	100%		

1.4.- Cuartiles, Deciles, Percentiles

Cuartiles para datos agrupados:

Para hallar los tres Cuartiles (Q) para datos agrupados se aplica la siguiente fórmula:

$$Q_k = L_i + a_i \cdot \frac{k \cdot \frac{n}{4} - F_{i-1}}{f_i}$$

Donde:

Qk : Cuartil

Li: Límite inferior del intervalo seleccionado.

k : Debe ser 1 ; 2 ó 3

n: Número total de datos

f : frecuencia absoluta del intervalo seleccionado.

Fi-1 : Frecuencia absoluta Acumulada (pero anterior a la clase cuartil)

a : Amplitud del intervalo (Restar los 2 valores: L sup - L inf)

Pasos a seguir:

1. Completar la tabla "llenando" la Frecuencia Absoluta Acumulada "F".

2. Encontrar la Clase Cuartil:

$$Clase\ cuartil = \frac{k.n}{4}$$

Luego en la Columna "F" escoger el primer valor mayor que la clase cuartil encontrada.

3. Aplicar la fórmula.

EJEMPLOS:

1. Encuentra los tres cuartiles de la siguiente Tabla de frecuencias que muestra el tiempo de servicio de trabajadores de la empresa ABC.

Años de servicio	f_i
1 – 8	9
8 – 15	7
15 – 22	12
22 – 29	16
29 – 36	8
36 – 93	3

Solución:

Completamos la tabla hallando F:

Años de servicio	f_i	F_i
1 – 8	9	**9**
8 – 15	7	**16**
15 – 22	12	**28**
22 – 29	16	**44**
29 – 36	8	**52**
36 – 93	3	**55**
TOTAL	**n=55**	

Encontramos las clase cuartil y la pintamos.

Años de servicio	f_i	F_i
1 – 8	9	9
8 – 15	**7**	**16**
15 – 22	**12**	**28**
22 – 29	**16**	**44**
29 – 36	8	52
36 – 93	3	55
TOTAL	n=55	

Cuartil 1:

$$Clase\ cuartil = \frac{1\ (55)}{4} = 13{,}75$$

El primer valor mayor a 13,75 es 16 y lo pintamos de anaranjado.

Aplicamos la fórmula reemplazando sus valores:

$$Q_1 = 8 + 7\left(\frac{13{,}75 - 9}{7}\right) = 12{,}7$$

Respuesta: El 25% de empleados tiene 12,7 años de servicio o menos.

Cuartil 2:

$$Clase\ cuartil = \frac{2\ (55)}{4} = 27{,}5$$

El primer valor mayor a 27,5 es 28 y lo pintamos de azul.

Aplicamos la fórmula reemplazando sus valores:

$$Q_2 = 15 + 7\left(\frac{27{,}5 - 16}{12}\right) = 21{,}7$$

Respuesta: El 50% de empleados tiene 21,7 años de servicio o menos.

Cuartil 3:

$$Clase\ cuartil = \frac{3\ (55)}{4} = 41{,}25$$

El primer valor mayor a 41,25 es 44 y lo pintamos de amarillo.

Aplicamos la fórmula reemplazando sus valores:

$$Q_3 = 22 + 7\left(\frac{41{,}25 - 28}{16}\right) = 27{,}8$$

Respuesta: El 75% de empleados tiene 27,8 años de servicio o menos.

2. De la siguiente tabla que muestra los salarios (en dólares) de 100 trabajadores en medio mes, calcula el cuartil 1, el cuartil 2 y el cuartil 3.

Solución:

DECILES

Los Deciles son los nueve valores de la variable que dividen a un conjunto de datos ordenados en 10 partes iguales (de 10% cada parte). De manera que para resolver un problema sobre deciles solamente tenemos que hallar D1; D2 ; D3 ; D4 ; ... D9

Representación de los deciles

Para hallar los Deciles, se sigue igual procedimiento que los cuartiles.

DECILES PARA DATOS AGRUPADOS

Para hallar los Deciles (D) para datos agrupados se aplica la siguiente fórmula:

$$D_k = L_i + a_i \cdot \frac{k \cdot \frac{n}{10} - F_{i-1}}{f_i}$$

Donde:

Dk : Decil

Li: Límite inferior del intervalo seleccionado.

k : Debe ser 1 ; 2 ; 3 ; 4; ... ; 9

n: Número total de datos

f : frecuencia absoluta del intervalo seleccionado.

Fi-1 : Frecuencia absoluta Acumulada (pero anterior a la clase decil)

a : Amplitud del intervalo (Restar los 2 valores: L sup - L inf)

Pasos a seguir:

1. Completar la tabla "llenando" la Frecuencia Absoluta Acumulada "F".

2. Encontrar la Clase Decil:

$$Clase\ Decil = \frac{k.n}{10}$$

Luego en la Columna "F" escoger el primer valor mayor que la clase Decil encontrada.

3. Aplicar la fórmula.

EJEMPLOS:

I. La tabla muestra el peso (en Kg) de los estudiantes de la I.E. "J. M. ARGUEDAS", Calcula e interpreta los cuatro primeros Deciles:

Peso (Kg)	f_i
50 – 60	8
60 – 70	10
70 – 80	16
80 – 90	14
90 – 100	10
100 – 110	5
110 – 120	2

Solución:

Completamos la tabla hallando F:

Peso (Kg)	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
TOTAL	n=65	

Encontramos las clase Decil y seleccionamos el intervalo del cual tomaremos los datos.

Decil 1:

$$Clase\ Decil = \frac{1\ (65)}{10} = 6{,}5$$

El primer valor mayor a 6,5 es 8 y seleccionamos el intervalo (50-60).

$$D_1 = 50 + 10\left(\frac{6{,}5 - 0}{8}\right) = \mathbf{58,16}$$

Respuesta: El 10% de los estudiantes tiene 58,16 Kg de peso o menos.

Decil 2:

$$Clase\ Decil = \frac{2\ (65)}{10} = 13$$

El primer valor mayor a 13 es 18 y seleccionamos el intervalo (60-70).

$$D_2 = 60 + 10\left(\frac{13 - 8}{10}\right) = \mathbf{65}$$

Respuesta: El 20% de los estudiantes tiene 65 Kg de peso o menos.

Decil 3:

$$Clase\ Decil = \frac{3\ (65)}{10} = 19{,}5$$

El primer valor mayor a 19,5 es 34 y seleccionamos el intervalo (70-80).

$$D_3 = 70 + 10\left(\frac{19{,}5 - 18}{16}\right) = \mathbf{70,94}$$

Respuesta: El 30% de los estudiantes tiene 570,94 Kg de peso o menos.

Decil 4:

$$Clase\ Decil = \frac{4\ (65)}{10} = 26$$

El primer valor mayor a 26 es 34 y seleccionamos el intervalo (70-80).

$$D_4 = 70 + 10\left(\frac{26 - 18}{16}\right) = \mathbf{75}$$

Respuesta: El 40% de los estudiantes tiene 75 Kg de peso o menos.

2. Problema sobre Deciles.

PERCENTILES

Los Percentiles son los 99 valores de la variable que dividen a un conjunto de datos ordenados en 100 partes iguales (de 1% cada parte). De manera que para resolver un problema sobre percentiles solamente tenemos que hallar P1; P2 ; P3 ; P4; ... ; P99

NOTA: Para hallar los Percentiles se sigue el mismo procedimiento que los Cuartiles.

PERCENTILES PARA DATOS AGRUPADOS

Para hallar los Percentiles (P) para datos agrupados se aplica la siguiente fórmula:

$$P_k = L_i + a_i \cdot \frac{k \cdot \frac{n}{100} - F_{i-1}}{f_i}$$

Donde:

Pk : Percentil

Li: Límite inferior del intervalo seleccionado.

k : Debe ser 1 ; 2 ; 3 ; 4; 5 ; ... ; 99

n: Número total de datos

f : frecuencia absoluta del intervalo seleccionado.

Fi-1 : Frecuencia absoluta Acumulada (pero anterior a la Clase Percentil)

a : Amplitud del intervalo (Restar los 2 valores: L sup - L inf)

Pasos a seguir:

1. Completar la tabla "llenando" la Frecuencia Absoluta Acumulada "F".

2. Encontrar la Clase Percentil:

$$Clase\ Percentil = \frac{k.n}{100}$$

Luego en la Columna "F" escoger el primer valor mayor que la clase Percentil encontrada.

3. Aplicar la fórmula-

EJEMPLOS:

1. La tabla muestra el consumo semanal de fruta de los pacientes de un hospital, Calcula e interpreta el Percentil 60 y 90.

Consumo de frutas	f_i
0 – 1,5	8
1,5 – 3,0	10
3,0 – 4,5	16
4,5 – 6,0	14
6,0 – 7,5	10

Solución:

Completamos la tabla hallando F:

Consumo de frutas	f_i	F_i
0 – 1,5	83	**15**
1,5 – 3,0	10	**41**
3,0 – 4,5	16	**61**
4,5 – 6,0	14	**74**
6,0 – 7,5	10	**80**
TOTAL	**n=80**	

Encontramos las Clase Percentil y seleccionamos el intervalo del cual tomaremos los datos.

Percentil 60:

$$Clase\ Percentil = \frac{\mathbf{90}\ (80)}{100} = 72$$

El primer valor "F" mayor a 72 es 74 y seleccionamos el intervalo (4,5-6,0).

$$P_{90} = 4{,}5 + 1{,}5\left(\frac{72 - 61}{13}\right) = \mathbf{5,8}$$

$$Clase\ Percentil = \frac{\mathbf{60}\ (80)}{100} = 48$$

El primer valor "F" mayor a 48 es 61 y seleccionamos el intervalo (3,0-4,5).

$$P_{60} = 3 + 1{,}5\left(\frac{48 - 41}{20}\right) = \mathbf{3,5}$$

Respuesta: Semanalmente el 60% de los pacientes consume como máximo 3.5 Kg de fruta y el 40% restante consume más de 3,5 Kg de fruta.

UNIDAD II

Medidas de tendencia central para datos agrupados

2.- Introducción a la media, mediana moda

Media

La media es el valor promedio de un conjunto de datos numéricos, calculada como la suma del conjunto de valores dividida entre el número total de valores. La media, a diferencia de la esperanza matemática, es un término matemático.

Mediana

La mediana es un estadístico de posición central que parte la distribución en dos, es decir, deja la misma cantidad de valores a un lado que a otro. Para calcular la mediana es importante que

los datos estén ordenados de mayor a menor, o al contrario de menor a mayor. Esto es, que tengan un orden

Moda

La moda es el valor que tiene mayor frecuencia absoluta. Se puede hallar la moda para variables cualitativas y cuantitativas. 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.

Varianza

En términos de estadística descriptiva, la varianza puede ser definida como la media de los cuadrados de las desviaciones sobre la media. A partir de esta definición, nos puede surgir la duda de por qué calculamos una media de cuadrados de las desviaciones y no de las desviaciones en sí

Desviación estándar

La desviación estándar es la medida de dispersión más común, que indica qué tan dispersos están los datos con respecto a la media. Mientras mayor sea la desviación estándar, mayor será la dispersión de los datos.

El símbolo σ (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que s se utiliza para representar la desviación estándar de una muestra. La variación que es aleatoria o natural de un proceso se conoce comúnmente como ruido.

La desviación estándar se puede utilizar para establecer un valor de referencia para estimar la variación general de un proceso.

2.1.- Media

Se conocen como Medidas de Tendencia Central y para esta explicación vamos a retomar el ejemplo que utilizamos para la elaboración de la tabla de *Distribución de Frecuencias para Datos Agrupados*

Vamos directo al punto con el ejemplo: Se consultó a 50 personas sobre su edad y estos fueron los resultados que representamos en una tabla de frecuencias para datos agrupados.

Datos Agrupados - Distribución de Frecuencias

Edades de 50 personas: 38 - 15 - 10 - 12 - 62 - 46 - 25 - 56 - 27 - 24 - 23 - 21 - 20 - 25 - 38 - 27 - 48 - 35 - 50 - 65 - 59 - 58 - 47 - 42 - 37 - 35 - 32 - 40 - 28 - 14 - 12 - 24 - 66 - 73 - 72 - 70 - 68 - 65 - 54 - 48 - 34 - 33 - 21 - 19 - 61 - 59 - 47 - 46 - 30 - 30

Intervalos $n = 50$
- $= \sqrt{n}$
- $= 1 + 3.322\ \text{Log}(n)$

Valor máximo: 73 años
Valor mínimo: 10 años

Intervalos $= \sqrt{50} = 7.07 \sim 7$

Rango $= 73 - 10 = 63$ años

Amplitud $= R \div I = 63 \div 7 = 9$

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		Frecuencia relativa acumulada (F_r)	
[10 - 19)	14.5	5	5	0.1	10%	0.1	10%
[19 - 28)	23.5	11	16	0.22	22%	0.32	32%
[28 - 37)	32.5	8	24	0.16	16%	0.48	48%
[37 - 46)	41.5	5	29	0.1	10%	0.58	58%
[46 - 55)	50.5	8	37	0.16	16%	0.74	74%
[55 - 64)	59.5	6	43	0.12	12%	0.86	86%
[64 - 73]	68.5	7	50	0.14	14%	1	100%
	Total	50	**Total**	1	100%		

Media Aritmética para Datos Agrupados

La media aritmética también se conoce como PROMEDIO, y básicamente se calcula como la suma de todos los datos dividida entre el número total de datos. Pero esto aplica para datos sueltos… es decir… NO AGRUPADOS…

Para los datos agrupados debemos considerar con un valor REPRESENTATIVO de cada intervalo que se denomina MARCA DE CLASE y asumir que TODAS las cantidades de la frecuencia absoluta se ven representadas por ese valor.

Analicemos el primer intervalo de nuestro ejemplo: Debemos asumir que esas 5 personas tienen 14.5 años

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)
[10 - 19)	14.5	5
[19 - 28)	23.5	11
[28 - 37)	32.5	8
[37 - 46)	41.5	5
[46 - 55)	50.5	8
[55 - 64)	59.5	6
[64 - 73]	68.5	7
	Total	50

Analicemos el segundo intervalo de nuestro ejemplo: Debemos asumir que esas 11 personas tienen 23.5 años

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)
[10 - 19)	14.5	5
[19 - 28)	23.5	11
[28 - 37)	32.5	8
[37 - 46)	41.5	5
[46 - 55)	50.5	8
[55 - 64)	59.5	6
[64 - 73]	68.5	7
	Total	50

Y así para todos los intervalos de la tabla.

La formula para calcular la media aritmética en datos agrupados es la siguiente:

$$\bar{x} = \sum \frac{x_i \cdot f_i}{n}$$

La media se calcula sumando todos los datos y dividiendo entre el total de ellos. Pero para datos agrupados asumimos que por ejemplo en el primer intervalo esas 5 personas todas tienen 14.5 años… entonces queda más práctico multiplicar 5×14.5 o lo que es lo mismo 14.5+14.5+14.5+14.5+14.5.

Vamos a realizar ese mismo procedimiento para cada intervalo, multiplicar marca de clase (x_i) por frecuencia absoluta (f_i) y colocamos el resultado en una nueva columna a la derecha:

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		Frecuencia relativa acumulada (F_r)		Marca de Clase por frecuencia absoluta $(x_i \cdot f_i)$
[10 - 19)	14.5	5	5	0.1	10%	0.1	10%	72.5
[19 - 28)	23.5	11	16	0.22	22%	0.32	32%	258.5
[28 - 37)	32.5	8	24	0.16	16%	0.48	48%	260
[37 - 46)	41.5	5	29	0.1	10%	0.58	58%	207.5
[46 - 55)	50.5	8	37	0.16	16%	0.74	74%	404
[55 - 64)	59.5	6	43	0.12	12%	0.86	86%	357
[64 - 73]	68.5	7	50	0.14	14%	1	100%	479.5
	Total	50	**Total**	1	100%			

El siguiente paso será sumar todos los datos de esa columna porque la fórmula indica que es una sumatoria con el símbolo Σ. Por lo tanto añadimos una fila con el total correspondiente:

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		Frecuencia relativa acumulada (F_r)		Marca de Clase por frecuencia absoluta $(x_i \cdot f_i)$
[10 - 19)	14.5	5	5	0.1	10%	0.1	10%	72.5
[19 - 28)	23.5	11	16	0.22	22%	0.32	32%	258.5
[28 - 37)	32.5	8	24	0.16	16%	0.48	48%	260
[37 - 46)	41.5	5	29	0.1	10%	0.58	58%	207.5
[46 - 55)	50.5	8	37	0.16	16%	0.74	74%	404
[55 - 64)	59.5	6	43	0.12	12%	0.86	86%	357
[64 - 73]	68.5	7	50	0.14	14%	1	100%	479.5
	Total	50	**Total**	1	100%	**Total**		2039

Ya tenemos la sumatoria de todas las multiplicaciones entre marca de clase y frecuencia absoluta. Reemplacemos en nuestra fórmula, recordando que n corresponde al total de datos (en este caso fueron 50 datos, por lo tanto n = 50)

$$\bar{x} = \sum \frac{x_i \cdot f_i}{n}$$

$$\bar{x} = \frac{2039}{50}$$

$$\bar{x} = 40.78 \text{ años}$$

Eso es todo, sólo debes multiplicar cada marca de clase (x_i) por su frecuencia absoluta (f_i), luego sumar todos esos resultados… y por último dividir entre el total de datos.

En este caso ya podemos afirmar que de las 50 personas encuestadas, el promedio de edad es de 40.78 años.

2.2.- Mediana

Mediana para Datos Agrupados

De nuestro ejemplo sabemos que las 50 personas se mueven en un rango de edad que va desde 10 años el más joven y hasta 73 años el más adulto.

La mediana sería esa edad hasta la cual acumulo el 50% de las personas y después de la cuál tengo el otro 50%.

Entonces, desde los 10 años hasta la Mediana hay 25 personas…. y desde la Mediana hasta los 73 años están las otras 25 personas…

La Mediana (M_e) la calculamos con la siguiente fórmula:

$$M_e = L_i + \left(\frac{\frac{N}{2} - F_{i-1}}{f_i} \right) \cdot A$$

L_i es el límite inferior del intervalo de la mediana.

f_i es la frecuencia absoluta del intervalo de la mediana.

F_{i-1} es la frecuencia absoluta acumulada anterior al intervalo de la mediana.

N es el número total de datos del ejercicio, en este caso vale 50.

A es la amplitud de los intervalos y en este caso vale 9 años.

Vamos a identificar el intervalo de la mediana para poder obtener los datos que necesitamos.

La idea es partir mitad y mitad la cantidad de personas en un valor… lo primero es obtener esa mitad:

$$\frac{N}{2} = \frac{50}{2} = 25$$

Vamos a apoyarnos en la columna de frecuencias absolutas acumuladas para descubrir en cuál intervalo tenemos metida a la persona número 25

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		Frecuencia relativa acumulada (F_r)	
[10 - 19)	14.5	5	5	0.1	10%	0.1	10%
[19 - 28)	23.5	11	16	0.22	22%	0.32	32%
[28 - 37)	32.5	8	24	0.16	16%	0.48	48%
[37 - 46)	41.5	5	29	0.1	10%	0.58	58%
[46 - 55)	50.5	8	37	0.16	16%	0.74	74%
[55 - 64)	59.5	6	43	0.12	12%	0.86	86%
[64 - 73]	68.5	7	50	0.14	14%	1	100%
	Total	50	**Total**	1	100%		

En el tercer intervalo teníamos hasta la persona número 24, en cambio en el cuarto intervalo tenemos a las personas 25, 26, 27, 28 y 29, por lo tanto ese es el que nos sirve.

Identificamos datos y reemplazamos en la fórmula:

$$M_e = L_i + \left(\frac{\frac{N}{2} - F_{i-1}}{f_i}\right) \cdot A$$

$N = 50$	$L_i = 37$	$A = 9$

$f_i = 5$	$F_{i-1} = 24$

$$M_e = 37 + \left(\frac{\frac{50}{2} - 24}{5}\right) \cdot 9$$

$$M_e = 37 + \left(\frac{25 - 24}{5}\right) \cdot 9$$

$$M_e = 37 + \left(\frac{1}{5}\right) \cdot 9$$

$$M_e = 37 + 1.8$$

$$M_e = 38.8 \textbf{ años}$$

Esto significa que desde los 10 años hasta los 38.8 años hay 25 personas…. y desde los 38.8 años hasta los 73 años están las otras 25 personas…

2.3.- Moda

Moda para Datos Agrupados

Su mismo nombre lo indica… ¿Cuál es la tendencia? ¿Cuál edad estará de moda en nuestro ejemplo?

Si fuesen datos NO AGRUPADOS, fácilmente diríamos que la moda es el dato que más se repite sin realizar ningún cálculo ni operación matemática.

Pero como nuestro interés es calcular la moda para datos agrupados… debemos utilizar la siguiente fórmula:

$$M_o = L_i + \left(\frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \right) \cdot A$$

La moda se simboliza como M_o y nuestro primer paso será identificar el intervalo modal.

Es muy sencillo, el intervalo modal corresponde a aquel que posee la frecuencia absoluta más alta.

Para nuestro ejemplo el modal sería el segundo intervalo ya que tiene frecuencia absoluta de 11

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		Frecuencia relativa acumulada (F_r)	
[10 - 19)	14.5	5	5	0.1	10%	0.1	10%
[19 - 28)	23.5	11	16	0.22	22%	0.32	32%
[28 - 37)	32.5	8	24	0.16	16%	0.48	48%
[37 - 46)	41.5	5	29	0.1	10%	0.58	58%
[46 - 55)	50.5	8	37	0.16	16%	0.74	74%
[55 - 64)	59.5	6	43	0.12	12%	0.86	86%
[64 - 73]	68.5	7	50	0.14	14%	1	100%
	Total	50	**Total**	1	100%		

Teniendo identificado el intervalo modal, vamos a analizar cada término de la fórmula para calcular la moda

$$M_o = L_i + \left(\frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \right) \cdot A$$

L_i es el límite inferior del intervalo modal, en este caso vale 19.

f_i es la frecuencia absoluta del intervalo modal, en este caso vale 11.

f_{i-1} es la frecuencia absoluta anterior al intervalo modal, en este caso vale 5.

f_{i+1} es la frecuencia absoluta siguiente al intervalo modal, en este caso vale 8.

A es la amplitud del intervalo modal, en este caso vale 9 porque el intervalo va de 19 a 28 años… es decir hay una distancia de 9 años allí.

Por si no te quedó claro lo de la frecuencia absoluta anterior y siguiente, así se identifican:

Frecuencia absoluta (f_i)
5
11
8
5
8
6
7

Listo, ahora reemplacemos los datos en la fórmula y calculemos la edad de moda

$$M_o = L_i + \left(\frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})}\right) \cdot A$$

$L_i = 19$	$f_i = 11$	$A = 9$	$f_{i-1} = 5$	$f_{i+1} = 8$

$$M_o = 19 + \left(\frac{11 - 5}{(11 - 5) + (11 - 8)}\right) \cdot 9$$

$$M_o = 19 + \left(\frac{6}{(6) + (3)}\right) \cdot 9$$

$$M_o = 19 + \left(\frac{6}{9}\right) \cdot 9 = 19 + 6 = 25 \textbf{ años}$$

Todo parece indicar que para nuestro ejemplo, está de moda tener 25 años.

2.4.- varianza y desviación estándar

Varianza y desviación estándar para datos agrupados por intervalos

Veamos como calcular la varianza y la desviación estándar a partir de una tabla de frecuencias con datos agrupados por intervalos, para la población y para la muestra.

Si necesitamos calcular la varianza y la desviación estándar de un conjunto de datos agrupados por intervalos en un tabla de frecuencias, usaremos las fórmulas que revisaremos en esta clase.

Fórmulas para la varianza y desviación estándar de datos agrupados

Varianza y desviación estándar para datos agrupados

	Varianza	Desviación estándar	Media	Número de elementos
Población	$\sigma^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \mu)^2}{N}$	$\sigma = \sqrt{\sigma^2}$	$\mu = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N}$	$N = \sum_{i=1}^{k} f_i$
Muestra	$s^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{n-1}$	$s = \sqrt{s}$	$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n}$	$n = \sum_{i=1}^{k} f_i$

Donde:

k: número de clases.

f_i: frecuencia absoluta de cada clase, es decir, el número de elementos que pertenecen a dicha clase.

x_i: marca de clase. Es el punto medio del límite inferior y del límite superior.

σ^2: varianza de la población.

σ: desviación estándar de la población.

μ: media de la población.

s^2: varianza de la muestra.

s: desviación estándar de la muestra.

$\bar{x}$: media de la muestra.

Tenemos siempre que fijarnos si estamos trabajando con datos que forman una población o con datos que forman una muestra, pues las fórmulas son diferentes.

En los problemas, seguiremos los siguientes pasos:

Calculamos el número de elementos.

Calculamos las marcas de clase.

Calculamos la media.

Calculamos la varianza.

Calculamos la desviación estándar, que es la raíz cuadrada de la varianza.

Ejemplo 1:

Calcular la varianza y la desviación estándar de una población de niños a partir de la siguiente tabla:

Edad (años)	Frecuencia f_i
[0 - 2)	7
[2 - 4)	8
[4 - 6)	8
[6 - 8]	7

Solución:

En este caso, nos dicen que los datos pertenecen a una población de niños, por lo tanto, usaremos las fórmulas de la población.

Primero calculamos el número de elementos de la población N:

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

Con ayuda de la tabla, calculamos la suma de las frecuencias f_i.

Edad (años)	Frecuencia f_i
[0 - 2)	7
[2 - 4)	8
[4 - 6)	8
[6 - 8]	7
$\sum$	30

Ahora sí, calculamos N.

$$N = \sum_{i=1}^{k} f_i = 30$$

Como segundo paso, calcularemos las marcas de clase. Recordemos que la marca de clase x_i, es el punto medio del límite inferior y el límite superior de cada intervalo. Se calcula con la siguiente fórmula:

$$x_i = \frac{L_i + L_s}{2}$$

Agregamos una columna más a nuestra tabla para la marca de clase x_i:

Edad (años)	Marca de clase x_i	Frecuencia f_i
[0 - 2)	1	7
[2 - 4)	3	8
[4 - 6)	5	8
[6 - 8]	7	7
	Σ	30

Como tercer paso, calculamos la media poblacional μ:

$$\mu = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N}$$

Agregamos una columna más a nuestra tabla, dónde colocaremos los valores de $x_i \cdot f_i$:

Edad (años)	Marca de clase x_i	Frecuencia f_i	$x_i \cdot f_i$
[0 - 2)	1	7	7
[2 - 4)	3	8	24
[4 - 6)	5	8	40
[6 - 8]	7	7	49
	Σ	30	120

Aplicamos la fórmula:

$$\mu = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N} = \frac{120}{30} = 4\ años$$

La media poblacional μ tiene un valor de 4 años.

Como cuarto paso, calculamos la varianza de la población:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^2}{N}$$

Agregamos más columnas a nuestra tabla, buscando la forma de la fórmula de la varianza:

Edad (años)	Marca de clase x_i	Frecuencia f_i	$x_i \cdot f_i$	$x_i - \mu$	$(x_i - \mu)^2$	$f_i(x_i - \mu)^2$
[0 - 2)	1	7	7	-3	9	63
[2 - 4)	3	8	24	-1	1	8
[4 - 6)	5	8	40	1	1	8
[6 - 8]	7	7	49	3	9	63
	Σ	30	120			142

Aplicamos la fórmula de la varianza de la población:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^2}{N} = \frac{142}{30} = 4{,}73\ (años)^2$$

Recuerda que la varianza queda expresada en unidades al cuadrado, por ello, nos queda en años al cuadrado.

Como último paso, calculamos la desviación estándar, recordando que es la raíz cuadrada positiva de la varianza.

$$\sigma = \sqrt{\sigma^2} = \sqrt{4{,}73\ (años)^2}$$

$$\sigma = 2{,}175\ años$$

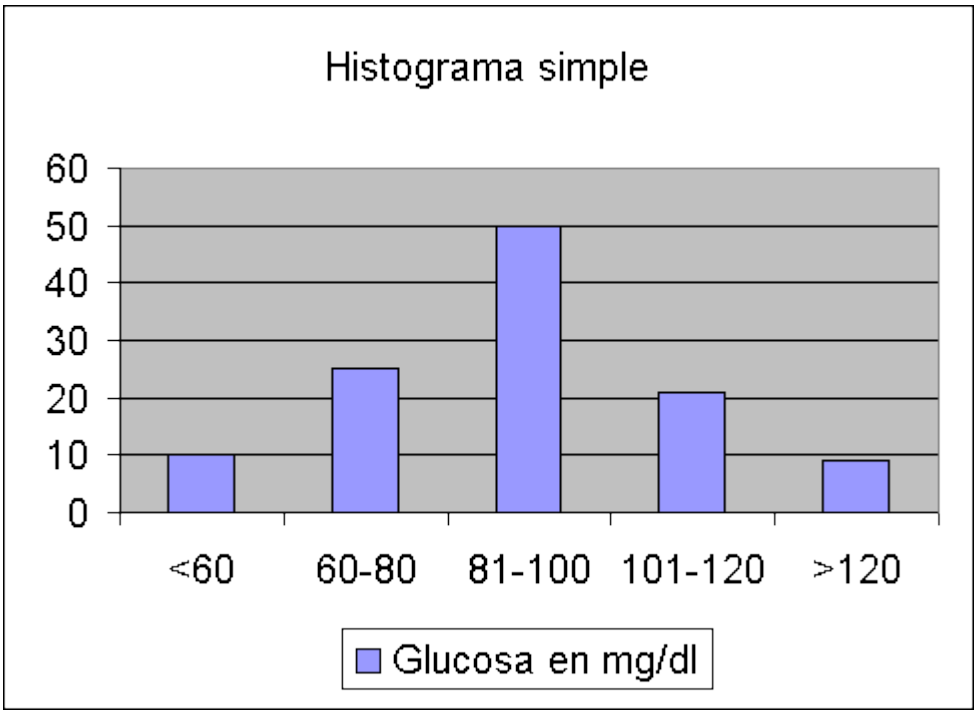
El valor de la desviación estándar poblacional σ es de 2,175 años.

2.5.- Graficas para representar datos agrupados

A continuación, se presentan algunos gráficos para representar los resultados de los datos agrupados. Cabe mencionar que estas graficas son algunas de las más usadas, aunque existen muchas más que se pueden ocupar para este tipo de problemas.

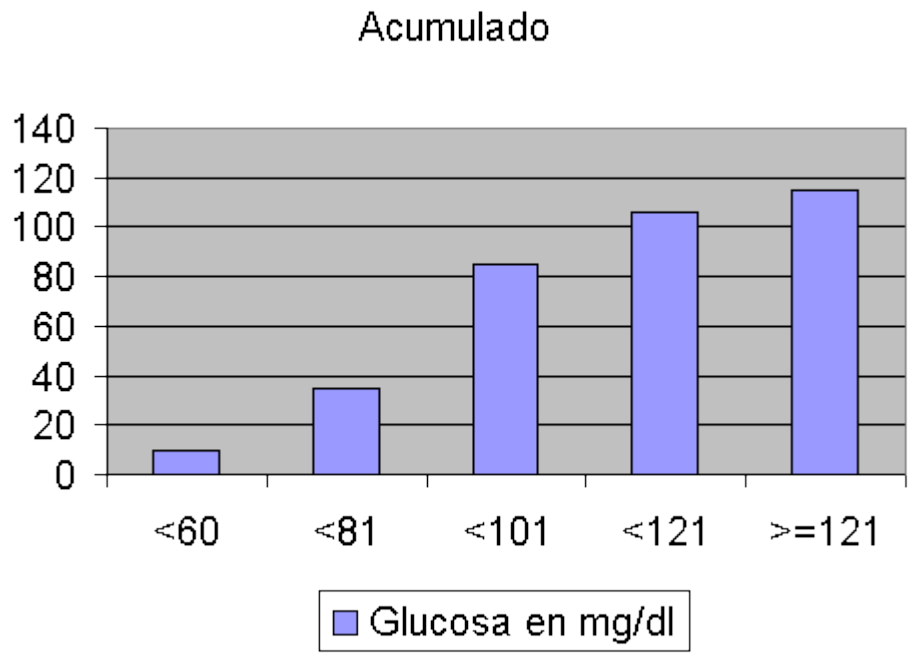
Ejemplos de tipos de representaciones gráficas

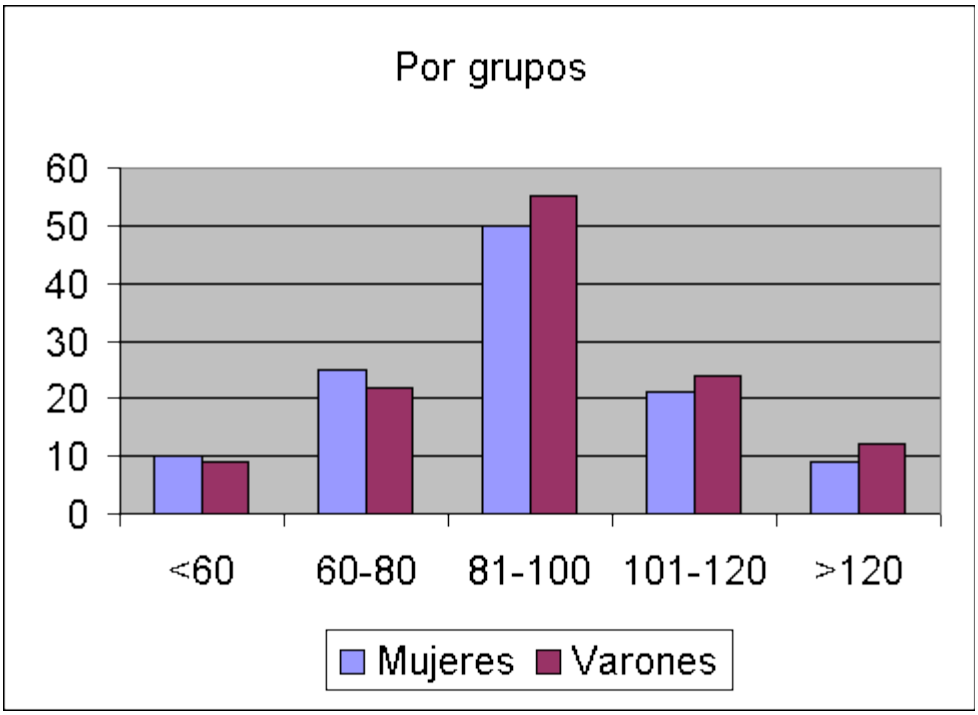
Histogramas: Se agrupan los datos en *clases*, y se cuenta cuántas observaciones (*frecuencia absoluta*) hay en cada una de ellas. En algunas variables (*variables cualitativas*) las clases están definidas de modo natural, p.e *sexo* con dos clases: mujer, varón o *grupo sanguíneo* con cuatro: A, B, AB, O. En las variables cuantitativas, las clases hay que definirlas explícitamente (*intervalos de clase*).



Se representan los intervalos de clase en el eje de abcisas (eje horizontal) y las frecuencias, absolutas o relativas, en el de ordenadas (eje vertical).

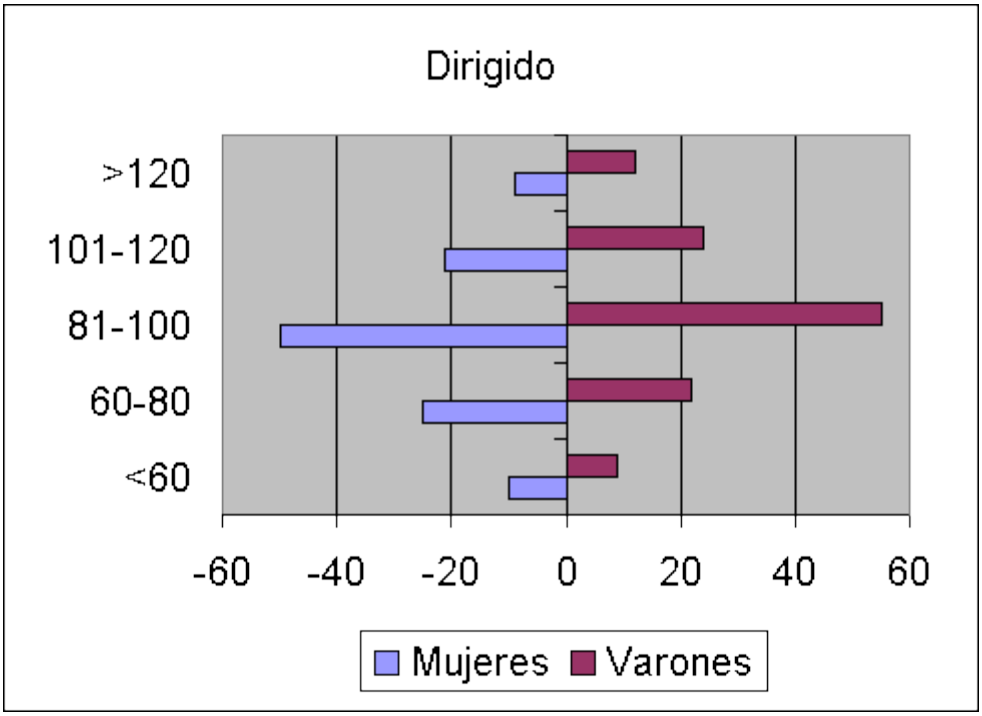
A veces es más útil representar las frecuencias acumuladas.

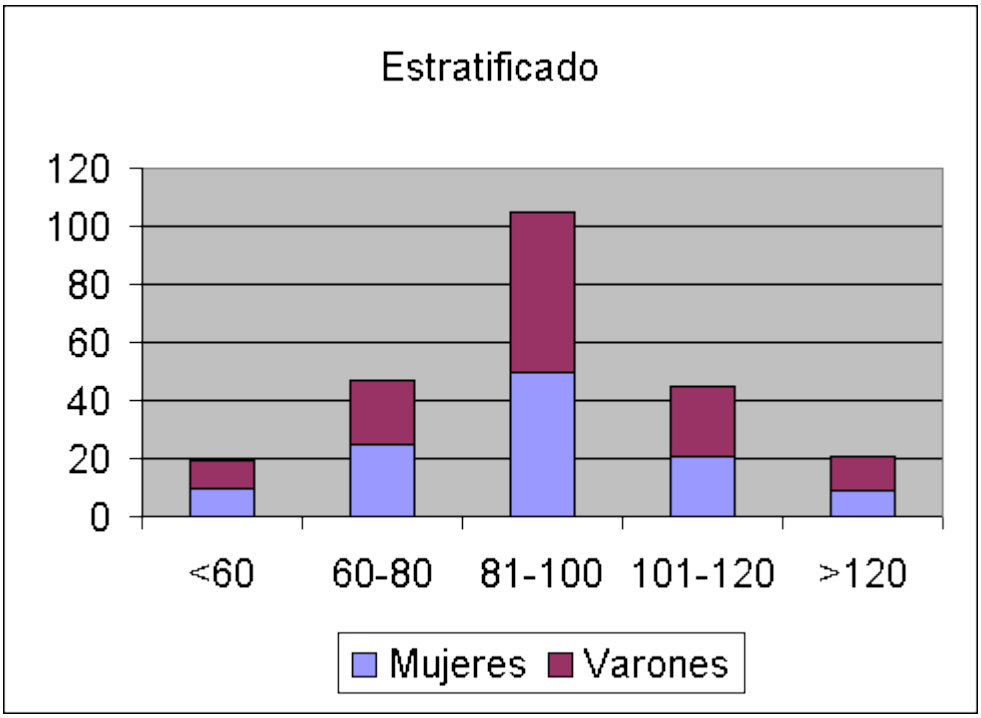




O representar simultáneamente los histogramas de una variable en dos situaciones distintas.

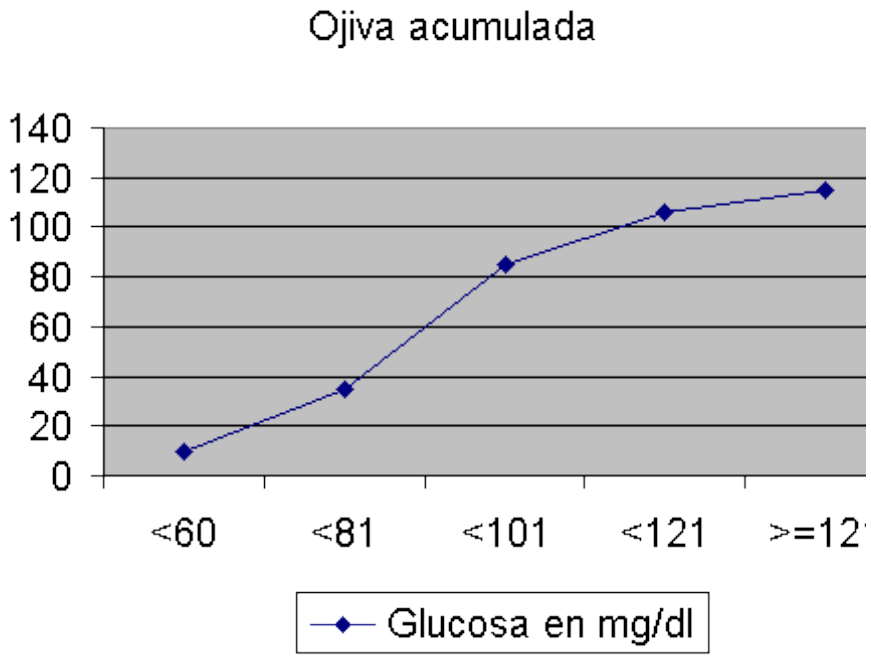
Otra forma muy frecuente, de representar dos histogramas de la misma variable en dos situaciones distintas.



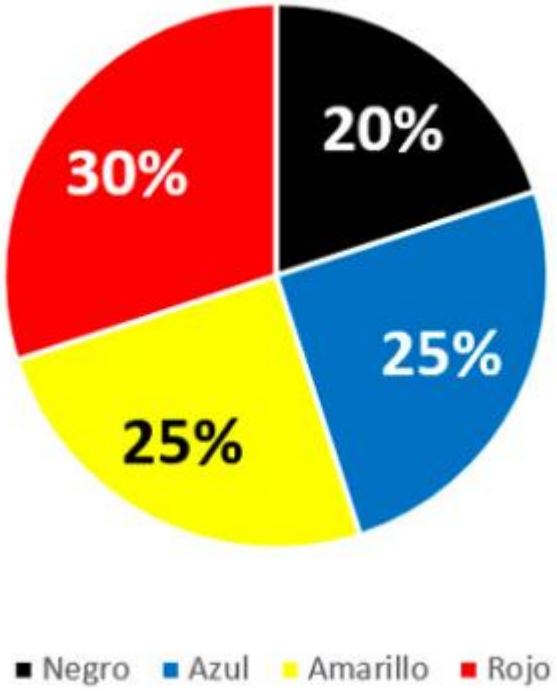


Otra más

En las variables cuantitativas o en las cualitativas ordinales se pueden representar *polígonos* de frecuencia en lugar de histogramas, cuando se representa la frecuencia acumulativa, se denomina *ojiva*.



Grafica de pastel



Grafica ojiva

UNIDAD III

Modelos de pronósticos

3.1.- Importancia de los pronósticos

Existen técnicas estadísticas que nos permiten estimar cuál será el comportamiento de diversos factores que pudieran ser importantes en el desempeño o en el desarrollo de nuestras empresas. Por ejemplo, podemos estimar cuál será el comportamiento de las ventas de nuestra empresa, cuál será el comportamiento de la demanda, del tipo de cambio, de la inflación, del crecimiento del país, del PIB, y prácticamente de cualquier indicador interno de la empresa, o cualquier indicador macroeconómico de nuestro país o de cualquier otro país. Se han llevado a cabo estudios en Estados Unidos y en México, en los cuales se ha comprobado que

prácticamente en todas la carreras profesionales del área de negocios, tienen en su plan curricular alguna materia donde se les enseña a los estudiantes los conceptos básicos con los cuales podrían llevar a cabo estos pronósticos empresariales; pero al momento de ir a checar a las empresas, los investigadores se han dado cuenta que los profesionistas que fueron preparados con dichas técnicas de pronósticos, no hacen prácticamente uso de dichas técnicas. Esto es algo muy triste porque en especial en países en desarrollo como el nuestro, algo que nos hace mucha falta es la utilización de herramientas que ayuden a hacer mejores estimaciones, ya que una buena estimación de ventas o de la demanda, es definitivamente la mejor manera de poder llevar a cabo una adecuada planeación de la producción, de recursos humanos, de las finanzas de la empresa, de entregas, o de compras, entre otros. Aunque he asesorado la aplicación de técnicas de pronósticos en distintas áreas, uno de los que más recuerdo es el de una empresa petroquímica de la región que tenía problemas muy fuertes ya que no podía estimar de manera adecuada la demanda mensual de su producto más importante, por lo que mes a mes iba sufriendo de la imposibilidad de cumplir con la demanda de sus clientes en algunos meses, y sufría de exceso de inventario en otros meses. Después de hacer un análisis estadístico de la demanda en los años previos, se determinaron las herramientas de pronósticos más adecuadas según el caso y se tuvieron resultados sobresalientes al planear la producción, consiguiendo impresionantes reducciones de costos por faltantes y por exceso de inventarios.

3.2.- Tipos de pronósticos

¿Qué es un pronóstico? Características y métodos

Un pronóstico, en el plano empresarial, es la predicción de lo que sucederá con un elemento determinado dentro del marco de un conjunto dado de condiciones. Se diferencia del presupuesto porque este último es el resultado de decisiones encaminadas a generar las condiciones que propiciarán un nivel deseado de dicho elemento.

Por qué se necesitan los pronósticos en la empresa

El objetivo básico de un pronóstico consiste en reducir el rango de incertidumbre dentro del cual se toman las decisiones que afectan el futuro del negocio y con él a todas las partes

involucradas. Aunque, el pronóstico no sustituye el juicio administrativo en la toma de decisiones, simplemente es una ayuda en ese proceso.

Los pronósticos se emplean en el proceso de establecimiento de objetivos tanto de largo como de corto plazo, constituyéndose así en bases para el desarrollo de planes, a nivel general y en las distintas áreas o unidades. Los planes basados en dichos pronósticos, no sólo atenderán a ellos sino que establecerán estrategias y acciones que los puedan contrarrestar, corregir o impulsar.

Por ejemplo, si el pronóstico de ventas para el siguiente ejercicio fiscal muestra una tendencia desfavorable, entonces el plan estratégico de ventas deberá encaminarse a revertir dicha tendencia a través de acciones que impulsen el crecimiento o que no permitan que las ventas decaigan o, el el peor de los casos, que simplemente se reduzcan en un nivel mínimo.

Usos de los pronósticos en la empresa

Todas las organizaciones que planifican las condiciones de su futuro, el cual no conocen a ciencia cierta, emplean pronósticos en sus diferentes áreas funcionales. Algunos casos de uso de pronósticos en la empresa son:

En el área de marketing se pronostica cómo va a crecer el mercado, cuál va a ser la participación propia y de los competidores, cuál será la tendencia de precios, cuáles serán los nuevos productos que sacudirán el mercado…

En el área de producción se hacen pronósticos sobre el costo y la disponibilidad de la materia prima, el costo y la disponibilidad de la mano de obra, cuándo se requerirá mantenimiento para los equipos, cuál será la capacidad de planta necesaria para atender la demanda…

En el área financiera se pronostica cuál será la tasa de interés de referencia para los créditos, cuál será el nivel de cuentas incobrables, cuánto capital se requerirá para ampliar la capacidad propia…

En recursos humanos se requieren pronósticos sobre el número de trabajadores, la rotación de personal, las tendencias de ausentismo, las necesidades de capacitación…

En el plano estratégico se pronostica acerca de factores económicos, cambios de precios, costos, crecimiento de líneas de productos…

Características del pronóstico empresarial

Es consistente con las demás áreas del negocio. Si marketing pronosticó un crecimiento del 25% de unidades vendidas entonces producción y recursos humanos deben estar en capacidad de cumplir.

Se basa en el conocimiento adecuado del pasado relevante. Aunque hay excepciones, la regla es que comportamientos ocurridos en el pasado son fuente de predicción del futuro.

Tiene en cuenta el entorno político y económico. Un cambio en las condiciones de estos factores puede traer consecuencias enormes en cualquier sector económico.

Es oportuno. Ya sea para ganar cuota de mercado introduciendo un nuevo producto o para retirar otro y evitar una crisis, el más preciso de los pronósticos pierde toda su utilidad si se ha dejado pasar la oportunidad correcta de aplicarlo en la planeación.

Clasificación de los modelos de pronósticos

Según el marco de tiempo al que atienden se clasifican en:

De corto plazo. Se usan para diseñar estrategias inmediatas, son empleados entre mandos medios y gerencias de primera línea.

De mediano plazo. Conjunta al corto y al largo plazo, útil para decisiones de todos los niveles.

Pronósticos de largo plazo. Requeridos para establecer el rumbo general de la organización, generalmente se hacen para que la alta dirección los use en los procesos de planeación estratégica.

Según su atención al detalle se clasifican en:

Micropronósticos. Involucran pequeños detalles e interesan a los niveles medios y de primera línea.

Macropronósticos. Se realizan a gran escala y son del interés de la alta dirección.

Según la intensidad del uso de datos se clasifican en:

Pronósticos cualitativos. Se basan en el juicio de individuos o grupos de individuos, se pueden presentar en forma numérica pero generalmente no están basados en series de datos históricos.

Pronósticos cuantitativos. Emplean cantidades significativas de datos previos como base de predicción. Pueden ser:

Simples (no formales): proyectan datos pasados hacia el futuro sin explicar las tendencias futuras.

Causales (explicativos): intentan explicar las relaciones funcionales entre la variable a ser estimada (variable dependiente) y la variable o variables que explican los cambios (variables independientes).

Métodos cualitativos

Las técnicas cualitativas se usan cuando los datos son escasos, por ejemplo cuando se introduce un producto nuevo al mercado. Estas técnicas usan el criterio de la persona y ciertas relaciones para transformar información cualitativa en estimados cuantitativos. Algunos son:

Jurado de opinión ejecutiva. Un grupo de ejecutivos corporativos se reúnen, sus opiniones se promedian para generar el pronóstico.

Composición de la fuerza de ventas. Combina estimaciones de los vendedores sobre las compras esperadas de los clientes.

Método Delphi. Empleada predominantemente en la predicción de tendencias y cambios tecnológicos. Emplea un panel de expertos que no se reúnen sino que el proceso se lleva a cabo mediante una serie secuencial de preguntas y respuestas escritas.

Encuestas de opinión. Permiten identificar cambios en las tendencias, se llevan a cabo en muestras de la población.

Investigación de mercado. Se usa para evaluar y probar hipótesis acerca de mercados reales.

Evaluación de clientes. Combina estimaciones de los clientes habituales.

En el siguiente video, de la UPV, se presentan las características de cuatro de los métodos cualitativos más empleados en la elaboración de pronósticos.

Métodos cuantitativos

Se basan en procedimientos mecánicos o modelos matemáticos que se apoyan en datos históricos o en variables causales para producir resultados cuantitativos. Algunos son:

Análisis de series temporales. Establece una ecuación para una tendencia y la proyecta al futuro

Modelos de regresión. Pronostica una variable a partir de lo que se sabe o supone de otras.

Modelos econométricos. Simula con ecuaciones de regresión segmentos de la economía.

Indicadores económicos. Pronostica con uno o más indicadores el estado futuro de la economís

Efecto de sustitución. Predice con una fórmula matemática cómo, cuándo y en qué circunstancias un nuevo producto o tecnología sustituirá al actual.

A través del siguiente video tutorial (8 videos), de INCAE, podrás aprender más sobre los métodos cuantitativos de pronóstico.

Cómo elegir el método de pronóstico adecuado

La principal consideración para seleccionar un método de pronóstico es que sus resultados deben orientar, de la mejor manera, la toma de decisiones administrativa, de lo contrario, el uso cualquier método, por sofisticado que este sea, no será conveniente. Algunas de las variables a considerar, al momento de seleccionar la técnica o método de pronóstico más adecuada, son:

El contexto del pronóstico

La relevancia y disponibilidad de datos históricos

El grado de exactitud deseado

El periodo de tiempo que se va a pronosticar

El análisis de costo-beneficio del pronóstico

El punto del ciclo de vida en que se encuentra el producto.

A fin de seleccionar adecuadamente la técnica conveniente de pronósticos, el pronosticador debe ser capaz de:

Definir la naturaleza del problema de pronóstico.

Explicar la naturaleza de los datos que se investigan.

Describir las capacidades y limitaciones de técnicas de pronósticos potencialmente útiles.

Desarrollar algunos criterios predeterminados sobre los que se pueda tomar la decisión de selección.

Un factor importante que influye en la selección de una técnica de pronóstico es identificar y entender los patrones históricos de los datos. Si se pueden reconocer patrones de tendencia, cíclicos o estacionales, pueden seleccionarse técnicas capaces de extrapolarlos de manera eficaz.

El proceso del pronóstico

Generalmente un pronóstico se elabora siguiendo los pasos que se indican a continuación:

Formulación del problema y recolección de datos. Estos dos elementos se tratan como un único paso porque el problema determina los datos adecuados. Si no se dispone de los datos adecuados el problema tendría que redefinirse o se tendría que acudir a un método puramente cualitativo.

Manipulación y limpieza de datos. Es posible tener muchos o pocos datos, datos irrelevantes, datos desactualizados, etc., todos ellos requerirán de cierto procesamiento para obtener los datos necesarios y adecuados.

Construcción y evaluación del modelo. Implica emplear los datos en un modelo de pronósticos que sea adecuado en términos de minimización del error de pronóstico.

Aplicación del modelo (el pronóstico real). Consiste en los pronósticos reales del modelo que se generan una vez que se han recolectado y quizás reducido a sólo los datos adecuados, tan pronto se ha elegido un modelo adecuado de pronósticos.

Evaluación del pronóstico. Implica comparar los valores del pronóstico con los valores históricos reales. Frecuentemente, el examen de los patrones de errores lleva al analista a modificar el procedimiento de pronósticos.

3.3.- Pronostico móvil simple

Ejemplo y fórmula promedio móvil (pronóstico)

El promedio móvil, al igual que el último dato, se utiliza para pronosticar series de tiempo estables serie estable, que no presente tendencia ni estacionalidad. Su nombre se debe a que conforme avanza el tiempo, se descarta el dato más antiguo y se considera el más reciente.

Fórmula

$$F_t = \frac{D_{t-1}+D_{t-2}+D_{t-3}+...+D_{t-n}}{n}$$

Notación:

F_t – pronóstico del siguiente periodo t

D_t – valor observado de la demanda en el periodo t

n – número de periodos a considerar en el promedio móvil

El promedio móvil siempre se mantendrá entorno a los datos históricos.

Ejemplo

Vamos a considerar los valores que utilizamos para el ejemplo del último dato:

Semana	Ventas
1	58
2	60
3	44
4	46
5	54
6	52
7	44
8	48
9	52
10	42
11	46
12	43
13	58
14	58
15	53
16	58

Esta serie histórica de ventas no presenta estacionalidad ni tendencia. Se trata de una serie estable y por ello podemos utilizar el promedio móvil.

Promedio móvil con n=2

El promedio móvil más simple que podemos obtener es con dos datos históricos, es decir, n=2. No podemos pronosticar las ventas de la semana 1 ni las ventas de la semana 2 porque no tenemos dos datos anteriores en ninguno de los dos casos. Debemos comenzar con el pronóstico F_3 de la semana 3 utilizando los datos históricos D_1 y D_2. correspondientes a las ventas de la semana 1 y semana 2.

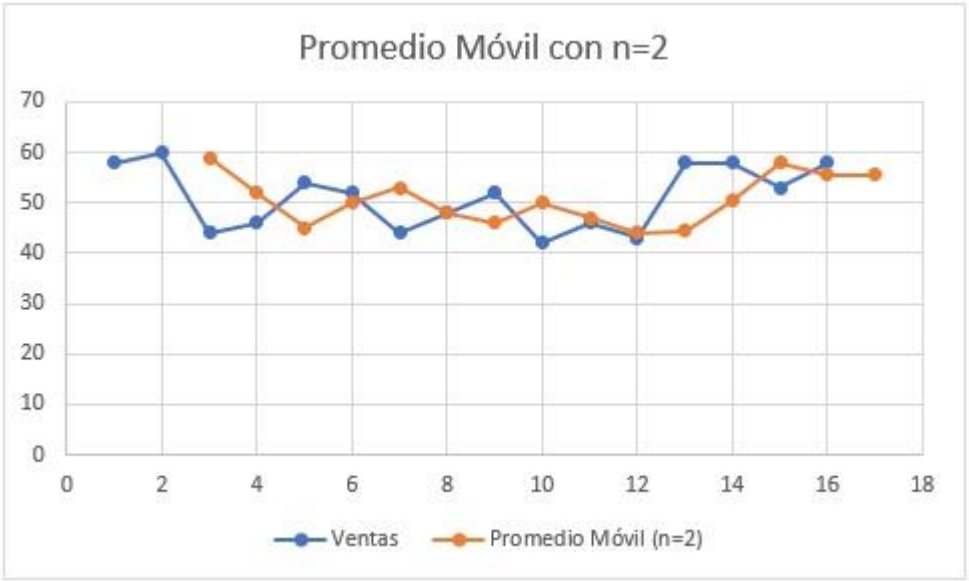
$$F_3 = \frac{D_1+D_2}{2} = 59$$

$$F_4 = \frac{D_2+D_3}{2} = 52$$

$$F_5 = \frac{D_3+D_4}{2} = 45$$

Realizamos estas operaciones hasta calcular el pronóstico de la semana 17:

$$F_{17} = \frac{D_{16}+D_{15}}{2} = 55.5$$



Promedio móvil con n=3

En este caso no podemos pronosticar las ventas de las semanas 1, 2 y 3, ya que necesitamos 3 datos históricos para hacerlo. El primer pronóstico que podemos realizar es el correspondiente a la semana 4.

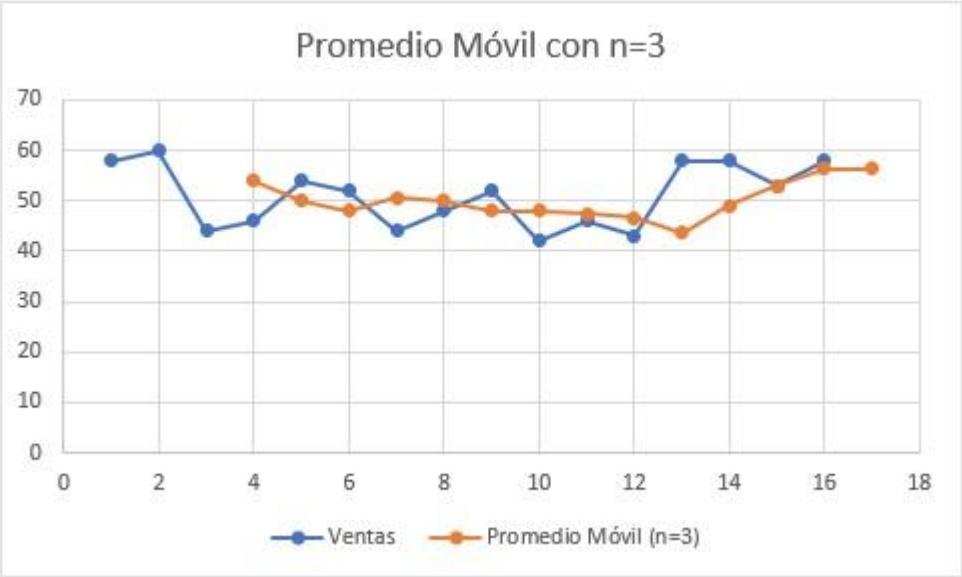
$$F_4 = \frac{D_1+D_2+D_3}{3} = 54$$

$$F_5 = \frac{D_2+D_3+D_4}{3} = 50$$

$$F_6 = \frac{D_3+D_4+D_5}{3} = 48$$

Continuamos hasta calcular el pronóstico de la semana 17:

$$F_{17} = \frac{D_{16}+D_{15}+D_{14}}{3} = 56.33$$



Promedio móvil con n=4

El primer pronóstico que podemos realizar es el correspondiente a la semana 5, ya que necesitamos 4 datos históricos para el pronóstico.

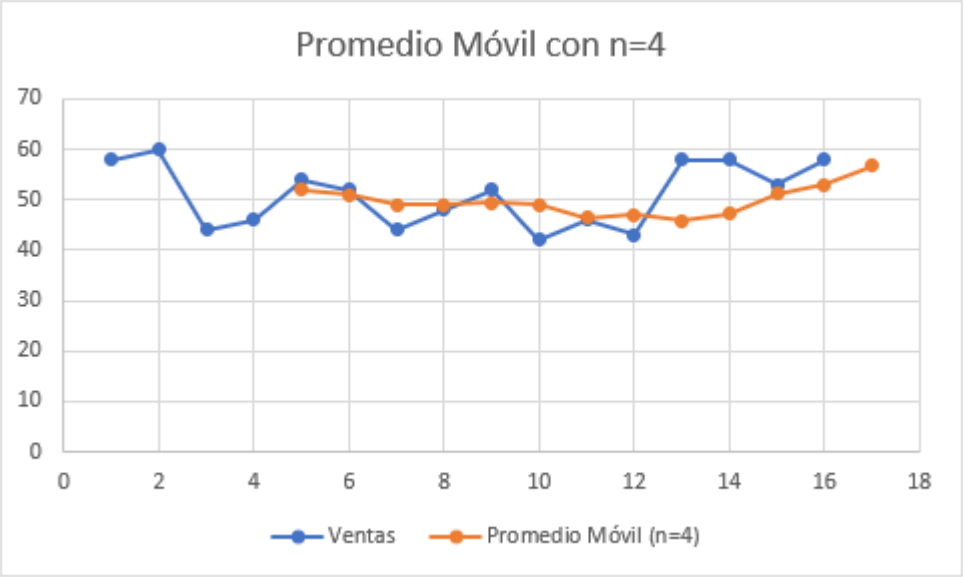
$$F_5 = \frac{D_1+D_2+D_3+D_4}{4} = 52$$

$$F_6 = \frac{D_2+D_3+D_4+D_5}{4} = 51$$

$$F_7 = \frac{D_3+D_4+D_5+D_6}{4} = 49$$

Hasta calcular el pronóstico de la semana 17:

$$F_{17} = \frac{D_{16}+D_{15}+D_{14}+D_{13}}{4} = 56.75$$



Se sigue el mismo procedimiento con cualquier n. La siguiente tabla resume los cálculos para los tres promedios móviles mencionados en el ejemplo.

Semana	Ventas	Promedio Móvil (n=2)	Promedio Móvil (n=3)	Promedio Móvil (n=4)
1	58	-	-	-
2	60	-	-	-
3	44	59.00	-	-
4	46	52.00	54.00	-
5	54	45.00	50.00	52.00
6	52	50.00	48.00	51.00
7	44	53.00	50.67	49.00
8	48	48.00	50.00	49.00
9	52	46.00	48.00	49.50
10	42	50.00	48.00	49.00
11	46	47.00	47.33	46.50
12	43	44.00	46.67	47.00
13	58	44.50	43.67	45.75
14	58	50.50	49.00	47.25
15	53	58.00	53.00	51.25
16	58	55.50	56.33	53.00
17	-	55.50	56.33	56.75

Las medidas de precisión, como es la desviación media absoluta, pueden ayudar a decidir entre los tres promedios móviles. Generalizando, se puede decir que el mejor pronóstico es el que muestra menor desviación respecto a los datos históricos.

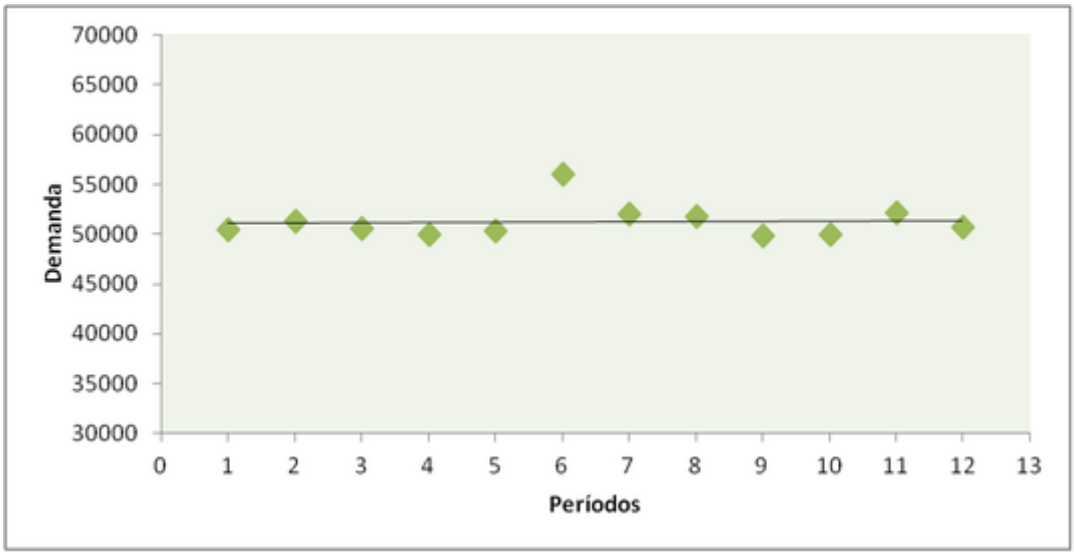
3.4.- Pronostico móvil ponderado

Promedio móvil ponderado

Este método de pronóstico es una variación del *promedio móvil*. Mientras, en el promedio móvil simple se le asigna igual importancia a cada uno de los datos que componen dicho promedio, en el promedio móvil ponderado podemos asignar cualquier importancia (peso) a cualquier dato del promedio (siempre que la sumatoria de las ponderaciones sean equivalentes al 100%). Es una práctica regular aplicar el factor de ponderación (porcentaje) mayor al dato más reciente.

¿Cuándo utilizar un pronóstico de promedio móvil ponderado?

El pronóstico de promedio móvil ponderado es óptimo para patrones de demanda aleatorios o nivelados donde se pretende eliminar el impacto de los elementos irregulares históricos mediante un enfoque en períodos de demanda reciente, dicho enfoque es superior al del promedio móvil simple.



Modelo de Promedio Móvil Ponderado

Fórmula

$$\hat{X}_t = \sum_{t=1}^{n} C_i * X_{t-1}$$

Período	Ventas (unidades)	Ponderación
Mes 1	100000	10%
Mes 2	90000	20%
Mes 3	105000	30%
Mes 4	95000	40%

$\hat{X}\ t$	Promedio de ventas en unidades en el período t
Σ	Sumatoria de datos
C_i	Factor de ponderación
X_{t-1}	Ventas o demandas reales en unidades de los períodos anteriores a t
n	Número de datos

Ejemplo de aplicación de un pronóstico de Promedio Móvil Ponderación

Un almacén ha determinado que el mejor pronóstico se encuentra determinado con 4 datos y utilizando los siguientes factores de ponderación (40%, 30%, 20% y 10%). Determinar el pronóstico para el período 5.

Solución

En este caso el primer paso consiste en multiplicar a cada período por su correspondiente factor de ponderación, luego efectuar la sumatoria de los productos.

$$\hat{X}_t = (100000 * 0{,}1) + (90000 * 0{,}2) + (105000 * 0{,}3) + (95000 * 0{,}4)$$

$$\hat{X}_t = 10000 + 18000 + 31500 + 38000$$

$$\hat{X}_t = 97500\ unidades$$

Ejemplo y fórmula promedio móvil ponderado (pronóstico)

El promedio móvil ponderado descarta el dato histórico más antiguo y considera el más reciente, tal y como se trabaja con los promedios móviles simples, la diferencia es que los datos históricos se ponderan, es decir, se les asignan pesos diferentes.

Otra similitud con el último dato y con los promedios móviles simples es que el promedio móvil ponderado se utiliza para pronosticar valores futuros de series estables, que no presenten tendencia ni estacionalidad.

Fórmula

Notación:

F_t – pronóstico del siguiente periodo t

D_t – valor observado de la demanda en el periodo t

w_i – peso o ponderación para el valor observado de la demanda en el periodo t-i

Desafortunadamente no existe una fórmula para calcular las ponderaciones que se deben asignar a cada datos histórico, pero se debe probar una conjunto de ponderaciones y seleccionar las que nos proporcionan un pronóstico más preciso.

Ejemplo

Consideremos la siguiente tabla de ventas semanales de un cierto producto:

Semana	Ventas
1	58
2	60
3	44
4	46
5	54
6	52
7	44
8	48
9	52
10	42
11	46
12	43
13	58
14	58
15	53
16	58

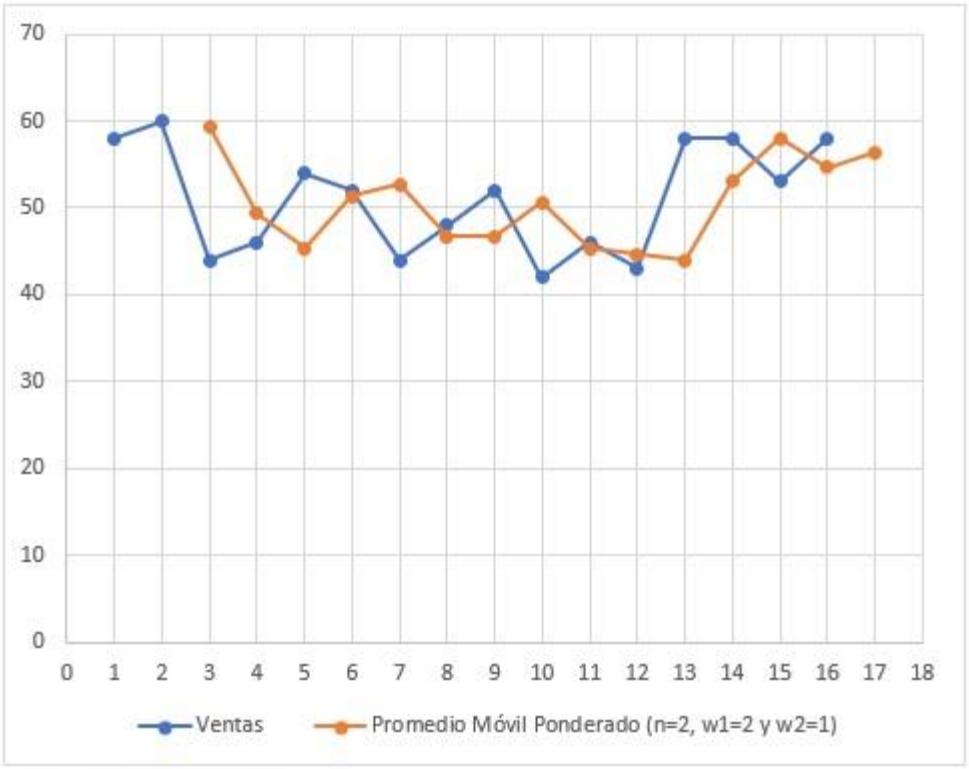
celeberrima.com

En esta serie histórica no se observa patrón de tendencia ni de estacionalidad, los datos históricos se mantienen entorno a un valor.

celeberrima.com

Promedio móvil con n=2, w_1=2 y w_2=1

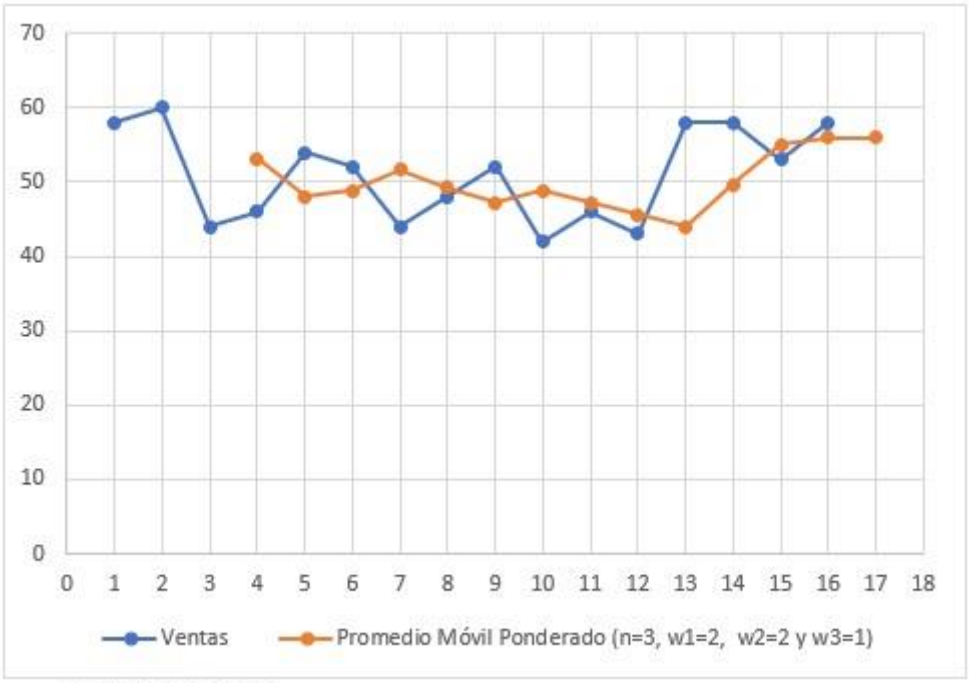
Ahora vamos a realizar un pronóstico considerando solo dos datos históricos (n=2) y dando un peso de 2 al más reciente w_1=2 y de 1 al más antiguo w_2=1. Dado que necesitamos dos datos para obtener un pronóstico vamos a comenzar con el pronóstico de la semana 3 usando las observaciones de las semanas 1 y 2.



celeberrima.com

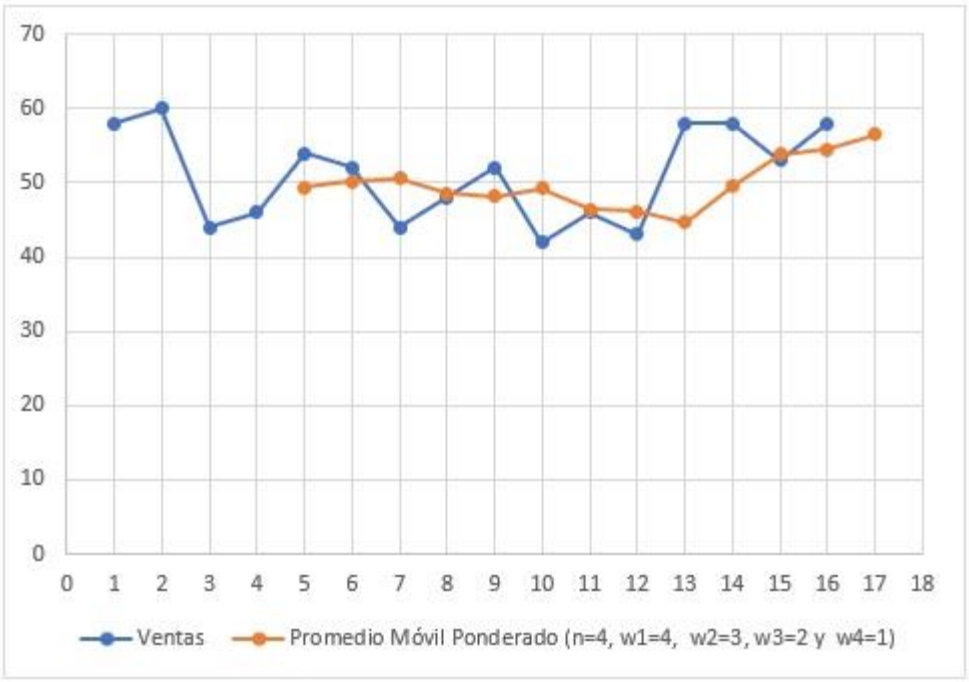
Promedio móvil con n=3, w_1=2, w_2=2 y w_3=1

Ahora vamos a probar un promedio móvil ponderado que considera tres datos históricos (n=3) y se pondera con 2 el dato más reciente, con 2 la segunda observación más reciente y con 1 el valor más antiguo. Como necesitamos tres valores para obtener el primer pronóstico comenzamos con F_4.



Promedio móvil con n=4, w_1=4, w_2=3, w_3=2 y w_4=1

Finalmente, vamos a realizar un pronóstico con 4 observaciones y asignando un peso de 4 al dato más reciente, de 3 al segundo dato más reciente, de 2 al tercer dato más reciente y de 1 al dato más antiguo. Como es de suponer comenzamos con el pronóstico F_5 correspondiente a la semana



celeberrima.com

En la siguiente tabla se muestran los resultados obtenidos para cada uno de los promedios móviles ponderados:

Semana	Ventas	Promedio Móvil Ponderado (n=2, w_1=2 y w_2=1)	Promedio Móvil Ponderado (n=3, w_1=2, w_2=2 y w_3=1)	Promedio Móvil Ponderado (n=4, w_1=4, w_2=3, w_3=2 y w_4=1)
1	58	-	-	-
2	60	-	-	-
3	44	59.33	-	-
4	46	49.33	53.20	-
5	54	45.33	48.00	49.40
6	52	51.33	48.80	50.20
7	44	52.67	51.60	50.60
8	48	46.67	49.20	48.60
9	52	46.67	47.20	48.20
10	42	50.67	48.80	49.20
11	46	45.33	47.20	46.40
12	43	44.67	45.60	46.20
13	58	44.00	44.00	44.60
14	58	53.00	49.60	49.50
15	53	58.00	55.00	53.80
16	58	54.67	56.00	54.50
17	-	56.33	56.00	56.50

Es natural que se obtengan valores diferentes para el pronóstico al utilizar diferentes valores de n y de w_i. Los promedios móviles ponderados pueden ser más sensibles a cambios recientes que los promedios móviles simples, siempre y cuando la ponderación de las observaciones recientes sea mayor, pero en esta situación se corre el riesgo de confundir una variación aleatoria con un cambio a u nivel de demanda superior o inferior.

Nuevamente, la elección de las ponderaciones dependerá del encargado del pronóstico pero la desviación media absoluta más pequeña es un buen indicador de que conjunto de ponderaciones elegir. Se pudo haber utilizado cualquier otro conjunto de ponderaciones, la elección dependerá del pronóstico que resulte más preciso.

3.5.- Pronostico regresión lineal

Cómo utilizar una Regresión Lineal para realizar un Pronóstico de Demanda

$$y = \beta_0 + \beta_1 x$$

$$\beta_0 = \bar{y} - b\bar{x}$$

$$\beta_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

x	y	xy	x^2	y^2
1	600	600	1	360.000
2	1.550	3.100	4	2.402.500
3	1.500	4.500	9	2.250.000
4	1.500	6.000	16	2.250.000
5	2.400	12.000	25	5.760.000
6	3.100	18.600	36	9.610.000
7	2.600	18.200	49	6.760.000

El Método de Mínimos Cuadrados o Regresión Lineal se utiliza tanto para pronósticos de *series de tiempo* como para pronósticos de relaciones causales. En particular cuando la variable dependiente cambia como resultado del tiempo se trata de un análisis de serie temporal.

En el siguiente artículo desarrollaremos un Pronóstico de Demanda haciendo uso de la información histórica de venta de un producto determinado durante los últimos 12 trimestres (3 años) cuyos datos se observan en la siguiente tabla resumen:

Trimestre	Ventas
1	600
2	1.550
3	1.500
4	1.500
5	2.400
6	3.100
7	2.600
8	2.900
9	3.800
10	4.500
11	4.000
12	4.900

La ecuación de mínimos cuadrados para la regresión lineal es la que se muestra a continuación donde β0 y β1 son los parámetros de *intercepto* y *pendiente*, respectivamente:

$$y = \beta_0 + \beta_1 x$$

$$\beta_0 = \bar{y} - b\bar{x}$$

$$\beta_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

Estimar los valores de dichos parámetros es sencillo haciendo uso de una planilla Excel tal como muestra la tabla a continuación:

	x	y	xy	x^2	y^2
	1	600	600	1	360.000
	2	1.550	3.100	4	2.402.500
	3	1.500	4.500	9	2.250.000
	4	1.500	6.000	16	2.250.000
	5	2.400	12.000	25	5.760.000
	6	3.100	18.600	36	9.610.000
	7	2.600	18.200	49	6.760.000
	8	2.900	23.200	64	8.410.000
	9	3.800	34.200	81	14.440.000
	10	4.500	45.000	100	20.250.000
	11	4.000	44.000	121	16.000.000
	12	4.900	58.800	144	24.010.000
PROMEDIO	**6,5**	**2.779,17**			
SUMA			**268.200**	**650**	
n	12				

Luego evaluamos en las ecuaciones presentadas anteriormente para obtener los valores de β0 y β1:

$$\beta_1 = \frac{268.200 - 12 * 6{,}5 * 2.779{,}17}{650 - 12 * 6{,}5^2} \cong 359{,}61$$

$$\beta_0 = 2.779{,}17 - 359{,}61 * 6{,}5 \cong 441{,}71$$

Una vez obtenido los parámetros de la regresión lineal se puede desarrollar un pronóstico de demanda (columna color naranja) evaluando en la ecuación de la regresión para los distintos valores de la variable independiente (x).

Por ejemplo, para el primer trimestre el pronóstico es: Y(1)=441,71+359,61*1=801,3.

Observación: *los valores de los pronósticos han sido redondeados* arbitrariamente *a un decimal para mayor comodidad.*

	x	y	xy	x^2	y^2	Y
	1	600	600	1	360.000	**801,3**
	2	1.550	3.100	4	2.402.500	**1.160,9**
	3	1.500	4.500	9	2.250.000	**1.520,5**
	4	1.500	6.000	16	2.250.000	**1.880,2**
	5	2.400	12.000	25	5.760.000	**2.239,8**
	6	3.100	18.600	36	9.610.000	**2.599,4**
	7	2.600	18.200	49	6.760.000	**2.959,0**
	8	2.900	23.200	64	8.410.000	**3.318,6**
	9	3.800	34.200	81	14.440.000	**3.678,2**
	10	4.500	45.000	100	20.250.000	**4.037,8**
	11	4.000	44.000	121	16.000.000	**4.397,4**
	12	4.900	58.800	144	24.010.000	**4.757,0**
PROMEDIO	**6,5**	**2.779,17**				
SUMA			**268.200**	**650**		
n	12					
β_0	441,71					
β_1	359,61					

Notar que con la información que hemos obtenido podemos calcular el MAD y la Señal de Rastreo y utilizar estos indicadores para validar la conveniencia de utilizar este procedimiento como dispositivo de pronóstico.

Adicionalmente puede resultar de interés consultar el artículo Ejemplo de una Regresión Lineal Múltiple para un Pronóstico con Excel y Minitab que muestra cómo abordar el caso de realizar una regresión lineal con más de una *variable independiente* (explicativa).

Siguiendo con nuestro análisis a continuación podemos desarrollar un pronóstico de demanda para los próximos 4 trimestres (un año) que corresponden a los trimestres 13, 14, 15 y 16:

- Y(13)=441,71+359,61*13=5.116,64

- Y(14)=441,71+359,61*14=5.476,25
- Y(15)=441,71+359,61*15=5.835,86
- Y(16)=441,71+359,61*16=6.195,47

UNIDAD IV

Relaciones entre variables

4.1.- Introducción

En el análisis de los estudios clínico-epidemiológicos surge muy frecuentemente la necesidad de determinar la relación entre dos variables cuantitativas en un grupo de sujetos. Los objetivos de dicho análisis suelen ser:

a. Determinar si las dos variables están correlacionadas, es decir si los valores de una variable tienden a ser más altos o más bajos para valores más altos o más bajos de la otra variable. b. Poder predecir el valor de una variable dado un valor determinado de la otra variable. c. Valorar el nivel de concordancia entre los valores de las dos variables

4.2.- Correlación

En este artículo trataremos de valorar la asociación entre dos variables cuantitativas estudiando el método conocido como correlación. Dicho cálculo es el primer paso para determinar la relación entre las variables. La predicción de una variable. La predicción de una variable dado un valor determinado de la otra precisa de la regresión lineal que abordaremos en otro artículo.

La cuantificación de la fuerza de la relación lineal entre dos variables cuantitativas, se estudia por medio del cálculo del coeficiente de correlación de Pearson. Dicho coeficiente oscila entre –1 y +1. Un valor de –1 indica una relación lineal o línea recta positiva perfecta. Una correlación próxima a cero indica que no hay relación lineal entre las dos variables.

El realizar la representación gráfica de los datos para demostrar la relación entre el valor del coeficiente de correlación y la forma de la gráfica es fundamental ya que existen relaciones no lineales.

El coeficiente de correlación posee las siguientes características:

a. El valor del coeficiente de correlación es independiente de cualquier unidad usada para medir las variables. b. El valor del coeficiente de correlación se altera de forma importante ante la presencia de un valor extremo, como sucede con la desviación típica. Ante estas situaciones conviene realizar una transformación de datos que cambia la escala de medición y modera el efecto de valores extremos (como la transformación logarítmica). c. El coeficiente de correlación mide solo la relación con una línea recta. Dos variables pueden tener una relación curvilínea fuerte, a pesar de que su correlación sea pequeña. Por tanto cuando analicemos las relaciones entre dos variables debemos representarlas gráficamente y posteriormente calcular el coeficiente de correlación. d. El coeficiente de correlación no se debe extrapolar más allá del rango de valores observado de las variables a estudio ya que la relación existente entre X e Y puede cambiar fuera de dicho rango. e. La correlación no implica causalidad. La causalidad es un juicio de valor que requiere más información que un simple valor cuantitativo de un coeficiente de correlación.

El coeficiente de correlación de Pearson (r) puede calcularse en cualquier grupo de datos, sin embargo la validez del test de hipótesis sobre la correlación entre las variables requiere en sentido estricto: a) que las dos variables procedan de una muestra aleatoria de individuos. b) que al menos una de las variables tenga una distribución normal en la población de la cual la muestra procede. Para el cálculo válido de un intervalo de confianza del coeficiente de correlación de r ambas variables deben tener una distribución normal. Si los datos no tienen una distribución normal, una o ambas variables se pueden transformar (transformación logarítmica) o si no se calcularía un coeficiente de correlación no paramétrico (coeficiente de

correlación de Spearman) que tiene el mismo significado que el coeficiente de correlación de Pearson y se calcula utilizando el rango de las observaciones.

El cálculo del coeficiente de correlación (r) entre peso y talla de 20 niños varones se muestra en la tabla 1. La covarianza, que en este ejemplo es el producto de peso (kg) por talla (cm), para que no tenga dimensión y sea un coeficiente, se divide por la desviación típica de X (talla) y por la desviación típica de Y (peso) con lo que obtenemos el coeficiente de correlación de Pearson que en este caso es de 0.885 e indica una importante correlación entre las dos variables. Es evidente que el hecho de que la correlación sea fuerte no implica causalidad. Si elevamos al cuadrado el coeficiente de correlación obtendremos el coeficiente de determinación (r2=0.783) que nos indica que el 78.3% de la variabilidad en el peso se explica por la talla del niño. Por lo tanto existen otras variables que modifican y explican la variabilidad del peso de estos niños. La introducción de más variable con técnicas de análisis multivariado nos permitirá identificar la importancia de que otras variables pueden tener sobre el peso.

4.3.- Coeficiente de correlación de Pearson

Tabla 1. Cálculo del Coeficiente de correlación de Pearson entre las variables talla y peso de 20 niños varones

Y	X			
Peso (Kg)	Talla (cm)	$X - \overline{X}$	$Y - \overline{Y}$	$(X - \overline{X})*(Y - \overline{Y})$
9	72	5.65	1.4	7.91
10	76	9.65	2.4	23.16
6	59	-7.35	-1.6	11.76
8	68	1.65	0.4	0.66
10	60	-6.35	2.4	-15.24
5	58	-8.35	-2.6	21.71
8	70	3.65	0.4	1.46
7	65	-1.35	-0.6	0.81

Tabla 1. Cálculo del Coeficiente de correlación de Pearson entre las variables talla y peso de 20 niños varones

4	54	-12.35	-3.6	44.46
11	83	16.65	3.4	56.61
7	64	-2.35	-0.6	1.41
7	66	-0.35	-0.6	0.21
6	61	-5.35	-1.6	8.56
8	66	-0.35	0.4	-0.14
5	57	-9.35	-2.6	24.31
11	81	14.65	3.4	49.81
5	59	-7.35	-2.6	19.11
9	71	4.65	1.4	6.51
6	62	-4.35	-1.6	6.96
10	75	8.65	2.4	20.76
				$\sum$ 290.8

$X\left(Media\ de\ \overline{X} = 66.35\right)$

$Y\left(Media\ de\ \overline{Y} = 7.6\right)$

4.4.- covarianza

$$Covarianza = \frac{\sum\left(\overline{X} - X\right)*\left(\overline{Y} - Y\right)}{n-1} = \frac{290.8}{19} = 15.30$$

$$r = \frac{covarianza}{S_x * S_y} = \frac{15.30}{8.087 * 2.137} = 0.885$$

S_x = Desviación típica x = 8.087

S_y = Desviación típica y = 2.137

4.5.- Test de hipótesis de r

Tras realizar el cálculo del coeficiente de correlación de Pearson (r) debemos determinar si dicho coeficiente es estadísticamente diferente de cero. Para dicho calculo se aplica un test basado en la distribución de la t de student.

$$Error\ estandard\ de\ r = \sqrt{\frac{1-r^2}{n-2}}$$

Si el valor del r calculado (en el ejemplo previo r = 0.885) supera al valor del error estándar multiplicado por la t de Student con n-2 grados de libertad, diremos que el coeficiente de correlación es significativo.

El nivel de significación viene dado por la decisión que adoptemos al buscar el valor en la tabla de la t de Student.

En el ejemplo previo con 20 niños, los grados de libertad son 18 y el valor de la tabla de la t de student para una seguridad del 95% es de 2.10 y para un 99% de seguridad el valor es 2.88.

$$Error\ estandard\ de\ r = \sqrt{\frac{1-0.885^2}{20-2}} = 0.109$$

Como quiera que r = 0.885 > a 2.10 * 0.109 = 2.30 podemos asegurar que el coeficiente de correlación es significativo ($p<0.05$). Si aplicamos el valor obtenido en la tabla de la t de Student para una seguridad del 99% (t = 2.88) observamos que como r = 0.885 sigue siendo > 2.88 * 0.109 = 0.313 podemos a su vez asegurar que el coeficiente es significativo ($p<0.001$). Este proceso de razonamiento es válido tanto para muestras pequeñas como para muestras grandes. En esta última situación podemos comprobar en la tabla de la t de student que para una seguridad del 95% el valor es 1.96 y para una seguridad del 99% el valor es 2.58.

Intervalo de confianza del coeficiente de correlación

La distribución del coeficiente de correlación de Pearson no es normal pero no se puede transformar r para conseguir un valor z que sigue una distribución normal (transformación de Fisher) y calcular a partir del valor z el intervalo de confianza.

La transformación es:

$$z = 1/2 L_n \frac{1+r}{1-r}$$

L_n representa el logaritmo neperiano en la base e

$$El\ error\ standard\ de\ z\ es = \frac{1}{\sqrt{n-3}}$$

donde n representa el tamaño muestral. El 95% intervalo de confianza de z se calcula de la siguiente forma:

$$z_1(limite\ inf\ erior) = z - 1.96/\sqrt{n-3}$$

$$z_2(limite\ sup\ erior) = z + 1.96/\sqrt{n-3}$$

Tras calcular los intervalos de confianza con el valor z debemos volver a realizar el proceso inverso para calcular los intervalos del coeficiente r

$$\frac{e^{2z1}-1}{e^{2z1}+1} \quad a \quad \frac{e^{2z2}-1}{e^{2z2}+1}$$

Utilizando el ejemplo de la Tabla 1, obtenemos r = 0.885

$$z = 1/2 L_n \frac{1+0.885}{1-0.885} = 1.398$$

95% intervalo de confianza de z

$$z_1 = 1.398 - 1.96/\sqrt{20-3} = 0.922$$

$$z_2 = 1.398 + 1.96/\sqrt{20-3} = 1.873$$

Tras calcular los intervalos de confianza de z debemos proceder a hacer el cálculo inverso para obtener los intervalos de confianza de coeficiente de correlación r que era lo que buscábamos en un principio antes de la transformación logarítmica.

$$\frac{e^{2*0.922}-1}{e^{2*0.922}+1} \quad a \quad \frac{e^{2*1.873}-1}{e^{2*1.873}+1}$$

0.726 a 0.953 son los intervalos de confianza (95%) de r.

Presentación de la correlación

Se debe mostrar siempre que sea posible la gráfica que correlaciona las dos variables de estudio (Fig 1). El valor de r se debe mostrar con dos decimales junto con el valor de la p si el test de hipótesis se realizó para demostrar que r es estadísticamente diferente de cero. El número de observaciones debe a su vez estar indicado.

Figura 1. Correlación entre Peso y Talla

Figura 1. Correlación entre Peso y Talla

4.6.- Interpretación de la correlación

El coeficiente de correlación como previamente se indicó oscila entre –1 y +1 encontrándose en medio el valor 0 que indica que no existe asociación lineal entre las dos variables a estudio. Un coeficiente de valor reducido no indica necesariamente que no exista correlación ya que las variables pueden presentar una relación no lineal como puede ser el peso del recién nacido y el tiempo de gestación. En este caso el r infra estima la asociación al medirse linealmente. Los métodos no paramétricos estarían mejor utilizados en este caso para mostrar si las variables tienden a elevarse conjuntamente o a moverse en direcciones diferentes.

La significancia estadística de un coeficiente debe tenerse en cuenta conjuntamente con la relevancia clínica del fenómeno que estudiamos ya que coeficientes de 0.5 a 0.7 tienden ya a ser significativos como muestras pequeñas. Es por ello muy útil calcular el intervalo de confianza del r ya que en muestras pequeñas tenderá a ser amplio.

La estimación del coeficiente de determinación (r2) nos muestra el porcentaje de la variabilidad de los datos que se explica por la asociación entre las dos variables.

Como previamente se indicó la correlación elevada y estadísticamente significativa no tiene que asociarse a causalidad. Cuando objetivamos que dos variables están correlacionadas diversas razones pueden ser la causa de dicha correlación: a) pude que X influencie o cause Y, b) puede que influencie o cause X, c) X e Y pueden estar influenciadas por terceras variables que hace que se modifiquen ambas a la vez.

El coeficiente de correlación no debe utilizarse para comparar dos métodos que intentan medir el mismo evento, como por ejemplo dos instrumentos que miden la tensión arterial. El coeficiente de correlación mide el grado de asociación entre dos cantidades pero no mira el nivel de acuerdo o concordancia. Si los instrumentos de medida miden sistemáticamente cantidades diferentes uno del otro, la correlación puede ser 1 y su concordancia ser nula.

Coeficiente de correlación de los rangos de Spearman

Este coeficiente es una medida de asociación lineal que utiliza los rangos, números de orden, de cada grupo de sujetos y compara dichos rangos. Existen dos métodos para calcular el coeficiente de correlación de los rangos uno señalado por Spearman y otro por Kendall. El r de Spearman llamado también rho de Spearman es más fácil de calcular que el de Kendall. El coeficiente de correlación de Spearman es exactamente el mismo que el coeficiente de correlación de Pearson calculado sobre el rango de observaciones. En definitiva la correlación estimada entre X e Y se halla calculado el coeficiente de correlación de Pearson para el conjunto de rangos apareados. El coeficiente de correlación de Spearman es recomendable utilizarlo cuando los datos presentan valores externos ya que dichos valores afectan mucho el coeficiente de correlación de Pearson, o ante distribuciones no normales.

El cálculo del coeficiente viene dado por:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

en donde di = rxi – ryi es la diferencia entre los rangos de X e Y.

Los valores de los rangos se colocan según el orden numérico de los datos de la variable.

Ejemplo: Se realiza un estudio para determinar la asociación entre la concentración de nicotina en sangre de un individuo y el contenido en nicotina de un cigarrillo (los valores de los rangos están entre paréntesis) .

X	Y
Concentración de Nicotina en sangre (nmol/litro)	Contenido de Nicotina por cigarrillo (mg)
185.7 (2)	1.51 (8)
197.3 (5)	0.96 (3)
204.2 (8)	1.21 (6)
199.9 (7)	1.66 (10)
199.1 (6)	1.11 (4)
192.8 (6)	0.84 (2)
207.4 (9)	1.14 (5)
183.0 (1)	1.28 (7)
234.1 (10)	1.53 (9)
196.5 (4)	0.76 (1)

Si existiesen valores coincidentes se pondría el promedio de los rangos que hubiesen sido asignado si no hubiese coincidencias. Por ejemplo si en una de las variables X tenemos:

Para el cálculo del ejemplo anterior de nicotina obtendríamos el siguiente resultado:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6\left[(2-8)^2 + (5-3)^2 + (8-6)^2 + \ldots + (4-1)^2\right]}{10(10^2-1)} = 1 - \frac{6(120)}{10(99)} = 0.27$$

Si utilizamos la fórmula para calcular el coeficiente de correlación de Pearson de los rangos obtendríamos el mismo resultado

$$r_s = \frac{n\sum r_x r_y - \sum r_x \sum r_y}{\sqrt{\left[n\sum r_x^2 - \left(\sum r_x\right)^2\right]\left[n\sum r_y^2 - \left(\sum r_y\right)^2\right]}}$$

$$\sum r_x = \sum r_y = 55 \qquad \sum r_x^2 = \sum r_y^2 = 385$$

$$\sum r_x r_y = 2\,(8) + 5\,(3) + 8\,(6) + \ldots + 4\,(1) = 325$$

$$r_s = \frac{10(325) - 55(55)}{\sqrt{\left[10(385) - 55^2\right]\left[10(385) - 55^2\right]}} = 0.27$$

La interpretación del coeficiente rs de Spearman es similar a la Pearson. Valores próximos a 1 indican una correlación fuerte y positiva. Valores próximos a –1 indican una correlación fuerte y negativa. Valores próximos a cero indican que no hay correlación lineal. Así mismo

el tiene el mismo significado que el coeficiente de determinación de r2. La distribución de rs es similar a la r por tanto el cálculo de los intervalos de confianza de rs se pueden realizar utilizando la misma metodología previamente explicada para el coeficiente de correlación de Pearson.

MEDIDAS DE ASOCIACIÓN ENTRE DOS VARIABLES Las medidas de asociación tratan de estimar la magnitud con la que dos fenómenos se relacionan. Se emplean: Covarianza: Es una medida de asociación entre dos variables y se calcula: Muestral:

$$Sxy = \frac{\Sigma(xi - X)(yi - Y)}{n-1}$$

Poblacional: $Sxy = \frac{\Sigma(xi - \mu x)(yi - \mu y)}{N}$

Coeficiente de correlación: Puede utilizarse para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas.

Muestral: $rxy = \frac{Sxy}{Sx\ Sy}$

Poblacional: $Pxy = \frac{\tilde{O}xy}{\tilde{O}x\ \tilde{O}y}$

Coeficiente de regresión: Indica el número de unidades en que se modifica la variable dependiente "Y" por efecto del cambio de la variable independiente "X" o viceversa en una unidad de medida. Clases de coeficiente de Regresión: El coeficiente de regresión puede ser: Positivo, Negativo y Nulo. Es positivo cuando las variaciones de la variable independiente X son directamente proporcionales a las variaciones de la variable dependiente "Y".

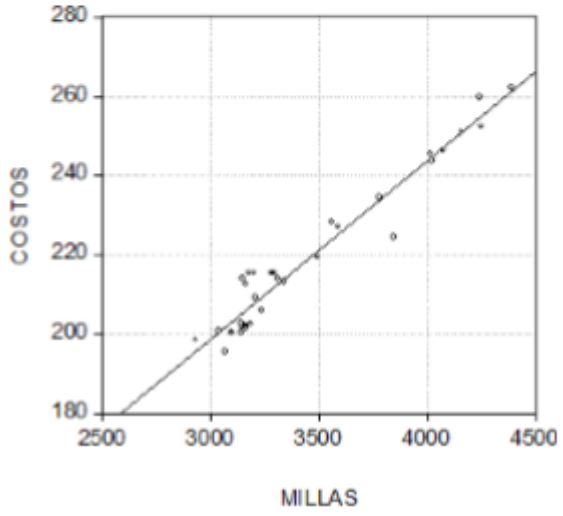
Es negativo, cuando las variaciones de la variable independiente "X" son inversamente proporcionales a las variaciones de las variables dependientes "Y". Es nulo o cero, cuando entre las variables dependientes "Y" e independientes "X" no existen relación alguna.

Se calcula:

$$y - Y = \frac{Sxy}{S^2 y}$$

$$Y - Yl = m\,(x - xi)$$

Gráfico de dispersión:



Bibliografía básica y complementaria:

Probabilidad y estadística de George Canavos

Estadística de Murray R. Spiegel

Videos academicos

https://www.youtube.com/watch?v=1myBo87lYyU

profe Alex medidas de dispersión

https://www.youtube.com/watch?v=Eju_9eM4PZg

profe Alex medidas de posición

https://www.youtube.com/watch?v=2Y68-BfSdbI

https://www.youtube.com/watch?v=EE2a2Cr-JfY

coeficiente de Pearson prof Alex