1 ESTADÍSTICA DESCRIPTIVA

La ciencia describe, explica y predice.
Stephen Hawking, en "Historia del tiempo".

Objetivo de la unidad: En el desarrollo de la presente Unidad de Aprendizaje (UA), el estudiante sintetizará un conjuntos de datos, tomados de una situación real o simulada, mediante el uso de tablas, gráficas y el cálculo de medidas de tendencia central y de dispersión, con la finalidad de comprender e interpretar la información estadística que se le presente en diferentes contextos de su vida académica.

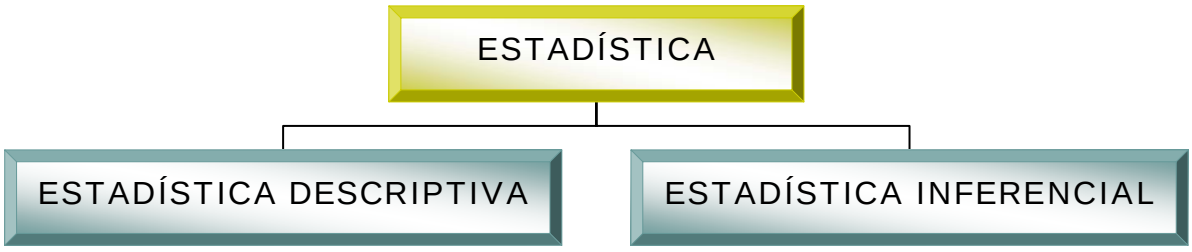
En este capítulo se dará un repaso a los conocimientos sobre **Estadística Descriptiva** estudiados en bachillerato o secundaria. Se explica la importancia de la Estadística en el mundo contemporáneo, su concepción y clasificación y el papel que desempeña la teoría de la probabilidad como puente entre la estadística descriptiva y la estadística inferencial. Enseguida se presentan algunos conceptos que frecuentemente aparecerán a lo largo del curso y que son importantes para la comprensión de otros. Posteriormente se desarrolla la notación sumatoria a través de sus propiedades y ejemplos numéricos para dar lugar a la revisión de las medidas descriptivas y de dispersión, tanto para datos no-agrupados como para datos agrupados (tablas). También se enumeran los pasos para construir tablas de frecuencias. Al final se revisan los principales métodos gráficos para resumir conjuntos de datos, con base en las tablas de frecuencia construidas; de los métodos gráficos, se caracterizan y se describe la forma de construirlos, de forma que el estudiante pueda identificarlas, diferenciarlas, construirlas e interpretar, en situaciones del mundo real, lo que representan.

1 Introducción

Usualmente relacionamos con la palabra "estadística" grandes conjuntos de datos (números) que se han colectado con un fin determinado. La estadística nace como respuesta a una necesidad de los estados de "tener idea" cuánto gastarán en rubros específicos para el siguiente año en cuestiones como el número de defunciones de personas desamparadas o el número de vacunas que deben tenerse disponibles para los recién nacidos. La estadística descriptiva se relaciona con las primeras técnicas utilizadas en la organización de datos, resumiéndolos en tablas, gráficas o a través del cálculo de medidas de tendencia central y de dispersión.

En la actualidad, la mayor parte del uso de la estadística, particularmente en la ciencia y en la ingeniería, se dirige a la Inferencia más que a la descripción; sin embargo, es indudable que la estadística descriptiva tiene aun una utilidad considerable. Los intervalos de confianza, por ejemplo, se usan en la construcción de las bandas de especificación y de procesos, en el control estadístico de la calidad.

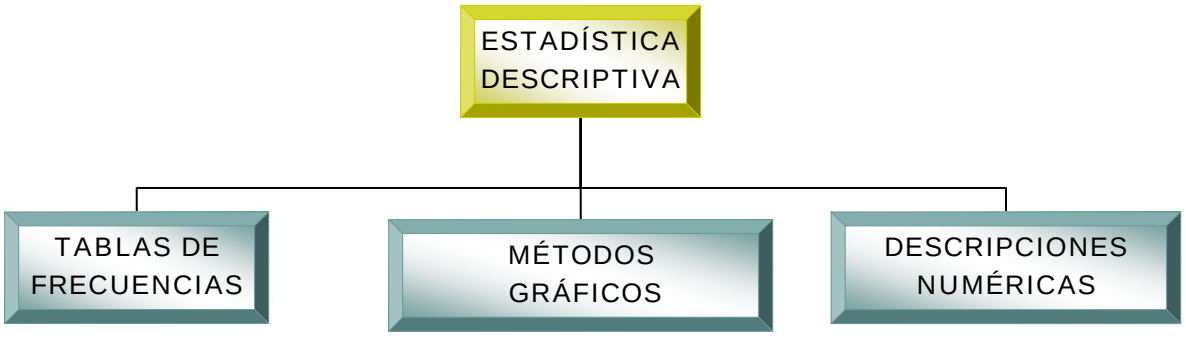
PROBABILIDAD: PUENTE ENTRE LA DESCRIPCIÓN Y LA INFERENCIA



La **Probabilidad** es el "puente" entre la estadística descriptiva y la estadística inferencial; provee los fundamentos teóricos que hacen posible las conjeturas a universos más amplios que la muestra. Mientras que la estadística descriptiva nos sirve para resumir información, la estadística inferencial (o simplemente Inferencia) se refiere prácticamente a la estimación de intervalos de confianza y a realizar pruebas de hipótesis sobre los parámetros poblacionales.

CONCEPTOS DE ESTADÍSTICA DESCRIPTIVA E INFERENCIAL

La estadística descriptiva se refiere a los métodos diseñados para resumir y organizar datos de tal forma que podamos captar la información esencial de los mismos o sus patrones de comportamiento a través de medidas descriptivas, gráficas y tablas de frecuencias. Podemos visualizar estas ramas de la estadística descriptiva en el siguiente diagrama:



Los métodos tabulares y gráficos nos permiten organizar y presentar datos de tal forma que los aspectos sobresalientes (y esenciales) de los mismos se pueden entender rápida y fácilmente. En ocasiones estos métodos nos ayudan a establecer hipótesis iniciales sobre la naturaleza del fenómeno que queremos estudiar o investigar.

La Estadística es la ciencia que se ocupa de la ordenación y análisis de datos procedentes de muestras y de la realización de inferencias sobre las poblaciones de las que éstas proceden. Generalmente se pueden distinguir dos fases en la realización de cualquier experimento o trabajo de investigación. La primera consiste en la observación y análisis de los hechos que acontecen (colecta de la información) y la segunda, en la interpretación y obtención de conclusiones.

La **estadística descriptiva** es la primera herramienta para el manejo de los datos y proporciona métodos para resumirlos y organizarlos. Tiene como objetivo caracterizar, describir y extraer conclusiones sobre una muestra de datos. Puede ser útil en la primera fase de una investigación.

La **estadística inferencial** implica obtener conclusiones sobre los caracteres de una población a partir de los datos muestrales y requiere el cálculo de probabilidades, que nos den el grado de certeza de dichas conclusiones.

POBLACIÓN Y MUESTRA

Población. Es el conjunto de referencia sobre el que van a recaer las observaciones, todos los elementos que tengan información sobre el fenómeno que se estudia (por ejemplo, si estudiamos el precio de la vivienda en una ciudad, la población será el total de viviendas de dicha ciudad). Generalmente este conjunto viene definido por comprensión, es decir, citando la propiedad que lo caracteriza (habitantes del sexo femenino de Yucatán con edades de 17 a 23 años, por ejemplo). Las poblaciones pueden ser **finitas** o **infinitas**. Son poblaciones finitas si es posible contar sus elementos y son infinitas en caso contrario.

Individuo. Es cada uno de los elementos que componen la población en estudio. Así, si estudiamos la altura de las niñas de una clase, cada alumna es un individuo; si estudiamos el precio de la vivienda, cada vivienda es un individuo. Si estudiamos el peso de unos cerdos, cada animalito es un individuo. Un individuo es cualquier persona, animal, planta u objeto observable.

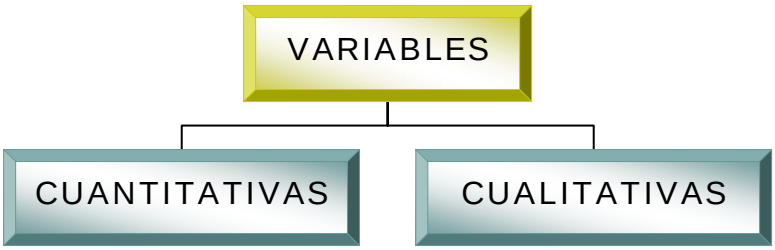
Muestra. Es un grupo de individuos que seleccionamos de la población. Se suelen tomar muestras cuando es difícil o costosa la observación de todos los elementos de la población. El número de elementos de la misma se llama **tamaño de la muestra**. Se deben escoger los individuos de la muestra de manera que sean representativos de la población de la que proceden, es decir, que conserven las propiedades de aquélla. Así, si se estudia el precio de la vivienda de una ciudad, usualmente no se colectará información sobre todas las viviendas de la ciudad (sería una labor muy compleja), sino que suele seleccionarse un subgrupo (muestra) que se entienda que es suficientemente representativo. Las **técnicas de muestreo** nos proveen la metodología en cada caso particular, dependiendo de la variabilidad de la población y del nivel de confianza que se desea. Existen así entre otros, el muestreo aleatorio simple, el muestreo estratificado, el muestreo por conglomerados, el muestreo sistemático y el muestreo en dos etapas.

CLASIFICACIÓN DE VARIABLES

Caracteres o variables estadísticas. El caracter es cualquier cualidad o propiedad inherente al individuo. Por ejemplo, si el individuo observado es un libro, podremos describirlo mediante los caracteres peso, tamaño, número de hojas, color de las pastas, etc. A los individuos de la población estudiantil, podemos observarles la estatura, el peso, el color de ojos, los idiomas que hablan o el sexo; cada una de estas características la

llamamos *variable estadística* y la representamos normalmente por las letras mayúsculas X, Y, Z.

Hay caracteres que son medibles, esto es, se pueden cuantificar, tal como la edad, el peso y la estatura de las personas, el precio de un producto, los ingresos anuales, etc. Sin embargo existen caracteres que no se pueden cuantificar, como el color de los ojos, el estado civil, el sexo, la nacionalidad, el nivel de felicidad declarado, etc. A los primeros se les llama caracteres cuantitativos (y a las variables que los representan variables cuantitativas) y a los segundos caracteres cualitativos o categóricos (y variables cualitativas a las variables que los representan). De este modo, podemos clasificar las variables en cuantitativas y cualitativas como se muestra en el siguiente esquema:



El carácter o variable estadística cualitativa estado civil puede tomar los valores o modalidades: casado/a, soltero/a o viudo/a (hay una categoría no formal llamada "ejerce sin título). La variable edad, que es cuantitativa puede tomar los valores: 10 años, 12 años, 15 años, etc. Los valores particulares de una variable se representa con las letras minúsculas x_1, x_2, …, x_n. El subíndice indica el número de individuo sobre el que se tomó la observación. Una variable puede tomar distintos valores y cada uno de éstos puede aparecer repetidas veces en la muestra.

A su vez, las variables cuantitativas se pueden clasificar en discretas y continuas.

Variable discreta. Toma valores aislados y no pueden tomar ningún valor entre dos valores consecutivos. Sólo puede tomar valores enteros (1, 2, 8).

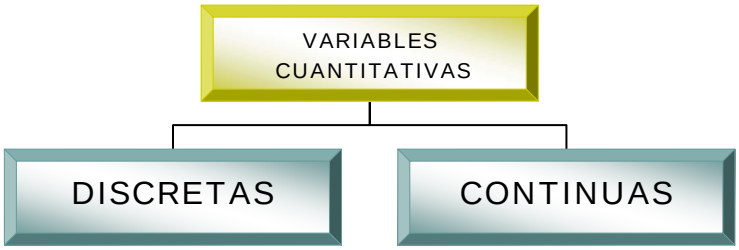
Ejemplo: número de hermanos (puede ser 1, 2, 3, ..., pero nadie podrá decir que tiene 3.45 hermanos).

Ejemplo: número de monedas que una persona lleva en la bolsa (0, 1, 2, 3, …)

Variable continua. Puede tomar cualquier valor real dentro de un intervalo real. Siempre puede tomar valores entre dos consecutivos, por muy próximos que los fijemos.

Ejemplo: la velocidad de un vehículo puede ser 80.3 km/h, 94.57 km/h.
Otros ejemplos: estatura de las personas, medida del tiempo,...etc.

A continuación el esquema de clasificación de las variables cuantitativas.



Las variables cualitativas (o categóricas) pueden clasificarse en nominales y ordinales.

ACTIVIDAD DE APRENDIZAJE

I. Conteste las siguientes preguntas.
   1. Escriba con sus propias palabras qué estudia la estadística.
   2. Escriba las diferencias entre población y muestra.
   3. Escriba dos ejemplos de población y dos ejemplos de muestra.
   4. Escriba dos ejemplos de: variable cualitativa, variable cuantitativa, variable discreta y variable continua.
   5. Escriba dos diferencias entre estadística descriptiva y estadística inferencial.
   6. Mencione un ejemplo de la vida real donde se utilice la estadística descriptiva.
   7. Mencione un ejemplo práctico donde se utilice la estadística inferencial.
   8. Investigue la definición de variables cualitativas nominales y ordinales.
   9. Escriba dos ejemplos de variables nominales y dos de variables ordinales.

II. Construya un mapa conceptual de la clasificación de las variables. Incluya ejemplos.

III. Construya un crucigrama de 10 filas por 10 columnas, usando los conceptos revisados en esta sección y otros relacionados que usted investigue.

IV. Investigue la biografía de un estadístico famoso. Enumere sus principales contribuciones y las posibles aplicaciones que tienen esas contribuciones en el mundo actual.

2 Notación sumatoria

- Hacer cálculos numéricos usando la notación sumatoria.
- Usar las propiedades del operador sumatoria para deducir algunos resultados.

En ocasiones necesitamos sumar las observaciones de una serie de datos. Las observaciones son representadas por $x_1, x_2, \ldots, x_n$. El último subíndice, *n*, representa el número total de observaciones. Por ejemplo, una serie de datos con $n = 5$ observaciones 2.1, 3.2, 4.1, 5.6, y 3.7 son representados por los símbolos x_1, x_2, x_3, x_4, x_5 donde $x_1 = 2.1$, $x_2 = 3.2$, $x_3 = 4.1$, $x_4 = 5.6$ y $x_5 = 3.7$. Podemos indicar la suma de las observaciones en la serie de datos o algunos otros números derivados de la misma, por medio de la notación sumatoria representada por la letra griega sigma mayúscula, Σ. A continuación la definición de esta notación.

Significado del operador $\sum$

$$\sum_{i=a}^{b} x_a + x_{a+1} + x_{a+2} + \ldots + x_{b-1} + x_b$$

donde *a* y *b* son números enteros, con $a \leq b$. La definición indica que se suman los valores que representan las x_i desde *i=a* hasta *i=b*. El valor de *a* se conoce como **extremo** (o límite) **inferior** de la sumatoria y *b* es el **extremo superior**. El símbolo *i* es el **índice de la sumatoria** o contador. Este índice puede ser representado por diferentes letras, aunque las más usadas son *i*, *j*, *k*, *l* y *h*. El extremo inferior nos indica el primer valor que toma el índice de la sumatoria y el extremo superior, su último valor.

Ejemplos ilustrativos.

.) $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots x_n$

..) $\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$

…) $\sum_{i=1}^{4} (x_i - 2) = (x_1 - 2) + (x_2 - 2) + (x_3 - 2) + (x_4 - 2)$

Ejemplos numéricos.

Suponga que las cuatro observaciones en una serie de datos son $x_1 = 3$, $x_2 = 5$, $x_3 = 4$, $x_4 = 3$.

i) $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 3 + 5 + 4 + 3 = 15$

ii) $\sum_{i=1}^{4} 3x_i = 3x_1 + 3x_2 + 3x_3 + 3x_4 = 3\sum_{i=1}^{4} x_i = 3\ (15) = 45$

iii) $\sum_{i=1}^{4}(x_i-2)=(x_1-2)+(x_2-2)+(x_3-2)+(x_4-2)=\sum_{i=1}^{4}x_i-4(2)=15-8=7$

iv) $\sum_{i=1}^{4}x_i^2=x_1^2+x_2^2+x_3^2+x_4^2=3^2+5^2+4^2+3^2=59$

v) $\sum_{i=1}^{4}(x_i-2)^2=(x_1-2)^2+(x_2-2)^2+(x_3-2)^2+(x_4-2)^2$

$$=(3-2)^2+(5-2)^2+(4-2)^2+(3-2)^2$$
$$=1+9+4+1=15$$

A continuación se presentan algunas propiedades de la operación sumatoria para su uso en la deducción de resultados usados con frecuencia.

Propiedades de la sumatoria

Considere que *b* y *c* son constante y que *x* y *y* son variables, entonces:

1) $\sum_{i=1}^{n}c=nc$ La sumatoria de 1 hasta n de una constante es n veces la constante

2) $\sum_{i=1}^{n}b\,x_i=b\sum_{i=1}^{n}x_i$ La sumatoria de una constante por una variable es igual a la constante por la sumatoria de la variabl

3) $\sum_{i=1}^{n}(x_i+y_i)=\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}y_i$ La sumatoria de una suma de variables es igual a la suma de las sumatorias respectivas.

*Demostrar algebraicamente las propiedades en el pizarrón y hablar de su importancia.

ACTIVIDAD DE APRENDIZAJE

I. Resuelva las siguientes sumatorias.
(Copiar los datos acá para resolver las sumatorias)

i) $\sum_{i=1}^{6}x_i=$

ii) $\sum_{i=3}^{7}3x_i=$

iii) $\sum_{i=1}^{4}(x_i-2)=$

iv) $\sum_{i=1}^{7}x_i^2=$

3 Datos no agrupados

- Calcular medidas de tendencia central y de dispersión para datos no-agrupados.
- Aplicar las propiedades de la notación sumatoria para exhibir que las suma de todas las desviaciones respecto al promedio de un conjunto de datos, siempre es cero.
- Aplicar las propiedades de la notación sumatoria para deducir la fórmula de operación de la varianza muestral a partir de su fórmula de definición.

En esta sección estudiamos las medidas de tendencia central y de dispersión para datos no agrupados. Los datos no-agrupados son aquéllos sin procesar, esto es, que aún están enlistados en la forma como se fueron haciendo los registros; su contraparte son los datos agrupados en tablas de distribución de frecuencias. Las medidas de tendencia central que abordaremos sobre estos datos sin procesar son la media aritmética, la moda y la mediana; y revisaremos las siguientes medidas de dispersión o de variabilidad: el rango o amplitud, la varianza, la desviación estándar y la desviación media absoluta.

Ejemplo de datos no agrupados. Un ejemplo de datos no-agrupados es el que se presenta enseguida y que corresponde a las calificaciones que 50 jóvenes y señoritas del Tecnológico obtuvieron en una unidad de Probabilidad.

Cuadro 1. Calificaciones obtenidas por 50 alumnos en un examen de Probabilidad

71	52	58	60	66	67	91	70	75	83
88	89	82	93	72	71	61	74	76	61
57	64	62	74	64	77	87	62	85	80
68	76	80	82	31	85	62	97	72	69
57	87	73	72	79	84	81	79	81	73

Como puede observar, las calificaciones en el Cuadro 1, están "revueltas"; aparecen en el orden como fueron registradas en los estudiantes. Puede usted hacer otros ejemplos de datos no agrupados registrando las estaturas, pesos, edades o número de hermanos de sus compañeros del salón.

MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central más frecuentes son la mediana, la moda y la media aritmética; ésta última es la más usada por su practicidad y sus buenas propiedades estadísticas.

LA MEDIANA

La mediana, representada por $\tilde{m}$, es el valor medio de una serie cuando los valores se han ordenado ascendentemente. Para la serie 3, 4, 5, 8 y 9, la mediana es el tercer valor, 5. Si hay seis valores en una serie, por ejemplo 3, 4, 5, 8, 9 y 10, cualquier valor entre 5 y 8 dividiría la serie en dos partes iguales; por tanto, cualquiera de tales valores podría ser la mediana. En la práctica, para un número par de datos, suponemos que la mediana se encontrará entre los dos valores centrales. Por tanto, en nuestro ejemplo, la mediana sería

6.5. La mediana puede tener valores idénticos con el suyo a la izquierda y a la derecha. Por ejemplo, en la serie 1, 2, 3, 4, 5, 5, 5, 6, 7, 8, 9, la mediana es 5.

Debido a estas características, puede definirse formalmente la *mediana* como aquel valor que divide una serie de tal forma que por lo menos 50 por 100 de los valores son iguales a él o menores que él, y por lo menos 50 por 100 de los valores son iguales o mayores que él. O bien, la mediana de una colección de datos ordenados en orden de magnitud es el valor medio o media aritmética de los dos valores centrales.

Característica de la mediana. Una característica sobresaliente de la mediana es su insensibilidad hacia las clasificaciones extremas. Considere el siguiente conjunto de calificaciones: 2, 5, 8, 11, 48. La mediana es 8.
Esto es verdad, aunque el conjunto tiene una calificación extrema de 48. Si en lugar de 48 tuviésemos una calificación de 97, la mediana seguiría siendo la misma.

LA MODA

La *moda*, denotada por Mo, es aquel valor de una serie de datos que aparece más frecuentemente que cualquier otro. Este valor puede ser descubierto inmediatamente cuando se ordenan los datos. Por ejemplo, en la serie 1, 2, 4, 4, 5, 6, y 7, la moda es 4. Por consiguiente, podemos considerar la moda como típica en el sentido en el sentido de que es el valor más "probable" de una serie. La moda para una serie de datos no agrupados siempre coincide con un valor real en la serie.

Aunque la moda es un concepto sencillo y útil, su aplicación presenta muchos aspectos engorrosos. Primero, una distribución puede revelar que dos o más valores repiten un número igual de veces, y en tal situación no hay forma lógica de determinar qué valor debe ser escogido como la moda. Hablando en sentido riguroso, cualquier valor se llama moda si aparece más a menudo que cualquiera de los valores adyacentes. Sin embargo, mientras las frecuencias de los valores modales no sean iguales, podríamos decidir escoger el valor con la frecuencia más alta como la moda para la serie.

Segundo, puede que no hallemos ningún valor que aparezca más de una vez.
Tercero, la moda es un valor muy inestable. Puede cambiar radicalmente con el método de redondeo de los datos.

Finalmente, la moda podría ser un valor extremo, como en el caso de una distribución triangular (una distribución en la que la densidad de frecuencias disminuye o aumenta, continuamente y a un ritmo de izquierda a derecha), y entonces difícilmente podría ser considerada como una medida de tendencia central.

Ejemplo 1. Las serie de datos 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18 tiene de moda 9.
Ejemplo 2. La serie 3, 5, 8, 10, 12, 15, 16 no tiene de moda.
Ejemplo 3. La serie 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9 tiene dos modas, 4 y 7, y se dice que es un conjunto de datos bimodal.

LA MEDIA ARITMÉTICA

La media aritmética, por su facilidad de cálculo, largo uso y propiedades matemáticas convenientes es promedio mejor conocido y de uso más común. A veces, se conoce sencillamente como "la media" o el "promedio", pero deben usarse siempre adjetivos apropiados cuando el contexto incluye varios tipos de medias.

La *media aritmética*, representada por $\bar{x}$, es la suma de los valores individuales de una muestra dividida por el número de observaciones de la muestra:

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

Propiedades de la media aritmética.

- Primero, es un valor típico porque es el centro de gravedad un punto de equilibrio. También es típica porque su valor puede substituir al valor de cada dato de la serie sin cambiar el total.
- La suma algebraica de las desviaciones con relación a la media es cero. Esto es, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.
- La tercera propiedad de la media aritmética es que la suma de las desviaciones elevada al cuadrado de los datos respecto a la media es menor que la suma de las desviaciones elevada al cuadrado de cualquier otro punto.

CUARTILES, DECILES Y PERCENTILES

Si una serie de datos se colocan en orden de magnitud, el valor medio (o la media aritmética de los dos valores medios) que divide el conjunto de datos en dos partes iguales es la mediana. Por extensión, de esta idea se puede pensar en aquellos valores que dividen a los datos en cuatro partes iguales. Estos valores, representados por Q_1, Q_2 y Q_3 se llaman primero, segundo y tercer *cuartil*, respectivamente; el valor de Q_2 es igual al de la mediana.

Análogamente los valores que dividen los datos en diez partes iguales se llaman *deciles* y se representan por D_1, $D_2, ..., D_9$, mientras que los valores que dividen los datos en cien partes iguales se llaman *percentiles* y se representan por $P_1, P_2, ..., p_{99}$ El quinto decil y el quincuagésimo percentil, se corresponden con la mediana. Los percentiles P_{25} y P_{75} se corresponden con el primer y tercer cuartil respectivamente.

ACTIVIDAD DE APRENDIZAJE

Calcule las medidas de tendencia central que se piden en los siguientes ejercicios.

1. Calcule la mediana, la moda y la media aritmética de las calificaciones del Cuadro 1.

2. Un artículo publicado en 1988 en una revista especializada describe el cálculo de los coeficientes de arrastre para la superficie aerodinámica NASA 0012. Para ello se utilizaron diferentes algoritmos computacionales con $M_\infty = 0.7$, obteniéndose los siguientes resultados (los coeficientes de arrastre están dados en unidades de conteos de arrastre; esto es equivalente a un coeficiente de arrastre de 0.0001): 79, 100, 74, 83, 81, 85, 82, 80 y 84. Calcule

   a). La media muestral

   b). La mediana muestral

3. Las siguientes mediciones corresponden a las temperaturas de un horno registradas en lotes sucesivos de un proceso de fabricación de semiconductores (las unidades son ºF): 953, 950, 948, 955, 951 949, 957, 954, 955. Calcule:

   a). La media muestral de estos datos.

   b). La mediana muestral de estos datos.

   c). ¿En cuánto puede incrementarse la mayor medición de temperatura sin que cambie la mediana muestral?

4. Haga un cuadro comparativo de las ventajas y desventajas de la mediana, la moda y la media aritmética.

MEDIDAS DE VARIABILIDAD

Al caracterizar una población por el estudio de un atributo, no sólo es necesario conocer el valor alrededor del cual tienden a presentarse con más frecuencias los valores de *x*, sino también el grado de dispersión de estos valores.

Un "promedio" sin salvedades puede carecer virtualmente de significado. Un factor que aumenta la confusión es que con algunas distribuciones todos los promedios importantes están estrechamente reunidos, mientras que con otras están muy separados. En otras palabras, un promedio puede ser muy engañoso, a menos que sea identificado y vaya acompañado de otra información que nos diga la amplitud de cosas o sus desviaciones con relación al promedio.

EL RANGO O AMPLITUD

El rango[1] es la diferencia entre la puntuación mayor y la puntuación menor. Indica el número de unidades necesarias en la escala de medición para incluir los valores máximo y mínimo. Se calcula así: $x_M - x_m$ (puntuación mayor menos puntuación menor). También se le puede llamar "amplitud" o "recorrido".

Ejemplo. Si tenemos los siguientes valores: 17, 18, 20, 20, 24, 28, 28, 30, 33. El rango será: $33 - 17 = 16$.

Cuanto más grande sea el rango, se espera mayor dispersión de los datos.

DESVIACIÓN MEDIA ABSOLUTA

La búsqueda de una medida de variabilidad que tome en cuenta todos los valores observados y que caracterizaría la dispersión de los valores individuales partiendo de la tendencia central nos conduce a la idea de calcular una medida como:

$$\sum_{i=1}^{n}(x_i - \bar{x})$$

pero esta medida será siempre igual a cero, porque

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\frac{\sum_{i=1}^{n} x_i}{n} - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

y, por tanto, difícilmente puede ser considerada como una medida de algo.

Una forma obvia de superar la dificultad es hallar una media de las desviaciones ignorando la dirección y el signo algebraico correspondiente. Al hacerlo así, obtendríamos lo que se llama la *desviación media absoluta*, o simplemente la desviación media, de la muestra. La desviación media absoluta, *dm*, de una serie de datos está dada por:

$$dm = \sum_{i=1}^{n} |x_i - \bar{x}|$$

La desviación media es útil para tratar situaciones en las que no se requiere un análisis minucioso.

[1] Dependiendo del contexto, el término "rango" tiene significados diferentes. En Álgebra lineal, por ejemplo, el rango de una matriz corresponde al número de sus columnas linealmente independientes; en Funciones, se le llama rango al contradominio (también llamado imagen); el rango de una variable aleatoria, está formado por el conjunto de sus posibles valores. Cuando se habla del rango de un conjunto de números se hace referencia a la distancia que hay entre el valor más grande y el más pequeño. El término rango proviene de una traducción literal de la palabra "range" del idioma inglés; aunque puede, tal vez, usarse de forma más apropiada la palabra **amplitud** para referirse a esta medida descriptiva, el término "rango" es ampliamente aceptado.

LA VARIANZA Y LA DESVIACIÓN ESTÁNDAR

El promedio de las desviaciones al cuadrado de la media se llama *varianza* muestral, designada por s^2. Simbólicamente,

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

donde $n - 1$ se llama "n − 1 grados de libertad".

El valor de la varianza, desde el punto de vista práctico, es un poco complicado de entender, porque las unidades asignadas a ella son cuadradas, tales como *metros*^{2}, *kg*^{2}, *personas*^{2}, etc. Para convertir esta medida de variabilidad en unidades originales, podemos tomar la raíz cuadrada (positiva) de s^2, obteniendo la *desviación estándar* de una muestra. La desviación estándar sirve como medida básica de variabilidad.

La desviación estándar, denotada por *s*, está dado por:

$$s = \sqrt{\text{var}\,ianza} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Fórmula de operación de la varianza. Falta agregar demostración algebraica.

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

La fórmula de operación de la varianza puede escribirse también como (sólo multiplicamos la fórmula anterior por n tanto en el numerador como en el denominador):

$$s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$$

COEFICIENTE DE VARIACIÓN

A menudo nos interesamos por comparar las variabilidades entre dos o más conjuntos de datos. Puede hacerse esto fácilmente con sus respectivas varianzas o desviaciones estándares cuando las variables se dan en las mismas unidades y cuando sus medias son

aproximadamente iguales. Cuando faltan están condiciones, puede que deseemos usar alguna medida relativa de dispersión. Una medida relativa de variabilidad frecuentemente usada se llama *coeficiente de variación,* designado *cv*, que es simplemente la razón de la desviación estándar a la media:

$$cv = \frac{s}{\overline{x}}$$

ACTIVIDAD DE APRENDIZAJE

1. Con los datos de las calificaciones del Cuadro 1, calcule:
   a). el rango
   b). la varianza
   c). la desviación estándar
   d). la desviación media absoluta
   e). el coeficiente de variación

2. Con los datos de inciso 2 de la Actividad de aprendizaje de la página 11, calcule:
   a). el rango
   b). la varianza
   c). la desviación estándar
   d). la desviación media absoluta
   e). el coeficiente de variación

4 Construcción de tablas de frecuencias (Datos agrupados)

- Construir tablas de distribución de frecuencias a partir de un conjunto de observaciones sin procesar (en bruto).
- Determinar el número de clases en una tabla de frecuencias.
- Calcular las frecuencias absoluta y relativa, las frecuencias absoluta y relativa acumuladas y la marca de clase (o punto medio).

En esta sección se presentan las características de las tablas de frecuencias, tanto por dato como por intervalos de clase. A partir de un conjunto de datos no procesados y con base en criterios definidos se puede construir una tabla de distribución de frecuencias. Sobre las tablas de frecuencias se pueden calcular medidas descriptivas y de dispersión así como generar gráficas que facilitan la interpretación de la información que contienen. El cálculo de estas medidas y las gráficas se revisan en las siguientes secciones.

Reescribimos los datos del Cuadro 1, de las calificaciones, que utilizaremos en esta sección.

Cuadro 1. Calificaciones obtenidas por 50 alumnos en un examen de Probabilidad

71	52	58	60	66	67	91	70	75	83
88	89	82	93	72	71	61	74	76	61
57	64	62	74	64	77	87	62	85	80
68	76	80	82	31	85	62	97	72	69
57	87	73	72	79	84	81	79	81	73

Si se desea, a partir de las calificaciones del Cuadro 1, resumir la información sobre el número de estudiantes que obtuvieron una determinada calificación, esto puede hacerse en una tabla de frecuencias por dato o por intervalos de clase.

Tablas de frecuencias por dato (datos ordenados)

Los datos en el Cuadro 2 es un ejemplo de distribución de frecuencias por dato (se le conoce también como datos ordenados o tablas de dos columnas). A cada calificación se le asocia una frecuencia, que en este ejemplo, es el número de estudiantes que obtuvieron una calificación.

Por regla general, como en el Cuadro 2, los datos (en este caso las calificaciones) se ordenan de menor a mayor, aunque alguien puede decidir hacerlo de forma descendente.

Frecuencia de dato. Es el número de veces que un dato se repite en una colección. Para el ejemplo que se sigue, "frecuencia de dato" se refiere al número de veces que una determinada calificación aparece entre las 50 reportadas.

La utilidad de las tablas de *distribución de frecuencias por dato* es máxima cuando el número de datos es pequeño y se escribe algún texto que indique el tipo de observación referenciado.

Cuadro 2 Frecuencias de calificaciones obtenidas por estudiantes del Tecnológico en un examen ordinario del curso de Probabilidad.

Calificación	Número de estudiantes	Calificación	Número de estudiantes
31	1	75	1
52	1	76	2
57	2	77	1
58	1	79	2
60	1	80	4
61	2	81	2
62	2	82	2
64	2	83	1
66	1	84	1
67	1	85	1
68	1	87	2
69	1	88	1
70	1	89	1
71	2	91	1
72	3	93	1
73	2	97	1
74	2		
		Total	50

Del cuadro 2, podemos observar lo siguiente:

- La calificación mínima obtenida fue de 31 puntos.
- La calificación máxima obtenida fue de 97 puntos.
- Nadie obtuvo 63 ni 78 puntos.
- 18 estudiantes, de los de 50, obtuvieron una calificación igual o mayor que 80.
- 16 estudiantes se negaron a aprobar, es decir, obtuvieron menos de 70 puntos.

Tablas de frecuencias por intervalo de clases

Cuando el valor del rango es grande, es posible que se tenga un gran número de datos diferentes lo que impediría resumir la información adecuadamente, como es deseable. En estos casos puede ser más conveniente agrupar los datos en categorías conocidos como intervalos de clase; a cada intervalo de clase se le asocia un número llamado "la frecuencia del intervalo de clase" que se define como "el número de observaciones que pertenecen a la clase". La frecuencia del *k*-ésimo intervalo de clase lo denotaremos con f_k.

Pasos para construir una tabla de distribución de frecuencias

a) Encuentre la diferencia de los valores máximo y mínimo de la serie de datos (rango o amplitud).
b) Elija un número de intervalos de igual tamaño que cubra el rango entre el mínimo y el máximo de los datos. Éstos son llamados *intervalos de clase*, y sus puntos

extremos se conocen como *límites de clase* [*límite de clase inferior* (LCI) y *límite de clase superior* (LCS)]

c) La diferencia entre el límite superior y el límite inferior de la clase se conoce como *anchura o tamaño de clase* (LCS – LCI).

d) Cuente el número de observaciones en los datos que corresponde a cada intervalo de clase. La cantidad de observaciones en cada clase se conoce como *frecuencia de clase*. La frecuencia del *k*-ésimo intervalo de clase la denotaremos por f_k.

e) Determine la frecuencia relativa de cada clase, (p_k), dividiendo la frecuencia de clase entre el número total de observaciones. En símbolos $p_k = f_k / n$

f) Cálculo de la *marca de clase* o punto medio de clase (m_k). Si l_k es el límite inferior de la *k*-ésima clase y L_k el límite superior, entonces. $m_k = \frac{l_k + L_k}{2}$

g) Frecuencias acumuladas. La frecuencia acumulada de la *k*-ésima clase es la suma de las frecuencias de las primeras *k* clases. Aplica tanto para las frecuencias absolutas como para las relativas. Por ejemplo, para la clase 3, la frecuencia absoluta acumulada es $f_1 + f_2 + f_3$; su frecuencia relativa acumulada es $p_1 + p_2 + p_3$

h) Debe observarse que la suma de todas las frecuencias absolutas da como resultado el número de observaciones (*n*) y la sumatoria de las frecuencias relativas es 1.

¿Cuántas clases tiene una tabla de distribución de frecuencias?

El número de clases depende del número de observaciones y de la dispersión de los datos. En general resulta satisfactorio usar entre 5 y 20 clases, y que el número de clases aumente en función del número total de observaciones, representado por *n*. En la práctica, puede usarse aproximadamente la raíz cuadrada del número de datos ($\sqrt{n}$) para definirse el número de clases a usarse.

Si hay menos clases que las necesarias, la pérdida de información es seria. Si hay muchas clases y la serie de datos es pequeña, las frecuencias entre celdas tienden a subir y a bajar de una manera caótica y no produce un modelo de distribución de los datos. Como un paso inicial, las frecuencias pueden ser determinadas con un número grande de intervalos que pueden después ser combinados con el deseo de obtener una elección de modelo de distribución más visible (Montgomery & Runger, 1996).

¿Dónde empieza y dónde termina un intervalo de clase?

Diremos que un número pertenece a una determinada clase si su valor numérico es mayor o igual que el límite inferior y estrictamente menor que el límite superior. Usando la notación de intervalos, estudiada en el curso **Matemáticas I**, tenemos que:

$$[a,b) = \{x / a \le x < b\}, \quad \text{con } a < b$$

En el Cuadro 3, la notación $[a,b)$ indica que la clase incluye números mayores o iguales que *a* y estrictamente menores que *b*, tal como lo manejan Montgomery & Runger (1996). La notación de intervalos tiene la ventaja de que no hay que preocuparse por establecer los

límites reales y nominales de los intervalos de clase, como a menudo se hace, para calcular las medidas descriptivas sobre la tabla de frecuencias; para tal efecto, sólo nos interesará identificar los extremos (o fronteras) inferior y superior de cada clase. Además, la notación de intervalos asegura que cada observación de la base de datos quede, sin ambigüedad, en sólo una clase, independientemente de la aproximación usada en la medición de los datos.

EJEMPLO. Con base en los datos del Cuadro 1, se construye la distribución de frecuencias siguiendo los pasos enumerados arriba.

Cuadro 3. Distribución de frecuencias por intervalos de clase de las calificaciones del Cuadro 1.

Intervalo de clase	Marca de clase	Frecuencia Absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
[30,40)	35	1	1/50	1/50	1/50
[40,50)	45	0	0/50	1/50	1/50
[50,60)	55	4	4/50	5/50	5/50
[60,70)	65	11	11/50	16/50	16/50
[70,80)	75	16	16/50	32/50	32/50
[80,90)	85	15	15/50	47/50	47/50
[90,100)	95	3	3/50	50/50	50/50
Total		$\sum_{k=1}^{7} f_k = 50$	$\sum_{k=1}^{7} p_k = 1$		

EJEMPLO

(*Construcción de una tabla de distribución de frecuencias*)

Los datos del Cuadro 4 representan la resistencia a la tensión, en libras por pulgada cuadrada (psi), de 80 muestras de una nueva aleación de litio y aluminio, que está siendo evaluada como posible material para la fabricación de elementos estructurales de aeronaves.

Cuadro 4 Resistencia a la tensión de 80 muestras de aleación de aluminio-litio.

105	224	183	186	124	181	180	145
97	154	153	174	120	168	167	144
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Fuente: Montgomery & Runger (1996, p.5)

Con los datos anteriores, construya una tabla de frecuencias. Utilice como límite inferior del primer intervalo de clase, el valor 70 y como límite superior del último intervalo, 250; con tamaño de clase de 20 para cada uno de los intervalos.

Cuadro 5. Distribución de frecuencias a partir de los datos del Cuadro 4.

Intervalo de clase	Marca de clase (m_k)	Frecuencia absoluta (f_k)	Frecuencia relativa (p_k)	Frecuencia relativa acumulada (P_K)	Frecuencia absoluta acumulada (F_K)
[70,90)	80	2	2/80	2/80	2
[90,110)	100	3	3/80	5/80	5
[110,130)	120	6	6/80	11/80	11
[130,150)	140	14	14/80	25/80	25
[150,170)	160	22	22/80	47/80	47
[170,190)	180	17	17/80	64/80	64
[190,210)	200	10	10/80	74/80	74
[210,230)	220	4	4/80	78/80	78
[230,250)	240	2	4/80	80/80	80

ACTIVIDAD DE APRENDIZAJE

Parte A.

Con la siguiente información, construya una tabla de distribución de frecuencias.

Uno de los mayores indicadores de contaminación del aire en las grandes ciudades y en los cinturones industriales es la concentración de ozono en la atmósfera. Las 78 observaciones del cuadro siguiente fueron recolectadas por las autoridades de Los Angeles, sobre la concentración de ozono en esa localidad durante los veranos de 1996 y 1997. Cada observación es un promedio de lecturas tomadas cada cuarto día.

Cuadro 6. Medidas de concentración de ozono en la atmósfera de la localidad de Los Ángeles durante los veranos de 1966 y 1967.

3.5	1.4	6.6	6.0	4.2	4.4	5.3	5.6
6.8	2.5	5.4	4.4	5.4	4.7	3.5	4.0
2.4	3.0	5.6	4.7	6.5	3.0	4.1	3.4
6.8	1.7	5.3	4.7	7.4	6.0	6.7	11.7
5.5	1.1	5.1	5.6	5.5	1.4	3.9	6.6
6.2	7.5	6.2	6.0	5.8	2.8	6.1	4.1
5.7	5.8	3.1	5.8	1.6	2.5	8.1	6.6
9.4	3.4	5.8	7.6	1.4	3.7	2.0	3.7
6.8	3.1	4.7	3.8	5.9	3.3	6.2	7.6
6.6	4.4	5.7	4.5	3.7	9.4		

Comentario: Para construir una distribución de frecuencias, primero debemos notar que las lecturas mínima y máxima son 1.1 y 11.7, respectivamente; por lo que se sugiere los siguientes intervalos de clase.

Cuadro 7. Distribución de frecuencia de los datos en el Cuadro 6

Intervalo de clase	Conteo	Marca de clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
[1.0, 2.0)						
[2.0, 3.0)						
[3.0, 4.0)						
[4.0, 5.0)						
[5.0, 6.0)						
[6.0, 7.0)						
[7.0, 8.0)						
[8.0, 9.0)						
[9.0, 10.0)						
[10.0, 11.0)						
[11.0, 12.0)						
TOTAL			78	1		

Comentario: En una distribución de frecuencias, las frecuencias relativas siempre deben sumar 1; las frecuencias absolutas, el número de datos.

Parte B.

Con los datos agrupados en la tabla de distribución de frecuencias anterior, calcule la media aritmética y la varianza. (Le puede auxiliar la siguiente tabla)

Cuadro 8

Intervalo de clase	m_k	f_k	$m_k f_k$	$(m_k - \bar{x})^2$	$(m_k - \bar{x})^2 f_k$	$m_k^2 f_k$	
[1.0, 2.0)							
[2.0, 3.0)							
[3.0, 4.0)							
[4.0, 5.0)							
[5.0, 6.0)							
[6.0, 7.0)							
[7.0, 8.0)							
[8.0, 9.0)							
[9.0, 10.0)							
[10.0, 11.0)							
[11.0, 12.0)							
TOTAL							

5 Cálculo de medidas descriptivas en tablas de frecuencias

- Calcular las siguientes medidas de tendencia central en tablas de frecuencias: media aritmética, moda y mediana.
- Calcular las siguientes medidas de dispersión en tablas de frecuencias: varianza y desviación estándar.

Las tablas de frecuencias nos proveen tendencias o patrones del comportamiento de un conjunto de datos. Podemos usar la tabla de frecuencias para calcular las medidas descriptivas o sumarias, como las de tendencia central o las de dispersión. Básicamente tenemos necesidad de calcular estas medidas descriptivas sobre la tabla de frecuencias cuando no tenemos acceso a los datos originales como el caso de los datos reportados en las publicaciones académicas o de investigación.

Por ejemplo, podemos determinar la media y la desviación estándar, para los datos del Cuadro 9, correspondientes a la distribución de frecuencia de la edad de 200 hombres casados.

Cuadro 9
Distribución de frecuencia de la edad de 200 estudiantes casados

Intervalos de clase (edad en años)	Marca de clase m_k	Frecuencia f_k
[15, 20)	17.5	18
[20, 25)	22.5	74
[25, 30)	27.5	62
[30, 35)	32.5	26
[35, 40)	37.5	20
Total	----------	200

CÁLCULO DE LA MEDIA ARITMÉTICA

Esta tabla muestra que 18 observaciones se encuentran en el intervalo 15-20, 74 observaciones están localizadas en el intervalo 20-25, y así sucesivamente. Sin saber la exacta posición de las observaciones en cada intervalo, estimamos que todas las observaciones están localizadas en la mitad. El punto medio de las clases están dadas en la segunda columna de la tabla. De acuerdo a esto, con esto la observación 17 es repetida 18 veces, 22 es repetida 74 veces, y así sucesivamente. La suma de estas observaciones puede ser calculada como

$$(17.5 \times 18) + (22.5 \times 74) + (27.5 \times 62) + (32.5 \times 26.5) + (37.5 \times 20)$$

Podemos dividir este total entre $n = 200$ y así obtener la media muestral $\bar{x}$. Los cálculos se muestran en el Cuadro 10.

Cuadro 10

Cálculo de la $\bar{x}$ y *s* para la distribución de frecuencias del Cuadro 9

Intervalo de clase (edad)	Marca de clase m_k	Frecuencia f_k	$m_k f_k$	$m_k^2 f_k$
[15, 20)	17.5	18	315	
[20, 25)	22.5	74	1,665	
[25, 30)	27.5	62	1,705	
[30, 35)	32.5	26	845	
[35, 40)	37.5	20	750	
Total	---	200	5,280	

Esto es, la media aritmética es $\bar{x} = \frac{5280}{200} = 26.4$

A continuación se establece la fórmula de la media aritmética para datos agrupados:

Media muestral para datos agrupados

Si la distribución de frecuencias tiene *r* intervalos de clase con puntos medios $m_1, m_2, \ldots, m_r$ y correspondientes frecuencias $f_1, f_2, \ldots, f_r$, entonces la media aritmética está dada por:

$$\bar{x} = \frac{\sum_{k=1}^{r} m_k f_k}{n}$$

CÁLCULO DE LA VARIANZA MUESTRAL

Supongamos que todas las observaciones están localizadas en el punto medio del intervalo. El cuadrado de las desviaciones $(m_1 - \bar{x})^2, (m_2 - \bar{x})^2, \ldots, (m_k - \bar{x})^2$, repetidas con sus respectivas frecuencias $f_1, f_2, \ldots, f_r$. La suma del cuadrado de las desviaciones es:

$$\sum_{k=1}^{r} (m_k - \bar{x})^2 f_k$$

que dividida entre *n* - 1 obtenemos la varianza de la muestra s^2. Puesto que la frecuencia total *n* es usualmente grande para una serie de datos, dividiendo la suma del cuadrado de las desviaciones entre *n* - 1 es equivalente a dividirla entre *n*, obteniendo la siguiente fórmula:

Fórmula de la varianza para tablas de frecuencias (datos agrupados)

$$s^2 = \frac{\sum_{k=1}^{r} (m_k - \bar{x})^2 f_k}{n}$$

Como hicimos con la varianza para datos no-agrupados, podemos obtener una fórmula alternativa para s^2, expandiendo $(m_k - \bar{x})^2$ en la expresión anterior y reduciendo términos semejantes. La fórmula de operación para el cálculo de la varianza en datos agrupados es:

$$s^2 = \frac{\sum_{k=1}^{r} m_k^2 f_k}{n} - \bar{x}^2$$

Esta fórmula de operación de la varianza puede utilizarse auxiliándose de la columna 5 de la tabla 10.

Como siempre, la desviación estándar se obtiene con la raíz cuadrada positiva de la varianza.

ACTIVIDAD DE APRENDIZAJE

1. Calcule la media y la desviación estándar de la edad de los estudiantes casados con la distribución de frecuencia dada en el Cuadro 9. Utilice el Cuadro 10 para auxiliarse en los cálculos.

2. Calcule la media y la desviación estándar en la distribución de frecuencias de las calificaciones del Cuadro 1.

clase	m_k	f_k			
[30,40)	35	1			
[40,50)	45	0			
[50,60)	55	4			
[60,70)	65	11			
[70,80)	75	16			
[80,90)	85	15			
[90,100)	95	3			
Total		$\sum_{k=1}^{7} f_k = 50$			

6 Métodos gráficos

Una gráfica dice más que mil palabras

- Describir algunos métodos gráficos.
- Identificar el tipo de gráfica en aplicaciones específicas.
- Construir gráficas a partir de una tabla de frecuencias o sobre datos en bruto (o sin procesar)
- Explicar el significado práctico y la información que resumen las gráficas que se publican en diferentes medios como periódicos, revistas y libros.

ACTIVIDAD DE APRENDIZAJE

Parte A. (Realizar en binas con entrega individual en el salón)
El instructor proporciona a las binas 2 recortes recientes de periódico que contenga algún tipo de gráfica. Para cada recorte:

1. Escriba el tipo de gráficas que aparecen.

2. Enumere los principales puntos del tópico que aborda.

3. Con base en la información contenida en las gráficas del recorte, escriba un ensayo entre media y una cuartilla.

Parte B. (Realizar y entregar individual en horario extraclase)

4. Investigue las ventajas de los métodos gráficos.

5. Con los datos de las calificaciones de probabilidad, construya un histograma, un polígono de frecuencias, una ojiva.

6. Busque recortes en revistas o periódicos que contenga información en esquemas gráficos y aplíqueles los puntos 1, 2 y 3 de esta Actividad. Los recortes originales debe adherirlos a su libreta. Nota: evite publicaciones de nota roja y tampoco recorte los libros de texto de hermanos menores.

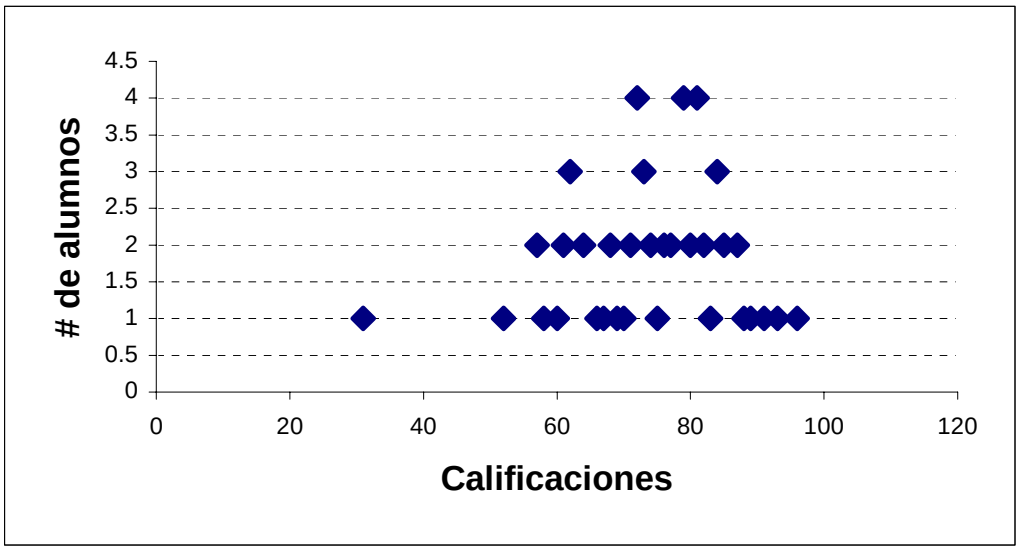
TERMINA LA ACTIVIDAD DE APRENDIZAJE

DIAGRAMA DE PUNTOS

El diagrama de puntos es una gráfica muy útil para visualizar un conjunto pequeño de datos; por ejemplo, 20 observaciones. La gráfica permite ver con rapidez y facilidad la ubicación o tendencia central de los datos, así como su dispersión o variabilidad. A menudo, los diagramas de puntos son útiles para comparar dos o más conjuntos de datos.

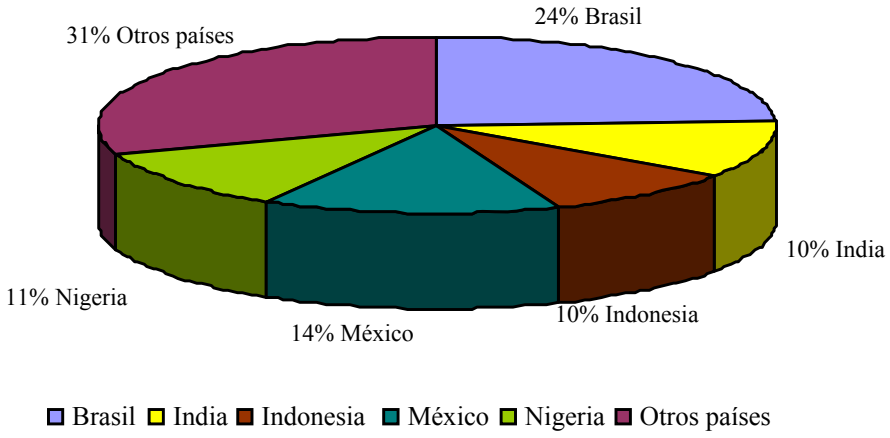
Si el número de observaciones es pequeño, a menudo es difícil identificar algún patrón de variación específico; sin embargo, con frecuencia el diagrama de puntos es útil y puede proporcionar información sobre características poco usuales de los datos. Cuando el

número de observaciones es moderadamente grande, pueden ser más útiles otros tipos de gráficas.



GRÁFICA DE PASTEL

(*Cultivo local e internacional de la papaya maradol*). Enseguida se ejemplifica una gráfica de pastel, para representar la producción mundial de la papaya maradol, con base en datos de la FAO.



HISTOGRAMA DE FRECUENCIAS RELATIVAS

Después de sumar los datos en la forma de distribución de frecuencias, puede ser presentada gráficamente a través de un *histograma de frecuencias relativas*, el cual es una representación visual del modelo de distribución.

Histograma de frecuencias relativas

Para dibujar un histograma de frecuencias relativas, los intervalos de clase son marcados en el eje horizontal de la gráfica. En cada intervalo es dibujado un rectángulo cuya área es equivalente a la frecuencia relativa del intervalo.

Altura de un rectángulo es igual a:

$$\frac{\text{Frecuencia relativa de una clase}}{\text{Base del intervalo de clase}}$$

Esto es porque el área de un rectángulo es igual a base por altura

El área de cada rectángulo en un histograma representa la proporción de las observaciones ocurridas en cada intervalo de clase. Además, el área total de todos los rectángulos en un histograma es 1. Lo convencional de usar el área de los rectángulos más que sus alturas. Las frecuencias relativas tiene distintas ventajas: instintivamente se compararan áreas cuando en dos partes de un histograma o dos diferentes histogramas. Cuando dos histogramas se basan en intervalos de clase de diferente base, la propiedad de tener un área total igual a 1 los hace comparables.

Ejemplo. Con referencia a los datos de concentración de ozono en el Cuadro 7, los histogramas correspondientes a la distribución de frecuencias en los cuadros 2, 3 y 4, son mostradas en las figuras 1: a, b, y c. En cada caso note que la altura de un rectángulo es obtenido dividiendo la frecuencia relativa entre la base del intervalo de clase. Por ejemplo, las alturas de los primeros rectángulos en las figuras 1: a, b, y c son 0.051 / 0.5 = 0.102, 0.09 / 1 = 0.09, y 0.09 / 2 = 0.045, respectivamente. El histograma en la figura 1 (a) parece delgado porque los intervalos de clase son muy cortos. Los intervalos hacen que el histograma de la figura 1: b y c sean más regulares y lisos en su forma, aunque más información es perdida entre más intervalos de clase son usados.

La regla de que los intervalos de clase sean iguales es inconveniente cuando los datos se extienden sobre determinado rango y se concentran altamente en un pequeña parte con relativamente pocos números en otras partes. Usar intervalos de clase más pequeños donde los datos son altamente concentrados y más grandes donde los datos están esparcidos ayuda a reducir la pérdida de información debido al agrupamiento. Tabulaciones de ingreso, edad, y otras características en reportes oficiales regularmente se hacen con intervalos de clase

diferentes. Cuando todos los intervalos no son iguales, el histograma puede ir de acuerdo a la convención de usar el área de los rectángulos para representar las frecuencias relativas de un modelo de distribución.

DIAGRAMA DE LÍNEA DE FRECUENCIAS RELATIVAS

Algunas veces los datos consisten en cantidades, como el número de niños en familia o el número de accidentes de tránsito por día, son observaciones sobre una escala continua. Si el número de los distintos valores no es demasiado grande, una distribución de frecuencias es construido usando los valores individuales como las clases en lugar de usar intervalos de clase. Los datos son luego presentados en forma de diagrama de líneas de frecuencias relativas.

Diagrama de línea de frecuencia relativa

Los distintos valores son colocados sobre el eje horizontal, Líneas verticales con alturas equivalentes a las frecuencias relativas son luego dibujadas con esos valores.

Las líneas reemplazan a los rectángulos acentuando que las frecuencias no son realmente extensiones sobre los intervalos. Un histograma puede también ser construido dibujando rectángulos centrados sobre los distintos valores de los datos, proveyendo las frecuencias relativas sobre esos puntos medios. Aunque ambos diagramas de línea y rectangulares son usados para contar datos, el diagrama de línea nunca debe ser dibujado para observaciones de escala continua.

POLÍGONO DE FRECUENCIAS

El polígono de frecuencias es una representación en línea se obtiene a partir del histograma. Su construcción se lleva a cabo uniendo los puntos medios de los datos superiores de los rectángulos del histograma.

OJIVAS

Si en lugar de frecuencias absolutas utilizamos sus correspondientes acumuladas obtendremos en vez del histograma, una representación gráfica en forma de línea creciente que se conoce con el nombre de *ojiva.*

ACTIVIDAD DE APRENDIZAJE

(Nota: En esta actividad, el docente puede seleccionar algunos de los siguientes ejercicios para que los estudiantes resuelvan de forma individual o por equipos en el salón, bajo su supervisión)

1. Se toman ocho mediciones del diámetro de los anillos para los pistones del motor de un automóvil. Los datos (en mm) son: 74.001, 74.003, 74.015, 74.000, 74.005, 74.002, 74.005 y 74.004. Construya un diagrama de puntos y haga comentarios al respecto.

2. Construya una distribución de frecuencias y un histograma para los datos siguientes, utilizando ocho clases.

Cuadro 12. Octanaje de varias mezclas de gasolina

88.5	87.7	83.4	86.7	87.5	91.5	88.6	100.3	96.5	93.3
94.7	91.1	91.0	94.2	87.8	89.9	88.3	87.6	84.3	86.7
84.3	86.7	88.2	90.8	88.3	98.8	94.2	92.7	93.2	91.0
90.1	93.4	88.5	90.1	89.2	88.3	85.3	87.9	88.6	90.9
89.0	96.1	93.3	91.8	92.3	90.4	90.1	93.0	88.7	89.9
89.8	89.6	87.4	88.4	88.9	91.2	89.3	94.4	92.7	91.8
91.6	90.4	91.1	92.6	89.8	90.6	91.1	90.4	89.3	89.7
90.3	91.6	90.5	93.7	92.7	92.2	92.2	91.2	91.0	92.2
90.0	90.7								

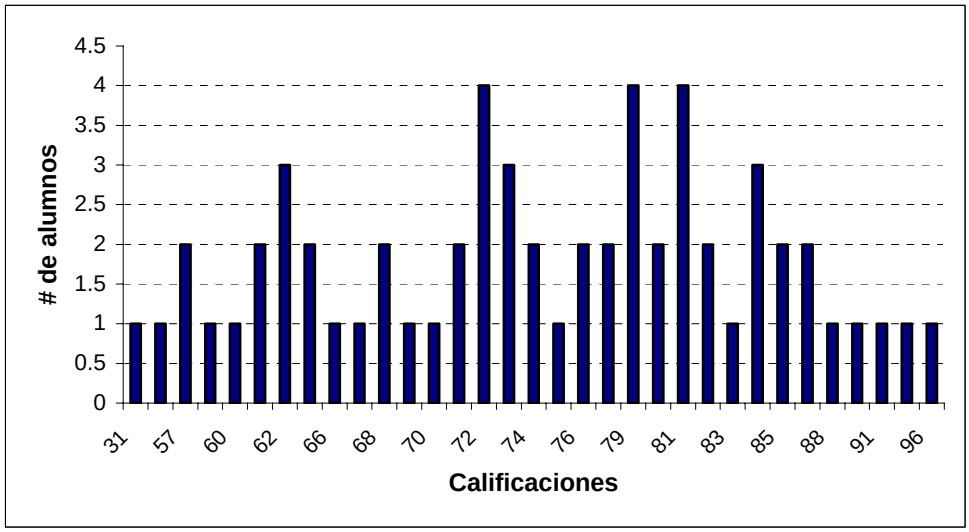
3. Los datos siguientes son mediciones de intensidad solar directa (en watts/m²) realizadas en una localidad del sur de México. Construya un histograma para estos datos.

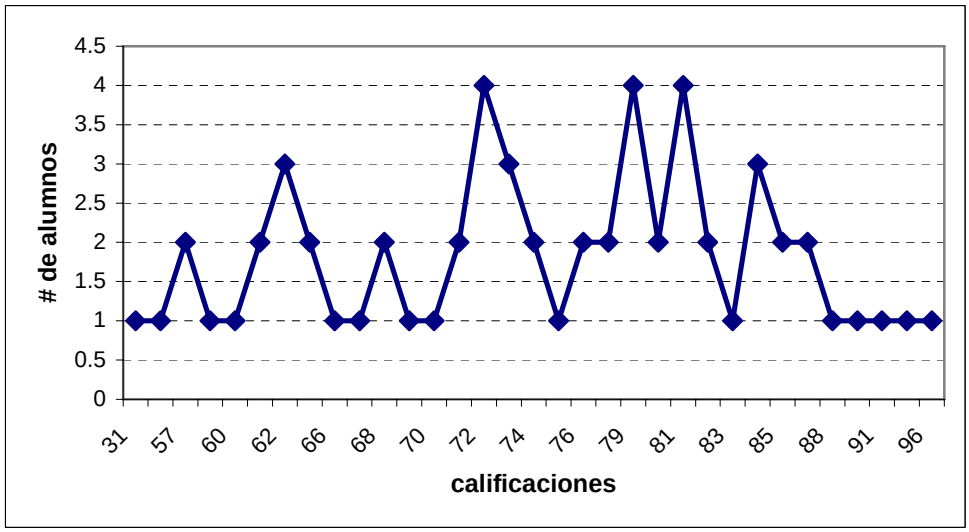
562	909	870	960	809	952	820
869	918	661	498	856	957	898
708	558	955	653	655	693	935
775	768	940	730	806	835	753
704	946	918	775	878	905	939

4. Los datos siguientes representan el rendimiento de 90 lotes consecutivos de un sustrato cerámico, en el que se ha aplicado un recubrimiento metálico mediante un proceso de depositación por vapor. Construya una distribución de frecuencias y un histograma para estos datos.

94.1	87.3	94.1	92.4	84.6	85.4
93.2	94.1	92.1	90.6	83.6	86.6
90.6	90.1	96.4	89.1	85.4	91.7
91.4	95.2	88.2	88.8	89.7	87.5
88.2	86.1	86.4	86.4	87.6	84.2
86.1	94.3	85.0	85.1	85.1	85.1
95.1	93.2	84.9	84.0	89.6	90.5
90.0	86.7	78.3	93.7	90.0	95.6
92.4	83.0	89.6	87.7	90.1	88.3
87.3	95.3	90.3	90.6	94.3	84.1
86.6	94.1	93.1	89.4	97.3	83.7
91.2	97.8	94.6	88.6	96.8	82.9
86.1	93.1	96.3	84.1	94.4	87.3
90.4	86.4	94.7	82.6	96.1	86.4
89.1	87.6	91.1	83.1	98.0	84.5

Gráfica de barras





REFERENCIAS

Devore, Jay. (1998). *Probabilidad y Estadística para Ingeniería y Ciencias*. 4ta edicion. México, Internacional Thomson Editores

Johnson, R. A. (1997). *Probabilidad y Estadística para Ingenieros de Miller y Freud*. 5ta edición. México, Prentice Hall Hispanoamericana.

Lipschutz, S. & J. **Schiller**. (2000). *Introducción a la Probabilidad y Estadística*. Serie Schaum. Madrid, McGraw Hill.

Mendenhall. Probabilidad y Estadística.

Montgomery, D. & G. **Runger**. (1996). *Probabilidad y Estadística Aplicadas a la Ingeniería*. México, McGraw Hill.

Levin, R. I. & D. S. **Rubin**. (1996). *Estadística para Administradores*. 6ta edición. México, Prentice Hall.

Torres León, R. (1999). *Introducción a la Probabilidad y Estadística*. Mérida,Yucatán, México, Ediciones de la Universidad Autónoma de Yucatán.

Walpole, R. E.; R. H. **Myers**; S. L. **Myers**. (2001). *Probabilidad y Estadística para Ingenieros*. 4ta edición. México, McGraw. (La sexta edición es de Prentice Hall Hispanoamericana)