Conferencia 1.

Tema 1. Introducción a Estadística. Estadística Descriptiva.

Sumario:

- Definiciones Fundamentales.
- Variables Cuantitativas y Cualitativas.
- Medidas de Tendencia Central.
- Medidas de Dispersión.
- Tablas de Distribución Empíricas de Frecuencia.
- Medidas de Posición.
- Gráficos

Bibliografía:

- Introductory Statistics Prem S. Mann, Christopher Jay Lacke, 2010 (Capítulos 1-3)
- Introductory Statistics. Barbara Illowsky, Susan Dean 2018 (Capítulos 1, 2).
- An Introduction to Categorical Data Analysis, Alan Agresti, 2007 (Apoyo para trabajo con datos categóricos)

Definiciones Fundamentales.

Realmente no es necesario hacer hincapié en la importancia de recolectar y analizar datos para tomar decisiones. En la actualizad es cosa de todos los días, incluso gigantes de la tecnología como Google y Facebook por solo citar dos, obtienen gran parte de sus ganancias gracias al análisis y venta de los datos de los usuarios.

Los estudios estadísticos tienen un amplio uso en la actualidad, desde el análisis de datos de pruebas de laboratorio para saber si un fármaco es efectivo, para saber qué productos colocar juntos a cuáles en que estantes y a que altura en la tienda para maximizar el consumo hasta ser capaces de predecir comportamientos y gustos. Por tanto, podemos plantear la siguiente definición.

Definición 1. Estadística: Es una rama de las matemáticas que engloba un conjunto de métodos usados para recolectar, analizar, presentar e interpretar datos para tomar decisiones.

Digamos entonces queremos saber si entre los estudiantes de la universidad tendrá aceptación un conjunto de opciones recreativas. Por supuesto si hablamos de los estudiantes de la facultad ofrecer un bailable de reggaetón, puede que no sea la mejor idea y tampoco lo seria ofrecer una simultánea de ajedrez a los estudiantes de Historia. Sin embargo, ambos estudiantes están en la categoría de ser estudiantes de la universidad de la habana. Si nos fijamos además en los estudiantes de cursos para trabajadores es posible que ninguna de las dos opciones anteriores sea de su agrado o quizás que las horas de las actividades no sean buenas para ellos. De aquí que presentamos las siguientes definiciones.

Definición 2. Población, Muestra: Una *población* es una colección específica de objetos de interés. Una *muestra* consiste en un subconjunto de la población, incluyéndose el caso de que sea la población total, este último lo denominamos *censo*.

Definición 3. Elemento: Un elemento o miembro de una muestra o de una población es un sujeto u objeto específico (persona, empresa, carro, estado, país, etc.) sobre la cual se recolecta la información.

Definición 4. Medición: Por tanto, una *medición* es un número o atributo calculado para cada miembro de una población o muestra. Todas las mediciones realizadas a una muestra se llaman *datos muestrales*.

Definición 5. Parámetro: Un parámetro es un número que resume algunos aspectos de una población.

Definición 6. Estadístico: Un estadístico es un número calculado a partir de los datos muestrales.

Continuando con el ejemplo anterior, supongamos que nos interesa saber la edad promedio con que se gradúan los estudiantes de carreras de computación de la habana. Podríamos revisar uno a uno los expedientes de todos los graduados de la UH, la UCI y la CUJAE y anotar con qué edad se graduaron y luego calcular el promedio. Este procedimiento es infactible, porque de la facultad solo el curso pasado hubo casi 100 graduados y la UCI tiene graduaciones masivas. Tenga en cuenta que no necesariamente la edad fluctúa entre 21-22 años pues se puede hacer prueba de ingreso hasta los 30 años. Si a todo esto se le suma la cantidad de años que le toma aun estudiante graduarse, que no es la misma para estas universidades, más las repitencias sería complicado acotar la edad de graduación de los estudiantes, y aun acotando el número, lo que queremos es la edad promedio.

Por tanto, para realizar el cálculo tomaríamos una muestra ***representativa*** de egresados de estas carreras y promediamos sus edades, entonces sería razonable concluir que el valor calculado es la edad promedio de graduación de los estudiantes universitarios.

Un detalle clave en la forma de conducir el experimento es utilizar una ***muestra representativa proporcionada***, en nuestro caso la población son todos los estudiantes de la universidad (Incluye las facultades de todas las universidades del país, ciencias médicas, UCI, el pedagógico, la Cujae y todas las universidades de provincias más la de la habana). Para seleccionar una muestra representativa los sujetos seleccionados de cada universidad deben aparecer en la misma proporción que en la población. O sea, si la cantidad de egresados de la facultad de Ingeniería Informática de la Cujae supera a la cantidad de egresados de MATCOM 3 a 1 y la UCI supera a Ing. Informática 10 a 1 entonces por cada graduado de MATCOM en la muestra debe haber 3 de Ing. Informática y 10 de la UCI.

Definición 6. Muestra Representativa: Una muestra que representa las características de la población tan cercano a la realidad como sea posible.

Definición 7. Muestra Aleatoria: Una muestra seleccionada de forma tal que cada elemento de la población tiene la misma probabilidad de ser escogidos.

Digamos que nos interesa en este caso la edad promedio de los estudiantes que están en el aula recibiendo esta conferencia. Todos los presentes escriben en un papel su nombre y edad, lo ponen en una caja y de ahí se selecciona a 10 papeles al azar, así todos los estudiantes tienen la misma probabilidad de ser escogidos, y los diez seleccionados constituyen una **muestra aleatoria**.

Sin embargo, si se organizan los estudiantes de forma alfabética y se toman los 10 primeros entonces la probabilidad de escoger a los estudiantes del lugar 11 al n-ésimo es 0 y sería una muestra no aleatoria, porque no importa cuántas veces se realice la selección siempre se escogerán los mismos estudiantes.

Siendo más precisos, si al seleccionar al azar a los 10 estudiantes cada vez que selecciono uno anoto su edad y lo regreso a la caja la **muestra** seria **aleatoria con remplazo**, si en vez de regresarlo a la caja lo boto entonces es una **muestra aleatoria sin remplazo**.

La acción de calcular el valor promedio para resumir los datos es una operación de la estadística descriptiva, sin embargo, cuando se utiliza este número para sacar conclusiones acerca de la población a partir de este valor, entonces estamos hablando de la estadística inferencial.

Definición 8. Estadística Descriptiva: Es una rama de la estadística que involucra organizar, mostrar y describir datos.

Es la que comúnmente les dieron en el preuniversitario como estadística a secas y que por lo general incluye muchos más detalles.

Definición 9. Estadística Inferencial: Es la rama de las estadísticas con la que se llega a conclusiones acerca de una población basados en la información contenida en una muestra extraída de esa misma población.

Pero qué pasa si en vez de querer saber edad promedio queremos saber el sexo predominante en una muestra. Este tipo de datos no es numérico y muchas veces los datos recogidos en una muestra no lo son. Por tanto, tenemos dos tipos de datos:

Definición 10. Datos Cualitativos: Medias que consisten en atributos u otros caracteres no numéricos.

Definición 11. Datos Cuantitativos: Medias numéricas que nacen de una escala numérica natural.

Existen dos escalas de medidas para las variables en dependencia del tipo de variable y de la naturaleza de los datos.

<table>
<tr><td>Variables</td><td colspan="3">Escala de Medida</td></tr>
<tr><td rowspan="3">Cuantitativas</td><td colspan="2">Razón</td><td rowspan="2">Intervalo</td></tr>
<tr><td colspan="2">El 0 no tiene significado.</td></tr>
<tr><td>Continua
Ej: Altura, Talla, Peso</td><td>Discreta
Ej: Cantidad de Estudiantes en un grupo</td><td>Continua, con distancia predeterminada en ella.
Ej: IQ, Nota de un Examen.</td></tr>
<tr><td>Cualitativas</td><td colspan="2">Nominal
Una variable que se divide en categorías.
Ej.: Pelo puede ser rubio, rojo, negro, etc.</td><td>Ordinal
Las categorías en las que se divide la variable están sujetas a un tipo de orden
Ej.: Estado de Salud (B, R, M)</td></tr>
</table>

Tabla 1. Tipos de Variables y sus Escalas de Medida.

En el caso de las variables cuantitativas con escala de medida por intervalo la distancia da una idea particular de diferencia entre los elementos, por ejemplo, en el caso del Coeficiente intelectual si alguien tiene 120 es más inteligente que alguien que tiene 60, pero no podemos decir que el de 120 sea el doble de inteligente que el de 60. En el caso de la nota de un examen nos referimos a la nota en base a 100 La forma de ver la distancia es que mientras más puntos tengan en ese examen, más habilidades tienen en la asignatura que examinaron.

Existen Mediciones que pueden ser medidas usando diferentes tipos de variables. Por ejemplo, las notas de un examen pueden ser en base a 100 y serian una Variable Cuantitativa de Intervalo o pueden ser medidas en una escala de 2-5 y en ese caso serían Cualitativas Ordinales.

Existen características a la que pueden asociarse diferentes tipos de variables según la escala con la que se mida. Por ejemplo, la edad, podríamos asociarla a una variable cuantitativa (la edad exacta), una variable cualitativa ordinal (las clases de edad) o una variable cualitativa nominal (jóvenes y viejos).

Los datos constituyen toda la información que obtenemos a partir de una encuesta o estudio, podemos usarlos de dos vías, en su forma simple o agrupada. Los usaremos de forma simples cuando $n \leq 15$. De otra forma los agrupamos en intervalos de clase usando tablas de Distribución Empíricas de Frecuencia y los analizamos teniendo en cuenta las Medidas de Tendencia Central, Dispersión y Posición.

Cálculo de medidas de tendencia central y Dispersión.

Medidas de Tendencia Central

Media Aritmética:

Es el promedio de los datos.

Agrupados:

Cualitativas $$\bar{X} = \frac{1}{n}\left(\sum_{i=1}^{n} V_i n_i\right)$$

Cuantitativas. $$\bar{X} = \frac{1}{n}\left(\sum_{i=1}^{n} MC_i n_i\right)$$

Simples:

$$\bar{X} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

Moda:

Puede que la muestra no tenga moda. Para que esto suceda todos los valores son únicos. Ejemplo (1,2,3,4,5)

Agrupados.

1. Primero hay que hallar la clase modal. La clase modal es la de mayor frecuencia absoluta n_j.
2. Si hay más de una clase modal se selecciona una moda por cada clase modal.
3. La fórmula para calcular la moda es la siguiente:

$$M_o = L_{j-1} + \frac{n_j - n_{j-1}}{(n_j - n_{j-1}) + (n_j - n_{j+1})} a_j$$

- Donde j es la clase modal,
- L_{j-1} límite inferior del intervalo de la clase modal,
- a_j es la amplitud,
- n_j es la frecuencia de la clase modal,
- n_{j-1} es la frecuencia de la clase anterior a la modal,
- n_{j+1} es la frecuencia de la clase posterior a la modal.

Simples.
Tomamos el o los valores que más se repiten. En caso de que se tengan dos valores decimos que la muestra es bimodal, tres modas es trimodal, etc.

Mediana:
Es el valor que se encuentra en el 50% de la muestra. Siempre existe la mediana.
Agrupados.
Para hallar la clase de la mediana tenemos que encontrar la clase que cumpla que $N_j > \frac{n}{2}$.

$$M_e = L_{j-1} + \frac{\frac{n}{2} - N_{j-1}}{N_j - N_{j-1}} a_j$$

- L_{j-1} límite inferior del intervalo de la clase de la mediana,
- a_j es la amplitud,
- N_j es la frecuencia absoluta acumulada hasta j,
- N_{j-1} es la frecuencia absoluta acumulada hasta j – 1,
- N_{j+1} es la frecuencia absoluta acumulada hasta j + 1.

Simples.
Ordenamos los datos y escogemos el del medio. En caso de que n sea par tomamos los dos del medio los sumamos y los dividimos entre dos.

Medidas de dispersión.

Varianza:
Simples.

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

Agrupados:

Cuantitativas $$\sigma^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} n_i(MC_i - \bar{X})^2\right)$$

Cualitativas $$\sigma^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} n_i(V_i - \bar{X})^2\right)$$

Desviación Típica: $\sigma = \sqrt{\sigma^2}$

Coeficiente de Variación $CV = \frac{\sigma}{\bar{X}} = \frac{DT}{\bar{X}}$

Construcción de Tablas de Distribución Empírica de Frecuencia.

Para variables Continuas, Cuantitativas.

La cantidad de datos recogidos es n. Para construir la TDEF Primero debemos tener en cuenta la cantidad de Intervalos de Clase para formar la tabla estos deben estar entre 5 y 20 intervalos pues la información estaría demasiado reducida si son menos de 5 y si son más de 20 muy dispersa.

Para Construir los intervalos seguiremos la siguiente convención

- Tomamos del conjunto de datos el de menor valor que llamaremos Min y el de mayor valor que llamaremos Max.
- Calculamos: $Max - Min = Recorrido$ al que llamaremos recorrido. Para calcular la cantidad de intervalos y la amplitud buscamos dos múltiplos del recorrido a, b
- Donde a y b son la amplitud y la cantidad de intervalos respectivamente con $a \in \mathbb{Z}^+$
- Los intervalos de clase serán cerrados en el límite inferior y abiertos en el superior.
- El primer intervalo será $[Min,\ Min + amplitud)$. Los demás se definen en forma análoga.
- El límite superior del último intervalo siempre llegara hasta el valor Max. No importa si la amplitud del intervalo es un poco más grande que la dada.

Intervalos de Clase IC	Marca de Clase MC	Frecuencia absoluta n_i	Frecuencia absoluta Acumulada N_i	Frecuencia Relativa $f_i = \frac{n_i}{n}$	Frecuencia Relativa Acumulada F_i
$[Min, Min + a)$					
$[Min + a, Min + 2a)$					
$\vdots$					
$[Min + b * a, Max]$			n		≈ 1

- Marca de Clase (MC): Punto medio del intervalo.
- Frecuencia absoluta (n_i): Es la cantidad de valores del conjunto de datos que pertenecen al intervalo.
- La frecuencia absoluta acumulada se calcula sumando todos los valores de la frecuencia relativa hasta el intervalo i. La suma del último intervalo debe dar la cantidad de datos del problema.
- La frecuencia relativa se calcula dividiendo la frecuencia absoluta entre el total de datos.
- La frecuencia relativa acumulada se calcula sumando los valores de la frecuencia relativa hasta el intervalo i. El valor en el último intervalo debe ser lo más cercano posible a 1.
- A partir de la Tabla podemos construir histogramas de frecuencia.

Ejemplo 1. Datos Agrupados.

Se estudia cómo se adaptan a su nueva situación los alumnos extranjeros que se incorporan a estudios de doctorado en la universidad. Se toma una muestra de los mismos y se les pide que califiquen su adaptación asignándole una puntuación entre 1 y 100 al impacto de las dificultades en su rendimiento. Los resultados fueron:

17, 35, 30, 10, 14, 3, 21, 28, 24, 8, 6, 7, 18, 14, 35, 12, 11, 23, 19, 17, 57, 61, 45, 23, 32.

Un valor de 30 puntos es considerado sensitivo, que un alto por ciento por encima de este valor establece una situación crítica, con la población de estudiantes.

Identifique el tipo de variable, Construya una Tabla empírica de frecuencia, y los histogramas de frecuencia. Calcule las medidas de tendencia central y de dispersión.

Datos:

$X = calificacion\ \ del\ impacto\ de\ las\ dificultades\ de\ rendimiento$

- Tipo de Variable: Cuantitativa

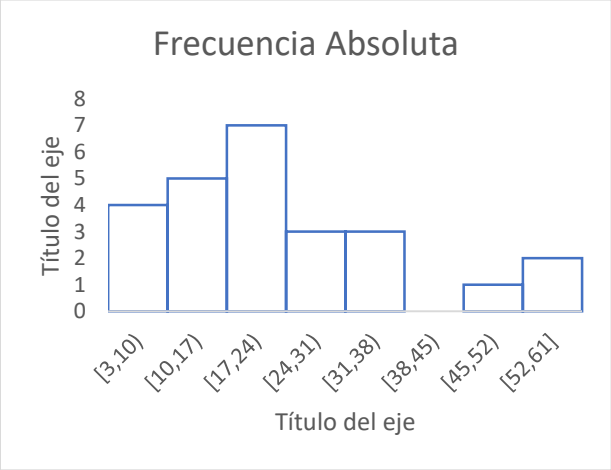
Construyendo la Tabla empírica de frecuencia.

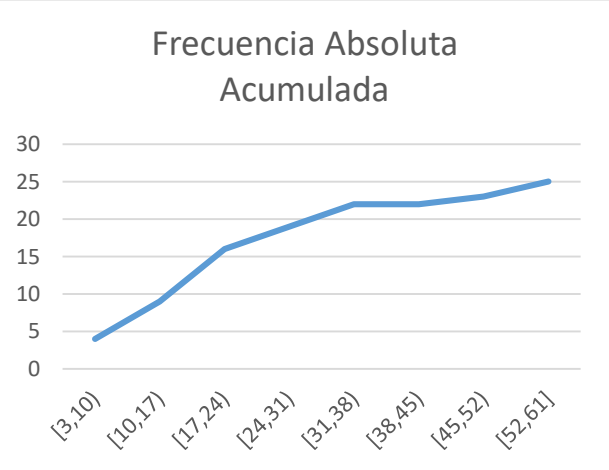
- $n = 25$ como la cantidad de datos es mayor que 15 los agrupamos.
- $Max = 61\ Min = 3$
- $Recorrido\ = 61 - 3 = 58$
- La multiplicación más cercana al valor del recorrido es $7 * 8 = 56$
- Quedando agrupados de la siguiente forma.

Intervalos de Clase IC	**Marca de Clase MC**	**Frecuencia absoluta n_i**	**Frecuencia absoluta Acumulada N_i**	**Frecuencia Relativa $f_i = \frac{n_i}{n}$**	**Frecuencia Relativa Acumulada F_i**
[3,10)	6.5	4	4	0.16	0.16
[10,17)	13.5	5	9	0.20	0.36
[17,24)	20.5	7	16	0.28	0.64
[24,31)	27.5	3	19	0.12	0.76
[31,38)	34.5	3	22	0.12	0.88
[38,45)	41.5	0	22	0	0.88
[45,52)	48.5	1	23	0.04	0.92
[52,61]	56.5	2	25	0.08	1

La elección entre la cantidad de intervalos y la amplitud es una cuestión de práctica, de analizar que es más importante, que los elementos de los grupos se parezcan entre si(menos amplitud, mayor cantidad) o si el estudio que se realiza requiere de menos cantidad de grupos y no importa tanto las diferencias entre los individuos.

Histogramas de Frecuencia.





Cálculo de medidas de tendencia central.

Media Aritmética

$$\bar{X} = \frac{1}{n}\left(\sum_{i=1}^{n} MC_i n_i\right) = \frac{584.5}{25} = 23.38$$

Interpretación: Un estudiante promedio tiene un resultado en el test de 23 puntos aproximadamente lo que quiere decir que la adaptación de un estudiante extranjero promedio no tiene impactos grandes en su rendimiento.

Moda.

La clase modal es $j = 3$ ya que $n_j = 7$. Tenemos que

$L_2 = 17$	$n_2 = 5$	$n_4 = 3$	$a_3 = 7$
Límite inferior del intervalo	Frecuencia relativa anterior	Frecuencia relativa posterior	Amplitud del intervalo

Sustituyendo esos valores en la ecuación de la moda obtenemos:

$$M_o = L_{j-1} + \frac{n_j - n_{j-1}}{(n_j - n_{j-1}) + (n_j - n_{j+1})} a_j$$

$$M_o = 17 + 7\left(\frac{7-5}{(7-5)+(7-3)}\right)$$

$$M_o = 17 + \frac{14}{6} = 19.33$$

Interpretación: la mayoría de los estudiantes tienen un resultado en el test de 19 puntos aproximadamente lo que quiere decir que la adaptación para la mayoría tiene bajo impacto en su rendimiento.

Mediana

En este caso necesitamos la clase que agrupe a la mitad de los estudiantes, por tanto

$$\frac{n}{2} = \frac{25}{2} = 12.5 < N_3 = 15$$

La clase de la mediana es $j = 3$ (La clase de la mediana y la moda no tiene por qué ser la misma.) En esta ecuación en vez de trabajar con la frecuencia absoluta trabajamos con la frecuencia absoluta Acumulada por tanto necesitamos N_{j-1} en este caso $N_2 = 9$.

Sustituimos los valores en la ecuación de la mediana:

$$M_e = L_{j-1} + \frac{\frac{n}{2} - N_{j-1}}{N_j - N_{j-1}} a_j$$

$$M_e = 17 + \frac{12.5 - 9}{16 - 9} * 7$$

$$M_e = 17 + \frac{3.5}{7} * 7$$

$$M_e = 17 + 3.5 = 20.5$$

Interpretación: La media de los estudiantes tiene un resultado en el test de 21 ptos aprox. lo que quiere decir que para el 50% de la muestra de estudiantes la adaptación tiene bajo impacto en su rendimiento. Análisis que es consistente con los de la Moda y Media.

Cálculo de medidas de Dispersión

Varianza

$$\sigma^2 = \frac{1}{n}\left(\sum_{i=1}^{n} n_i(MC_i - \bar{X})^2\right) = \frac{4932.64}{25} = 197.31$$

Desviación típica

$$DT = \sigma = \sqrt{\sigma^2} = \sqrt{197.31} = 14.05$$

Coeficiente de Variación.

$$CV = \frac{DT}{\bar{X}} = \frac{14.05}{23.38} = 0.6$$

Interpretación: Las tres medidas se interpretan utilizando el Coeficiente de Variación y se da en términos de porciento. Por tanto, en este caso se puede observar que el valor del CV es del 60% así que los datos se pueden considerar Muy Heterogéneos.

Ejemplo 2. Datos Simples.

Se mide el tiempo semanal dedicado a buscar información relacionada con él estudio en internet en 10 estudiantes de MATCOM. Se desea conocer las medidas de tendencia central, y las medidas de dispersión de estos datos.

Los resultados fueron en horas:

2.3 5.8 9.3 5.6 7 12.6 23.5 6.4 21.8 14.3

Datos:

$X = Tiempo\ semanal\ dedicado\ a\ buscar\ información\ relacionada\ con\ él\ estudio\ en\ internet$

- Tipo de Variable: Cuantitativa.
- No se pueden trabajar como datos agrupados porque $n = 10 < 15$.

Cálculo de medidas de tendencia central.

Media.

$$\bar{X} = \frac{\sum x_i}{n} = \frac{2.3 + 5.8 + 9.3 + 5.6 + 7 + 12.6 + 23.5 + 6.4 + 21.8 + 14.3}{10} = \frac{108.6}{10} = 10{,}86$$

Interpretación: El tiempo semanal promedio dedicado a la consulta de internet es de 11 horas aproximadamente lo que quiere decir teniendo en cuenta que una semana tiene 168 horas es realmente muy poco tiempo es menos de un día a la semana.

Moda.

1. Ordenar los Datos.

   2.3 5.6 5.8 6.4 7 9.3 12.6 14.3 21.8 23.5

   Como no hay ningún dato repetido esta muestra no tiene moda.

Mediana.

2. Ordenar los Datos.

   2.3 5.6 5.8 6.4 7 9.3 12.6 14.3 21.8 23.5

3. Calcular.
   a. $n = 10$ es par, tomamos los valores del medio. 7 y 9.3 y los sumamos 16.3
   b. Dividimos la suma entre dos.

$$M_e = \frac{7 + 9.3}{2} = \frac{16.3}{2} = 8.15$$

Interpretación: El tiempo semanal dedicado a la consulta de internet por el 50% de los estudiantes es de 8 horas aproximadamente lo que quiere decir teniendo en cuenta que una semana tiene 168 horas es muy poco tiempo.

Cálculo de medidas de Dispersión

Varianza.

$$\sigma^2 = \frac{1}{n-1}\left(\sum_{i=1}^{10}(x_i - \bar{X})^2\right) = \frac{458}{9} = 50.88$$

Desviación Estándar.

$$\sigma = \sqrt{50.88} = 7.13$$

Coeficiente de Variación.

$$CV = \frac{DT}{\bar{X}} = \frac{7.13}{10.86} = 0.66$$

Es Heterogénea la muestra.

Para variables Discretas.

La construcción y el llenado de la tabla es similar, solo que como estamos tratando con variables discretas, no tendremos intervalos de clases, si no que codificaremos las variables usando números enteros.
Por ejemplo, si tuviéramos una variable *Color del Pelo* los posibles valores pudieran ser rubio, rojo, castaño y negro por lo que la codificación seria:

Rubio	1
Rojo	2
Castaño	3
Negro	4

Al no tener intervalos de clase tampoco tenemos la columna marca de clase, Ahora las demás columnas, frecuencia absoluta y relativa y sus acumulados, se calculan de la misma forma.

Valores V	**Frecuencia absoluta n_i**	**Frecuencia absoluta Acumulada N_i**	**Frecuencia Relativa $f_i = \frac{n_i}{n}$**	**Frecuencia Relativa Acumulada F_i**
V_1				
…				
V_k		n		≈ 1

Ejemplo 3.

Al implantar un nuevo servicio informático para el control del desvío de recursos en una entidad procesadora de productos cárnicos, se detectaron al cabo de un semestre las substracciones siguientes:

Substracciones	0	3	5	7
Semanas	10	5	4	6

Determine la media, las medidas de tendencia central y las de dispersión, que caracteriza las substracciones detectadas.

Datos:

$X = Sustracciones\ detectadas\ en\ una\ semana$

- Tipo de Variable: Cualitativa. $X = \{0,\ 3,\ 5,\ 7\}$

De acuerdo a la tabla anterior la cantidad de sustracciones caracterizan una semana en particular. Por tanto, hay que construir la tabla de distribución empírica de frecuencia usando como valores la codificación de la variable substracciones de forma tal que:

1	0 substracciones
2	3 substracciones
3	5 substracciones
4	7 substracciones

y las semanas serian la frecuencia absoluta, por tanto, la cantidad de datos es $n = 25$, luego los debemos trabajarlos como datos agrupados.

V	n_i	N_i	$f_i = \frac{n_i}{n}$	F_i
1	10	10	0.4	0.4
2	5	15	0.2	0.6
3	4	19	0.16	0.76
4	6	25	0.24	1

Cálculo de medidas de tendencia central.

Media.

$$\overline{X} = \frac{\sum ni * V_i}{n} = \frac{1 * 10 + 2 * 5 + 3 * 4 + 4 * 6}{25} = \frac{56}{25} = 2.24$$

Interpretación: El promedio de substracciones realizadas en una semana es de 2 aproximadamente. Teniendo en cuenta que se midieron 25 semanas entonces a un promedio de 2 substracciones por semana eso haría un total de 50 substracciones en un semestre lo que es una cantidad considerablemente grande.

Mediana.

En este caso

$$\frac{n}{2} = \frac{25}{2} = 12.5 < N_2 = 15$$

La clase de la mediana es $j = 2$

$$N_1 = 10.$$
$$a_2 = 1$$
$$L_{j-1} = 2$$

En este caso el límite inferior es el propio valor $L_{j-1} = V_2 = 2$.

Sustituimos los valores en la ecuación:

$$M_e = L_{j-1} + \frac{\frac{n}{2} - N_{j-1}}{N_j - N_{j-1}} a_j$$

$$M_e = 2 + \frac{12.5 - 10}{15 - 10} * 1$$

$$M_e = 2 + \frac{2.5}{5}$$

$$M_e = 2 + 0.5 = 2.5$$

Interpretación: Como la mediana $M_e = 2.5$ esto significa que la mitad de las semanas 50% de las semanas fueron hechas 3 substracciones aproximadamente.

Cálculo de medidas de Dispersión

Varianza.

$$\sigma^2 = \frac{1}{n-1}\left(\sum_{i=1}^{10} n_i(V_i - \bar{X})^2\right) = \frac{36.56}{24} = 1.46$$

Desviación Estándar.

$$\sigma = \sqrt{1.46} = 1.21$$

Coeficiente de Variación.

$$CV = \frac{DT}{\bar{X}} = \frac{1.21}{2.24} = 0.54$$

Interpretación: Como se puede observar el valor del CV es del 54% por lo que los datos son Muy Heterogéneos.

Medidas de Posición.

Las medidas de posición se utilizan para describir la posición de un dato específico con respecto al resto de los datos cuando están en orden por categorías. Cuartiles y Percentiles son dos de las medidas más populares.

Cuartiles

Son valores de la variable que dividen los datos ordenados en cuartos; cada conjunto de datos tiene 3 cuartiles. El primer cuartil Q_1, es un numero tal que a lo sumo el 25% de los datos son menores que él y a lo sumo el 75% de los datos son mayores que él. El segundo cuartil Q_2 es la mediana. El tercer cuartil Q_3, es un numero tal que a lo sumo el 75% de los datos es son menores que él y a lo sumo el 25% de los datos son mayores que él.

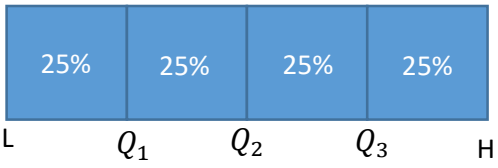


Figura 1. Datos Clasificados Orden Creciente.

El cálculo de los cuartiles se realiza de la misma forma que el de los percentiles y este se explica a continuación.

Percentiles

Son los valores de la variable que dividen el conjunto de datos clasificados en 100 subconjuntos iguales; cada conjunto de datos tiene 99 percentiles (vea la figura 2). El k-esimo percentil P_k es un valor que a lo sumo $k\%$ de los datos son menores que el valor P_k y a lo sumo $(100-k)\%$ de los datos son mayores (ver figura 2.)

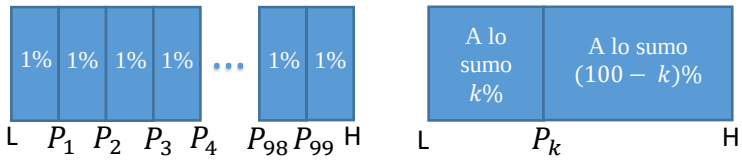


Figura 2. Percentiles y k-esimo Percentil.

- El primer cuartil y el 25 percentil son lo mismo; es decir $Q_1 = P_{25}$. También $Q_3 = P_{75}$
- La mediana, el segundo cuartil y el percentil 50 son todos lo mismo: $M_e = Q_2 = P_{50}$, por lo que para hallar Q_2 o P_{50} usamos el procedimiento para hallar la mediana.

El procedimiento para determinar el valor de cualquier k-esimo percentil (o cuartil) comprende 4 pasos básicos:

Paso 1 -Ordenar los Datos de menor a mayor.

Paso 2 -Determinar la profundidad del k-esimo $d(P_k)$ percentil, proceso que se realiza en tres pasos

a. Calcular el valor $\frac{nk}{100}$ donde n es el tamaño de la muestra y k es el valor del percentil

b. Si $\frac{nk}{100}$ es un número entero, entonces $d(P_k) = \frac{nk}{100} + 0.5$

   Ejemplo: $\frac{nk}{100} = 23 \Rightarrow d(P_k) = 23 + 0.5 = 23.5$

c. Si $\frac{nk}{100}$ no es un número entero, $d(P_k)$ es igual el siguiente entero más grande que $\frac{nk}{100}$

   Ejemplo: $\frac{nk}{100} = 17.2 \Rightarrow d(P_k) = 18$

Paso 3 - Finalmente Determinar el Valor de P_k, esto se hace contando desde el menor valor de la muestra de datos ordenados hasta el $d(P_k)$-esimo número.

a. Si el valor de $d(P_k)$ es un entero, el valor de P_k será el valor encontrado.

b. Si el valor de $d(P_k)$ termina en $.5$, o sea contiene la fracción $\frac{1}{2}$ entonces el valor de P_k esta entre el valor encontrado en la posición $d(P_k) - 0.5$ y $d(P_k) + 0.5$, por lo que el promedio de estos valores es el valor de P_k

Ejemplo 4. Utilizando datos del Ejemplo 1.

Los datos que tenemos son los siguientes

17, 35, 30, 10, 14, 3, 21, 28, 24, 8, 6, 7, 18, 14, 35, 12, 11, 23, 19, 17, 57, 61, 45, 23, 32.

Calculando el Primer Cuartil.

Paso 1 - Ordenamos los datos:

3, 6, 7, 8, 10, 11, 12, 14, 14, 17, 17, 18, 19, 21, 23, 23, 24, 28, 30, 32, 35, 35, 45, 57, 61.

Paso 2 - Para hallar el primer cuartil necesitamos determinar la profundidad de Q_1 o sea $d(Q_1)$

$n = 25$ – Porque tenemos 25 datos.

$k = 25$ dado que $Q_1 = P_{25}$

$$\frac{nk}{100} = \frac{25 * 25}{100} = \frac{625}{100} = 6.25$$

Por tanto $d(Q_1) = 7$

Paso 3 - Por tanto Q_1 sería el séptimo elemento de los datos ordenados. O sea $Q_1 = 12$

Calculando el Percentil 60.

Paso 1 - Los datos están ordenados.

Paso 2 - Se necesita determinar la profundidad de P_{60} o sea $d(P_{60})$

$n = 25$ – Porque tenemos 25 datos.

$k = 60$ $\frac{nk}{100} = \frac{25*60}{100} = \frac{1500}{100} = 15$ Por tanto $d(Q_1) = 15.5$

Paso 3 - Por tanto P_{60} sería el promedio del decimoquinto y decimosexto elemento de los datos ordenados. O sea $P_{60} = \frac{23+23}{2} = 23$

Gráficos de Caja y Bigotes.

Son muy usados por los estadísticos debido a la división que realiza para analizar los datos. Es una gráfica que da un sumario basado en 5 números de un conjunto de datos. Los 5 valores numéricos (smallest, lower hinge (bisagra), median, upper hinge and largest) están localizados en una escala que puede ser tanto horizontal como vertical. La "**caja**" es usada para representar la parte del medio de los datos o sea la que está dentro de los **hinges** o **bisagras**, los bigotes son segmentos de línea que representan la otra parte de los datos, un segmento representa el cuarto de los datos que son menores en valor que el valor del **lowest hinge** y el otro segmento de línea representa el cuarto de los datos que son mayores que el valor del **upper hinge**. Las bisagras y el valor medio separan el conjunto de datos ordenados en 4 subconjuntos. Los valores de las bisagras son usualmente los valores del primer y tercer cuartil, pero dependiendo de la cantidad de datos estos pueden diferir un poco. Una forma muy sencilla de calcularlos es ordenar los datos de menor a mayor y formar en 2 subconjuntos que ambos incluyan a la mediana. Para el **lower hinge** hallan la mediana de los datos desde la posición 1 hasta la posición de la mediana y para el **upper hinge** hallan la mediana de los datos desde la posición de la mediana hasta la posición n.

Ejemplo 4. Utilizando datos del Ejemplo 1.

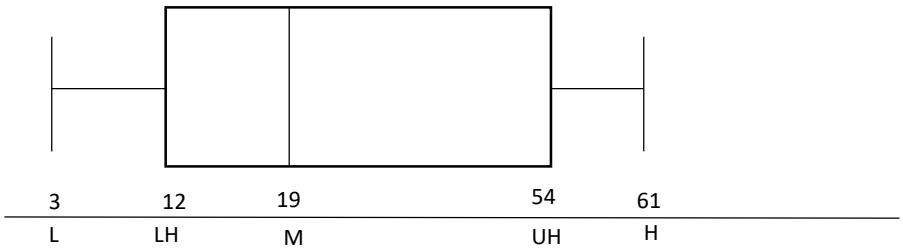
Los datos que tenemos ordenados son los siguientes

3, 6, 7, 8, 10, 11, 12, 14, 14, 17, 17, 18, 19, 21, 23, 23, 24, 28, 30, 32, 35, 35, 45, 57, 61

Por tanto, el menor valor o smallest es 3 y el mayor o largest es 61. La mediana es $M = 19$

Como la mediana es 19, para calcular el lower hinge calculamos la mediana del primer subconjunto 3, 6, 7, 8, 10, 11, **12**, 14, 14, 17, 17, 18, 19 en este caso 12. Para calcular el upper hinge calculamos la mediana del subconjunto 19, 21, 23, 23, 24, 28, **30**, 32, 35, 35, 45, 57, 61

Por tanto, el diagrama quedaría de la siguiente forma



Si los datos siguen una distribución normal (Eso quiere decir que la variable X que mide estos datos, $X \sim N(\mu, \sigma^2)$) entonces este grafico se verá simétrico y centrado en los datos, mientras más se desvíen los datos de la distribución normal más se deformara la caja de una

Resumen de Interpretación de las Medidas de Tendencia Central, Dispersión y Posición.

Son medidas descriptivas que nos permite interpretar el comportamiento de las muestras.

- La Media es el dato que equilibra la distribución de los datos. O sea, el promedio.
- La Mediana es el valor intermedio de la distribución de los datos. Donde se agrupan alrededor del 50% de los datos.
- La moda es el dato que se repite con mayor frecuencia en la distribución de los datos.
- La Varianza mide las desviaciones de los datos con respecto al promedio del conjunto de datos. Está expresada en unidades elevadas al cuadrado.
- La Desviación Estándar mide cuánto se separan los datos dados, del promedio del conjunto de datos.
- Coeficiente de variación mide la dispersión en términos de porcentaje, señala qué tan grande es la magnitud de la desviación estándar respecto al promedio del conjunto de datos que se examina. Para interpretar el coeficiente de variación, podemos usar las apreciaciones de la Tabla que se muestra a continuación:

CV	Interpretación
$CV \geq 26\%$	Muy Heterogéneo
$16\% \leq CV < 26$	Heterogéneo
$11\% \leq CV < 16$	Homogéneo
$0\% \leq CV \leq 11$	Muy Homogéneo

- Los cuartiles dividen a la distribución de los datos en 4 partes iguales.
- Los deciles dividen a la distribución de los datos en 10 partes iguales
- Los percentiles dividen a la distribución de los datos en 100 partes iguales.