Probabilidad y Estadística. Unidad 2: Medidas de resumen.

Año de cursada: 1°

Clase N.º 3: Resumiendo información.

Contenido: Medidas de tendencia central. Media aritmética. Propiedades. Mediana. Moda. Relación empírica entre la media, mediana y moda. Medidas de orden. Cuartiles. Deciles. Percentiles. Medidas de variabilidad. Rango. Varianza, Desvío Estándar. Coeficiente de variación. Datos atípicos. Diagramas de cajas.

1. **Presentación:**

Bienvenidos y Bienvenidas a la tercera clase de Probabilidad y Estadística. En las clases anteriores hemos comenzado a tratar con la definición de estadística, las dos grandes ramas sobre las que opera esta disciplina del campo de la matemática, así también, los conceptos básicos que nos permiten precisar algunos de los aspectos indispensables para el trabajo en este campo.

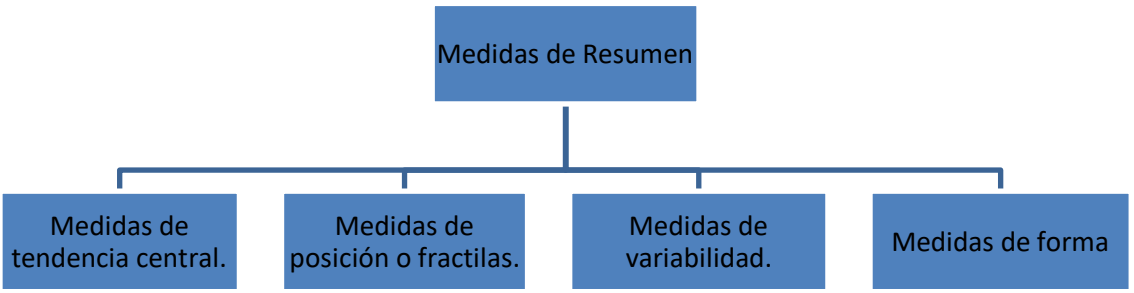
Avanzando, hemos dado los primeros pasos para el tratamiento de datos obtenidos por algún método de recolección. Además, hemos podido comenzar a trabajar con algunas herramientas para la organización y presentación de información.

En esta clase, profundizaremos sobre las medidas de resumen como una forma de poner de manifiesto algunas de las características más prominentes de un conjunto de datos. Esta unidad podemos relacionarla con la etapa de síntesis que habíamos descrito con describimos las etapas del método estadístico en la primera clase. Dentro de las medidas de resumen vamos a

encontrar las medidas de tendencia central, de posición, de variabilidad, de forma, entre otras.

2. Desarrollo:

Como habíamos dicho en la presentación de la clase, trabajaremos con las medidas de resumen. Éstas, representan valores numéricos que tienen la propiedad de caracterizar a un conjunto de datos. Existen varias medidas de resumen que se utilizan para medir magnitudes o características particulares de cada conjunto. Veamos entonces, cuáles son las que vamos a considerar para este curso:



Comencemos a trabajar con las medidas de tendencia central. Estas corresponden a valores numéricos que nos permiten caracterizar a un conjunto de datos, es decir, resumir o sintetizar poniendo de manifiesto sus particularidades más prominentes, pues si queremos referirnos a un conjunto de datos resultaría poco preciso elegir los valores extremos (mínimo o máximo) ya que no representaría a los valores típicos. Entonces, adquiere sentido realizar la elección de algún valor intermedio. Por ello, se denominan de tendencia central porque usualmente se ubican en el centro de la distribución. No se preocupen, cuando representemos gráficamente a las medidas veremos con mayor claridad a qué nos referimos con la ubicación en el centro.

Entonces, veremos tres medidas dentro de este grupo: **la media aritmética, la mediana y la moda**, por el momento vamos a suponer que representan a un conjunto de datos muestrales.

Media Aritmética

Esta medida es una de las más conocida, comúnmente suele llamarse *promedio* y suele representarse con la $\underline{x}$. Para su obtención debemos sumar todos y cada uno de los valores de la variable, luego, el resultado de dicha suma se lo divide por el tamaño de la muestra[1].

Su cálculo difiere de acuerdo con el tipo de variable con la que se trabaje y presenta algunas modificaciones según se trabaje con datos agrupados o no.

Para datos sin agrupar:

Simbólicamente se representa de la siguiente manera: $\underline{x} = \frac{\sum \quad x_i}{n}$

x_i: representa cada uno de los valores que asume la variable en el conjunto de datos sobre el cual realizamos el estudio. El símbolo $\sum$ se denomina sumatoria, y representa la suma de los valores observados.

Para datos agrupados:

Simbólicamente se representa de la siguiente manera: $\underline{x} = \frac{\sum \quad x_i * FA}{n}$

Los símbolos significan lo mismo, aunque en la fórmula podemos ver que se incluye a la frecuencia absoluta. Vale una aclaración, en el caso de trabajar con datos agrupados en intervalos de clase, x_i representaría la marca de clase.

[1] El tamaño de la muestra corresponde a la cantidad de observaciones realizadas, en el caso de estar trabajando con datos agrupados en tablas de frecuencia coincide con la sumatoria de las FA.

Esta medida presenta algunas ventajas que consisten en que involucra en su cálculo todos los valores el conjunto de datos que se está estudiando (no se pierde información), es de cálculo sencillo y es única para cada serie o grupo de datos. También presenta algunas desventajas que consiste en que no puede computarse en el caso de intervalos abiertos, es muy sensible a los valores extremos y no puede calcularse para variables cualitativas.

Propiedades de la media aritmética

1. La suma de las desviaciones[2] con respecto a la media aritmética es nula (cero). Significa que la media aritmética compensa las desviaciones positivas con negativas.
2. La suma de los cuadrados de las desviaciones es igual a un mínimo. Es decir, que cualquier suma de los cuadrados de las desviaciones respecto de otro valor distinto de la media aritmética siempre será mayor.
3. La media de un conjunto de datos constante (siempre aparece el mismo valor), es la misma constante.
4. Si sumamos cada valor de la variable por una constante y calculamos la media aritmética en el nuevo conjunto de datos, el valor es igual a la media del conjunto original pero aumentada la constante que se sumó.
5. Si multiplicamos cada valor de la variable por una constante, entonces al calcular la media, ésta queda multiplicada por dicha constante.

Ahora bien, conozcamos una nueva medida.

Mediana

[2] Observación: llamaremos desviación a una diferencia (resta) entre un valor de la variable y una medida de posición.

La medida es el valor situado en el centro del conjunto ordenados por magnitud (podría pensarse de mayor a menor), se representa como $\tilde{x}$, Me o Md. El 50% de los valores de la variable son inferiores a la mediana, mientras que el otro 50% son mayores a él. Tiene como particularidad que no depende de los valores numéricos del conjunto de datos, sino que depende de la posición que ocupan, es decir, para este caso la variable necesariamente debe poder ordenarse.

Al igual que el caso de la media aritmética, distinguimos dos casos: uno en el que los datos no presentan agrupación frecuencial y en el caso en los que están organizados en una tabla de distribución de frecuencias.

Vamos por partes, primero analicemos el caso donde los datos se presentan sin agrupar. La primera acción por realizar sería ordenar nuestra serie de datos. Puede haber dos posibilidades respecto de la cantidad de datos que posea el conjunto sea par o bien impar. En cada caso que la cantidad de datos se impar, tendremos un solo valor central y para seleccionarlo debemos ubicar la posición $\frac{n+1}{2}$, la cual nos indica cual es la posición de la mediana. En el caso de que estudiemos un conjunto de datos y el valor que resulte de aplicar la fórmula anterior sea 10, no quiere decir que la mediana es el valor 10 sino que la media corresponde al valor que se encuentra en la posición número 10.

Veamos ahora cuando la cantidad de datos es par (siempre para datos ordenados, pero no agrupados en tablas), lógicamente tendremos dos valores centrales y para hallarlos debemos ubicar las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$. Entonces, valiéndonos de dichas fórmulas podremos encontrar los valores que se encuentran en dicha posición. Como la mediana es un valor único, lo que debemos hacer es promediarlos, es decir sumarlos y dividirlo por dos.

Ahora veamos qué sucede cuando los datos se encuentran en tablas de frecuencias. Veamos en primer lugar, cuando tenemos datos agrupados en tablas de distribución de frecuencias simples. Para este caso, tomamos el tamaño de la muestra y lo dividimos por 2, con el valor resultante vamos a operar sobre la columna de frecuencia acumuladas, vemos qué valor lo contiene[3]. Seguidamente, ubicamos el valor de la variable que se corresponde con dicha FAA, y éste será el valor de la mediana.

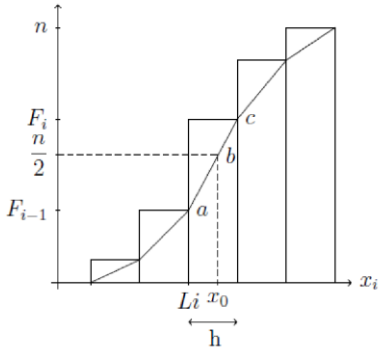
Por otra parte, se presenta el último caso para datos agrupados en tablas de distribución de frecuencias con intervalos de clase. Aquí la fórmula es bastante distinta, puesto que en la primera columna de la tabla tenemos como elementos a intervalos. Entonces, la fórmula es la siguiente:

$\tilde{x} = LI + \frac{\frac{n}{2}+FAA_{-1}}{FA}.h$ → En esta fórmula, la mayoría de los elementos son conocidos excepto FAA_{-1}, lo cual simboliza la frecuencia acumulada anterior.

Para poder identificar cada uno de los elementos de la fórmula, en primer lugar, vamos a calcular el cociente entre el tamaño de la muestra (n) y 2. Luego, con ese valor vamos a buscar en la columna de las frecuencias acumuladas cuál es el valor en el que está contenido y, consecuentemente, identificamos cuál es el intervalo en el que se encuentra la mediana, para utilizar LI, FA y h que son los valores que intervienen en la fórmula.

Para representar gráficamente la mediana, utilizamos un histograma para frecuencias acumuladas.

[3] Hace referencia al valor superior inmediato. Por ejemplo, el 11 está contenido por el 12, y no así por el 10.



Trazamos una proyección perpendicular al eje de las frecuencias acumuladas pasando por $\frac{n}{2}$ hasta cortar a la ojiva[4], y desde dicha intersección trazamos una proyección perpendicular al eje de la variable hasta cortarla en x_0.

Seguidamente, avanzamos sobre la definición y cálculo de la moda.

Moda o Modo

Esta medida de tendencia central corresponde al valor con mayor frecuencia[5] de ocurrencia en el conjunto de datos en cuestión, y se representa mediante $\hat{x}$ o Mo. Puede ocurrir que un conjunto de datos tenga una sola moda o modo y, ese caso se denomina *unimodal*. De manera análoga, el conjunto de datos que posee dos modas o modos se denomina *bimodal*, mientras que el conjunto de datos que posee más de una moda se denomina *multimodal.*

Para el caso de esta medida, podemos distinguir (como en los casos anteriores) entre datos no agrupados, datos agrupados en TdDF[6] y TdDF con intervalos. Para el primer caso, solo basta con organizar el conjunto de datos de modo que podamos contar de manera ordenada y determinar cuál es el

[4] Ojiva se denomina a la curva que resulta al construir un histograma para frecuencias acumuladas y unir mediante segmentos los extremos inferiores y superiores de cada barra ubicada sobre los distintos intervalos.
[5] Entendemos a la expresión “mayor frecuencia” como la mayor frecuencia absoluta.
[6] Tabla de distribución de frecuencias.

valor de la variable que más se repite. En el segundo caso, solo basta con posicionarnos en la columna correspondiente a las frecuencias absolutas e identificar la mayor de ellas. Para el último caso, TdDF con intervalos de clase debemos recurrir a una fórmula que nos permita determinar cuál es el valor de la moda, les presento la fórmula y, posteriormente, la describimos.

$$\hat{x} = LI + \frac{\Delta_1}{\Delta_1 + \Delta_2}.h$$

En esta fórmula lo que tenemos de nuevo serían Δ_1 y Δ_2, son en realidad diferencias entre frecuencias absolutas.

Δ_1: Es la diferencia entre la FA del intervalo modal y $FA_{-1.}$

Δ_2: Es la diferencia entre FA de intervalo modal y $FA_{+1.}$

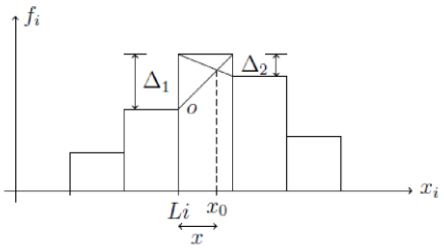
Cabe aclarar que para encontrar el intervalo que contiene a la moda, debemos identificar el mayor valor en la columna que corresponde a la frecuencia absoluta.

Cabe aclarar que, en el InfoStat, no encontraremos la sección que nos permita realizar el cálculo de forma automática, pero se puede sencillamente hallar este valor que corresponde a la moda o modo.

Representación gráfica de la moda

Para representar la moda o modo de manera gráfica vamos a utilizar el histograma de frecuencias absolutas. Una vez realizado el histograma, se debe ubicar la barra (que se extiende por sobre los intervalos) con mayor altura. Posteriormente, trazar una línea que una el extremo superior derecho de la barra con mayor altura y el extremo superior derecho de la barra correspondiente al intervalo posterior al intervalo modal. De la misma

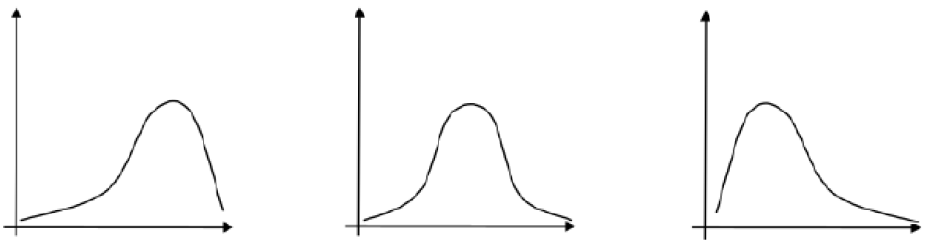
manera se procede para trazar la línea que va desde el extremo superior izquierdo.



Relación entre media, mediana y moda.

¿Cuál es la mejor medida de tendencia central? Desafortunadamente, esta pregunta no tiene una sola respuesta óptima porque no existen criterios objetivos para determinar cuál es la medida más representativa para todos los conjuntos de datos.

Una distribución se dice que es sesgada, si no es simétrica[7] y se extiende más hacia un lado que hacia el otro. Una comparación de la media, la mediana y la moda puede revelar información acerca de la característica del sesgo que se define e ilustra a continuación.



a) Sesgada a la izquierda (sesgo negativo): La media y la mediana están a la izquierda de la moda.

b) Simétrica (sesgo cero): La media, la mediana y la moda son iguales.

c) Sesgada a la derecha (sesgo positivo): La media y la mediana están a la derecha de la moda.

[7] Una distribución se dice simétrica cuando se la puede dividir en dos partes que resulten idénticas.

Tipos de distribución según la relación entre media y mediana.

- Simétrica: $\underline{x} = \tilde{x}$
- Sesgada a izquierda: $\underline{x} > \tilde{x}$
- Sesgada a derecha: $\underline{x} < \tilde{x}$

Para curvas de frecuencias unimodales, que sean moderadamente sesgadas, se tiene la siguiente relación empírica. $\underline{x} - \hat{x} = 3(\underline{x} - \tilde{x})$

Les dejo a continuación un link para que puedan ver algunos ejemplos sobre el cálculo de las medidas de tendencia central. Además, en el siguiente link podrán observar el cálculo de medidas de tendencia central y otras medidas de resumen que comenzaremos a trabajar en la próxima clase.

I. Los siguientes datos corresponden al número de hijos de mujeres mayores a 20 años que asistieron a un centro de salud.

0 2 3 3 4 0 2 1 3 4 1 3 3 0 1 3 4 3 0 1 1 2

a) Presente los datos en una tabla de distribución de frecuencias.
b) Complete las frecuencias acumuladas y las relativas.
c) ¿Cuántas mujeres no tienen hijos?
d) ¿Cuántas mujeres constituyen la muestra?
e) ¿Qué porcentaje de mujeres tiene menos de 3 hijos?
f) ¿Qué porcentaje de mujeres tiene 4 hijos?
g) ¿Cuántas mujeres tienen al menos un hijo?
h) Represente gráficamente las frecuencias absolutas y las acumuladas.

II. Los siguientes datos corresponden al tiempo, en segundos, que necesitó cada alumno de un grupo para realizar una actividad.

57 69 65 75 60 72 67 79 57 71 65 75 61 73 67 81 58 71 65 76 64 74 68 82 60 71 67 79 65 75 69 86

a) Identifique y clasifique la variable.
b) Construya una tabla de distribución de frecuencias con 6 intervalos cerrados a la izquierda y que el límite inferior del primer intervalo sea igual a 57 seg.
c) Construya un histograma de frecuencias absolutas y el de frecuencias acumuladas.
d) Responda las siguientes cuestiones:

1. ¿Cuántos alumnos han demorado entre 62 y 67 segundos en realizar la actividad?
2. ¿Cuántos han demorado entre 72 y 77 seg?
3. ¿Qué porcentaje demoró entre 57 y 62 seg?
4. ¿Cuántos demoraron menos de 72 seg?
5. ¿Qué porcentaje demoró menos de 82 seg?
6. ¿Cuántos emplearon 72 seg o más?

Medidas de orden

Las medidas de orden o de posición no centrales permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Entre otros indicadores, se suelen utilizar una serie de valores que dividen la muestra en tramos iguales. También se las conoce como fractilas.

Cuartiles[8] (Q_1, Q_2, Q_3)

Son 3 valores que distribuyen la serie, ordenada de forma creciente, en cuatro tramos con igual cantidad de datos, cada uno de ellos concentra el 25% de los resultados.

- **Primer cuartil** (Q_1). Es un valor tal que el 25% de las observaciones son menores o iguales que dicho cuartil y el 75% de las observaciones son mayores o iguales. En una tabla simple se ordenan los datos y se toma como

[8] Los cuartiles, al igual que la mediana, responden a una posición dentro del conjunto de datos, es decir, que el cociente i4n, encuentra la posición del valor que corresponde al cuartil buscado.

Q_1 el valor que acumula el 25% de los datos. Para calcularlo en una tabla con intervalos se debe utilizar una fórmula similar a la de la mediana.

- **Segundo cuartil** (Q_2). Coincide con la mediana.
- **Tercer cuartil** (Q_2). Es un valor tal que el 75% de las observaciones son menores o iguales a dicho cuartil y el 25% son mayores o iguales. En una tabla simple se ordenan los datos y se toma como Q3 el valor que acumula el 75% de los datos. Para calcularlo en una tabla con intervalos se debe utilizar la fórmula similar a la del primer cuartil.

Ahora bien, la fórmula general para el cálculo de percentiles es la siguiente:

$Q_i = LI + \frac{\frac{i}{4}n - FAA_{-1}}{FA}.h$ → De esta fórmula, a diferencia de la que habíamos utilizado para la media, la única diferencia que encontramos es el cociente $\frac{i}{4}n$, el índice i varía de acuerdo con el cuartil que deseamos calcular, por ejemplo, si queremos calcular el cuartil 1 (Q_1), en dicho caso asume el valor 1. Ahora si queremos calcular el cuartil 2, entonces i valdrá 2.

Entonces, para poder utilizar la fórmula, lo primero que debemos hacer es encontrar aquel intervalo que contenga al cuartil, para ello usamos el cociente $\frac{i}{4}n$ que nos indicará (de manera similar al caso de la mediana) cuál es el intervalo con el que debemos trabajar. Una vez obtenido el valor de dicho cociente, buscamos en la columna de las FAA un valor que lo contenga, y posteriormente se aplica la fórmula.

En el caso que estemos trabajando con datos sin agrupar, solo debemos identificar el valor en la posición $\frac{i}{4}n$, asignándole un valor a i de acuerdo con el cuartil a calcular.

Deciles

Son 10 valores que distribuyen la serie, ordenada en forma creciente, en diez tramos iguales en cantidad de datos, cada uno de ellos concentra el 10% de los resultados. Los deciles se denotan $D_1; D_2; \dots; D_9$.

El procedimiento de trabajo es muy similar a lo que vimos anteriormente, para los cuartiles.

La fórmula general es: $D_i = LI + \frac{\frac{i}{10}4 - FAA_{-1}}{FA}.h \rightarrow$ La fórmula es similar a la que habíamos trabajado para los cuartiles. En este caso, i asume valores entre 1 y 9, en función del decil que se quiera calcular. Cabe aclarar que el decil 10 coincide con el total de observaciones, por lo que no tiene mucho sentido incorporarlo.

En el caso que se trabaje con datos sin agrupar o agrupados en tablas simples, se procede de manera análoga a la anterior.

Percentiles.

Son 100 valores que distribuyen la serie, ordenada en forma creciente, en cien tramos iguales en cantidad de datos, en los que cada uno de ellos concentra 1% de los resultados. La fórmula de trabajo y la manera de utilizarla es similar al caso de los cuartiles y deciles.

Entonces, la fórmula general es: $P_i = LI + \frac{\frac{i}{100}n - FAA_{-1}}{FA}.h \rightarrow$ En este caso, i puede asumir valores entre 1 y 99.

Vamos a estudiar ahora otro tipo de medidas de resumen.

Medidas de Variabilidad.

Con estas medidas se estudia la distribución de los valores de la serie, analizando si estos se encuentran más o menos concentrados, o más o menos dispersos.

Existen dos tipos de medidas de dispersión, las absolutas y las relativas. Las primeras llevan unidad de medida y las últimas no. Entre las más utilizadas podemos destacar las siguientes: el rango, el rango intercuartílico, la varianza, el desvío estándar (absolutas) y el coeficiente de variación (relativa).

- **Rango.** Es la diferencia entre el mayor valor y el menor valor en un conjunto de observaciones. El rango que denotamos por R tiene la ventaja de que es fácil de calcular y sus unidades son las mismas que las de la variable que se mide. El rango no toma en consideración el número de observaciones de la muestra estadística, sino solamente la observación del valor máximo (x_{max}) y la del valor mínimo (x_{min}). Sería deseable utilizar también los valores intermedios del conjunto de observaciones. $R = x_{max} - x_{min}$
- **Rango intercuartílico.** Es la diferencia entre el tercer cuartil y el primero. Es una medida de variabilidad que supera la dependencia sobre los valores extremos y lo denotaremos por RIC.

$$RIC = Q_3 - Q_1$$

- **Desviación Media.** Esta medida es más acorde que la de amplitud, ya que involucra a todos los valores del conjunto de observaciones corrigiendo la desviación. Esta medida que denotamos por DM se obtiene calculando la media de la muestra, y luego realizando la sumatoria de las diferencias (positivas, para evitar la anulación de los

desvíos) de todos los valores de la variable X con respecto de la media. Luego se divide por el número de observaciones.

Una medida como esta tiene la ventaja de utilizar cada observación y corregir la variación en el número de observaciones al hacer la división final. Y por último también se expresa en las mismas unidades que las observaciones mismas.

- **Varianza.** Existe otro mecanismo para solucionar el efecto de

Datos sin agrupar	Datos agrupados
$D_M = \dfrac{\sum_{i=1}^{n} \lvert x_i - \overline{x} \rvert}{n}$	$D_M = \dfrac{\sum_{i=1}^{n} \lvert x_i - \overline{x} \rvert f_i}{n}$

Datos sin agrupar	Datos agrupados
$s_X^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$	$s_X^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 f_i$

cancelación entre diferencias positivas y negativas. Si elevamos al cuadrado cada diferencia antes de sumar, desaparece la cancelación. Si denotamos por s_x^2 la varianza muestral (sesgada)[9] de la variable X tenemos las fórmulas:

[9] La varianza muestral presenta un sesgo que más adelante será corregido.

Fórmula de trabajo para datos sin agrupar: $s_x^2 = \frac{\sum x_i^2}{n} - \underline{x}^2$, mientras que para datos agrupados resulta ser $s_x^2 = \frac{\sum x_i^2 . FA_i}{n} - \underline{x}^2$.

Esta medida tiene una desventaja, y es que sus unidades no son las mismas que las de las observaciones, ya que son unidades elevadas al cuadrado. Por lo que recurriremos a otra medida para solucionar este inconveniente, pero antes veamos algunas propiedades de la varianza.

Propiedades de la varianza.

1. $s_x^2 \geq 0$
2. La varianza de una constante es cero, es decir, vale lo exactamente cero solo cuando todos los valores de la variable (x_i) son iguales.
3. Si se multiplica a un conjunto de datos por una constante, se obtiene un nuevo conjunto de datos, entonces, la varianza del último es igual a la varianza del primero multiplicada por el cuadrado de la constante que se multiplicó al conjunto de datos, en símbolos resultaría: si $y_i = k.x_i \rightarrow s_y^2 = k^2.s_x^2$
4. Si se suma o se resta una constante a un conjunto de datos, se obtiene un nuevo conjunto de datos y la varianza de este último es igual a la varianza del primero.

- **Desvío Estándar.** La dificultad anterior se soluciona, tomando la raíz cuadrada de la ecuación anterior. Dada la variable X, su desvío estándar s_x es la raíz cuadrada de la varianza s_x^2.

Datos sin agrupar	Datos agrupados
$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$	$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 f_i}$

El desvío estándar es una medida de dispersión absoluta porque su valor numérico está expresado en la misma unidad de medida de la variable. Esta medida, además, es la adecuada para establecer la variabilidad de los valores observados con respecto a la media aritmética. En distribuciones normales, resulta que:

a) El 68% de los datos están entre un desvío estándar a la izquierda de la media y un desvío estándar a la derecha.
b) El 95% de los datos están entre dos desvíos estándar a la izquierda de la media y dos desvíos estándar a la derecha.
c) El 99% de los datos están entre tres desvíos estándar a la izquierda de la media y tres desvíos estándar a la derecha.

- **Coeficiente de variación.** Se calcula como cociente entre el desvío estándar y la media, y lo denotamos por CV. El coeficiente de variación es un número puro desprovisto de magnitud. Es una medida de dispersión relativa. Su valor numérico permite establecer criterios generales acerca de la homogeneidad de los datos, de la representatividad de la media aritmética y la comparación de variabilidad de otras variables, aunque las unidades de medidas o las magnitudes sean distintas. $CV\% = \frac{s_x}{\underline{x}}.100$

Hemos visto que las medidas de centralización y dispersión nos dan información sobre una muestra. Nos podemos preguntar si tiene sentido usar estas magnitudes para comparar dos poblaciones. Por ejemplo, si nos piden comparar la dispersión de los pesos de las poblaciones de estudiantes de dos escuelas diferentes, el desvío estándar nos dará información útil.

¿Pero qué ocurre si lo que comparamos es la altura de unos estudiantes con respecto a su peso? Tanto la media como la desviación estándar, $\underline{x}$ y s_x, se

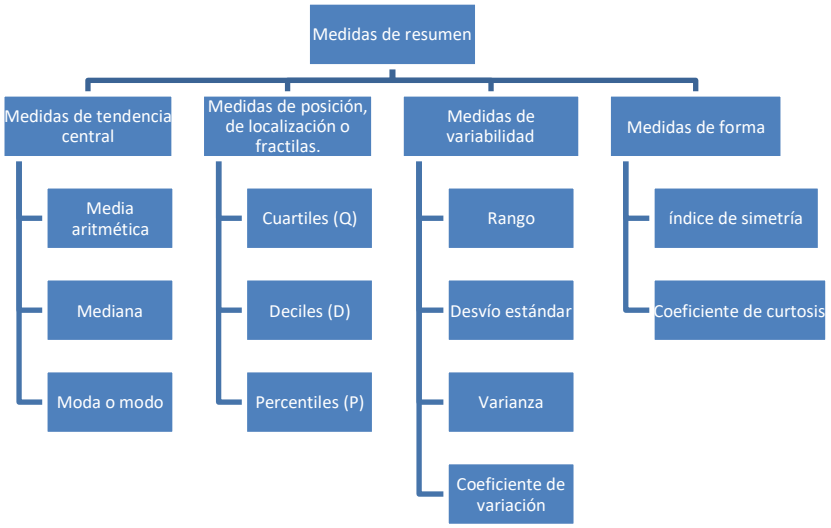
expresan en las mismas unidades que la variable. Por ejemplo, en la variable altura podemos usar como unidad de longitud el metro y en la variable peso, el kilogramo. Comparar una desviación (con respecto a la media) medida en metros con otra en kilogramos no tiene ningún sentido.

El problema no deriva sólo en que una de las medidas sea de longitud y la otra sea de masa. El mismo problema se plantea si medimos cierta cantidad, por ejemplo, la masa de dos poblaciones, pero con distintas unidades. Por ejemplo, es el caso en que comparamos el peso en toneladas de una población de 100 elefantes con el correspondiente en miligramos de una población de 50 hormigas.

El problema no se resuelve tomando las mismas escalas para ambas poblaciones. Por ejemplo, se nos puede ocurrir medir a las hormigas con las mismas unidades que los elefantes (toneladas). Si la ingeniería genética no nos sorprende con alguna novedad, lo lógico es que la dispersión de la variable peso de las hormigas sea prácticamente nula (¡Aunque haya algunas que sean 1.000 veces mayores que otras!).

En los dos primeros casos mencionados anteriormente, el problema viene de la dimensionalidad de las variables, y en el tercero de la diferencia enorme entre las medias de ambas poblaciones. El coeficiente de variación es lo que nos permite evitar estos problemas, pues elimina la dimensionalidad de las variables y tiene en cuenta la proporción existente entre media y desviación estándar.

En el siguiente cuadro vamos a encontrar la categorización de todas las medidas que hemos estudiado.



Actividad.

En el software estadístico InfoStat, abra una nueva tabla para analizar la variable de la base de datos "Colesterol en niños" disponible en: https://docs.google.com/spreadsheets/d/1FJTq6dNgyQzuotjWEjXIBH-hdqQwUNGF/edit?usp=drive_link&ouid=112081569025043459784&rtpof=true&sd=true

i. Construya las tablas de frecuencia para cada caso, distinguiendo entre variables cualitativas y cuantitativas.
ii. Proponga los gráficos que se correspondan con las tablas según el tipo de variable para cada caso.
iii. Calcule e interprete las medidas de resumen.
iv. Realice un Boxplot e interprete.
v. ¿Existe algún dato atípico dentro del conjunto de datos? Fundamente su respuesta.

Para no extendernos más, les dejo dos links para que puedan indagar acerca de las medidas de asimetría y curtosis, por un lado y por el otro, acerca de la construcción del gráfico de cajas o también conocido como Boxplot.

3. Actividad de Integración:

Para esta clase, les proponemos realizar el cuestionario que se propone el campus donde se recuperan los contenidos que hemos desarrollado y se vinculan con algunas actividades prácticas. Sugerimos realizar las actividades que proponen para luego realizar el cuestionario que proponemos en el campus.

4. Cierre:

En la tercera clase, estudiamos las medidas de resumen en general (vinculadas con la etapa del método estadístico que hemos denominado síntesis). Hemos profundizado sobre aquellas medidas que son más usuales cuando queremos caracterizar a un conjunto de datos. Si bien son varias particularidades, sería recomendable armar una red que donde puedan ubicar los casos de cada una de las medidas según las consideraciones necesarias.

Aprovecho este espacio para saludarlos y alentarlos para el desarrollo del trabajo que se propone para esta clase. Nos seguimos encontrando en el aula. Saludos cordiales.

5. Bibliografía:

- García, J; López, N; Calvo, J. (2011). Estadísticas Básicas para Estudiantes de Ciencias. Facultad de Ciencias Físicas Universidad Complutense de Madrid. España.
- Wackerly, D; Mendenhall, W; Schaeffer, R. (2010). Estadística Matemática con Aplicaciones. 7ma Ed. Cengage Learning. Santa Fe, México.