FOLLETO DE:

ESTADÍSTICA Y PROBABILIDAD

MATEMÁTICA-PEA

REALIZADO POR:
Emmanuel Ulloa Campos
Andrés Jiménez Aragón

VARIABILIDAD

La variabilidad es una parte fundamental del análisis de datos, ya que nos permite entender cómo los valores en un conjunto de datos se distribuyen y cómo pueden cambiar en diferentes situaciones. La variabilidad puede ser medida y cuantificada a través de varias herramientas estadísticas, como el rango, la desviación estándar, la varianza y otros métodos que estudiaremos en este folleto.

La noción de variabilidad es esencial para comprender la diversidad, la incertidumbre y las diferencias inherentes en una amplia gama de fenómenos, desde el comportamiento humano hasta la naturaleza misma.

INTRODUCCIÓN A LAS MEDIDAS DE VARIABILIDAD

Las medidas de variabilidad son herramientas estadísticas que se utilizan para cuantificar y describir la dispersión o extensión de un conjunto de datos. Estas medidas proporcionan información sobre cuán diferentes o dispersos están los valores individuales dentro del conjunto de datos.

EJEMPLO

Imagina que estas investigando el rendimiento de dos grupos musicales en una serie de conciertos. Tienes dos grupos, Grupo Firme y Grupo Frontera, y has recopilado la cantidad de comentarios positivos que recibió cada banda al final de sus actuaciones en diferentes conciertos.

La siguiente tabla muestra los datos recopilados en los diferentes conciertos de ambas bandas:

GRUPOS MUSICALES	COMENTARIOS POSITIVOS AL FINAL DE LAS ACTUACIONES EN DIFERENTES CONCIERTOS			
	CONCIERTO 1	CONCIERTO 2	CONCIERTO 3	CONCIERTO 4
GRUPO FIRME	320	450	290	510
GRUPO FRONTERA	280	620	310	530

Tu objetivo es analizar la variabilidad en el rendimiento de ambas bandas utilizando las medias aritméticas y determinar cuál banda es más balanceada o estable en términos de su éxito en los conciertos.

¿DE QUÉ FORMA PODEMOS MEDIR LA VARIABILIDAD?

¿De que forma podríamos medir la variabilidad que hay en la cantidad de comentarios positivos en los conciertos de Grupo Frontera y Grupo Firme?

¿A caso las medidas de posición nos pueden dar una noción de la variabilidad? ¿La media aritmética, la moda, la mediana o los cuartiles son útiles para medir la variabilidad?

Las medidas de posición, como la media aritmética, la moda, la mediana y los cuartiles, **son útiles para comprender la distribución de datos en un conjunto, pero no proporcionan una medida directa de la variabilidad de los datos**.

1. **Media aritmética:** La media aritmética es sensible a los valores extremos en los datos. Si hay una gran variabilidad en los datos, esto se puede reflejar en una mayor distancia promedio de cada punto de datos a la media. Sin embargo, la media por sí sola no proporciona una medida directa de la variabilidad, ya que puede ser engañosa si los datos están muy dispersos.

2. **Moda:** La moda representa el valor más frecuente en un conjunto de datos. Si la moda está bien definida y hay poca variabilidad en los datos, esto puede indicar que los datos tienden a agruparse alrededor de ciertos valores. Sin embargo, la moda tampoco proporciona una medida directa de la variabilidad.

3. **Mediana:** La mediana es el valor que se encuentra en el centro de un conjunto de datos ordenados. A diferencia de la media, la mediana no se ve afectada por valores extremos y proporciona una medida de la tendencia central que puede ser más robusta ante la presencia de valores atípicos. Si hay una gran variabilidad en los datos, la mediana podría estar más alejada de la media, lo que indica una distribución sesgada o una dispersión asimétrica de los datos.

4. **Cuartiles:** Los cuartiles dividen un conjunto de datos ordenados en cuatro partes iguales. Conocer los cuartiles puede ser útil para entender la variabilidad en los datos. Los cuartiles dividen un conjunto de datos ordenados en cuatro partes iguales, lo que proporciona una idea de cómo se distribuyen los datos en diferentes rangos. Sin embargo, sigue sin ser una medida exacta de la variabilidad

En resumen, aunque estas medidas de posición no proporcionan una medida directa de la variabilidad, pueden dar cierta indicación sobre la dispersión de los datos en torno a la medida central. Para una comprensión más completa de la variabilidad, es útil complementar estas medidas con medidas de dispersión específicas, como la desviación estándar, el rango o el rango intercuartílico.

MEDIDAS DE VARIABILIDAD

Las medidas de posicicón, como su nombre lo dice, representan la posición de los datos, ellas por sí solas no son tan útiles para medir la variabilidad, pero podemos definir algo llamado **medidas de variabilidad** o **fórmulas de variabilidad** para conocer la variabilidad de los datos usando las medidas de posición.

La **VARIABILIDAD** es el nombre que reciben las diferencias en el comportamiento de un fenómeno observable que se repite bajo iguales condiciones. Estas diferencias pueden ser casi imperceptibles, como en el caso de experimentos de laboratorio, donde hay un alto grado de control sobre los factores que influyen sobre el fenómeno; y pueden ser grandes, como en el caso de fenómenos en que está involucrado el comportamiento humano.

RECORRIDO O RANGO

El **rango** o **recorrido**, el cuál denotaremos como R, es la **diferencia** entre el **valor máximo y el valor mínimo** en un conjunto de datos (corresponde al valor más alto obtenido al restar dos datos). Aunque simple, el rango no proporciona una comprensión completa de cómo están distribuidos los datos.

$$R = max - min$$

En otras palabras, es la extensión total de los valores observados en los datos (indica la longitud del intervalo en el que se hallan todos los datos.). El rango proporciona información sobre cuánto se dispersan los valores alrededor de los extremos.

EJEMPLO

En un colegio hay dos secciones de undécimo año y en los exámenes aplicados de diagnóstico en la asignatura de Estudios Sociales obtuvieron las siguientes calificaciones:

Sección 11-A

90	62	74	56	96	34	56
45	80	86	54	40	80	74
86	52	78	60	90	40	56

Halle el recorrido

Recorrido = 96 – 34 = 62

Una interpretación es que el recorrido de 62 indica que la diferencia más grande entre dos calificaciones cualesquiera de la sección 11-A es de 62 puntos.

Sección 11-B

Calificaciones	Frecuencia
[60, 70[	5
[70, 80[	9
[80, 90[	4
[90, 100]	3
Total	n = 21

Halle el recorrido

Recorrido = 100 – 60 = 40

Una interpretación es que el recorrido de 40 indica que la diferencia más grande entre dos calificaciones cualesquiera de la sección 11-B es de 40 puntos.

Podríamos comparar los recorridos de ambas secciones para determinar cuál es más variable, podríamos definir, que según el recorrido, la sección más variable es la $11 - A$. A pesar de que el recorrido resulta ser una medida de variabilidad útil, no es la más confiable, ya que se puede ver afectada por datos atipícos.

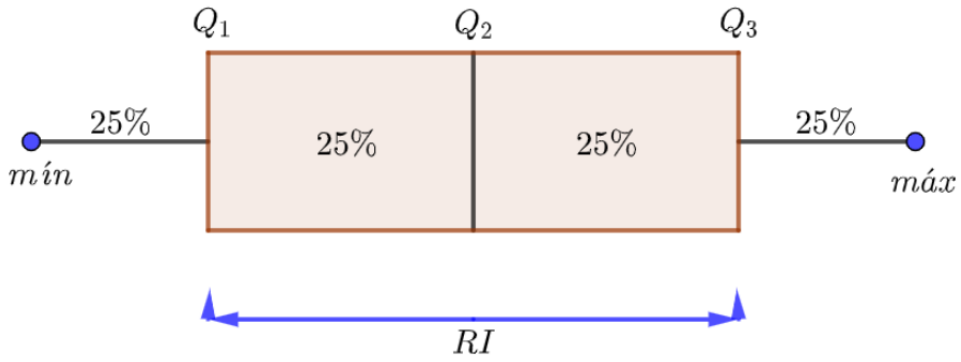
RECORRIDO INTERCUARTÍLICO

El **recorrido intercuartílico** proporciona información sobre la variabilidad de los datos en la mitad central de la distribución, **excluyendo los valores extremos**. Una interpretación común es que el recorrido intercuartílico representa la amplitud del rango de los datos donde se encuentra el 50 % central de los valores. Cuanto mayor sea el recorrido intercuartílico, mayor será la dispersión en esa porción central de los datos

El **recorrido intercuartílico**, el cuál denotaremos como RI, es la **diferencia** entre el **primer cuartil y el tercer cuartil** en un conjunto de datos. El recorrido intercuartílico, al igual que el rango, no proporciona una comprensión completa de cómo están distribuidos los datos.

$$RI = Q_3 - Q_1$$

En un diagrama de cajas:



EJEMPLO

Durante cierto fin de semana, la duración en minutos que tuvieron las películas en carteleras en todos los cines del país fueron:

120	**89**	**136**	**93**	**121**	**107**	**105**	**100**
148	**107**	**121**	**111**	**107**	**122**	**98**	

El estudiante puede notar que, al acomodar los datos:

89	**93**	**98**	**100**	**105**	**107**	**107**	**107**	**111**	**120**	**121**	**121**	**122**	**136**	**148**

Hay 15 datos, el estudiante podrá determinar que $Q_1 = 100$, $Q_2 = 107$ y $Q_3 = 121$. Por lo tanto, el recorrido intercuartílico corresponde a:

$$RI = Q_3 - Q_1 = 121 - 100 = 121$$

Significa que la diferencia entre el tercer cuartil y el primer cuartil es de 21 unidades. Esto indica que el 50 % de los datos se encuentran dentro de este rango y proporciona información sobre la dispersión de los datos centrales en el conjunto de datos.

VARIANZA

La **varianza** es una medida estadística que indica **la dispersión o variabilidad de un conjunto de datos en relación con su media aritmética**. En otras palabras, la varianza mide cuánto se alejan los valores individuales del conjunto de datos de su media, proporcionando información sobre qué tan dispersos están los datos alrededor de la media. Una varianza alta indica que los datos están más dispersos alrededor de la media, mientras que una varianza baja indica que los datos están más agrupados alrededor de la media.

La varianza es un indicador estadístico de medida de dispersión, es decir, mide la cantidad de variación de una variable estadística. Es equivalente a **la suma de los cuadrados de las diferencias de los datos y de una lista con respecto a su media y dividida por el número de datos n (si se calcula para una muestra se divide entre "$n-1$").**

La varianza utiliza todos los valores de datos, esto compensa el inconveniente del recorrido que sólo utiliza los valores máximo y mínimo del conjunto de datos, por tanto la varianza es más representativa.

VARIANZA POBLACIONAL

Si la varianza se hace con los datos de una población, entonces, la representamos con la letra griega σ y está dada por:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{X})^2}{N}$$

A veces, en vez de aparecer "N" aparece "n". Los datos son los siguientes:

- σ^2 := se refiere a la **varianza**.
- x_i := son los datos en la posición i.
- $\overline{X}$:= es el promedio de la población.
- $N = n$:= tamaño de la población.

VARIANZA MUESTRAL

Si la varianza se hace con los datos de una muestra, entonces, la representamos con la letra S y está dada por:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n-1}$$

Los datos son los siguientes:

- S^2 := se refiere a la **varianza**.
- x_i := son los datos en la posición i.
- $\overline{X}$:= es el promedio de la muestra.
- n := tamaño de la muestra.

Nosotros, por lo general, vamos a usar la varianza muestral, a excepción que en el enunciado se indique que sean datos de toda la población que se estudia.

EJEMPLO

La administración de un colegio está preocupada por el rendimiento de un grupo de estudiantes en el I parcial de Estudios Sociales, pues el promedio del grupo fue de $63,8$. Las notas fueron:

72	**60**	**65**	**18**	**76**	**73**	**70**	**60**	**74**	**70**

Un estudiante, en defensa del grupo, aclaró que ese promedio no refleja el rendimiento de la mayoría, puesto que existe una nota muy baja, de un compañero que no estudió para la prueba porque se iría a vivir a otro país. Realice un análisis de variabilidad, con la varianza, tomando en cuenta las notas, luego otro, en el que se elimina la nota más baja.

Vamos a explicar un proceso para hallar la varianza de manera ordenada y es la siguiente:

Posición	$\boldsymbol{n}$	$\boldsymbol{n-1}$	$\boldsymbol{x_i}$	$\boldsymbol{\overline{x}}$	$\boldsymbol{x_i - \overline{x}}$	$\boldsymbol{(x_i - \overline{x})^2}$
1	10	9	18	$63,8$	$-45,8$	$2097,64$
2	10	9	60	$63,8$	$-3,8$	$14,44$
3	10	9	60	$63,8$	$-3,8$	$14,44$
4	10	9	65	$63,8$	$1,2$	$1,44$
5	10	9	70	$63,8$	$6,2$	$38,44$
6	10	9	70	$63,8$	$6,2$	$38,44$
7	10	9	72	$63,8$	$8,2$	$67,24$
8	10	9	73	$63,8$	$9,2$	$84,64$
9	10	9	74	$63,8$	$10,2$	$104,04$
10	10	9	76	$63,8$	$12,2$	$148,84$
–					$\boldsymbol{\sum_{i=1}^{n} (x_i - \overline{X})^2 \Longrightarrow}$	$2609,6$

PD: cuando aparece el x_i se deben ordenar los datos de menor a mayor, x_1 es el mínimo, x_2 es el segundo dato, y así consecutivamente, hay que hacer tantas filas como número de datos. La sumatoria se hace hasta haber recolectado todos los cuadrados. Finalmente, se haya la varianza:

$$(S_1)^2 = \frac{\boldsymbol{\sum_{i=1}^{n} (x_i - \overline{X})^2}}{\boldsymbol{n-1}} = \frac{2609,6}{9} = 289,96$$

Hemos puesto S_1 debido a que tendremos que calcular dos varianzas. Ahora, el estudiante podría verificar que al eliminar el dato más pequeño, el promedio cambia y debe hacer una tabla respectiva, en la cuál debe obtener que:

$$(S_2)^2 = \frac{\boldsymbol{\sum_{i=1}^{n} (x_i - \overline{X})^2}}{\boldsymbol{n-1}} = \frac{278,89}{8} = 34,86$$

Note que es evidente que el dato $x_1 = 18$ es un valor atipíco que afecta el medir la variabilidad utilizando la varianza. Es decir, el estudiante que salió en defensa del grupo acertó, el promedio no refleja el rendimiento de la mayoría.

DESVIACIÓN ESTÁNDAR

La **desviación estándar** es una medida estadística que **indica la dispersión o variabilidad de un conjunto de datos en relación con su media aritmética**. Es una de las medidas más utilizadas para comprender **cuánto se alejan los valores individuales del conjunto de datos de su media**, de una manera más intuitiva que la varianza. La desviación estándar se expresa en las mismas unidades que los datos originales, lo que facilita su interpretación. Una desviación estándar alta indica que los datos están más dispersos alrededor de la media, mientras que una desviación estándar baja indica que los datos están más agrupados alrededor de la media.

Como la varianza está **contituida por la suma de cuadrados de las desviaciones, las unidades de medida que tienen los datos quedan al cuadrado; para simplificar esto se acostumbra obtener la raíz cuadrada de la varianza**. Esa nueva medida, se llama **desviación estándar**.

Recordemos que hay **varianza poblacional** y **varianza muestral**, esto implica que hay dos tipos de desviaciones estándar:

DESVIACIÓN ESTÁNDAR POBLACIONAL

Si la varianza se hace con los datos de una población, entonces, la representamos con la letra griega σ y, por ende, la desviación estándar está dada por:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \overline{X}\right)^2}{N}}$$

A veces, en vez de aparecer "N" aparece "n". Los datos son los siguientes:

- $\sigma :=$ se refiere a la **desviación estándar**.
- $\sigma^2 :=$ se refiere a la **varianza**.
- $x_i :=$ son los datos en la posición i.
- $\overline{X} :=$ es el promedio de la población.
- $N = n :=$ tamaño de la población.

DESVIACIÓN ESTÁNDAR MUESTRAL

Si la varianza se hace con los datos de una muestra, entonces, la representamos con la letra S y está dada por:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}{n-1}}$$

Los datos son los siguientes:

- $S :=$ se refiere a la **desviación estándar**.
- $S^2 :=$ se refiere a la **varianza**.
- $x_i :=$ son los datos en la posición i.
- $\overline{X} :=$ es el promedio de la muestra.
- $n :=$ tamaño de la muestra.

Nosotros, por lo general, vamos a usar la desviación estándar muestral, a excepción que en el enunciado se indique que sean datos de toda la población que se estudia.

EJEMPLO

A un grupo de estudiantes de la clase de matemática, se les preguntó cuántas horas dedican por semana para el estudio de la materia. Sus respuestas fueron:

2	**4**	**7**	**8**	**9**

Determine la desviación estándar del grupo.

Debemos hallar la varianza, para ello, procedemos de la misma manera que anteriormente hemos explicado:

Posición	$\boldsymbol{n}$	$\boldsymbol{n-1}$	$\boldsymbol{x_i}$	$\boldsymbol{\overline{x}}$	$\boldsymbol{x_i - \overline{x}}$	$\boldsymbol{(x_i - \overline{x})^2}$
1	5	4	2	6	-4	16
2	5	4	4	6	-2	4
3	5	4	7	6	1	1
4	5	4	8	6	2	4
5	5	4	9	6	3	9
$-$					$\boldsymbol{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2 \Longrightarrow}$	34

De este modo, tenemos que la varianza corresponde a:

$$S^2 = \frac{\boldsymbol{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}}{\boldsymbol{n-1}} = \frac{34}{4} = 8,5$$

Finalmente, la desviación estándar, es la raíz de la varianza, por lo que:

$$S = \sqrt{S^2} = \sqrt{8,5} = 2{,}92$$

La desviación estándar de la muestra calculada es aproximadamente 2,92 horas. Esto significa que, en promedio, las horas dedicadas al estudio de matemáticas por semana varían alrededor de 2,92 horas con respecto a la media de 6 horas.

¿QUÉ MEDIDA DE VARIABILIDAD ES MÁS CONFIABLE O ÚTIL?

Si estás buscando una medida rápida y simple de la dispersión, el rango puede ser útil. Si deseas considerar tanto la dispersión como la relación con la media, la desviación estándar es una buena opción. La varianza es útil para cálculos más avanzados. En última instancia, la elección de la medida depende de tus necesidades analíticas y del nivel de detalle que requieras en tu análisis.

ESTANDARIZACIÓN

La **estandarización** es un proceso en estadística que se utiliza para **transformar una variable de tal manera que tenga una media de 0 y una desviación estándar de 1**. Este proceso es útil en varios contextos, incluyendo el análisis de datos, la comparación entre diferentes conjuntos de datos y la aplicación de ciertos modelos estadísticos.

También se le llama **tipificación** y se utiliza para comparar datos procedentes de diferentes muestras o poblaciones. Se recurre a la tipificación para transformar el dato original a uno estandarizado, llamado valor "z".

$$z = \frac{\text{Dato} - \text{Promedio}}{\text{Desviación Estándar}} = \frac{x - \overline{x}}{s}$$

Después de aplicar este proceso, los valores resultantes tendrán una media de 0 y una desviación estándar de 1. Esto simplifica la interpretación y comparación de los datos, ya que los valores estandarizados representan cuántas desviaciones estándar están alejados de la media original.

EJEMPLO

Alberto y María realizan una prueba y obtienen los siguientes resultados.

Sección	Estudiante	Puntuación	Media	Desviación Estándar
$7-A$	Alberto	84	78	$8,1$
$7-B$	María	75	71	$4,9$

¿Cuál de los alumnos obtuvo una puntuación mejor al compararla con el resto de los compañeros?

Esta es una simple pregunta de estandarización, aquí nos dan la media y la desviación estándar de cada sección, por lo que el proceso no será tan largo, de modo que:

$$z_{\text{Alberto}} = \frac{84 - 78}{8,1} = 0,74$$

$$z_{\text{María}} = \frac{75 - 71}{4,9} = 0,82$$

Al comparar las puntuaciones estandarizadas, vemos que María tiene una puntuación estandarizada ligeramente más alta que Alberto. Esto significa que María obtuvo una puntuación relativamente mejor en comparación con el resto de sus compañeros en su respectiva sección.

COEFICIENTE DE VARIACIÓN

El **Coeficiente de Variación** **realiza comparaciones útiles respecto a la dispersión de los datos entre dos o más poblaciones diferentes en relación con el nivel general de los valores y la media de cada conjunto**.

El Coeficiente de Variación se denota con el símbolo CV. A mayor valor del CV, mayor heterogeneidad de los valores de la variable; y a menor CV, mayor homogeneidad en los valores de la variable.

Es una medida de la dispersión relativa de un conjunto de datos en relación con su media. **Se calcula como la desviación estándar dividida por la media, y se expresa como un porcentaje**.

$$CV = \left(\frac{\text{Desviación Estándar}}{\text{Media Muestral}}\right) \cdot 100\,\% = \left(\frac{s}{\overline{x}}\right) \cdot 100\,\%$$

El coeficiente de variación es útil para comparar la variabilidad de diferentes conjuntos de datos que pueden tener diferentes unidades o escalas.

EJEMPLO

El profesor de matemática de un colegio técnico desea hacer un estudio comparativo con respecto al rendimiento de los estudiantes en las materias de informática y logística. Los datos se resumen en la siguiente tabla

	Informática	**Logística**
Varianza	$25,15$	$22,18$
Promedio	$88,22$	$87,01$

Determine:

(a) Los coeficientes de variación para cada uno de los grupos.

(b) ¿Cuál grupo presenta menos variabilidad relativa?

En este caso, nos dan la varianza y el promedio, hay que hallar la desviación estándar de ambas materias, pero recordemos que es la raíz cuadrada de la varianza, de modo que:

$$CV_{\text{Informática}} = \left(\frac{\sqrt{25,15}}{88,22}\right) \cdot 100\,\% = 5,68\,\%$$

$$CV_{\text{Logística}} = \left(\frac{\sqrt{22,18}}{87,01}\right) \cdot 100\,\% = 5,41\,\%$$

Para determinar cuál grupo presenta menos variabilidad relativa, simplemente comparamos los coeficientes de variación. En este caso, el grupo de Logística tiene un coeficiente de variación ligeramente menor (5,41 %) que el grupo de Informática (5,69 %). Por lo tanto, el grupo de Logística presenta menos variabilidad relativa en comparación con el grupo de Informática.