B. MEDIDAS DE DISPERSIÓN O VARIABILIDAD

Supongamos que en nuestra planta fabricamos engranajes que, entre otras especificaciones, deben cumplir con cierto diámetro interno (en milímetros) para que cumplan el propósito para el cual fueron fabricados. Supongamos también que disponemos de tres máquinas (A, B, C) que se encargan de fabricarlos.

Una forma de controlar que los engranajes cumplan con las especificaciones del diámetro interno es tomar muestras periódicas de los engranajes fabricados por cada máquina, calcular los diámetros en esas muestras, es decir

$$\bar{x}_A, \bar{x}_B, \bar{x}_C$$

Y luego determinar si hay razones para pensar que se encuentra fuera de cierto intervalo de control (pero esto es tema de Estadística II)

Por ahora se asumirá que cada máquina debe estar ajustada para que el diámetro tenga el menor error posible y sea cierto valor D. Como es el mismo engranaje fabricado por tres máquinas distintas, este diámetro D debe ser el mismo en los tres casos. Siempre habrá mínimas variaciones (en el mejor de los casos) y tendremos piezas con un diámetro menor y otras con un diámetro mayor que mientras estén dentro del margen de tolerancia serán piezas útiles. Dejemos de lado este detalle y simplemente pensemos que habrá engranajes con diámetros menores a D y engranajes con diámetros mayores a D. Pero ¿qué es D? No es otro que el diámetro PROMEDIO que debe representar al diámetro de TODOS los engranajes fabricados por cada máquina.

Así D = μ y

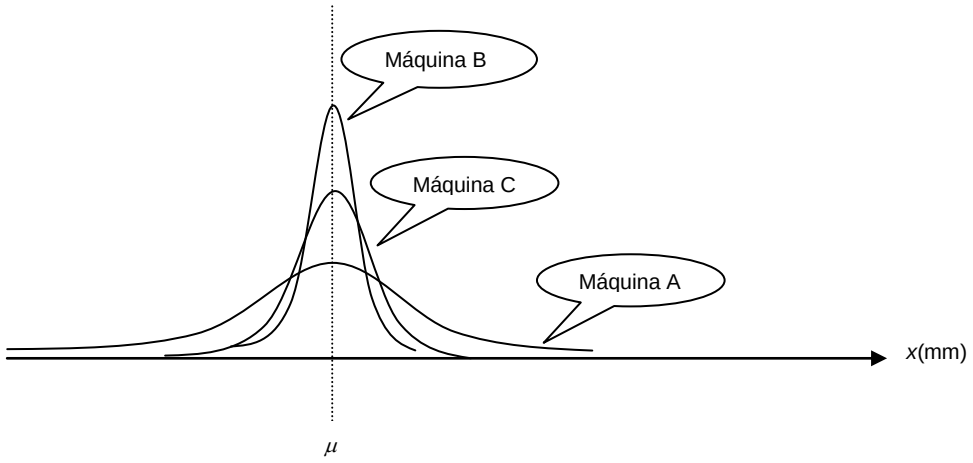
$\mu = \mu_A = \mu_B = \mu_C$

Nos interesa el diámetro promedio pues en su papel de medida de tendencia central nos dice alrededor de qué valor se concentran los diámetros de

todos los engranajes fabricados (nótese que hemos usado el símbolo de media poblacional)

Si esto se cumple, es decir, si las tres máquinas fabrican engranajes cuyo diámetro promedio μ se ajusta a las especificaciones ($\mu_A = \mu_B = \mu_C = \mu$) ¿podemos darnos por satisfechos de que las máquinas fabrican engranajes adecuados?

Para responder esto imaginemos una representación del comportamiento de cada máquina mediante histogramas para cada una (simplificaremos los histogramas con curvas suaves las cuales siguen representando la frecuencia con la que aparecen los engranajes según su diámetro)



Para simplificar aun más la ejemplificación se han elegido distribuciones aproximadamente simétricas pero esto no necesariamente tiene que ser así.

En el eje x se encuentra la variable diámetro (en milímetros) y en él todos los posibles diámetros que tendría cada uno de los engranajes fabricados. Las tres máquinas fabrican piezas con el mismo diámetro promedio

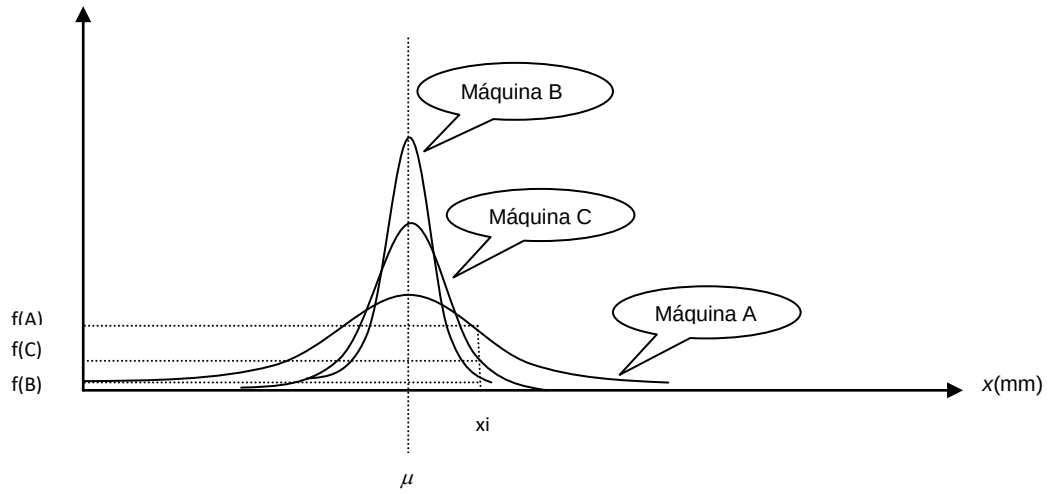
$$(\mu_A = \mu_B = \mu_C = \mu)$$

lo cual se refleja en la imagen en la línea vertical que pasa por el centro de las tres distribuciones (¿Por qué la media pasa por el centro de simetría en una distribución simétrica? Para responder esto recordemos que la media es

el punto de equilibrio del conjunto de datos y preguntémonos dónde se encontraría si la distribución fuera sesgada)

¿Las tres máquinas, A, B y C operan adecuadamente? Si solamente nos guiáramos por el diámetro promedio μ diríamos que sí, dado que todas fabrican engranajes con el diámetro según especificaciones. Sin embargo, puede notarse que las gráficas son distintas para cada una, exceptuando la simetría. Poniéndolo en palabras simples, algunas curvas son más "puntiagudas" y otras más "achatadas".

¿Qué implica esto? Supongamos que nos preguntamos con qué frecuencia aparecerán engranajes cuyo diámetro está por encima de la media. Este sería un valor x_i cualquiera en el eje x, tal como se muestra a continuación:



En el eje vertical se miden las frecuencias de aparición de los resultados de la variable; entonces para un diámetro x_i cualquiera, por encima del diámetro promedio, la cantidad de veces o frecuencia con que esto sucede será mayor en la máquina A, un poco menor en la C y menos (la menor) en la B. Esto podría indicar que la máquina A tendrá una mayor cantidad de desechos de piezas fabricadas por exceso de diámetro y la B habrá la menor.

Idéntica situación ocurriría si hubiésemos estudiado qué ocurriría con un diámetro x_i cualquiera por debajo de la media. La máquina B tendría la menor cantidad de piezas de desecho por diámetros pequeños.

Esta información no nos la dio la media. Si solamente nos quedáramos con la información proporcionada por ella podríamos haber tomado la mala decisión de no corregir la operación de la máquina A (y eventualmente la de la C) con la consecuente pérdida económica por una gran cantidad de desechos.

Lo que no refleja la media (ni ninguna otra medida de tendencia central) es qué ocurre con los datos alrededor de cierto valor de interés central, en este caso la media misma. En la curva A los datos parecen estar, en conjunto, mucho más alejados del centro (de la media) que en B y en C. Parecen estar menos concentrados o más **dispersos**. La curva es más "achatada". Por el contrario, en la curva B los datos parecen estar más concentrados o menos **dispersos**, en conjunto, con respecto al centro (la media). La curva es más "puntiaguda". En la máquina A hay una gran variabilidad de diámetros, lo cual no es para nada bueno, mientras que en la máquina B ocurre todo lo contrario, la variabilidad es mucho menor.

A continuación veremos un conjunto de medidas de variabilidad o dispersión que complementan a las de tendencia central para describir adecuadamente el comportamiento de un conjunto de datos.

I. Rango o recorrido

Esta es una medida de dispersión muy sencilla pero también muy limitada. El rango se define como la diferencia entre el máximo valor de la variable y el mínimo.

$$R = x_{max} - x_{min}$$

Es la distancia que recorre la variable entre sus extremos y nos da una idea de cuán dispersos están los valores de la misma.

Por ejemplo, si en la comisión 1 un profesor tiene estudiantes cuyas edades van de 18 a 22 años y en la comisión 2 de 19 a 25 los respectivos rangos para la variable "Edad de los estudiantes de la comisión" serán

R_1 = 22-18 = 4 (años)

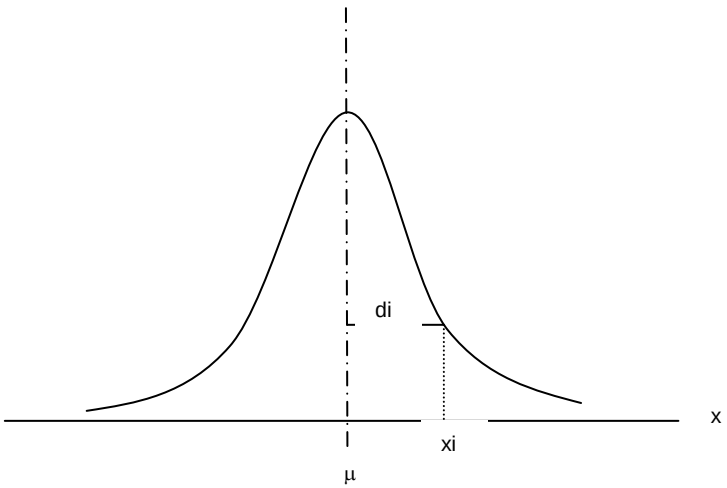
R_2 = 25 – 19 = 6 (años)

En la comisión 2 la variable edad se extiende más que en la 1 y las edades estarán seguramente más alejadas en conjunto de la edad promedio, cualquiera sea esta.

Sin embargo, no tenemos una buena idea de cómo están los datos (las edades en este caso) realmente agrupados alrededor de la media. El rango no nos ofrece esta posibilidad. Para ello definiremos la siguiente medida.

II.<u>Varianza</u>:

Nuestra intención es averiguar si los datos en una distribución de frecuencias se encuentran alejados o cerca, en conjunto, del centro. Para nosotros ese centro será siempre la media (ya sea que se encuentre verdaderamente en el centro de la distribución o no). Veámoslo gráficamente.



Por simplicidad nuevamente se elige una distribución simétrica pero esto no es necesario y el concepto tanto como la fórmula que encontraremos se aplican en cualquier caso. Como nos interesa cuán alejados de la media μ están los datos, medimos la distancia d_i a ella (una distancia genérica para cualquier dato x_i). Es decir:

$$d_i = x_i - \mu$$

El paso lógico siguiente sería promediar todas estas distancias, pues no nos interesan las distancias una a una sino un valor que las represente a todas y ya sabemos que la media es una buena herramienta para ello, por lo que sería útil tener un promedio de las distancias (o desvíos) con respecto a la media, así que a la suma de todas las distancias la dividimos por la cantidad de ellas. Es decir :

$$\sum_{i=1}^{N} d_i = \sum_{i=1}^{N} (x_i - \mu)$$

$$\frac{\sum_{i=1}^{N} d_i}{N} = \frac{\sum_{i=1}^{N} (x_i - \mu)}{N}$$

Tendríamos entonces un "promedio de los desvíos o distancias". Cuanto mayor fuera este promedio, mayor sería la dispersión de los datos. Pero esta expresión presenta un inconveniente importante: la suma de los desvíos siempre es igual a cero; esto es así porque al ser la media el centro de equilibrio del sistema, todas las diferencias con ella se cancelan entre sí (las negativas con las positivas), o sea que

$$\sum_{i=1}^{N} d_i = \sum_{i=1}^{N} (x_i - \mu) = 0$$

por lo que no podríamos calcular la dispersión.

Para ilustrar esto recordemos el ejemplo de las calificaciones en el tema de Medidas de Tendencia Central

nº	xi	xi - $\bar{x}$	$(xi - \bar{x})^2$	
x_1	9	2,25	5,0625	
x_2	5	-1,75	3,0625	
x_3	7	0,25	0,0625	
x_4	6	-0,75	0,5625	
x_5	4	-2,75	7,5625	
x_6	4	-2,75	7,5625	
x_7	10	3,25	10,5625	
x_8	9	2,25	5,0625	
$\bar{x}$	6,75	$\Sigma = 0$	$\Sigma = 39,5$	
			$s^2 = 39,5/7 =$	5,64

La tercera columna de esta tabla muestra las diferencias entre cada calificación y el promedio. Al final de la misma la suma de todas estas diferencias da cero. Esto ocurre siempre, sin importar cuál sea la media y si la distribución de datos es simétrica o sesgada; la media siempre es el punto de equilibrio.

Una simple demostración comprueba esto para cualquier caso.

$$\sum_{i=1}^{N}(x_i-\mu)=x_1-\mu+x_2-\mu+x_3-\mu+\cdots+x_N-\mu$$

$$=x_1+x_2+x_3+\cdots+x_N-\mu-\mu-\mu-\cdots-\mu$$

$$=\sum_{i=1}^{N}x_i-N\mu$$

y recordando que

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

entonces despejando y reemplazando queda

$$=N\mu-N\mu=0$$

(este procedimiento es igualmente válido con la media muestral)

Una manera de evitar este inconveniente es elevando al cuadrado estas diferencias

$$\sum_{i=1}^{N}(d_i)^2=\sum_{i=1}^{N}(x_i-\mu)^2$$

logrando de este modo que las diferencias sean siempre positivas y no se cancelen (teniendo además que las diferencias grandes se notarán más, o sea, reflejarán mejor la dispersión). Entonces, al dividir la suma de todas estas diferencias elevadas al cuadrado por la cantidad total de ellas tenemos:

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

Esta magnitud encontrada (se lee "sigma cuadrado"), llamada varianza, refleja muy bien la dispersión de un conjunto de datos, y es de hecho **el promedio del cuadrado de las diferencias de los datos con respecto a su media**. La expresión aquí obtenida es en realidad la varianza poblacional (notar el símbolo griego utilizado para ella), es decir, la varianza de toda la población con respecto a la media poblacional y es única y es la verdadera varianza.

Al igual que con la media, también definimos la varianza muestral, la cual es

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

Nótese que se ha reemplazado la media poblacional μ por "equis raya" y el tamaño de la población N por el tamaño muestral n. También el símbolo deja de ser una letra griega para ser una letra latina. Al igual que la media muestral, hay tantas varianzas muestrales como muestras y todas serán distintas entre sí e independientes (y si hemos hecho bien las cosas serán una buena estimación de la verdadera, la poblacional)

Pero hay un cambio muy importante además con respecto a la varianza poblacional: la varianza muestral se obtiene promediando por las n observaciones de la muestra menos 1 (observar el denominador). La razón estadística para esto será explicada en inferencia estadística (habrá que esperar hasta la UT-5) pero por ahora podemos tomar como explicación suficiente que es necesario proceder de esta manera si esperamos que la varianza de la muestra sea una buena estimación de la varianza poblacional.

Así, volviendo al ejemplo de las calificaciones, la cuarta columna muestra el cuadrado de las diferencias de cada nota con respecto a la media (o sea el resultado de las filas de la columna tres elevado al cuadrado). Luego la suma de esta columna da 39,5 y al dividirlo por n – 1 observaciones (8 – 1 =7) obtenemos la varianza 5,64.

La interpretación de este resultado es "el promedio de los desvíos cuadráticos de las calificaciones con respecto a la calificación 6,75 (la media) es 5,64 puntos^2"

Como se ve, tiene el inconveniente de tener sus unidades elevadas al cuadrado (así, si estamos midiendo alturas de personas, la unidad será en metros cuadrados), lo cual no tiene interpretación en el contexto y sólo logra confundir. Veremos ahora una solución a este problema.

III. **Desvío estándar**

Si calculamos la raíz cuadrada de la varianza, una vez hallada ésta, la expresión queda

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

siendo ésta el **desvío estándar poblacional** (se lee simplemente "sigma") y se interpreta como *el promedio de las distancias o desvíos de los valores con respecto a la media del conjunto de datos*, y las unidades vuelven a ser las de la variable.

En el caso de la muestra

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

El símbolo "s" indica que se trata del desvío estándar muestral.

Ejemplo:

Se hacen cinco mediciones de la longitud de una pieza de aluminio (en mm) luego de calentarla, registrándose los siguientes datos:

Intento		Valor medido		Desviación		Desviación al cuadrado
i		x_i		d_i		$(d_i)^2$
1		71		-0,8		0,64
2		72		0,2		0,04
3		72		0,2		0,04
4		73		1,2		1,44
5		71		-0,8		0,64
	$\Sigma x_i=$	359	$\Sigma d_i=$	0	$\Sigma d_i^2=$	2,8
	$\Sigma x_i/N=$	71,8			$\Sigma d_i^2/N=$	0,56

Se ha asumido valores poblacionales pues es sólo un ejemplo ilustrativo.

μ = 71,8 mm "el promedio de las mediciones de longitud de pieza de aluminio luego de calentarla es de 71,8 mm"

Me = 72 mm "la mediana de las mediciones de longitud de la pieza de aluminio luego de calentarla es 72mm" (recordar que este cálculo se efectúa luego de ordenar todos los datos de menor a mayor y encontrar la observación que está en la mitad del conjunto de datos).

Mo_1 = 71 mm; Mo_2 =72 mm (esta se puede considerar una distribución bimodal pero esto no aporta mucha información, principalmente debido a que son pocas observaciones)

σ^2 = 0,56 mm^2 "el promedio de los desvíos cuadráticos de las longitudes de la barra de aluminio después de calentada con respecto a la longitud media es 0,56mm^2"

$\sigma = \sqrt{0,56} = 0,75$ mm "el promedio de los desvíos de las longitudes de la barra de aluminio después de calentada con respecto a la longitud media es 0,75mm"

IV. Coeficiente de variación

Si tuviéramos que estudiar dos ríos de Argentina, por ejemplo el río Mendoza y el río Paraná para comparar su variabilidad estacional en el caudal, calcularíamos la media y el desvío estándar de cada uno.

Río	Caudal medio (m^3/s)	Desvío estándar (m^3/s)	
Mendoza	50	15	
Paraná	17.300	1.500	

Si nos preguntamos cuál de los dos ríos presenta mayor variabilidad estacional en su caudal, mirando el desvío estándar estaríamos tentados en responder que el río Paraná puesto que su desvío estándar es (mucho) mayor. Sin embargo esto no sería correcto porque estaríamos pasando por alto el hecho de que la media (el caudal promedio) también es (mucho) mayor que el del río Mendoza. En un caudal mayor, es esperable también una variación mayor. Necesitamos mirar el desvío estándar de otra manera para que la comparación que deseamos hacer sea correcta (o más "justa"). Para ello encontraremos cuánto representa el desvío estándar si se lo compara con la media mediante la razón entre ellos de la siguiente manera:

Río Mendoza: (15/50)x100=30%

Río Paraná: (1.500/17.300)x100=8,7% (aprox.)

Completando la tabla tenemos:

Río	Caudal medio (m^3/s)	Desvío estándar (m^3/s)	CV(%)
Mendoza	50	15	30
Paraná	17.300	1.500	8,7

Mientras que el desvío estándar del río Mendoza representa el 30% del caudal medio, el desvío estándar del río Paraná representa sólo el 8,7% del caudal medio, algo esperable y lógico si se tiene en cuenta la enorme diferencia de caudales entre un río y el otro.

Definimos entonces el coeficiente de variación porcentual como:

$$CV(\%) = \frac{\sigma}{\mu} \cdot 100 \text{ (para datos poblacionales)}$$

$$CV(\%) = \frac{s}{\bar{x}} \cdot 100 \text{ (para datos muestrales)}$$

Características

- Mientras que el desvío estándar es una medida de dispersión absoluta, el coeficiente de variación es **relativa** (porque está referida a la media) y es el apropiado para comparar la variabilidad de dos o más conjuntos de datos.
- A diferencia del desvío estándar, el coeficiente de variación es adimensional (no tiene unidades de medida) pues es el cociente entre dos medidas con las mismas unidades (el desvío estándar y la media)
- Puede tomar valores desde cero (cuando el desvío estándar es cero) en adelante, pasando incluso 1 (el 100%) cuando el desvío estándar es mayor que la media aunque rara vez ocurre.
- También es útil para tener una idea de si el desvío estándar es grande dentro de un mismo grupo de datos. En el caso de la barra de aluminio, CV(%)=(0,75/71,8)x100=1% y en el de las calificaciones CV(%)=(2,37/6,75)x100=35%. En el primero el desvío estándar sólo representa el 1% de la media; podemos darnos cuenta que es muy pequeño pues sólo representa el 1% de la media (por supuesto esto podría no ser así si la naturaleza del proceso indicara que este valor incluso es alto). En el segundo caso el desvío estándar represente el 35% de la calificación promedio, lo cual es un valor bastante más alto (de nuevo, esto podría tener otro peso si por ejemplo el docente que evalúa considera que tal variabilidad no es alta)

V. Puntuación Z

(Ver filmina en clase)