Análisis de Datos

Eje temático: Datos y Azar

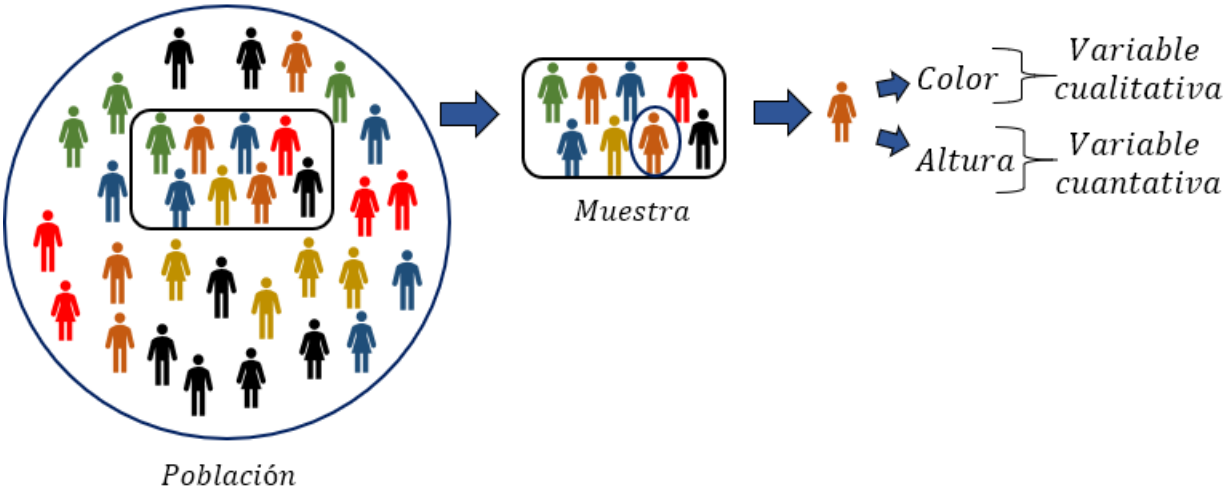
Estadística Descriptiva

Motivación

La estadística es una rama de la matemáticas que comprende métodos y técnicas que se emplean en la recolección e interpretación de datos. Esta es de gran utilidad para distintas áreas del conocimiento (como las ciencias naturales, políticas y sociales) dado que sirve para cuantificar un panorama general de la realidad, facilitando su análisis. Sin embargo, es importante conocer las limitaciones de la estadística, porque cuando mejoran las tendencias generales, no significa que todos los casos particulares hayan mejorado (pueden haber algunos que no).

¿Cuántas veces has querido calcular qué notas necesitas sacarte para obtener el promedio que quieres? Esta clase te servirá para esto y mucho más.

Definiciones



Se entiende por **población** a un conjunto cuyos elementos tienen **una o más características en común** (no necesariamente tienen que ser personas; pueden ser, por ejemplo, bolitas de colores). Además, ¡el número de características de estos elementos puede ser infinito!

Para estudiar una característica puede resultar complejo analizar a toda la población. Para ello, se define la **muestra** como un **subconjunto de la población**. Este subconjunto puede ser representativo y a la vez aleatorio. Sin embargo, siempre habrá un **error muestral**, que es la diferencia asociada al hecho de que si se considera una muestra, siempre se dejarán de lado elementos que permiten saber sobre la configuración completa de la población.

A modo de ejemplo, si se quiere saber el **promedio de altura de las personas de Chile**, podríamos ir a alguna estación de metro y promediar la altura de 1.000 personas. Las **1.000 personas serían nuestra muestra** y es probable que nos acerquemos al promedio verdadero, pero siempre habrá un error, dado que no contamos la altura de todas las personas de Chile.

Los elementos de una población poseen características (o variables) que pueden ser de dos tipos: **cualitativas y cuantitativas**.

Las **variables cualitativas** son aquellas que hacen referencias a atributos **no numéricos**, como sexo, nacionalidad, colores, entre otros. Ojo, estos elementos al no ser números, **no se pueden ordenar en la recta numérica**, es decir, que **no hay un dato mayor que otro**. ¿Se te ocurre un ejemplo?

Las **variables cuantitativas** son aquellas que hacen referencias a **atributos numéricos**, donde existen dos tipos. Las **discretas**, solo pueden tomar ciertos valores enteros (por ejemplo: número de hijos), mientas que las **continuas**, pueden tomar valores reales (por ejemplo: estatura). Para este último, se suelen utilizar intervalos $[a, b]$ para clasificarlos.

Tabulación de Datos

Para ordenar datos estadísticos, muchas veces utilizamos tablas. A continuación se presentan algunos de las definiciones y símbolos que se utilizan para este fin.

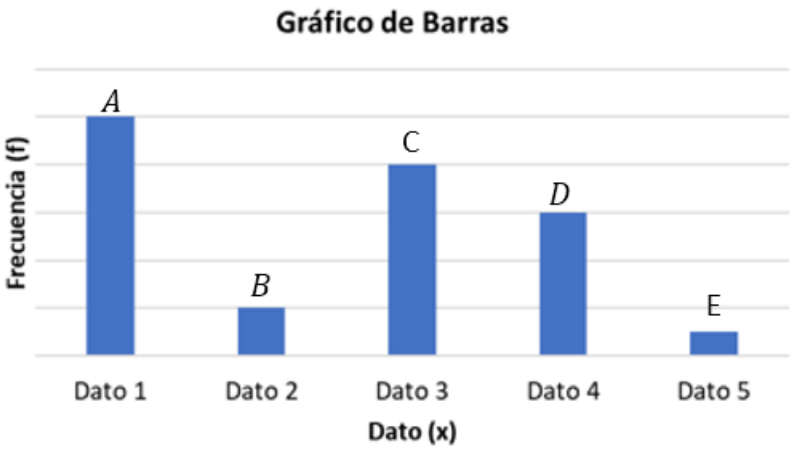
Dato	(x_i)	Es algún **valor específico** del criterio que se estudia. La i corresponde a los valores que puede tener.
frecuencia (absoluta)	(f)	Número de veces que se **repite** un dato.
Frecuencia acumulada	(F_{ac})	Es la que se obtiene al **sumar las frecuencias absolutas anteriores**.
Frecuencia relativa	(fr)	Es la división entre la frecuencia absoluta y el total de datos. ¡Es la **proporción**!
Marca de clase	(c_i)	Es el **valor medio de un intervalo**, se calcula promediando los extremos del intervalo.
Amplitud de Intervalo	-	Es la resta del extremo derecho con el extremo izquierdo del intervalo. Así se tiene el espectro del intervalo.

Representación Gráfica

Es una manera de representar gráficamente el estudio estadístico. Entre los gráficos más utilizados encontramos:

Gráfico de Barras

Dato	Frecuencia (f)
Dato 1	A
Dato 2	B
Dato 3	C
Dato 4	D
Dato 5	E



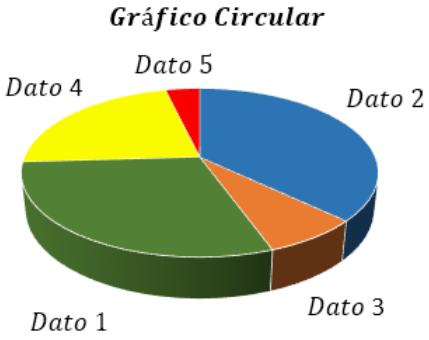
En este tipo de gráfico, se puede observar que **las barras indican la frecuencia de cada una de las variables**, permitiéndonos **comparar** las variables a partir de su altura. Estos tipos de gráficos suelen utilizarse para representar **datos cualitativos**.

Gráfico Circular

Dato	Frecuencia (f)	Frecuencia relativa (fr)
Dato 1	A	a %
Dato 2	B	b %
Dato 3	C	c %
Dato 4	D	d %
Dato 5	E	e %

$$\frac{f}{Total} = \frac{x°}{360°}$$

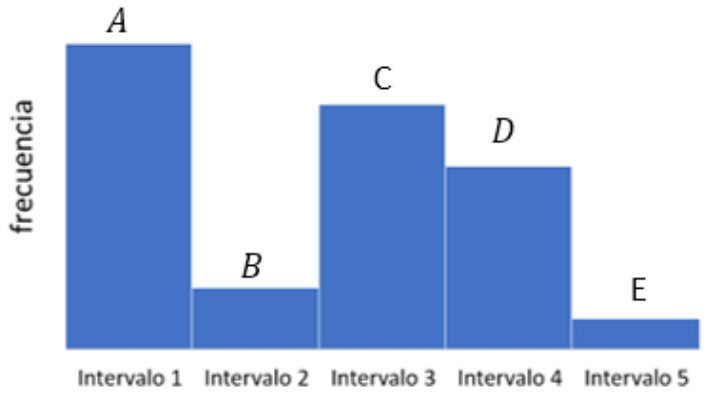
$$\frac{f}{Total} = \frac{fr}{100\ \%}$$



Los gráficos circulares nos permiten observar **cómo se reparten las diferentes variables en el total de de datos**. Es decir, mientras más grande sea una sección circular (o el trozo de la torta), mayor será la proporción que esta tendrá con respecto al total (si un dato cubriese la mitad, entonces la mitad del total serían de ese dato).

Histograma

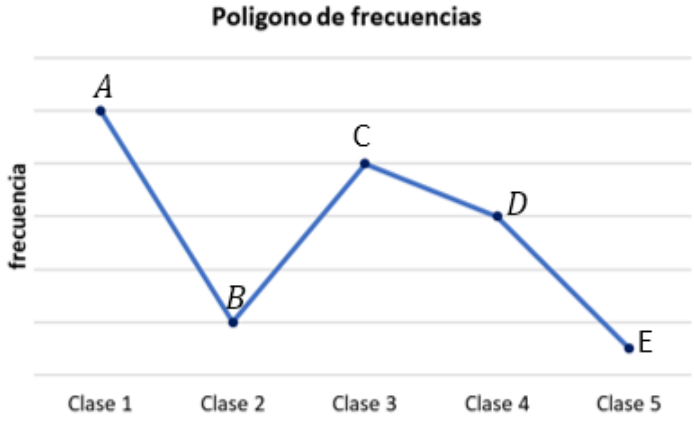
Dato	Frecuencia (f)
Intervalo 1	A
Intervalo 2	B
Intervalo 3	C
Intervalo 4	D
Intervalo 5	E



Los histogramas son muy similares a los gráficos de barras, pero son utilizados para **variables continuas**, donde las variables son **intervalos de datos**. Esto puede ser útil para representar estaturas, pesos y más.

Polígono de frecuencias

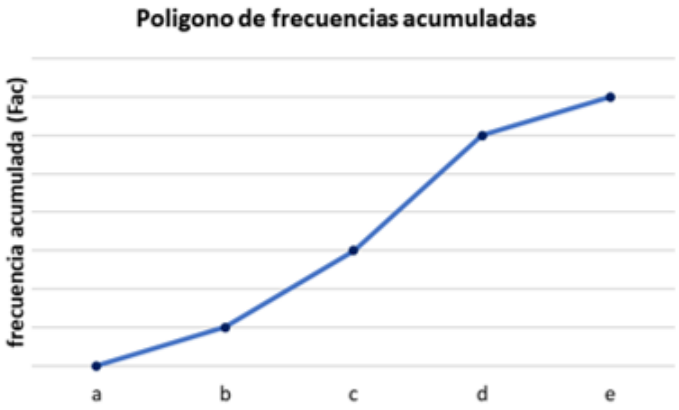
Intervalo	Dato	Frecuencia (f)
Intervalo 1	Clase 1	A
Intervalo 2	Clase 2	B
Intervalo 3	Clase 3	C
Intervalo 4	Clase 4	D
Intervalo 5	Clase 5	E



Los polígonos de frecuencia son una representación gráfica similar al gráfico de barras, que permite **comparar las frecuencias** de las diferentes variables rápidamente. En este gráfico se marcan los puntos que representan la frecuencia de los datos y después, simplemente, se trazan líneas que los unan ordenadamente.

Polígono de frecuencia acumulada u ojiva

Intervalo	Frecuencia acumulada (Fac)
[a, b [	A
[b, c [	B
[c, d [	C
[d,e [	D



De manera similar, puede **representarse la frecuencia acumulada de los datos** de la misma manera que el polígono de frecuencias, con la diferencia de que se debe considerar la frecuencia acumulada para cada punto. Hay que notar que este gráfico **nunca va a ser decreciente**, puesto que el número de datos siempre es mayor o igual a cero.

Medidas de Tendencia Central

Son indicadores de la tendencia general de los datos. Estas son números que pretenden **resumir el conjunto de datos a un solo valor**. Tenemos a:

Media Aritmética ($\overline{x}$)

Conocido también como **promedio**. Es el **cuociente entre la suma de los datos y el número de datos**. Si se tienen n datos, se determina de la siguiente manera: (las x_i corresponden a los datos)

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

En el caso de que se tengan los **datos organizados en tabla de frecuencias**, en vez de sumar uno por uno, aprovechamos el saber la frecuencia para determinar la media aritmética:

x_1	f_1
x_2	f_2
...	...
x_n	f_n

$$\overline{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + ... + x_n \cdot f_n}{f_1 + f_2 + ... + f_n} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i}$$

En el caso de que no se tengan los datos exactos, más bien, agrupados en intervalos, entonces se calcula primero la **marca de clase** y esa se considerará como los "datos que se tienen" para calcular la media aritmética: (las c corresponden a las marcas de clase)

$$\overline{x} = \frac{c_1 \cdot f_1 + c_2 \cdot f_2 + ... + c_n \cdot f_n}{f_1 + f_2 + ... + f_n} = \frac{\sum_{i=1}^{n} c_i \cdot f_i}{\sum_{i=1}^{n} f_i}$$

Veamos un ejemplo:

x_i	f_i
4	3
2	5
1	4
3	2

El promedio se calcula utilizando la formula anterior:

$$\frac{4\cdot 3+2\cdot 5+1\cdot 4+3\cdot 2}{3+5+4+2}=\frac{12+10+4+6}{14}=\frac{32}{14}\sim 2,28$$

Luego la media es: $\overline{x}\sim 2,28$. ¡Fíjate que los datos no están ordenados en orden creciente!

Mediana

Cuando se tienen los datos **ordenados de forma creciente** (si no están ordenados, ¡hay que hacerlo!), siempre existirá **uno o dos datos al medio**, ya sea si son impar o par respectivamente. Esa es la **mediana**.

Caso Impar: Si la cantidad de datos que se tiene es un número impar, entonces sólo se debe tomar la mediana como el **dato central**:

$$\begin{aligned}1,2,3,4,5 &\Longrightarrow \text{Mediana: } 3\\ 1,2,3,4,5,6,7,8,9 &\Longrightarrow \text{Mediana: } 5\\ 1,2,3,..,97,98,99 &\Longrightarrow \text{Mediana: } 50\end{aligned}$$

Caso Par: Si la cantidad de datos que se tiene es un número par, entonces se debe calcular el **promedio de los dos datos centrales** que quedan expuestos:

$$\begin{aligned}1,2,3,4 &\Longrightarrow \text{Mediana: } 2,5\\ 1,2,3,4,5,6,7,8,9,10 &\Longrightarrow \text{Mediana: } 5,5\\ 1,2,3,..,97,98,99,100 &\Longrightarrow \text{Mediana: } 50,5\\ 2,4,5,42,43,90 &\Longrightarrow \text{Mediana: } 23,5\end{aligned}$$

Cuando se tiene una **tabla de frecuencias** se puede proceder con la siguiente **técnica**:

x_i	f_i
4	3
2	5
1	4
3	2

- **Paso 1:** Los datos se deben ordenar de forma creciente

x_i	f_i
1	4
2	5
3	2
4	3

- **Paso 2:** Se restan los extremos de la columna de las frecuencias, sobreviviendo el término mayor.

x_i	f_i
1	~~4~~ 1
2	5
3	2
4	~~3~~

- **Paso 3:** Se itera el procedimiento con los nuevos valores de los extremos hasta llegar a tener uno o dos valores.

x_i	f_i
1	~~4~~ ~~1~~
2	5
3	~~2~~ 1
4	~~3~~

$\Longrightarrow$

x_i	f_i
1	~~4~~ ~~1~~
2	~~5~~ **4**
3	~~2~~ ~~1~~
4	~~3~~

Por ende la mediana corresponde al 2.

Moda

La moda corresponde al **dato que más se repite en una muestra**. En el problema anterior, el dato con mayor frecuencia corresponde a 2, ya que su frecuencia es 5.

Por ejemplo, si todos los datos poseen la misma frecuencia, se les dice **amodal**. En el caso de que existan **dos o más datos con la mayor frecuencia** se les llama **bimodal** o **polimodal**, respectivamente. En el caso de poseer intervalos, se elige que posee mayor frecuencia, a ese se le dice **intervalo modal**.

¡OJO! ¡En datos cualitativos solo se puede calcular la moda! Estos no poseen valores cuantificables que nos permitan estimar medias y como no se pueden ordenar es imposible calcular una mediana.

Ejercicios Propuestos:

1. Si las notas de esteban en una asignatura son: 3, 4, 6, 3, 5, 5, 6, 3, 4 y de estas notas se cambia un 6 por un 7, ¿cuál(es) de las siguientes medidas de tendencia central cambia(n)?

   (i) La moda
   (ii) La mediana
   (iii) La media

   a) Solo II
   b) Solo III
   c) Solo I y II
   d) Solo II y III
   e) Ninguna de ellas

2. La siguiente tabla resume las edades de un grupo de niños. ¿Cuál es la media de las edades?

   a) 8 años
   b) 10 años
   c) 12 años
   d) 16 años
   e) 30 años

Edad	f	F
$[0, 4[$		7
$[4, 8[$	6	
$[8, 12[$		
$[12, 16]$	5	30

Medidas de Posición

Tal y como el nombre lo indica, las medidas de posición **dividen la distribución de datos en partes iguales**, clasificando los elementos dentro de una muestra, siempre ordenando los datos de manera creciente. Así se pueden categorizar los individuos, por ejemplo, en sectores socioeconómicos es muy común usar estas medidas según su ingreso per cápita. Entre los más importantes están los:

Cuartiles

Son 3 cuartiles que dividen a los datos en 4 partes. En el Q_1, Q_2 y Q_3, se acumulan el 25 %, 50 % y 75 % de los datos respectivamente. Además, Q_2 = Mediana.

El rango intercuartil se define como la diferencia entre los cuartiles extremos:

$$\text{Rango Intercuartil: } Q_3 - Q_1$$

Quintiles

Son 4 quintiles que dividen a los datos en 5 partes. En el q_1, q_2, q_3, y q_4 se acumulan el 20 %, 40 %, 60 % y 80 % de los datos respectivamente.

Deciles

Son 9 deciles que dividen a los datos en 10 partes. En el d_1, d_2, ..., y d_9 se acumulan el 10 %, 20 %, ..., y 90 % de los datos respectivamente.

Ejemplo: Deciles de ingreso en Chile

Los deciles se utilizan para definir sectores socioeconómicos según el ingreso promedio considerando la cantidad de personas que conforman un hogar. La lista de los deciles según el nivel de ingresos es el siguiente:

Deciles	Ingresos
1°	$48.750
2°	$74.969
3°	$100.709
4°	$125.558
5°	$154.166
6°	$193.104
7°	$250.663
8°	$352.743
9°	$611.728

Percentiles

Son 99 percentiles que dividen a los datos en 100 partes.

En el p_1, p_2, ..., y p_{99} se acumulan el 1 %, 2 %, ..., y 99 % de los datos respectivamente.

Si los datos están tabulados, se debe proceder de la misma forma que se utiliza para determinar la mediana si se quiere saber una medida de posición, utilizando la frecuencia acumulada. Si son intervalos se dice este último simplemente.

Gráficos de Caja y Bigotes

El diagrama de caja y bigotes es una **representación gráfica de los cuartiles**, que ilustra estas particularidades cuando los datos son **ordenados de forma creciente**. Para graficarlo se necesitan: el **valor mínimo, el valor máximo y los cuartiles**.

Veamos un ejemplo de cómo se grafican:

Imaginemos una generación de un colegio que está compuesta por tres cursos (A, B y C) de 32 alumn@s cada uno. Ellos dan una prueba de matemáticas (con nota del 1 al 100) obteniendo resultados diferentes y variados.

Digamos que el curso A es el que obtuvo mejores notas, teniendo a más de la mitad obteniendo un puntaje mayor a 60 puntos. El curso B es el que mantuvo una respuesta promedio y en el curso C, más de la mitad obtuvieron un puntaje menor a 40 puntos.

Imaginémonos que el curso B tuvo una distribución en el gráfico de caja y bigotes como se ilustra a continuación:

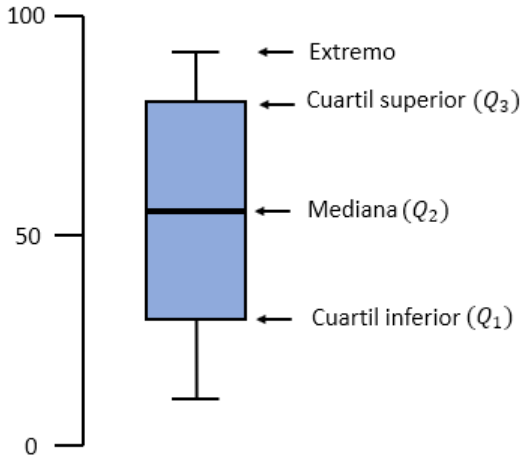


Figura 1: Distribución de los puntajes del curso B para una evaluación.

A esta muestra la llamamos **Muestra Simétrica**, que quiere decir que los valores están igualmente dispersos.

Apreciemos un poco más el gráfico de caja y bigotes. Notemos que hay dos extremos, uno inferior y otro superior. **Estos corresponden a los datos extremos, es decir, más bajo y alto** respectivamente cuando se ordenan. Estos son conectados a la caja, y corresponden a los "bigotes". Los **límites de la caja corresponden a los cuartiles** Q_1 y Q_3, mientras que la **línea central es el cuartil** Q_2 (o mediana).

En el caso del curso A, donde se obtuvieron mejores resultados, podríamos esperar algo como el gráfico que se aprecia a continuación:

Vemos que en este gráfico los datos están distribuidos diferentes, y se ve un desplazamiento

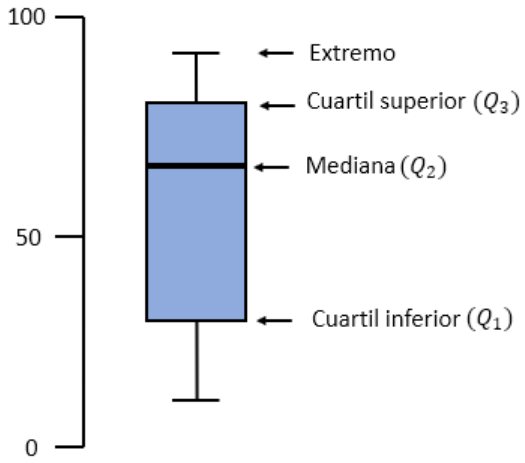


Figura 2: Distribución de los puntajes del curso A para una evaluación.

general hacia arriba (hacia puntajes más altos). Es importante notar que, **entre los cuartiles y los extremos, siempre existirán la misma cantidad de datos**. Es decir, que entre Q_2 y Q_3 hay un total de 8 alumn@s (ya que $\frac{32}{4} = 8$).

A este gráfico se le llama **Muestra Negativamente Asimétrica**, ya que los valores menores están más dispersos. Eso implica que l@s estudiantes con peores puntajes del curso A están más dispersos que los que poseen notas más altas.

El curso C finalmente tendría un gráfico extendido que posee esta forma:

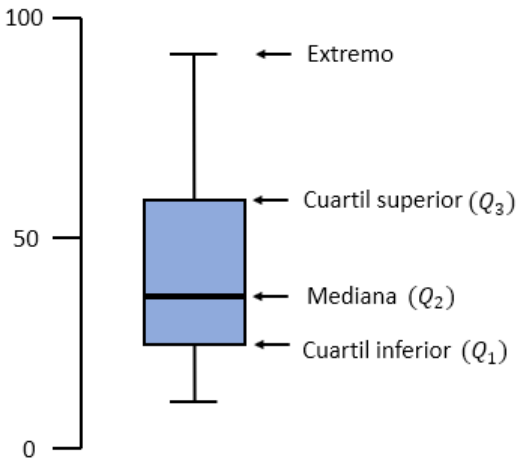


Figura 3: Distribución de los puntajes del curso C para una evaluación.

A este gráfico se le llama **Muestra Positivamente Asimétrica**, porque los valores mayores están más dispersos y se ve un desplazamiento general hacia abajo. Eso implica que l@s estudiantes con mejores puntajes del curso C están más dispersos que los que poseen notas más bajas.

Medidas de Dispersión

Como su nombre lo indica, determinan qué tan dispersos están los datos con respecto a su valor central. Mientras haya menos dispersión, más homogénea será la muestra.

Existen 3 medidas ampliamente utilizadas: el Rango, la Desviación Estándar o Desviación Típica (σ) y la Varianza (σ^2). En esta oportunidad solo veremos la primera ya que es la única que se encuentra en el temario Admisión 2022 DEMRE.

Rango

Es la diferencia entre el mayor y el menor de los datos.

Ejemplo: En un curso de 20 estudiantes de IV medio la mínima edad es de 16 años y la máxima de 19, entonces el rango es: $19 - 16 = 3$ años.

Ejemplo resuelto: Realizar un diagrama de caja con los datos de casos nuevos (PCR positivo COVID-19) por día correspondientes a una comuna del sur de Chile a partir del 18 de Septiembre hasta el 16 de Noviembre del año 2020. Además calcular rango.

Fecha	Casos nuevos	Fecha	Casos nuevos
18/09	2	18/10	8
19/09	6	19/10	14
20/09	2	20/10	1
21/09	6	21/10	15
22/09	9	22/10	7
23/09	15	23/10	42
24/09	1	24/10	18
25/09	4	25/10	8
26/09	9	26/10	6
27/09	4	27/10	6
28/09	6	28/10	24
29/09	1	29/10	0
30/09	3	30/10	4
01/10	12	31/10	11
02/10	17	01/11	40
03/10	13	02/11	20
04/10	2	03/11	1
05/10	39	04/11	1
06/10	3	05/11	18
07/10	19	06/11	2
08/10	32	07/11	5
09/10	28	08/11	3
10/10	11	09/11	1
11/10	4	10/11	4
12/10	3	11/11	4
13/10	25	12/11	3
14/10	8	13/11	7
15/10	10	14/11	7
16/10	23	15/11	3
17/10	10	16/11	9

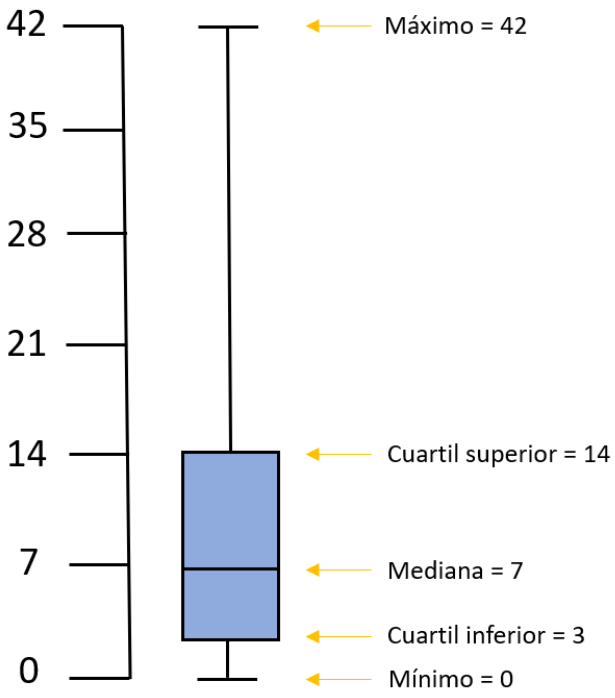
Solución: Ordenando los datos de menor a mayor

0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 6, 6, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 12, 13, 14, 15, 15, 17, 18, 18, 19, 20, 23, 24, 25, 28, 32, 39, 40 y 42.

Como son 60 datos, lo divimos por 4 y nos da el valor de 15. Esto quiere decir que al ordenar los números de menor a mayor, las cifras que se ubican en las posiciones 15, 30 y 45 corresponden a los cuartiles $Q_1 = 3$, $Q_2 = 7$ y $Q_3 = 14$ representados por los colores verdes.
El mínimo es 0 y el máximo es 42 son representados por el color naranjo.

0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 6, 6, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 12, 13, 14, 15, 15, 17, 18, 18, 19, 20, 23, 24, 25, 28, 32, 39, 40, 42

Hacemos el diagrama de caja:



Observamos que existe un desplazamiento general hacia abajo (números más bajos).

Por último, se define el **rango intercuartil** como la resta entre el cuartil superior Q_3 y el cuartil inferior Q_1:

$$\text{Rango Intercuartil} = Q_3 - Q_1$$

Para este caso en particular:

$$\text{Rango Intercuartil} = Q_3 - Q_1 = 14 - 3 = 11$$

Finalmente el rango sería: $42 - 0 = 42$.

Preguntas resueltas de problemas publicados por el DEMRE

1. El gráfico circular de la figura adjunta muestra los resultados de una encuesta aplicada a 300 estudiantes sobre su nivel de acuerdo sobre la implementación de salas de computación en su colegio.

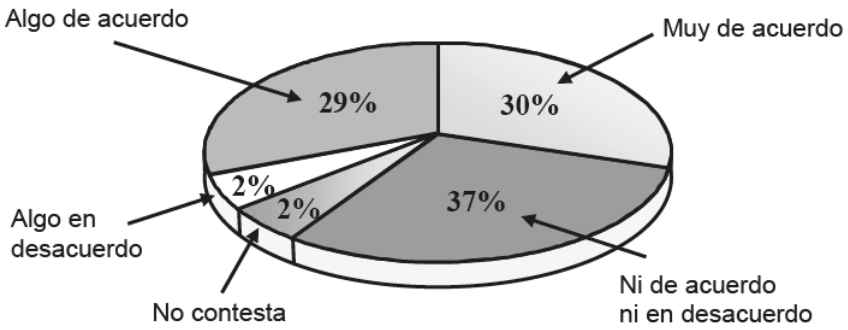


Figura 4: Imagen extraída ensayo DEMRE oficial admisión 2021.

¿Cuál de las siguientes afirmaciones es verdadera?

A) La frecuencia relativa de los que contestan "Muy de acuerdo" es 3/10
B) La frecuencia de los que contestaron "Ni de acuerdo ni en desacuerdo" supera en 8 estudiantes a los que contestaron "Algo de acuerdo".
C) El nivel de acuerdo de la encuesta es bimodal.
D) 2 estudiantes no contestan la encuesta.

Solución: La frecuencia relativa se calcula como el cociente de la frecuencia absoluta de algún valor de la población/muestra y el total de datos. La alternativa A) afirma que la frecuencia relativa de "Muy de acuerdo" es 3/10. Veamos si es cierta esta afirmación o no:

La frecuencia absoluta de "Muy de acuerdo" es el número de veces que se repite un dato: 30 % de 300 es $\underbrace{0,3}_{30\,\%} \cdot 300 = 90$.

La frecuencia relativa es:

$$fr = \frac{\text{frecuencia absoluta}}{\text{total de datos}} = \frac{90}{300} = \frac{\cancel{30} \cdot 3}{\cancel{30} \cdot 10} = \frac{3}{10}$$

Entonces es correcta la alternativa A).

Para estar seguros, debemos corroborar que el resto de las alternativas son incorrectas.

La frecuencia absoluta de "Ni de acuerdo ni en desacuerdo" es $\underbrace{0,37}_{37\,\%} \cdot 300 = 111$. Por otro lado,

La frecuencia absoluta de "Algo de Acuerdo" es $\underbrace{0,29}_{29\,\%} \cdot 300 = 87$. Y al restarlas: $111 - 87 = 24$.

La alternativa B) es falsa.

También es falsa la afirmación dada en C), porque "Ni de acuerdo ni en desacuerdo" es la categoría que tiene la mayor frecuencia, por lo tanto, solo esta categoría sería la moda. Bimodal significa dos modas, pero hay solo una.

La afirmación dada en D) también es falsa, pues $\underbrace{0,02}_{2\,\%} \cdot 300 = 6$. Entonces 6 estudiantes no contestaron la encuesta.

2. En la tabla adjunta se muestra la distribución de las edades, en años, de un grupo de personas.

Intervalo	**Frecuencia**	**Frecuencia relativa porcentual**
$[12, 18[$	8	16
$[18, 24[$	14	
$[24, 30[$		
$[30, 36[$		18
$[36, 42[$	3	

Según los datos de la tabla, ¿cuál de las siguientes afirmaciones es FALSA?

A) La marca de clase del intervalo de mayor frecuencia es 27 años.
B) Un 44 % de las personas tiene menos de 24 años.
C) El grupo en total tiene 50 personas.
D) Exactamente, un 38 % de las personas tiene menos de 30 años.
E) 28 personas tienen a lo menos 24 años.

Solución: Con la información de la primera fila podemos deducir la cantidad de personas total ya que el 16 % son 8 personas.

$$16\,\% = 8\,\text{personas}$$
$$100\,\% = X\,\text{personas}$$

$$X\,\text{personas} = \frac{100\,\% \cdot 8\,\text{personas}}{15\,\%} = 50\,\text{personas}$$

Con este dato podemos completar la tabla:

Intervalo	Frecuencia	Frecuencia relativa porcentual
$[12, 18[$	8	16
$[18, 24[$	14	$\frac{14}{50} \cdot 100 = 28$
$[24, 30[$		
$[30, 36[$	$\frac{18}{100} \cdot 50 = 9$	18
$[36, 42[$	3	$\frac{3}{50} \cdot 100 = 6$
Total	**50**	**100**

La fila que quedó en blanco se calcula con el total de ambas columnas: La columna de frecuencia suma 50 y la frecuencia relativa porcentual suma 100:

Intervalo	Frecuencia	Frecuencia relativa porcentual
$[12, 18[$	8	16
$[18, 24[$	14	28
$[24, 30[$	$50 - (8 + 14 + 9 + 3) = 16$	$100 - (16 + 28 + 18 + 6) = 32$
$[30, 36[$	9	18
$[36, 42[$	3	6
Total	**50**	**100**

Ahora que completamos la tabla, queda analizar cada afirmación. La marca de clase del intervalo de mayor frecuencia es $[24, 30[= (24 + 30)/2 = 27$ es correcta.

La alternativa B) es correcta ya que al sumar la frecuencia de los 2 primeros intervalos es $8 + 14 = 22$, y 22 de 50, en porcentaje es:

$$\frac{22 \cdot 100\,\%}{50} = 44\,\%$$

La alternativa C) es correcta.

La alternativa D) es incorrecta ya que al sumar la frecuencia de los 3 primeros intervalos es $8 + 14 + 16 = 38$, y 38 de 50, en porcentaje es:

$$\frac{38 \cdot 100\,\%}{50} = 76\,\%$$

La alternativa E) es correcta ya que bajo 24 años son 22 personas (alternativa B), entonces: $50 - 22 = 28$ personas al menos tienen 24 años.

3. La distribución de los sueldos, en pesos, de los trabajadores de una empresa se muestra en el diagrama de caja de la figura adjunta.

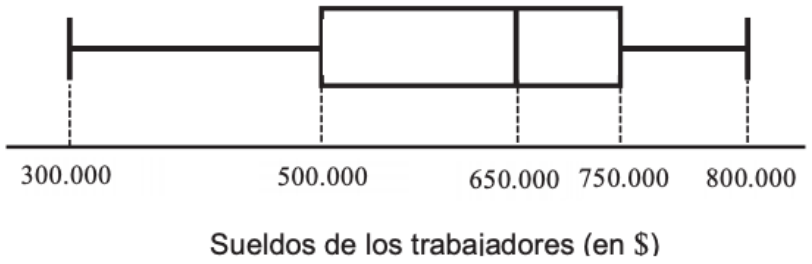


Figura 5: Imagen extraída ensayo DEMRE oficial admisión 2021.

Según este diagrama, ¿cuál de las siguientes afirmaciones es siempre verdadera?

A) El rango intercuartil de los sueldos de los trabajadores es \$250.000.
B) El promedio de los sueldos de los trabajadores es \$ 650.000.
C) La cantidad de trabajadores que ganan entre \$300.000 y \$500.000 es mayor que la cantidad de trabajadores que gana entre \$650.000 y \$750.000.
D) Exactamente un 50 % de los trabajadores gana \$650.000.
E) Un 62,5 % de los sueldos de los trabajadores es igual o menor a \$700.000.

Solución: Del diagrama de caja tenemos que el tercer cuartil de los sueldos de los trabajadores es \$750.000 y el primer cuartil es \$500.000 , luego el rango intercuartil de los sueldos es \$750.000 - \$500.000 = \$250.000, por lo que la afirmación en A) es siempre verdadera.

Como el diagrama de caja no entrega información sobre el promedio de los datos, puede que el promedio de los datos sea distinto a \$650.000, por lo que la afirmación en B) no es siempre verdadera.

Como la cantidad de trabajadores que gana entre \$300.000 y \$500.000 y la cantidad de trabajadores que gana entre \$650.000 y \$750.000 son ambas aproximadamente un 25 % de los datos, puede que la cantidad de trabajadores que gana entre \$650.000 y \$750.000 sea igual que la cantidad de trabajadores que gana entre \$300.000 y \$500.000, por lo que la afirmación en C) no es siempre verdadera.

Ahora, el segundo cuartil de los sueldos es \$650.000 por lo que, aproximadamente, un 50 % de los trabajadores gana menos que esa cantidad, pero dicha información no es suficiente para asegurar que exactamente un 50 % de los trabajadores gana \$650.000, por lo que la afirmación en D) no es siempre verdadera.

Por último, el diagrama de caja solo entrega información referente al dato mayor, el dato menor y los cuartiles de la distribución de los sueldos de los trabajadores, por lo que la afirmación en E) no es siempre verdadera.

Por el desarrollo anterior, la clave es A).

Figura 6: Jennifer A. Doudna

Científica Destacada: Jennifer A. Doudna

Jennifer Anne Doudna nació el 19 de febrero de 1964 en Washington, DC. Estudió Licenciatura en Química en la Universidad de Pomona en California. Hizo su doctorado en Bioquímica en la Universidad de Harvard, y su trabajo postdoctoral en la Universidad de Colorado. Durante su doctorado logró demostrar que el ARN no solo recibe instrucciones del ADN para la síntesis de proteínas sino que también acelera el proceso. Más adelante, logró determinar la estructura de distintos tipos de ARN, uno de ellos ligado al virus de la Hepatitis B. Desde 1994 hasta el año 2000 fue profesora asistente de la Universidad de Yale y luego fue promovida a profesora Henry Ford II de Biofísica y Bioquímica Molecular. Desde el año 2002 trabaja en la Universidad de California, Berkeley, como profesora de Bioquímica y Biología Molecular. Esto le permitió el acceso al Laboratorio Nacional de Lawrence Berkeley, en particular a un tipo de acelerador de partículas llamado ciclotrón, que permite estudiar de mejor manera moléculas complejas.

Es reconocida principalmente por su trabajo en descifrar los mecanismos moleculares del sistema inmunológico bacteriano CRISPR-Cas9 y su aplicación como herramienta en ingeniería genética. Junto con Emanuelle Charpentier descubrieron como Cas9 (junto con moléculas sintécticas de ARN guía) puede insertar, suprimir y modificar ADN. Este descubrimiento le ha otorgado numerosos premios incluyendo el Premio Nobel de Química 2020, en conjunto con Emanuelle Charpentier.