Medidas Descriptivas

M.Sc. Helen Guillén Oviedo

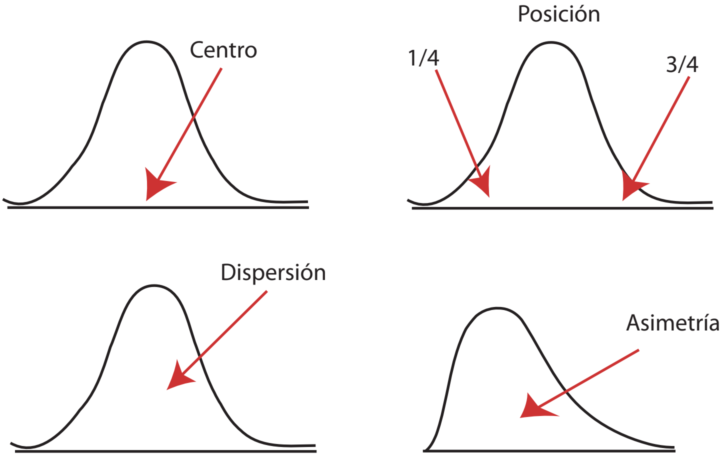
MAT006 Probabilidad y Estadística

Escuela de Matemática
Universidad Nacional

Hasta aquí nos hemos preocupado de la descripción de un conjunto de observaciones, ya sea gráficamente o por medio de una distribución de frecuencias. En muchos casos, es preferible trabajar con una o más medidas descriptivas que resuman en forma cuantitativa ciertas características de los datos. Tales medidas reciben el nombre de Estadísticos o Estadígrafos.
En este sentido pueden examinarse varias características de interés en una muestra, siendo las más comunes:

1. La tendencia central de los datos.
2. La dispersiíon o variación con respecto a este centro.
3. Los datos que ocupan ciertas posiciones.
4. La forma y simetría con que los datos se agrupan.

Medidas Descriptivas



Medidas de tendencia central

Las medidas de tendencia central son valores numéricos que quieren mostrar el centro de un conjunto de datos, nos interesan especialmente: la moda, la media y la mediana. Si los datos son una muestra, la moda, la media(o promedio) y la mediana se llamarán estadísticas. Si los datos son una población entonces estas medidas de tendencia central se llamarán parámetros.

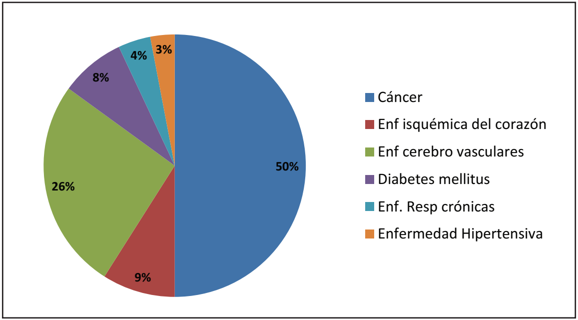
Medidas y Representaciones Estadísticas

- **Posición**
  - Moda
  - Mediana
  - Media Aritmética
  - Mínimo
  - Máximo
  - Cuartiles
- **Variabilidad**
  - Recorrido ó Amplitud
  - Recorrico Intercuartílico
  - Varianza
  - Desviación estándar
- **Medidas Relativas**
  - Posición Relativa
  - Coeficiente de Variación
- **Representaciones Gráficas**
  - Diagramas de Cajas

Ejemplo 1

¿Que información podría resaltar del siguiente gráfico?

Gráfico Nº 1.
Distribución porcentual de muertes prematuras según enfermedades crónicas no transmisibles.
Costa Rica, 2012.



Fuente: Dirección de vigilancia de la salud, base del INEC.

Ejemplo 2

Observe el siguiente cuadro

Cuadro Nº 1.
Cantidad de estudiantes de quinto año según número de caries.
Liceo San Francisco, Costa Rica. III trimestre, 2016.

Cantidad de Caries	Cantidad de estudiantes
0	12
1	10
2	8
3	4
4	5
5	3
6	2
7	1
total	45

Fuente: Directora del Liceo.

¿Que información podría resaltar?

Moda

- La moda es el valor que más se repite en el conjunto de datos, es decir el que tiene mayor frecuencia absoluta
- La moda siempre es uno de los datos que hay en el conjunto.
- Se puede utilizar para resumir variables cualitativas y cuantitativas.
- Si sucede que dos valores tienen la misma frecuencia, siendo ésta la más alta, se dice que el conjunto es bimodal (o sea, tiene dos modas).

Moda

- Si ocurre que en un conjunto de datos hay tres o más valores que presentan la mayor frecuencia, se dice que la distribución de datos es multimodal y en este caso, la moda no es útil como medida de tendencia central.
- Se debe tener cuidado con su interpretación, debido a que el concepto de valor que más se repite o de mayor frecuencia, no significa que la mayoría de las observaciones tome ese valor, pues la mayoría representa más de la mitad, pero el valor modal no siempre incluye mas de la mitad de los datos.
- La moda si existe la representaremos con M_o.

Ejemplo 3: Moda para datos no agrupados

En el conjunto de datos

$$1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 5 \quad 5 \quad 5 \quad 6 \quad 6 \quad 9 \quad 10$$

Se observa que la moda es 2, dado que es el valor que aparece mayor cantidad de veces en el conjunto. Pero sería un error afirmar que la mayoría de los valores son 2, debido a que hay doce datos en total y solamente cuatro de ellos son 2. Por lo que lo correcto es decir, que el valor más frecuente en el conjunto de datos es 2.

Ejemplo 4: Moda para datos no agrupados

En la muestra

$$3 \quad 3 \quad 3 \quad 3 \quad 4 \quad 4 \quad 5 \quad 5 \quad 5 \quad 5 \quad 7$$

se tiene que el 3 y el 5 presentan la mayor frecuencia. Por lo tanto, se puede decir que ambos valores son la moda del conjunto.
Si ocurre que en un conjunto de datos hay más de dos valores que presentan la mayor frecuencia, se dice que la distribución de datos es multimodal y en este caso, la moda no es útil.

Moda para datos no agrupados

Si los datos no están agrupados, se define la Moda como aquel valor que aparece el mayor número de veces. Es decir

$$f_l = max\{f_i\} \forall i = 1, \cdots, k \Rightarrow M_o = \text{el dato con } f_l$$

Mo = Valor que tiene la mayor frecuencia absoluta.

Ejemplo 6

Retomando el ejemplo de la cantidad de caries, dado en el Cuadro No.1, escribamos las 45 respuestas en una lista ordenada de menor a mayor, así:

0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
3	3	3	3	4	4	4	4	4	5	5	5	6	6	7

¿Si preguntáramos por la medida central que diríamos?...

Ejemplo 7

¿Y para este cuadro que podemos decir acerca del valor central?

Cuadro Nº 1.
Cantidad de estudiantes de quinto año según número de caries.
Liceo San Francisco, Costa Rica. III trimestre, 2016.

Cantidad de Caries	Cantidad de estudiantes	Cantidad Acumulada de estudiantes
0	12	12
1	10	22
2	8	30
3	4	34
4	5	39
5	3	42
6	2	44
7	1	45
total	45	

Fuente: Directora del Liceo.

Mediana

- Es el valor que ocupa la posición central en el conjunto de datos, cuando se encuentran ordenados de menor a mayor de acuerdo con su magnitud.
- Es un valor tal que la mitad de las observaciones son menores o iguales que él y la otra mitad, son mayores o iguales.
- La mediana puede estar o no estar en el conjunto de datos. La mediana la representaremos con M_e.
- La mediana siempre es un valor mayor o igual que el valor mínimo de los datos y menor o igual que el valor máximo. No necesariamente corresponde a un dato en el conjunto.
- Esta medida de posición se puede utilizar para resumir datos cuantitativos o cualitativos con nivel de medición ordinal, pues para poder calcular la mediana, los datos se deben ordenar en una lista de menor a mayor.

Mediana

Si tenemos un conjunto de datos no agrupados, la mediana se puede encontrar mediante:

$$M_e \;=\; M_e = x_{\frac{n+1}{2}}$$

Donde:
n= número total de observaciones

Ejemplo 8: Mediana para datos no agrupados

Considere la siguiente muestra de 15 hogares, donde se les pregunto a cada familia la cantidad de niños por hogar.

$$2 \quad 3 \quad 0 \quad 1 \quad 4 \quad 0 \quad 3 \quad 0 \quad 1 \quad 2 \quad 2 \quad 3 \quad 4 \quad 3 \quad 2$$

Calculemos la mediana del conjunto de datos, para esto primero se deben ordenar los datos de menor a mayor

$$0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 3 \quad 4 \quad 4$$

Ahora usamos la formula para calcular la mediana

$$M_e = X_{\frac{15+1}{2}} = X_8 = 2$$

Interpretación: el 50% de los hogares entrevistados tienen 2 o menos niños y el otro 50% tienen 2 o más niños.

Ejercicio 1

Calcule e interprete la moda y mediana para el siguiente conjunto de datos, correspondientes a edad en años cumplidos de 20 empleados del ebais de San Antonio de Belén:

40	22	37	25	28	29	30	30	30	34
35	38	30	38	25	40	40	42	43	50

Media aritmética o Promedio

La medida de posición central de mayor uso teórico y práctico es la media aritmética, que también suele llamarse promedio o simplemente media. El símbolo que se usa para la media de una población es la letra griega μ, y el símbolo para la media de la muestra es $\overline{X}$(x-barra).

El promedio se define como el punto de equilibrio, el punto donde la distribución se balancea.

Media para datos sin agrupar

Media de la población

$$\mu = \frac{x_1 + x_2 + \ldots + x_n}{N} = \frac{\sum_{i=1}^{n} x_i}{N}$$

Media de la muestra

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

Donde:

x_i=elemento i

n= número total de datos

Ejemplo 10: Media para datos sin agrupar

A continuación se presentan los datos correspondientes a la cantidad de materias que tienen que presentar un grupo de 20 estudiantes de octavo año del Colegio Santísima Trinidad durante las convocatorias del año 2012.

$$1 \quad 2 \quad 3 \quad 4 \quad 3 \quad 1 \quad 2 \quad 2 \quad 2 \quad 5 \quad 2 \quad 2 \quad 4 \quad 3 \quad 3 \quad 3 \quad 1 \quad 5 \quad 2 \quad 4$$

Calcule el promedio.

$$\overline{X} = \frac{1+2+3+4+3+1+2+2+2+5+2+2+4+3+3+3+1+5+2+4}{20} = 2.7$$

Interpretación: La cantidad de material promedio que tiene que presentar un grupo de 20 estudiantes de octavo año del Colegio Santísima Trinidad durante las convocatorias del año 2012 es de 3.

Ejercicio 2:

Sofía, una estudiante de décimo año obtuvo las siguientes calificaciones en cinco exámenes correspondientes a la materia de Español: 94, 92, 90, 93 y 92. Carlos, un compañero de clase de Sofía, obtuvo las siguientes calificaciones en esos mismos exámenes: 95, 96, 45, 98 y 100. Si usted fuera su profesor y tuviera que escoger a uno de ellos para ir a un concurso representando al colegio, ¿A cuál de ellos escogería? ¿Por qué?

Propiedades de la media aritmética:

- Todo conjunto de datos derivado de una variable cuantitativa tiene una media.
- Si un conjunto de datos $x_1, x_2, \ldots, x_n$ posee media $\overline{X}$, entonces $x_1 + b, x_2 + b, \ldots, x_n + b$ tiene media $\overline{X} + b$.
- Si un conjunto de datos $x_1, x_2, \ldots, x_n$ posee media $\overline{X}$, entonces $a \cdot x_1, a \cdot x_2, \ldots, a \cdot x_n$ tiene media $a \cdot \overline{X}$.

La media aritmética es un valor que no necesariamente pertenece al conjunto de datos, pero siempre es mayor o igual que el valor mínimo y menor o igual que el valor máximo. Esta medida de posición es un valor representativo de los datos que se promedian, y me indica la cantidad de la variable estudiada que correspondería a cada observación (persona, animal, país, aula, etc.), si se dividiera el total suponiendo que cada una tiene la misma cantidad.

A veces a los valores $x_1, x_2, \ldots, x_n$ de una variable se les asocia ciertos factores o pesos $w_1, w_2, \ldots, w_n$ que dependen de la importancia de cada uno de los valores. En este caso la media aritmética se llama media aritmética ponderada y se determina de la siguiente manera:

$$\overline{X} = \frac{\displaystyle\sum_{i=1}^{n} (x_i \cdot w_i)}{\displaystyle\sum_{i=1}^{n} w_i}$$

Donde:

x_i=Elemento i

n= Número total de datos

w_i= Ponderador del elemento i

Ejemplo 12: Media aritmética Ponderada:

La evaluación de un curso está distribuida como sigue 10% quices, 20% tareas y 70% exámenes. Al final del curso un estudiante obtuvo las siguientes notas: 85 en tareas, 81 en quices y 91 en exámenes.

Ejemplo 12: Media aritmética Ponderada:

Si queremos calcular la nota promedio de este estudiante, debemos usar la fórmula del promedio ponderado, ya que cada nota tiene un peso diferentes, el cual corresponde al porcentaje que se le asigno en la evaluación del curso. Así tenemos que

$$\overline{X} = \frac{85 \cdot 20 + 81 \cdot 10 + 91 \cdot 70}{10 + 20 + 70} = 88,8$$

Interpretación: La nota promedio del estudiante en el curso es de un 88,8.

Ejercicio 3

Suponga que la nota del curso de Probabilidad y Estadística está distribuida de la siguiente manera

Rubro	Porcentaje
I Parcial	20 %
II Parcial	25 %
III Parcial	30 %
Pruebas Cortas	10%
Tareas	15%

Si un estudiante obtiene las siguientes calificaciones I parcial: 50, II parcial: 70; III parcial: 80, Pruebas cortas: 60, tareas: 90.

¿Cuál es su nota promedio?

Ejercicio 4

Suponga que se tienen los siguientes datos acerca del dinero gastado por cuatro colegios de cierto país en una actividad deportiva:

Colegio	Cantidad de estudiantes	gastos en colones
A	150	200000
B	350	400000
C	300	300000
D	250	200000

Calcule el gasto promedio por colegio

Ejercicio 5

La enfermedad de la Cirrosis es la cicatrización y el funcionamiento deficiente del hígado. La cirrosis es el resultado final del daño crónico al hígado causado por enfermedad hepática crónica. Las causas comunes de la enfermedad hepática crónica son: Infección por hepatitis B o hepatitis C o Alcoholismo. Dentro de las pruebas para detectar si una persona padece de cirrosis está un examen de sangre que mide la albúmina. La albúmina es una proteína producida por el hígado. El examen de albúmina en suero mide la cantidad de esta proteína en la parte líquida y transparente de la sangre. La albúmina también se puede medir en la orina. El rango normal es de $3,4$ a $5,4$ gramos por decilitro (g/dl).Un nivel de albúmina en la sangre por debajo de lo normal puede ser un signo de cirrosis.

Ejercicio 5

- Sofía, una estudiante de uńdecimo año asistió al médico y le realizaron la prueba de albúmina cada semana por 7 días y obtuvo los siguientes resultados:

$$2,89 \quad 2,9 \quad 2,8 \quad 2,9 \quad 2,4 \quad 2,79 \quad 2,93$$

- Carlos, un compañero de clase de Sofía, también asistió al médico y le realizaron las mismas pruebas, obteniendo los siguientes resultados:

$$2,5 \quad 2,28 \quad 2,8 \quad 2,8 \quad 8 \quad 2,9 \quad 2,7$$

¿Qué considera usted que debería el doctor recomendar a cada uno de ellos al ver estos resultados?

Desventajas de la Media

1. Es muy sensible a los valores extremos de la variable, ya que todas las observaciones intervienen en el cálculo de la media, la aparición de una observación extrema, hará que la media se desplace en esa dirección.
2. No es recomendable usar la media como medida central en las distribuciones muy asimétricas.
3. Depende de la división en intervalos en el caso de variables continuas.
4. Si consideramos una variable discreta, por ejemplo, el número de hijos en las familias de Limón el valor de la media puede no pertenecer al conjunto de valores de la variable; Por ejemplo: $\bar{x} = 2.5$ hijos.

Ejercicio

A continuación se ofrecen los datos correspondientes al tiempo de espera (en minutos) de 50 usuarios de una biblioteca hasta que son atendidos por algún miembro del personal de ésta.

1	3	5	20	21	4	7	9	10	12
20	18	6	4	13	11	10	13	15	9
4	20	2	22	8	6	11	4	8	6
5	18	19	20	7	15	16	13	12	14
7	10	5	24	11	8	9	10	11	7

Determine e interprete en términos del problema, la moda, la mediana y la media aritmética.

Medidas de posición

- Las medidas de posición equivalen a los valores que puede tomar una variable caracterizados por agrupar a cierto porcentaje de observaciones en la muestra o población.
- Las medidas de posición son ideales para obtener información adicional a partir de datos resumidos, es decir, que presentan perdida de información por agrupamiento en intervalos de clase.
- Podemos decir que las medidas de posición son indicadores estadísticos que muestran la frecuencia acumulada hasta un valor k cualquiera.
- Dentro de las medidas de posición mas usadas, están los percentiles, de los cuales podemos describir los cuartiles y los deciles. describiremos cada una de estas medidas.

Los Percentiles

- Los percentiles son tales que dividen al conjunto de datos en 100 partes iguales que representan cada una el 1% de los valores. Por ejemplo, el Percentil 90 es un valor tal que el 90% de todos los valores son menores y el 10% son mayores que él.
- El percentil de orden m lo representaremos como P_m.
- Se define el percentil de orden m como la observación i-ésima, que queda por debajo de el m % de los datos.

- **Para NO datos agrupados**
  Para calcular el percentil en datos no agrupados, debemos primero ordenar los datos de menor a mayor y luego ubicar la posición en la que se ubica el percentil a encontrar, para eso calculamos $\frac{m(n+1)}{100}$.
  Luego buscamos la posición que deseamos en los datos y este representa el percentil que queremos encontrar.

$$P_m = X\left[\frac{m(n+1)}{100}\right]$$

Donde:

n= Número total de observaciones
m= Percentil de orden m.

Ejemplo 13: Medidas de posición

Considere el siguiente conjunto de datos que corresponden a la nota de 20 estudiantes del curso de Cálculo de la UNA:

$$\begin{array}{cccccccccc} 40 & 50 & 40 & 60 & 35 & 45 & 35 & 70 & 40 & 70 \\ 87 & 70 & 56 & 70 & 40 & 70 & 63 & 35 & 70 & 50 \end{array}$$

Si queremos calcular P_{50} debemos ordenar los datos de menor a mayor.

Ejemplo 13: Medidas de posición

35	35	35	40	40	40	40	45	50	**50**
56	60	63	70	70	70	70	70	70	87

Y luego calculamos

$$P_{50} \; = X_{\left[\frac{50(20+1)}{100}\right]} = X_{10.5}$$

Es decir buscamos los valores que están en la posición 10 y 11 y los promediamos, pues la posición 10.5 esta justo en medio de la posición 10 y 11, Así

$$P_{50} \; = X_{10.5} = \frac{50+56}{2} = 53$$

Interpretación: El 50% de los estudiantes tienen notas menores o iguales 53 y el otro 50% tiene notas mayores o iguales a 53.
Observe que el percentil 50, coincide con la mediana del conjunto de datos.

Ejemplo 13: Medidas de posición

Nota: Que pasa si queremos calcular el P_{60} en el ejemplo anterior,

$$P_{60} = X_{\left[\frac{60(20+1)}{100}\right]} = X_{12.6}$$

Podemos observar que en este caso la posición que deseamos no esta justo en medio de la posición 12 y 13, por lo cual para poder determinar el valor exacto que ocupa esa posición, debemos realizar una interpolación lineal.

35	35	35	40	40	40	40	45	50	50
56	**60**	**63**	70	70	70	70	70	70	87

Para realizar la interpolación lo podemos hacer con el uso de la calculadora o siguiendo la siguiente fórmula.

Ejemplo 13: Medidas de posición

Posición	Valor
A	C
$\frac{m(n+1)}{100}$	$P_m = X_{\frac{m(n+1)}{100}}$
B	D

Así $$\frac{D-C}{B-A} = \frac{P_m - C}{\frac{m(n+1)}{100} - A} \Longrightarrow P_m = (D-C)\left(\frac{m(n+1)}{100} - A\right) + C$$

Es decir,

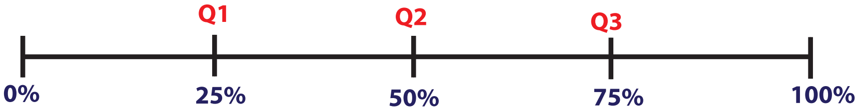
Posición	Valor
12	60
12.6	P_{60}
13	63

Así $$P_{60} = (63-60)\,(12,6-12) + 60 = 61,8$$

Los Cuartiles

Son casos particulares de los percentiles y se define en forma análoga a la de los percentiles.
Los cuartiles dividen al conjunto de datos en cuatro partes iguales, y solo existen tres de ellos: el cuartil 1: Q_1, el cuartil 2: Q_2 y el cuartil 3: Q_3
La forma de calcular los cuartiles es por medio de los percentiles con las siguientes igualdades:

1. $\mathbf{P_{25} = Q_1}$
2. $\mathbf{P_{50} = Q_2}$
3. $\mathbf{P_{75} = Q_3}$



Representación de los cuartiles.

Ejercicio 8

El siguiente conjunto de datos corresponde a edad en años cumplidos de 20 empleados del EBAIS de San Antonio de Belén:

22	25	25	28	29	30	30	30	30	34
35	37	38	38	40	40	40	42	43	50

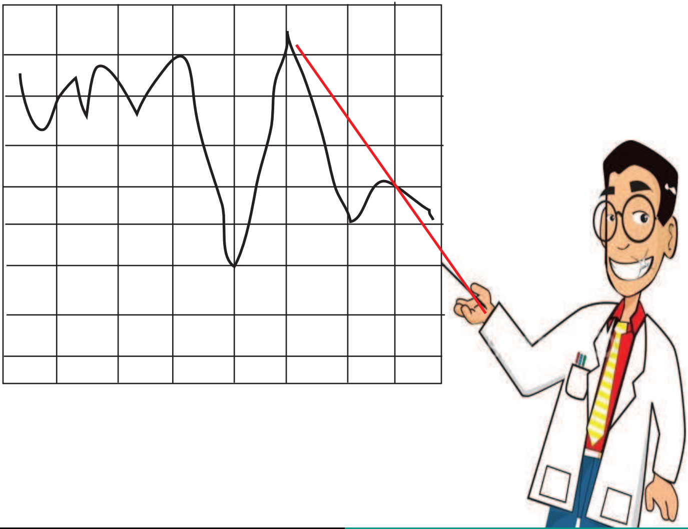
Calcule e interprete Q_1 y Q_3.

Los Deciles

Son casos particulares de los percentiles y se definen en forma análoga a la de los percentiles. Los deciles dividen al conjunto de datos en diez partes iguales, y solo existen nueve de ellos.
La forma de calcular los deciles al igual es los cuartiles es por medio de los percentiles con las siguientes igualdades:

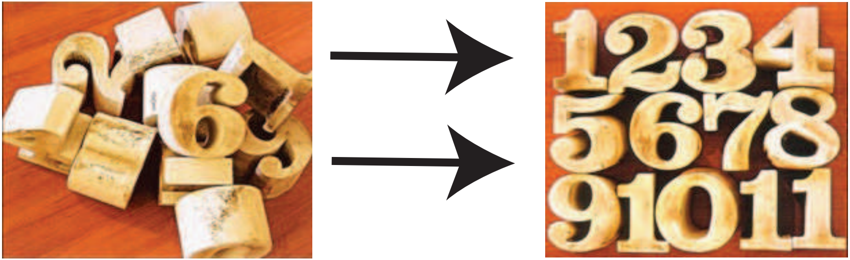
1. $\mathbf{P_{10} = D_1}$
2. $\mathbf{P_{20} = D_2}$
3. $\mathbf{P_{30} = D_3}$
4. $\mathbf{P_{40} = D_4}$
5. $\mathbf{P_{50} = D_5}$
6. $\mathbf{P_{60} = D_6}$
7. $\mathbf{P_{70} = D_7}$
8. $\mathbf{P_{80} = D_8}$
9. $\mathbf{P_{90} = D_9}$

Moda

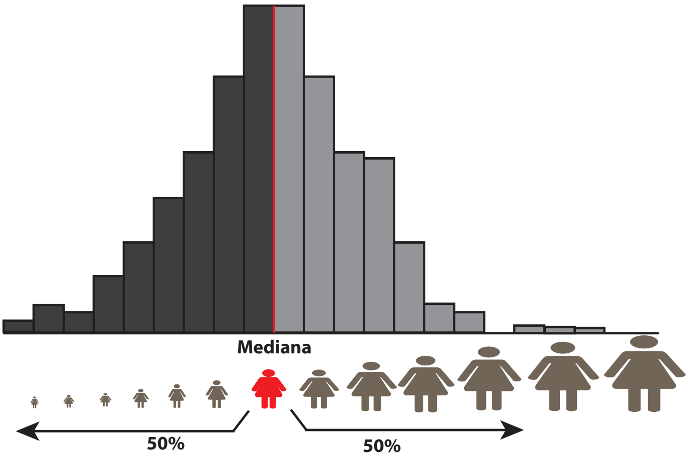


Resumen de medidas de posición

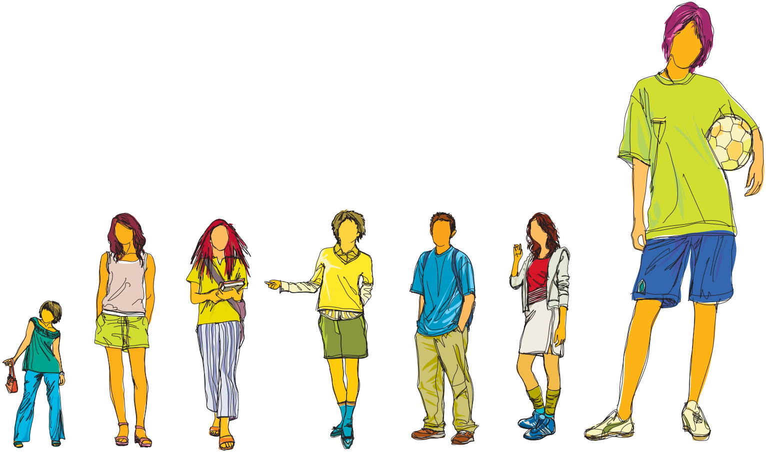
Cuartiles y Mediana



Resumen de medidas de posición



Media Aritmética




Media
Mediana
Cuartiles

Utilidad de las medidas de Posición

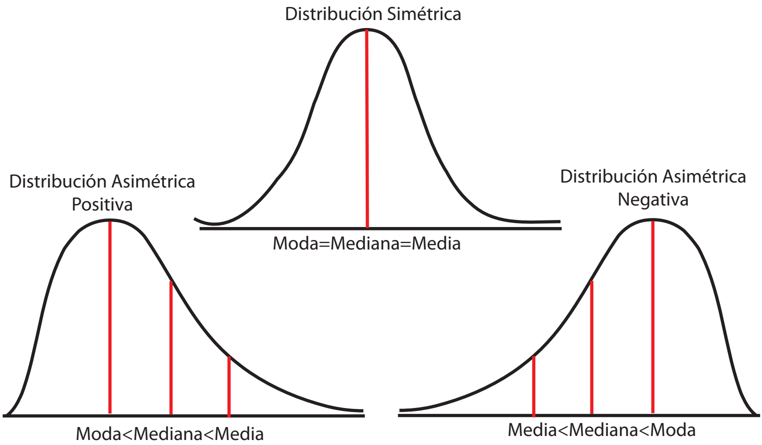
- Para describir una serie de datos, es necesario ubicarla dentro de un rango de valores posibles; el valor máximo y mínimo del conjunto pueden servir para este propósito, pero esto no aporta nada acerca de la forma en que se distribuyen los datos.

Utilidad de las medidas de Posición

- Se debe apoyar la caracterización con algunas medidas que ayuden a describir la distribución de los datos de manera más completa. Así, se puede complementar el resumen con uno o más valores numéricos que se encuentren ubicados hacia el centro de la distribución de datos, como es el caso de las medidas de posición: moda, mediana y media aritmética, también conocidas como medidas de tendencia central. Estas medidas son valores alrededor de los cuales se encuentran distribuidos los datos del conjunto estudiado.

- Medidas como los cuartiles, que dividen al conjunto de datos en cuatro partes iguales, agregan información importante acerca de la serie. Entre el primer cuartil y el tercero, se encuentran el 50% de los datos centrales.
- Las medidas de posición se complementan unas a otras, de manera que al usarlas conjuntamente se logra una mejor descripción del conjunto de datos.

Distribución de los datos



Medidas de Variabilidad o dispersión

Cuando hablamos de variabilidad o dispersión, nos referimos al grado de semejanza o diferencia que existe entre los valores que toma la variable de estudio. De esta forma, un conjunto de datos que posea una alta variabilidad es un conjunto de datos cuyo valores son muy diferentes unos de otros (muy dispersos). Por el contrario, si un conjunto posee una variabilidad baja estaremos frente a un conjunto de datos cuyo valores son muy similares entre sí (poco dispersos).

Rango o recorrido

Una medida razonable de la dispersión podría ser el Rango o Recorrido, que se obtiene restando el valor mas bajo del conjunto de observación del valor mas alto, es decir. El rango lo representaremos con: R

Desventajas:

- No utiliza todas las observaciones.
- Es muy sensible a alguna observación extrema.
- El rango aumenta cuando el número de observaciones también aumenta, o bien se queda igual. En cualquier caso, nunca disminuye.

1. **Datos no agrupados**
   $R = X_{max} - X_{min}$
   Donde:
   X_{max}= valor mayor del conjunto de datos
   X_{min}= valor menor del conjunto de datos

El recorrido intercuartílico

- Esta medida corresponde a la diferencia entre el tercer y el primer cuartil, es decir corresponde al rango de variabilidad de 50% de los valores centrales.
- Al recorrido intercuartílico se representa mediante el símbolo RI y su formula viene dada por:

$$RI = Q_3 - Q_1$$

Ejemplo 15: El recorrido intercuartílico

A continuación se presentan los datos correspondientes a la cantidad de materias que tienen que presentar los 20 estudiantes de un grupo de octavo año del Colegio Santísima Trinidad durante las convocatorias del año 2012.

$$1 \;\; 2 \;\; 3 \;\; 4 \;\; 3 \;\; 1 \;\; 2 \;\; 2 \;\; 2 \;\; 5 \;\; 2 \;\; 2 \;\; 4 \;\; 3 \;\; 3 \;\; 3 \;\; 1 \;\; 5 \;\; 2 \;\; 4$$

Para determinar el recorrido intercuartílico, primero se debe calcular el Q_3 y el Q_1, para lo cual debemos ordenar los datos de menor a mayor

$$1 \;\; 1 \;\; 1 \;\; 2 \;\; 2 \;\; 2 \;\; 2 \;\; 2 \;\; 2 \;\; 2 \;\; 3 \;\; 3 \;\; 3 \;\; 3 \;\; 3 \;\; 4 \;\; 4 \;\; 4 \;\; 5 \;\; 5$$

Ejemplo 15: El recorrido intercuartílico

Así,

$$Q_1 = P_{25} \ = X_{\left[\frac{25(20+1)}{100}\right]} = X_{5,25} = 2$$

$$Q_3 = P_{75} \ = X_{\left[\frac{75(20+1)}{100}\right]} = X_{18,75} = 4,75$$

Con lo que,

$$RI = 4,75 - 2 = 2,75$$

Interpretación: el 50% central de la cantidad de materias que debe presentar un estudiante tiene una variación de 3 materias, es decir esta entre y y 5 materias.

El recorrido intercuartílico

Aunque esta medida es mucho más precisa que el recorrido total pues elimina la influencia de los valores extremos, para mayor precisión se requiere de una medida que utilice para su cálculo todos los datos del conjunto.

Desviación media

En estadística la desviación absoluta promedio o, sencillamente desviación media o promedio de un conjunto de datos es la media de las desviaciones absolutas y es un resumen de la dispersión estadística. Se expresa, de acuerdo a esta fórmula:

1. Datos no agrupados

$$DM = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|$$

2. Datos agrupados

$$DM = \frac{1}{n}\sum_{i=1}^{k} (|x_i - \overline{x}| \cdot f_i)$$

Varianza

La varianza de un conjunto de datos, $x_1, x_2, \ldots, x_n$ se define como la media de las desviaciones cuadráticas de las n observaciones:

Varianza de la población

1. **Datos no agrupados en tablas**

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

Varianza de la muestra

1. **Datos no agrupados en tablas**

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

Ejemplo 16: Varianza

Basados en el Ejemplo 15, donde se tienen los datos correspondientes a la cantidad de materias que tienen que presentar los 20 estudiantes de octavo año del Colegio Santísima Trinidad durante las convocatorias del año 2012.

$$1 \quad 2 \quad 3 \quad 4 \quad 3 \quad 1 \quad 2 \quad 2 \quad 2 \quad 5 \quad 2 \quad 2 \quad 4 \quad 3 \quad 3 \quad 3 \quad 1 \quad 5 \quad 2 \quad 4$$

Para determinar la varianza del conjunto de datos, se debe determinar primero la media, la cual es $2,7$, así,

Ejemplo 16: Varianza

$$(1-2,7)^2+(2-2,7)^2+(3-2,7)^2+(4-2,7)^2+(3-2,7)^2+(1-2,7)^2+$$

$$(2-2,7)^2+(2-2,7)^2+(2-2,7)^2+(5-2,7)^2+(2-2,7)^2+(2-2,7)^2+(4-2,7)^2$$

$$+(3-2,7)^2+(3-2,7)^2+(3-2,7)^2+(1-2,7)^2+(5-2,7)^2+(2-2,7)^2+(4-2,7)^2$$

$$=28,2$$

$$S^2 = \frac{28,2}{20-1} = 1,4842$$

Es importante mencionar, que el calculo de la varianza es muy extenso, pero con el uso de la calculadora resulta ser más eficiente.

Desviación estándar

1. **Desviación estándar de la población**

$$\sigma = \sqrt{\sigma^2}$$

2. **Desviación estándar de la muestra**

$$S = \sqrt{S^2}$$

Ejemplo 17: Desviación estándar

Supongamos que tenemos tres poblaciones cualquiera, con los siguientes datos

$$\{0, 0, 14, 14\},\ \{0, 6, 8, 14\} \text{ y } \{6, 6, 8, 8\}$$

donde cada una tiene una media aritmética igual a 7. Si calculamos sus desviaciones estándar obtenemos que son 7, 5 y 1 respectivamente, por lo que observamos que la tercera población tiene una desviación mucho menor que las otras dos porque sus valores están más cerca de 7.

Observaciones sobre medidas de variabilidad y dispersión

- Las medidas de variabilidad dependen mucho de la unidad de medida, por lo que no es sencillo dar su interpretación.
- Si la varianza y la desviación es 0 es porque no existe ninguna variabilidad entre los datos.
- cuando crece la dispersión de un conjunto de datos, también crece el valor de S^2 y S
- S^2 y S son sensibles a la variación de cada una de las observaciones, es decir, si una observación cambia, cambia con ella la varianza.
- Si se calculan a través de los datos agrupados en una tabla, S^2 y S dependen de la amplitud y del numero de intervalos elegidos.
- No es recomendable el uso de ellas cuando tampoco lo sea el uso de la media.

Se utiliza para comparar datos aislados entre diferentes conjuntos. Se denomina estandarización o tipificación, y lo que hace es relativizar el valor de manera que se pueda realizar la comparación correspondiente. La estandarización o tipificación para un dato se determina por:

Para la población:

$$estandarizacion = \frac{x_i - \mu}{\sigma}$$

Para la muestra:

$$estandarizacion = \frac{x_i - \overline{X}}{S}$$

Ejemplo 19: Estandarización o tipificación de mediciones

Roberto entró a la Universidad ABC en el año 2010 y obtuvo una calificación de 94 en su examen de admisión, Rocío entró en el año 2012 a esa misma Universidad y su calificación en el examen de admisión fue de 92. Se sabe que en el año 2010 las calificaciones obtenidas en el examen de admisión a esta universidad tuvieron una media aritmética de 75 con desviación estándar de 5,23; mientras que en el año 2012 la media fue de 72 con desviación estándar de 4,5. ¿Cuál de ellos logró el mejor puntaje relativo a su generación?

Ejemplo 19: Estandarización o tipificación de mediciones

Si se observa solo la nota de admisión de cada uno, se pensaría que Roberto tuvo mejor puntaje, pero se debe pensar que el examen es en años diferentes, por lo que su dificultad podría ser distinta y esto se evidencia con las notas promedio de cada año, ademas de considerar la desviación estándar para darse cuenta cuan largo esta cada estudiante de la media de su correspondiente año, aunque es evidente que ambos están por encima del promedio, pero en cuanto a la desviaci'on estándar esta mas cerca Rocio del promedio, así que estandarizar y no tener este problema se calcula lo siguiente:

Roberto

$$\frac{94-75}{5,23} = 3,6329$$

Rocío

$$\frac{92-72}{4,5} = 4,4444$$

Variabilidad relativa (Coeficiente de variación)

El coeficiente de variación es la relación entre la desviación típica de una muestra y su media. Suele expresarse en porcentajes. El coeficiente de variación permite comparar las dispersiones de dos distribuciones distintas. Se calcula para cada una de las distribuciones y los valores que se obtienen se comparan entre sí. La mayor dispersión corresponderá al valor del coeficiente de variación mayor.

Coeficiente de variación para la población

$$CV = \frac{\sigma}{|\mu|} \cdot 100$$

Coeficiente de variación para la muestra

$$CV = \frac{S}{|\overline{X}|} \cdot 100$$

Variabilidad relativa (Coeficiente de variación)

El CV es usado como medida de la homogeneidad de los datos:

1. De 0 a menos de 11% se dice los datos son muy homogéneos
2. De 11% a menos de 16% se dice que los datos son homogéneos,
3. De 16% a menos de 26% se dice que los datos son heterogéneos,
4. De 26% o más se dice que los datos son muy heterogéneos.

Ejemplo 20: Variabilidad relativa (Coeficiente de variación)

Los siguientes datos corresponden a la distribución de una muestra aleatoria de 40 empleados de la Fábrica XYZ, ubicada en la provincia de Guanacaste, según estatura en centímetros y según peso en kilogramos. Los datos fueron recolectados por el Departamento de Personal de dicha Fábrica el 15 de marzo de 2007.

Estatura	Cantidad de empleados
155	4
165	9
175	18
185	6
195	3
Total	40

Peso	Cantidad de empleados
52,5	6
57,5	10
62,5	12
67,5	8
72,5	4
Total	40

¿Cuál de los dos conjuntos de datos presenta mayor variabilidad?

Ejemplo 20: Variabilidad relativa (Coeficiente de variación)

Estatura

$$\overline{X} = 173,75$$

$$S_X = 10,4237$$

Peso

$$\overline{X} = 61,75$$

$$S_X = 6,0500$$

Estatura

$$CV = \frac{10,4237}{173,75} \cdot 100 = 6,00\%$$

Peso

$$CV = \frac{6,0500}{61,75} \cdot 100 = 9,80\%$$

Interpretación: El conjunto de datos que presenta mayor variabilidad relativa es el peso de los empleados dela fabrica XYZ. Además, se tiene que la estatura y el peso son datos muy homogéneos.

Ejercicio

El departamento de Estadísticas del Hospital de Niños en Costa Rica, realizó un estudio para determinar cuáles pesos en kilogramos son relativamente más variables si los de los niños con edades entre los 2 y 4 años o los de los niños con edades entre los 10 y 12 años, para eso recolectó la siguiente información de dos muestras de 14 niños.

Niños de 2 a 4 años	**Niños de 10 a 12 años**
9	32
15	32
16	33
14	42
11	40
13	44
16	32
10	32
15	27
19	28
15	38
16	38
10	28
15	27

Ejercicio

¿Cuál conjunto de niños presenta mayor variabilidad relativa en cuanto a la característica **Peso** en Kilogramos, los **niños de 2 a 4** años o los **niños de 10 a 12**?

Ejercicio

Medida	**Niños de 2 a 4 años**	**Niños de 10 a 12 años**
$\overline{X}$	13,8571	33,7857
S_X	2,8785	5,6729
CV	20,77	16,79

Interpretación: Se puede observar que el peso de los niños con edades de 2 a 4 años, presentan una variabilidad relativa de un 20,77%, es decir es un conjunto de datos heteregéneo, lo mismo con los pesos de los ni;os con edades de 10 a 12, con una variabilidad de 16,79%. Los pesos mas variables son los de los niños con edades de 2 a 4 años.