Estadística
IDEI

Ingeniería Industrial

2020

Análisis Descriptivo Univariado

Medidas de posición central

Las medidas de posición central están diseñadas para brindar al investigadora algunos valores cuantitativos respecto de sus datos, por ejemplo saber cuál es la ubicación central en su muestra.
Algunas de estas medidas son:

- Moda
- Mediana
- Media aritmética

Antes de comenzar con el desarrollo se van a plantear tres ejemplos de datos:

Ejemplo variable cualitativa: Se encuestó a 20 personas acerca del tipo de vivienda donde residen. Se obtuvo la siguiente información:

Tipo de vivienda	n_i
Casa	10
Departamento	7
Duplex	3

Ejemplo variable discreta: Se encuestó a 20 personas acerca de la cantidad de hijos que tenían:

N° de hijos	n_i	N_i
0	3	3
1	7	10
2	8	18
3	2	20

Ejemplo variable continua: Se registró el tiempo de secado (en horas) de cierta marca de pintura esmaltada:

Clases	n_i	m_i	N_i
(2 - 3]	3	2,5	3
(3 - 4]	5	3,5	8
(4 - 5]	7	4,5	15
(5 - 6]	5	5,5	20

Moda (Mo)

La moda es el valor de la variable que ocurre con mayor frecuencia en un conjunto de datos. En el caso en que todos los datos se repiten una única vez, es práctico decir que el conjunto de datos no presenta moda. Si hay dos valores que se repiten igual cantidad de veces es bimodal. A su vez, cabe aclarar que la moda es una medida que se puede utilizar para datos cualitativos.
Para hallar la moda es útil ordenar los datos para ver cuál es el que se repite una mayor cantidad de veces.
Supongamos el conjunto de datos es {0,0,0,0,1,1,2,2,3,4}, entonces el 0 es la moda, porque se repite un total de 4 veces. Si tuviesemos en cambio {0,1,2,2,3,4,4} habría dos modas porque el 2 y el 4 son los más frecuentes y se repiten la misma cantidad de veces, entonces en este caso el conjunto de datos sería bimodal.
Cuando resumo los datos en una tabla de distribución de frecuencias agrupando por intervalos o clases, se habla de clase modal, y un valor estimativo para la misma es el punto medio de esa clase modal.

En los ejemplos planteados al comienzo la moda sería:
Ejemplo variable cualitativa: Mo=Casa
Ejemplo variable cuantitativa discreta: Mo=2
Ejemplo variable cuantitativa continua: clase modal=4,5

Mediana (Md)

La mediana de un conjunto de n observaciones ordenadas de menor a mayor, es un valor tal que la mitad de las observaciones es menor o igual a ese valor, y la otra mitad de las observaciones es mayor o igual a ese valor.

Si n es par:

2 8 4 9 50 14

Como 1º paso debemos ordenar el conjunto de datos:

2 4 8 9 14 50

Luego, en 2º paso la Md se la calcula como:

$$Md = \frac{X(q) + X(q+1)}{2}$$

donde q representa la posición de los datos y se calcula como $q = \frac{n}{2}$, siendo n el número de observaciones, y $X(q)$ representa los valores de nuestras observaciones asociados a esa posición. Entonces con nuestros datos $q = \frac{6}{2} = 3$, entonces:

$$Md = \frac{X(3) + X(4)}{2} = \frac{8+9}{2} = 8,5$$

Si n es impar:

2 8 4 9 14

1º ordenamos los datos:

$$2 \qquad 4 \qquad 8 \qquad 9 \qquad 14$$

Luego como 2º paso, calculamos la Md:

$$Md = X(q)$$

donde $q = \frac{n+1}{2}$. Entonces con nuestros datos, $q = \frac{5+1}{2} = 3$, y por consiguinte:

$$Md = X(3) = 8$$

Cuando los datos están agrupados por clases o por intervalos, la mediana es el punto medio del intervalo que acumula el 50 % de los datos.

En los ejemplos planteados al comienzo la mediana sería:
Ejemplo variable cuantitativa discreta: Como n es par, $q = 10$, y la $Md = \frac{X(10)+X(11)}{2} = \frac{7+8}{2} = 7,5$
Ejemplo variable cuantitativa continua: La clase que acumula el 50 % de los datos en (4 - 5], entonces $Md = 4,5$

Media aritmética ($\overline{x}$)

La media aritmética es simplemente un promedio de los valores observados. Se obtiene sumando los valores observados y dividiendo por el número de observaciones (n).

Supongamos que tenemos los siguientes datos:

$$2 \qquad 8 \qquad 4 \qquad 9 \qquad 50 \qquad 14$$

Entonces:

$$\overline{x} = \frac{2+8+4+9+50+14}{6} = 14,5$$

En los ejemplos planteados al comienzo la mediana sería:
Ejemplo variable cuantitativa discreta:

$$\overline{x} = \sum_{i=1}^{k=4} \frac{x_i n_i}{n} = \frac{0*3+1*7+2*8+3*2}{20} = \frac{29}{20} = 1,45$$

donde k es la cantidad de categorías.
Ejemplo variable cuantitativa continua:

$$\overline{x} = \sum_{i=1}^{k=4} \frac{m_i n_i}{n} = \frac{2,5*3+3,5*5+4,5*7+5,5*5}{20} = \frac{84}{20} = 4,2$$

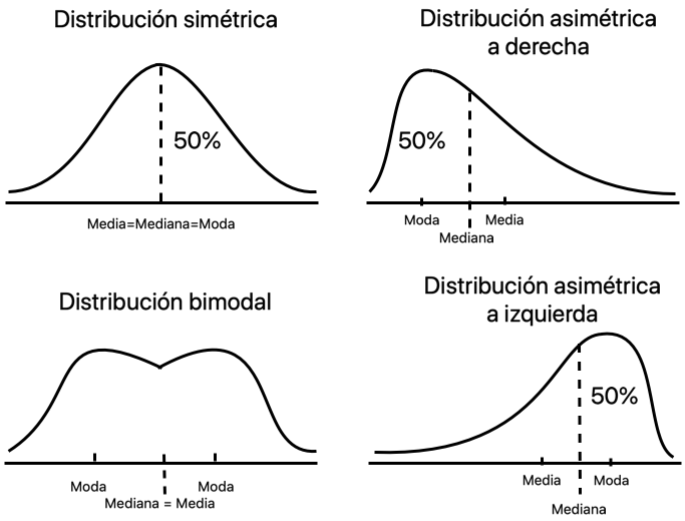
donde k es la cantidad de intervalos o clases y m_i es la marca de clase (punto medio del intervalo).

Entonces, ¿qué medida de posición utilizar?. La *moda* no es muy utilizada, dado que el valor mas frecuente puede encontrarse lejos del centro de la distribución. La *media* es el centro de gravedad de la distribución, el punto en el cual la distribución balancearía, pero entonces esta medida está influenciada por valores que se encuentren alejados de los valores comunes. En cambio, esto no ocurre con la *mediana*, dado que esta medida tiene en cuenta la posición de los datos pero no sus valores.

Veamos un ejemplo: Supongamos que se tienen los valores

$$2 \qquad 5 \qquad 7 \qquad 9 \qquad 35$$

La media es $\overline{x} = 11,6$ y la mediana $Md = 7$. La media dista de la mediana dado que la última observación, que vale 35, la está influenciando. Entonces se podría decir que le mediana es una medida robusta, dado que no está influenciada por los valores extremos.
Los siguientes gráficos muestran la relación entre la media, mediana y moda.



Para distribuciones asimétricas o distribuciones con valores extremos, la mediana tiende a ser la mejor elección como medida de tendencia central.

Medidas de dispersión

Estas medidas sirven para dar una idea de la variabilidad que tienen los datos.

Medida de dispersión para la media

Supogamos que la media es un valor representativo de posición central de conjunto de datos. Entonces veamos cuánto se desvía cada observación de su media y calculemos un promedio de ellas. Para ello:

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

Pero el problemas aquí es que estas desviaciones sumadas dan 0, porque

$$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = n\overline{x} - n\overline{x} = 0$$

Supongamos que 5 grupos de alumnos se someten a un test obteniendo los siguientes valores (notas):

Grupo A:	3	4	5	6	7	(Variable X)
Grupo B:	1	3	5	7	9	(Variable Y)
Grupo C:	5	5	5	5	5	(Variable Z)
Grupo D:	3	5	5	7		(Variable W)
Grupo E:	3,5	5	6,5			(Variable V)

Calculando las medianas y las medias llegamos a que:

$$Md(X) = Md(Y) = Md(Z) = Md(W) = Md(V) = 5$$
$$\overline{x} = \overline{y} = \overline{z} = \overline{w} = \overline{v} = 5$$

Las medidas de posición central son iguales en todos los grupos pero no estamos diciendo nada de la variabilidad de los datos. Entonces podemos considerar la sumatoria en valor absoluto de los desvíos y sino podríamos considerar la suma de los cuadrados de los desvíos.

Grupo	$\frac{1}{n}\sum \lvert x_i - \overline{x}\rvert$	$\frac{1}{n}\sum (x_i - \overline{x})^2$
Grupo A	$\frac{2+1+0+1+2}{5} = \frac{6}{5} = 1,2$	$\frac{4+1+0+1+4}{5} = \frac{10}{5} = 2$
Grupo B	$\frac{4+2+0+2+4}{5} = \frac{12}{5} = 2,4$	$\frac{16+4+0+4+16}{5} = \frac{40}{5} = 8$
Grupo C	$\frac{0}{5} = 0$	$\frac{0}{5} = 0$
Grupo D	$\frac{2+0+0+2}{4} = \frac{4}{4} = 1$	$\frac{4+0+0+4}{4} = \frac{8}{4} = 2$
Grupo E	$\frac{1,5+0+1,5}{3} = \frac{3}{3} = 1$	$\frac{2,25+0+2,25}{3} = \frac{4,5}{3} = 1,5$

Entonces resumiendo:

- **Desvío medio**: $\frac{1}{n}\sum_{i=1}^{n} \lvert x_i - \overline{x}\rvert$
- **Varianza**: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- **Desvío estándar**: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

Observaciones:

- Usualmente se define y se usa $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$, veremos más adelante la conveniencia (por propiedades teóricas) de esta definición.
- Cuando tenemos los **datos agrupados por categorías** (datos discretos), el cálculo es

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i$$

- Cuando tenemos los **datos agrupados por clases o intervalos**, el cálculo es

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k}(m_i - \overline{x})^2 n_i$$

Medida de dispersión para la mediana

Por el contrario, si la medida utilizada para representar la posición central de los datos fue la mediana, entonces una medida de variabilidad asociada a ella es el MAD.
El MAD se calcula como:

$$Md|x_i - Md| \quad \text{con} \quad 1 \leq i \leq n$$

.

Los pasos para calcularlo son:

1. Calcular los desvíos: $x_i - Md$
2. Calcular los valores absolutos de los desvíos: $|x_i - Md|$
3. Ordenar los valores absolutos de los desvíos $|x_i - Md|$
4. Hallar la mediana de estos valores.

Veamos un ejemplo: Supongamos que tenemos los siguientes datos:

5 7 12 32 8

Para hallar la mediana, primero debemos ordenar el conjunto de datos

5 7 8 12 32

Ahora como n es impar, $q = \frac{5+1}{2} = 3$, la mediana será el valor que se encuentra en la posición 3, entonces $Md = 8$.

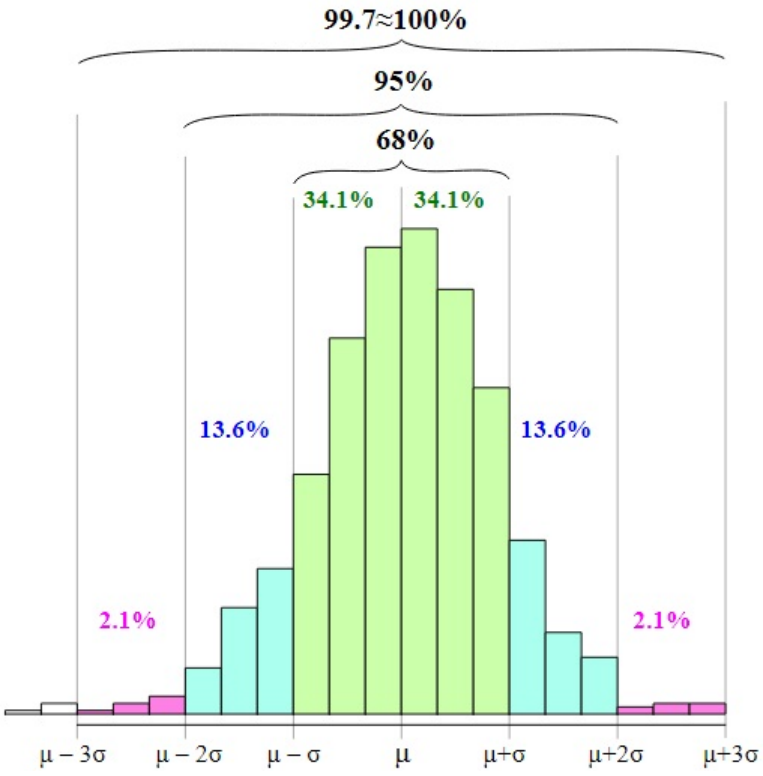
Ahora pasemos a calcular el MAD:

1. Calculemos los desvíos:
   -3 -1 4 26 0
2. En valores absolutos quedan:
   3 1 4 26 0
3. Ordenados:
   0 1 3 4 24
4. Y la mediana de estos es el valor que se encuentre en la posición 3, entonces el $MAD = 3$.

Regla empírica

La regla empírica sirve para tener una idea de la simetría del conjunto de datos con el cual se está trabajando.
Para corroborar si el comportamiento de la variable es simétrico se debe verificar que:

- el **68,27 %** de los valores (u observaciones) están comprendidas en $\overline{x} \pm s$.
- el **95,45 %** de los valores (u observaciones) están comprendidas en $\overline{x} \pm 2s$.
- el **99,73 %** de los valores (u observaciones) están comprendidas en $\overline{x} \pm 3s$.



Si se verifica la regla empírica entoences el comportamiento de la variable es simétrico y puedes utilizar como medida de posición central la media y el desvío estándar.

Medidas de Asimetría

Una distribución puede no ser simétrica. El sesgo es el grado de asimetría de una distribución. Si la curva de la variable tiene una curva más larga a la derecha del máximo central que a la izquierda, se dice que la distribución está **sesgada a derecha** o que tiene **sesgo positivo** o **asimetría positiva**. En caso contrario, se dice que está **sesgada a la izquierda** o que tiene **desgo negativo** o **asimetría**.
Una importante medida de estimación de la asimetría utiliza el momento de tercer orden, m_3, con respecto a la media, expresado en forma adimensional:

$$m_3 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n}$$

$$\text{Asimetría} = a_3 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{ns^3} = \frac{m_3}{s^3}$$

Medidas de Curtosis

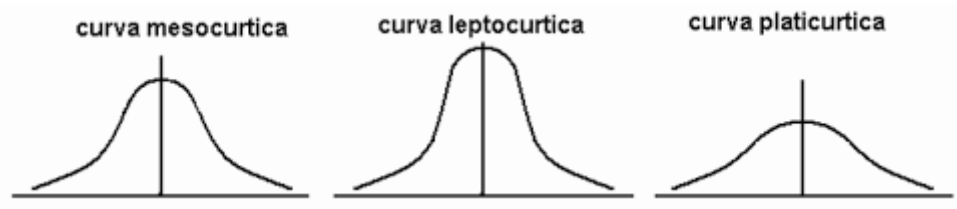
Es el grado de apuntamiento o curtosis de una distribución, se puede calcular empleando el momento de cuarto orden:

$$m_4 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n}$$

$$\text{Curtosis} = a_4 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{ns^4} - 3 = \frac{m_4}{s^4} - 3 = \frac{m_4}{(m_2)^2} - 3$$

En a_4 se restan **3** con objeto de generar un coeficiente que valga 0 para una curva simétrica y tome a ésta como referencia de apuntamiento. Tomando la distribución simétrica como referencia, una distribución puede ser:

- **Leptocúrtica:** Más apuntada y con colas más anchas que la simétrica.
- **Platicúrtica:** Menos apuntada y con colas menos anchas que la simétrica.
- **Mesocúrtica:** Es la distribución simétrica respecto de su media.



Coeficiente de Variación (C.V. %)

Un problema que se plantea para el desvío estándar (ó desviación típica) es la dependencia de las unidades de medida de la variable. Para evitar este problema, introducimos el Coeficiente de Variación de Pearson que se define como:

$$CV\,\% = \frac{s}{\overline{x}} \times 100$$

En caso de que la media sea muy próxima a cero no debe usarse ya que el denominador es muy pequeño y puede dar un grado erróneo de la dispersión.
Cuanto menor sea el coeficiente de variación menor será la dispersión en el comportamiento del conjunto de datos, y de esa manera, la media será más representativa.

Boxplot

El boxplot es un gráfico que sirve para vizualizar el comportamiento de los datos de una muestra. El mismo se construye con 5 puntos o medidas:

- $min(x)$: mínimo valor observado en el conjunto de datos.

- q_1: 1º cuartil. Es la mediana de las observaciones que caen por debajo de la mediana.
- $Md = q_2$: 2º cuartil. Es el valor por sobre el cual se encuentre el 50 % de los datos.
- q_3: 3º cuartil. Es la mediana de las observaciones que caen por encima de la mediana.
- $max(x)$: máximo valor observado en el conjunto de datos.