PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Métodos y técnicas de investigación cuantitativa

César Antonio Aguilar

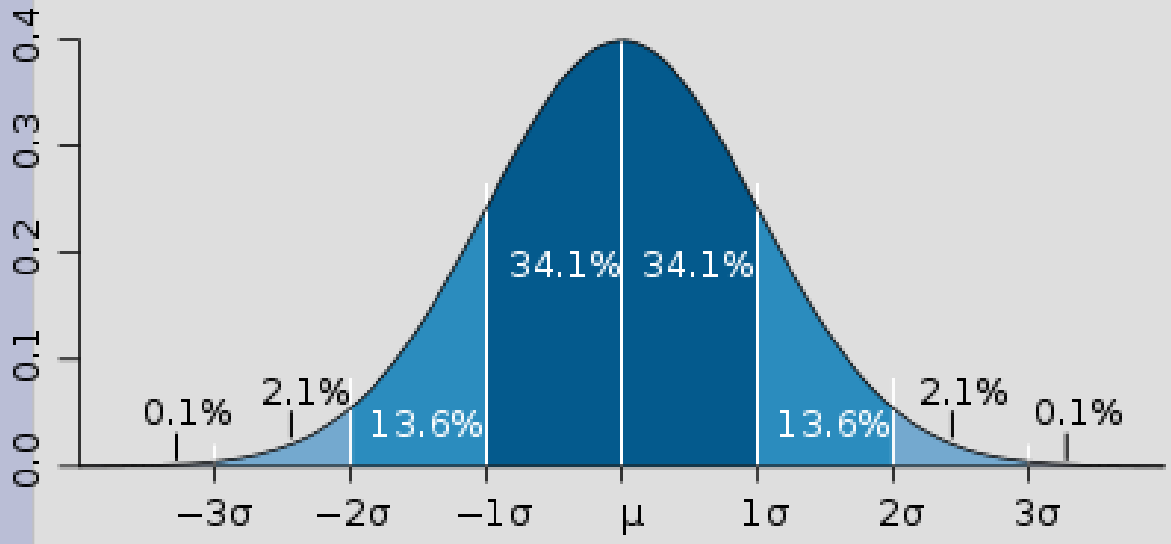
Facultad de Lenguas y Letras

27/05/2015

Cesar.Aguilar72@gmail.com

Mediana (1)

Vamos a retomar en esta clase nuestro análisis de tendencias centrales, y este caso analizaremos lo que se conoce como **mediana**. Así, podemos definirla como el valor de la variable de posición central en un conjunto de datos ordenados.



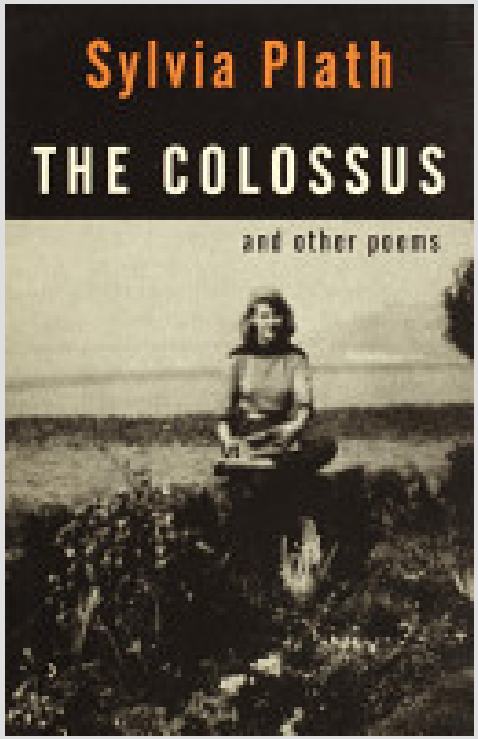
Nota: dada esta definición, podríamos pensar que tanto la media y la mediana son el mismo valor. Sin embargo, esto no es así. Recordemos entonces que la media representa un porcentaje promedio detectado en una población.

Mediana (2)

Para hacer esta distinción, veamos el siguiente ejemplo. Vamos a contabilizar las palabras que hay en un poema de Silvia Plath (1932-1963), titulado “Strumpet Song”: veamos el poema:

With white frost gone
And all green dreams not worth much,
After a lean day's work
Time comes round for that foul slut:
Mere bruit of her takes our street
Until every man,
Red, pale or dark,
Veers to her slouch.

Mark, I cry, that mouth
Made to do violence on,
That seamed face
Askew with blotch, dint, scar
Struck by each dour year.
Walks there not some such one man
As can spare breath
To patch with brand of love this rank grimace
Which out from black tarn, ditch and cup
Into my most chaste own eyes
Looks up.



Mediana (3)

La secuencia que observamos en este poema es la siguiente:

4, 7, 5, 7, 7, 3, 4, 4, 5, 5, 3, 5, 5, 7, 4, 9, 8, 6, 2.

Si organizamos la secuencia de palabras de manera ascendente, tenemos:

2, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 7, 7, 7, 7, 8, 9.

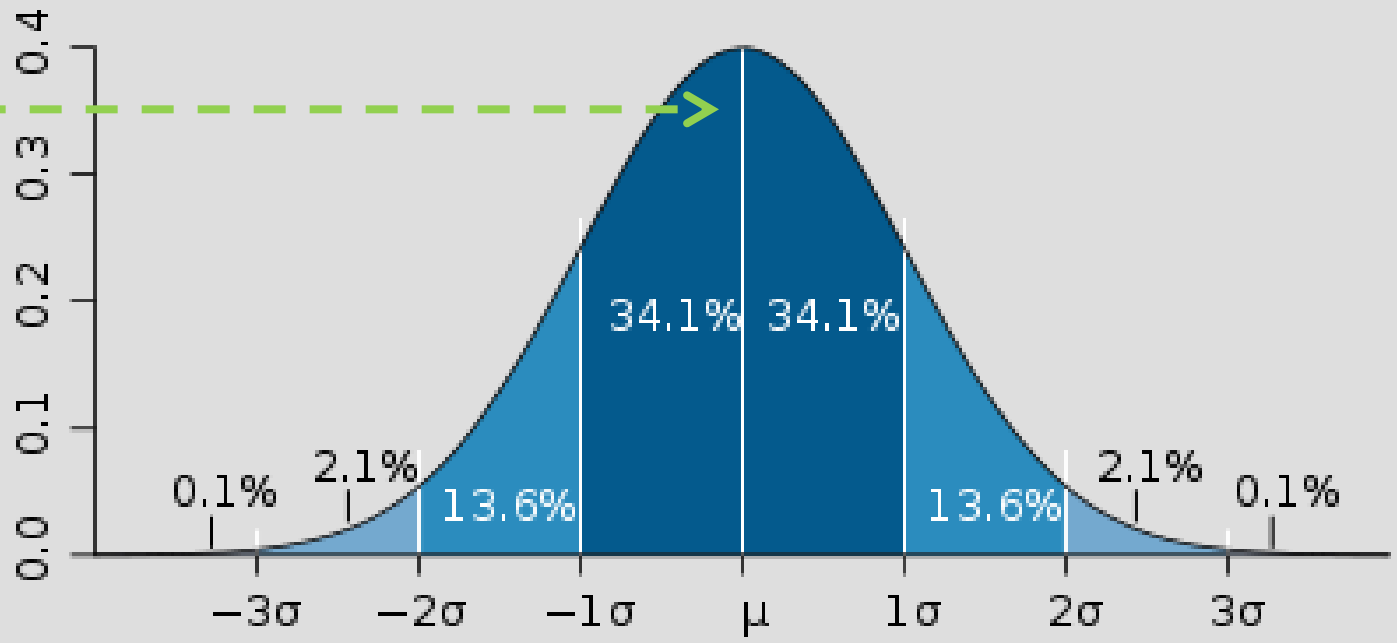
Preguntas:

1. ¿Cuál es el rango máximo de palabras que puede tomar un verso de este poema?
2. ¿Cuál es el valor medio de cada verso, y cómo lo obtendrían?

Mediana (4)

La mediana puede entenderse como el valor medio de un conjunto de datos ordenados. Este valor corresponde al al percentil 50 o segundo cuartil (P50 o Q2). Retomando la imagen que aparece en la primera lámina, esto quiere decir:

Si nuestros datos estuvieran ordenados, la mediana se ubicaría justo aquí:



Mediana (5)

Respondiendo a las preguntas anteriores, tenemos entonces que:

1. El valor máximo que podemos observar en cada uno de los versos del poema es cercano a las 10 palabras (o, en todo caso, 9).

2. El número central de palabras es 5 (o si quieren algo más preciso, 4.5), el cual lo deducimos calculando cuál es el valor medio de 10, además de que debe coincidir con la posición que ocuparía nuestro percentil 50 (o nuestro segundo cuartil 25).

3. Con estas coordenadas, deberíamos localizar la mediana justo a la mitad de nuestra curva de Gauss.

Calculando la mediana (1)

Vamos a trabajar en esta sesión considerando conjuntos de datos sin agrupar, haciendo la distinción entre valores impares y pares. Veamos los impares:

Supongamos que $X_1, X_2, X_3\ldots, X_n$, son los datos de una muestra ordenada en orden creciente, de modo que si designamos la mediana como M_e, distinguimos dos casos:

Si n es impar, la mediana es el valor que ocupa la posición $(n + 1) / 2$, una vez que los datos han sido ordenados (en orden creciente o decreciente), ya que ahí se da el valor central, es decir: $M_e = X(n +1)/2$.

Veamos entonces: si tenemos 5 datos, que ordenados son: $X_1 = 3$, $X_2 = 6$, $X_3 = 7$, $X_4 = 8$, $X_5 = 9$. En esta escala, el valor central es el tercero, y lo deducimos aplicando la fórmula: $X(5+ 1)/2 = X3 = 7$. Este valor, que es la mediana de ese conjunto de datos, deja dos datos por debajo (X_1, X_2) y otros dos por encima de él (X_4, X_5).

Calculando la mediana (2)

Si n es par, la mediana es la media aritmética de las dos observaciones centrales. Cuando n es par, los dos datos que están en el centro de la muestra ocupan las posiciones $n/2$ y $n/2+1$, esto es: $Me = X(n/2 +(n/2+1))/2$.

Un ejemplo: si tenemos 6 datos, que ordenados son: $X_1 = 3$, $X_2 = 6$, $X_3 = 7$, $X_4 = 8$, $X_5 = 9$, $X_6 = 10$. En esta escala, hay dos valores que están que están por debajo del $X_{6/2} = X_3 = 7$, y otros dos que quedan por encima del siguiente dato $X_{6/2 +1} = X_4 = 8$.

Por tanto, la mediana de este grupo de datos es la media aritmética de estos dos datos: $M_e = X_3 + X_4 / 2 = 7+8/2 = 7.5$.

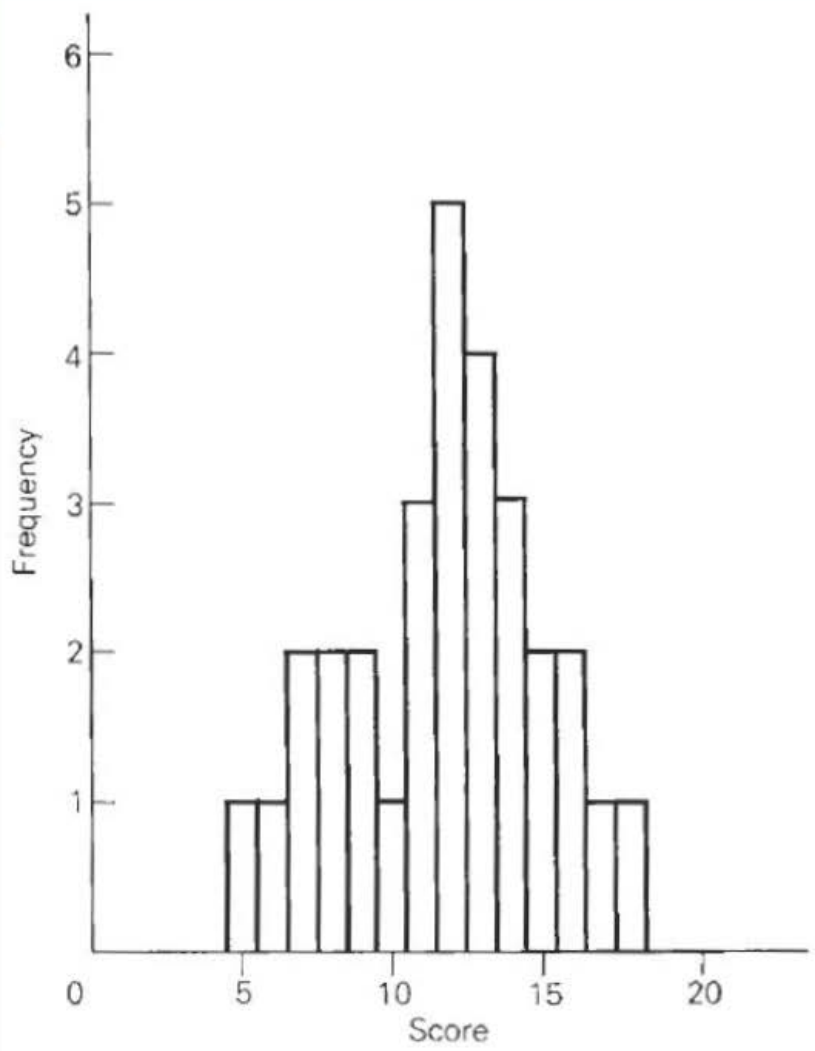
Calculando la mediana (3)

Apliquemos estos cálculos al ejemplo que vimos sobre los resultados de los estudiantes de inglés, a quienes aplicamos una prueba.

Retomando este ejemplo, ¿cuál sería el valor medio para los estudiantes que conforman el grupo B?:

Group B (Traditional) (N = 30)
11
16
14
18
6
8
9
14
12
12
10
15
12
9
13
16
17
12
8
7
15
5
14
13
13
12
11
13
11
7

Calculando la mediana (4)



Language test scores: group B

¿Recuerdan la distribución que mostraban ambos grupos conforme a nuestro polígono de frecuencias?

¿Cuál era el rango medio en donde se ubicarían la mayoría de los resultados generados por nuestros estudiantes?

Calculando la mediana (5)

De acuerdo con Butler, podemos deducir que la mediana entre ambos grupos es de 12, ya que a partir de este número (contando de abajo hacia arriba) se agrupan las calificaciones más bajas, en contraste con la ubicación de los puntajes más altos, situados después de 13.

Score	*Group A*		*Group B*	
5			/	1
6			/	1
7			//	2
8			//	2
9	/	1	//	2
10			/	1
11	//	2	///	3
12	//	2	~~////~~	5
13	///	3	////	4
14	////	4	///	3
15	~~////~~ /	6	//	2
16	///	3	//	2
17	////	4	/	1
18	///	3	/	1
19	//	2		

Calculando la mediana (6)

Ahora, para hacer un cálculo más exacto, podemos desarrollar la siguiente fórmula:

$$\text{median} = L + \left(\frac{N/2 - F}{f_{\text{m}}}\right)h$$

En donde:

L = El límite de intervalos más bajo considerado para la mediana

F = La suma de todas las frecuencias situadas por debajo del límite *L*.

f_m = Las frecuencias que se ubican en el rango de la mediana

N= Número de observaciones hechas (esto es, 30 resultados obtenidos)

h = La amplitud de los intervalos de clase (si los datos están agrupados)

Calculando la mediana (7)

Aplicando la fórmula, obtenemos:

$$
\begin{aligned}
L &= 11.50 \\
F &= 12 \\
f_m &= 5 \\
N &= 30 \\
h &= 1 \text{ (since the data are ungrouped, so that the 'class interval' is just 1 mark)}
\end{aligned}
$$

so that

$$\text{median} = 11.50 + \left(\frac{30/2 - 12}{5}\right) \times 1 = 11.50 + 0.6 = 12.10.$$

El resultado exacto debe ser 12.10, que es en donde estará ubicada la mediana del grupo B.

Moda (1)

La última medida que describe la tendencia central de un conjunto de datos es la **moda**, la cual podemos definir como el valor con una mayor frecuencia en una distribución de datos. Tomando de nuevo como ejemplo los grupos que participaron el examen de inglés, podemos intuir cuál es la moda de ambos grupos:

Así, en el caso del Grupo A, el valor más frecuente es 15, ya que hay 6 sujetos que caen en el rango de tal valor.

En el caso del Grupo B, sería 12, pues contamos 5 sujetos.

Score	*Group A*		*Group B*	
5			/	1
6			/	1
7			//	2
8			//	2
9	/	1	//	2
10			/	1
11	//	2	///	3
12	//	2	~~////~~	5
13	///	3	////	4
14	////	4	///	3
15	~~////~~ /	6	//	2
16	///	3	//	2
17	////	4	/	1
18	///	3	/	1
19	//	2		

Moda (2)

Tomando en cuenta la distribución de datos que observamos en un conjunto, podemos hablar de distribuciones **bimodales** (esto es, dos datos que tengan la misma frecuencia absoluta máxima), **trimodal** (si tenemos tres datos), y así de manera subsecuente.

Ahora bien, si en nuestro conjunto todos los datos muestran la misma distribución, se considera entonces que no existe ninguna moda.

Finalmente, al intervalo que tenga mayor frecuencia absoluta se le conoce como **intervalo modal**. Cuando tratamos con datos agrupados antes de definir la moda, se debe definir primero el intervalo modal.

Moda (3)

La moda, cuando los datos están agrupados, equivale a un punto que divide al intervalo modal en dos partes de la forma *p* y *c-p*, siendo c la amplitud del intervalo, el cual se puede verificar usando la siguiente fórmula:

$$\frac{p}{c-p} = \frac{n_i - n_{i-1}}{n_i - n_{i+1}}$$

Moda (4)

Ahora bien, la fórmula para obtener la moda es la siguiente:

$$Moda = L_{i_1} + \left(\frac{D_1}{D_1 + D_2}\right) + i$$

Donde:

L_{i-1}= Límite inferior de la clase modal.

D_1 = es el delta de frecuencia absoluta modal y la frecuencia absoluta premodal.

D_2 = es el delta de frecuencia absoluta modal y la frecuencia absoluta postmodal.

i = intervalo

Moda (5)

Para entender mejor esto, veamos este ejemplo: tratemos de encontrar la estatura modal de un grupo que se encuentra distribuido de la siguiente forma:

Entre 1 y 1.10 hay 1 estudiante
Entre 1.10 y 1.15 hay 1 estudiantes
Entre 1.20 y 1.25 hay 2 estudiantes
Entre 1.30 y 1.35 hay 2 estudiantes.
Entre 1.45 y 1.55 hay 3 estudiantes.
Entre 1.50 y 1.60 hay 4 estudiantes.
Entre 1.60 y 1.70 hay 10 estudiantes.
Entre 1.70 y 1.80 hay 8 estudiantes.

Clase modal = 1.60 y 1.70 (es la que tiene frecuencia absoluta más alta, 10)

Moda (6)

Haciendo nuestros cálculos, tenemos:

Li-1 = 1.60

D1 = 6

D2 = 2

i = 0.10

D1 y D2 los deducimos restando sus frecuencias absolutas, de tal modo que para D1 tenemos: 10 – 4 = 6; y para D2 tenemos 10 – 8 = 2.

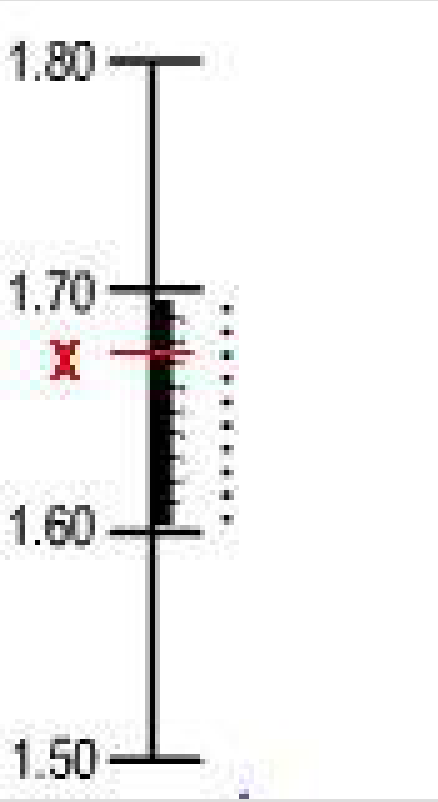
Ahora bien, el intervalo *i* lo deducimos corroborando las diferencias de valores con el intervalo modal, esto es 1.60/1.70, respecto a 1.50/1.60 (esto es, D1), y 1.70/1.80 (es decir, D2). Esto nos da como resultado: 0.10.

Moda (7)

Finalmente, obtenemos:

Moda = 1.60 + (6/8) * 0.1 = 1.675

Esto quiere decir entonces la estatura 1.67 es el valor más recurrente respecto a la frecuencia en una distribución de estaturas que vemos con en esta población de estudiantes, considerando el contraste entre los que miden 1.50 y 1.80.



Ejercicio (1)

Vamos a resolver un ejemplo hipotético con una búsqueda de **colocaciones** dentro de un corpus.

En este sentido, entendemos por *colocación* una expresión formada por dos o más palabras, la cual equivale a tipo de expresión convencional. Al respecto, podemos hacer la siguiente distinción:

- **Colocaciones fijas**, p. e.: *strong tea* vs *powerful tea*.

- **Colocaciones flexibles**, las cuales se fijan a través de relaciones más laxas entre palabras:

*She **knocked** on his **door***
*They **knocked** at the **door***
*100 women **knocked** on Donaldson's **door***
*A man **knocked** on the metal front **door***

Ejercicio (2)

Las palabras que aparecen entre ***knocked*** y ***door*** varían, de modo que la distancia entre las dos no es constante.

Sin embargo, existen suficiente regularidad en los patrones para determinar que ***knock*** es el verbo correcto en inglés para esta situación y no *hit*, *beat* o *rap*.

En el caso de ***door***, podemos considerar también que esta entidad forma parte de la estructura argumental que proyecta el evento ***knock***, y que en algunos casos puede ser mucho más recurrente que otras posibles entidades seleccionadas por el mismo verbo, tales como *table*, *metal*, *pocket*, etc.

Ejercicio (3)

1. Calcular la media, la mediana y la moda de los desplazamientos entre dos palabras en un corpus. En este caso, la media simplemente será el desplazamiento promedio.

2. Para el caso del ejemplo entre *knock* y *door* mostrado en las 4 oraciones anteriores la media será:

$$\frac{1}{4}(2+2+4+4)=3.0$$

Ejercicio (4)

Finalmente, la secuencia que vamos a analizar es de 20 oraciones, distribuidas de la siguiente forma:

Grupos de oraciones	Intervalo de palabras entre *knock* y *door*
3	1
4	2
9	3
2	4
1	5
1	6
20 oraciones	

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/meacutetodos-y-teacutecnicas-de-investigacioacuten-cuantitativa.html