MEDIDAS DE RESUMEN PARA DATOS CUANTITATIVOS[1]

Cuando los datos corresponden a la medición de variables en escalas cuantitativas, la forma de resumirlos es mediante estadísticos de posición central, no central y de dispersión. En esta sección vamos a estudiarlos en profundidad. Recuerde que éste documento se complementa con los videos que se encuentran disponibles en https://www.youtube.com/watch?v=_2x5dI5hR2Q&feature=youtu.be y en https://www.youtube.com/watch?v=e-vA2cG4wAk&feature=youtu.be

MEDIDAS DE POSICION CENTRAL

Los estadísticos de posición central son la media, la mediana y el modo.

Media

La media es la medida de resumen de posición central más conocida y utilizada en investigaciones médicas. Se obtiene mediante la suma de cada valor dividido la cantidad de observaciones. La notación para la media muestral es $\bar{x}$ y la fórmula matemática de la media es:

$$\bar{x} = \frac{\sum x_i}{n}$$

Donde la *letra* griega $\sum$. significa sumatoria, x_i representa cada valor de la variable "x" y "n" hace referencia al total **de** observaciones. Esta fórmula supone que se dijera "sume cada valor de x y divídalo por la cantidad de observaciones"

Suponga que recolecta la edad, expresada en años, de 5 consultantes a un servicio de salud, las cuales son:

2-3-5-5-6

[1] Documento adaptado y elaborado por el Prof. Mgter. Sebastián Genero, en el año 2018

La media se obtiene sumando cada valor de las observaciones, es decir, 2+3+5+5+6=21 dividido por la cantidad de observaciones, que en este caso son 5 personas. Entonces la media es =21/5=4,2 años.

Como observa, el valor de la media corresponde a uno del centro de la serie de datos (que en este caso presenta valores de 2 a 6) y nos retribuye una imagen de la edad de los consultantes. No es lo mismo la imagen que se nos presenta si la media de edad es de 4,2 años a que si fuera de 65,4 años. Esta es la utilidad de la media, y de todos los estadísticos de resumen, la de justamente, sintetizar en un valor alguna característica de la población de estudio a través de la cual, podemos tener una idea general de la misma. A su vez, cada vez que recurrimos a una medida de resumen, se pierde en detalle, dada la característica de la misma (resumir, no detallar, la información).

En nuestro ejemplo, la media es de 4,2 años, pero observe que ningún consultante tenía 4 años. Es decir, la media representa el valor típico, promedio, de los datos aunque a veces, ningún dato sea su valor.
Las medidas de posición central deben ser acompañadas de una medida de dispersión. En el caso de utilizar la media debemos presentarla con el Desvío Estándar, que estudiaremos más adelante.

La media solo es posible de ser utilizada para serie de datos numéricos, y es la medida de posición central preferida para los mismos, por diferentes razones (es de fácil cálculo e interpretación y porque la mayoría de las pruebas estadísticas la utilizan) sin embargo, no es recomendable cuando existen valores aberrantes en la serie de datos, porque **la media es sensible a valores extremos**.
Observe lo que ocurre con la media de nuestra serie de datos, si se introduce un valor extremo:

2-3-5-5-**86**

La media resulta en 20,2 años. Observe que ahora aunque la mayoría de los consultantes son niños, la media nos retribuye una imagen mental de una población adulta joven, que no se corresponde con la realidad. Ello ocurre porque en el cálculo de la media se utiliza cada uno de los valores de la serie de datos, y por eso, la introducción de un valor aberrante desplaza la media hacia su dirección (si el valor es muy alto, como en este caso, la media pasa de 4,2 a 20,2, es decir, aumenta y si el valor fuera muy bajo, la media disminuye)[2]

Cuando en la serie de datos se presentan valores extremos la medida de posición central más recomendable es la mediana que estudiaremos en la sección siguiente.

Mediana

Habíamos mencionado que, cuando en la serie de datos existen valores extremos, no es conveniente utilizar la media, en este caso es más recomendable utilizar la mediana.

[2] Todos aprendimos el efecto de valores extremos en la media cuando veníamos con estas calificaciones 10, 10, 10, 9 y 9 en nuestra carrera (media=9,6) y en una materia obtuvimos un 4 (media=8,6; aunque hubiera sido peor obtener un 1, porque la media sería=8,1)

La mediana (se representa con "m" o "Ma") de un conjunto de n mediciones es el valor de "x" que corresponde al centro espacial de la serie, una vez ordenadas los datos.[3]

En esta serie de datos es fácil observar cual es la mediana (en este ejemplo es el 5 que hemos resaltado en negrita)

2-3-**5**-5-86

Como en el cálculo de la Ma no se tienen en cuenta los valores de cada observación, sino su posición, no se desplaza en dirección a los valores extremos, como ocurre con la $\overline{X}$. Observe lo que ocurre en estas dos series:

2-3-5-5-6	2-3-5-5-86
$\overline{X}$=4,2	$\overline{X}$=20,2
Ma=5	Ma=5

Mientras que la $\overline{X}$ se ha desplazado hacia la dirección del valor extremo introducido en la segunda serie, la Ma no se ha modificado.

La mediana es el valor del centro espacial de la serie, y por lo tanto, existen antes y después del mismo, la misma cantidad de observaciones. La **posición del valor** de la mediana de una serie de datos se obtiene mediante la fórmula:

$$Ma = \frac{n+1}{2}$$

Donde "n" es la cantidad de observaciones. En esta serie de datos: 2-3-5-5-6, la Mn= 5+1=6 y 6/2=3. Recuerde que el resultado de la fórmula es la posición en la cual se encuentra el valor de la Ma, por lo tanto el resultado 3 debe ser interpretado como si se dijera "en el tercer lugar se encuentra la mediana", como observa, el número que está en el tercer lugar, una vez ordenada la serie de datos es el 5. En el caso de series impares, como la del ejemplo, el valor de la mediana corresponde a aquel que deja antes y después la misma cantidad de observaciones. Esto es un problema si la serie es par como en esta:

2-3-**5-6**-7-8

Aplicando la fórmula obtenemos 6+1=7 y 7/2=3,5. Este 3,5 se interpreta "entre la 3º y 4º posición", los valores que ocupan esos lugares son 5 y 6. Para obtener el valor de la mediana en este caso se debe calcular la media entre esos valores, que en este caso, resulta en 6,5 (5+6=11, y 11/2=6,5).

[3] Aunque es indiferente si el ordenamiento es de menor a mayor, o de mayor a menor, en general se utiliza el primero.

Por último, mientras que la media se utiliza solo con datos numéricos, la mediana es posible de utilizarla para variables medidas en escalas cualitativas ordinales. Considere el grado de satisfacción de 5 usuarios de un centro de salud:

Poco satisfecho Poco satisfecho Satisfecho Muy satisfecho Muy satisfecho

Ma=5+1/2)= 6/2=3

La mediana de esta serie de datos es "Satisfecho" lo cual implica que la mitad se encuentra menos que Satisfecho y la otra mitad, más que Satisfecho.

Modo

El Modo o moda (Mo) es la categoría (si la escala es cualitativa) o el valor (si es numérica) que se presenta con más frecuencia en una serie de datos. Por lo tanto, en la serie 2-3-5-5-6 el Mo es=5 y en la serie "Poco satisfecho/Poco satisfecho/Satisfecho/Muy satisfecho/Muy satisfecho" existen dos modos: Poco satisfecho y Muy Satisfecho.

Cuando dos valores ocurren con la misma frecuencia y ésta es la más alta, ambos valores son los modos, por lo que el conjunto de datos es bimodal; si más de dos valores ocurren con la misma frecuencia y ésta es la más alta, todos esos valores son modos, por lo que el conjunto de datos es multimodal y cuando ningún valor se repite, se dice que no hay moda (la distribución es uniforme).

El modo no es una medida de resumen muy utilizada en ciencias de la salud, aunque entre las distintas medidas de tendencia central que consideramos, es el único que puede usarse con datos medidos en escalas nominales, aunque como hemos estudiado, se disponen de otras medidas para resumir datos cualitativos.

MEDIDAS DE POSICION NO CENTRAL

En ocasiones es necesario ubicar la posición de un valor respecto del resto de los valores de una serie de datos. Esto se logra mediante los cuantiles. Se denomina cuantil de orden α, al valor de la variable por debajo del cual existe una frecuencia acumulada de x. Los cuantiles surgen de la división en partes iguales de una serie ordenada de datos. Veamos los siguientes ejemplos en base a una serie de 40 observaciones sobre la edad de los pacientes atendidos en un centro de salud:

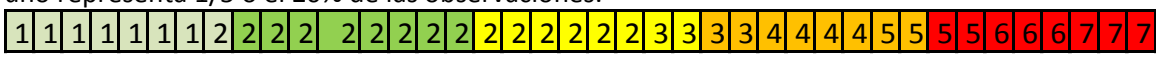
Al dividir la serie de datos en tres partes iguales, denominaremos a cada parte "tercil", cada tercil representa 1/3 del total de las observaciones, o, lo que es lo mismo, el 33,3%.

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

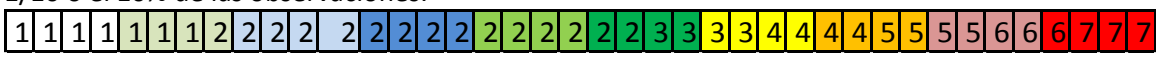
Si dividimos la serie de datos en 4 partes iguales, cada parte se denomina "cuartil". Cada cuartil representa 1/4 o 25% del total de las observaciones.



Si dividimos la serie de datos en 5 partes iguales, obtenemos los "quintiles", en cuyo caso cada uno representa 1/5 o el 20% de las observaciones.



Si dividimos la serie de datos en 10 partes iguales, obtenemos los "deciles". Cada uno representa 1/10 o el 10% de las observaciones.



Finalmente, si la serie de datos se dividiera en 100 partes iguales, cada parte se llamaría "percentil" y representaría el 1% del total de las observaciones. No se realiza el esquema de los percentiles, porque tiene sentido en caso de series de más de 100 unidades, pero son muy utilizados en las ciencias de la salud.

En estadística, también son ampliamente utilizados los cuartiles, así que volveremos a ese ejemplo. Además de la utilidad del uso de los cuartiles por las propiedades descriptivas que presentan, entender la división de una serie de datos en cuartiles también lo ayudará a interpretar el gráfico de cajas o "box plot" que trataremos en otro capítulo de este encuentro.

Previamente hemos definido a un cuantil de orden α, al valor de la variable por debajo del cual existe una frecuencia acumulada de x.

Considere el esquema presentado de los cuartiles. Cada una de las cuatro partes en que se ha dividido la serie de datos representa el 25% de las observaciones. Los cuartiles se denotan con la Q, y se reconocen entonces Q_1, Q_2, Q_3 y Q_4 Por lo tanto, el Q_1 es aquel valor por debajo del cual se encuentra el 25% de las observaciones. Visualmente, en este ejemplo, Ud. puede identificar el Q_1 como el último "2" del segmento verde.

El Q_2 es el último "2" del segmento amarillo, y hasta ahí se encuentra el 50% de las observaciones. Note que el Q_2 es el valor que se encuentra en el centro espacial de la serie, tal que por encima y por debajo del mismo existe un 50% de valores mayores y menores que él, respectivamente, por lo tanto el Q_2 coincide con la mediana. Finalmente, al llegar al Q_3 ya se han recorrido el 75% de los valores de la serie, en este ejemplo es Q_3 corresponde al último "4" del segmento naranja.

Q_1 Q_2 Q_3 Q_4



La fórmula para encontrar la posición del Q_1 es 0,25(n+1) y para encontrar la del Q_3 es 0,75(n+1). En este ejemplo: Q1=0,25(40+1)= 0,25*41=9,02 (lo que se interpreta como el valor que se

encuentre entre la 9º y 10º posición, lo que corresponde al último 2 de la serie verde y Q_3=0,75(40+1)=0,75*41=30,7. Ud. ya sabe cómo interpretar este resultado, así como también como calcular el Q_2 dado que como dijimos, corresponde a la mediana.

Determinar estos estadísticos nos permite responder preguntas como las siguientes:

Considerando el total de los consultantes, cuál es la edad del 25% de los más jóvenes?
Considerando el total de los consultantes, cuál es la edad del 25% de los más grandes?
Considerando el total de los consultantes, por debajo de que edad se encuentra la del 75% de ellos?

Antes de finalizar esta sección tenga en cuenta que la manera de interpretar los cuantiles es la misma en todos los casos, teniendo en cuenta la división que de los datos de haya hecho. Si una serie de datos de pesos de niños, expresada en kilogramos, se ha dividido en 100 partes iguales y, por lo tanto, cada parte es un percentil. Un niño que tenga un peso que se encuentre por debajo del percentil 3 debe ser interpretado como "el peso de este niño se encuentra en menos del 3% de los niños medidos", y por lo tanto, Ud. debe suponer que el peso de ese niño representa un valor "infrecuente". De hecho, esa es la base estadística del diagnóstico de desnutrición en los niños así como de otras situaciones en las que importa saber que tan raro o no, es un valor observado. Sobre este tema volveremos más adelante cuando estudiemos la distribución normal.

MEDIDAS DE DISPERSION

Las series de datos pueden tener la misma media pero con los datos pueden distanciarse de manera diferente a la media. Observe las dos distribuciones que se muestran en las figuras Nº3 y 4 que simulan calificaciones de dos exámenes de estadística realizadas a la misma cantidad de personas.

La media de ambas distribuciones es 6 pero hay una gran diferencia en la forma en que las mediciones se dispersan o varían. Mientras que las de la Fig. Nº3 varían de 5 a 7, las de la Fig. Nº4 varían de 2 a 10. Suponga que Ud. puede elegir el examen a rendir, cual elegiría?

Fig. Nº 3 Calificaciones de un examen de estadistica, diciembre de 2020, Facultad de Medicina, Corrientes, n=20, $\bar{x}$=6

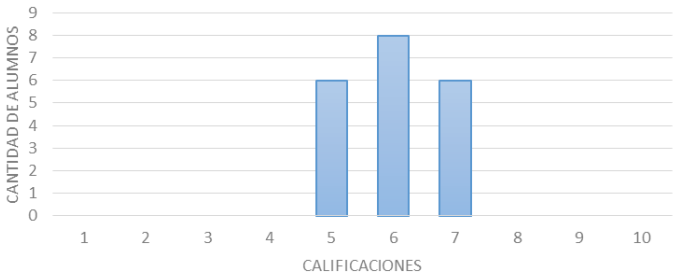
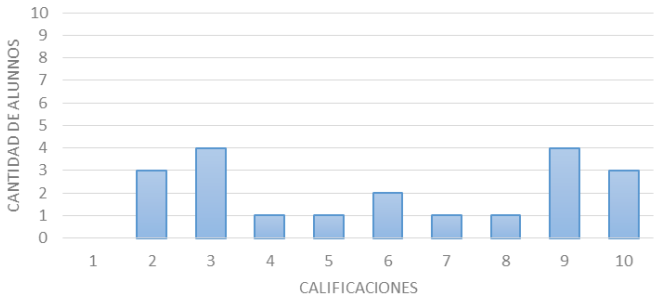


Fig. Nº 4 Calificaciones un examen de estadistica, dieciembre de 2020, Facultad de Medicina, Corrientes, n=20, $\bar{x}$=6



Dado que la imagen mental que se nos presentaría si solo conociéramos la media, sería insuficiente, es necesario que las medidas de centralización se acompañen de medidas de dispersión o variabilidad, con el propósito de mejorar la descripción de los datos.

Rango

El rango de un conjunto de mediciones se define como la diferencia entre el valor máximo y el mínimo. El rango puede ser una medida adecuada de variación para series de datos pequeñas. Pero, para series grandes, el rango no es una medida adecuada de variabilidad. En la serie de datos de la Fig. Nº3 el valor máximo es 7 y el mínimo es 5, y entonces el rango es=7-5=2, en cambio, en la Fig. Nº 4 en rango es=10-2=8.

Dado que el rango no ofrece mayor información no es muy utilizado. Existen otras medidas de variabilidad que retribuyen mayor utilidad.

Varianza

La varianza se mide en términos del cuadrado de las unidades originales de medición. Si las mediciones originales se refieren a años de edad, la varianza se expresa en "años cuadrados". Es evidente que la expresión 3,4 años cuadrados, como medida de variabilidad puede ser más precisa que el rango, pero es muy poco intuitiva. Que son los años cuadrados? Dado que este estadístico es complejo y se estudia en profundidad más adelante en nuestra carrera, en esta asignatura no lo vamos a desarrollar, pero es importante que recuerde que si se aplica la raíz cuadrada a la varianza se obtiene el **desvío estándar**, que es la medida de dispersión más utilizada en las ciencias de la salud.

Desvío estándar

El desvío estándar, se denota con "s" es la medida de dispersión más utilizada y se calcula aplicando la raíz cuadrada a la varianza. Por lo tanto:

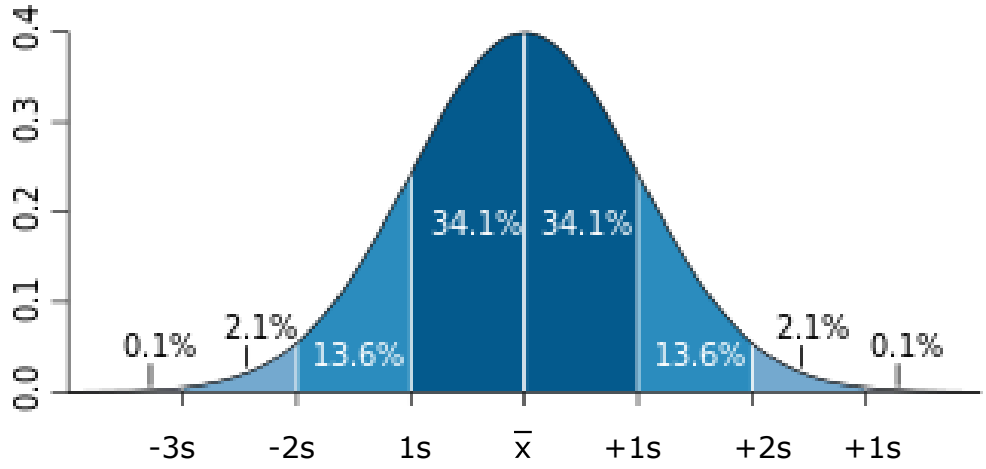
$$s = \sqrt[2]{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Donde: $\bar{x}$ = media aritmética, x= valor de cada una de las observaciones, Σ= sumatoria de los valores de las observaciones y n = número de observaciones.

Una manera de estimar qué proporción de observaciones se encuentran en un intervalo dado considerando la distancia de los valores respecto a su media, es aplicable cuando el conjunto de datos presenta una distribución similar a la curva normal de Gauss. Esta curva representa un tipo de distribución de frecuencias simétrica a la que se adaptan (por aproximación) muchos de los hechos y fenómenos biológicos y sociales. En ella, las tres medidas de tendencia central (media,

mediana y modo) coinciden. La curva de Gauss es simétrica a un lado y otro de la media, comprendiendo, cada una, el cincuenta por ciento de las observaciones. En la curva normal de Gauss, se ha estimado el porcentaje de observaciones que se encuentran a partir de la media, según la cantidad de desvió estándar que se consideren. La Fig. Nº 5 ilustra este concepto.

Fig. Nº5 Esquema de la curva de Gauss y el porcentaje de observaciones que se ubican en alrededor de la media según la cantidad s considerados



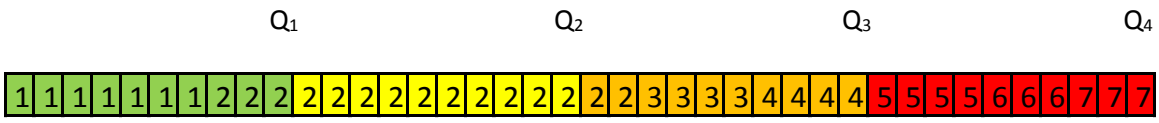
Entre la $\overline{x}$ + 1s se encuentra un 34% de las observaciones. Aplicado al ejemplo de la edad al diagnóstico de HIV/SIDA, en el cual la edad $\bar{x}$ del diagnóstico fue =33 años y el s=11, esto implica que un 34% de las personas con HIV/SIDA, se realizaron el diagnóstico entre los 33 y 44 años de edad (porque 33 que es la media + 11 que es el valor de 1s resulta =44) Como la curva es simétrica, esto también implica que otro 34% de realizó el diagnóstico de HIV/SIDA entre los 33 y 22 años de edad (porque 33 es la media -11 que es el valor de s resulta =22) Sumando ambos porcentajes podemos estimar que entre los 22 y 44 años de edad se realizó el diagnostico el 68,2% de los pacientes.

Usualmente, conviene recordar que en la curva de Gauss: el valor del desvío estándar sumado y restado a la media abarca el 68,2% de las observaciones. Si se utilizan dos desvíos estándar se abarca el 95,4 % y con tres el 99,6 % de las observaciones.

Antes de terminar recuerde que **siempre que se utilice la media como medida de tendencia central debe acompañarse con el desvío estándar como medida de dispersión.** Si no se ha utilizado la media como medida de tendencia central, se puede utilizar el rango que ya hemos estudiado antes, o el intervalo intercuartílico, que es el último tema que trataremos en esta clase.
Intervalo o rango intercuartílico

La última medida de dispersión que estudiaremos es el intervalo intercuartílico (IIQ). Esta es una medida de variabilidad adecuada cuando la medida de posición central empleada ha sido la

mediana. Se define como la diferencia entre el tercer cuartil (Q_3) y el primer cuartil (Q_1), es decir: $IIQ = Q_3 - Q_1$.
Como cada cuartil representa el 25% de las observaciones, la diferencia entre el Q_3 y Q_1 representa el 50% de las observaciones. Recuerde nuestro esquema, presentado previamente:



Si Q_3 es= 4 y Q_1 es=2, el IIQ= 4-2=2, esto implica que el 50% de las observaciones tienen una dispersión de 2 años de edad. Mientras más amplio el IIQ mas variabilidad de las observaciones.

BIBLIOGRAFIA

Palladino A C, Estadística, Diplomatura Superior en Epidemiología, Facultad de Medicina Universidad Nacional del Nordeste, año 2017

Hernández, R., Fernández, C. y Baptista, P. (2014). Metodología de la investigación (6a ed.). México, D.F., México: McGraw-Hill Interamericana.

Altman DG, Bland JM. The normal distribution. BMJ 1995; 310: 298.

Altman DG, Bland JM. Quartiles, quintiles, centiles, and other quantiles. BMJ 1994; 309: 996.