**UNIDAD 1 - PROBABILIDAD**

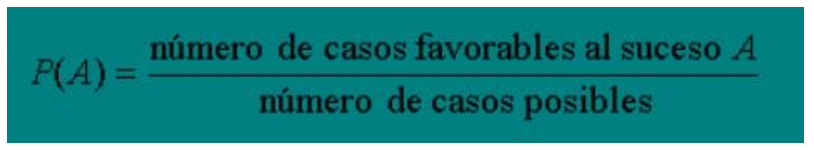
Para poder comprender que es la probabilidad primero debemos definir algunos concepto como suceso aleatorio, experimento y espacio muestral.

Un suceso o evento aleatorio es un acontecimiento que ocurrirá o no, dependiendo del azar. Los experimento o fenómenos aleatorios son los que pueden dar lugar a varios resultados, pero sin tener la certeza de cual de estos va a ocurrir.

El espacio muestral es el conjunto formado por todo los posibles resultados de un experimento aleatorio.

Por ejemplo, si el experimento consiste en lanzar un dado, el espacio muestral será el siguiente {1,2,3,4,5,6}, y un suceso aleatorio podría ser obtener un 3 en la tirada. Si el experimento consiste e lanzar una moneda, el espacio muestral será {cara, seca}, y obtener una cara seria un evento aleatorio particular-

Dicho esto, podemos definir la probabilidad de un suceso “A” como el coeficiente entre el número de resultados favorables a que ocurra dicho suceso en el experimento y el numero de resultados posible sal experimento. Volviendo al anterior ejemplos del dado, asumiendo que el mismo no esta trucado (todos los sucesos elementales del espacio muestral son equiprobables, es decir, tiene la misma probabilidad de ocurrir) la probabilidad de que se obtenga algún numero en particular es P(A) = 1/6



**PROBABILIDAD CONJUNTA**

Decimos que dos sucesos son independientes cuando la ocurrencia de uno no afecta o modifica la probabilidad de ocurrencia del otro. Si dos sucesos M y N son independientes, la probabilidad de ocurrencia de ambos sucesos simultáneamente será igual al producto de las probabilidad de cada suceso. Matemáticamente:

*P(M y N )=P(M∧N)=P(M)∙(N)*

Nos referimos a un suceso independientes cuando la ocurrencia de uno no afecta o modifica la probabilidad de ocurrencia del otro.

Si se tiene mas de dos suceso independientes, la probabilidad conjunta será igual al producto de las probabilidad de cada uno de los sucesos.

Ejemplo:

Supongamos que contamos con una mazo de 52 cartas, elegimos una de ellas al azar y anotamos su número. Luego, volvemos a mezclar esa carta en el mazo y repetimos el experimento. ¿Cuál es la probabilidad de haber obtenido dos ases consecutivamente?

🡪Definimos los sucesos:

M = Obtenemos un AS en la primera extracción.

N = Obtenemos un as en la segunda extracción.

Si la primera carta se repone, es decir, se coloca nuevamente en el mazo antes de hacer la segunda extracción, los suceso son independientes, por lo que podemos calcular la probabilidad conjunta de la siguiente manera:

*P(MyN) = P(M∩N) = 4/52 x 4/52 = 0.006 = 0.6%*

**DIAGRAMA DE ARBOL**

El diagrama de árbol es una manera de poder visualizar distintas combinaciones de sucesos – junto la probabilidad asociada.

Ejemplo:

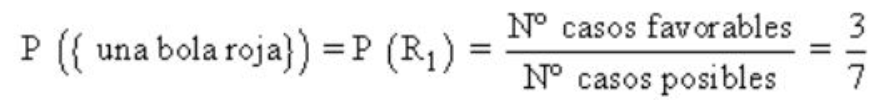
Tenemos una caja con 2 bolas verdes, 3 bolas rojas y 2 bolas azules. Vamos a sacar dos bolas al azar, devolviendo la primera antes de sacar la segunda. Este tipo de extracción se dice que es con reposición y asegura la independencia entre cada extracción., Queremos conoce la probabilidad de sacar dos bolas rojas.

En primer lugar definamos nuestros eventos:

🡪R1 = { sacar una bola roja en la primera extracción}

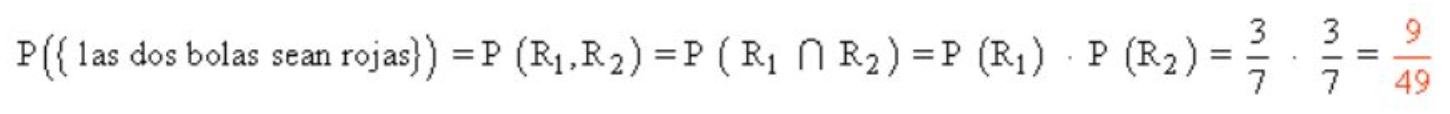
🡪R2 = { sacar una bola roja en la segunda extracción}

La probabilidad de suceso de R1 es la siguiente:

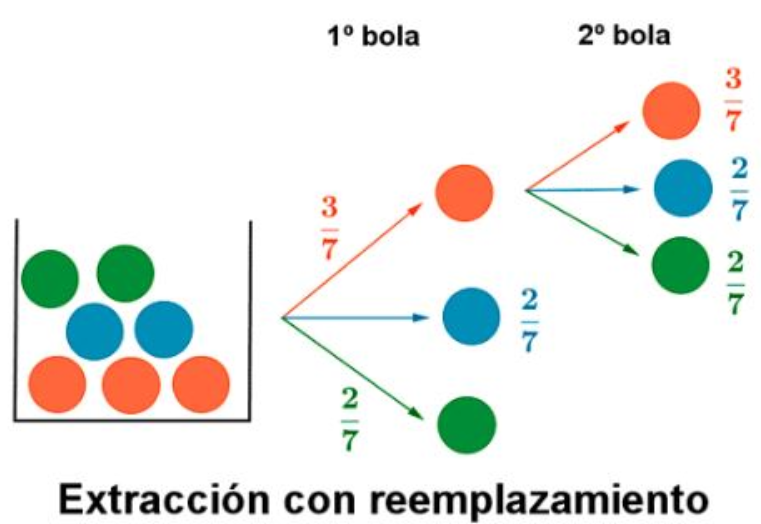


Esta probabilidad es la misma para la segunda extracción, por la reposición que mencionamos antes.

Entonces la probabilidad de dos bolas rojas es:



Veamos el razonamiento en el diagrama de árbol. Cada vez que avancemos por una rama, se multiplican las probabilidades.



**UNIDAD 2 - ESTADISTICA DESCRIPTIVA**

El uso de la estadística en el análisis de datos es fundamental, por distintos motivos, entre ellos:

* Favorece un metodología de trabajo sistemática
* Evita dar por validas afirmaciones que en realidad no tienen fundamentos científicos.

A grandes rasgos, podemos distinguir dos tipo de análisis estadísticos distintos:

* La estadística descriptiva, nos permite tener una primera aproximación a los datos al caracterizarlos a partir de ciertas medidas como las de tendencia central o dispersión.
* La estadística inferencial, que nos permite poner a prueba ciertas hipotesis y sacar conclusiones sobre las características de una población (por ejemplo, el ingreso promedio de todos los habitantes de Caba) a partir de los datos de una muestra (por ejemplo, una encuesta realizada de forma aleatoria a tan solo dos mil individuos que viven en la ciudad).

*“La estadística es la gramática de la ciencia”*

*Karl Pearson (1857 – 1936)*

TIPOS DE VARIABLES

Cuando recolectamos datos acerca de un determinado fenómeno, en general deseamos capturar información vinculada a distintas aristas que hacen a este fenómeno. Volviendo al ejemplo mencionado anteriormente , es probable que al encuestar a una persona, le preguntemos cuál es su género, cuantos años tiene, cuál es su máximo nivel educativo alcanzado, si esta empleado o no, entre otros. Cada uno de estos campos se considera una variable distinta. Es fundamental distinguir los tipos de variables que existen, ya que tienen gran influencia durante el análisis y nos condicionan durante el tratamiento de los datos.

**VARIABLE CUANTITAIVAS O NUMERICAS**

* Discretas: Dados dos valores, no toma valores intermedios. Por ejemplo, la lista de años de la década del ’90 (1990, 1991, 1992, etc.)
* Continuas: Dados dos valores, puede tomar cualquier valor numérico dentro de se intervalo. Por ejemplo, las ventas mensuales de una empresa ($1.500.000, $2.375.689,50)

**VARIABLE CUALITATIVAS**

* Categóricas: Se refieren a características o atributos que presenta una observación. Por ejemplo, genero (masculino/femenino) o el listado de las provincias argentina (Buenos Aires, Córdoba, Santa Fe, etc.)
* Ordinales: Se refieren a una secuencia de categorías ordenadas. Por ejemplo, los resultados de una encuesta de satisfacción con cierto servicio prestado a clientes (muy insatisfecho, insatisfecho, neutral, satisfecho, muy satisfecho) o las edades agrupadas por rangos etarios (0-14 años, 15-29 años, 30-59 años, 60 o más años)

**Ejemplo:** ¿Qué tipo de variables son?

1. Rango horario = Ordinal
2. Nombre = Categórica
3. $45 = Continua
4. $55,78 = Continua

MEDIDAS DE TENDENCIA CENTRAL

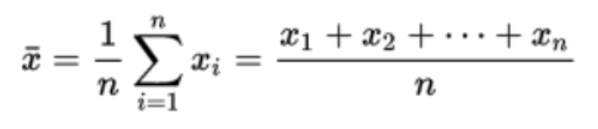
Las medidas de tendencia central son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro al cual se están ubicando los datos. La idea de dentro no es única y por eso hay mas de un criterio para definirlo. Las mas utilizadas son la media, moda y mediana,

Para cada tipo de variables se pueden usar ciertas medidas:

|  |  |
| --- | --- |
| **VARIABLES CUANTITATIVAS** | **VARIABLES CUALITATIVAS** |
| Media | Moda |
| Moda |  |
| Mediana |  |

MEDIA

El promedio o media aritmética se calcula haciendo la suma de todos los valores, dividida por la cantidad de valores. La interpretabilidad es muy intuitiva. La media se define de la siguiente forma:

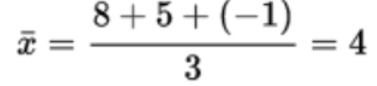


Donde:

x: los valores a promediar.

n: cantidad total de veces.

Por ejemplo, para el conjunto de valores [8, 5, -1 ] el cálculo de la media es el siguiente



Nota: Recordemos que la media solo se puede aplicar a variables cuantitativas. Resulta imposible calcularles la media a variables cualitativas.

Ejemplo:

1. ¿Cuánto vale la media de la siguiente muestra?

**{4,49,5,2,55,22,64,78,38,65}**

R : Media = 4 + 49 + 5 + 2 + 55 + 22+ 64 + 78 + 38 + 65 / 10 = 38,22

MEDIANA

La mediana puede pensarse de manera simple como el valor central o “del medio” de una lista ordenada de datos (o el valor que separa la primera mitad y a la segunda mitad de una distribución). En términos mas técnicos, la mediana es aquel valor de una distribución que acumula a ambos lados un 50% del total de elementos de esa distribución.

Para una lista ordenada, la mediana se calcula de diferente manera dependiendo de la cantidad de elementos que contiene.

* Para distribuciones con una cantidad total de elementos impar la mediana es el valor del elemento que separa en dos mitades de igual tamaño a nuestra lista ordenada de valores. Si tenemos [2,5,8,10,14,19], la mediana es 9.

Ejemplo: ¿Cuánto vale la mediana de nuestra muestra: [36,35,36,39.5,40,44,53,25,80,42]

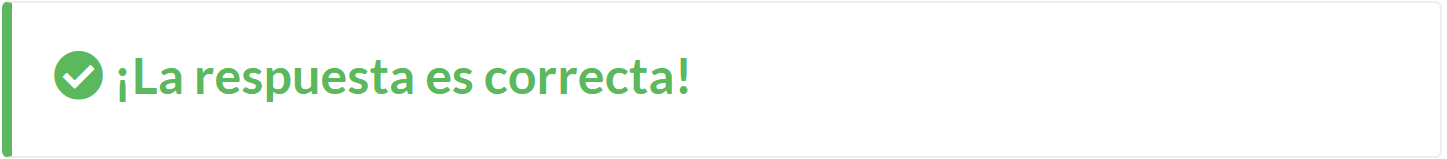
Lo primero que tenemos que hacer es ordenar los datos:

Entonces 🡪 [25,35,36,36,39.5,40,42,44,53,80]

Contamos la cantidad de elementos que tiene nuestra lista, en este caso nuestra lista tenemos 10 elementos.

Como es par, dividimos entre dos y sumamos uno, entonces los valores centrales serian 🡪 [39.5,40]

Sumamos y dividimos entre dos 🡪 (39.5 + 40) / 2 = 39.75

La mediana de nuestra muestra es = 39.75 

Moda

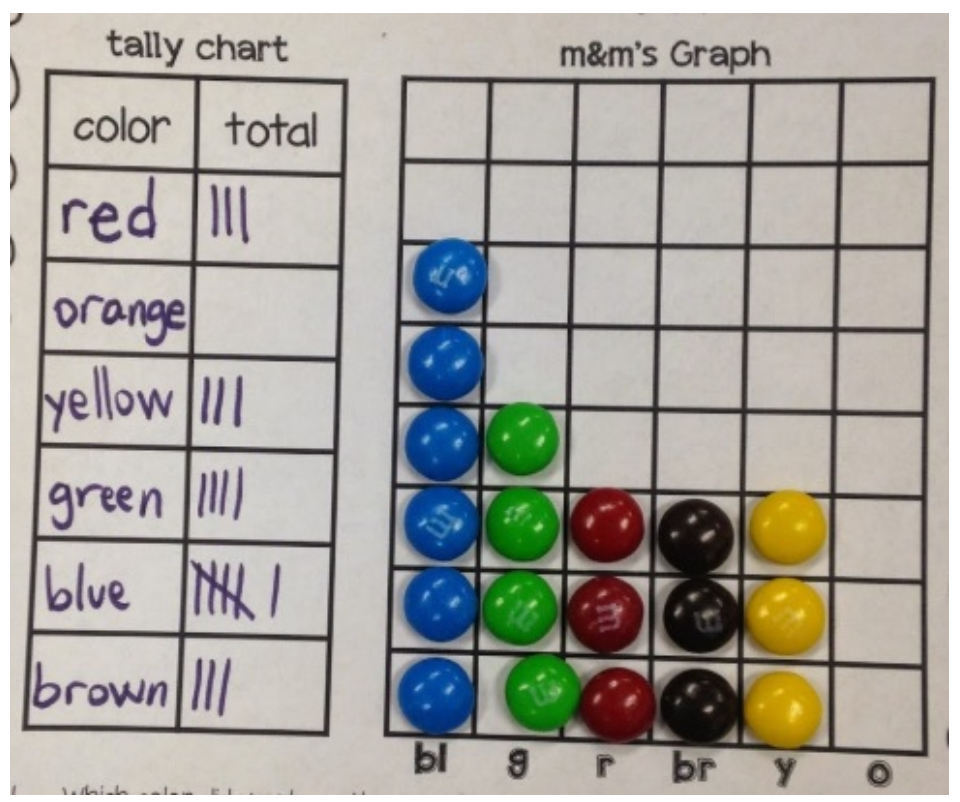
La moda es el valor que aparece con mayor frecuencia o el que mas se repite en una distribución. Esta medida se aplica tanto a variables cualitativa como cuantitativas, ya que cuenta repeticiones del elemento pero no opera con él.

Por ejemplo, la moda de [0,1,1,2,2,2,2,2,3,3,4,4,4,5] es 2.

A diferencia de las otras medidas, la moda no es necesariamente única. Puede ocurrir que haya dos valores diferentes que se repitan la misma cantidad de veces.

Por ejemplo, la moda de [10,10,13,13,20] son tanto 13 como 10. En este caso, decimos que se trata de una distribución bimodal.

Ejemplo:

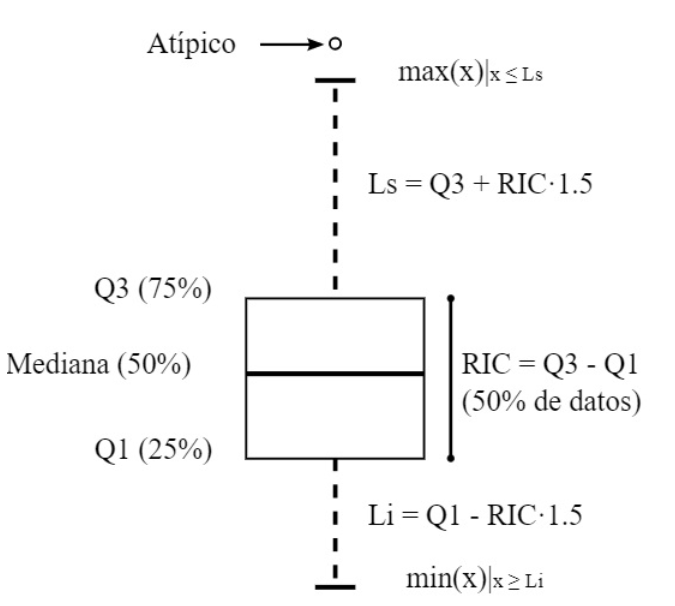


VISUALIZACIONES: BOXPLOT

En estadística, suele ser de mucha utilidad representar la información a través de gráficos, ya que permite condensar la información en una sola visualización, relacionar los conceptos y darle cierta interpretabilidad en algunos casos.

**BOXPLOT**

El *boxplot* o *diagrama de caja con bigotes* nos da una distribución de los datos y nos indica los cortes de los *cuartiles* y la *mediana*, e incluso permite visualizar *casos atípicos*.

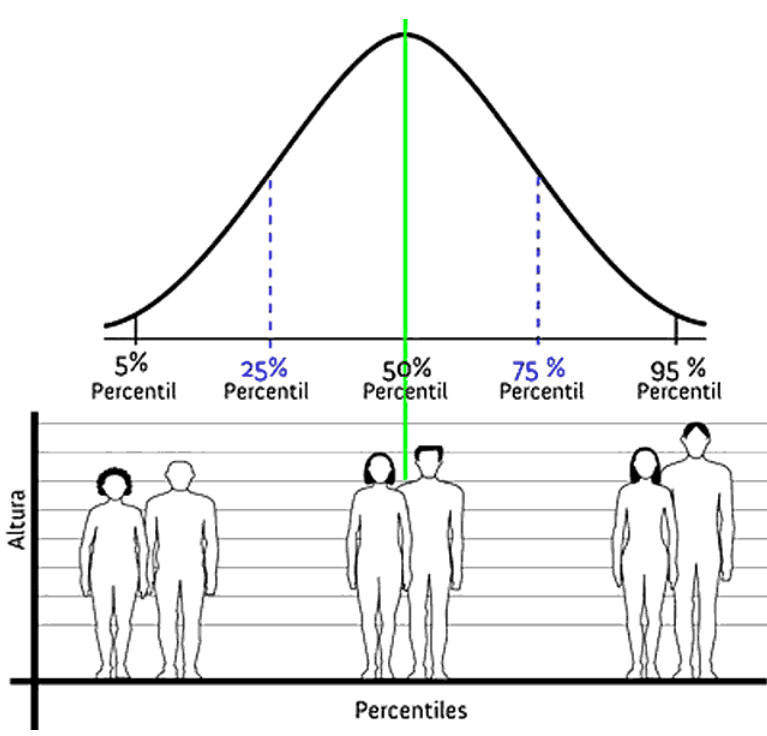


**PERCENTILES Y CUARTILES**

El percentil es una medida de posición que indica, una vez ordenados los datos de menor a mayor, el porcentaje de observaciones que se acumulan hasta ese valor. Veámoslo con un ejemplo simple:

Supongamos que tenemos una muestra de las alturas de 100 alumnos y alumnas de una universidad y las ordenamos de menor a mayor; el percentil 75 incluye las 75 primeras observaciones independientemente de cuál sea su valor.

Si quisiéramos separar nuestras observaciones en *cuatro grupos de igual tamaño*, podríamos definir *tres puntos de corte: el percentil 25, percentil 50 y el percentil 75*. Estos puntos de corte, justamente, reciben el nombre *cuartiles*, y son representados en un *boxplot* por la línea inferior, la línea central y la line superior de la caja, respectivamente. De esta forma, vemos que la caja los valores propios del 50% central de nuestros datos.



**RANGO INTERCUARTILICO, BIGOTES Y CASOS ATIPICOS**

Ya vimos que los limites de la caja vienen dados por los valores del primer cuartil (percentil 25) y del tercer cuartil (percentil 75). A la diferencia entre ambos se llama comúnmente *rango intercuartílico*:

*RIC = Q3(75%) – Q1(25%)*

Los *bigotes*( *whiskers* , en inglés) son las líneas que se extienden por fuera de la caja. Para definir su largo, debemos *multiplicar al* *RIC por algún valor arbitrario*, típicamente 1,5 (más adelante veremos que este valor tiene estrecha relación con las características propias de la *distribución normal*). Restando y sumando este valor a los valores de los bordes inferior y superior de la caja, respectivamente, obtenemos los limites inferior y superior de nuestro *boxplot*:

*Límite inferior = Q1 – 1.5 \* RIC*

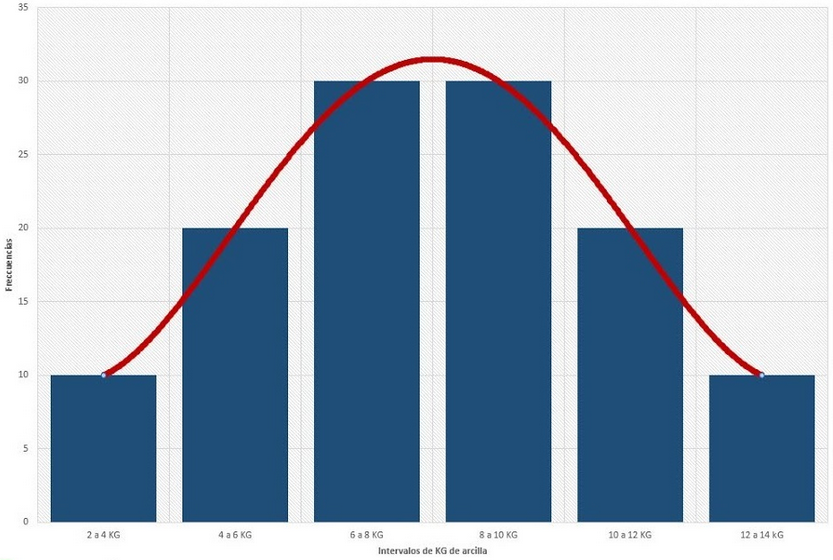
*Límite superior = Q3 + 1.5 \* RIC*

Todas las observaciones que queden por fuera de los bigotes serán consideradas *casos atípicos* (o *outliers*, en inglés) es decir, aquellos registros que presentan valores muy por encima o por debajo de la mayora parte de nuestros datos. Estas observaciones deberán ser tratadas con especial atención, ya que van a tener suma influencia sobre los cálculos de las medidas de tendencia central y dispersión mas frecuentes. Los outliers, lejos de representar valores de la realidad, muchas veces son el resultado de errores de registro típicos de cualquier proceso de recolección de datos (ya sea manual o automática).

HISTOGRAMAS

Un histograma es una representación grafica en forma de barras de frecuencias que poseen los valores de una variables determinada. En este tipo de visualización, el alto de las barras representa que tanto se repiten los valores (en frecuencia absoluta o relativa), mientras que el ancho queda definido por los rango de valores que entres en consideración.

A continuación se muestra un histograma correspondiente al peso de conjunto de bolsas de arcilla:



Resulta útil recorrer el paso a paso del armado de un histograma para terminar de comprenderlo:

1. Si se trata de datos cuantitativos, debemos definir intervalos para sementarlos. Si se trata de datos cualitativos, no hace falta este paso.
2. Si se trata de datos cuantitativos, indicamos cuantas observaciones hay en cada intervalo. Si se trata de valores datos cualitativos, simplemente contamos la cantidad de ocurrencias de cada categoría. En este punto, estamos definiendo las frecuencias de cada valor o rango de valores.
3. Graficamos con barras los resultados, indicando los intervalos de valores de la variable en el eje horizontal y sus respectivas frecuencias en el eje vertical.

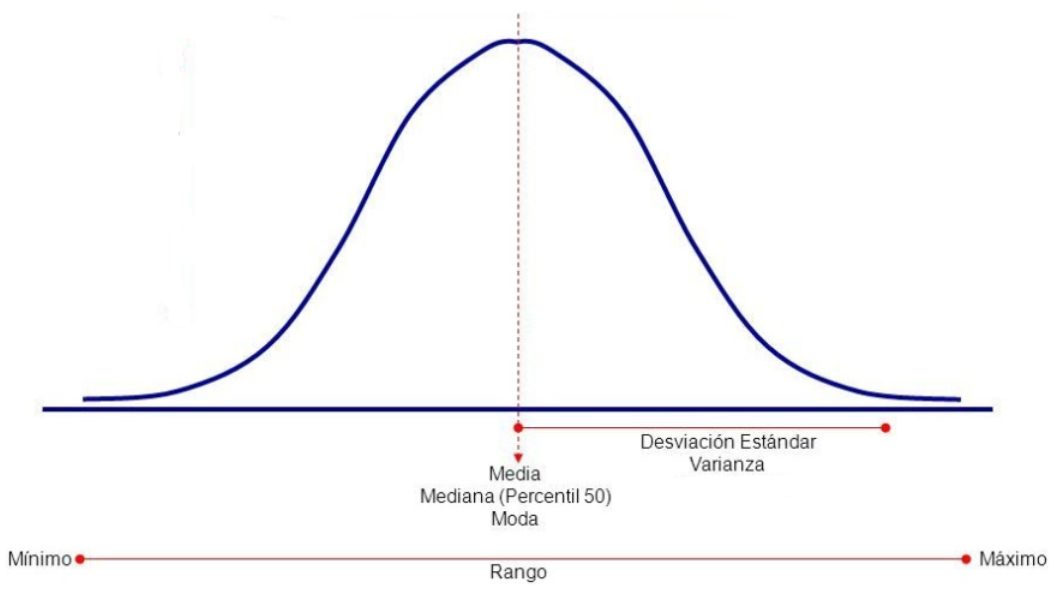
DISTRIBUCIONES ESTADISTICAS: NORMAL

La línea roja de la imagen anterior, representa la *distribución normal* esta representa y nos muestra como se distribuyen las observaciones. Podemos decir con toda certeza que la mayor cantidad de las observaciones se encuentran en los intervalos centrales y que los intervalos de los extremos parecen ser menos frecuentes, Esta distribución de frecuencia es muy común y se da en datos de distinta naturaleza; se llama *distribución normal.*

En el próximo grafico se puede observar una distribución normal típica, pero lo interesante es poder conectar lo que vimos hasta ahora. Veremos que están indicadas la media, la moda y la mediana, y que las tres medidas coinciden en el mimo punto central Por este motivo, decimos que la norma es un distribución simétrica. Repasemos las definiciones de medidas de tendencia central y comparemos con el grafico de distribución normal:

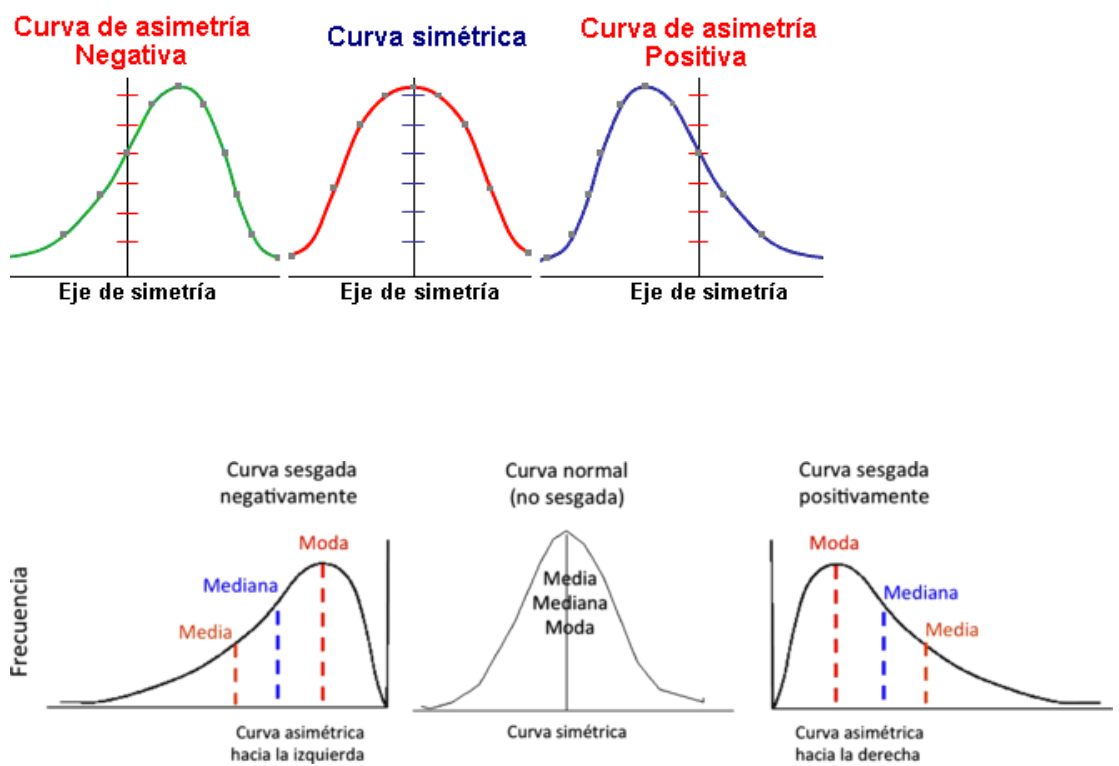
1. La moda representa el valor que mas se repite o el de mayor frecuencia. Siempre coincidirá con el pico o máximo del gráfico.
2. La mediana es aquel valor que deja el 50% de los datos de un lado y el 50% restante del otro lado. Siendo que la distribución normal es simétrica, o tiene otra opción que estar en la mitad del gráfico.
3. La media no es mas que el promedio de los valores de una variables. Si bien no es interpretable visualmente en un gráfico, si podemos decir que por propiedad de las distribuciones simétricas, es coincidente con la moda y la mediana.

En el siguiente gráfico, a parte de las medidas descriptas anterior mente observaremos la desviación estándar y la varianza:



DISTRIBUCIONES ASIMETRICAS

Las distribuciones no siempre son asimétricas, de hecho, la mayoría de las veces no lo son. La media, moda, y median se ven afectadas por la asimetría, es decir, por el lado haca el cual presentan un mayor es respecto de una distribución simétrica.



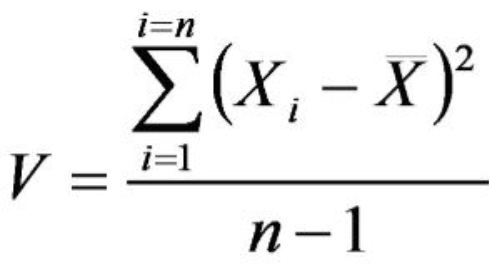
MEDIDAS DE DISPERSION: VARIANZA

Las medidas de dispersión son parámetros estadísticos que indican cuanto se alejan los datos respecto de la media. Sirven como indicadores de la variabilidad de los datos. Las medidas de dispersión mas utilizadas son el desvío estándar y la varianza.

**VARIANZA**

La varianza es un medida de dispersión y se calcula comparando las observaciones contra su media.

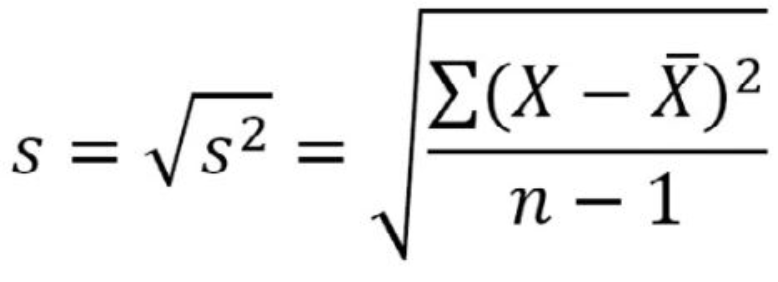
Esta es su expresión matemática:



Su unidad de medida corresponde al cuadrado de la unidad de medida de la variables: por ejemplo, si la variable mide una distancia en metros, la varianza se expresa en metros al cuadrado. La varianza tiene como valor mínimo 0.

DESVIO ESTANDAR

La expresión matemática en este caso es la siguiente:



MEDIDAS DE DISPERSION: COEFICIENTE DE VARIACION

El coeficiente de variación (CV) permite compara la dispersión de distintas variables. Como dijimos antes, varianza y el desvío comparan las observaciones contra la media, por lo tanto depende de ella. Ahora bien, si dividimos el desvío por la media nos estamos independizando de ella. De este modo, podemos s comprar variables con distintas medias.

Otras característica relevante del CV es que no tiene unidades ya que el desvío y la media tienen las mismas unidades, la unidad de la variable, y se cancelan al dividirlos. Esto también nos permite comprar variables que no tienen las mismas unidades.

Su expresión matemática es la siguiente:

