ESTADÍSTICA BÁSICA

1.) Conceptos básicos:

Estadística: Es una ciencia que analiza series de datos (por ejemplo, edad de una población, altura de un equipo de baloncesto, temperatura de los meses de verano, etc.) y trata de extraer conclusiones sobre el comportamiento de estas variables. Es una de las ciencias que permite conocer, o al menos entender, la realidad en la que nos desenvolvemos. A través de la estadística podemos obtener información de gran valor que nos ayudará en la toma de decisiones en cualquier ámbito de nuestra vida. El análisis de la información pasada para tomar la decisión más correcta, de cara al futuro, es el objeto de la estadística.

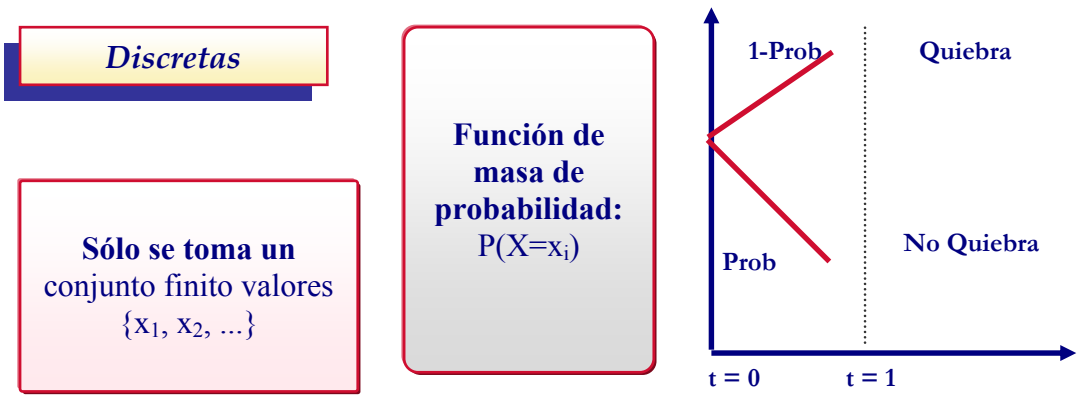
Variable aleatoria: Conjunto de distintos valores numéricos que adoptan un carácter cuantitativo. Es aquel dato susceptible de tomar diferentes valores en determinadas circunstancias. La estadística es el estudio cuantitativo de las variables, por lo que podemos considerar éstas como la materia prima de los estudios estadísticos. Toda variable que tiene asociada una determinada ley de probabilidad; cada uno de los valores que puede tomar le corresponde una probabilidad específica.

Las variables pueden ser cualitativas o cuantitativas,

Variables cualitativas (o categóricas): aquellas que no aparecen en forma numérica, sino como categorías o atributos (sexo, profesión, color de ojos).

Variables cuantitativas: las que pueden expresarse numéricamente (temperatura, salario, número de goles en un partido). Variables cuantitativas según el tipo de valores que pueda tomar pueden:

- **Discretas**: Aquellas que toman valores aislados (números naturales), y que no pueden tomar ningún valor intermedio entre dos consecutivos fijados.
  Por ejemplo; nº de goles marcados, nº de hijos, nº de discos comprados, nº de pulsaciones,...



- **Continuas**: Aquellas que toman infinitos valores (números reales) en un intervalo dado, de forma que pueden tomar cualquier valor intermedio, al menos teóricamente, en su rango de variación. Por ejemplo; talla, peso, presión sanguínea, temperatura, ..

Frecuencia: Número de veces en que se repite un dato. Distinguimos dos clases de frecuencias:

- **Frecuencia absoluta:** La frecuencia absoluta de una variable estadística es el número de veces que aparece en la muestra dicho valor de la variable.

- **Frecuencia relativa:** La frecuencia absoluta, es una medida que está influida por el tamaño de la muestra, al aumentar el tamaño de la muestra aumentará también el tamaño de la frecuencia absoluta. Esto hace que no sea una medida útil para poder comparar. Para esto es necesario introducir el concepto de *frecuencia relativa*, que es el cociente entre la frecuencia absoluta y el tamaño de la muestra.

Cuando se estudia el comportamiento de una variable hay que distinguir los siguientes conceptos:

(Componentes de una Investigación Estadística)

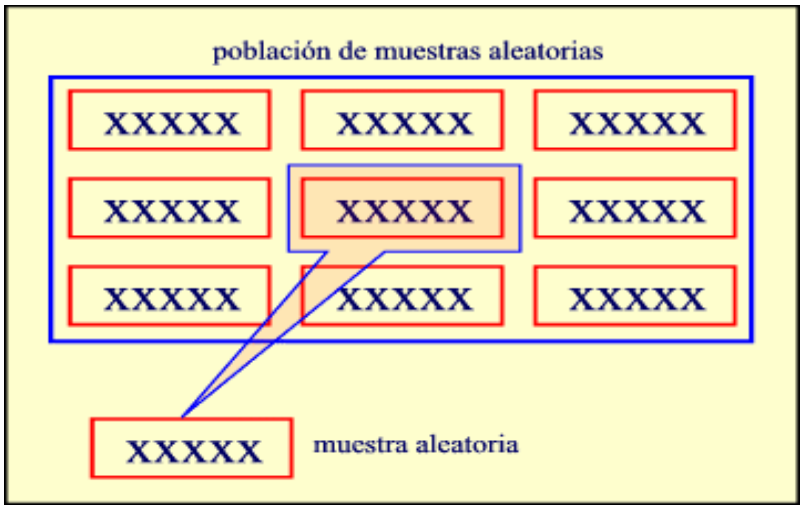
Población: Es el conjunto de todos los elementos que cumplen ciertas propiedades y entre los cuales se desea estudiar un determinado fenómeno (pueden ser hogares, número de tornillos producidos por una fábrica en un año, lanzamientos de una moneda, etc.). Llamamos población estadística o universo al conjunto de referencia sobre el cual van a recaer las observaciones.

Individuo: Se llama unidad estadística o individuo a cada uno de los elementos que componen la población estadística. El individuo es un ente observable que no tiene por qué ser una persona, puede ser un objeto, un ser vivo, o incluso algo abstracto.

Muestra: es el subconjunto de la población que es estudiado y a partir de la cual se sacan conclusiones sobre las características de la población. La muestra debe ser representativa, en el sentido de que las conclusiones obtenidas deben servir para el total de la población.
Las muestras pueden ser probabilísticas o no probabilísticas. Una muestra probabilística se elige mediante reglas matemáticas, por lo que la probabilidad de selección de cada unidad es conocida de antemano. Por el contrario, una muestra no probabilística no ser rige por las reglas matemáticas de la probabilidad. De ahí que, mientras en las muestras probabilísticas es posible calcular el tamaño del error muestral, no es factible hacerlo en el caso de las muestras no probabilísticas.
La modalidad más elemental de muestra probabilística es la muestra aleatoria simple, en la que todos los componentes o unidades de la población tienen la misma oportunidad de ser seleccionados.

Censo: Decimos que realizamos un censo cuando se observan todos los elementos de la población estadística.

Parámetro: Característica de una población, resumida para su estudio. Se considera como un valor verdadero de la característica estudiada.



Probabilidad: Es el conjunto de posibilidades de que un evento ocurra o no en un momento y tiempo determinado.

Dichos eventos pueden ser medibles a través de una escala de 0 a 1 (o la expresamos en tanto por ciento, entre 0% y 100%), donde el evento que no pueda ocurrir tiene una probabilidad de 0 y uno que ocurra con certeza es de 1, y el resto de sucesos tendrá probabilidades entre "cero y uno" que será tanto mayor cuanto más probable sea que dicho suceso tenga lugar.

Ejemplo: Cuando se lanza una moneda, se desea saber cual es la probabilidad de que se sello o cara, es decir existe un 0,5 (50%) de que sea cara o 0,5 (50%) de que sea sello.

El experimento tiene que ser aleatorio, es decir, que pueden presentarse diversos resultados, dentro de un conjunto posible de soluciones, y esto aún realizando el experimento en las mismas condiciones. Por lo tanto, a priori no se conoce cual de los resultados se va a presentar. Ejemplo: Lotería de Navidad.

Hay experimentos que no son aleatorios y por lo tanto no se les puede aplicar las reglas de la probabilidad.

Modelo de distribución de probabilidad: especificación de los valores de la variable aleatoria con sus probabilidades respectivas

2.) Medidas de variables aleatorias

En muchas ocasiones es mucho más eficaz, sencillo y preciso el estudio de una variable utilizando valores numéricos que la descripción visual de la distribución de una variable mediante tablas y gráficos, ya que los valores numéricos dan una idea de la ubicación o del centro de los datos (medidas de posición), y usando cantidades que informen de la concentración de las observaciones alrededor de dicho centro (medidas de dispersión).

a) Medidas de posición central:

Informan sobre los valores medios de la serie de datos. Una medida de centralización es un valor, que es representativo de un conjunto de datos y que tiende a situarse en el centro del conjunto de datos, ordenados según su magnitud.

Media: Es el valor medio ponderado de la serie de datos o valores que toma la variable estadística. La media no es más que la suma de todos los valores de una variable dividida entre el número total de datos de los que se dispone. Y se calcula como;

$$\overline{X} = \frac{x_1 + x_2 + x_3 + ...x_{n-1} + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Si el valor xi de la variable X se repite n_i veces, aparece en la expresión de la media aritmética de la forma:

$$\overline{X} = \frac{\sum x_i n_i}{n}$$

Siendo x_i las variables, n_i las veces que aparece la variable x_i y N la suma de todas las n_i. Es decir;

$$N = \Sigma n_i$$

A la media aritmética se la denomina también CENTRO DE GRAVEDAD de la distribución.

Mediana: Es uno de los cálculos más representativos de la muestra. La mediana es el valor del elemento intermedio cuando todos los elementos se ordenan. La mediana se calcula ordenando los datos de menor a mayor y tomando el valor del medio que es el que deja un 50% de observaciones a su izquierda y un 50% a su derecha.

El lugar que ocupa se determina dividiendo el nº de valores entre 2: $\frac{n}{2}$

Cuando hay un número impar de valores de la variable, la mediana será justo el valor de orden central, aquel cuya frecuencia absoluta acumulada coincida con $\frac{n}{2}$. Por tanto la mediana coincide con un valor de la variable.

El problema está cuando haya un número par de valores de la variable. Si al calcular $\frac{n}{2}$ resulta que es un valor menor que una frecuencia absoluta acumulada, el valor de la mediana será aquel valor de la variable cuya frecuencia absoluta cumpla la siguiente condición: $N_{i-1} < \frac{n}{2} \leq N_i \Rightarrow Me = x_i$.

Por el contrario si coincide que $\frac{N}{2} = N_i$, para obtener la mediana realizaremos el siguiente cálculo: $Me = \frac{x_i + x_{i+1}}{2}$

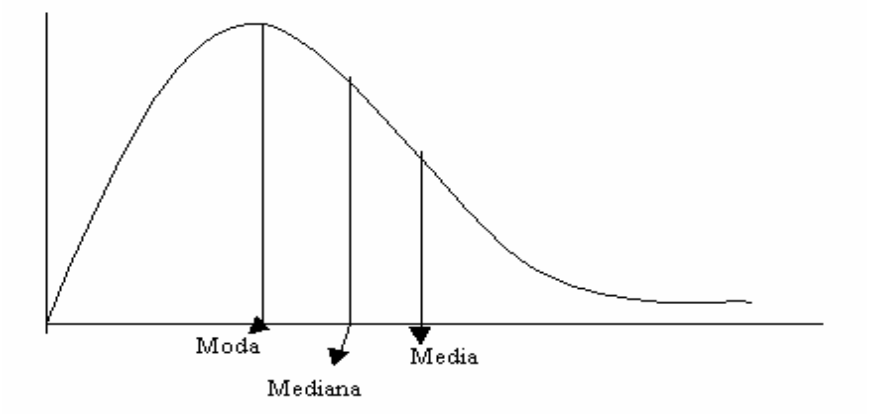
Moda: Es el valor más frecuente de la variable estadística; valor que se

corresponde al máximo del histograma.

Ejemplo: El conjunto 2,2,5,7,9,9,9,10,10,11,12 y 18 tiene moda 9.
Ejemplo: El conjunto 3,5,8,10,12,15 y 16 no tiene moda.

Ejemplo: El conjunto 2,3,4,4,4,5,5,7,7,7 y 9 tiene dos modas, 4 y 7 y se llama bimodal.

Una distribución con moda única se dice *unimodal.*



b) Medidas de posición no central:

Informan de cómo se distribuye el resto de los valores de la serie.

Los **Cuantiles** (cuartiles, deciles, percentiles) son medidas de localización, su función es informar del valor de la variable que ocupará la posición (en tanto por cien) que nos interese respecto de todo el conjunto de variables.

Podemos decir que los Cuantiles son unas medidas de posición que dividen a la distribución en un cierto número de partes de manera que en cada una de ellas hay el mismo de valores de la variable.

Las más importantes son:

CUARTILES, dividen a la distribución en cuatro partes iguales (tres divisiones). Q_1,Q_2,Q_3, correspondientes a 25%, 50%,75%.

DECILES, dividen a la distribución en 10 partes iguales (9 divisiones). D_1,...,D_9, correspondientes a 10%,...,90%

PERCENTILES, cuando dividen a la distribución en 100 partes (99 divisiones). P_1,...,P_{99}, correspondientes a 1%,...,99%.

Existe un valor en cual coinciden los cuartiles, los deciles y percentiles es cuando son iguales a la Mediana y así veremos:

$$\frac{2}{4} = \frac{5}{10} = \frac{50}{100}$$

Cuartiles: Los cuartiles son los tres valores que dividen al conjunto de datos ordenados en cuatro partes porcentualmente iguales.
Hay tres cuartiles denotados usualmente Q1, Q2, Q3:

El primer cuartil Q1, es el menor valor que es mayor que una cuarta parte de los datos; es decir, aquel valor de la variable que supera 25% de las observaciones y es superado por el 75% de las observaciones

El segundo cuartil Q2, (coincide, es idéntico o similar a la mediana, Q2 = Md), es el menor valor que es mayor que la mitad de los datos, es decir el 50% de las observaciones son mayores que la mediana y el 50% son menores.

El tercer cuartil Q3, es el menor valor que es mayor que tres cuartas partes de los datos, es decir aquel valor de la variable que supera al 75% y es superado por el 25% de las observaciones.

Deciles: Los deciles son ciertos números que dividen la sucesión de datos ordenados en diez partes porcentualmente iguales. Son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales, son también un caso particular de los percentiles, ya que podemos definir Decil como"percentil cuyo valor que indica su proporción es un múltiplo de diez. Percentil 10 es el primer decil, percentil 20 el segundo decil, etc".

El primer decil D1: indica que sólo existe un 10% de probabilidad de que el valor de la variable esté por debajo de esa cifra.

Quinto decil D5 o denominado también "Caso Base": indica que existe igualmente un 50% de probabilidad de que el valor esté por encima como por debajo de esa cifra. Representa la Mediana de la distribución.

Percentiles o centiles: Los percentiles son, tal vez, las medidas más utilizadas para propósitos de ubicación o clasificación de las personas cuando atienden características tales como peso, estatura, etc.

Los percentiles son ciertos números que dividen la sucesión de datos ordenados en cien partes porcentualmente iguales. Estos son los 99 valores que dividen en cien partes iguales el conjunto de datos ordenados. Sencillamente Percentil es el valor del recorrido de una variable, bajo el cual se encuentra una proporción determinada de la población.

Los percentiles (P1, P2,... P99), leídos primer percentil,..., percentil 99, muestran la variable que deja detrás una frecuencia acumulada igual al valor del percentil:

Primer percentil, que supera al uno por ciento de los valores y es superado por el noventa y nueve por ciento restante.

El 60 percentil, es aquel valor de la variable que supera al 60% de las observaciones y es superado por el 40% de las observaciones.

El percentil 99 supera 99% de los datos y es superado a su vez por el 1% restante.

c) Medidas de dispersión:

Son aquellas que permiten retratar la distancia de los valores de la variable a un cierto valor central, o que permiten identificar la concentración de los datos en un cierto sector del recorrido de la variable. Estudian la distribución de los valores de la serie, analizando si estos se encuentran más o menos concentrados, o más o menos dispersos.

Rango: Mide la amplitud de los valores de la muestra y se calcula por diferencia entre el valor más elevado y el valor más bajo.

$$R_e = x_{max} - x_{min}$$

Varianza: Mide la distancia existente entre los valores de la serie y la media. Se calcula como sumatorio de las diferencias al cuadrado entre cada valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. El resultado obtenido se divide por el tamaño de la muestra.

$$S_x^2 = \sigma_x^2 = \frac{\sum_{i=1}^{r}\left(x_i - \bar{x}\right)\cdot n_i}{N}$$

La varianza siempre será mayor que cero. Mientras más se aproxima a cero, más

$$\sigma = std(X) = +\sqrt{\mathrm{var}(X)}$$

concentrados están los valores de la serie alrededor de la media. Por el contrario, mientras mayor sea la varianza, más dispersos están.

Desviación Típica: es la raíz cuadrada de la varianza. Expresa la dispersión de la distribución y se expresa en las mismas unidades de medida de la variable. La desviación típica es la medida de dispersión más utilizada en estadística.

3.) Distribuciones de probabilidad

Como ya hemos mencionado anteriormente una **variable aleatoria** es aquella que toma diversos valores o conjuntos de valores con distintas probabilidades. Existen 2 características importantes de una variable aleatoria, sus valores y las probabilidades asociadas a esos valores.
Una tabla, gráfico o expresión matemática que dé las probabilidades con que una variable aleatoria toma diferentes valores, se llama **distribución de la variable aleatoria.**

La inferencia estadística (es decir, el proceso que realiza la herramienta Riesgómetro) se relaciona con las conclusiones que se pueden sacar acerca de una población de observaciones basándose en una muestra de observaciones. Entonces intervienen las probabilidades en el proceso de la selección de la muestra; en este caso se desea saber algo sobre una distribución con base en una muestra aleatoria de esa distribución.

De tal manera vemos que trabajamos con **muestras aleatorias de una población** que es más grande que la muestra obtenida; tal muestra aleatoria aislada no es mas que una de muchas muestras diferentes que se habrían podido obtener mediante el proceso de selección, por ello es de gran relevancia el uso de **distribuciones de probabilidad**.

Distribuciones discretas:

Son aquellas en las que la variable puede pude tomar un número determinado de valores. Existen diversos tipos, entre los que destacan:

Bernouilli; Es aquel modelo que sigue un experimento que se realiza una sola vez y que puede tener dos soluciones: acierto o fracaso:

Cuando es acierto la variable toma el *valor 1*

Cuando es fracaso la variable toma el *valor 0*

Al haber únicamente dos soluciones se trata de sucesos complementarios:

A la probabilidad de éxito se le denomina "p"

A la probabilidad de fracaso se le denomina "q"

Verificándose que: $p + q = 1$

La distribución de Bernouiili se aplica cuando se realiza una sola vez un experimento que tiene únicamente dos posibles resultados (éxito o fracaso), por lo que la variable sólo puede tomar dos valores: el 1 y el 0.

Ejemplo: lanzamiento de una moneda.

Binomial; La distribución binomial parte de la distribución de Bernouilli, se aplica cuando se realizan un número "n" de veces el experimento de Bernouiili, siendo cada ensayo independiente del anterior. La variable puede tomar valores entre:

0: si todos los experimentos han sido fracaso

n: si todos los experimentos han sido éxitos

Ejemplo: lanzar repetidamente una moneda.

Poisson; La distribución de Poisson parte de la distribución binomial.

Cuando en una distribución binomial se realiza el experimento un número "n" muy elevado de veces y la probabilidad de éxito "p" en cada ensayo es reducida, entonces se aplica el modelo de distribución de Poisson**.** Se tiene que cumplir que:

$$" p " < 0{,}10$$

$$" p * n " < 10$$

Ejemplo: cantidad de erratas por página en un libro.

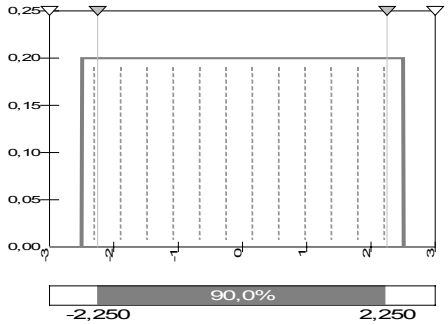
Distribuciones continuas:

Son que presentan un número infinito de posibles soluciones.

Tipos distribuciones:

Uniforme; es aquella que puede tomar cualquier valor dentro de un intervalo, todos ellos con la misma probabilidad.

Características:



• La totalidad de los posibles valores a tomar por la variable, situados entre las cantidades máximas y mínimas, presentan las mismas posibilidades de ser alcanzados
• El emprendedor identifica un rango de valor para las variables
• Variables exógenas
• Parámetros de carga identificables y cuantificables por el emprendedor.

Normal; Se utiliza para medir y representar multitud de variables como el peso, la altura, la calificación de un examen..., cuya distribución es simétrica con respecto a un valor central, alrededor del cual toma valores con gran probabilidad, sin existir apenas valores extremos.

Es el modelo de distribución más utilizado en la práctica. La importancia de la distribución normal se debe principalmente a que hay muchas variables asociadas a fenómenos naturales que siguen el modelo de la normal (tallas, pesos, envergaduras, consumo de cierto producto, puntuaciones de examen, grado de adaptación a un medio, etc.), multitud de fenómenos se comportan según una distribución normal.

Esta distribución de caracteriza porque los valores se distribuyen formando una **campana de Gauss**, en torno a un valor central que coincide con el valor medio de la distribución

Un 50% de los valores están a la derecha de este valor central y otro 50% a la izquierda.

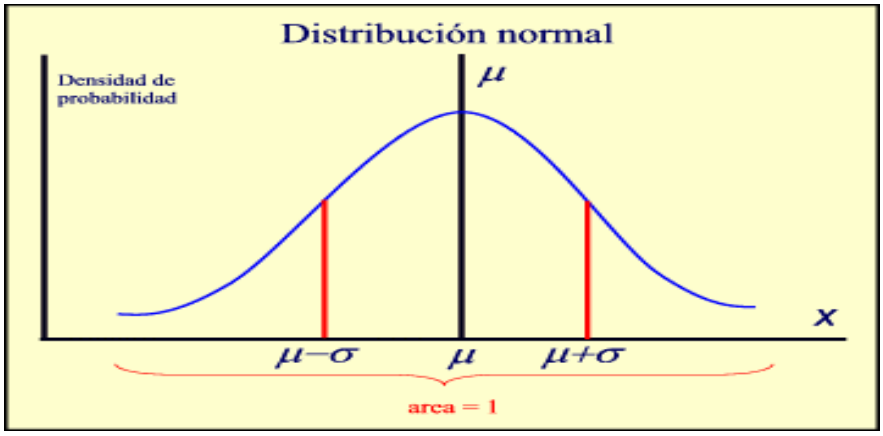
Esta distribución viene definida por dos parámetros:

X: N (□□□□ ²)

□□□□es el valor medio de la distribución y es precisamente donde se sitúa el centro de la curva (de la campana de Gauss).

□□² : es la varianza. Indica si los valores están más o menos alejados del valor central: si la varianza es baja los valores están próximos a la media; si es alta, entonces los valores están muy dispersos.

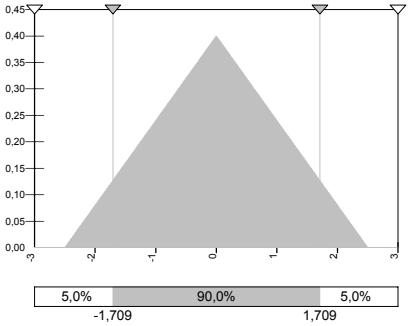
Cuando la media de la distribución es 0 y la varianza es 1se denomina "normal tipificada", y su ventaja reside en que hay tablas donde se recoge la probabilidad acumulada para cada punto de la curva de esta distribución.



Características:

- Mínimo predeterminado
- Máximo predeterminado
- Todos los valores entre el mínimo y el máximo de la distribución son igualmente probables.

Triangular; La distribución triangular es útil como una aproximación inicial en situaciones par las que no se dispone de datos confiables.
Nos permite estimar las duraciones de las actividades de un proyecto usando las tres estimaciones: optimista, muy pesimista, y pesimista.



Características:

• Función de distribución comúnmente aplicada a las variables de ventas y costes de mercado
• Variables endógenas, el emprendedor dispone de poder negociador sobre las mismas
• Parámetros de carga identificables y cuantificables por el emprendedor.

Ejemplo práctico:

Valor de la variable (X_i)	**Frecuencia absoluta**	**Frecuencia relativa**	**Frecuencia acumulada**
1.20	5	5/20 = 25%	5
1.7	4	4/20 = 20%	9
2.35	3	3/20 = 15%	12
2.01	7	7/20 = 35%	19
0.94	1	1/20 = 5%	20
Total	*20*	*100%*	

Medidas de posición central:

Media ⟹ $\bar{X} = \frac{(1.20*5)+(1.7*4)+(2.35*3)+(2.01*7)+(0.94*1)}{20} = 1.743$

Mediana ⟹ Ordenamos: 0.94 – 1.20 – 1.7 – 2.01 – 2.35
Med.= 1.7

Moda ⟹ Moda = 2.01 (es el valor que más se repite, teniendo en cuenta que tiene la mayor frecuencia)

Medidas de posición no central:

P_{75} = Tercer cuartil (Q_3)

Percentil ⟹ $P_{75} = \frac{3 * n}{4} = \frac{3 * 20}{4} = 15$

OObservando en la tabla las frecuencias acumuladas nos damos cuenta de que para X_i=2.01 dejamos por debajo el 75% de las observaciones y por encima el 25%.

Medidas de dispersión:

Varianza ⟹ $\square\square^2 = \frac{[(1.2-1.743)^2 * 5]+[(1.7-1.743)^2 * 4]+\ldots+[(0.94-1.743)^2 *1]}{20} = 0.586$

Desviación típica ⟹ □□□□0.7655

Rango ⟹ R= 2.35 – 0.94 = 1.41