Apunte teórico I – Probabilidad y estadística

Introducción a la estadística:

La estadística es el proceso matemático de recogida, descripción, organización, análisis e interpretación de datos.

Durante la cursada veremos la *estadística descriptiva*, esta tabula, representa y describe una serie de datos sin sacar conclusiones.

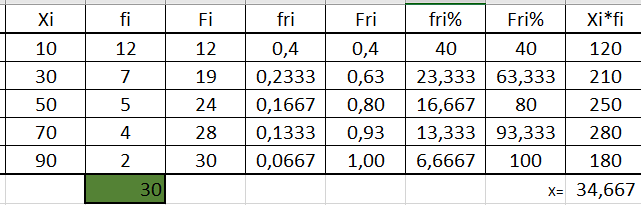
Estos datos serán cuantitativos, a su vez pueden ser discretos o continuos.

Si son discretos se pueden contar, serán valores enteros.

Si son continuos pueden encontrarse en un intervalo.

Para realizar esta interpretación realizaremos una ***tabla estadística***.

Un ejemplo de esta es la siguiente, la misma es en no intervalos:



Referencias:

Xi: variable de estudio

fi: frecuencia, cantidad de veces que aparece Xi

Fi: frecuencia acumulada, sumatoria de fi hasta el momento dado.

fri: frecuencia relativa, será Xi/n , siendo ***n*** la cantidad de veces que aparece la variable, esta es la sumatoria de fi.

Fri: frecuencia relativa acumulada, sumatoria de fri.

fri%: fri \* 100

Fri%: sumatoria de fri

Hasta acá acomodamos correctamente los datos básicos, sin embargo, se importante tener en cuenta que según la consigna quizás debamos hacer más operaciones, es por esto que nunca esta demás reservar un espacio para seguir agregando valores correspondientes casi siempre a ciertas multiplicaciones junto, o no, a otras operaciones que contienen una sumatoria, la cual será facilitada con la tabla.

***Cuando hablamos de una sumatoria debemos tener cuidado con discriminar valores.***

De esta manera tendremos resumida la información y podremos pasar a realizar ciertas operaciones.

Estás operaciones nos permitirán calcular ciertos datos y características de la muestra o población a analizar, algunas de estas son:

* Medidas de centralización y posición: Las medidas de centralización o de tendencia central expresan el valor en torno al cual se sitúan los datos de una muestra. Estos son:
* Media
* Mediana
* Moda

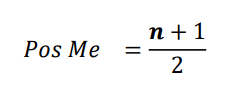
A su vez dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos en los ***cuantiles***:

* Percentiles
* Cuartiles
* Deciles
* Medidas de dispersión o variabilidad: Indica la mayor o menor concentración de los datos con respecto a las medidas de centralización. Estos son:
* Rango
* Varianza
* Desviación típica
* Rango intercuartílico y el coeficiente de variación

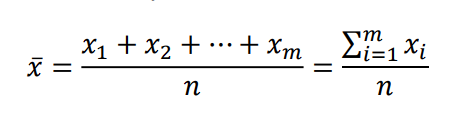
Medidas de centralización:

La **moda** nos habla del valor que más se repite, por ejemplo, si tenemos en la consigna, ¿cuál es el valor que más se repite? Debemos ir a *fi*, este es un ejemplo genérico, pero se repite en todas las ocasiones.

La **media** representa el valor de la variable de posición central en un conjunto de datos ordenados. Para esto debemos ir a *Fi* y ver cual es el valor en el que entra. Acá ya si que debemos usar formulas:



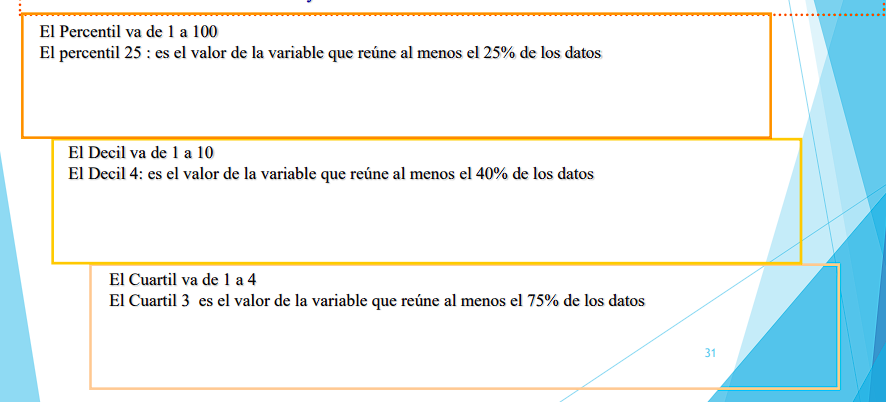
La **media aritmética** es el cociente entre la suma de todos los valores de la muestra y el número total de ellos:



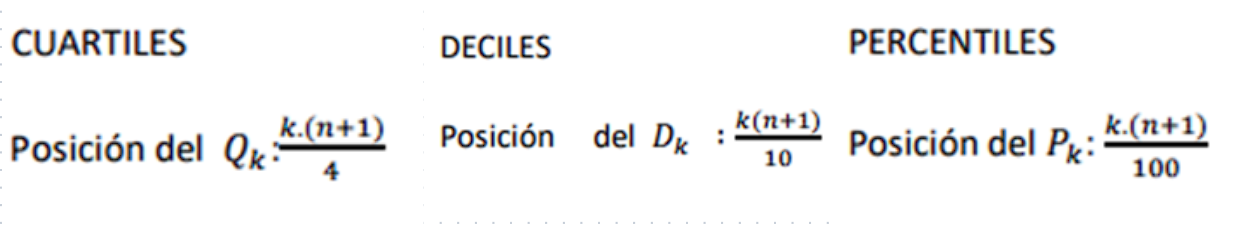
Donde *m* es la cantidad de valores distintos de la muestra.

También existen los **percentiles**, estos son medidas estadísticas que indican la posición relativa de un valor dentro de un conjunto de datos.

Existen los siguientes percentiles *( copiado textualmente, esta muy resumido en el PDF )* :



Las fórmulas son las siguientes:



Medidas de dispersión o variabilidad:

El **rango** se obtiene restando el valor más alto del más chico:

***X = Xmáx – Xmin***

Aunque no es del todo eficiente por cuestiones que no vienen al caso de este apunte.

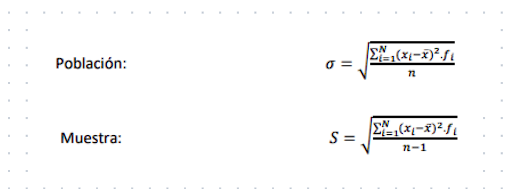
La **varianza** indica la dispersión o variabilidad de un conjunto de datos

respecto a su media. Su formula es una sumatoria algo complicada, no tanto por formula como tal que simplemente se copia sino por su complejidad a la hora de discriminar valores, si simplemente escribimos la sumatoria tal como elemento1 + elemento2 + …. + elementoN nos podremos equivocar muy fácilmente, ***debemos hacer esta cuenta tranquilos y con la tabla***.

**

Durante la clase ya vimos ejemplos, verlos y rehacerlos para practicar.

El **desvió estándar** es la raíz cuadrada de la varianza, mide la dispersión de los datos en las mismas unidades que los datos originales:

**

El **coeficiente de variación** es la relación entre el tamaño de la media y la variabilidad de la variable.

Recordemos que una *relación* es la manera en que dos o más variable están conectadas o afectan entre sí. *(Muy resumidamente, no viene al caso tampoco, simplemente viene bien saberlo para no dejar en la nada un concepto tan importante)*.

A mayor valor de *Cv*mayor heterogeneidad de los valores de la variable, a menor *Cv* menor heterogeneidad.

Se recomienda que no sea mayor a 0,2 (20%).

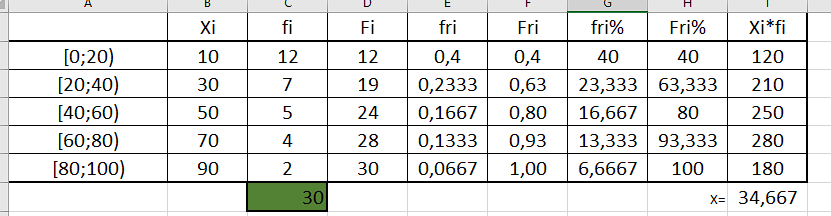
Entonces, si *Cv ≤ 0,2* los datos son homogéneos.

Datos agrupados en conjuntos

Los datos a su vez pueden estar agrupados en conjuntos, es por esto que debemos también saber como aplicar gran parte de los conceptos y formulas anteriormente mencionadas en este tipo de ejercicios, es muy importante interpretar la consigna.

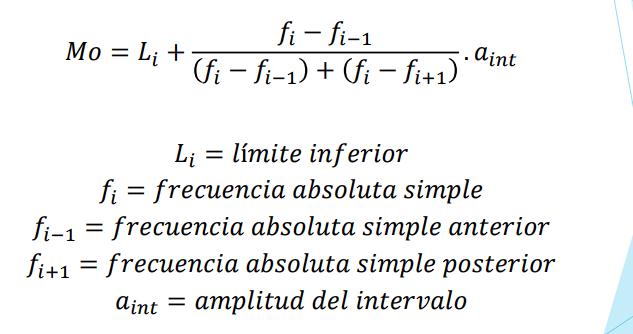
La tabla inicialmente será igual, lo que cambiará es que Xi será el resultado del ***(Xmin – Xmax) / 2*** , teniendo el intervalo expresado como *[Xmin;Xmax)*.

Por ejemplo:



Las medidas de centralización y posición serán las mismas, al igual que las de dispersión y variabilidad, la diferencia estará en las formulas y la interpretación de los ejercicios, para esto es muy importante no solo hacer los mimos mecánicamente sino que también entenderlos, pensarlos no como un ejercicio de la facultad, sino como un problema más que se nos presenta en el día a día y tendremos que saber como solucionarlo.

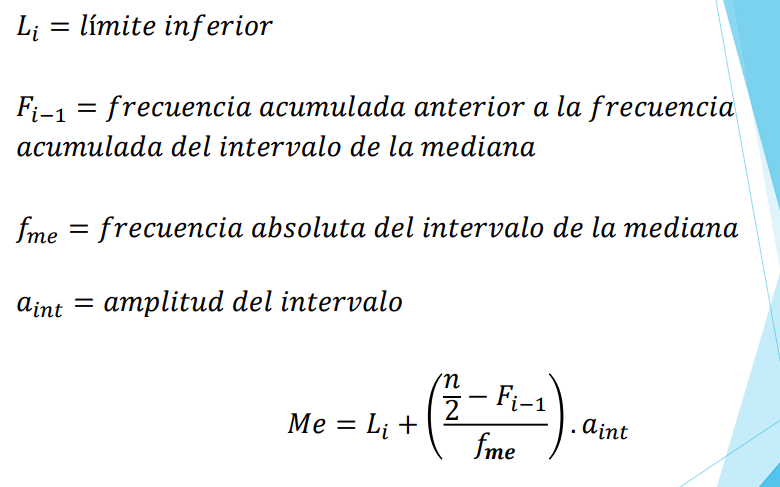
La **moda** se calculará de la siguiente manera:



Los valores a tomar corresponden al intervalo con mayor *fi*.

Prestar mucha atención a las referencias de *fi-1*y lo demás.

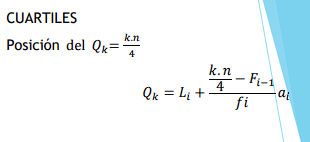
La **media** se calculará de la siguiente manera:

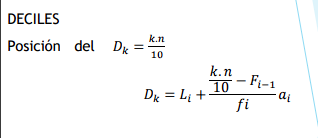


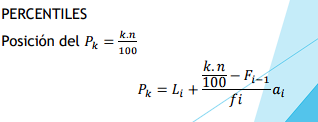
n/2 = Pme , una vez calculado esto buscamos en la frecuencia acumulada 𝐹𝑖 el intervalo al cual pertenece dicha frecuencia. Esta será la que tenga dicho valor más cerca. Una vez tengamos esto bastará con seguir la formula.

La **media aritmética** se calculará de la misma manera.

Los cuartiles, deciles y percentiles representarán lo mismo, pero se calcularán de una manera diferente:







Gráficos

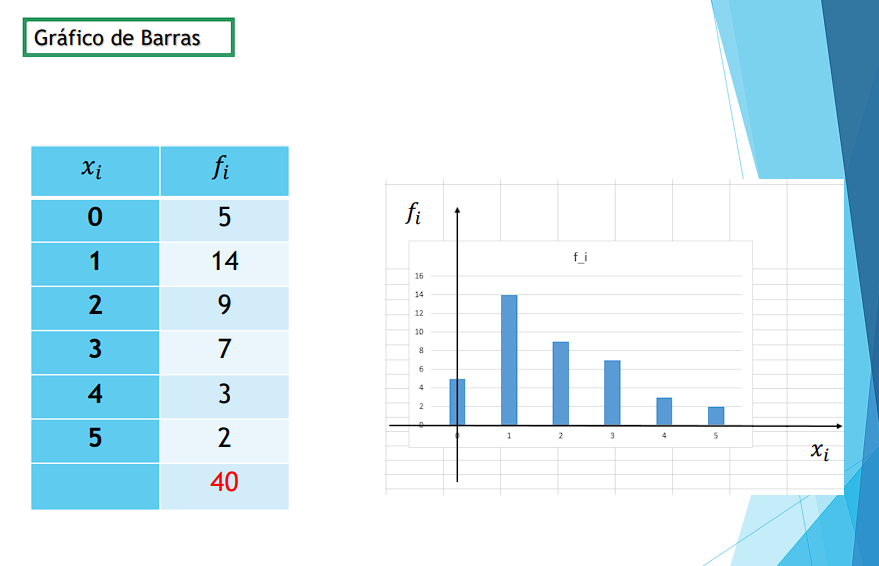
Una vez tengamos ciertos datos podremos realizar gráficos con los mismos, algunos son más tediosos que otros, sin embargo, todos en caso de estar bien hechos serán consistentes y nos dirán mucho o poco según el tipo y nos permitirán tener los datos de una forma más visualmente agradable a la vista.

Los gráficos que vamos a ver pueden ser:

* Gráfico de barras
* Gráfico de frecuencias acumuladas
* Histograma
* Polígono de frecuencia
* Diagrama de caja *( El más complejo )*

El más complejo y útil, independientemente de su dificultad para realizar el mismo, es el *diagrama de caja,* realizaré una pequeña pero útil demostración de cómo se hace cada uno de estos gráficos y por último ira el diagrama de caja, sin embargo, este será el más complicado.

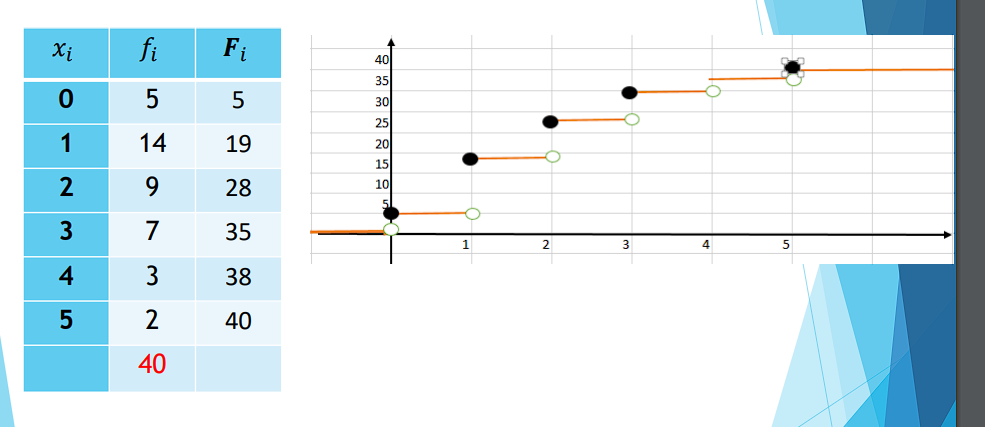
Gráfico de barras



El gráfico de barras se hará tomando de menor o mayor los valores de Xi para luego realizar una barra con la frecuencia del mismo.

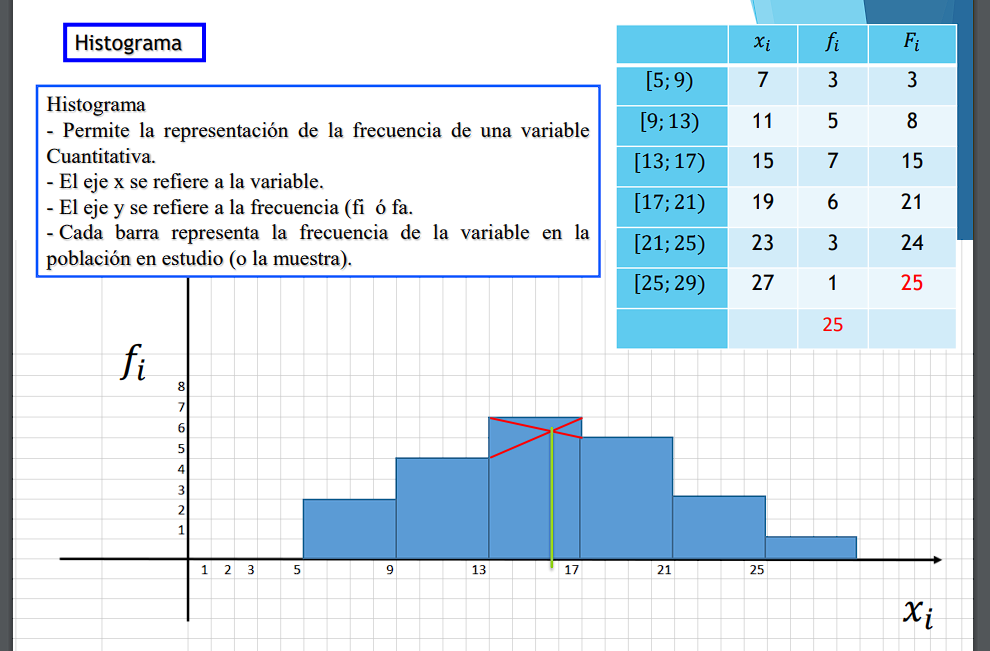
Los valores de *Xi* irán en el eje *x* y los de *fi* en el *y*.

Gráfico de frecuencia acumulada



El gráfico de frecuencia acumulada se hará tomando los valores de *Xi* en eje x, pero en el eje *y* la frecuencia acumulada *Fi*, debemos tener en cuenta que la distancia entre los puntos del eje x correspondientes a las variables será del tamaño de *Fi*.

Histograma



Polígono de frecuencia

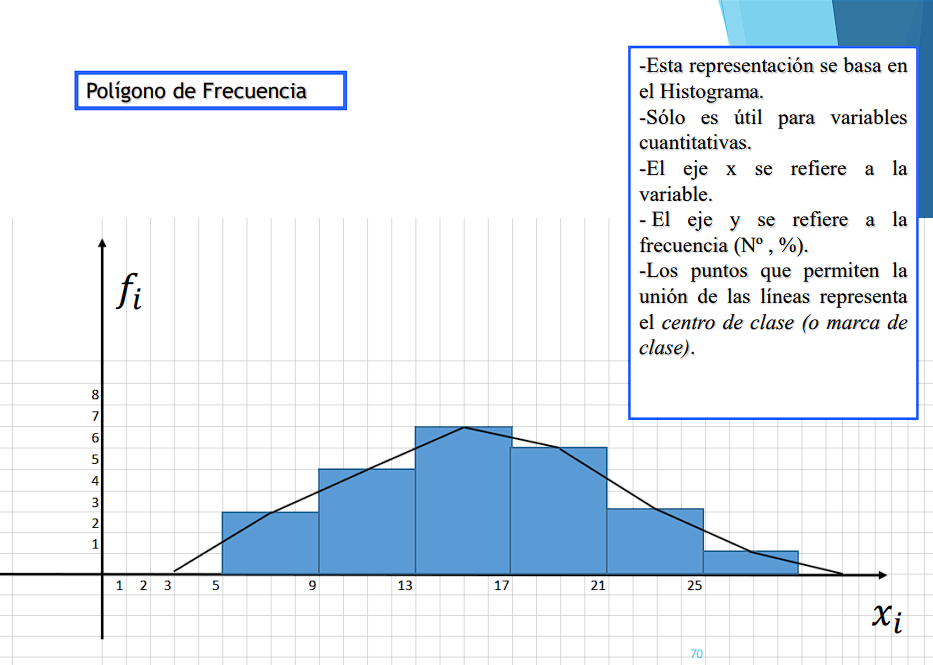


Diagrama de caja



El *Diagrama de Caja* se representa sobre el eje de la variable en análisis, construyéndose un rectángulo o “caja” con prolongaciones o “bigotes” teniendo en cuenta los siguientes valores:

* **Q1:** el Cuartilo de orden 1 marca el límite inferior del rectángulo.
* **Q3:** el Cuartilo de orden 3 marca el límite superior del rectángulo.
* **Me = Q2:** la Mediana se marca con una línea dentro del rectángulo.
* **Un valor mínimo:** que es el mayor entre el xMín (menor valor de la variable) y 1,5 veces el Q (Rango intercuartílico), restado al Q1. Desde el límite inferior del rectángulo, se extiende una línea o “bigote” hasta este valor mínimo.
* **Un valor máximo:** que es el menor entre el *xMáx* (mayor valor de la variable) y 1,5 veces el Q (Rango intercuartílico), sumado al Q3 . Desde el límite superior del rectángulo, se extiende una línea o “bigote” hasta este valor máximo. El *xMín* y/o el *xMáx* si es que no quedan comprendidos dentro del Diagrama del rectángulo con los bigotes, constituyen valores anómalos.