ESTADÍSTICA DESCRIPTIVA

La Estadística descriptiva se ocupa de los métodos y procedimientos para recolectar, organizar, resumir, analizar, interpretar y presentar la información. *Es decir, se reduce el conjunto de datos obtenidos por un pequeño número de valores descriptivos, como pueden ser: el promedio, la mediana, la moda, la varianza, la desviación típica, etc. Estas medidas descriptivas pueden ayudar a brindar las principales propiedades de los datos observados, así como las características clave de los fenómenos estudiados.*

CONCEPTOS BÁSICOS Y PROPIEDADES:

Población: conjunto de individuos o elementos que cumplen ciertas propiedades comunes y qué será objeto de estudio estadístico.

En relación al tamaño de la población, ésta puede ser:

- **Finita**, como es el caso del número de personas que llegan al servicio de urgencia de un hospital en un día;
- **Infinita.** En Estadística, población infinita se refiere a una población con un número tan grande de elementos que no le es posible al investigador someter a medida cada uno de ellos. Por ejemplo: número de estrellas en el universo, que se extinguirán en los próximos cien millones de años, número de gotas de agua que están presentes en el océano Atlántico.

Muestra:

Es un subconjunto representativo de la población, que se toma para realizar los estudios respectivos a partir de los cuales se sacan conclusiones generales válidas para la población.

¿Por qué estudiar una muestra y no a toda la población?

- Por costos no se puede llegar a todos los elementos de interés.
- Si la población es extremadamente grande (infinita) se debe estudiar una muestra.
- Por tiempo, la información se precisa en un momento determinado.
- Imposibilidad de acceso a toda la población.
- En ocasiones, el estudio de la población implica la destrucción de los elementos. Ej.: fabrica de fósforos.
- Si la muestra es representativa, permite disminuir errores.

Para obtener una muestra representativa se suele recurrir a criterios de aleatoriedad proporcionalidad.

- Una muestra es **aleatoria** cuando sus elementos se escogen al azar, mediante algún tipo de sorteo. Dentro de éstas existen dos tipos de muestras: muestra aleatorio simple y muestra aleatorio sistemático. La muestra **aleatorio simple:** consiste en elegir al azar los n individuos de una población, en el que cada uno tiene la misma probabilidad de salir (equiprobables).
  La muestra **aleatorio sistemático** consta de tomar cada k-ésima unidad de la población, una vez que las unidades de muestreo están numeradas o arregladas de alguna forma. La letra k es la razón de muestreo, esto es, la razón del tamaño de la población correspondiente al tamaño de la muestra. Ejemplo: al seleccionar una muestra de 40 unidades de entre una población de 1, 000, entonces k es 1000/40= 25, por lo que la muestra se obtiene tomando cada 25- ésima unidad de la población.
- Una muestra es **proporcional** cuando cada parte de la población está representada de acuerdo con su importancia en ella. Por ejemplo: si queremos elegir 5 alumnos entre los 36 de una clase, en la que dicha población está dividida en 21 chicas y 15 chicos; debemos obtener de

forma proporcional el tamaño de cada submuestra. Por lo que se realiza lo siguiente:
$\frac{21}{36} = \frac{n_1}{5} \Rightarrow n_1 \cong 2{,}92$ $\frac{15}{36} = \frac{n_2}{5} \Rightarrow n_2 \cong 2{,}08$
Esto nos indica que la muestra se debe de componer de 3 chicas y 2 chicos.

Variable estadística: es una característica común que tienen todos los individuos o elementos de una población y que va a ser objeto de estudio estadístico. Por ejemplo, en este grupo de estudiantes son variables estadísticas la edad, estatura, color de ojos, sexo, preferencias musicales, etc…
Dato: Cada valor observado de la variable.

Modalidades de una variable: diferentes situaciones posibles de una variable. Ejemplo: color de ojos: marrones, azules, verdes, otros.
Las modalidades deben ser a la vez ***exhaustivas*** (todo elemento tiene que pertenecer a alguna categoría) y ***excluyentes*** (cada elemento tiene que pertenecer a una sola categoría) de las modalidades posibles.
Por ejemplo, si en la variable: "lugar de residencia" se omite medio rural no es exhaustiva la clasificación.
"Otros" es una forma de hacer exhaustivas las categorías.
Lugar de residencia podría no ser excluyente, pues puede declarar vivir en Montevideo y en una ciudad del interior. Entonces se aclara que lugar de residencia se considera el lugar donde esté mas días en la semana.

Las variables estadísticas se clasifican del siguiente modo:

Variables cualitativas: son aquellas que únicamente pueden describirse, es decir, asociadas a características no cuantificables. Por ej.: sexo, lugar de residencia, …
Los diferentes estados en que se puede presentar una variable cualitativa se denominan modalidades, categorías o atributos.
Al conjunto de modalidades que puede tomar una variable cualitativa se le llama recorrido.
Por ejemplo, el grupo sanguíneo tiene por modalidades:
Grupos Sanguíneos posibles: A, B, AB, O.
A su vez las variables categóricas se clasifican en:
Variables cualitativas nominales: son aquellas donde las asignaciones no suponen ningún orden. Ej. En tipo de profesión, cualquier asignación es válida.
Variables cualitativas ordinales: Son aquellas en las que las asignaciones corresponden a un orden de preferencias.
Ejemplo: en una encuesta de opinión: las respuestas: muy mal, regular, bien, muy bien; a una pregunta determinada permiten asociar los números 1,2,3,4,5 cuyo orden sí supone una característica importante, a mayor número, mayor satisfacción.

Variables cuantitativas o medibles: son aquellas que los datos se miden en base a cantidades, la característica que poseen es susceptible de ser medida y obtener, por lo tanto, una cantidad. Ej.: edad de un individuo, ingreso de dinero en una casa, …
Éstas se pueden subdividir a su vez en:
Variables discretas: Aquellas que entre dos valores próximos puede tomar un número finito o a lo sumo numerable de valores. Ej. Nro. de obreros de una fábrica, el de alumnos de la universidad, el número de lanzamientos de una moneda hasta obtener cara puede tomar cualquier valor natural.
Variables continuas: Las que admiten modalidades intermedias, es decir, puede haber infinitos valores entre dos de éstas. Ejemplos, peso, estatura, distancias, etc.
El conjunto de valores que toma una variable cuantitativa se denomina ***recorrido***.

ORGANIZACIÓN Y TABULACIÓN DE LOS DATOS: TABLA DE DISTRIBUCIÓN DE FRECUENCIAS Y PORCENTAJE.

La tabla de distribución de frecuencias es una tabla donde los datos estadísticos aparecen bien organizados, distribuidos según su frecuencia, es decir, según las veces que se repite en la muestra. En esta tabla se representan los diferentes tipos de frecuencias, ordenados en columnas.

TIPOS DE FRECUENCIAS:

Frecuencia Absoluta: (Notación: n_i)

Es el número de veces que tiene lugar la observación de un determina fenómeno.
La suma de las frecuencias absolutas es igual al número total de observaciones. (N)

Frecuencia Relativa: (Notación: f_i)

Con respecto a un fenómeno, es la proporción que representa su frecuencia absoluta con respecto al número total de observaciones.

$$f_i = \frac{n_i}{N}$$

Observaciones:

1) La suma de las frecuencias relativas da 1. Es decir: $\sum f_i = 1$.

2) Como la frecuencia relativa es la proporción que representa su frecuencia absoluta con respecto al total de observaciones y la definición de probabilidad clásica de un suceso expresa que es la proporción que representan los casos favorables a que ocurra el suceso con respecto al total de casos posibles, entonces:

$f_i = p_i$, siendo p_i la probabilidad del suceso "i".

Entonces, $0 \leq \sum f_i \leq 1$ o lo que es lo mismo, $0 \leq \sum p_i \leq 1$.

3) ***Porcentaje***: se obtiene multiplicando por 100 la frecuencia relativa de cada modalidad.

Frecuencias acumuladas:

Generalmente en una tabla de distribución de frecuencias no sólo se muestran las frecuencias absolutas y relativas, sino que también se incluyen las **frecuencias acumuladas absolutas (N_i)** y **relativas (F_i).**

Frecuencia absoluta acumulada: es el resultado de ir sumando las frecuencias absolutas.

Frecuencia relativa acumulada: es el resultado de ir sumando las frecuencias relativas.

A continuación, se trabajarán en un ejercicio para ver esta terminología aplicada a un ejemplo:

Ejercicio 1):

Un experimento consistió en contar el número de flores por planta de una muestra de 50 plantas. Los valores resultantes del conteo fueron los siguientes: 10, 8, 6, 3, 9, 7, 5, 4, 6, 9, 8, 10, 7, 9, 10, 6, 8, 6, 3, 2, 4, 3, 2, 7, 5, 5, 4, 3, 7, 6, 6, 7, 8, 8, 6, 7, 7, 9, 8, 6, 5, 3, 2, 1, 4, 3, 6, 8, 7, 0.
Los datos así presentados son de difícil interpretación, por lo que conviene resumirlos como en la siguiente tabla de distribución de frecuencias para la variable "número de flores por planta":

x_i	n_i	N_i	f_i	F_i	%
0					
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

a) Especifica cuál es la muestra, y la variable de estudio.
b) Indica que tipo de variable es.
c) Identifica la columna que se refiere a cantidad de flores.
d) Identifica la columna que se refiere a cantidad de plantas.
e) ¿Qué valor o valores de la variable tuvo mayor frecuencia absoluta?
f) ¿Qué cantidad de plantas tienen exactamente 6 flores?
g) ¿Qué cantidad de plantas poseen menos de 3 flores?
h) ¿Qué porcentaje de plantas no tiene flores?
i) ¿Qué cantidad de plantas contienen más de 6 flores?

Estas preguntas, como algunas otras, pueden responderse fácilmente a partir de la lectura de una tabla de distribución de frecuencias, con los datos ya procesados.
Además, se pueden visualizar los datos en gráficos, que pueden ser de barra o circular.

Gráficos de barras:

Para su construcción se necesita un par de ejes cartesianos. El eje de abscisas se asigna a la categoría o valores de la variable, según corresponda, el eje de ordenadas a la escala de frecuencias que pueden ser absolutas o relativas.

Ejercicio 2):

Construye un gráfico de barras en base al ejercicio 1). Puedes utilizar para el eje "y", la frecuencia absoluta, o relativa o porcentaje.

Gráficos circulares:

Los sectores circulares se construyen representando cada categoría o valor de la variable estudiada mediante una parte del círculo. El ángulo correspondiente a cada categoría se calcula mediante una regla de tres simple, o lo que es lo mismo, el sector "x" correspondiente a cada categoría se obtiene multiplicando 360° por su frecuencia relativa, es decir: $x_i = 360°.f_i$

Ejercicio 3):

Un docente de 6to año de Ciencias Biológicas del Liceo N° 4, quiere observar el sexo de sus estudiantes y está interesado en saber las preferencias de sus estudios posteriores al liceo, en la población femenina. Dicha clase consta de 40 estudiantes; de los cuales el 50% es masculino y el resto femenino. De la población femenina se sabe qué; un 55% prefieren continuar los estudios de enfermería, un 15% odontología, un 20% medicina y un 10% bioquímica.
a) ¿Cuál es la población de estudio?
b) ¿Qué variables intervienen en el estudio realizado?
c) Clasifica dichas variables.

d) Completa la siguiente tabla, teniendo en cuenta que la variable x es: "preferencias de las chicas para sus estudios posteriores al Liceo".

x_i	n_i	f_i	%
Enfermería			
Odontología			
Medicina			
Bioquímica			

e) Construye dos gráficos circulares. Uno que muestre como se distribuyen los estudiantes según el sexo. Y el otro gráfico que muestre, de la población femenina, sus preferencias de estudios posteriores al Liceo.
f) Menciona características de la población estudiada que se puedan deducir de los gráficos presentados.

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS PARA DATOS AGRUPADOS:

Tabla de frecuencias para variable continua: recorrido, intervalo, amplitud, marca de clase.

Cuando nos encontramos con una distribución con un gran número de datos, se suelen agrupar en intervalos para facilitar la comprensión de los datos. Esta práctica tiene en cambio un inconveniente: se pierde información sobre la propia distribución.

INTERVALO DE NÚMEROS REALES:

Intervalo Cerrado [a ; b] es el conjunto de números reales formado por a, b y todos los números comprendidos entre a y b, siendo a ≤ b. En símbolos: $[a;b]=\{x \in lR / a \le x \le b\}$

Se representa en la recta real así:

Intervalo Abierto $(a;b)=\{x \in lR / a < x < b\}$

Intervalo Semiabierto a derecha o Semicerrado a izquierda

$(a;b]=\{x \in lR / a < x \le b\}$

Intervalo Semicerrado a izquierda o Semiabierto a derecha

$[a;b)=\{x \in lR / a \le x < b\}$

NOTACIÓN DE INTERVALO UTILIZADO EN LA TABLA: $[L_{i-1}, L_i)$

Se indica por L_{i-1} al extremo inferior del intervalo y por L_i al extremo superior. Cerramos el intervalo por la izquierda y abrimos por la derecha. Es una manera de organizarse, pudiendo ser al contrario.

MARCA DE CLASE: el punto medio de cada intervalo y es el valor que representa al intervalo para el cálculo de parámetros. Para calcularla: $c_i = \frac{L_{i-1}+L_i}{2}$.

La **amplitud del intervalo**, es la longitud del intervalo, se representa por:

$$a = L_i - L_{i-1}$$

¿Cómo obtener, a partir de los datos, una tabla de frecuencias agrupada?

- **Nº de intervalos:** A partir de la raíz cuadrada del número de datos, decidimos, redondeando el número de intervalos (se aproxima hacia arriba).
- **Recorrido**: Valor mayor, menos valor menor de los datos. Re= x_n-x_1
- **Amplitud**: División entre el Recorrido y el número de intervalos que hayamos decidido. Se puede redondear también. $a_i = \frac{\text{Re}}{N^o \text{ de intervalos}}$

Ejercicio 4): En CAMEDUR, en enero 2020, se realizó un estudio sobre: "peso de los recién nacidos en el centro de salud". Para ello, se recogieron los datos de 40 bebes obteniendo los siguientes datos:

3.2	3.7	4.2	4.6	3.7	3.0	2.9	3.1	3.0	4.5
4.1	3.8	3.9	3.6	3.2	3.5	3.0	2.5	2.7	2.8
3.0	4.0	4.5	3.5	3.5	3.6	2.9	3.2	4.2	4.3
4.1	4.6	4.2	4.5	4.3	3.2	3.7	2.9	3.1	3.5

a) Construye la tabla de distribución de frecuencias y porcentaje, teniendo en cuenta cómo calcular el número de intervalos de la tabla, la amplitud de cada intervalo, y la marca de clase de cada intervalo.
b) Si sabemos que los bebes que pesan menos de 3,1 kilos nacen prematuramente ¿Qué porcentaje de niños prematuros han nacido entre estos 40?
c) Normalmente los niños que pesan más de 3,4 kilos no necesitan estar en la incubadora, ¿qué porcentaje de niños están en esta situación?
Completa la tabla de frecuencias:

Intervalos	c_i	n_i	N_i	f_i	F_i	%
[.....,.....)						
Total						

REPRESENTACIÓN GRÁFICA: HISTOGRAMA

Un histograma es una representación gráfica de una variable continua, en forma de barras agrupadas, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. En el eje vertical se representan las frecuencias (absolutas o relativas), y en el eje horizontal los valores de la variable, normalmente señalando las marcas de clase, es decir, la mitad del intervalo en el que están agrupados los datos.
Intervalos de clase: Cuando se tiene un conjunto de datos continuo cabe la posibilidad de que los intervalos de clase tengan igual o diferente longitud.

Ejercicio 5):
Construye un histograma en base al ejercicio 4).

POLÍGONO DE FRECUENCIAS:

Un polígono de frecuencias es una representación gráfica lineal que se obtiene a partir de un histograma de frecuencias. Es la gráfica que se obtiene al unir en forma consecutiva con segmentos los puntos medios de cada barra del histograma, incluyendo el punto medio anterior a la primera clase y el punto medio posterior a la última clase, en el eje $\overrightarrow{Ox}$.

OJIVA:

Una representación gráfica para representar las frecuencias acumuladas es la ojiva. Se trata de una gráfica poligonal en la cual cada punto representa el límite superior de la clase marcada en el eje horizontal y la frecuencia acumulada en el eje vertical, después se une cada par de puntos consecutivos con un segmento de recta.

Ejercicio: Grafica la ojiva de las frecuencias acumuladas correspondiente al ejemplo anterior.

Ejercicio 6):

Construye la ojiva en base al ejercicio 4).