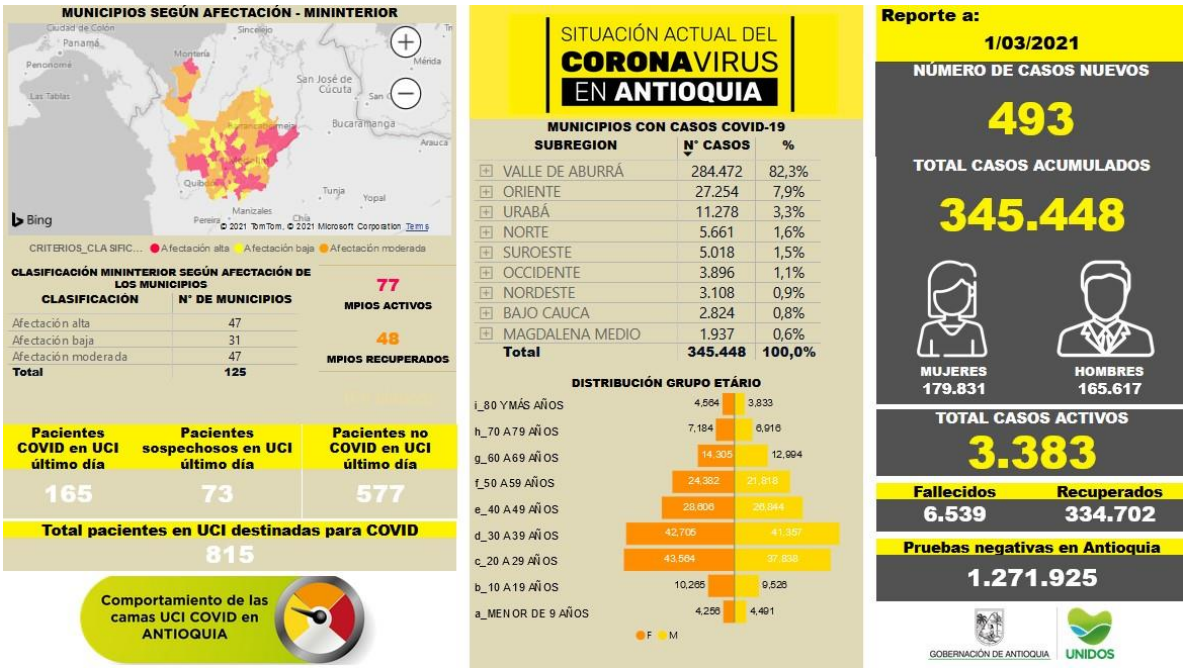
Resumen módulo 2 MOOC Ciencia e ingeniería de datos:

Introducción:

Podemos definir la estadística como una ciencia que se ocupa de la recopilación, análisis, interpretación y presentación de datos. Vemos y usamos datos en nuestra vida diaria.

Por ejemplo, en el siguiente enlace se muestra el avance de los casos semana a semana, por país, del COVID 19 a nivel mundial semana a semana hasta noviembre del 2020. Es un ejemplo de cómo la representación gráfica de unos datos de contagios puede permitirnos hacer comparaciones de su evolución temporal por cada país, muy práctico para que tengamos un panorama general del estado a nivel mundial de esta pandemia. Entonces podemos ver como las herramientas estadísticas de visualización de información nos pueden permitir organizar datos, mostrarlos de manera general, comparar diferentes países en este caso, y tomar decisiones de entidades como la organización mundial de la salud.



Gráfica 1. Cifras estadísticas del COVID 19 al día 1 de marzo de 2021. Fuente: Gobernación de Antioquia.

Tomando un ejemplo más cercano, en la gráfica 1 podemos ver diferentes cifras estadísticas del COVID 19 en Antioquia, actualizado a inicios del mes de marzo del 2021. En el infográfico podemos ver segmentado por colores los municipios de Antioquia con mayor afectación, información que permite a entidades gubernamentales como la gobernación de Antioquia priorizar acciones teniendo mayores argumentos, por ejemplo, destinar mayores recursos y personal a municipios con mayor afectación. En la gráfica también podemos ver cifras relacionadas al uso y ocupación de camas UCI, información que ayudó a tomar muchas medidas para tratar de reducir la cantidad de personas en cuidados intensivos con diagnósticos no COVID y destinarlas para tratar a personas infectadas, también podemos ver cifras distribuidas por subregiones y en porcentajes, datos de contagiados por

rango de edad y género lo que ayuda a tener un panorama general de la evolución de contagios y ayuda a las instituciones públicas a tomar mejores decisiones.

En estadística, generalmente queremos estudiar una población. Podemos pensar en una población como una colección de personas, cosas u objetos en estudio. Para estudiar la población, seleccionamos una muestra. La idea del muestreo es seleccionar una porción (o subconjunto) de la población más grande y estudiar esa porción (la muestra) para obtener información sobre la población. Los datos son el resultado del muestreo de una población. Por ejemplo en un proceso democrático, sean las elecciones presidenciales, de congreso y asamblea, de alcaldes o gobernadores vemos que en el país se realizan muchas encuestas de opinión que se toman muestras de entre 1.000 y 2.000 personas, en estos ejercicios se supone que la encuesta de opinión representa las opiniones de la gente en todo el país, por tanto el proceso de selección de muestra o muestreo debe ser imparcial y aleatorio para que los resultados no estén sesgados y la muestra represente en realidad la población, en este ejemplo la población son todas las personas en todo el país y la muestra son por lo general de 1000 a 2000 personas. Otro ejemplo es si se desea calcular el promedio general de calificaciones en su colegio, tendría sentido seleccionar una muestra de estudiantes que asisten a la escuela. Los datos recopilados de la muestra serían los promedios de calificaciones de los estudiantes. En este caso la población son todos los estudiantes de tu colegio y la muestra sería seleccionar 50 o 100 estudiantes.

Cuando primero se nos presenta un conjunto de mediciones, ya sea una muestra o una población, necesita encontrar una forma de organizarlo y resumirlo. La rama de la estadística que presenta técnicas para describir conjuntos de mediciones se denomina estadística descriptiva. Nos hemos enfrentado con la estadística descriptiva en muchas ocasiones en forma de: gráficas de barras, gráficas de pastel y gráficas de líneas presentadas por un candidato político; tablas numéricas en el periódico; o el promedio de cantidad de lluvia informado por el Sistema de Alerta Temprana del valle de Aburrá SIATA. Las gráficas y resúmenes numéricos son comunes en día a día. Supongamos que deseamos estudiar la altura promedio de los estudiantes en un salón de clase, en estadística descriptiva registraría las alturas de todos los estudiantes de la clase y de ella se sacarían algunos datos importantes como el promedio, el máximo, el mínimo, la moda, la mediana, la desviación estándar, etc. Que nos permiten describir como se comportan los datos.

Por otro lado, tenemos la estadística inferencial que aplica la probabilidad para llegar a una conclusión. Le permite inferir parámetros de la población basados en estadísticas descriptivas de una muestra y construir modelos a partir de ellos. Su papel es interpretar, hacer proyecciones y comparaciones.

En conclusión, la estadística descriptiva pretende describir los datos, y la inferencial utilizando los datos descriptivos se encarga de hacer proyecciones, predicciones y comparaciones con otros datos.

Variables y tipos de variables:

En estadística, una variable tiene tres características muy bien definidas:

- Una variable es un atributo que describe a una persona, lugar, cosa o idea.
- Su valor se puede representar por un dato.
- El valor de la variable puede "variar" de una entidad a otra.

Por ejemplo, el color de pelo de una persona es un buen prospecto de variable ya que describe un atributo como el color de pelo de alguien, su valor se puede describir mediante un dato, por ejemplo, el pelo puede ser “mono”, “negro”, “castaño”, “rojo”, y puede variar de una persona u otra, ya que para una persona puede ser “rojo” y para otra “castaño”. Otros tipos de ejemplos son el género (masculino, femenino, o no determinado), las calificaciones finales de una materia (Excelente, sobresaliente, insuficiente), el peso en kilogramos de una persona (por ejemplo, yo peso 57 kg) o el número de hermanos de alguien, cada una de estas variables pertenecen a dos tipos diferentes.

Las variables se pueden clasificar en dos grandes ramas, la primera son las que se conocen como variables categóricas o cualitativas (ya que representar con cualidades) o cuantitativas (ya que sus valores se pueden representar con números).

- Las variables cualitativas: Toman valores que son nombres o etiquetas. El color de una pelota (por ejemplo, rojo, verde, azul) o la raza de un perro (por ejemplo, ´pastor alemán, pitbull, labrador, french poodle, criollito). Las variables cualitativas se dividen también en dos:
    - Variable cualitativa nominal: Se representa con datos no numéricos, pero no tienen un orden especifico. Ejemplo: El estado civil, con las siguientes modalidades: soltero, casado, separado, divorciado y viudo, ninguna es más importante que otra.
    - Variable cualitativa ordinal: Se representan con datos no numéricos, pero existe un orden. Ejemplo: La nota en un examen: Excelente, Aceptable, Insuficiente, o el puesto conseguido en una prueba deportiva: primero, segundo, tercer o medallas de una prueba deportiva: oro, plata, bronce.
- Las variables cuantitativas: Toman valores numéricos. pueden distinguirse entre discretas y continuas:
    - Discreta: Se dice que una variable es discreta cuando no puede tomar ningún valor entre dos números consecutivos, es decir toma sólo uno de una lista de valores posibles, por ejemplo, el número de hijos, puedes tener 0, 1, 2, 3, etc. pero no podemos decir que tenemos medio hijo y si lo dices eres muy cruel, porque aun siendo un hijo de baja estatura (como yo) sigue contando como uno. En otras palabras, las variables discretas pueden tomar valores numéricos específicos. Ejemplo el número de empleados de una fábrica; número de hermanos; número de árboles en un parque, numero de mascotas, etc.
    - Continua: Se dice que una variable es continua puede tomar cualquier valor dentro de un intervalo. Algunos ejemplos son: temperaturas registradas en un observatorio (puede tomar valores de 24,2 °C o 31 °C o cualquier otro valor en un rango de temperatura), el tiempo en recorrer una distancia en una carrera (10,5 segundos, 75 segundos, etc.), la estatura (puede ser 1.65 m o 1.82 m etc.)

Medidas de tendencia central:

Las medidas de tendencia central son un conjunto de medidas que se usan en estadística descriptiva para ayudarnos a conocer un poco el comportamiento de un conjunto de datos, y nos va a facilitar su comprensión.

Veamos un ejemplo. Supongamos que estas son las edades de 5 estudiantes de un salón: 16, 16, 15, 19, 17. Vamos a organizarla en una tabla de esta manera. Primero colocamos los valores de la edad que nos aparecieron, por ejemplo, vimos que apareció el 15, el 16, el 17 y el 19. Ahora pongamos las veces que se repite cada uno de estos datos: el 15 se repitió 1 vez, el 16 se repitió 2 veces, el 17 se repite 1 veces, y el 19 se repite 1 vez.

Lo primero que vamos a calcular es el promedio o la media de los datos, para esto usamos una formula muy sencilla, es sumar todos los datos y dividirlo por el número total de datos, entonces tenemos que el promedio de los datos es de 16, 6 años.

La moda se define como el valor que más se repite, por lo tanto, decimos que la edad de 16 años es la moda de los datos.

La mediana es el valor que me parte los datos en dos grupos iguales, para sacarla debemos organizar los datos de mayor a menor por nos quedaría 15, 16, 16, 17, 19. En este caso como el número de datos es impar es muy fácil ver la mediana, el valor de la mitad (16). Si tuviéramos por ejemplo otra edad de 20 años, la mediana la sacamos como el promedio de los dos datos que hay en el medio.

La desviación y la varianza son pequeños significa que los datos están muy cercanos entre ellos, de lo contrario si son grande significa que los datos están muy dispersos.

Recurso para videojuego de práctica evaluativa.

Tabla 1. Valores de estudio para cada uno de los casos, se debe calcular la varianza según la fecha propuesta en el videojuego.

	Fecha				
Subregión	19/04/2021	07/06/2021	11/06/2021	30/06/2021	01/07/2021
VALLE DE ABURRÁ	362.772	460.091	469.445	518.411	520.777
ORIENTE	34.834	44.394	45.519	52.883	53.330

URABÁ	13.279	20.576	21.179	23.979	24.087
NORTE	6.250	8.451	8.729	10.404	10.504
SUROESTE	5.913	8.007	8.244	10.010	10.157
OCCIDENTE	4.495	6.340	6.479	7.889	7.954
NORDESTE	3.462	5.495	5.786	7.418	7.560
BAJO CAUCA	3.152	6.117	6.685	8.896	9.021
MAGDALENA MEDIO	2.107	2.514	2.550	3.198	3.225

Fuente: Gobernación de Antioquia.