



IES VU20
INGRESO 27

MÓDULO
SABERES ESPECÍFICOS

INTRODUCCIÓN A LA
ESTADÍSTICA Y ANÁLISIS
DE DATOS

Tec. Sup. en Estadística y
Análisis de Datos



ÍNDICE

MÓDULO SABERES ESPECÍFICOS: INTRODUCCIÓN A LA ESTADÍSTICA Y ANÁLISIS DE DATOS

SABERES ESPECÍFICOS EJE N° 1

EJERCITACIÓN DE LA EJE N°1

CORRECCIÓN DE LA EJERCITACIÓN - EJE N° 1

SABERES ESPECÍFICOS EJE N° 2

EJERCITACIÓN DE LA EJE N° 2

CORRECCIÓN DE LA EJERCITACIÓN - EJE N° 2

SABERES ESPECÍFICOS EJE N° 3

EJERCITACIÓN DE LA EJE N° 3

CORRECCIÓN DE LA EJERCITACIÓN - EJE N° 3

MÓDULO SABERES ESPECÍFICOS INTRODUCCIÓN A LA ESTADÍSTICA Y ANÁLISIS DE DATOS

TEMARIO

INTRODUCCIÓN AL MÓDULO

EJE 1: FUNDAMENTOS DE ESTADÍSTICA

- Conceptos centrales de la estadística.
- Tipo de datos y variables.
- Tablas de frecuencias.

EJE 2: HERRAMIENTAS DIGITALES BÁSICAS

- Introducción a hojas de cálculo.
- Organización e interpretación de bases de datos.
- Representación gráfica de resultados.

EJE 3: APLICACIONES DE LA ESTADÍSTICA Y EL ANÁLISIS DE DATOS

- Introducción a Big Data: Análisis de datos para la toma de decisiones
- Lectura crítica de datos estadísticos.

SABERES ESPECÍFICOS INTRODUCCIÓN A LA ESTADÍSTICA Y ANÁLISIS DE DATOS

Introducción al Módulo

En este módulo abordaremos los saberes específicos vinculados a la **Estadística**, disciplina fundamental para el análisis, la interpretación y la toma de decisiones fundamentadas.

A través de la bibliografía seleccionada y una serie de actividades iniciales, podrás recorrer conceptos, métodos y herramientas esenciales para comprender cómo se producen, organizan y analizan los datos. Trabajaremos procedimientos básicos de sistematización, lectura crítica de información y criterios técnicos que orientan una práctica rigurosa, ética y fundamentada.

Estos contenidos son base de tu formación como futuro/a profesional en el campo de la estadística, y serán profundizados a lo largo de tu trayectoria académica.

¡Bienvenido/a!

EJE 1: Fundamentos de Estadística

En este eje abordaremos los conceptos fundamentales que permiten iniciar el estudio de la Estadística, disciplina clave para el análisis y la interpretación de datos en múltiples campos del conocimiento. Trabajaremos:

- **Conceptos centrales de la estadística**
- **Tipo de datos y variables**
- **Tablas de frecuencias**

Estos contenidos brindarán las bases necesarias para comprender cómo se organizan, describen y sintetizan conjuntos de datos, favoreciendo el desarrollo de habilidades para realizar análisis estadísticos.

¿Qué es la estadística?

Para comenzar, es importante precisar qué entendemos por **Estadística**. En términos amplios, puede definirse como la ciencia que estudia las regularidades presentes en conjuntos de fenómenos y desarrolla procedimientos para su descripción, análisis e interpretación. En este sentido, constituye “la colección de métodos científicos que permiten el análisis e interpretación de la información numérica” (García Ferrando, 1992, p. 23).

Una definición particularmente breve y significativa sostiene que la Estadística es la **ciencia de aprender con los datos**. La idea central de esta definición es precisamente *aprender*: la ciencia constituye la forma de adquirir conocimiento de manera sistemática, mientras que los datos representan el material a partir del cual ese conocimiento puede construirse (Ambrosi, 2008, p. 17). Como señala el autor, la ciencia implica una forma ordenada de producir conocimiento mediante procedimientos como la observación y la experimentación, a través de los cuales es posible obtener información sobre la realidad. Los datos, por su parte, son representaciones de las características observadas de los fenómenos que se busca conocer y analizar.

Desde esta perspectiva, la Estadística no se reduce a un conjunto de cálculos o técnicas matemáticas. Se trata de una disciplina que permite **transformar datos en información y, a partir de su análisis, generar conocimiento para comprender fenómenos y fundamentar decisiones**. Por ello, adquiere una importancia central tanto en la investigación científica como en ámbitos profesionales y sociales diversos, entre ellos el mercado, las organizaciones y el diseño, implementación y evaluación de políticas públicas.

Los datos pueden ser recopilados de distintas fuentes y según su naturaleza pueden ser:

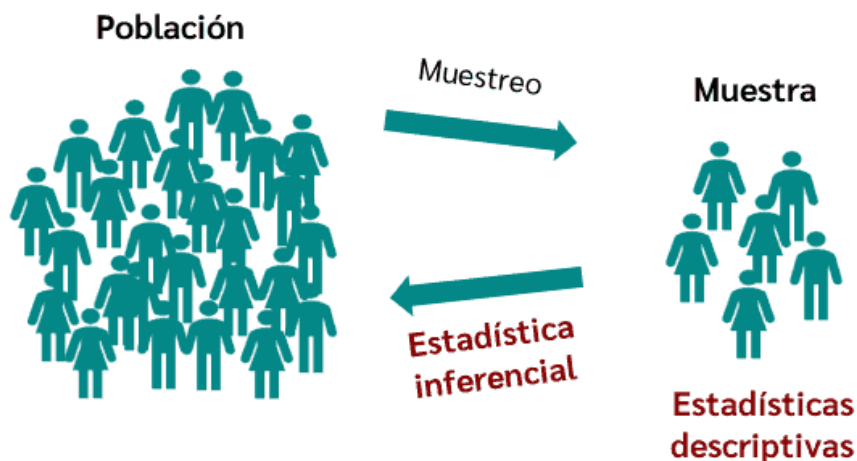
- **Poblacionales:** cuando se recopila información para todas las unidades de interés. Por ejemplo, los censos de población, hogares y viviendas, censo agropecuario, etc.
- **Muestrales:** cuando se recopila información de un subconjunto de las unidades de interés. Por ejemplo, encuestas políticas para conocer la intención de voto.
- **Experimentales:** cuando se obtiene información de un diseño experimental. Por ejemplo, con tratamientos diferenciales.

A la vez, dentro de esta disciplina se desarrollan dos tipos de estadística y cada una tiene objetivos diferentes y herramientas diferenciadas.

Estadística descriptiva: trata del recuento, ordenamiento y clasificación de los datos obtenidos por las observaciones realizadas. Este tipo de estadística ayuda a describir e ilustrar los conjuntos de datos. Para ello construye tablas y representa gráficos que permiten simplificar la complejidad de los datos que intervienen en la distribución. Además, se calculan parámetros estadísticos que caracterizan la distribución de los datos.

Estadística inferencial: es una rama de la estadística que estudia el comportamiento y propiedades de las muestras y la posibilidad, y límites, de la generalización de los resultados obtenidos, a partir de aquellas, a las poblaciones que representan. Dicho de otro modo, esta rama estadística utiliza

diversas herramientas analíticas para extraer conclusiones sobre la población en estudio, a partir de datos de la muestra seleccionada.



La estadística descriptiva proporciona herramientas para describir una muestra (subconjunto limitado extraído de la población). A partir de la muestra, la estadística inferencial puede utilizarse para hacer una declaración sobre la población.

Tipos de datos y escalas de medición

Los tipos de datos dependen de las variables que se utilizarán en un estudio. Las unidades de análisis son los elementos que componen el conjunto estudiado, el cual tiene que estar delimitado con precisión y exactitud.

Las unidades de análisis, siguiendo a López-Roldán y Fachelli (2015), pueden ser personas, hogares-familias, grupos, aulas, empresas, zonas geográficas, instituciones, documentos.

Como en todo el proceso de investigación, es la problemática de interés la que guía la determinación de las unidades, contando con las limitaciones tanto de índole técnica como presupuestaria. Por otro lado, la determinación de las unidades conlleva la especificación del ámbito geográfico o espacial, así como la referencia temporal sobre los que se definen las unidades.

Una vez definido el conjunto de las unidades de análisis, se puede describir las características que tiene cada una.

La variable es el resultado de la medición de un fenómeno a través de un procedimiento de observación que traduce en términos numéricos o categóricos los distintos aspectos, características o propiedades que se miden y que se asignan a cada unidad, objeto o acontecimiento.

Así podríamos definir la variable como cualquier característica o propiedad de un fenómeno que contenga dos o más categorías posibles con las que se pueden identificar las unidades.

Por ejemplo, si definimos que las unidades de análisis son todas aquellas personas que residen en el departamento de Guaymallén, quedan excluidas todas las personas que viven en otros departamentos. Y si queremos describir algunas de sus características podemos indagar sobre las edades que tienen, el nivel educativo alcanzado, sexo, entre otras variables.

El dato, por su parte, constituye la materia prima con la cual opera la estadística. Es una medida obtenida de una unidad de análisis, es el registro de una información. “Los datos finalmente son realizaciones concretas de la operación de observar una variable en una unidad de análisis” (López-Roldán y Fachelli, 2015, p.6).

El modo en que se articulan las unidades de análisis, variables y valores es bajo la forma de una **matriz de datos**.

Por ejemplo, si estamos realizando un análisis de 5 estudiantes del ingreso que pertenecen al departamento de Junín (unidad de análisis) y les preguntamos a cada uno/a cuál es su edad (variable 1), el género (variable 2) y la cantidad de cuestionarios aprobados en la plataforma (variable 3); la matriz de datos que obtenemos es la siguiente:

Unidad de análisis	Variables		
Ingresantes	Edad	Género	Cantidad de cuestionarios aprobados
Ingresante A	30	Mujer	5
Ingresante B	42	Varón	2
Ingresante C	18	Mujer	3
Ingresante D	20	Varón	2
Ingresante E	28	Varón	1

Clasificación de variables

Esas características que llamamos variables pueden ser cualitativas o cuantitativas, según su nivel de medición.

Variable cuantitativa: su valor sobre cada unidad resulta de una medida **numérica** realizada sobre la misma,

Variable cualitativa: su valor sobre cada unidad resulta de una clasificación que determina si la unidad pertenece o no a una modalidad (aunque ésta pueda identificarse con un código numérico).

Es importante tener presente que una variable tiene que poder tomar al menos dos valores u opciones, de lo contrario no sería una variable. Cada una de esas opciones es una categoría y, en su conjunto, conforman el sistema de categorías de cada variable.

Entonces recalamos que las variables **cualitativas** no se miden numéricamente, sólo se **clasifican**, mientras que las **cuantitativas** tienen una medida **numérica** (Casado, 2007, pág. 40)

Variables cuantitativas

Las variables cuantitativas pueden ser, según su naturaleza, discretas o continuas, según que la medida de cada modalidad sea un número entero o un número real, respectivamente.

El número de goles por partido de fútbol, el tamaño de una familia o el número de habitantes de los municipios son **variables discretas**: ninguna familia está formada por 3,45 miembros y ningún partido de fútbol termina con 2,5 goles.

La edad o el peso de una persona, el gasto en alimentación de una familia en un año o la hora de entrada al trabajo de los empleados de una empresa son **variables continuas**, que pueden tomar cualquier valor en un cierto intervalo: un empleado puede tener registrada su entrada al trabajo a las 8:32 horas, pero puede haber entrado en cualquier instante entre las 8:32 y las 8:33 horas. En la práctica las variables continuas se tratan como discretas al limitar el número de los decimales en su medición. (Casado, 2007, pág. 40)

Según su escala de medición:

Escala intervalar: se caracteriza porque su punto de origen o punto cero es un “punto de acuerdo” o “punto convencional” o “arbitrario”. En estas escalas, el valor cero no indica la ausencia de la característica

El ejemplo que mejor nos permite interpretarla es la temperatura, en la cual es fácil observar que el punto cero es convencional. Con esta escala podemos hacer comparaciones de distancia (valores de la variable) pero no respecto a sus magnitudes proporcionales. Por ejemplo, si ayer tuvimos una temperatura de 10 grados y hoy tenemos 20 grados, podemos afirmar que hoy hacen 10 grados (comparar distancias) más que ayer, pero no que hace el doble de calor (magnitudes proporcionales).

Escala de razón: presenta todas las propiedades de orden y distancia de la escala intervalar pero su origen es el cero absoluto. La existencia del cero absoluto implica que podemos realizar comparaciones proporcionales. Esta escala satisface todas las propiedades de los números reales, por lo tanto, podemos aplicar cualquier operación factible para este tipo de números.

La idea de cero absoluto es fácil de comprender pensando que, si al medir un elemento y obtenemos el valor cero, estamos frente a la ausencia del atributo de la variable. Si la variable es km. recorridos y obtengo el valor de 0 Km., implica que no se ha recorrido nada (ausencia de variable). Algunos ejemplos de estas variables son:: ingreso, estatura, peso, metros construidos, años de escolaridad, edad.

Variables cualitativas

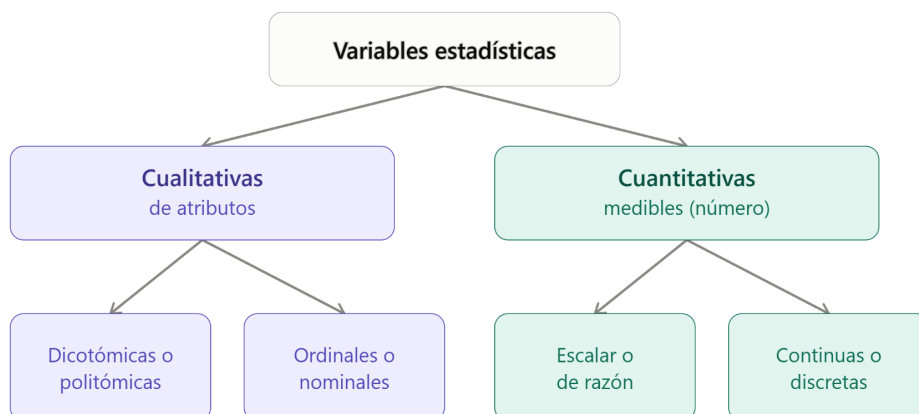
Cuando las variables categóricas se refieren solamente a la presencia o ausencia de una categoría se las llama dicotómicas o binarias, por ejemplo, tenencia de obra social (sí o no); buscó trabajo en los últimos doce meses (sí o no), es de nacionalidad argentino (sí o no).

Cuando se enumeran, o se hace referencia a todas las modalidades posibles, se dice que estamos frente a variables cualitativas policotómicas. Por ejemplo, si nos referimos al estado civil de una persona, mencionamos las posibles categorías que puede asumir: soltero/a, casado/a, viudo/a, divorciado/a, separado/a de hecho, otros; máximo nivel educativo alcanzado: primario, secundario, terciario universitario

Además, según los tipos de datos que contienen, la escala de medición de las variables cualitativas son:

Escalas nominales: no tienen ningún orden jerárquico, dado lo cual podríamos decir que ordenamos las categorías como queremos. Un ejemplo de variable nominal es “Gustos musicales”, cuyas categorías pueden ser “Folclore”, “Rock”, “Pop”, “Cumbia”, “Bachata”, “Otros” y si las ubicamos en diferente orden será exactamente lo mismo.

Escala ordinal: son aquellas cuyas categorías presentan un orden. Por ejemplo, un sistema de categorías con: Muy alto, Alto, Mediano, Bajo, Muy Bajo, para la variable salario, cada categoría esconde un valor numérico.



Distribución de frecuencias

Toda investigación va a buscar conocer determinadas características (variables) sobre un fenómeno o temática específica. Esa información se organiza para poder ser procesada y analizada. En tal sentido, dice Hernández Sampieri (1998) una distribución de frecuencia es un conjunto de puntuaciones ordenadas en sus respectivas categorías (pág. 345).

Además, en las distribuciones de frecuencias podemos agregar las frecuencias relativas, que nos permiten expresar en porcentaje la representación de cada categoría o valor en el todo y las frecuencias acumuladas. Las primeras son los porcentajes de casos en cada categoría; y las frecuencias acumuladas, son lo que se va acumulando en cada categoría, desde la más baja hasta la más alta.

Veamos el siguiente ejemplo:

Supongamos que nos ubicamos en la entrada de una de las sedes del instituto y anotamos en una planilla el nivel educativo de cada persona, sean estudiantes, docentes o personal del instituto que va ingresando. Finalizamos nuestra medición al ingresar 100 personas y disponemos entonces de una planilla con anotaciones puramente cualitativas: superior completo, superior en curso, secundario completo, secundario incompleto, primario completo, primario incompleto.

Sin embargo, podemos realizar un análisis estadístico de los resultados al observar las frecuencias de ocurrencia de cada nivel educativo. Al procesar los datos, los tabulamos y obtenemos la siguiente tabla:

NIVEL EDUCATIVO	FRECUENCIAS OBSERVADAS (n)	FRECUENCIA RELATIVA (fr)	FRECUENCIA RELATIVA PORCENTUAL (fr%)	FRECUENCIA ACUMULADA (fa)
Primario incompleto	7	0,007	0,07%	7
Primario completo	8	0,08	8%	15
Secundario incompleto	12	0,12	12%	27
Secundario completo	14	0,14	14%	41
Superior en curso o incompleto	33	0,33	33%	74
Superior completo	26	0,26	26%	100
Total (N)	100	1	100%	

En la columna “Frecuencia relativa”, computamos el cociente $f = n / N$, es decir, el número de personas observadas de cada nivel educativo, dividido por el total de observaciones, en nuestro ejemplo 100. Como en este ejemplo las observaciones son justo 100, la columna f puede parecer redundante, pero en general con un número cualquiera de observaciones totales, resulta práctica para determinar el peso de cada categoría.

En la columna “Frecuencia relativa porcentual”, multiplicamos la frecuencia relativa por 100 y obtenemos (f%). Por ejemplo, en el caso del valor superior en curso o incompleto, la frecuencia absoluta, es decir la cantidad de personas observadas en ese tiempo y espacio delimitados que tienen ese nivel educativo

fueron 33 personas, la frecuencia relativa es $33/100$, es decir 0,33, por lo tanto su frecuencia relativa porcentual es $0,33 \times 100 = 33\%$. Entonces, el 33% de las personas que ingresaron ese día tenían como nivel educativo alcanzado superior en curso o incompleto.

En la última columna se encuentra la “Frecuencia acumulada”, la cual permite observar el acumulado de categoría en categoría. Por ejemplo, si tomamos la frecuencia acumulada hasta la categoría secundario incompleto, sabemos que 27 de las 100 personas observadas tenían como máximo nivel educativo hasta el secundario incompleto (incluyendo entonces a quienes tenían primaria incompleta y completa).

EJERCITACIÓN EJE 1

Aquí encontrarás actividades de autoejercitación del eje N° 1 Estas actividades no las debes entregar, te servirán para que practiques tus conocimientos luego de la lectura del material y para estar más preparado/a para realizar el Ejercicio de Aplicación de este módulo.

Actividad N° 1: Indica si las siguientes afirmaciones son V (verdaderas) o F (falsas).

Afirmación	V/F
Si se estudia a todas las personas que residen en un departamento, los datos obtenidos pueden considerarse poblacionales.	
Una encuesta realizada a un subconjunto de personas para conocer una característica de toda una población produce datos poblacionales.	
La estadística descriptiva permite organizar y sintetizar los datos observados, mientras que la estadística inferencial permite realizar declaraciones sobre una población a partir de una muestra.	
Una característica que solo puede asumir un único valor constituye una variable, porque igualmente describe a las unidades de análisis.	
En una escala intervalar, una diferencia de 10 unidades tiene significado, pero no puede afirmarse que un valor sea proporcionalmente el doble que otro.	
La escala de razón permite realizar comparaciones proporcionales porque posee un cero absoluto.	
Una variable ordinal y una variable nominal se diferencian por la cantidad de categorías posibles de respuesta	

Actividad N° 2: Complete cada situación con uno de los siguientes conceptos

unidad de análisis – variable – dato – categoría – matriz de datos – estadística inferencial

1. En un estudio sobre estudiantes de nivel superior, cada estudiante seleccionado constituye una _____.
2. Si se registra la edad de cada estudiante, la edad constituye una _____.
3. Si uno de los estudiantes tiene 28 años, “28 años” constituye un _____ correspondiente a esa variable.
4. En la variable “nivel educativo alcanzado”, “secundario completo” constituye una _____.
5. La organización conjunta de las unidades de análisis, las variables y sus valores se presenta mediante una _____.
6. Cuando se utiliza información de una muestra para realizar una declaración sobre la población, se está aplicando _____.

Actividad N° 3: Marcar con una cruz la opción correcta

- **Un dato muestral proviene de:**
 - Un censo nacional
 - Un subconjunto de unidades de interés
 - Un experimento de laboratorio exclusivamente
- **Un ejemplo de variable continua es:**
 - Cantidad de hijos
 - Número de goles
 - Peso corporal

- **Los colores de autos registrados en un estacionamiento corresponden a una variable:**
 - Cualitativa nominal
 - Cuantitativa discreta
 - Cuantitativa continua
- **Se registra el estado civil de un grupo de personas con categorías como soltero/a, casado/a, viudo/a, divorciado/a. Se trata de una variable:**
 - Cuantitativa de razón
 - Cualitativa nominal
 - Cualitativa ordinal

CORRECCIÓN DE LA EJERCITACIÓN - EJE 1

Te sugerimos que antes de pasar a esta sección de los resultados intentes tener todas las actividades resueltas previamente para que puedas contrastar tus conocimientos de manera más efectiva. Recuerda que estas actividades son muy parecidas a las que tendrás en el Ejercicio de Aplicación Obligatorio de este módulo por eso es importante que las realices atentamente.

Actividad N° 1

Afirmación	V/F
Si se estudia a todas las personas que residen en un departamento, los datos obtenidos pueden considerarse poblacionales.	V
Una encuesta realizada a un subconjunto de personas para conocer una característica de toda una población produce datos poblacionales .	F
La estadística descriptiva permite organizar y sintetizar los datos observados, mientras que la estadística inferencial permite realizar declaraciones sobre una población a partir de una muestra.	V
Una característica que solo puede asumir un único valor constituye una variable, porque igualmente describe a las unidades de análisis.	F
En una escala intervalar, una diferencia de 10 unidades tiene significado, pero no puede afirmarse que un valor sea proporcionalmente el doble que otro.	V
La escala de razón permite realizar comparaciones proporcionales porque posee un cero absoluto.	V
Una variable ordinal y una variable nominal se diferencian por la cantidad de categorías posibles de respuesta	F

Actividad N° 2

1. En un estudio sobre estudiantes de nivel superior, cada estudiante seleccionado constituye una **unidad de análisis**.
2. Si se registra la edad de cada estudiante, la edad constituye una **variable**.
3. Si uno de los estudiantes tiene 28 años, “28 años” constituye un **dato** correspondiente a esa variable.
4. En la variable “nivel educativo alcanzado”, “secundario completo” constituye una **categoría**.
5. La organización conjunta de las unidades de análisis, las variables y sus valores se presenta mediante una **matriz de datos**.
6. Cuando se utiliza información de una muestra para realizar una declaración sobre la población, se está aplicando **estadística inferencial**.

Actividad N° 3

- **Un dato muestral proviene de:**
 - ☐ Un censo nacional
 - ☒ **Un subconjunto de unidades de interés**
 - ☐ Un experimento de laboratorio exclusivamente
- **Un ejemplo de variable continua es:**
 - ☐ Cantidad de hijos
 - ☐ Número de goles
 - ☒ **Peso corporal**

- Los colores de autos registrados en un estacionamiento corresponden a una variable:
 - ✓ Cualitativa nominal
 - ✗ Cuantitativa discreta
 - ✗ Cuantitativa continua
- Se registra el estado civil de un grupo de personas con categorías como soltero/a, casado/a, viudo/a, divorciado/a. Se trata de una variable:
 - ✗ Cuantitativa de razón
 - ✓ Cualitativa nominal
 - ✗ Cualitativa ordinal

EJE 2: Herramientas Digitales Básicas

En este eje exploraremos herramientas digitales esenciales para el manejo y análisis de información en entornos académicos y profesionales. Nos centraremos en el uso básico de hojas de cálculo como recurso fundamental para organizar, procesar y visualizar datos. Trabajaremos:

- **Introducción a hojas de cálculo.**
- **Organización e interpretación de bases de datos.**
- **Representación gráfica de resultados.**

Estos contenidos permitirán desarrollar habilidades iniciales para gestionar información de manera eficiente, automatizar cálculos y elaborar representaciones visuales claras, facilitando así la comprensión y comunicación de resultados.

Introducción a las bases de datos

El siglo XXI nos ha traído como consecuencia del avance tecnológico, la generación de grandes cantidades de datos, en grandes volúmenes. Si bien hay datos que siempre han estado disponibles para ser analizados, como los censos, hoy se suman los datos que generamos en línea, ya sea las compras online, las consultas en Google, los likes en app, los viajes en Uber.

Sin embargo, el manejo de volúmenes grandes de datos tiene algunas complicaciones. Para comenzar es importante tener en claro qué son las bases de datos. Las bases de datos con estructura de registro que tienen las siguientes características:

*“1. Una base de datos representa algún aspecto o porción del mundo real, llamado **minimundo** o el **universo del discurso**. Los cambios, en dicha porción del mundo, serán reflejados en la base de datos.*

2. Una base de datos es una colección lógicamente coherente de datos con cierto significado inherente. Un rejunte aleatorio de datos sin conexión no puede considerarse una base de datos.

3. Una base de datos está diseñada, construida y poblada por datos con un propósito específico. Además, está destinada a un grupo específico de usuarios y tiene un conjunto preconcebido de aplicación en las que dichos usuarios están interesados.” (Perez, Perez y Seca, 2020).

Las bases de datos nos permiten el análisis de los datos, pero para ello es muy relevante el trabajo de limpieza, ordenamiento y consistencia de cada base de datos.

A medida que las dimensiones de la base de datos son mayores es necesario trabajar con sistemas de bases de datos. Estos sistemas se construyen de forma interdisciplinaria que implica la interacción de varios actores:

- Proveedor: diseñadores y administradores de bases de datos.

“Su trabajo consiste, en primer lugar, en recibir los datos del equipo que los haya relevado, uniformizar su sistema de registro (p. ej., tabulación) y limpiar los registros cuya corrupción pueda generar problemas al usuario final. Además, deben generar una estructura funcional para la base, es decir, transformarla a uno o más formatos que permitan la interacción mediante el uso de programas y lenguajes de computación de uso común, los llamados sistemas de gestión de bases de datos”. (Perez, Perez, y Seca, 2020, pág. 337)

- Usuarios: existen tres tipos
 - a. Ocasional
 - b. Paramétrico (también llamado usuario ingenuo): Son usuarios cuyas búsquedas no son muy sofisticadas, pero para las cuales los administradores de la base deben generar una interfaz accesible y potente.

- c. Sofisticados: científicos, ingenieros y analistas que hacen un uso a la vez intenso y complejo de las bases.

Es importante que tengamos en cuenta que, al interactuar con este tipo de bases de datos, estamos interiorizados en lo que es la programación de *scripts* y actualizado en cuanto a las políticas de las distintas API involucradas en la extracción de datos de diferentes sitios web.

Organización e interpretación de bases de datos

Para organizar una base de datos de manera eficaz, se debe seguir un proceso de diseño estructurado que garantice la integridad de los datos y la facilidad de su uso. Algunos de los pasos claves para organizar una base de datos, especialmente una de tipo relacional, que es la más común:

1. Definir el propósito de la base de datos

Tener claridad sobre la finalidad de la base de datos. Para ello se puede considerar:

¿Qué información necesita almacenar?

¿Para qué se utilizará la información?

¿Qué tipos de datos necesitará, como texto, números, fechas o imágenes?

2. Dividir la información en tablas

Una vez que sepa lo que quiere almacenar, debe dividir la información en temas o entidades principales. Por ejemplo, en una base de datos de ventas, las entidades podrían ser Clientes, Productos y Pedidos. Crear una tabla por entidad: Cada entidad se convertirá en una tabla independiente en la base de datos.

3. Convertir los datos en columnas

Dentro de cada tabla, las categorías de datos se convierten en columnas o campos.

Campos descriptivos: Cada columna debe representar un atributo específico del tema de la tabla. Por ejemplo, la tabla Clientes podría tener columnas como Nombre, Dirección y Correo electrónico.

Tipos de datos: Asigne el tipo de datos adecuado para cada columna (texto, número, moneda, etc.).

4. Revisión y consistencia de datos

En este momento se realiza un proceso para organizar los datos de manera que se evite la duplicación de información y se garantice la coherencia.

Se busca:

- evitar la redundancia: eliminar los datos duplicados, que ocupan espacio innecesario y pueden llevar a inconsistencias.
- mejorar la integridad: asegurar que los datos sean lógicos y estén relacionados correctamente.

5. Establecer nombres a los diferentes datos

Una convención de nombres consistente y clara es crucial para la legibilidad y el mantenimiento de la base de datos.

Nombres consistentes: Use nombres claros y descriptivos para tablas y columnas (por ejemplo, ventas, clientes en lugar de ventas cl).

Documente su diseño: Mantenga una documentación completa del esquema para futuras referencias y actualizaciones.

6. Realizar copias de seguridad y mantenimiento

Una vez que la base de datos esté organizada, el mantenimiento constante es necesario.

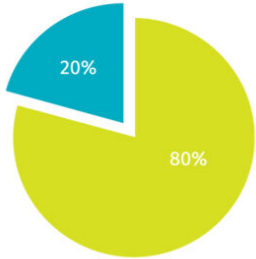
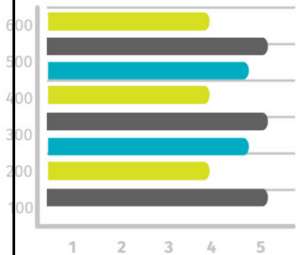
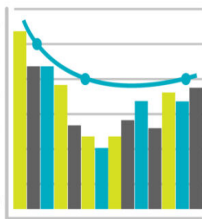
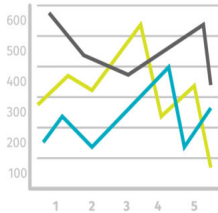
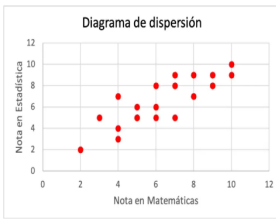
Copias de seguridad regulares: Mantenga copias de seguridad periódicas para evitar la pérdida de datos.

Revisión y optimización: Revise y actualice la organización de la base de datos de manera periódica, especialmente a medida que evoluciona su uso.

Representación gráfica de resultados.

La representación gráfica es una etapa crucial dentro de la generación de datos, porque implica generar información a partir de los datos generados. Transformar los datos procesados en información visual facilita la interpretación y comunicación a diferentes públicos, sean especializados o no (ciudadanos, políticos, medios).

Se pueden identificar diferentes gráficos según el tipo de variable y el objetivo del análisis de esos datos. A continuación presentamos algunos ejemplos:

Tipo de Datos	Gráfico Recomendado	Objetivo/ Uso	Imágenes
Distribución de Frecuencias absolutas o porcentuales (Variable Categórica)	Gráfico de Barras o Gráfico de Torta	-Mostrar la proporción de la población según los atributos de interés -Comparar partes en un TODO	 
Distribución de Frecuencias (Variable Numérica)	Histograma	Analizar la distribución de la variable .	
Evolución Temporal (Series de Tiempo)	Gráfico de Líneas	Mostrar la tendencia de los fenómenos a medir a lo largo de los años.	
Relación entre dos Variables Numéricas	Diagrama de Dispersión	Visualizar si existe correlación entre dos variables numéricas, es decir si el movimiento de una explica el comportamiento de la otra	

EJERCITACIÓN EJE 2

Aquí encontrarás actividades de auto ejercitación del eje N° 2 de este módulo. Estas actividades no las debes entregar, te servirán para que practiques tus conocimientos luego de la lectura del material y para estar más preparado/a para realizar el Ejercicio de Aplicación de este módulo, el cual tendrá actividades muy parecidas.

Actividad N°1: Marcar con una cruz la opción correcta en cada caso.

- Una base de datos representa un aspecto del mundo real llamado:
Metalenguaje
Archivo general
Minimundo
- El gráfico adecuado para mostrar cómo se distribuye un presupuesto entre diferentes áreas es:
Gráfico de líneas
Diagrama de dispersión
Gráfico de torta
- Las columnas dentro de una tabla representan:
Figuras
Registros
Atributos o campos
- Para la distribución de frecuencias de una variable categórica se usa:
Gráfico de barras
Histograma
Diagrama de dispersión

- **Para analizar la relación entre dos variables numéricas se utiliza un:**
 Histograma
 Gráfico de torta
 Diagrama de dispersión

Actividad N° 2: Marcar con una cruz si Verdadero o Falso

Afirmación	V/F
Las hojas de cálculo permiten organizar, procesar y visualizar datos.	
Las bases de datos son simples conjuntos aleatorios de datos sin relación.	
Los usuarios sofisticados realizan búsquedas simples y poco frecuentes.	
Dividir la información en tablas es un paso clave para organizar bases de datos relacionales.	
La redundancia de datos mejora la consistencia de las bases de datos.	
La representación gráfica facilita la interpretación de los resultados para diferentes públicos.	
Un gráfico de líneas es adecuado para mostrar la evolución temporal de una variable.	
Para una variable categórica, el gráfico recomendado es un histograma.	

Actividad N° 3: Completar con palabra clave (elegir una de tres opciones)

- Las hojas de cálculo permiten organizar, procesar y _____ datos.
(visualizar, ocultar, exportar)
- Una base de datos debe ser una colección _____ de datos.
(aleatoria, coherente, temporal)

- El minimundo también se denomina _____. (sistema raíz, base original, universo del discurso)
- Los proveedores son quienes diseñan y administran _____. (bases de datos, gráficos, interfaces web)
- Dividir la información en temas principales permite crear _____. (tablas, gráficos, algoritmos)
- Dentro de cada tabla, los atributos se expresan en _____. (metadatos, columnas, filas especiales)
- Un gráfico adecuado para mostrar partes de un todo es el gráfico de _____. (torta, líneas, cajas)
- Para observar la tendencia del desempleo a lo largo del tiempo se utiliza un gráfico de _____. (líneas, barras apiladas, dispersión)

CORRECCIÓN DE LA EJERCITACIÓN - EJE 2

Te sugerimos que antes de pasar a esta sección de los resultados intentes tener todas las actividades resueltas previamente para que puedas contrastar tus conocimientos de manera más efectiva. Recuerda que estas actividades son muy parecidas a las que tendrás en el Ejercicio de Aplicación Obligatorio de este módulo por eso es importante que las realices atentamente.

Actividad N°1: Marcar con una cruz la opción correcta en cada caso.

- Una base de datos representa un aspecto del mundo real llamado:
 - ☐ Metalenguaje
 - ☐ Archivo general
 - ☒ **Minimundo**

- El gráfico adecuado para mostrar cómo se distribuye un presupuesto entre diferentes áreas es:
 - ☐ Gráfico de líneas
 - ☐ Diagrama de dispersión
 - ☒ **Gráfico de torta**

- Las columnas dentro de una tabla representan:
 - ☐ Figuras
 - ☐ Registros
 - ☒ **Atributos o campos**

- Para la distribución de frecuencias de una variable categórica se usa:
 - ☒ **Gráfico de barras**
 - ☐ Histograma
 - ☐ Diagrama de dispersión

- Para analizar la relación entre dos variables numéricas se utiliza un:

☒ Histograma

☒ Gráfico de torta

☒ **Diagrama de dispersión**

Actividad N° 2: Marcar con una cruz si Verdadero o Falso

Afirmaciones	
Las hojas de cálculo permiten organizar, procesar y visualizar datos.	V
Las bases de datos son simples conjuntos aleatorios de datos sin relación.	F
Los usuarios sofisticados realizan búsquedas simples y poco frecuentes.	F
Dividir la información en tablas es un paso clave para organizar bases de datos relacionales.	V
La redundancia de datos mejora la consistencia de las bases de datos.	F
La representación gráfica facilita la interpretación de los resultados para diferentes públicos.	V
Un gráfico de líneas es adecuado para mostrar la evolución temporal de una variable.	V
Para una variable categórica, el gráfico recomendado es un histograma.	F

Actividad N° 3: Completar con palabra clave (elegir una de tres opciones)

- Las hojas de cálculo permiten organizar, procesar y **visualizar** datos.
- Una base de datos debe ser una colección **coherente** de datos.
- El minimundo también se denomina **universo del discurso**.

- Los proveedores son quienes diseñan y administran **bases de datos**.
- Dividir la información en temas principales permite crear **tablas**.
- Dentro de cada tabla, los atributos se expresan en **columnas**.
- Un gráfico adecuado para mostrar partes de un todo es el gráfico de **torta**.
- Para observar la tendencia del desempleo a lo largo del tiempo se utiliza un gráfico de **líneas**.

EJE 3: Aplicaciones de la Estadística y Análisis de datos

En este eje abordaremos el análisis de datos masivos como un valor estratégico para la toma de decisiones dentro de las organizaciones. Para ello se trabaja:

- **Introducción a Big Data: Análisis de datos para la toma de decisiones**
- **Lectura Crítica de Datos Estadístico**

Nos introduciremos en las dimensiones del Big Data y su soporte técnico, para comenzar a comprender la diversidad de saberes y técnicas que convergen en el análisis de datos en la actualidad, además proporcionaremos un marco mínimo para desarrollar una lectura crítica de la información estadística con el fin de detectar sesgos.

Introducción a Big Data: Análisis de datos para la toma de decisiones

Como trabajamos en el apartado anterior, a principio del siglo XXI el crecimiento desproporcionado de los datos almacenados provocó el colapso de las capacidades de análisis manual, convirtiendo la transición hacia la gestión automatizada en un imperativo estratégico, más que en una opción técnica.

Sosa Escudero (2020, p.36) nos propone que pensemos en todo lo que hicimos en las últimas dos horas. *Si salimos a caminar con el celular en el bolsillo, es muy probable que hayamos dejado un rastro de nuestra ubicación geográfica, y esto se vuelve aún más evidente si encendimos el GPS para manejar. Algo parecido ocurre si salimos a trotar con un reloj inteligente que registra el ritmo cardíaco y la cantidad de pasos. Lo mismo sucede si pagamos con tarjeta de crédito, viajamos en subte, miramos una serie en Netflix, le dimos "me gusta" a una foto de un familiar en Facebook, enviamos o recibimos*

un correo electrónico, o buscamos un par de zapatillas en Amazon. Cada una de estas acciones deja datos como huella.

El volumen de información generada por estos medios supera cualquier idea de magnitud que podamos. Sin embargo, el tamaño es apenas una parte del fenómeno. A diferencia de una encuesta tradicional, los datos del big data surgen de manera desordenada y espontánea, sin seguir un diseño metodológico previo. (Sosa Escudero, 2020)

El desbordamiento de las bases de datos transaccionales y científicas exigió una respuesta arquitectónica vital: el surgimiento del KDD (Knowledge Discovery in Databases), este representa el proceso completo y sistemático de extracción de conocimiento, que abarca desde la preparación crítica de los datos hasta la interpretación de los resultados para hallar relaciones y patrones (García y otros, 2018)

La extracción de conocimiento efectiva requiere una convergencia multidisciplinar que integra:

- Áreas de Conocimiento: Computación, estadística, matemáticas, ingeniería y adquisición de conocimiento.
- Técnicas de Procesamiento: Aprendizaje automático (machine learning), redes neuronales, razonamiento basado en casos, razonamiento aproximado y visualización avanzada.
- Tareas Operativas:
 - Inducción de reglas y detección de dependencias.
 - Clasificación de grandes volúmenes.
 - Reconocimiento de patrones y modelado predictivo.

Transforma registros masivos en activos estratégicos. García y otros (2018) ejemplifican en distintos sectores. En el sector bancario, esto se traduce en sistemas de detección de fraude en tiempo real; en salud y biología, permite avanzar en genómica y medicina de precisión; y en el gobierno electrónico, optimiza la gestión de ciudades inteligentes.

La capacidad de reducir la complejidad de búsqueda mediante estas técnicas permite a las organizaciones actuar sobre modelos descriptivos que el ojo humano no podría detectar. Esta necesidad de procesar patrones complejos bajo una metodología estructurada en entornos de datos masivos es lo que define la era del Big Data.

El éxito de una estrategia de datos se apoya en cuatro pilares fundamentales, que constituyen el **sustento técnico del Big Data**

1. Bases de Datos: Los mayores éxitos actuales se atribuyen a avances donde los ordenadores superan a los humanos en capacidad de procesamiento. Esto implica el uso de bases de datos no relacionales y procesamiento en clúster a nivel de hardware para una gestión estructurada y eficiente.
2. Visualización: El entendimiento humano depende de representaciones gráficas que permitan filtrar la complejidad y entender tendencias de forma intuitiva.
3. Estadística: Aporta el rigor analítico necesario para desarrollar modelos que expliquen el comportamiento de los datos con validez científica.
4. Aprendizaje Automático (Machine Learning): Permite la creación de modelos conceptuales predictivos, automatizando el descubrimiento de patrones que serían invisibles mediante métodos convencionales.

Esta combinación tecnológica es la única vía para reducir la complejidad de búsqueda y resolver problemas de representación del conocimiento que las organizaciones enfrentan hoy. La tecnología no es un fin, sino el medio para suministrar conocimiento útil a los tomadores de decisiones en sistemas reales.

Para garantizar la calidad del éxito, se deben aplicar medidas cuantitativas rigurosas a los patrones obtenidos. Un patrón sólo es valioso si cumple con criterios de precisión, utilidad y simplicidad; un modelo demasiado complejo para ser ejecutado por un humano carece de valor estratégico. En última instancia, la madurez de una organización se mide por su capacidad de

transformar la explosión de datos en factores estratégicos que impulsen el beneficio y el crecimiento institucional.

Lectura Crítica de Datos Estadístico

En el campo de la Estadística, no basta con saber calcular promedios y/o desviaciones; la habilidad crucial es la lectura crítica. Esto significa ir más allá de los números y gráficos superficiales para comprender el contexto, la metodología y las posibles manipulaciones o sesgos que subyacen a los datos. La crítica de datos se basa en cuatro preguntas fundamentales que todo analista de datos debe plantearse:

1. ¿Quién produce el dato?

- La fuente es vital. No es lo mismo un dato sobre empleo proveniente de un organismo oficial, que el de una consultora privada o un partido político.
- Preguntarnos ¿Cuál es el interés de la fuente al publicar este resultado? ¿Tiene una agenda (política, comercial, ideológica) que podría influir en la presentación o interpretación?

2. ¿Cómo se produjo el dato?

Esta pregunta hace referencia a la metodología que existe detrás del dato, es decir que sustenta el dato final. Cada método tiene su propio sesgo

3. ¿Qué dice el dato realmente y qué no dice?

La interpretación es un elemento clave para el análisis de datos.

- Causa vs. Correlación: Este es el error más común. Si dos variables se mueven juntas (correlación), no implica que una sea la causa de la otra. Por ejemplo, el aumento del consumo de helados se correlaciona con el aumento de crímenes, pero la causa es la llegada del verano (la temperatura).
- Contexto: Los números aislados no tienen valor. Es clave compararlos: ¿Cómo era este indicador el año pasado? ¿Cómo

se compara Mendoza con otras provincias o con el promedio nacional?

- Medidas de Tendencia Central: Nunca se conforme solo con el promedio (media); debe considerar las otras medidas de tendencia central y de dispersión, para conocer la variabilidad de la distribución

4. ¿Cómo se está presentando el dato?

La visualización y el lenguaje utilizado para presentar los datos es muy relevante a la hora de analizar datos:

- Los gráficos pueden ser engañosos. Es importante estar atentos a
 - Ejes cortados:
 - Escalas inapropiadas: Usar escalas que no empiezan en cero para variables de magnitud, distorsionando la percepción de la proporción.
 - Gráficos 3D o Complejos: A menudo añaden confusión en lugar de claridad.
- El lenguaje utilizado en los titulares o informes también es crucial. Títulos como "El empleo se dispara" pueden esconder una leve mejoría dentro del margen de error muestral.

EJERCITACIÓN DEL EJE 3

Aquí encontrarás actividades de auto ejercitación del eje N° 3 de este módulo. Estas actividades no las debes entregar, te servirán para que practiques tus conocimientos luego de la lectura del material y para estar más preparado/a para realizar el Ejercicio de Aplicación de este módulo.

Actividad N° 1: Marcar con una cruz si es Verdadero o Falso

Afirmaciones	Verdadero/ Falso
La lectura crítica de datos implica analizar únicamente los valores numéricos presentados.	
La fuente que produce el dato puede influir en la interpretación de los resultados.	
Los cambios en las definiciones de indicadores pueden modificar significativamente los resultados estadísticos.	
El surgimiento del KDD (Knowledge Discovery in Databases) representa únicamente la etapa final de interpretación de resultados, sin incluir la preparación de los datos.	
La extracción de conocimiento efectiva requiere la convergencia de varias disciplinas, como computación, estadística, matemáticas e ingeniería	

Actividad N° 2: Marcar la respuesta adecuada.

- **El Knowledge Discovery in Databases (KDD) representa:**

Solo la etapa final de interpretación de los resultados de un análisis

El proceso completo y sistemático de extracción de conocimiento, desde la preparación de los datos hasta la interpretación de los resultados

Una técnica exclusiva de visualización de datos

- **Los cuatro pilares fundamentales del Big Data mencionados en el texto, son:**

Hardware, software, redes y seguridad

Bases de datos, visualización, estadística y aprendizaje automático

Encuestas, censos, registros administrativos y big data

- **Un patrón hallado en los datos carece de valor estratégico**

Cuando es demasiado complejo para ser ejecutado o comprendido por un humano

Cuando se obtiene mediante aprendizaje automático

Cuando proviene de una base de datos no relacional

- **La primera pregunta fundamental para hacer una lectura crítica de un dato es:**

¿Qué dice el dato realmente y qué no dice?

¿Quién produce el dato?

¿Cómo se está presentando el dato?

- **El aumento del consumo de helados se correlaciona con el aumento de crímenes en verano. Esto es un ejemplo de:**

Una relación de causa y efecto directa

Un error de escala inapropiada en un gráfico

Correlación que no implica causalidad (ambas se explican por la llegada del verano)

- **En un gráfico, una escala que no empiece en cero para una variable de magnitud:**

Distorsiona la percepción de la proporción entre los valores

Hace que el gráfico sea imposible de leer

No genera ningún problema si el gráfico es 3D

- **Cuando dos variables se mueven juntas, hablamos de:**
 - Muestra
 - Causalidad
 - Correlación
- **La confidencialidad es especialmente importante porque:**
 - Mejora la estética del informe
 - Protege la identidad de personas encuestadas
 - Asegura gráficos más simples

Actividad N° 3: Completar con palabra clave (elegir una de tres opciones).

- El pilar de _____ aporta el rigor analítico necesario para desarrollar modelos que expliquen el comportamiento de los datos con validez científica. (estadística, visualizaciones, Big Data)
- El _____ permite crear modelos predictivos y automatizar el descubrimiento de patrones invisibles para los métodos convencionales. (aprendizaje automatizado, software, correlaciones)
- La tecnología no es un fin en sí misma, sino el _____ para suministrar conocimiento útil a los tomadores de decisiones. (cluster, medio, gráfico)
- La lectura crítica de datos comienza preguntándose quién _____ el dato. (fuente, produce, calcula)
- Un cambio en la _____ de un indicador puede modificar completamente su valor. (escala, definición, frecuencia)

CORRECCIÓN DE LA EJERCITACIÓN - EJE 3

Te sugerimos que antes de pasar a esta sección de los resultados intentes tener todas las actividades resueltas previamente para que puedas contrastar tus conocimientos de manera más efectiva. Recuerda que estas actividades son muy parecidas a las que tendrás en el Ejercicio de Aplicación Obligatorio de este módulo por eso es importante que las realices atentamente.

Actividad N° 1: Marcar con una cruz si es Verdadero o Falso

Afirmaciones	Verdadero/ Falso
La lectura crítica de datos implica analizar únicamente los valores numéricos presentados.	F
La fuente que produce el dato puede influir en la interpretación de los resultados.	V
Los cambios en las definiciones de indicadores pueden modificar significativamente los resultados estadísticos.	V
El surgimiento del KDD (Knowledge Discovery in Databases) representa únicamente la etapa final de interpretación de resultados.	F
La extracción de conocimiento efectiva requiere la convergencia de varias disciplinas, como computación, estadística, matemáticas e ingeniería	V

Actividad N° 2: Marcar la respuesta adecuada.

- **El Knowledge Discovery in Databases (KDD) representa:**
 - ☒ Solo la etapa final de interpretación de los resultados de un análisis
 - ☒ El proceso completo y sistemático de extracción de conocimiento, desde la preparación de los datos hasta la interpretación de los resultados
 - ☒ Una técnica exclusiva de visualización de datos

- **Los cuatro pilares fundamentales del Big Data mencionados en el texto, son:**
 - ✗ Hardware, software, redes y seguridad
 - ✓ Bases de datos, visualización, estadística y aprendizaje automático
 - ✗ Encuestas, censos, registros administrativos y big data
- **Un patrón hallado en los datos carece de valor estratégico**
 - ✓ Cuando es demasiado complejo para ser ejecutado o comprendido por un humano
 - ✗ Cuando se obtiene mediante aprendizaje automático
 - ✗ Cuando proviene de una base de datos no relacional
- **La primera pregunta fundamental para hacer una lectura crítica de un dato es:**
 - ✗ ¿Qué dice el dato realmente y qué no dice?
 - ✓ ¿Quién produce el dato?
 - ✗ ¿Cómo se está presentando el dato?
- **El aumento del consumo de helados se correlaciona con el aumento de crímenes en verano. Esto es un ejemplo de:**
 - ✗ Una relación de causa y efecto directa
 - ✗ Un error de escala inapropiada en un gráfico
 - ✓ Correlación que no implica causalidad (ambas se explican por la llegada del verano)
- **En un gráfico, una escala que no empieza en cero para una variable de magnitud:**
 - ✓ Distorsiona la percepción de la proporción entre los valores
 - ✗ Hace que el gráfico sea imposible de leer
 - ✗ No genera ningún problema si el gráfico es 3D

- Cuando dos variables se mueven juntas, hablamos de:

☐ Muestra

☐ Causalidad

☒ Correlación

Atividad N° 3: Completar con palabra clave (elegir una de tres opciones).

- El pilar de _____ aporta el rigor analítico necesario para desarrollar modelos que expliquen el comportamiento de los datos con validez científica. (estadística, visualizaciones, Big Data)
- El _____ permite crear modelos predictivos y automatizar el descubrimiento de patrones invisibles para los métodos convencionales. (aprendizaje automatizado, software, correlaciones)
- La tecnología no es un fin en sí misma, sino el _____ para suministrar conocimiento útil a los tomadores de decisiones. (cluster, medio, gráfico)
- La lectura crítica de datos comienza preguntándose quién _____ el dato. (lee, produce, calcula)
- Un cambio en la _____ de un indicador puede modificar completamente su valor. (escala, definición, frecuencia)

Bibliografía

Ambrosi, H. (2008). *La verdad de las estadísticas: aprender con los datos*. Lumiere.

García, J; Molina, J; Berlanga, Patricio, M; Bustamante, A y Padilla, W (2018). *Ciencia de datos. Técnicas analíticas y aprendizaje estadístico*. Alfaomega

García Ferrando, M (1992). *El análisis de la realidad social. Métodos y Técnicas de investigación*. Alianza Universidad.

Hernández Sampieri, R.; Mendoza Torres, C.; (2018) *Metodología de la Investigación. Las rutas cuantitativa, cualitativa y mixta*. Mc Graw-Hill

López-Roldán, P y Fachelli, S. (2015). *Metodología de la Investigación Social Cuantitativa*. Universitat Autònoma de Barcelona

Perez, L.; Perez, R; Seca, V. (2020). *Metodología de la investigación científica*. Maipue Editorial.

Sosa Escudero, W (2020). *Big data. Breve manual para conocer la ciencia de datos que ya invadió nuestras vidas*. Siglo XXI