

## **ANÁLISIS EXPLORATORIO DE DATOS Y SU DIDÁCTICA EN BACHILLERATO Y EDUCACIÓN SUPERIOR**

*Cuevas, Humberto*

jesus.humberto.cuevas@outlook.com

Instituto Tecnológico de Chihuahua II (México)

### **RESUMEN**

*En este documento se describen los principios básicos del Análisis Exploratorio de Datos [AED] y se plantea una actividad que parte de la identificación y delimitación de un problema que consiste en el incumplimiento de las dimensiones –longitudes y diámetros– de objetos físicos. Se espera que al final de la actividad, los participantes conozcan las fases del AED, sean capaces de realizar un análisis riguroso en consonancia con los lineamientos establecidos por una comunidad epistémica, den respuesta a interrogantes y conjeturas que se plantearán, examinen las ventajas de usar el AED y lo distinguan de la Estadística Descriptiva. Para la realización de cálculos y representaciones gráficas, se utilizará el entorno unificado y homogéneo de ejecución BioStatFLOSS, y el programa de cómputo estadístico Ststdisk.*

### **PALABRAS CLAVE**

Análisis Exploratorio de Datos, Investigación dirigida, Didáctica.

### **INTRODUCCIÓN**

El propósito central del taller es presentar a los participantes al menos una actividad que implique el Análisis Exploratorio de Datos [AED] y que pueda ser utilizada con sus estudiantes. No obstante, también se contemplan propósitos específicos:

1. Identificar, plantear y delimitar un problema relacionado con no-conformancias en términos de dimensiones de longitudes y diámetros de objetos físicos.
2. Proponer interrogantes de investigación.
3. Efectuar mediciones de longitudes y diámetros de los objetos físicos y registrarlos en una base de datos.
4. Presentar y examinar los orígenes, fases –exploratoria y confirmatoria– y virtudes del AED y la pertinencia de su utilización en el proceso de solución del problema planteado, especialmente en lo referido a las representaciones gráficas y los métodos resistentes y robustos.
5. Realizar un análisis de los datos a través del uso de software especializado.
6. Identificar patrones en los datos y adoptar una actitud crítica ante ellos.
7. Generar conjeturas susceptibles de ser probadas empíricamente y, en su caso, formular otro problema y continuar con el ciclo.

Es pertinente subrayar que para el proceso de cálculo y elaboración de gráficas se hará uso intensivo del entorno unificado y homogéneo de ejecución *BioStatFLOSS* y el programa de cómputo especializado *Ststdisk*. En la siguiente Imagen se muestra la interfaz de inicio de *BioStatFLOSS*.



**Imagen 1. BioStatFLOSS**  
Fuente: BioStatFLOSS

## MARCO DE REFERENCIA

La actividad a desarrollar se planeó desde una perspectiva racionalista-pragmática. Se promoverá entre los participantes:

- El análisis riguroso de problemas reales y contextuales.
- Utilizar el proceso de falsación en lugar del de verificación.
- Aplicar una taxonomía que permita evaluar desempeños de manera coherente y claramente diferenciada.
- Emplear un lenguaje denotativo.
- Establecer referentes empíricos.
- Se promoverá la adopción de una postura que busque evidencia para refutar teorías o conjeturas, en lugar de aceptarlas como válidas.

Habitualmente se reconoce al visionario matemático y estadístico John W. Tukey (1962, 1977) como el principal impulsor del AED y de la clarificación de las diferencias con el análisis de datos confirmativo.

Con el AED se pretende estudiar los datos desde todas las perspectivas posibles, apoyándose en diversas herramientas, especialmente las de cómputo estadístico. El objetivo



es obtener información valiosa; generar, modificar o ajustar hipótesis existentes; detectar patrones de comportamiento; “limpiar” los datos, entre otros.

La realización de un AED precede al desarrollo de actividades experimentales y cuasi experimentales. Una de sus principales virtudes es su aplicación en objetos de estudio categóricos.

Este taller parte de los siguientes supuestos:

1. Es necesario analizar rigurosamente los datos antes de iniciar una indagación de tipo experimental o cuasi experimental.
2. La introducción del AED en el aula promueve una actitud más positiva hacia la realización de investigación científica.
3. En virtud de la generación abundante de datos en la vida diaria y la necesidad de su análisis en la toma de decisiones, es indispensable que el currículo escolar asigne un espacio amplio en todos los niveles educativos, incluso si eso va en detrimento del tiempo y recursos asignados a los tópicos de estadística descriptiva.

## **DESCRIPCIÓN DE ACTIVIDADES**

En función de la cantidad y el perfil de los participantes, se crearán grupos de trabajo que no excederán los tres integrantes. Posteriormente se planteará un problema de investigación claramente delimitado y relacionado con inconsistencias en medidas representativas de un conjunto de piezas pequeñas. Seguido a ello se dedicará un periodo corto para proponer interrogantes de investigación en relación con el problema y acordes a un conjunto de estándares que permitirán discriminar entre las piezas que cumplen con ellos y las que no. Luego, se realizarán mediciones de longitudes y diámetros de los objetos físicos y se registrarán en una base de datos computacional.

Al final de la etapa, se presentará y examinarán los orígenes, fases –exploratoria y confirmatoria– y virtudes del AED y la pertinencia de su utilización en el proceso de solución del problema planteado, especialmente en lo referido a las representaciones gráficas y los métodos resistentes y robustos. En caso de ser requerido, se describirán las diferencias con los métodos de Estadística Descriptiva. Después se realizará un análisis de los datos a través del uso de software especializado y se procederá a detectar patrones de comportamiento. En la parte final, además de la interpretación de los resultados obtenidos, se examinará si es factible proponer hipótesis de trabajo que puedan probarse a la luz de evidencia objetiva, y de ser necesario, continuar con el proceso.

Es importante subrayar que el instrumento de medición será un calibrador vernier digital, y se dará entrenamiento para su uso en caso de ser necesario.

## **CONCLUSIONES**

La actividad propuesta en este taller se está desarrollando actualmente con estudiantes universitarios. En el estudio en curso, se está trabajando con una muestra de 59 estudiantes

de Ingeniería divididos en dos grupos, uno (A) con 30 estudiantes y el otro (B) con 29. En virtud de que los sujetos participantes no fueron asignados de manera aleatoria, es que se optó por un diseño cuasi experimental con un grupo de control y otro experimental.

En el de control se está utilizando una didáctica consistente en utilizar una base de datos de 250 registros con las siguientes variables: Sexo, Edad, Estatura y Peso Corporal (PC), Índice de Masa Corporal (IMC), Medida del Perímetro de la Cintura (MPC). En el grupo A se usa una didáctica típica que consiste en generar estadísticas y representaciones gráficas para cada variable y posteriormente interpretarlas de forma global y por sexo. Al final se les pedirá que elaboren un reporte escrito donde describan los resultados obtenidos y el análisis efectuado.

En el grupo experimental, inicialmente se planteó un problema de investigación relacionado con el sobrepeso y obesidad en la población estudiantil de la universidad. Se efectuó una revisión de literatura especializada en relación al problema en México, las perspectivas teórico-metodológicas desde las que se han realizado los estudios y su alcance.

Posteriormente se revisaron los estándares internacionales con relación a los límites inferior y superior de las variables PC, IMC y MPC en función del sexo del sujeto. Enseguida se plantearon de forma grupal nueve interrogantes y se procedió a diseñar un plan de trabajo para obtener una muestra representativa de estudiantes universitarios y obtener los datos necesarios para cada variable. Para el acopio de los datos se utilizó un estadímetro mecánico con báscula integrada y una cinta métrica para la medición del perímetro de la cintura. Todos los registros se almacenaron en una base de datos y se efectuó un análisis exploratorio de datos utilizando el Entorno y Lenguaje de Programación R incluido en BioStatFLOSS.

Al momento de redactar este documento, los participantes están en el proceso de detección de patrones en los datos. Para este entonces se han podido encontrar dos agrupamientos y se encuentran discutiendo sus características. Se espera que a través de sesiones de discusión realicen otros análisis y representaciones gráficas que permitan dar respuesta a las interrogantes planteadas, e incluso agregar otras o reformular las ya existentes y volver a iniciar el proceso. A la luz de los resultados obtenidos, se solicitará que redacten un primer borrador con sus hallazgos, valoren la pertinencia de formular conjeturas capaces de ser probadas empíricamente, y en su caso, formular un protocolo para un estudio de mayor envergadura.

Los resultados parciales señalan que los participantes del grupo experimental muestran un interés mayor en el AED a través del método de trabajo de investigación dirigida, y se muestran más receptivos a dedicar más tiempo en explorar los datos desde diferentes perspectivas, cálculos y representaciones gráficas, antes de planear un estudio más amplio. No obstante, es importante subrayar que estos resultados son parciales y subjetivos, en consecuencia no son concluyentes.



**REFERENCIAS**

Tukey, J. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33, 1-67.

Tukey, J. (1977). *Exploratory Data Analysis*, Reading. Massachussets: Addison Wesley.