LABORATORIO 1

Análisis estadístico computacional con R

Medidas de tendencia central y dispersión

Escuela de Informática y Computación

Información del laboratorio
Curso: EIY403 - Introducción al análisis de datos para otras carreras
Valor: 10 % de la nota final
Modalidad: Desarrollo guiado en clase
Duración: 90 minutos
Entregable: Archivo .Rmd compilado + archivo .html generado
Formato: R Markdown con especificaciones técnicas obligatorias
Entrega: Correo institucional a `jordyab00@gmail.com`
Asunto: `Lab1_Nombre_Apellido1_Apellido2`

1. Introducción

En este laboratorio implementaremos computacionalmente el análisis estadístico de datos de pH utilizando R. El objetivo es que comprendan la potencia de las herramientas computacionales para automatizar cálculos estadísticos y generar visualizaciones profesionales.

Contextualización:
Utilizaremos un dataset de 50 mediciones de pH de soluciones buffer de una empresa química industrial. Calcularemos todas las medidas estadísticas fundamentales y crearemos visualizaciones para comprender mejor el comportamiento de estos datos químicos.

2. Preparación del entorno de trabajo

2.1. Configuración obligatoria del R Markdown

Ejercicio 1: Configuración del documento (10 puntos)

Cree un nuevo archivo R Markdown (.Rmd) con la siguiente configuración en el encabezado YAML:

- **title:** "Laboratorio 1: Análisis de pH con R"
- **author:** Su nombre completo y carné
- **date:** Use `` `r Sys.Date()` `` para fecha automática
- **output:** html_document con las siguientes especificaciones:
  - `toc: true`
  - `toc_float: true`
  - `toc_depth: 3`
  - `number_sections: true`
  - `theme: flatly`
  - `highlight: tango`
  - `code_folding: show`
  - `df_print: paged`
  - `fig_caption: true`

2.2. Instalación y carga de librerías

Ejercicio 2: Carga de librerías (6 puntos)

Cree un chunk de código con **include=FALSE** para cargar las siguientes librerías:

- `knitr` - Para generar tablas profesionales
- `ggplot2` - Para gráficos profesionales

Si alguna librería no está instalada, use la función `install.packages()` antes de cargarla con `library()`.

3. Inserción de imagen y carga de datos

3.1. Imagen institucional

Ejercicio 3: Inserción de imagen (4 puntos)

Inserte una imagen relacionada con química o análisis de datos en su documento R Markdown. Puede usar una imagen de internet o una local. Use la sintaxis de markdown: `![Descripción](URL_o_ruta)` y agregue un caption descriptivo.

3.2. Importación del dataset

Ejercicio 4: Carga de datos (10 puntos)

a) Descargue el archivo `datos_ph_tarea1.txt` del repositorio del curso.
b) Use la función `read.csv()` para cargar los datos en R. Asigne el resultado a una variable llamada `datos_ph`.
c) Verifique la carga exitosa usando las funciones:

- `str()` para ver la estructura
- `head()` para ver las primeras filas
- `nrow()` para confirmar que hay 50 observaciones

d) Extraiga únicamente los valores de pH en un vector usando `datos_ph$ph` y asígnelo a una variable llamada `valores_ph`.

4. Organización y clasificación de datos

4.1. Exploración básica

Ejercicio 5: Exploración inicial (10 puntos)

a) Use la función `summary()` para obtener un resumen estadístico de la variable pH.
b) Use la función `length()` para confirmar que tiene 50 valores en `valores_ph`.
c) Ordene los valores usando `sort()` y guarde el resultado en `ph_ordenado`. Muestre los primeros 10 valores usando `head(ph_ordenado, 10)` y los últimos 10 valores usando `tail(ph_ordenado, 10)`.
d) Use las funciones `min()` y `max()` para encontrar los valores extremos y calcule el rango restando máximo menos mínimo.

5. Medidas de tendencia central

5.1. Media aritmética

Ejercicio 6: Cálculo de la media (10 puntos)

a) Calcule la media usando la función `mean()`.
b) Verifique el cálculo usando `sum()` y `length()`: divida la suma total entre el número de observaciones.
c) Use la función `round()` para redondear a 3 decimales.
d) Guarde el resultado en una variable llamada `media_ph`.

5.2. Mediana

Ejercicio 7: Cálculo de la mediana (10 puntos)

a) Calcule la mediana usando la función `median()`.
b) Para n=50 (par), verifique que es el promedio de las posiciones 25 y 26 en los datos ordenados. Use indexación: `ph_ordenado[25]` y `ph_ordenado[26]`.
c) Calcule la diferencia absoluta entre media y mediana usando `abs()`.
d) Guarde la mediana en una variable llamada `mediana_ph`.

6. Medidas de dispersión

6.1. Cuartiles y rango intercuartílico

Ejercicio 8: Cuartiles (15 puntos)

a) Use la función `quantile()` para calcular Q1 (percentil 25) y Q3 (percentil 75).
b) Calcule el rango intercuartílico usando la función `IQR()`.
c) Use `quantile()` con el argumento `probs = c(0.25, 0.5, 0.75)` para obtener los tres cuartiles simultáneamente.
d) Guarde Q1, Q3 e IQR en variables separadas.

6.2. Detección de outliers

Ejercicio 9: Outliers (6 puntos)

a) Calcule los límites para outliers:

- `limite_inferior = Q1 - 1.5 * IQR`
- `limite_superior = Q3 + 1.5 * IQR`

b) Use operadores lógicos para identificar outliers: `valores_ph <limite_inferior | valores_ph >limite_superior` y use `sum()` para contar cuántos outliers hay.

6.3. Varianza y desviación estándar

Ejercicio 10: Medidas de dispersión (12 puntos)

a) Calcule la varianza muestral usando `var()`.
b) Calcule la desviación estándar usando `sd()`.
c) Verifique que `sd()` es igual a `sqrt(var())`.
d) Calcule el coeficiente de variación: `(sd/mean) * 100`.

7. Visualización de datos

7.1. Histograma

Ejercicio 11: Histograma con ggplot2 (5 puntos)

Use el siguiente código y ejecútelo:

```
library(ggplot2)

# Crear un data.frame para ggplot
datos_grafico <- data.frame(ph = valores_ph)

# Crear histograma basico
ggplot(datos_grafico, aes(x = ph)) +
  geom_histogram(bins = 10, fill = "lightblue",
                 color = "black", alpha = 0.7) +
  labs(title = "Distribucion de valores de pH",
       subtitle = "Soluciones buffer industriales",
       x = "pH",
       y = "Frecuencia") +
  theme_minimal()
```

Única modificación requerida:

- Cambie el color de relleno de “lightblue” a “steelblue”

7.2. Boxplot

Ejercicio 12: Diagrama de caja (5 puntos)

Use el siguiente código y ejecútelo:

```
# Crear boxplot basico
ggplot(datos_grafico, aes(y = ph)) +
  geom_boxplot(fill = "lightgreen", alpha = 0.7) +
  labs(title = "Diagrama de caja: valores de pH",
       y = "pH") +
  theme_minimal()
```

Única modificación requerida:

- Cambie el color de relleno de “lightgreen” a “coral”

8. Interpretación y conclusiones

Ejercicio 13: Análisis final (8 puntos)

Escriba un párrafo breve (máximo 100 palabras) que incluya:
a) ¿La media de pH indica que las soluciones son ácidas, neutras o básicas?
b) ¿El coeficiente de variación indica alta o baja variabilidad en el proceso?
c) ¿Se encontraron outliers en los datos?
d) Basándose en los resultados, ¿considera que el proceso de producción tiene buen control de calidad?

9. Especificaciones técnicas de entrega

Requisitos obligatorios para la entrega:

Correo electrónico:

- Enviar desde su correo institucional UNA
- Destinatario: `jordyab00@gmail.com`
- Asunto: `Lab1_Nombre_Apellido1_Apellido2`
- Ejemplo: `Lab1_Maria_Rodriguez_Gonzalez`

Estructura del documento R Markdown:

- Header YAML completo con todas las especificaciones requeridas
- Índice flotante funcional (`toc_float: true`)
- Code folding habilitado (`code_folding: show`)
- Una imagen insertada con caption descriptivo
- Datos mostrados con formato paginado (`df_print: paged`)
- Librerías cargadas en chunk con `include=FALSE`

Archivos a entregar:

- Archivo .Rmd (código fuente)
- Archivo .html compilado
- Ambos archivos deben funcionar independientemente

10. Criterios de evaluación

Criterio	Puntos	Descripción
Configuración R Markdown (Ej. 1)	10	Header YAML completo y correcto
Carga de librerías (Ej. 2)	6	Chunk con include=FALSE
Imagen insertada (Ej. 3)	4	Imagen externa con caption
Carga de datos (Ej. 4)	10	Importación y verificación
Exploración básica (Ej. 5)	10	Summary, ordenamiento, extremos
Media aritmética (Ej. 6)	10	Cálculo y verificación
Mediana (Ej. 7)	10	Cálculo y verificación manual
Cuartiles (Ej. 8)	15	Q1, Q3, IQR correctos
Outliers (Ej. 9)	6	Límites y detección
Dispersión (Ej. 10)	12	Varianza, SD, CV
Histograma (Ej. 11)	5	Cambio simple de color
Boxplot (Ej. 12)	5	Cambio simple de color
Interpretación (Ej. 13)	8	Conclusiones apropiadas
Total	**100**	

11. Recursos de apoyo

Funciones R clave para este laboratorio:

Carga de datos: `read.csv()`, `str()`, `head()`, `tail()`, `nrow()`
Estadísticas básicas: `mean()`, `median()`, `var()`, `sd()`, `summary()`
Valores extremos: `min()`, `max()`, `quantile()`, `IQR()`
Ordenamiento: `sort()`, `length()`
Funciones lógicas: `sum()` con condiciones, `abs()`
Visualización: `ggplot()`, `geom_histogram()`, `geom_boxplot()`
Utilidades: `round()`, `sqrt()`

Este laboratorio desarrolla competencias fundamentales en análisis estadístico computacional, preparando a los estudiantes para el manejo eficiente de datos en sus disciplinas específicas.