MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA Y BIOESTADÍSTICA

PROGRAMARIO PARA EL ANÁLISIS DE DATOS (PAD)

Prueba de Evaluación Continua 2 (PEC2)

CONTENIDO Y ORIENTACIONES

Para realizar la PEC2, es necesario haber trabajado los siguientes laboratorios y recursos asociados:

- **LAB3. Fundamentos de programación y acceso a base de datos en R**
- **LAB4. Probabilidad y simulación en R**

La PEC2 consta de diversos ejercicios organizados en secciones que corresponden a cada uno de los contenidos de los diferentes laboratorios.

FORMATO DE ENTREGA

La solución de la PEC2 ha de realizarse en formato Rmarkdown.
Se entregará el documento generado por Rmarkdown en formato *.html o *.pdf.

FECHA LÍMITE DE ENTREGA

La fecha límite de entrega de la **PEC2** es el **25 de noviembre de 2025 hasta las 23:59 h.**

Sección 1. Fundamentos de programación y BBDD (5 puntos)

Ejercicio 1 (1 punto)

Interpretad el siguiente código que hace referencia a genes, su nivel de expresión (en TPM) y la categoría funcional a la que pertenecen (0.2 puntos). Mostrad cuál sería el resultado obtenido:

```
genes <- matrix(c(
   "BRCA1", 120, "Reparación de ADN",
   "TP53", 300, "Supresor tumoral",
   "MYC", 500, "Oncogén",
   "GAPDH", 1000, "Control interno",
   "EGFR", 250, "Receptor de membrana"
 ), nrow = 5, byrow = TRUE)
print("Matriz de genes:")
print(genes)
```

Cuestiones:

1. ¿Por qué es necesario usar *c()* para crear la matriz en R (0.2 puntos)?

2. ¿Qué estructura de datos en R resultaría más adecuada para representar información de genes con distintos tipos de datos (texto, número y categoría) (0.2 puntos)?

3. ¿Cómo cambiaría la visualización de la matriz si se eliminan los argumentos *nrow=5*, *byrow=TRUE* de la función *matrix()* (0.2 puntos)?

Escribid un código que recorra las filas de la matriz y muestre por pantalla la lista de genes que sean oncogenes o supresores tumorales, junto con su nivel de expresión (0.2 puntos).

El resultado esperado debe tener el siguiente formato:

```
Lista de genes relevantes en cáncer:
 Gen: TP53 - Expresión: 300 TPM
 Gen: MYC - Expresión: 500 TPM
```

Nota: Para resolver esta tarea, utilizad bucles (*for*) e instrucciones condicionales (*if*), no funciones predeterminadas de filtrado en R.

Ejercicio 2 (2 puntos)

Supongamos que queremos calcular el índice de expresión relativa de un gen en un experimento de qPCR, comparando el nivel de expresión de un gen de interés con el de un gen de referencia.

Se pide:

a) (0.5 puntos) Definir una función en R llamada *indice_expresion* que reciba dos parámetros:

- *exp_interes*: nivel de expresión del gen de interés.
- *exp_referencia*: nivel de expresión del gen de referencia.

La función debe calcular y devolver el índice de expresión relativa, definido como:

$$Indice = \frac{exp\_interes}{exp\_referencia}$$

b) (0.5 puntos) Mostrar el resultado por pantalla suponiendo, por ejemplo:
- *exp_interes = 250*
- *exp_referencia = 100*

c) (0.5 puntos) Probar la función con diferentes ejemplos para comprobar la consistencia del código:
- Valores numéricos distintos.
- Casos de error, por ejemplo cuando el valor de referencia sea 0 (evitar división entre cero).
- Casos en los que los parámetros no sean numéricos.

d) (0.5 puntos) Analizar cómo se pueden definir los parámetros:
- Introduciéndolos directamente en el código.
- Pidiéndolos al usuario mediante la función *readline()*.

Ejercicio 3 (2 puntos)

A partir del dataset ToothGrowth del paquete datasets, utilizad RSQLite y las instrucciones asociadas para explorar dicho conjunto de datos. Podéis encontrar más información de la descripción de este dataset en: https://www.rdocumentation.org/packages/datasets/topics/ToothGrowth.

Este dataset contiene:

- *len*: longitud del diente (respuesta).
- *supp*: tipo de suplemento (*VC* = ácido ascórbico, *OJ* = jugo de naranja).
- *dose*: dosis de vitamina C administrada (en mg/día).

a) Mostrad la longitud media del diente (*len*) en las cobayas que recibieron una dosis de al menos 1 mg/día (0.5 puntos).

b) Mostrad la longitud y la dosis de aquellas cobayas tratadas con jugo de naranja (*supp* = *"OJ"*) cuya longitud del diente sea superior a 20 (0.5 puntos).

c) Mostrad cuántos animales fueron tratados con cada tipo de suplemento (*supp*), agrupando los resultados por suplemento y ordenándolos de mayor a menor número de animales (0.5 puntos).

d) Mostrad la longitud media del diente agrupada por dosis y tipo de suplemento (*supp* y *dose*) (0.5 puntos).

Sección 2. Probabilidad y simulaciones (5 puntos)

Ejercicio 4 (2.5 puntos)

En este ejercicio estudiaremos el crecimiento de bacterias en un medio de cultivo usando distribuciones estadísticas y simulaciones en R. El número de horas que tarda una colonia de bacterias en duplicar su tamaño sigue aproximadamente una distribución log-normal con parámetros de media logarítmica (*meanlog*) igual a 1.2 y desviación típica logarítmica (*sdlog*) igual a 0.3. Si se añade un antibiótico, este tiempo de duplicación se ajusta bien a una distribución log-normal con *meanlog* = 1.5 y *sdlog* = 0.4.

a) Graficad las densidades de los tiempos de duplicación en ambos escenarios (con y sin antibiótico). Comentad qué diferencias observáis en los gráficos (0.6 puntos).

b) Calculad la probabilidad de que el tiempo de duplicación sea inferior a 2 horas en ambos casos. Después, calculad la probabilidad de que el tiempo sea superior a 5 horas. Realizad los cálculos exactos y comentad los resultados (0.6 puntos).

c) A través de 10000 simulaciones para cada escenario, estimad la media, varianza y percentiles 25%, 50% y 75% del tiempo de duplicación. Comparad estos valores con los obtenidos de forma exacta con la función *qlnorm* de R (0.8 puntos).

d) Escribid una breve reflexión sobre los dilemas éticos relacionados con el uso de antibióticos en exceso y la resistencia bacteriana (0.5 puntos)

Ejercicio 5 (2.5 puntos)

Queremos analizar el rendimiento de una vacuna contra una nueva enfermedad infecciosa. Supongamos que el 3% de la población no desarrolla inmunidad tras recibir la vacuna. Si una persona no desarrolla inmunidad, la probabilidad de infectarse al exponerse al virus es del 40%. Si una persona sí desarrolla inmunidad, la probabilidad de infectarse es solo del 2%.

a) Si seleccionamos 200 personas vacunadas, ¿cuál es la probabilidad de que al menos 5 se infecten? Calculadlo mediante un cálculo exacto (usando la binomial) y mediante 10000 simulaciones. Comparad los resultados (0.6 puntos).

b) Si seleccionamos 50 personas que no desarrollan inmunidad, ¿cuál es la probabilidad de que ninguna se infecte? Y si seleccionamos 150 personas que sí desarrollan inmunidad, ¿cuál es la probabilidad de que al menos 3 se infecten? Resolvedlo con cálculos exactos y simulaciones (0.6 puntos).

c) Graficad la distribución de probabilidad del número de infecciones esperadas en los grupos de personas con inmunidad y sin inmunidad. Comentad las diferencias entre ambas distribuciones (0.7 puntos).

d) A partir de vuestros resultados y del hecho de que un pequeño porcentaje de vacunados no desarrolla inmunidad, redactad una breve justificación sobre si es más eficaz centrarse en reforzar la inmunidad individual o en reducir la exposición general al virus (0.6 puntos).