

Práctica: Análisis Estadístico de Horas de Estudio

Objetivo:

El propósito de esta práctica es que implementes un programa en PHP que calcule **medidas estadísticas** clave a partir de un conjunto de datos, como son las horas de estudio diarias de un grupo de estudiantes. El programa deberá calcular la **media**, **mediana**, **desviación estándar** y **rango** de los datos proporcionados, y usar esos resultados para deducir un indicador de calidad de los hábitos de estudio.

Se considera lo siguiente:

Si la media de estudio son 2 horas -> la clase tiene buenos hábitos de calidad

Por debajo de dos horas tiene que mejorar

Descripción del problema:

Tienes un grupo de 30 estudiantes que han registrado la cantidad de horas que estudian por día durante una semana. Se te proporciona un arreglo con estos datos, y debes realizar lo siguiente:

Instrucciones:

1. Cálculo de la Media:

- Crea una función que reciba el arreglo de datos y calcule el promedio de horas de estudio por día. Para esto, deberás sumar todos los elementos del arreglo y dividir entre el número total de elementos.
- **Sugerencia:** Utiliza la función `array_sum()` para sumar los elementos del arreglo y `count()` para obtener el número de estudiantes.

```
$media = calcular_media($horas_estudio);
```

2. Cálculo de la Mediana:

- Implementa una función que ordene el arreglo y luego calcule la mediana, que es el valor central. Si el número de estudiantes es impar, selecciona el valor del medio; si es par, calcula el promedio de los dos valores centrales.
- **Sugerencia:** Usa `sort()` para ordenar los datos y luego encuentra el valor central. Recuerda que debes manejar de manera diferente los casos en los que el número de datos sea par o impar.

```
$media = calcular_mediana($horas_estudio);
```

1. Cálculo del Rango:

- Implementa una función que calcule el rango, que es la diferencia entre el valor máximo y el valor mínimo del arreglo de datos.
- **Sugerencia:** Usa `max()` y `min()` para obtener el valor máximo y mínimo del arreglo.

```
$rango = calcular_rango($horas_estudio);
```

1. Cálculo de la desviación estándar:

- Crea una función que calcule la desviación estándar, la cual mide la dispersión de los datos respecto a la media. Esto implica sumar las diferencias al cuadrado entre cada valor y la media, dividir por el número de estudiantes, y luego tomar la raíz cuadrada.
- **Sugerencia:** Utiliza `pow()` para elevar las diferencias al cuadrado y `sqrt()` para la raíz cuadrada.

2.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

```
$desviacion_tipica = calcular_desviacion_tipica($horas_estudio, $media);
```

3. Evaluación de la Calidad de los Hábitos de Estudio:

- Una vez calculadas las estadísticas, deberás deducir la **calidad de los hábitos de estudio** de los estudiantes.
 - Si la **media** está entre 3 y 5 horas por día, los estudiantes tienen buenos hábitos de estudio.
 - Si la **desviación estándar** es menor a 1.5 horas, los estudiantes tienen horarios consistentes.
- **Sugerencia:** Usa una simple estructura `if` para tomar decisiones basadas en los valores de las estadísticas.

```
if ($media >= 3 && $media <= 5) {
```

```
    echo "Los estudiantes tienen buenos hábitos de estudio.";
} else {
    echo "Los estudiantes tienen malos hábitos de estudio.";
}
```

Conjunto de datos de ejemplo:

```
$horas_estudio = [3.5, 4.0, 4.2, 3.8, 5.5, 2.9, 3.3, 4.1, 4.0, 4.2,
                 3.5, 4.8, 4.0, 5.4, 4.3, 3.9, 4.2, 3.7, 4.1, 4.6,
                 3.9, 4.1, 4.0, 4.4, 4.5, 3.6, 4.0, 4.9, 3.8, 4.2];
```

1. El programa deberá mostrar los siguientes resultados por pantalla:
 - **Media:** La media de las horas de estudio.
 - **Mediana:** La mediana del conjunto de datos.
 - **Desviación Estándar:** La desviación estándar de las horas de estudio.
 - **Rango:** La diferencia entre la cantidad máxima y mínima de horas de estudio.
 - **Conclusiones:** Evaluación sobre la calidad de los hábitos de estudio.

Solución

```
<?php

// Función para calcular la media
function calcular_media($datos) {
    $suma = array_sum($datos);
    $n = count($datos);
    return $suma / $n;
}

// Función para calcular la mediana
```

```
function calcular_mediana($datos) {  
    sort($datos);  
    $n = count($datos);  
    $mediana = ($n % 2 == 0) ? ($datos[$n / 2 - 1] + $datos[$n / 2]) / 2 :  
$datos[floor($n / 2)];  
    return $mediana;  
}
```

// Función para calcular la desviación estándar

```
function calcular_desviacion_tipica($datos, $media) {  
    $n = count($datos);  
    $suma_cuadrados = 0;  
  
    foreach ($datos as $valor) {  
        $suma_cuadrados += pow($valor - $media, 2);  
    }  
  
    return sqrt($suma_cuadrados / $n);  
}
```

// Función para calcular el rango

```
function calcular_rango($datos) {  
    return max($datos) - min($datos);  
}
```

// Datos de las horas de estudio por día de los estudiantes

```
$horas_estudio = [3.5, 4.0, 4.2, 3.8, 5.5, 2.9, 3.3, 4.1, 4.0, 4.2,  
3.5, 4.8, 4.0, 5.4, 4.3, 3.9, 4.2, 3.7, 4.1, 4.6,
```

```
3.9, 4.1, 4.0, 4.4, 4.5, 3.6, 4.0, 4.9, 3.8, 4.2];
```

```
// Calcular la media
```

```
$media = calcular_media($horas_estudio);
```

```
// Calcular la mediana
```

```
$mediana = calcular_mediana($horas_estudio);
```

```
// Calcular la desviación estándar
```

```
$desviacion_tipica = calcular_desviacion_tipica($horas_estudio, $media);
```

```
// Calcular el rango
```

```
$rango = calcular_rango($horas_estudio);
```

```
// Mostrar resultados
```

```
echo "Media: " . number_format($media, 2) . " horas<br>";
```

```
echo "Mediana: " . number_format($mediana, 2) . " horas<br>";
```

```
echo "Desviación Típica: " . number_format($desviacion_tipica, 2) . " horas<br>";
```

```
echo "Rango: " . number_format($rango, 2) . " horas<br>";
```

```
// Deducción del indicador de calidad de los hábitos de estudio
```

```
if ($media >= 3 && $media <= 5) {
```

```
    $calidad_estudio = "Buena";
```

```
} else {
```

```
    $calidad_estudio = "Mala";
```

```
}
```

```
echo "Calidad de los hábitos de estudio: $calidad_estudio<br>";
```

```
if ($desviacion_tipica < 1.5) {  
    echo "Los estudiantes tienen hábitos de estudio consistentes.<br>";  
} else {  
    echo "Los estudiantes tienen una variabilidad considerable en las horas de  
estudio.<br>";  
}  
?>
```

Vamos a suponer que queremos medir si existe una correlación entre **las horas de estudio** y **las calificaciones obtenidas** de un grupo de estudiantes. A continuación, implementaremos un programa en **PHP** para calcular el coeficiente de correlación entre estas dos variables.

Paso a paso:

1. Recopila los datos. Suponemos que tienes dos arreglos: uno con las horas de estudio y otro con las calificaciones.
2. Calcula el coeficiente de correlación de Pearson usando la fórmula.
3. Interpreta el resultado.

La **fórmula de correlación de Pearson** mide la fuerza y la dirección de la relación lineal entre dos variables. Su valor está entre -1 y 1, donde:

- 1 indica una correlación positiva perfecta.
- -1 indica una correlación negativa perfecta.
- 0 indica que no hay correlación.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

Donde:

- r es el coeficiente de correlación de Pearson.
- n es el número de pares de datos.
- $\sum xy$ es la suma del producto de x e y para cada par de datos.
- $\sum x$ es la suma de los valores de x .
- $\sum y$ es la suma de los valores de y .
- $\sum x^2$ es la suma de los cuadrados de los valores de x .
- $\sum y^2$ es la suma de los cuadrados de los valores de y .

Código PHP para calcular la correlación:

```
<?php

// Función para calcular el coeficiente de correlación de Pearson
function calcular_correlacion_pearson($x, $y) {
    $n = count($x);

    // Calcular las sumas necesarias
    $sum_x = array_sum($x);
    $sum_y = array_sum($y);
    $sum_xy = 0;
    $sum_x2 = 0;
    $sum_y2 = 0;

    for ($i = 0; $i < $n; $i++) {
```

```

    $sum_xy += $x[$i] * $y[$i];

    $sum_x2 += pow($x[$i], 2);

    $sum_y2 += pow($y[$i], 2);
}

// Aplicar la fórmula del coeficiente de correlación de Pearson

$numerador = $n * $sum_xy - $sum_x * $sum_y;

$denominador = sqrt(($n * $sum_x2 - pow($sum_x, 2)) * ($n * $sum_y2 -
pow($sum_y, 2)));

if ($denominador == 0) {
    return 0; // Evitar división por cero
}

$correlacion = $numerador / $denominador;

return $correlacion;
}

// Datos: horas de estudio (X) y calificaciones (Y) de los estudiantes
$horas_estudio = [2, 3, 5, 6, 8, 10];
$calificaciones = [60, 65, 70, 75, 80, 90];

// Calcular la correlación
$correlacion = calcular_correlacion_pearson($horas_estudio, $calificaciones);

// Mostrar el resultado

```



```
echo "El coeficiente de correlación de Pearson es: " . number_format($correlacion,
2) . "<br>";

// Interpretar el resultado
if ($correlacion > 0.7) {
    echo "Existe una fuerte correlación positiva entre las horas de estudio y las
calificaciones.<br>";
} elseif ($correlacion < -0.7) {
    echo "Existe una fuerte correlación negativa entre las horas de estudio y las
calificaciones.<br>";
} else {
    echo "No hay una correlación significativa entre las horas de estudio y las
calificaciones.<br>";
}

?>
```

Salida esperada:

El coeficiente de correlación de Pearson es: 0.97

Existe una fuerte correlación positiva entre las horas de estudio y las calificaciones.

Tareas para los alumnos:

1. **Modificar los datos:** Prueba cambiando los datos de entrada (por ejemplo, menos horas de estudio con mejores calificaciones) y observa cómo cambia el coeficiente de correlación.
2. **Interpretar el resultado:** Discute cómo las horas de estudio influyen en las calificaciones basándose en el coeficiente de correlación calculado.
3. **Añadir más datos:** Incrementa el tamaño de los arreglos para simular más estudiantes y evaluar cómo cambia la correlación.

La correlación es una herramienta muy útil para entender la relación entre dos variables, pero tiene sus limitaciones y puede dar lugar a problemas o malinterpretaciones si no se maneja adecuadamente.

1. Correlación no implica causalidad

- **Problema:** Una correlación positiva o negativa entre dos variables no significa que una cause la otra. A menudo, se comete el error de asumir que si dos variables están correlacionadas, entonces una es la causa de la otra.

Muestras pequeñas

- **Problema:** Con tamaños de muestra pequeños, una correlación puede ser significativa por pura casualidad. Esto aumenta la posibilidad de obtener correlaciones espurias, ya que hay menos datos para evaluar la verdadera relación.
- **Ejemplo:** Si solo observas a cinco personas, la correlación entre su altura y su salario podría ser alta simplemente porque seleccionaste un grupo pequeño que, por casualidad, sigue esa tendencia.

Sesgo por valores extremos (outliers)

- **Problema:** Los **outliers** o valores extremos en los datos pueden afectar drásticamente el valor de la correlación. Un solo punto de datos inusual puede distorsionar la relación entre las variables.
- **Ejemplo:** Si la mayoría de las personas estudian entre 2 y 5 horas al día y una persona estudia 20 horas, ese valor extremo puede aumentar o reducir artificialmente el coeficiente de correlación.

La correlación es una herramienta poderosa, pero requiere un análisis cuidadoso y crítico para evitar conclusiones erróneas.