Estad. Descriptiva e Inferencia

Práctica de el módulo de Estadística

Laura Rodríguez Ropero

Máster en Big Data, Data Science e IA

Estad. Descriptiva e Inferencia

Práctica de el módulo de Estadística

por

Laura Rodríguez Ropero

21/11/2024

Índice

Profesor:	Conrado
Facultad:	Facultad de Estudios Estadísticos
Universidad:	Universidad Complutense de Madrid
Máster:	Big Data, Data Science e Inteligencia Artificial

python

1

Análisis Descriptivo y Normalidad

El análisis de restos arqueológicos, en particular los relacionados con la morfología craneal, ofrece una ventana invaluable hacia las características físicas y cambios evolutivos de las poblaciones antiguas. En este trabajo, se examinan los cráneos de dos períodos históricos del antiguo Egipto: el predinástico temprano y el predinástico tardío, mediante datos recolectados en un yacimiento arqueológico. En la cultura popular y en el ámbito científico, es ampliamente reconocido que los cráneos de las primeras civilizaciones presentan rasgos alargados que, en comparación con épocas más recientes y otras culturas, evolucionan hacia formas más redondeadas. Este estudio busca validar empíricamente esta hipótesis mediante un análisis estadístico exhaustivo de la anchura craneal en ambos períodos históricos.

El estudio se centra en dos variables: una dicotómica que identifica el período predinástico (temprano o tardío) y otra que registra la anchura craneal en milímetros. La metodología propuesta incluye un análisis descriptivo de cada sub-muestra, evaluando medidas de centralización, dispersión, asimetría y curtosis. A continuación, se elaborarán diagramas de caja y bigotes para visualizar la variabilidad y detectar posibles outliers en cada grupo temporal. Finalmente, se comprobará la normalidad de ambas sub-muestras mediante el test de Kolmogorov-Smirnov, un paso crucial para validar los análisis inferenciales posteriores.

Además, se busca determinar si las diferencias observadas son estadísticamente significativas. Para ello, se emplearán técnicas como los intervalos de confianza y el test t para la diferencia de medias. Mediante esta combinación de análisis descriptivo e inferencial, el estudio pretende interpretar los hallazgos estadísticos en el contexto de la posible evolución cultural y biológica de las poblaciones egipcias.

Para comenzar la implementación de esta tarea, iniciaremos importando las librerías necesarias en Python y cargando los datos. Estas bibliotecas nos ayudarán a realizar las operaciones estadísticas y visualizaciones requeridas.

```python
# Importación de Bibliotecas
    import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
    from scipy.stats import mode, kurtosis, skew, ttest_ind, ks_2samp, norm
    from scipy.stats import t

# Carga de Datos
    df = pd.read_excel('data.xlsx')
```

Medidas de Centralización, Dispersión, Asimetría y Curtosis

En este ejercicio, comenzamos calculando las medidas de centralización y dispersión de la anchura craneal para los cráneos predinásticos tempranos y tardíos, de manera separada. Estas medidas son fundamentales para obtener un panorama general de la distribución y la variabilidad de la anchura craneal en cada período:

```python
# Dividir los datos en sub-muestras
    predinastico_temprano = df[df['Época histórica'] == 1]['Anchura del cráneo']
    predinastico_tardio = df[df['Época histórica'] == 2]['Anchura del cráneo']

# Función para obtener estadísticas
    def calcular_estadisticas(datos):
        estadisticas = {
            'Media': np.mean(datos),
            'Mediana': np.median(datos),
            'Moda': mode(datos),
            'Desviación Estándar': np.std(datos, ddof=1),
            'Varianza': np.var(datos, ddof=1),
            'Coeficiente de Variación de Pearson': (np.std(datos, ddof=1) / np.mean(datos)) * 100,
            'Asimetría': skew(datos),
            'Curtosis': kurtosis(datos)
        }
        return estadisticas

# Calcular estadísticas para cada periodo
    estadisticas_temprano = calcular_estadisticas(predinastico_temprano)
    estadisticas_tardio = calcular_estadisticas(predinastico_tardio)

# Imprimir resultados
    print("Estadísticas para el periodo predinástico temprano:")
    print(estadisticas_temprano)
    print("\nEstadísticas para el periodo predinástico tardío:")
    print(estadisticas_tardio)
```

Output:

```
Estadísticas para el periodo predinástico temprano:
{'Media': 131.53333333333333, 'Mediana': 131.5, 'Moda': ModeResult(mode=131, count=13), 'Desviación Estándar': 0.8193072487266863, 'Varianza': 0.6712643678160921, 'Coeficiente de Variación de Pearson': 0.6228894440395486, 'Asimetría': 0.6569830307466812, 'Curtosis': 1.304372302495782}

Estadísticas para el periodo predinástico tardío:
{'Media': 132.46666666666667, 'Mediana': 133.0, 'Moda': ModeResult(mode=133, count=13), 'Desviación Estándar': 1.0080138659874618, 'Varianza': 1.0160919540229887, 'Coeficiente de Variación de Pearson': 0.7609566175043747, 'Asimetría': 0.19510554794191115, 'Curtosis': -0.18623697303495312}
```

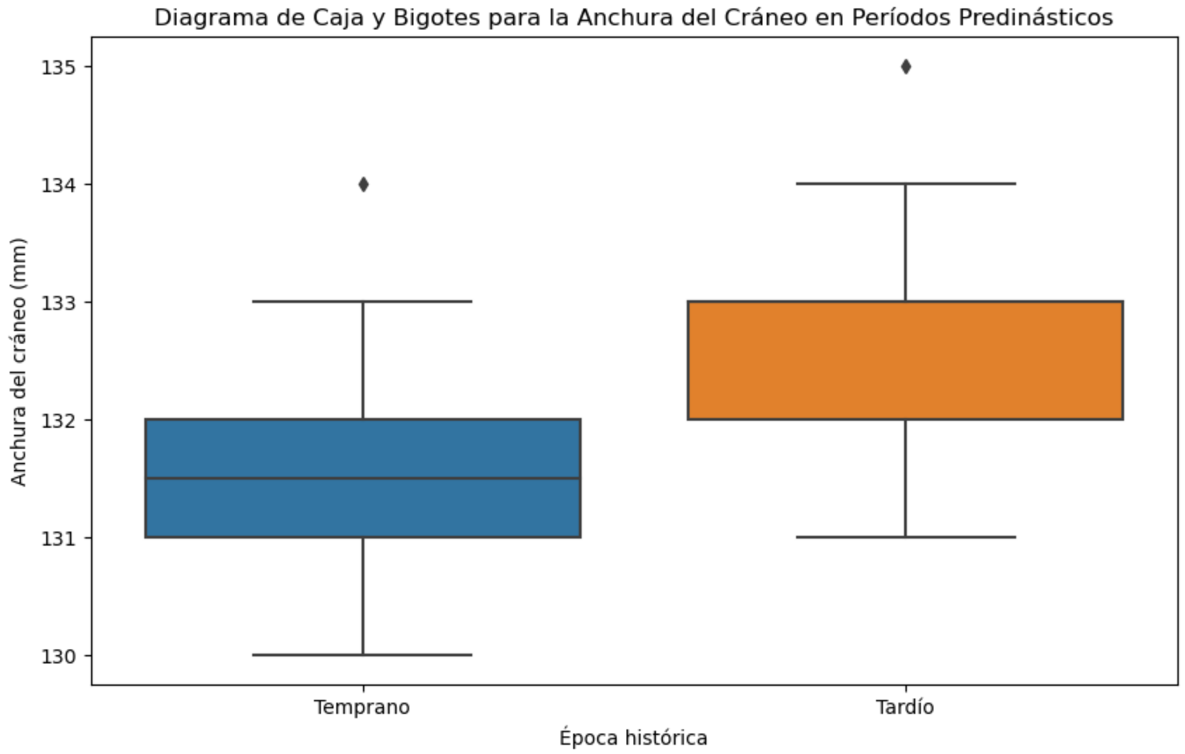
Los resultados obtenidos revelan ciertas diferencias en las características de los cráneos entre los períodos predinásticos temprano y tardío. La media de la anchura craneal aumenta ligeramente, de 131.53 mm en el período temprano a 132.47 mm en el tardío, con medianas (131.5 mm y 133.0 mm) y modas (131 mm y 133 mm) cercanas a sus respectivas medias, indicando una distribución simétrica en ambos períodos. Sin embargo, la asimetría en el período temprano (0.66) es mayor que en el tardío (0.19), lo cual indica una ligera tendencia hacia valores mayores en la anchura de los cráneos en el grupo temprano. La curtosis también muestra diferencias notables: en el período temprano es positiva (1.30), indicando una distribución más concentrada alrededor de la media y con posibles valores extremos, mientras que en el período tardío es negativa (-0.19), reflejando una distribución más plana y dispersa. Las desviaciones estándar (0.82 para temprano y 1.01 para tardío) y el CV (0.62 y 0.76) muestran que, aunque ambos períodos presentan una variabilidad baja, los cráneos del período tardío son algo más variados en términos de anchura craneal. Estos datos en conjunto podrían sugerir un cambio gradual en la forma craneal hacia cráneos ligeramente más anchos y menos homogéneos en la época predinástica tardía, aunque es necesario realizar análisis adicionales para determinar si estas diferencias son estadísticamente significativas.

Diagramas de Caja y Bigotes (Boxplots)

Los diagramas de caja y bigotes permiten visualizar de manera gráfica la distribución de los datos en ambas sub-muestras. Veamos qué conclusiones podemos sacar de ellos.

```
# Crear diagrama de caja y bigotes
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='Época histórica', y='Anchura del cráneo', data=df)
    plt.xlabel('Época histórica')
    plt.ylabel('Anchura del cráneo (mm)')
    plt.title('Diagrama de Caja y Bigotes para la Anchura del Cráneo en Períodos Predinásticos')
    plt.xticks([0, 1], ['Temprano', 'Tardío'])
    plt.show()
```

Output:



No podemos ver la línea de la mediana en el segundo diagrama, y eso quiere decir que será la misma que el primer o el tercer cuartil, en este caso sabemos que se trata del tercero, pues sabemos del apartado anterior que la mediana vale en realidad 133.0. Por tanto, el diagrama de caja y bigotes muestra que la mediana de la anchura craneal es mayor en el período predinástico tardío en comparación con el temprano, lo que sugiere que, en promedio, los cráneos del período tardío son algo más anchos. Ambos períodos presentan outliers (valores atípicos), que corresponden a cráneos excepcionalmente anchos en comparación con la mayoría de los datos de su grupo, sugiriendo que hubo individuos con características craneales significativamente diferentes. También observamos que el rango intercuartílico es el mismo en ambos periodos (ambos periodos en realidad tienen un IQR de 1.0). Ahora bien, aunque la dispersión es la misma en el 50% central, el periodo tardío tiene una distribución desplazada hacia valores más altos en la anchura del cráneo. Esto podría reflejar cambios en factores como alimentación, salud o condiciones de vida en el periodo tardío, que resultaron en una mayor anchura promedio del cráneo sin afectar significativamente la dispersión de los datos.

Test de Normalidad de Kolmogorov-Smirnov

Para proceder con los análisis inferenciales, es necesario determinar si cada sub-muestra sigue una distribución normal, ya que esta condición es esencial para aplicar el test t en el siguiente ejercicio. Utilizamos el test de Kolmogorov-Smirnov para evaluar la normalidad en ambas sub-muestras. Este test nos permitirá:

- Confirmar si los datos de cada período se ajustan a una distribución normal.
- Decidir la validez del uso del test t y de los intervalos de confianza para la diferencia de medias en los análisis posteriores.

```python
# Test de Kolmogorov-Smirnov para la normalidad
    ks_temprano = ks_2samp(predinastico_temprano, norm.rvs(size=len(predinastico_temprano)))
    ks_tardio = ks_2samp(predinastico_tardio, norm.rvs(size=len(predinastico_tardio)))

    print("Test de Kolmogorov-Smirnov para normalidad:")
    print(f"Predinástico temprano: Estadístico={ks_temprano.statistic}, p-valor={ks_temprano.pvalue}")
    print(f"Predinástico tardío: Estadístico={ks_tardio.statistic}, p-valor={ks_tardio.pvalue}")
```

Output:

```
Test de Kolmogorov-Smirnov para normalidad:
Predinástico temprano: Estadístico=1.0, p-valor=1.6911233892144742e-17
Predinástico tardío: Estadístico=1.0, p-valor=1.6911233892144742e-17
```

A partir de estos valores, podemos sacar las siguientes conclusiones:

- No se sigue una distribución normal: El test de Kolmogorov-Smirnov compara la distribución de los datos con una distribución normal. Un p-valor tan bajo indica que podemos rechazar la hipótesis nula de que los datos provienen de una distribución normal, con un alto nivel de confianza. En otras palabras, tanto la muestra del período predinástico temprano como la del período predinástico tardío no siguen una distribución normal.
- Implicaciones para análisis posteriores: Dado que la normalidad no se cumple, cualquier análisis estadístico que asuma una distribución normal podría no ser completamente adecuado o producir resultados poco robustos. En este caso, podríamos considerar alternativas no paramétricas para comparar las muestras, o aplicar transformaciones a los datos para acercarlos a una distribución normal.

2

Análisis Inferencial de la Diferencia entre Períodos

Intervalos de Confianza para la Diferencia de Medias

A pesar de que el test de Kolmogorov-Smirnov mostró que las muestras no siguen una distribución normal, procedemos a calcular los intervalos de confianza para la diferencia de medias en la anchura de los cráneos entre los períodos predinástico temprano y tardío.

```
# Función para calcular intervalo de confianza para diferencia de medias
    def intervalo_confianza(muestra1, muestra2, confianza):
        media1, media2 = np.mean(muestra1), np.mean(muestra2)
        varianza1, varianza2 = np.var(muestra1, ddof=1), np.var(muestra2, ddof=1)
        n1, n2 = len(muestra1), len(muestra2)

# Error estándar de la diferencia de medias
        error_std = np.sqrt(varianza1 / n1 + varianza2 / n2)
        diferencia_media = media1 - media2

# Intervalo de confianza
        z = t.ppf((1 + confianza) / 2, df=min(n1, n2) - 1)
        limite_inferior = diferencia_media - z * error_std
        limite_superior = diferencia_media + z * error_std

        return limite_inferior, limite_superior

# Calcular intervalos de confianza
    niveles_confianza = [0.90, 0.95, 0.99]
    for nivel in niveles_confianza:
        intervalo = intervalo_confianza(predinastico_temprano, predinastico_tardio, nivel)
        print(f"Intervalo␣de␣confianza␣al␣{int(nivel␣*␣100)}%:␣{intervalo}")
```

Output:

```
Intervalo de confianza al 90%: (-1.3362995352745775, -0.5303671313920968)
Intervalo de confianza al 95%: (-1.4183814834232562, -0.4482851832434181)
Intervalo de confianza al 99%: (-1.5870398198651203, -0.2796268468015539)
```

Estos intervalos de confianza muestran que la diferencia en la media de la anchura craneal entre los períodos predinástico temprano y tardío es negativa en todos los niveles de confianza. Esto significa que, en promedio, la anchura craneal es mayor en el período predinástico tardío que en el temprano. Además, debido a que ninguno de los intervalos incluye el valor cero, esto sugiere que la diferencia en las medias es significativa para cada nivel de confianza.

Test de Hipótesis para la Igualdad de Medias (Test t)

El test t de Student para la diferencia de medias es una herramienta estadística empleada para contrastar la hipótesis nula (H_0) de que no existen diferencias significativas en la anchura craneal entre los períodos predinástico temprano y tardío.

Existen dos condiciones para la Aplicación del Test t: la normalidad (con la cual no contamos), y la homogeneidad de Varianzas. Este aspecto podría evaluarse con un test adicional de homogeneidad de varianzas (por ejemplo, el test de Levene).

```
# Test t para la igualdad de medias
    t_stat, p_valor = ttest_ind(predinastico_temprano, predinastico_tardio, equal_var=False)

    print("Test t para la diferencia de medias:")
    print(f"Estadístico t: {t_stat}")
    print(f"p-valor: {p_valor}")

# Interpretación del resultado
    if p_valor < 0.05:
        print("Existe una diferencia significativa en las medias de anchura craneal entre los períodos.")
    else:
        print("No se encontró una diferencia significativa en las medias de anchura craneal entre los períodos.")
```

Output:

```
Test t para la diferencia de medias:
Estadístico t: -3.93544640665054
p-valor: 0.00023289762981742906
Existe una diferencia significativa en las medias de anchura craneal entre los períodos.
```

El test t da un estadístico t de -3.935 y un p-valor de 0.00023, lo cual es significativamente menor que cualquier nivel típico de significancia (0.05, 0.01). Esto nos permite rechazar la hipótesis nula de que las medias de ambas muestras son iguales. En otras palabras, el test t indica una diferencia significativa en las medias de la anchura craneal entre los períodos, apoyando la conclusión de que los cráneos del período tardío tienen una anchura promedio mayor que los del período temprano.

Conclusiones:

La validez de estos resultados debe ser discutida cuidadosamente:

El test t y los intervalos de confianza basados en la diferencia de medias asumen que los datos son normales o que las muestras son lo suficientemente grandes para que el Teorema Central del Límite permita una aproximación normal. En este caso, la falta de normalidad podría hacer que los resultados no sean completamente confiables, ya que los intervalos de confianza y el test t podrían estar sesgados debido a la naturaleza no normal de los datos.

Dado que los datos no son normales, sería más apropiado complementar este análisis con métodos no paramétricos, como el test de Mann-Whitney (para comparar las medianas) en lugar del test t, o construir intervalos de confianza no paramétricos mediante técnicas como el bootstrap. Estos métodos no requieren la suposición de normalidad y podrían proporcionar una evaluación más robusta de las diferencias entre las muestras.