

Casos de tuberculosis pulmonar

Gómez Pinzón Michelle Andrea - Ingeniería Financiera

1. Resumen

En este análisis se evaluó la distribución estadística de las variables **Género, Municipio de Residencia, Edad, Año, Semana y Nacionalidad**, provenientes del conjunto de datos **Casos de Tuberculosis Pulmonar**. El objetivo fue determinar si dichas variables siguen una distribución normal.

Se aplicaron pruebas de normalidad (Shapiro–Wilk, Kolmogorov–Smirnov, Anderson–Darling y Jarque–Bera) y se construyeron gráficos Q-Q para comparar cuantiles observados con teóricos.

2. Metodología

- Se extrajeron las variables de interés del dataset.
- Para cada variable, se aplicaron las pruebas:
 - Shapiro–Wilk
 - Kolmogorov–Smirnov
 - Anderson–Darling
 - Jarque–Bera
- Se construyeron gráficos Q-Q para verificar visualmente la normalidad.
- Se interpretaron los resultados utilizando un nivel de significancia $\alpha = 0.05$.
- Se analizó la naturaleza de cada variable y sus implicaciones.

3. Tratamiento de datos

Categorías: Nacionalidad, Género, Municipio de Residencia.

Numéricas: Edad, Año, Semana.

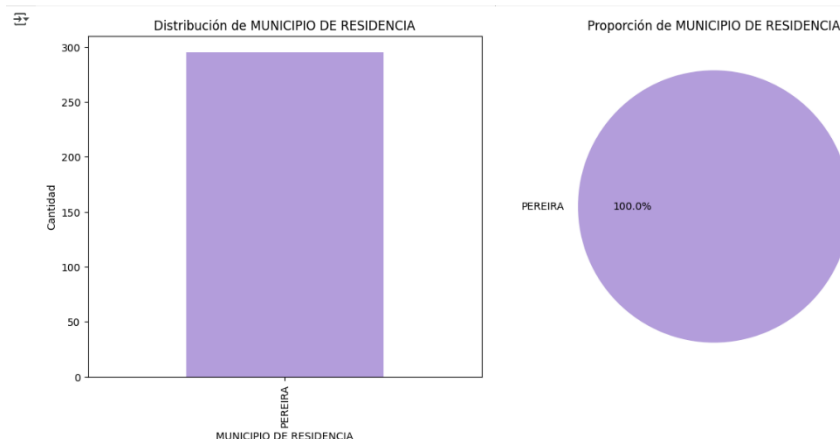
Google Collab

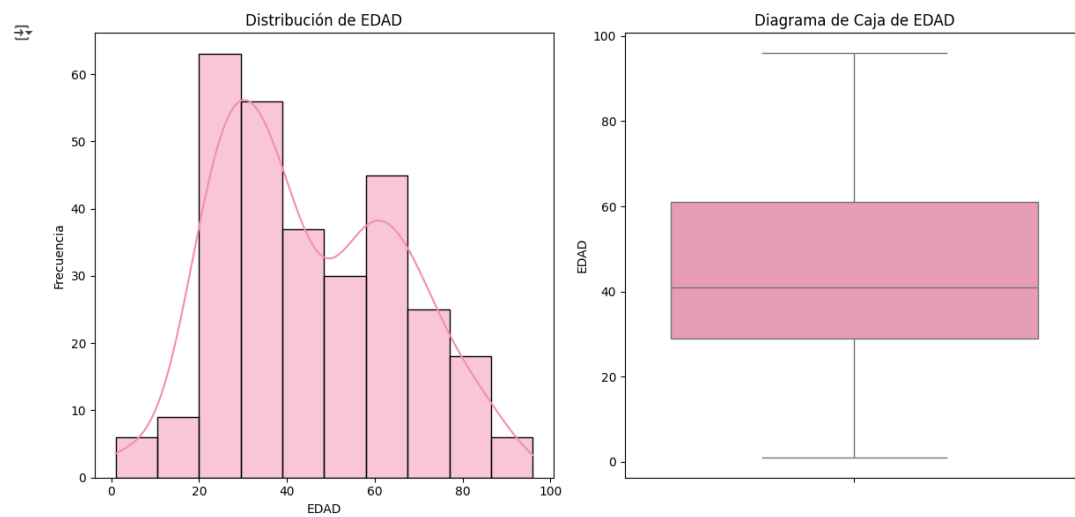
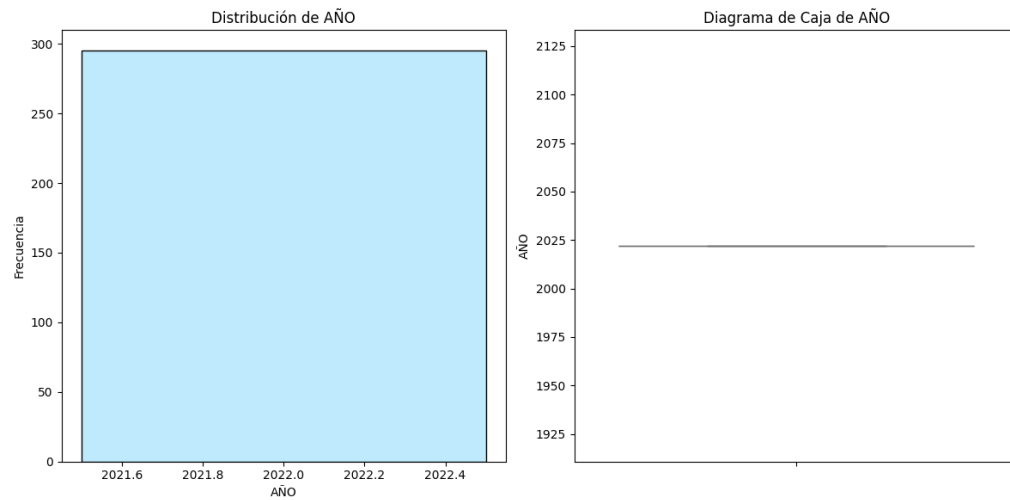
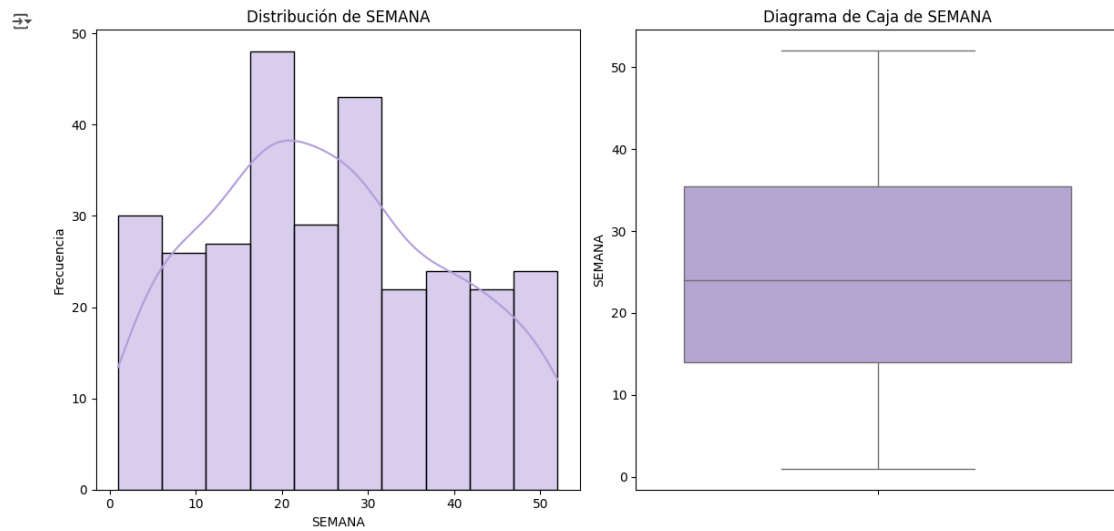
4. Análisis de resultados

```
Información del subset dataset
<bound method DataFrame.info of
0  FEMENINO      PEREIRA  64  2022  51  COLOMBIA
1  FEMENINO      PEREIRA  67  2022   6  COLOMBIA
2  FEMENINO      PEREIRA  42  2022  31  COLOMBIA
3  FEMENINO      PEREIRA  62  2022   5  COLOMBIA
4  FEMENINO      PEREIRA  63  2022   3  COLOMBIA
...
290 MASCULINO      PEREIRA  39  2022  38  COLOMBIA
291 MASCULINO      PEREIRA  21  2022  18  COLOMBIA
292 MASCULINO      PEREIRA  78  2022  18  COLOMBIA
293 MASCULINO      PEREIRA  71  2022  21  COLOMBIA
294 MASCULINO      PEREIRA  26  2022   8  VENEZUELA

[295 rows x 6 columns]>

Primeras del subset dataset
<bound method NDFrame.head of
0  FEMENINO      PEREIRA  64  2022  51  COLOMBIA
1  FEMENINO      PEREIRA  67  2022   6  COLOMBIA
2  FEMENINO      PEREIRA  42  2022  31  COLOMBIA
3  FEMENINO      PEREIRA  62  2022   5  COLOMBIA
4  FEMENINO      PEREIRA  63  2022   3  COLOMBIA
...
290 MASCULINO      PEREIRA  39  2022  38  COLOMBIA
291 MASCULINO      PEREIRA  21  2022  18  COLOMBIA
292 MASCULINO      PEREIRA  78  2022  18  COLOMBIA
293 MASCULINO      PEREIRA  71  2022  21  COLOMBIA
294 MASCULINO      PEREIRA  26  2022   8  VENEZUELA
```

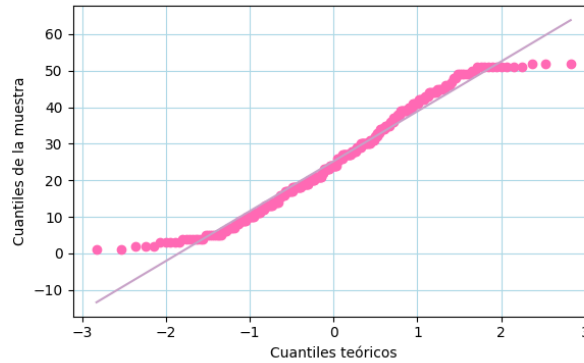






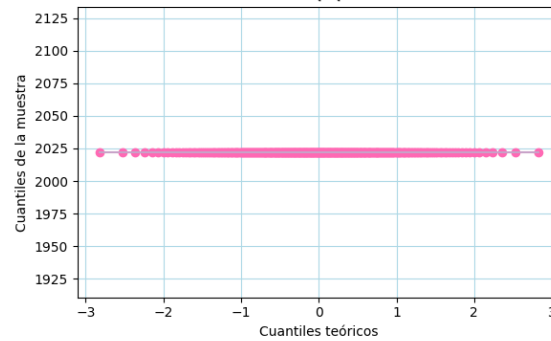
📊 Análisis de normalidad para: SEMANA
Shapiro-Wilk -> Estadístico: 0.9677, p-valor: 0.0000
Kolmogorov-Smirnov -> Estadístico: 0.0644, p-valor: 0.1656
Anderson-Darling -> Estadístico: 1.9159
Nivel 15.0% -> Valor crítico: 0.5680
Nivel 10.0% -> Valor crítico: 0.6470
Nivel 5.0% -> Valor crítico: 0.7770
Nivel 2.5% -> Valor crítico: 0.9060
Nivel 1.0% -> Valor crítico: 1.0780
Jarque-Bera -> Estadístico: 12.1511, p-valor: 0.0023

Gráfico Q-Q: SEMANA

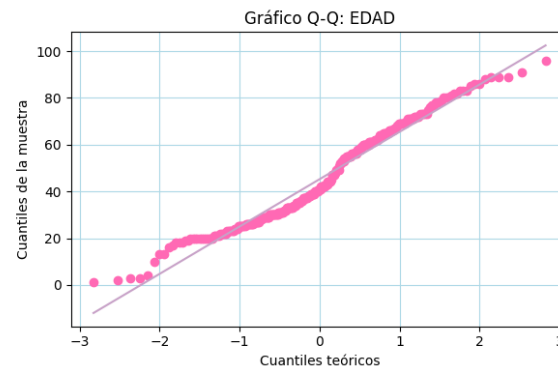


📊 Análisis de normalidad para: AÑO
Shapiro-Wilk -> Estadístico: 1.0000, p-valor: 1.0000
Kolmogorov-Smirnov -> Estadístico: nan, p-valor: nan
Anderson-Darling -> Estadístico: nan
Nivel 15.0% -> Valor crítico: 0.5680
Nivel 10.0% -> Valor crítico: 0.6470
Nivel 5.0% -> Valor crítico: 0.7770
Nivel 2.5% -> Valor crítico: 0.9060
Nivel 1.0% -> Valor crítico: 1.0780
Jarque-Bera -> Estadístico: nan, p-valor: nan

Gráfico Q-Q: AÑO



Análisis de normalidad para: EDAD
Shapiro-Wilk -> Estadístico: 0.9662, p-valor: 0.0000
Kolmogorov-Smirnov -> Estadístico: 0.0970, p-valor: 0.0072
Anderson-Darling -> Estadístico: 3.9278
Nivel 15.0% -> Valor crítico: 0.5680
Nivel 10.0% -> Valor crítico: 0.6470
Nivel 5.0% -> Valor crítico: 0.7770
Nivel 2.5% -> Valor crítico: 0.9060
Nivel 1.0% -> Valor crítico: 1.0780
Jarque-Bera -> Estadístico: 12.4451, p-valor: 0.0020



En el gráfico de **EDAD**, se observa una marcada curvatura en forma de S, que se aleja notablemente de la línea diagonal esperada bajo normalidad. Esta forma sugiere una distribución con colas pesadas y posible bimodalidad. En cuanto al gráfico de **AÑO**, se presenta una línea completamente horizontal en 2022, lo que confirma que todos los registros pertenecen al mismo periodo, sin variabilidad temporal. Por último, el gráfico de **SEMANA** muestra una curvatura suave con desviaciones moderadas respecto a la línea teórica, lo cual sugiere una distribución aproximadamente uniforme de los casos a lo largo del año. Sin embargo, las leves desviaciones en los extremos del gráfico podrían indicar ciertas variaciones estacionales.

5. Conclusiones

La existencia de casos en ambos géneros facilita la identificación de posibles variaciones en la distribución de la tuberculosis por género

Todos los casos analizados pertenecen al municipio de Pereira, lo que permite un enfoque geográfico focalizado.

Las variables numéricas *edad* y *semana* no presentan una distribución normal, por lo que se recomienda el uso de métodos estadísticos no paramétricos para el análisis de los datos.

La posible bimodalidad en la distribución de *edad* sugiere que los grupos de mayor riesgo podrían ubicarse en los extremos etarios, particularmente en niños pequeños y adultos mayores.

6. Referencias bibliográficas

Presidencia de la República de Colombia. (2022). *Casos de tuberculosis en Colombia - 2022* [Conjunto de datos]. Datos.gov.co. <https://www.datos.gov.co/>