# **Mariana Suárez Pacheco**

Práctica 1 Estadística Inferencial y muestreo

Profesor: Andrés Fabián Leal Archisla

Cuarto semestre - Marketing

Agosto 7 de 2025

**Informe Base de Datos de Nacimientos**

**1. Resumen**

Este informe presenta los resultados del análisis exploratorio realizado sobre una base de datos que contiene información sobre nacimientos, incluyendo características demográficas de los padres, características del nacimiento y datos de ubicación geográfica. El análisis permitió identificar patrones importantes en los datos, valores faltantes significativos y la necesidad de aplicar métodos estadísticos específicos para el análisis posterior.

**2. Descripción del Dataset**

La base de datos analizada contiene registros de nacimientos con las siguientes características principales:

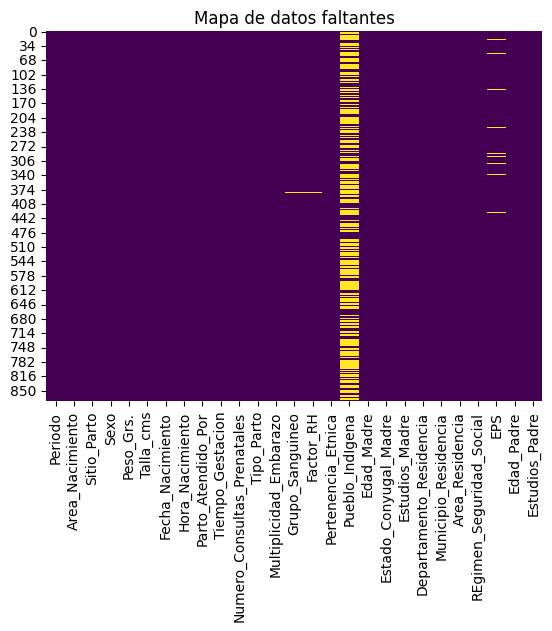
* **Información demográfica**: Datos sobre la madre y el padre
* **Características del nacimiento**: Peso, talla y sexo del recién nacido
* **Información geográfica**: Municipio de residencia
* **Datos médicos**: Grupo sanguíneo, factor RH, EPS

**3. Análisis de Calidad de Datos**

**3.1 Valores Faltantes**

Se identificaron valores faltantes en las siguientes variables:

* **Grupo\_Sanguíneo**: Presencia de valores nulos
* **Factor\_RH**: Presencia de valores nulos
* **Pueblo\_Indígena**: **Número significativo** de valores faltantes
* **EPS**: Presencia de valores nulos



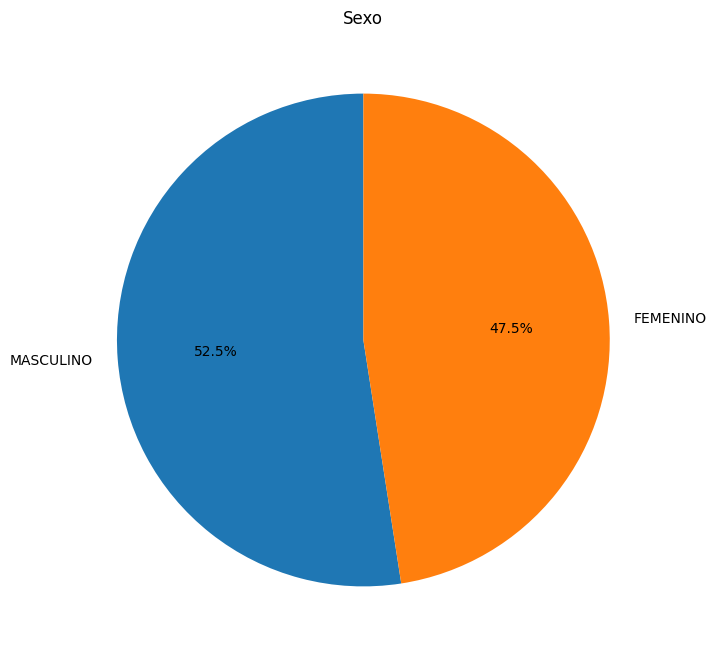
**4. Análisis Descriptivo**

Se eligieron 6 variables específicas para analizar: 3 categóricas y 3 númericas que se revisarán a continuación:

**4.1 Variables Categóricas**

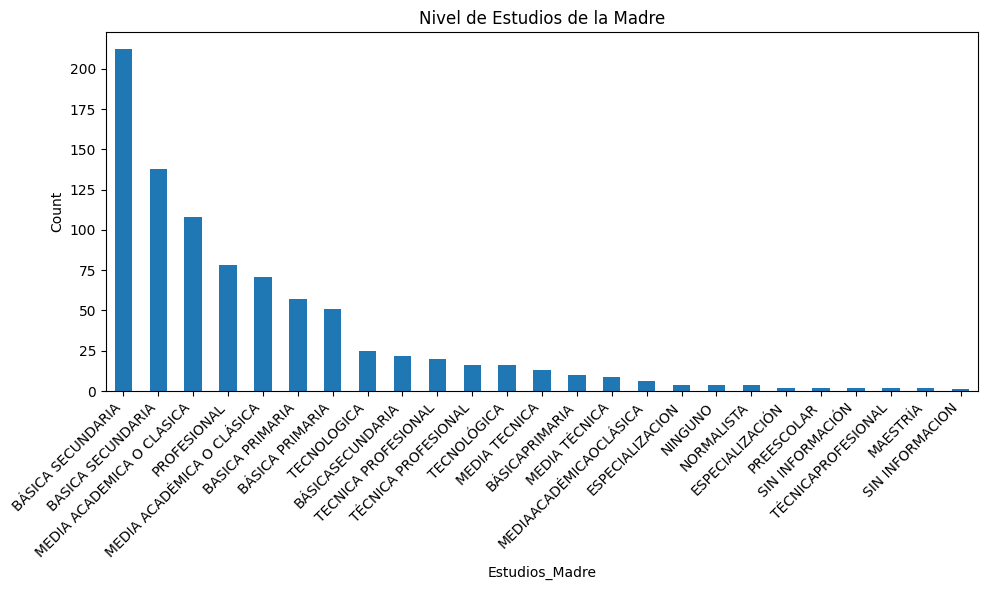
**Distribución por Sexo**

Se observa una **distribución relativamente equilibrada** entre nacimientos masculinos y femeninos, lo cual es consistente con patrones demográficos normales.



**Nivel Educativo de las Madres**

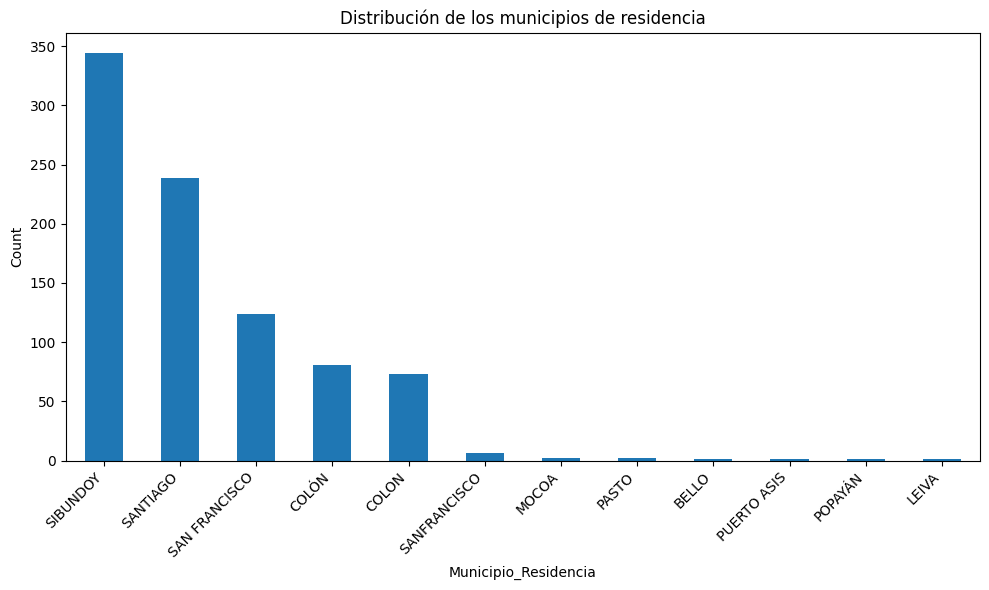
* **"BÁSICA SECUNDARIA"** es el nivel educativo más frecuente entre las madres
* Existe variabilidad en los niveles educativos, lo que sugiere diversidad socioeconómica en la población estudiada



**Distribución Geográfica**

Los nacimientos se concentran principalmente en tres municipios:

1. **SIBUNDOY**
2. **SANTIAGO**
3. **SAN FRANCISCO**



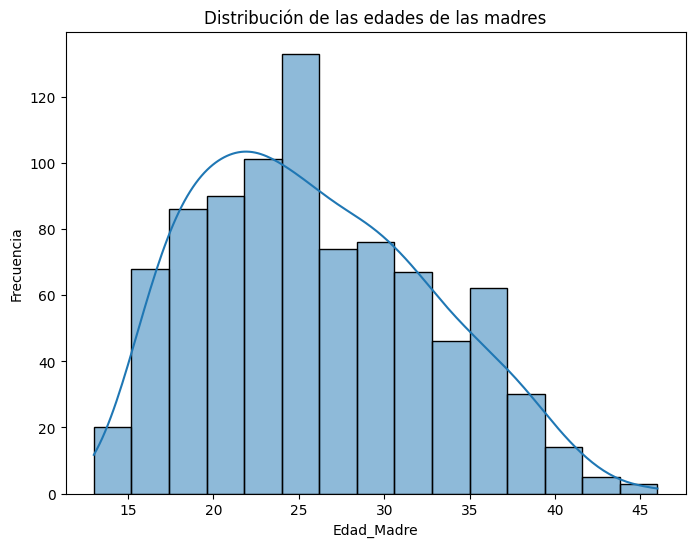
**4.2 Variables Numéricas**

Se analizaron tres variables numéricas clave:

**4.2.1. Histogramas**

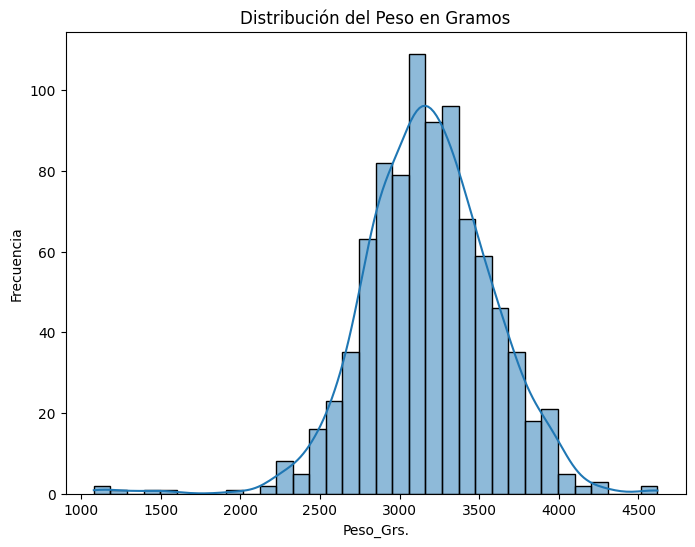
* **Edad\_Madre**: Edad de la madre al momento del nacimiento

La distribución muestra un patrón típico de edad reproductiva, donde la mayoría de nacimientos se concentran en la edad óptima de fertilidad (20-30 años).

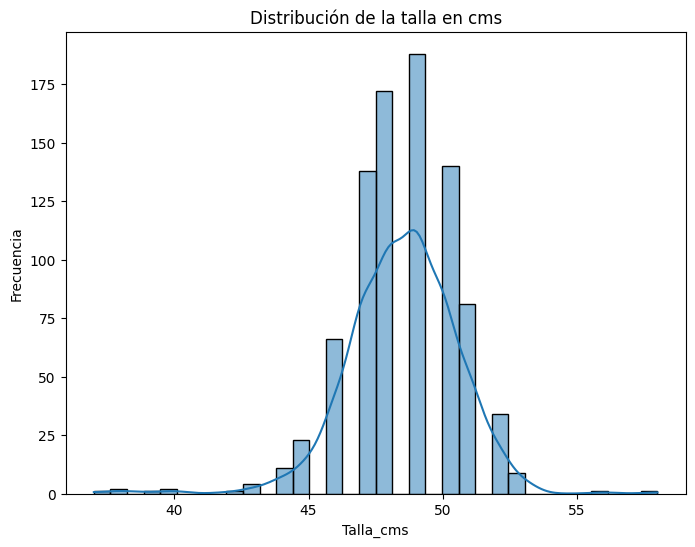


* **Peso\_Grs**: Peso del recién nacido en gramos

La distribución del peso presenta la forma más cercana a una distribución normal entre las tres variables analizadas. El peso promedio se encuentra dentro de rangos considerados normales para recién nacidos, con relativamente pocos casos en los extremos (bajo peso o sobrepeso al nacer).

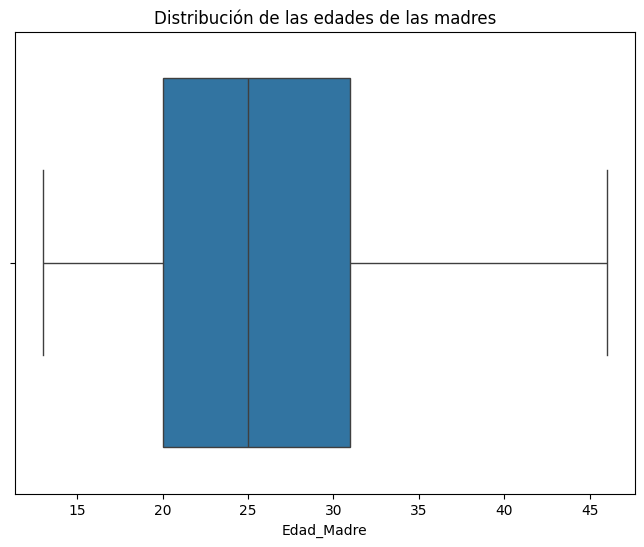


* **Talla\_cms**: Talla del recién nacido en centímetros

La talla muestra una distribución más concentrada alrededor del valor central, indicando menor variabilidad en esta medida.

## **4.2.2. Análisis de Diagramas de Caja (Box Plots)**

### **Edad de las Madres (Edad\_Madre)**

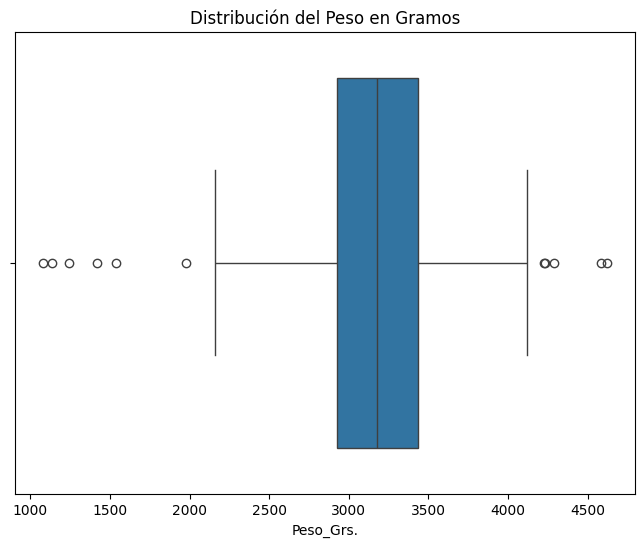


Características del diagrama de caja:

* Mediana: Aproximadamente 25-26 años
* Rango intercuartílico (IQR): Caja compacta entre ~21 y ~30 años
* Simetría: Distribución relativamente simétrica en el IQR
* Bigotes: Se extienden hasta aproximadamente 15 años (inferior) y 45 años (superior)
* Valores atípicos: Ausentes después del filtrado

La distribución filtrada muestra una edad materna concentrada en el rango reproductivo óptimo, sin casos extremos que pudieran indicar embarazos de muy alto riesgo.

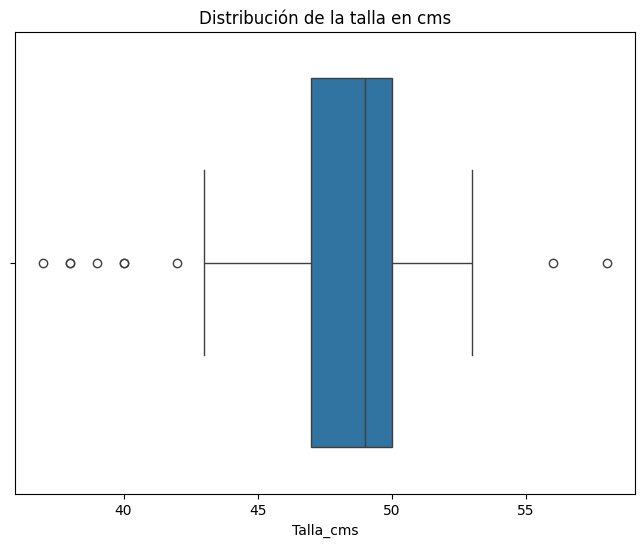
### **Peso de los Recién Nacidos (Peso\_Grs)**



Características del diagrama de caja:

* Mediana: Aproximadamente 3,200-3,300 gramos
* Rango intercuartílico: Caja entre ~2,900 y ~3,600 gramos
* Simetría: Excelente simetría en la caja y bigotes
* Bigotes: Rango de ~2,200 a ~4,200 gramos aproximadamente
* Valores atípicos: Múltiples outliers presentes en ambos extremos
  + Outliers inferiores: Casos de bajo peso (< 2,200g)
  + Outliers superiores: Casos de sobrepeso al nacer (> 4,200g)

### **Talla de los Recién Nacidos (Talla\_cms)**



Características del diagrama de caja:

* Mediana: Aproximadamente 49 centímetros
* Rango intercuartílico: Caja compacta entre ~47 y ~51 centímetros
* Simetría: Distribución simétrica en el IQR
* Bigotes: Rango de ~42 a ~56 centímetros aproximadamente
* Valores atípicos: Pocos outliers en ambos extremos
  + Outliers inferiores: Casos de talla baja (< 42cm)
  + Outliers superiores: Casos de talla alta (> 56cm)

**5. Tratamiento de Valores Atípicos**

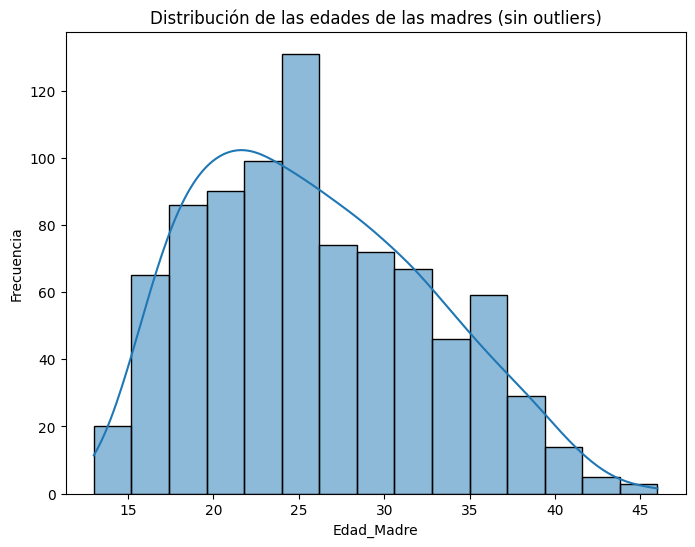
Se aplicó el **método del Rango Intercuartílico (IQR)** para identificar y remover valores atípicos en las tres variables numéricas principales.

* Se creó un dataset filtrado (df\_filtered) libre de valores extremos
* La remoción de outliers mejoró la calidad de las distribuciones
* Se ajustaron los rangos de valores a intervalos más realistas

**6. Análisis de Histogramas de las variables numéricas sin Outliers**

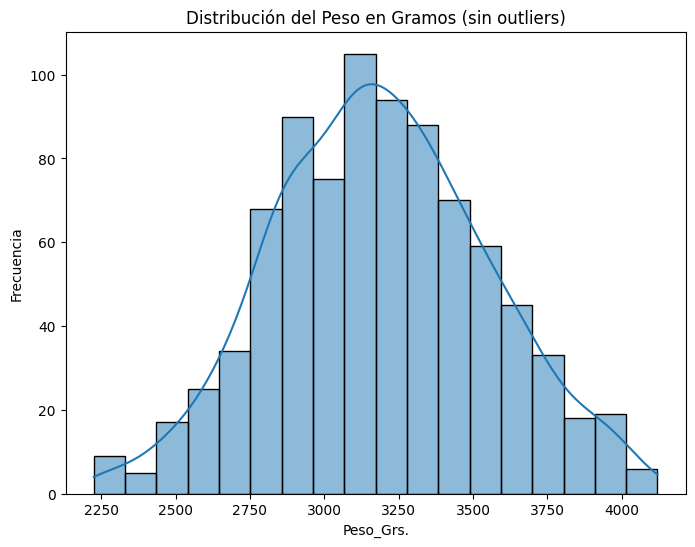
## **6.1. Edad de las Madres (Sin Outliers)**

* **Forma**: Asimétrica hacia la derecha, sin cambios significativos
* **Pico**: 25 años (~135 casos)
* **Rango**: 15-43 años
* El filtrado eliminó casos extremos pero **mantuvo la asimetría natural**
* Confirma que la no-normalidad es **característica poblacional**, no efecto de outliers
* Concentración en edad reproductiva óptima (20-30 años)



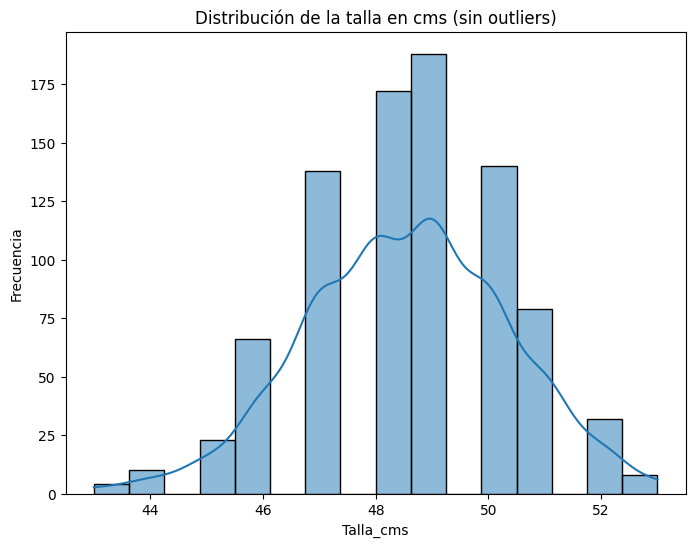
## **6.2. Peso en Gramos (Sin Outliers)**

* **Forma**: **Distribución perfectamente simétrica** tipo campana
* **Pico**: 3,100-3,200 gramos (~105 casos)
* **Rango**: 2,250-4,000 gramos
* Los outliers **enmascaraban la naturaleza normal** de la variable



## **6.3. Talla en Centímetros (Sin Outliers)**

* **Forma**: **Distribución bimodal** con dos picos distintos
* **Picos**: 48 cm (~170 casos) y 49 cm (~190 casos)
* **Rango**: 43-53 centímetros
* El filtrado descubrió estructura bimodal oculta



**7. Análisis de Normalidad**

**7.1 Pruebas Estadísticas Aplicadas**

Se realizaron múltiples pruebas de normalidad:

* **Shapiro-Wilk**
* **Kolmogorov-Smirnov**
* **Anderson-Darling**
* **Jarque-Bera**
* **Gráficos Q-Q**

**7.2 Resultados Detallados por Variable**

**Edad\_Madre**

* **Shapiro-Wilk**: Estadístico = 0.9707, p-valor = 0.0000
* **Kolmogorov-Smirnov**: Estadístico = 1.0000, p-valor = 0.0000
* **Anderson-Darling**: Estadístico = 6.9925
  + No normal en todos los niveles de significancia (15%, 10%, 5%, 2.5%, 1%)
* **Jarque-Bera**: Estadístico = 37.4323, p-valor = 0.0000

Todas las pruebas rechazan contundentemente la hipótesis de normalidad con p-valores extremadamente bajos (< 0.0001).

**Peso\_Grs**

* **Shapiro-Wilk**: Estadístico = 0.9965, p-valor = 0.0541
* **Kolmogorov-Smirnov**: Estadístico = 1.0000, p-valor = 0.0000
* **Anderson-Darling**: Estadístico = 0.4805
  + **Normal en todos los niveles de significancia** (15%, 10%, 5%, 2.5%, 1%)
* **Jarque-Bera**: Estadístico = 2.8115, p-valor = 0.2452

Las pruebas de Anderson-Darling y Jarque-Bera sugieren normalidad, mientras que Kolmogorov-Smirnov la rechaza fuertemente. Shapiro-Wilk está en el límite (p = 0.0541 > 0.05). Esta variable requiere **evaluación cuidadosa** antes de aplicar métodos paramétricos.

**Talla\_cms**

* **Shapiro-Wilk**: Estadístico = 0.9718, p-valor = 0.0000
* **Kolmogorov-Smirnov**: Estadístico = 1.0000, p-valor = 0.0000
* **Anderson-Darling**: Estadístico = 10.8802
  + No normal en todos los niveles de significancia (15%, 10%, 5%, 2.5%, 1%)
* **Jarque-Bera**: Estadístico = 2.7209, p-valor = 0.2565

Aunque Jarque-Bera sugiere normalidad (p = 0.2565), las pruebas más robustas (Shapiro-Wilk, K-S, Anderson-Darling) rechazan la normalidad. Se recomienda **asumir distribución no normal**.

**8. Análisis de los gráficos Q-Q**

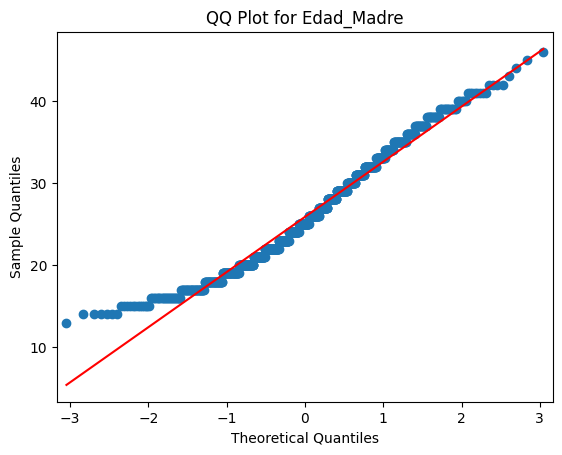
## **8.1. Q-Q Plot para Edad\_Madre**

**Descripción visual:**

* **Forma general**: Curva en "S" pronunciada
* **Extremo inferior**: Puntos **por debajo** de la línea roja (cola izquierda más pesada)
* **Centro**: Puntos **por encima** de la línea roja
* **Extremo superior**: Puntos **por debajo** de la línea roja (cola derecha más pesada)
* **Desviación**: **Sistemática y pronunciada** en toda la distribución

**Interpretación:**

* **No normalidad clara**: La curva en S es característica de distribuciones asimétricas
* **Sesgo positivo confirmado**: El patrón coincide con distribuciones sesgadas hacia la derecha
* **Colas pesadas**: Ambos extremos se desvían de la normalidad
* **Conclusión**: **Fuerte evidencia visual** de no normalidad



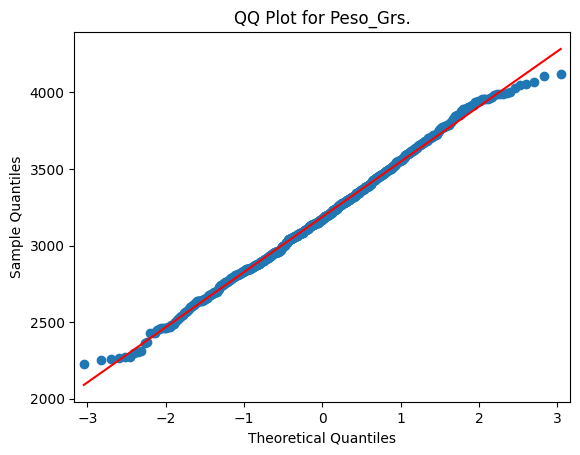
## **8.2. Q-Q Plot para Peso\_Grs**

**Descripción visual:**

* **Forma general**: Puntos **muy cercanos a la línea roja** en casi todo el rango
* **Alineación**: **Excelente seguimiento** de la línea teórica
* **Extremos**: Ligeras desviaciones menores en las colas
* **Centro**: **Alineación prácticamente perfecta**
* **Desviación**: **Mínima** comparada con las otras variables

**Interpretación:**

* **Normalidad visual**: Los puntos siguen muy de cerca la línea teórica
* **Confirmación del filtrado**: El proceso reveló efectivamente la naturaleza normal
* **Colas**: Las pequeñas desviaciones en extremos son normales y aceptables
* **Conclusión**: **Fuerte evidencia visual** de normalidad

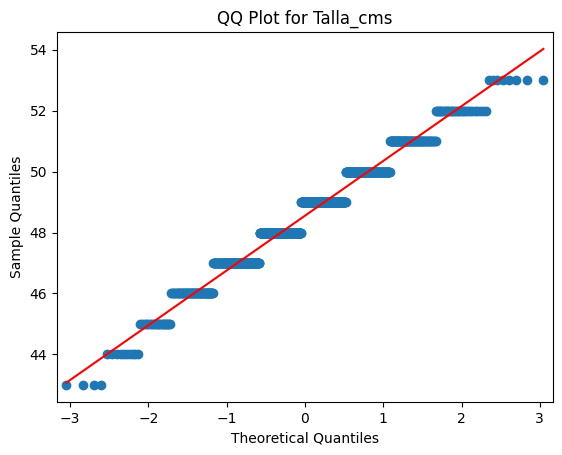


## **8.3. Q-Q Plot para Talla\_cms**

**Descripción visual:**

* **Forma general**: Patrón de **"escalones" o "mesetas"**
* **Estructura**: Grupos horizontales de puntos en diferentes niveles
* **Línea roja**: Los puntos se agrupan en **bandas horizontales** que no siguen la línea
* **Patrón distintivo**: **Discontinuidades claras** entre grupos de puntos
* **Distribución**: No sigue el patrón lineal esperado

**Interpretación:**

* **No normalidad evidente**: El patrón de escalones indica distribución no normal
* **Evidencia de discretización**: Los valores parecen estar agrupados en categorías
* **Posible bimodalidad**: El patrón escalonado puede reflejar la estructura bimodal observada
* **Medición discretizada**: Sugiere que la talla se registra en valores enteros o medios centímetros
* **Conclusión**: **Clara evidencia** de no normalidad y estructura compleja
* 

9. Análisis y conclusiones

* El conjunto de datos contiene información detallada sobre nacimientos, incluyendo datos demográficos de los padres, características del recién nacido y ubicación geográfica.
* Se identificaron valores faltantes, especialmente en *Pueblo\_Indígena*, lo que podría requerir tratamiento previo al análisis.
* La distribución de *Sexo* es equilibrada, el nivel educativo más común de las madres es “Básica Secundaria” y los nacimientos se concentran en Sibundoy, Santiago y San Francisco.
* Se detectaron y eliminaron valores atípicos en *Edad\_Madre*, *Peso\_Grs.* y *Talla\_cms* mediante el método IQR, obteniendo un conjunto de datos más limpio.
* Las pruebas de normalidad muestran que *Edad\_Madre* y *Talla\_cms* no siguen una distribución normal, y *Peso\_Grs.* presenta resultados mixtos, por lo que se recomienda considerar métodos no paramétricos o transformaciones de datos en el análisis futuro.