Estadística descriptiva de los nacidos vivos en el Hospital General de Medellín en el año 2021

Julián A. Angarita-Suárez*

2022-02-24

Abstract

En el siguiente artículo se presenta un análisis de estadística descriptiva de los nacidos vivos en el Hospital General de Medellín *(HGM)* en el año 2021. Para ello se describen las variables cualitativas y cuantitativas; para las cualitativas se lleva a cabo una distribución porcentual; para las cuantitativas se realiza un análisis de tendencia central; un análisis de dispersión; un análisis de posición y apuntalamiento para describir las principales características de su distribución y su variabilidad. Estos análisis nos arrojan que de manera marginal la mayoría de nacidos vivos en el HGM en el 2021 son de sexo masculino y tuvieron un parto de tipo espontáneo. Adicional a esto la mayor cantidad de los nacidos vivos tuvieron pesos y tallas inferiores a la media. En la sección de anexos se analiza la relación existente entre talla y peso y sus correspondientes características.

Palabras claves— estadística descriptiva, nacidos vivos, medidas de resumen.

*Estudiante del pregrado de estadística en la Universidad de Antioquia, julian.angarita@udea.edu.co

Introducción

En el siguiente texto se llevará a cabo una estadística descriptiva de las algunas características de los nacidos vivos en el Hospital General de Medellín en el año 2021. Para ello, se especificará, primero, de dónde provienen los datos, quién fue el encargado de llevar a cabor los registros, se presentarán las transformaciones hechas a los datos para los efectos del análisis, junto con otras transformaciones para depurar y alistar el conjunto de datos. También, se realiza una identificación de los casos que presentan valores perdidos, o no registrados, en alguna de sus características, y se presentan los métodos aplicados al análisis de las variables cuantativas. En la siguiente sección, se presentan los análisis y los resultados de las distribuciones de las variables cualitativas y cuantitativas que permiten describir el comportamiento y las características de los nacidos vivos. Por último, se presentan las principales conclusiones del trabajo. Y en la sección de anexos, se plantea un modelo de relación lineal entre la talla y el peso de los nacidos vivos.

Materiales y métodos

Fuente de los datos

La base de datos se obtuvo a través del portal de divulgación de datos abiertos del Estado de Colombia y se puede acceder a ellos dando click **aquí.** Contiene registros de todos los nacidos vivos en el Hospital General de Medellín en el año 2021. Dicha información fue recopilada por el Hospital.

Carga y transformación de los datos

El enlace del conjunto de datos se aloja en el objeto `ruta` para facilitar la carga de los datos. Posteriormente, en la creacion del objeto que contiene los datos (`bd_nacimientos`) se limpian los nombres de las variables para su posterior uso y se arregla el contenido de las variables tipo `character`. Adicionalmente, se descartan las variables cuyo nombre comienza con *apgar_*, pues en el diccionario de los datos no se especifica el contenido de estas variables. Lo anterior se hizo de la siguiente manera:

```
enlace <- "https://r-short.herokuapp.com/r1OWDogZq"

bd_nacimientos <- openxlsx::read.xlsx(enlace) %>%
  clean_names() %>%
```

```
  mutate_if(is.character, str_to_sentence) %>%
  select(-contains("apgar"), -n) %>%
  rename(fecha_nacimiento = fechana_cimiento)
```

Posteriormente, identificamos que las variables de la edad y la fecha de nacimiento contienen expresiones de tipo `character`. Para terminar de limpiar el conjunto de datos, extraemos el patrón numérico que es de nuestro interés, y damos formato tipo `date` a la variable de fecha de nacimiento, de la siguiente forma:

```
bd_nacimientos %<>%
  mutate(
    fecha_nacimiento = dmy(fecha_nacimiento),
    edad_padre = as.numeric(str_match(edad_padre, "\\d{2}")),
    edad_madre = as.numeric(str_match(edad_madre, "\\d{2}")),
    dia = str_to_title(wday(fecha_nacimiento,
      label = TRUE, abbr = FALSE,
      locale = "Spanish_Colombia.1252"
    )),
    mes = str_to_title(month(fecha_nacimiento,
      label = TRUE, abbr = FALSE,
      locale = "Spanish_Colombia.1252"
    ))
  ) %>%
  relocate(c(dia, mes), .after = fecha_nacimiento)
```

Resumen descriptivo del conjunto de datos

El conjunto de datos tiene 4559 filas y 29 columnas. Adicionalmente, la distribución por clase de las variables es la siguiente: 19 variables de tipo `categórico`, 9 de tipo `numérico`, y 1 de tipo `fecha`.

Para los propositos del curso, se selecciona un subconjunto de variables, donde hay 2 variables cualitativas, y 4 cuantitativas. Dentro de las cualitativas están:

- Sexo del nacido vivo;
- Tipo de parto.

Dentro del conjunto de cuantitativas, están:

- Peso del nacido vivo;
- Talla del nacido vivo;
- Tiempo de gestación de la madre.

- Edad de la madre.

```
selvar <- c(
  "sexo", "tipo_parto",
  "peso", "talla", "tiempo_gestacion",
  "edad_madre"
)


bd_subset <- bd_nacimientos %>%
  select(all_of(selvar))
```

Con relación a las variables cuantitativas, es necesario especificar la unidad de medida de cada una de ellas; las cuales son:

- La unidad de medida de la variable *peso del nacido vivo* es en **gramos**;
- La unidad de medida de la variable *talla del nacido vivo* es en **centímetros**;
- La unidad de medida de la variable *tiempo de gestación de la madre* es en **semanas**;
- La unidad de medida de la variable *edad de la madre* es en **años**.

En cuanto a la presencia de valores perdidos en el conjunto de datos, es de notar que sólo hay 1 registro que presenta ésta situación en una de las variables. La cantidad de valores pérdidos es detectada por medio del siguiente script `sum(is.na(bd_subset))`. Este valor perdido tiene como etiqueta `NA`; la cual significa no disponible (not available) y es asignada por defecto en R al detectar la ausencia del dato en la coordenada (m_i, n_i) del conjunto de datos, donde **m** referencia las filas y **n** las columnas.

Adicionalmente, identificamos e individualizamos el registro que presenta esta condición para conocer las características que tiene, y cuál es la variable que tiene el valor perdido.

```
bd_subset %>%
  filter_all(any_vars(is.na(.))) %>%
  kable(
    booktabs = TRUE,
    format = "latex",
    align = "c",
    caption = "Registros con valores perdidos."
  ) %>%
  footnote("elaboración propia.",
    footnote_as_chunk = T,
    title_format = c("italic", "underline"), general_title = "Fuente:"
  ) %>%
  kable_styling(latex_options = "HOLD_position")
```

Tabla 1: Registros con valores perdidos.

sexo	tipo_parto	peso	talla	tiempo_gestacion	edad_madre
Masculino	Espontáneo	3160	48	NA	28

Fuente: elaboración propia.

La `tabla 1` nos permite reconocer que el caso que presenta un valor perdido es el registro de un nacido vivo de sexo masculino, en el cual no se cuenta con el tiempo de gestación que tuvo la madre previo al parto.

Apuntes metodológicos y analíticos

Después de procesar y describir el conjunto de datos, se realiza una descripción de la población de los nacidos vivos, en el que se lleva a cabo el estudio de la distribución porcentual del `sexo`y el `tipo de parto` para los nacidos vivos. Adicionalmente, se realizará una distribución porcentual del `tipo de parto` por `sexo` de los nacidos vivos. En cuanto a las variables de `talla`, `peso`, `tiempo de gestación` y `edad de la madre`, se describirá su distribución; es decir, se analizará la forma en como se distribuyen los datos, su inclinación, su variación, tanto a nivel general como por el `sexo` de los nacidos vivos.

Con relación a las medidas de tendencia central usados en este artículo, por tener la condición de ser el registro total de los nacidos vivos en un espacio y tiempo determinado, se usara la media poblacional, la mediana y la moda, siendo la primera definida como:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

La mediana calculada es la correspondiente para conjuntos de datos impares, después de ordenar de menor a mayor los datos, es definida como:

$$m_e = \frac{n+1}{2}$$

En cuanto a las medidas de dispersión, se usará la desviación estándar poblacional, que se define como:

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

Y el rango, que se define como:

$$R = x_{max} - x_{min}$$

Por último, se usarán las medidas de forma para describir el comportamiento de las variables `peso` y `talla` de los nacidos vivos; las cuales son: el coeficiente de asimetria y el coeficiente de curtosis.

El primero es definido como:

$$CA = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3 N_i}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 N_i)^{\frac{3}{2}}}$$

En cuanto al coeficiente de curtosis, es definido como:

$$CC = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4 N_i}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 N_i)^2} - 3$$

Resultados y análisis

En esta sección se presentarán los resultados de la descripción de los nacidos vivos en el Hospital General de Medellín *(HGM)* en el año 2021. Para ello, presentaremos algunas gráficas y tablas que permitan visualizar los datos de una manera clara.

Distribuciones de la población de los nacidos vivos

De los nacidos vivos en el Hospital General de Medellín en el año 2021, como se aprecia en la `figura 1`, el 51.6% es de sexo masculino y el 48.4% es de sexo femenino; en cuanto al tipo de parto que tuvieron, como se aprecia en la `figura 2`, el 73.8% de los partos fue de manera espontánea, y, con relación a los partos asistidos, el 22.3% por cesárea, y, de manera muy marginal, el 3.9% fue instrumentado; es decir, un parto asistido por medio de ventosas o fórceps.

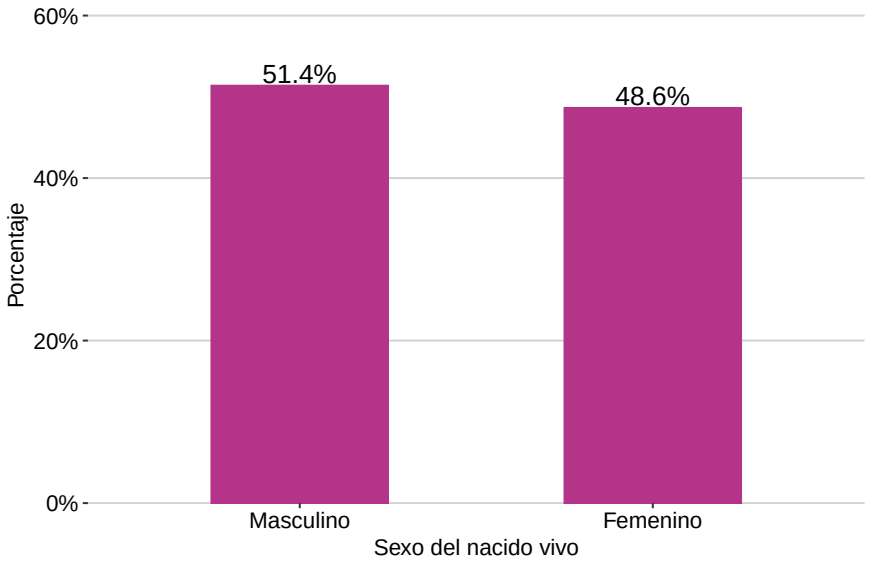

Figura 1: Distribución por sexo de los nacidos vivos en el Hospital General de Medellín en el año 2021.

En cuanto a la distribución del tipo de parto según el sexo del nacido vivo, se observan diferencias porcentuales marginales; pues para el 75.6% de los nacidos vivos de sexo femenino y para el 72.1% de los nacidos vivos de sexo masculino, el parto fue espontáneo, presentándose una diferencia de + 3.5 puntos porcentuales del sexo femenino respecto al masculino. En cuanto a los partos asistidos, para el 23.6% de los nacidos vivos de sexo masculino y para el 20.9% de sexo femenino, el parto fue por cesárea, presentándose una diferencia de -2.7 puntos porcentuales del sexo femenino respecto al masculino. Y, por último, para el 4.3%

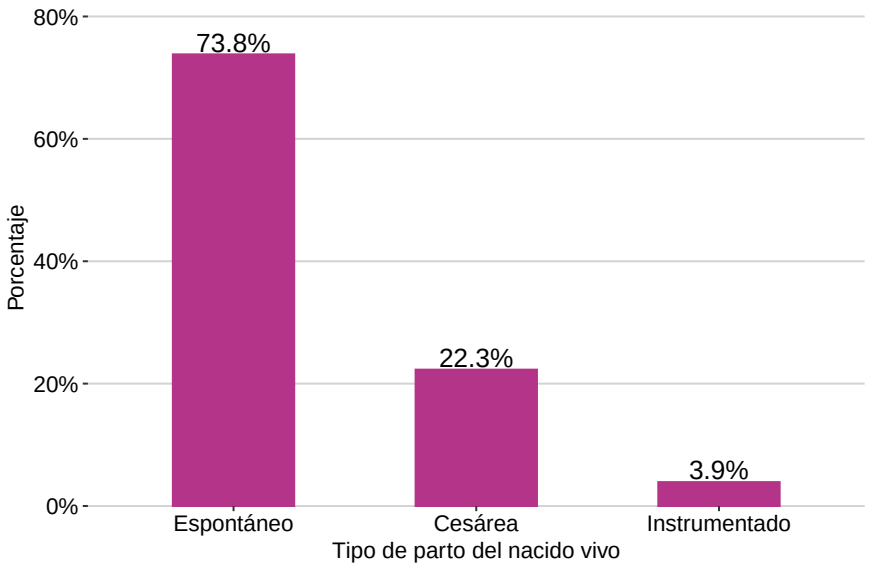

Figura 2: Distribución por tipo de parto de los nacidos vivos en el Hospital General de Medellín en el año 2021.

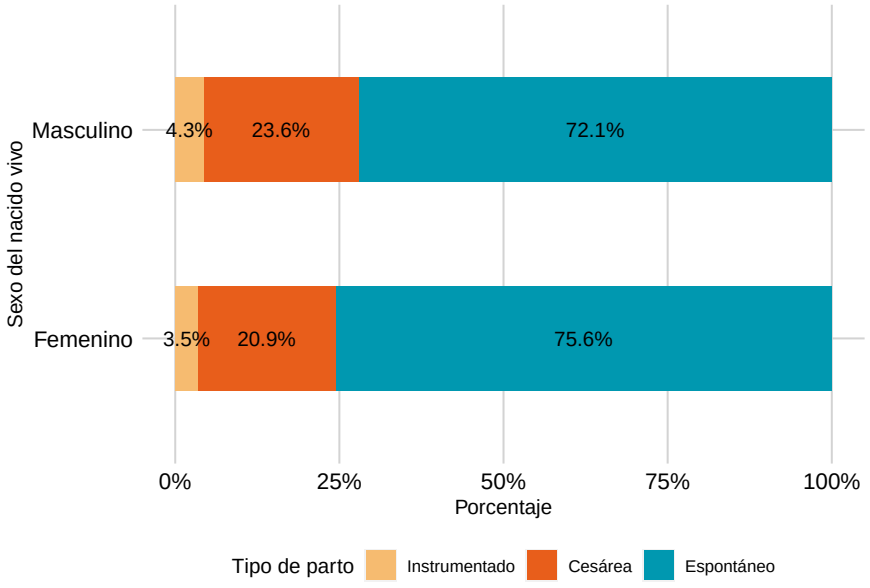

Figura 3: Distribución por tipo de parto según el sexo del nacido vivo en el Hospital General de Medellín en el año 2021.

de los masculinos, y el 3.5% de los femeninos, el parto fue instrumentado, presentándose una diferencia de +0.8 puntos porcentuales del sexo masculino respecto al sexo femenino (`ver figura 3`).

En lo que respecta al peso de los nacidos vivos en el Hospital General de Medellín para el año 2021, podemos observar que el peso mínimo registrado fue de 245 gramos, y el máximo fue de 4872 gramos. La media del peso de los nacidos vivos fue de 2941.93 gramos, la mediana fue de 3014 gramos, y la moda fue de 2890 gramos.

La desviación estandar del peso de los nacidos vivos fue de 594.61 gramos. Y la variación máxima presentada (rango) en el peso fue de 4627 gramos. Como se puede observar, la moda es inferior a la media y a la mediana, y la media es levemente superior a la moda pero levemente inferior a la media.

En lo relacionado al comportamiento de la curva de la distribución del peso, el valor del coeficiente de asimetría es -0.7836217 ($ca < 0$); lo cual nos indica que el sesgo de la curva es negativo, con una mayor concentración de los valores en la parte derecha superior a la media y un sesgo hacia la izquierda; esto también se puede apreciar por `sexo` como se evidencia en la `figura 6`.

En cuanto al apuntamiento de la curva que que concentra los valores en la zona central de la distribución, medido a través del coeficiente de curtosis, su valor de 1.2061648 ($ca > 0$) nos indica que la curva es leptocúrtica al presentarse una gran concentración de valores en la zona antes menciondada como se puede observar en la `figura 4`.

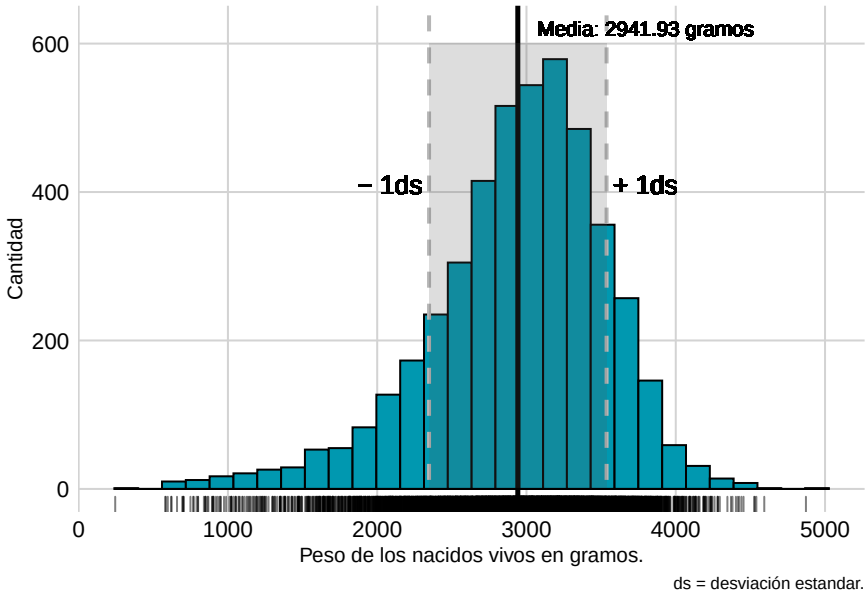

Figura 4: Distribucion del peso (en gramos) de los nacidos vivos.

En lo que respecta a la talla de los nacidos vivos en el Hospital General de Medellín para el año 2021, podemos observar que la talla mínima registrada fue de 22 centímetros, y la máxima fue de 58 centímetros. La media de la talla de los nacidos vivos fue de 47.59 centímetros, la mediana fue de 48 centímetros, y la moda fue de 48 centímetros.

La desviación estandar de la talla de los nacidos vivos fue de 3.17 centímetros. Y la variación máxima presentada (rango) en la talla fue de 36 centímetros. Como se puede observar, la media es inferior a la mediana y a la mediana, pero la mediana y la moda coinciden en su valor (48 cms).

En lo relacionado al comportamiento de la curva de la distribución de la talla, el valor del coeficiente de asimetría es -1.6818805 ($ca < 0$); lo cual nos indica que el sesgo de la curva es negativo, con una mayor concentración de los valores en la parte derecha superior a la media, con una mayor concentración de los valores en la parte derecha superior a la media y un sesgo hacia la izquierda; esto también se puede apreciar por `sexo` como se evidencia en la `figura 7`.

En cuanto al apuntamiento de la curva que que concentra los valores en la zona central de la distribución, medido a través del coeficiente de curtosis, su valor de 5.5494514 ($ca > 0$) nos indica que la curva es leptocúrtica al presentarse una gran concentración de valores en la zona antes menciondada como se puede observar en la `figura 5`.

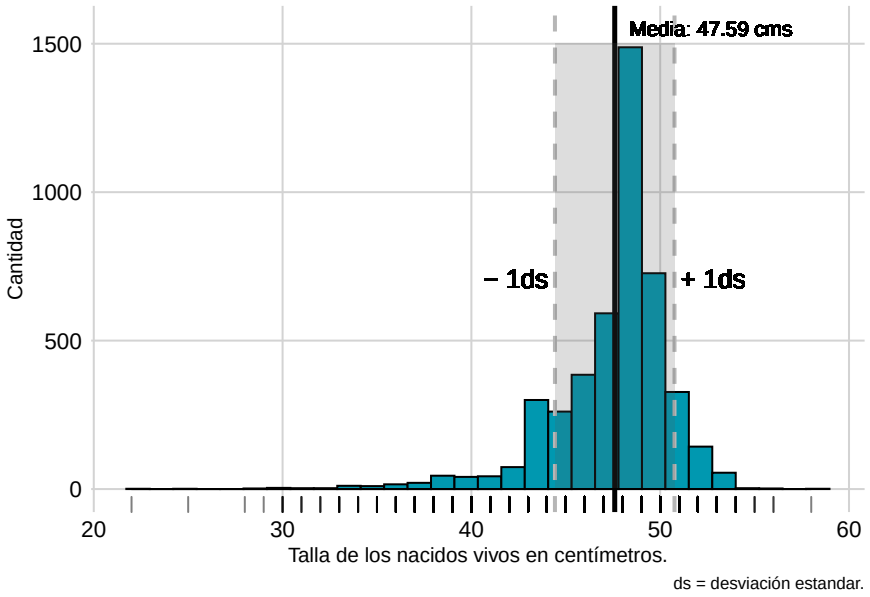

Figura 5: Distribucion de la talla (en centímetros) de los nacidos vivos.

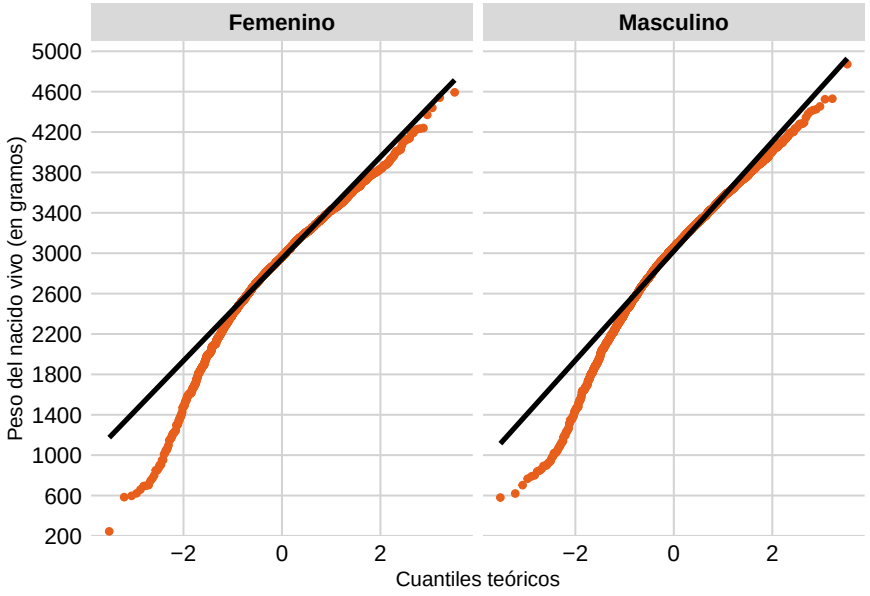

Figura 6: QQ-plot de la variable peso de los nacidos vivos según el sexo.

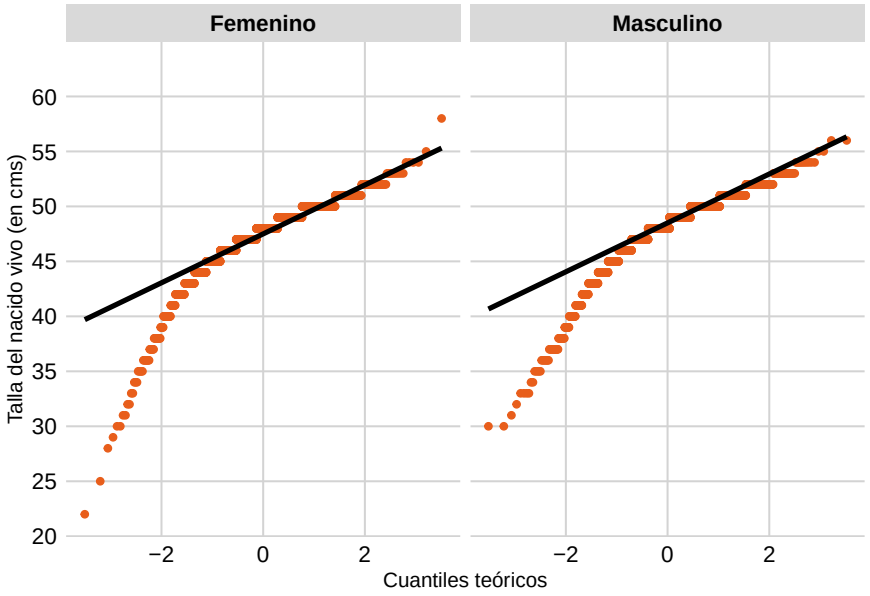

Figura 7: QQ-plot de la variable talla de los nacidos vivos según el sexo.

Por último, en lo relativo al tiempo de gestación de la madre según el sexo de los nacidos vivos, se presenta una mayor dispersión de las semanas de gestación en los nacidos vivos de género femenino (`ver figura 8`), pues el 50% de los registros tuvo entre 37 y 40 semanas de gestación, con una mediana de 38 semanas, y una distribución de las semanas con una asimetría hacia la izquierda; lo cual indica que la mayoría de registros de nacidos vivos tienen tiempos de gestación inferior a las 28 semanas. Una situación similar se presenta al observar el tiempo de gestación en los nacidos vivos de sexo masculino, pues, si bien el 50% de los registros tuvo entre 37 y 39 semanas de gestación, con una mediana de 28 semanas, también se presenta esta asimetría hacia la izquierda, es decir, la mayoría de registros de nacidos vivos tienen tiempos de gestación inferior a las 28 semanas. Sin embargo, al comparar las cajas según el sexo, en la de los nacidos vivos de sexo masculino se denota una menor dispersión. En este sentido, como un rasgo característico de lo anterior, el rango intercuartil en los nacidos vivos de sexo femenino es de 3 semanas de gestación, mientras que en los de sexo masculino es de 2 semanas de gestación.

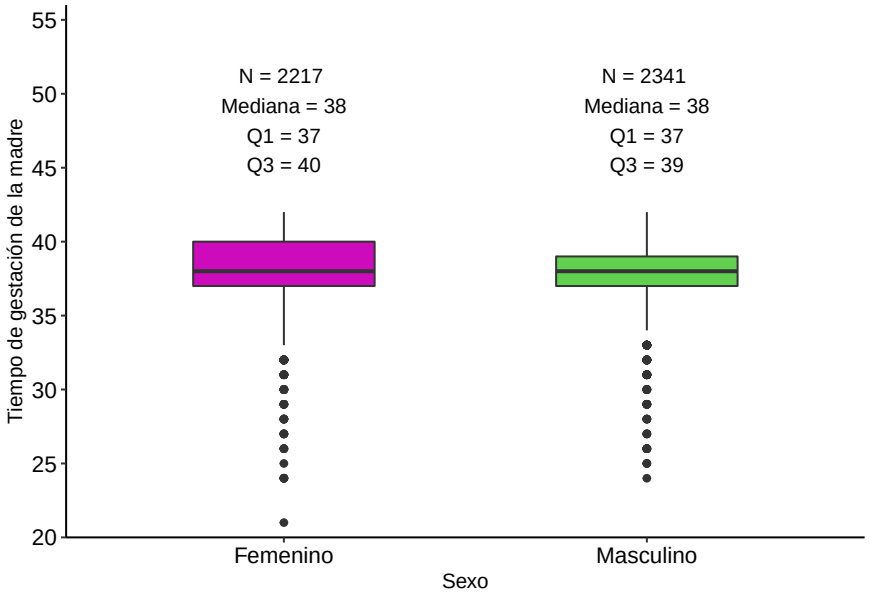

Figura 8: Diagrama de cajas y bigotes del tiempo de gestación de la madre según el sexo del nacido vivo.

Al observar la distribución de la edad de la madre según el tipo de parto que tuvo, se presenta una mayor dispersión de la edad en aquellas madres que tuvieron un parto por cesárea (`ver figura 9`), pues el 50% de las madres tiene una edad entre 22 y 32 años, con una mediana de 27 años, y una distribución de la edad con una asimetría hacia la derecha; lo cual indica que la mayoría registros son de madres con una edad superior a los 27 años que tuvieron un parto por cesárea. Sin embargo, cabe resaltar, en este caso, que el mínimo de la edad registrada por una madre que tuvo parto por cesárea fue de 14 años. En lo relativo a la

edad de las madres que tuvieron partos espontáneos y asistidos por instrumentado (fórceps o ventosas), las medianas de edad de las madres es igual para cada tipo (23 años) y el 50% de las madres que tuvieron un parto espontáneo tiene entre 20 y 28 años, y el 50% de las que tuvieron un parto instrumentado tiene entre 20 y 29 años; cabe resaltar que la edad mínima en las mujeres que tuvieron un parto de tipo espontáneo es de 12 años, y en las que tuvieron un parto de tipo instrumentado es de 14 años. En estos dos tipos de parto, se presenta una distribución de la edad de la madre con una asimetría hacia la derecha; lo que significa que hay una mayor cantidad de registros de madres con edad superior a los 23 años que tuvieron un parto espontáneo y para las que tuvieron un parto instrumentado desde 23 años en adelante. Cabe anotar que, al comparar las cajas según el tipo de parto, en la correspondiente a *espontáneo* se observa una menor dispersión. En este sentido, como un rasgo característico de lo anterior, el rango intercuartil de la edad de la madre cuyo parto fue por cesárea es de 10 años de edad, mientras que en los de tipo instrumentado es de 9 años de edad y en los de tipo espontáneo fue de 8 años de edad.

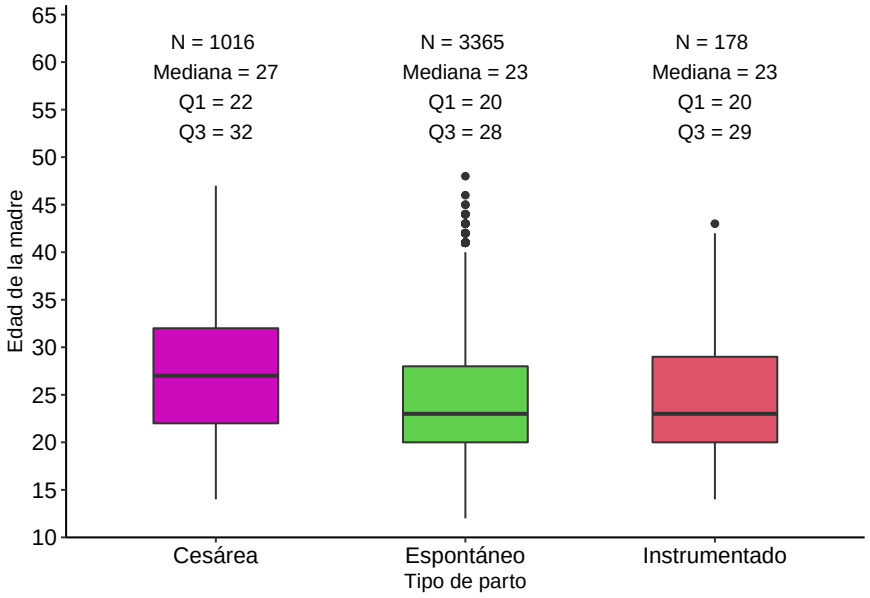

Figura 9: Diagrama de cajas y bigotes de la edad de la madre según el tipo de parto.

Conclusiones

Después de analizar las características de los nacidos vivos en el Hospital General de Medellín en el año 2021, podemos concluir que, de manera marginal, la mayor proporción es de sexo masculino y tuvo un parto de tipo espontáneo; sin presentarse diferencias significativas en el tipo de parto según el sexo del nacido vivo.

En cuanto al peso de los nacidos vivos, la media fue de 2942 gramos, con una desviación estandar de ± 595 gramos. Sin embargo el peso mínimo registrado por un nacido vivo fue de 245 gramos, y el máximo fue de 4872 gramos. En lo relativo a la talla de los nacidos vivos, la media fue de 47.6 centímetros, con una desviación estandar de ± 3.17 centímetros. No obstante, la talla mínima registrada por un nacido vivo fue de 22 centímetros y la máxima fue de 58. En ambos casos, se presentó una asimetría hacia la izquierda; es decir, la mayor cantidad de registros tuvieron pesos y tallas inferiores a la media.

En lo relativo al tiempo de gestación de las madres, se presenta una dispersión de los valores con sesgo negativo; es decir, la mayoría de las madres tuvieron tiempos de gestación inferior a las 37.9 semanas (media). No obstante, al observar el comportamiento por sexo, hay una mayor dispersión en el tiempo de semanas en los nacidos vivos de sexo femenino, a diferencia de los de sexo masculino. Sin embargo, las medianas del tiempo de gestación de la madre para ambos sexos del nacido vivo son la misma: 38 semanas.

Por último, al observar el comportamiento de la distribución de la edad de la madre según el tipo de parto que tuvo, llama la atención que la edad mínima registrada por tipo de parto fue de: 14 años para partos de tipo cesárea, 12 años para los de tipo espontáneo, y 14 años para partis de tipo instrumentado. En cuanto a la dispersión de la edad de la madre según el tipo de parto, comparten la característica de tener un sesgo positivo, es decir, la mayoría de las madres que tuvieron un parto por cesárea tiene más de 27 años, la mayoría de las que tuvieron un parto espontáneo tienen más de 24.5 años, y la mayoría de las que tuvieron un parto instrumentado tienen más de 24.9 años.

Anexo

Modelo de regresión

En este se anexo, se analizará el comportamiento de la variable talla en función de la variable peso, como se plantea en la siguiente ecuación:

$$\widehat{\text{talla}} = \hat{\beta}_0 + \hat{\beta}_1(\text{peso})$$

Es decir, la anterior ecuación, reemplazando los términos con los coeficientes estimados quedaría de la siguiente manera:

$$\widehat{\text{talla}} = 33.8 + 0.0047(\text{peso})$$

En donde $\hat{B}_0$ es 33.8 centímetros, y puede ser interpretado como la predicción de la talla del nacido vivo en el Hospital General de Medellín cuando el peso es igual a 0. Y $\hat{B}_1$ es igual a 0.0047;lo que significa, teniendo presente que la unidad de medida de la variable peso es en gramos, cuando el peso es igual a 1000 gramos, el incremento en la talla corresponde a 4.7 centímetros, tal como se observa en la siguiente ecuación:

$$\widehat{\text{talla}} = 33.8 + 0.0047(1000) = 38.5 \textit{centímetros}$$

Al calcular el error cuadrático medio, que indica la cantidad de error contenida en el modelo y cuyo resultado es interpretado en la unidad de medida de la variable predecida, el resultado es de 1.5004525 centímetros. Dicho error se obtiene a través de la siguiente ecuación:

$$ECM(RMSE) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

Al comparar el error cuadrático medio, que nos da la dispersión de los residuos del modelo al rededor del valor predecido, con la desviación estandar de la variable talla, que nos da la dispersión de los valores en torno a la media de la talla, la cual es 3.17 centímetros, se puede observar que el valor del error cuadrático medio es inferior $RMSE(1.50cms) < \sigma(3.17cms)$. Esto nos indica que el modelo tiene una buena capacidad predictiva.

Como se puede observar en la `figura 10`, hay una relación lineal directa o positiva entre el peso y la talla de los nacidos vivos, cuyo coeficiente de correlación es 0.880559; lo cual significa que, mientras más peso tiene el nacido vivo, es probable que tenga una mayor talla en centímetros. Adicionalmente, al calcular el coeficiente de determinación R^2, podemos decir que el 77.5% de la variabilidad en la talla del nacido vivo es explicada por el peso del nacido vivo. Por ende, la variabilidad no explicada por el modelo es del 22.5%.

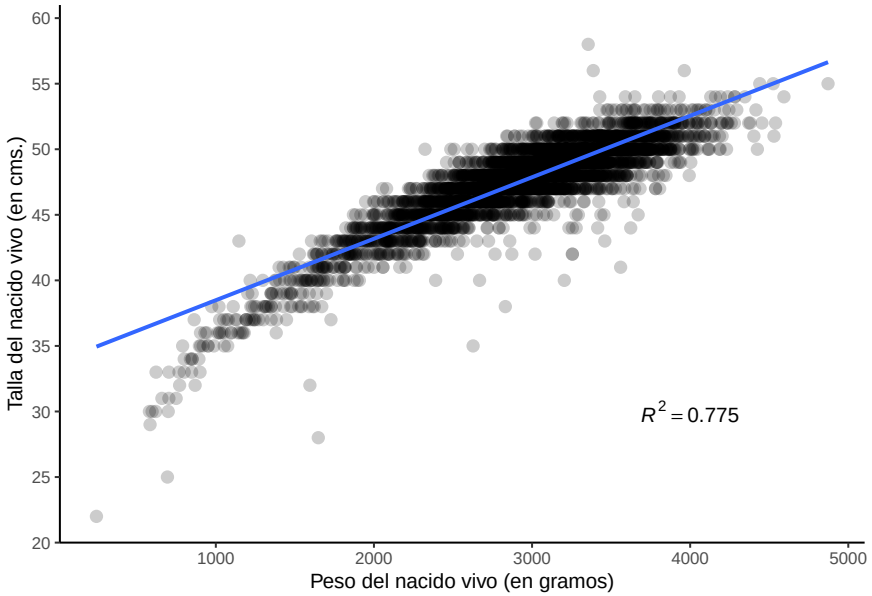

Figura 10: Gráfico de dispersión de la talla en función del peso de los nacidos vivos.

Para validar el modelo de relación lineal planteado, vamos a analizar la distribución de los residuos en función de los valores esperados para la talla de los nacidos vivos. En la figura 11, puede observarse la presencia de varios valores atípicos que afectan nuestro modelo lineal en el sentido en que afectan el error del modelo; especialmente en la franja que comprende las tallas de 35 a 40 centímetros, como se señala en la `figura 11`, en el rectángulo que señala dicha franja y contiente la linea azul. Sin embargo, si extraemos esta zona del gráfico, los residuos tienen un comportamiento de dispersión constante al rededor del cero, con un comportamiento homocedástico, presentando un patrón de canalización entre -5 y 5, como se observa en la gráfica; especialmente, en el rango de tallas comprendido entre 40 centímetros y 57 centímetros.

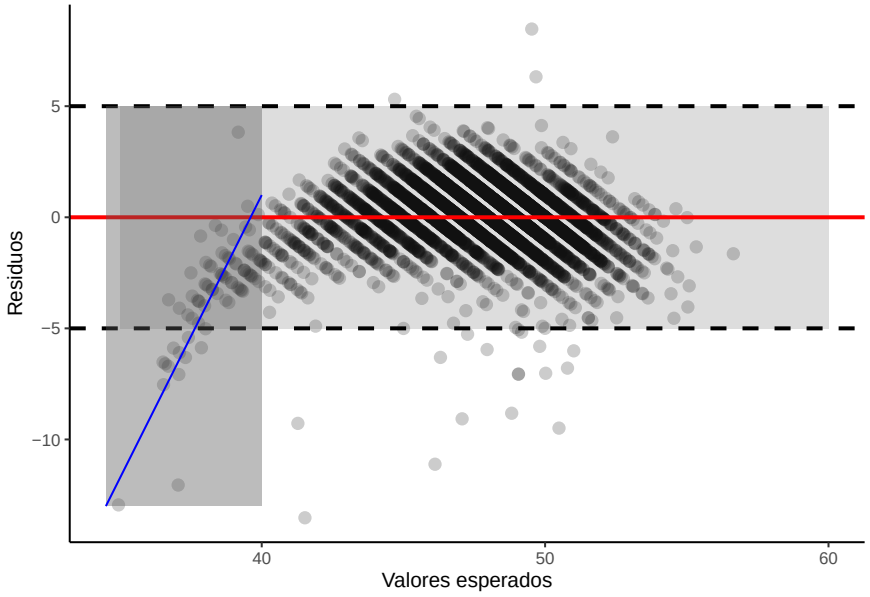

Figura 11: Distribución de los residuos del modelo.

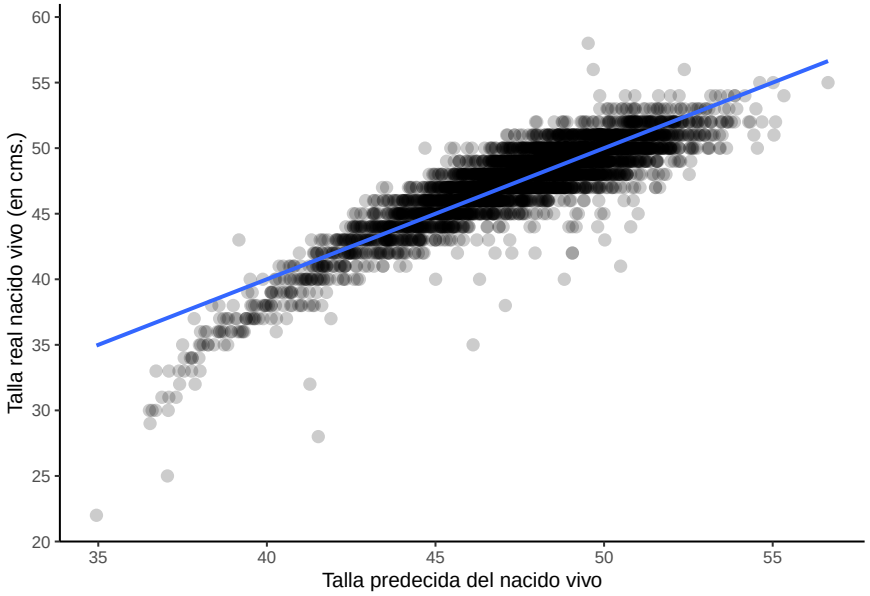

Figura 12: Gráfico de dispersión de la talla esperada y la talla observada en los nacidos vivos.

Referencias bibliográficas

Las fórmulas contenidas en este artículo provienen de los siguientes textos:

- Goos, P., & Meintrup, D. (2015). Statistics with JMP: graphs, descriptive statistics and probability. John Wiley & Sons.
- Meyer, T. (2012). Root Mean Square Error Compared to, and Contrasted with, Standard Deviation. Surveying & Land Information Science, 72(3).