CÓMO LEER Y GENERAR PUBLICACIONES CIENTÍFICAS. EXPLORACIÓN GRÁFICA DE DATOS CUANTITATIVOS: LA IMPORTANCIA DE MIRAR LA INFORMACIÓN

Article · December 2019

CITATIONS
0

READS
331

2 authors:

Mauricio Fuentes Alburquenque
University of Chile
24 PUBLICATIONS **37** CITATIONS
SEE PROFILE

Karla Yohannessen
University of Chile
48 PUBLICATIONS **147** CITATIONS
SEE PROFILE

CÓMO LEER Y GENERAR PUBLICACIONES CIENTÍFICAS
Exploración gráfica de datos cuantitativos: La importancia de mirar la información

HOW TO READ AND GENERATE SCIENTIFIC PUBLICATIONS
Graphical exploration of quantitative data: The importance of looking at the information

Mauricio Fuentes A.*
Programa de Bioestadística, Escuela de Salud Pública
Facultad de Medicina, Universidad de Chile

Karla Yohannessen V.
Programa de Salud Ambiental, Escuela de Salud Pública
Facultad de Medicina, Universidad de Chile
Departamento de Pediatría y Cirugía Infantil
Facultad de Medicina, Universidad de Chile

Publicado en *Neumología Pediátrica*, Vol. 14, No. 4: 194-199
(https://www.neumologia-pediatrica.cl/wp-content/uploads/2019/12/1.pdf)

Resumen

Una vez finalizada la recolección de datos de un estudio y se cuenta con la respectiva base de datos, es frecuente que el investigador esté impaciente por responder a la pregunta de investigación y se aventure a realizar los pasos finales del análisis. No obstante, una etapa clave, previa a un análisis estadístico más complejo o sofisticado, es la exploración de datos y la estadística descriptiva. Lamentablemente, el análisis exploratorio de los datos muchas veces es realizado sin mucha dedicación, o simplemente es "saltado", lo que puede tener consecuencias importantes en los resultados obtenidos y conducir al reporte de conclusiones erróneas. Por un lado, la exploración permite detectar errores en los datos y, si es posible, corregirlos desde la fuente de origen o tenerlos en cuenta para tomar decisiones respecto a qué hacer con ellos. Por otra parte, la exploración permite conocer el comportamiento de las variables evaluadas en términos de su distribución (concepto clave en Estadística) y posibles relaciones entre ellas, lo cual es fundamental para los análisis descriptivo e inferencial posteriores. El objetivo de este artículo es mostrar herramientas gráficas para la exploración de datos cuantitativos, con el fin de visualizar su distribución y comparar grupos según categorías de variables cualitativas.

Palabras clave: variable cuantitativa, análisis exploratorio, gráficos estadísticos, distribución de una variable

*Independencia 1027, Independencia, Santiago, Chile; +562 29786554; mauriciofuentes@med.uchile.cl.

Abstract

Once the collection of data from a study has been completed and the respective database is available, the researcher is often impatient to answer the research question and ventures into the final steps of the analysis. However, a key stage, prior to a more complex or sophisticated statistical analysis, is data exploration and descriptive statistics. Unfortunately, the exploratory analysis of the data is often performed without much dedication, or is simply "skipped", which can have important consequences on the results obtained and lead to the report of erroneous conclusions. On the one hand, exploration allows to detect errors in the data and, if possible, to correct them from the source of origin or take them into account to make decisions about what to do with them. On the other hand, exploration allows to know the behavior of the variables evaluated in terms of their distribution (key concept in Statistics) and possible relationships among them, which is essential for subsequent descriptive and inferential analysis. The objective of this article is to show graphic tools for the exploration of quantitative data, in order to visualize its distribution and compare groups according to categories of qualitative variables.

Keywords: quantitative variable, exploratory analysis, statistical graphics, variable distribution

Una vez que ha finalizado la etapa de recolección de datos de un estudio, el Análisis Exploratorio (AE) es la primera fase del análisis estadístico previo al análisis descriptivo e inferencial. El AE permite evaluar la calidad de los datos recogidos y digitados, si es posible corregir los datos erróneos o tenerlos en cuenta para análisis posteriores, resguardando un reporte de resultados y conclusiones adecuado. Por otro lado, en el caso de las variables cuantitativas, el AE permite evaluar la distribución de dichas variables. En el artículo *Rol y definición de las variables en una investigación: el protagonismo que se merecen* [1], el lector podrá revisar qué es una variable cuantitativa.

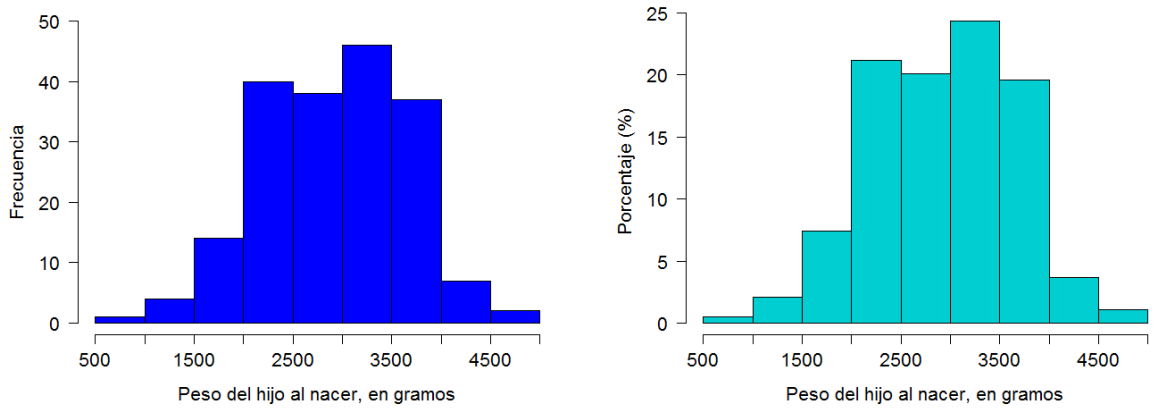
Existen diferentes herramientas gráficas para estudiar la distribución de variables cuantitativas, las cuales se presentan en las siguientes secciones de este artículo. Para ilustrar estas herramientas se utilizó una base de datos, presentada por Hosmer y Lemeshow [2], de 189 recién nacidos y sus madres atendidos en el Baystate Medical Center de Springfield, Massachusetts, EE.UU (disponible en el sitio `ftp://ftp.wiley.com/public/sci_tech_med/logistic`). Estos datos fueron recolectados dentro de un estudio cuyo objetivo era investigar si algunas características o comportamientos durante el embarazo (alimentación, consumo de tabaco, atención médica prenatal, entre otros) influían en el peso del recién nacido. Dentro de las variables registradas, que serán utilizadas en este artículo, se encuentran las siguientes:

- Peso de la madre a la fecha de su último período menstrual, en kilogramos (kg).
- Raza de la madre (blanca, negra, otra).
- Hábito tabáquico de la madre (fuma, no fuma).
- Peso al nacer, en gramos (g).

La variable de mayor interés en el estudio fue el peso al nacer, por lo que este artículo se enfocará principalmente en el comportamiento de dicha variable, intentando responder a través de herramientas gráficas las preguntas ¿cómo se distribuyeron los pesos al nacer de todos los niños? y ¿la distribución fue distinta, por ejemplo, si las madres fumaban o no fumaban?

Histogramas

Los números en una base de datos son los valores que toma cada variable para cada individuo de la muestra. En el caso de una variable cuantitativa la mayoría de los valores pueden tomar valores diferentes [3]. No obstante, habitualmente los valores se agrupan formando intervalos, determinando la distribución de la variable, que corresponde al patrón de ocurrencia observado en un conjunto de valores de una variable [4]. De las diversas formas para visualizar estos patrones, el gráfico usado por excelencia es el histograma, un gráfico con barras generalmente verticales donde la altura de cada barra representa la frecuencia absoluta (número de observaciones) o el porcentaje de ocurrencia del respectivo valor o intervalo de la variable [3, 4]. La Figura 1 muestra los histogramas de la variable peso al nacer y su interpretación. En general, en el eje horizontal (eje X) del histograma se ubican los intervalos de la variable analizada y en el eje vertical (eje Y) se ubica la frecuencia absoluta o el porcentaje de ocurrencia. La simetría, uno de los elementos más importantes que se evalúa en un histograma, corresponde a la forma que muestra el conjunto de barras, y se habla de una distribución simétrica cuando se observa aproximadamente la misma forma hacia ambos lados (izquierda y derecha) de su valor central, por ejemplo, similar a una campana. Por el contrario, si se observa una distribución de las barras más alargada hacia la derecha o hacia la izquierda, se hablará de una distribución asimétrica. Otro elemento que se puede observar en el histograma es el intervalo modal o más frecuente, que corresponde a la barra más alta, el cual usualmente es único. No obstante, algunas veces se puede encontrar más de un intervalo modal.

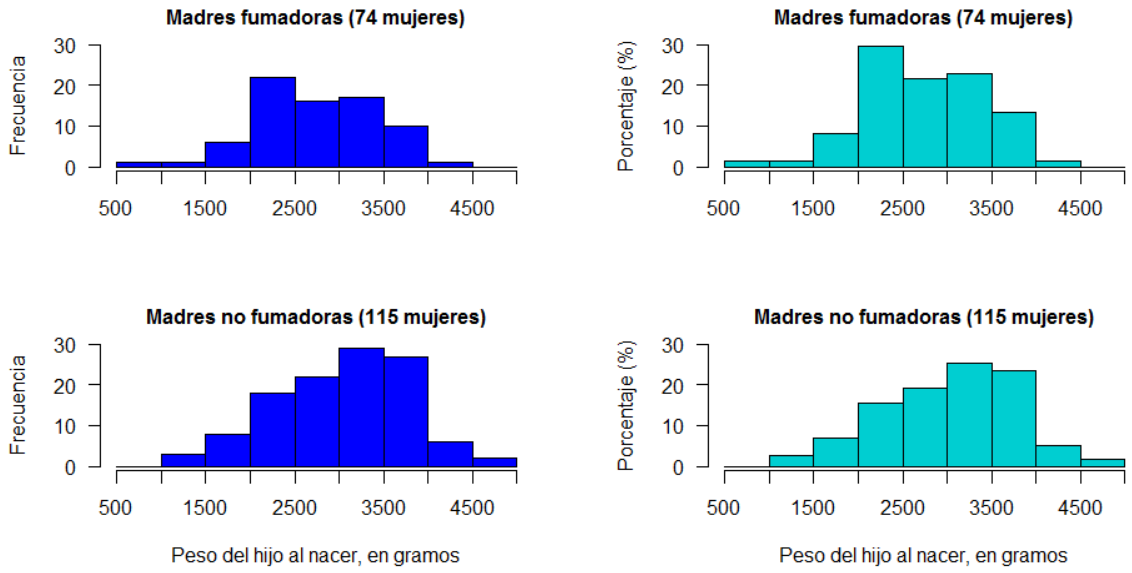


(a) **(b)**

Figura 1: Histogramas de la variable peso al nacer, cuyos valores se muestran en el eje horizontal (eje X), y en el eje vertical (eje Y) se muestra (a) la frecuencia absoluta (número de casos) o (b) el porcentaje de ocurrencia. En ambos gráficos se observa que los pesos al nacer tienen una distribución más o menos simétrica, el intervalo modal (barra de mayor altura) es el de 3000 a 3500 g y que la mayoría de los recién nacidos pesaron entre 2000 y 4000 g.

La Figura 2 compara los histogramas del peso al nacer de los hijos de madres fumadoras y no fumadoras utilizando la frecuencia absoluta y el porcentaje de ocurrencia. La comparación de dos histogramas representados con frecuencias absolutas se debe hacer con precaución, en especial si el número de sujetos en ambos grupos es distinto. En este caso resulta más apropiado comparar histogramas que representen los porcentajes de ocurrencia. Un aspecto importante a destacar es la necesidad de que los histogramas a comparar tengan las mismas escalas, tanto en el rango de los valores usados en el eje vertical como en los intervalos usados en eje horizontal. De esta manera, la visualización gráfica entregará una correcta impresión de la información, lo que es relevante dado que se espera que un gráfico permita interpretar de

manera rápida a través de una primera mirada. En la Figura 3 se comparan los histogramas del peso a la fecha del último período menstrual según la raza de las madres, utilizando distintas escalas (a) y las mismas escalas en los ejes (b).



(a) **(b)**

Figura 2: Histogramas de la variable peso al nacer según hábito tabáquico de la madre, usando (a) frecuencias absolutas y (b) porcentaje de ocurrencia. En los histogramas de frecuencias absolutas (a), se observa que el intervalo modal del peso al nacer en los hijos de madres fumadoras fue de 2000-2500 g, con aproximadamente 20 casos, mientras que en los hijos de madres no fumadoras el intervalo más frecuente fue 3000-3500 g, con alrededor de 30 casos. No obstante, resulta difícil conocer la importancia relativa de estos dos valores. Por otro lado, en los histogramas de porcentaje de ocurrencia (b) se puede observar que el intervalo modal en el grupo de madres fumadoras (2000-2500 g), representa aproximadamente un 30 % de los casos, mientras que el intervalo modal del grupo de madres no fumadoras (3000-3500 g) representa un porcentaje inferior al 30 %. Además, en estos gráficos los pesos ya no se muestran tan simétricos como en la muestra completa.

Polígonos de frecuencia

Los polígonos de frecuencia relacionan los valores de la variable con sus respectivas frecuencias, mostrando la distribución de los datos cuantitativos como una serie de puntos conectados por medio de líneas rectas, quedando representada una curva que resulta ser muy útil tanto para describir los datos [5, 6] como para comparar dos o más grupos. Similar a lo mencionado en los histogramas, para la comparación resulta más adecuado usar porcentajes de ocurrencia. La Figura 4 muestra los pesos al nacer utilizando polígonos de frecuencia, comparando el grupo de madres fumadoras con el de no fumadoras.

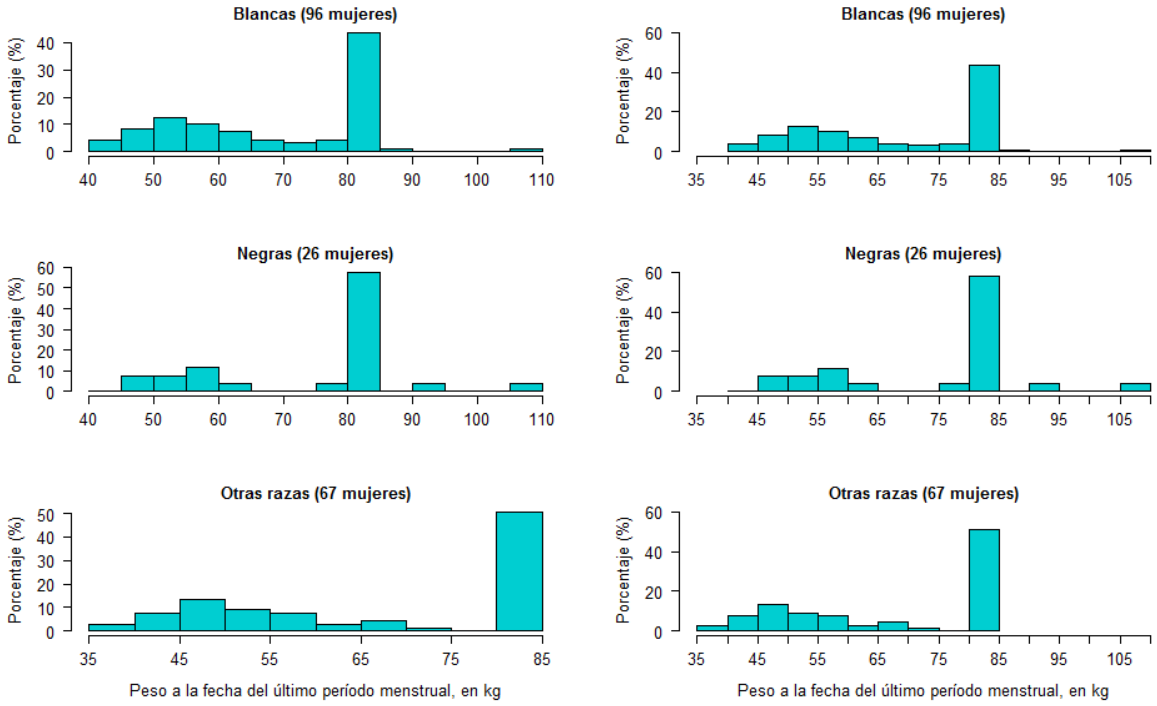
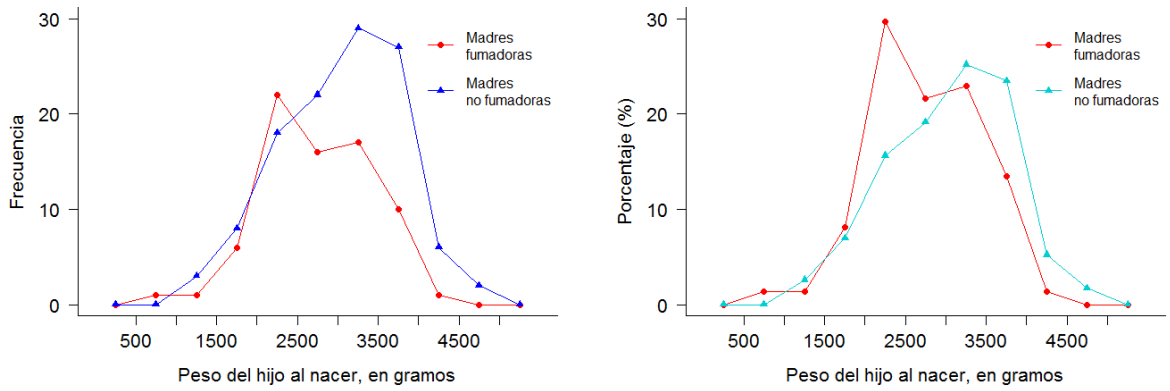


Figura 3: Histogramas de los pesos a la fecha del último período menstrual según raza de la madre usando los porcentajes de ocurrencia: (a) con distintas escalas en los ejes y (b) con la misma escala en los ejes. En (a), una primera impresión podría indicar que el intervalo modal es mayor en las madres de otras razas, dado que en el tercer histograma la barra más alta está más a la derecha. Sin embargo, esta interpretación errónea es producto de haber usado una escala distinta en el eje horizontal del último gráfico. Otra impresión errónea sería que los intervalos modales son similares en los tres grupos (alturas de las barras similares), sin embargo, los ejes verticales tienen distintas escalas. En (b) estos aspectos fueron corregidos, donde se observa que el intervalo modal es el mismo en las tres razas (80-85 kg) con una proporción menor en las madres blancas. Otra información que se observa es que ninguna madre del tercer grupo presentó un peso mayor 85 kg, y que ninguna madre blanca o negra tuvo un peso menor a 40 kg.

Diagramas de caja y bigote

El diagrama de caja y bigote o *box plot* (o *box-and-whisker plot*) es un gráfico comúnmente usado para comparar distribuciones entre grupos, incluso más que los histogramas y los polígonos de frecuencias. Son muy útiles para resumir visualmente la forma de una distribución y su grado de simetría [4]. Como se puede ver en la Figura 5, la caja muestra las posiciones de los percentiles 25 (extremo inferior), 50 (línea interior) y 75 (extremo superior), que corresponden a los cuartiles de la distribución o aquellos valores que dividen el conjunto de datos en cuatro partes iguales. Particularmente, el percentil 50 o segundo cuartil corresponde a la mediana, valor bajo el cual se encuentra la mitad de los datos. El extremo inferior (percentil 25) y superior (percentil 75) de la caja indican que ésta contiene la mitad central de los valores de las observaciones, lo que se conoce como el rango intercuartil de la distribución [4, 5]. Los bigotes se extienden, en ambos sentidos, hasta el dato más extremo que no esté más alejado de 1,5 veces el rango intercuartil desde el respectivo borde de la caja [7]. Cualquier dato fuera de dicho rango, es decir, más allá



(a) **(b)**

Figura 4: Polígonos de frecuencia de los pesos al nacer para madres fumadoras y no fumadoras usando (a) frecuencias absolutas y (b) porcentajes. Este tipo de gráfico es equivalente a superponer dos histogramas, pero en vez de representar las barras, se unen sus alturas en el punto medio (marca de clase del respectivo intervalo) a través de líneas rectas. De este modo, es fácil ver que ambas distribuciones presentan una dispersión similar, aunque los hijos de las madres fumadoras tendieron a tener menores pesos de nacimiento. Tal como en la Figura 2, las alturas de las gráficas son diferentes si se representan frecuencias o porcentajes, siendo mejor la segunda opción para comparar.

de los bigotes, se denomina "outlier" y es mostrado en el gráfico (Figura 5). Estos valores fuera de rango deben ser interpretados en el contexto de cada variable, identificando si corresponde a un valor plausible o a errores de registro o codificación. En este último caso, el investigador podrá corregir, si es posible, o tener en cuenta estos valores para tomar decisiones respecto a qué hacer con ellos en los análisis posteriores.

Se han mencionado y explicado brevemente algunas medidas de resumen como la mediana y los percentiles, las cuales serán abordadas con profundidad en un próximo artículo de esta serie. No obstante, aunque no se recuerden o manejen del todo dichos conceptos, una gran ventaja de los diagramas de caja y bigote para distintos grupos es que permiten identificar visualmente si hay aparentes diferencias entre ellos.

Otra característica de los datos que resulta fácil de visualizar en un diagrama de caja y bigote es su grado de simetría. Cuando la mediana se ubica cerca de la mitad de la caja, podemos decir que el 50 % central de los datos es aproximadamente simétrico, en cambio si la mediana se acerca al borde superior de la caja (percentil 75) o al inferior (percentil 25), lo más seguro es que la distribución sea asimétrica [4]. Una interpretación similar se hace en relación a la distancia de los bigotes desde la caja.

En el diagrama de caja y bigote también es posible observar fácilmente la dispersión de la variable. Mientras más alargada es la caja, más amplio es el rango en el cual se encuentra la mitad central de los datos (rango intercuartil). Asimismo, mientras más amplio es el rango entre ambos bigotes mayor es la dispersión o variabilidad total de los datos. Evaluar el grado de dispersión de una variable es importante en el análisis estadístico, ya que indica qué tan homogénea (menor dispersión) o heterogénea (mayor dispersión) es la muestra respecto a dicha variable. Cuando un grupo es más homogéneo, los individuos son más parecidos entre ellos (al menos en la variable de interés), y por lo tanto resulta más sencillo describir y generalizar sus características.

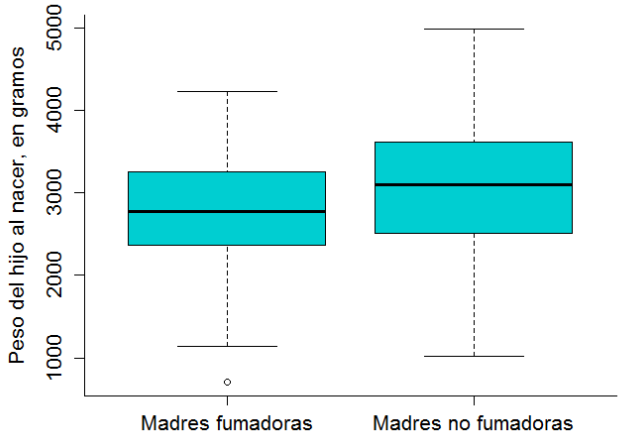


Figura 5: Comparación de pesos al nacer entre madres fumadoras y no fumadoras usando diagramas de caja y bigote. La variable cuantitativa (peso al nacer) se ubica en el eje vertical y los grupos a comparar (categorías de la variable cualitativa) en el eje horizontal. La línea ubicada dentro de las cajas indica la mediana de cada grupo, siendo éstas aproximadamente 2700 g en las madres fumadoras y 3100 g en las madres no fumadoras. El extremo inferior (percentil 25) y superior (percentil 75) de las cajas permiten estimar que la mitad de los hijos de las madres fumadoras nació con pesos entre 2300 y 3200 g, mientras que la mitad de los hijos de las madres no fumadoras nació con pesos entre 2500 y 3600 g. El punto ubicado bajo el bigote inferior del grupo de madres fumadoras corresponde a un "outlier", y representa un bebé que nació con un peso inusualmente bajo (¡aproximadamente 700 g!). En general, se observa que los hijos de madres fumadoras presentaron menores pesos al nacer que los de madres no fumadoras. Finalmente, los pesos al nacer en el grupo de madres fumadoras fueron algo más asimétricos que los pesos del otro grupo, dado que el bigote inferior está más alejado de la caja que el bigote superior. Además, se ve que la extensión tanto de la caja como de los bigotes en el grupo de madres fumadoras es menor, lo que indica que éste tiene menos dispersión que el otro grupo, es decir, es más homogéneo en términos del peso al nacer.

Distribución acumulada

Otro aspecto importante en la exploración de los datos es conocer la distribución de frecuencia acumulada o porcentaje acumulado. Para un valor determinado de la variable, la frecuencia acumulada indica el número de casos con valores menores o iguales a dicho valor, y el porcentaje acumulado indica el porcentaje de casos con valores menores o iguales a ese número. Esto se puede visualizar a través de un polígono de frecuencia acumulada u ojiva como el mostrado en la Figura 6. La forma de la ojiva depende de la simetría de la distribución, lo que se ilustra al comparar las ojivas (a) y (b) con la (d) de la Figura 6.

Conclusión

Se han mostrado los gráficos estadísticos utilizados con mayor frecuencia y que, a juicio de los autores, son los más útiles para explorar datos cuantitativos. Se visualizó gráficamente la distribución de todos los valores de una variable cuantitativa (análisis univariado), así como los valores separados en categorías de una variable cualitativa (análisis bivariado), intentando visualizar si hay una relación entre la variable cuantitativa de interés y esa variable cualitativa. Estos sencillos procedimientos, entre otros similares, permiten en primera instancia identificar errores de registro y codificación, aumentando la calidad y conocimiento de los datos en los que se basarán los siguientes análisis. En segunda instancia, también

permiten evaluar la distribución de las variables cuantitativas, lo cual constituye ineludiblemente la base de los siguientes análisis estadísticos, tanto el descriptivo como el uso de métodos de inferencia estadística (estimación de parámetros, test de hipótesis, modelos de regresión, entre otros), ya que éstos consideran la distribución de la variable para su aplicación. La omisión de esta primera etapa puede conducir al reporte de resultados sesgados y conclusiones erróneas.

Los autores declaran no tener conflictos de interés.

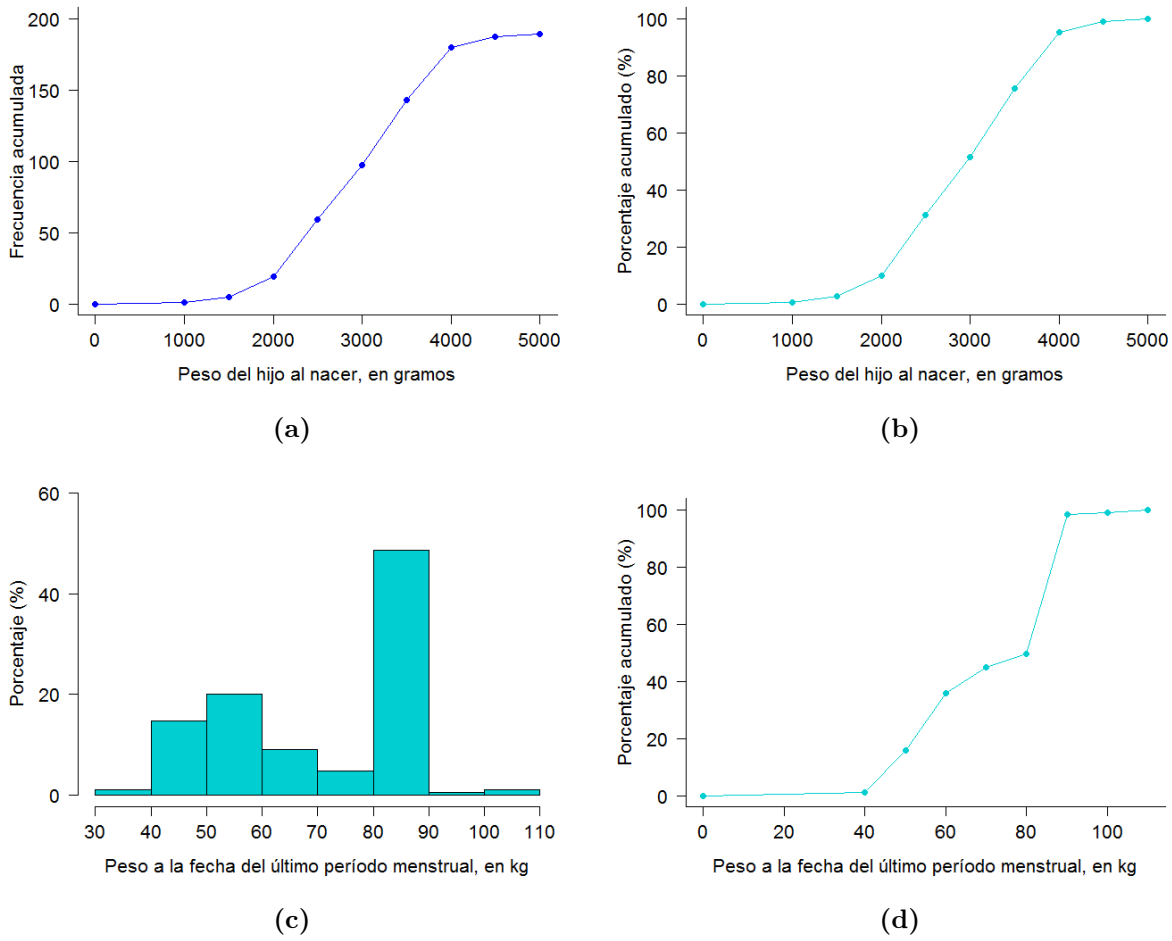


Figura 6: Arriba: Polígono de frecuencia acumulada u ojiva de los pesos al nacer usando (a) frecuencias absolutas y (b) porcentajes. Abajo: Distribución porcentual de los pesos de las madres a la fecha del último período menstrual: (c) histograma, (d) ojiva. En (a) cada punto indica cuántos niños nacieron con un peso menor o igual a dicho valor. Por ejemplo, el sexto punto de la curva indica que (aproximadamente) 100 niños tuvieron un peso al nacer menor o igual a 3000 g, que es equivalente a decir que (aproximadamente) 90 niños pesaron 3000 g o más. En (b) cada punto indica el porcentaje de niños que nacieron con un peso menor o igual a dicho valor, es decir, el sexto punto indica que aproximadamente el 50 % de los niños nacieron con 3000 g o menos. Nótese que incluir o no el valor puntual de 3000 g en una u otra porción de la distribución es irrelevante estadísticamente, es decir, da lo mismo decir que la mitad pesó 3000 g o menos o decir que pesó menos de 3000 g. En (c) se observa que la distribución del peso de la madre a la fecha del último período menstrual dista de ser simétrica, y su correspondiente ojiva en (d) tiene una forma más escalonada, con un cambio pronunciado en el valor modal.

Referencias

[1] K. Yohannessen V. and M. Fuentes A., "Cómo leer y generar publicaciones científicas. Rol y definición de las variables en una investigación: El protagonismo que se merecen," *Neumología Pediátrica*, vol. 14, no. 3, pp. 122–125, 2019.

[2] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York: John Wiley & Sons, Inc., 2nd ed., 2000.

[3] J. M. Argimon Pallás and J. Jiménez Villa, *Métodos de investigación clínica y epidemiológica*. Barcelona: Elsevier España, S.L., 4a ed., 2013.

[4] G. Dawson, *Interpretación fácil de la bioestadística*. Barcelona: Elsevier España, S.L., 1a ed., 2009.

[5] R. Hernández Sampieri, C. Fernández Collado, and P. Baptista Lucio, *Metodología de la investigación*. México D.F.: McGraw-Hill/Interamericana Editores, 6a ed., 2014.

[6] V. J. O. Bennet, W. L. Briggs, and M. F. Triola, *Razonamiento estadístico*. México D.F.: Pearson Educación de México, S.A., 1a ed., 2011.

[7] E. Taucher, *Bioestadística*. Santiago: Universidad de Chile, Escuela de Salud Pública, Ocho Libros Editores, 3a ed., 2014.