UNIVERSIDAD INDUSTRIAL DE SANTANDER.
ESCUELA DE MATEMÁTICAS. BUCARAMANGA
CÍRCULOS MATEMÁTICOS 2022

Nombre: _____ **Identificación:** _____

EL ANÁLISIS DE DATOS PARA LA TOMA DE DECISIONES EN LAS CIENCIAS SOCIALES Y DE LA SALUD.

Sesión 5, orientada por: Andrés Sebastián Ríos-Gutiérrez

El objetivo de la presente actividad es introducirnos en la labor de un analista de datos, mediante el análisis de gráficos estadísticos. Para ello tenemos dos conjuntos de datos:

- **La última encuesta longitudinal de Protección Social:**

  "*El DANE para desarrollar su objetivo misional de producir estadísticas oficiales que cumplan con los estándares internacionales y que sirvan para la toma de decisiones, en cooperación con el Departamento Nacional de Planeación (DNP), propuso realizar una encuesta longitudinal de carácter oficial para Colombia que permitiera observar la dinámica del ingreso y consumo de los hogares, sus dinámicas en el mercado laboral, factores de riesgo y vulnerabilidad frente a choques externos, la efectividad de las políticas de protección social, monitorear condiciones de calidad de vida de la población, así como las dinámicas de movilidad social."* [Da1].

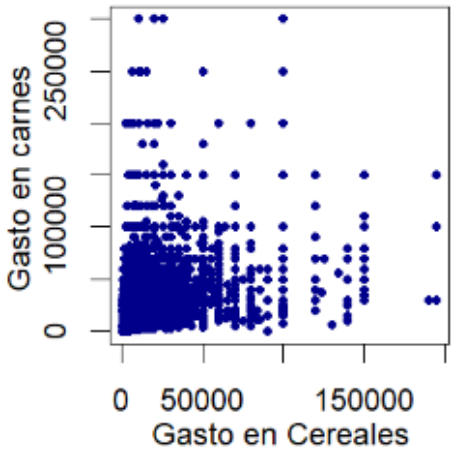
- **El Censo de Recién Nacidos vivos en Colombia del año 2019:**

  El Censo de Recién Nacidos vivos en Colombia corresponde a una de las bases de datos de las Estadísticas Vitales que reúne el DANE anualmente. Estas estadísticas "*son un proceso que reúne información mediante un registro y reporta la frecuencia o la ocurrencia de acontecimientos vitales específicos y definidos por el sistema (Nacimientos y defunciones en Colombia), así como las características propias de los hechos vitales. También integra procesos de compilación, procesamiento, análisis, evaluación y difusión de los datos de forma estadística*" [Da2].

Notemos que antes de analizar nuestros datos, es importante especificar de dónde fueros tomados, si la información contenida es pública o no, y si estos están *anonimizados* o no. Adicionalmente, es fundamental conocer qué variables fueron tomadas, cómo se hizo la toma de datos y toda la información contenida en la ficha técnica de cada encuesta o censo realizado.

Respecto a la **encuesta longitudinal de Protección Social** (Manual en [Da1]):

Vamos a considerar las variables "Gasto de cereales (incluido el arroz) por semana de una familia colombiana" y "Gasto de productos cárnicos (incluidos los embutidos) por semana de una familia colombiana". Observemos el *diagrama de dispersión*:



Interpretación: Si el gráfico no tiene un patrón definido se dice que las variables son independientes.

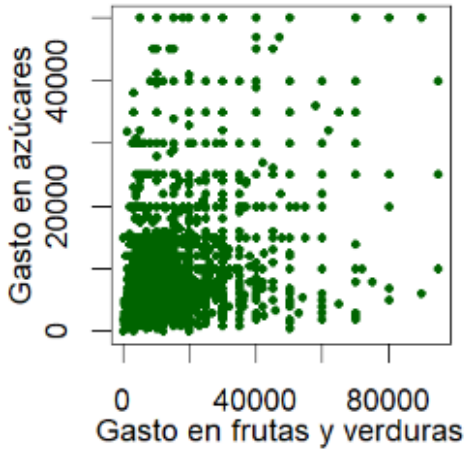
Describe el gráfico con la mayor cantidad de información que consideres pertinente: _____

_____

_____.

De acuerdo con el diagrama de dispersión el gasto mensual de productos cárnicos es (marca una, con una 'X') ❑ **dependiente** ❑ **independiente** del gasto mensual de cereales por semana para una familia colombiana. Esto quiere decir que hay familias con bajo gasto en cereales, pero con alto gasto en productos cárnicos; y hay familias con alto gasto en cereales y bajo gasto en productos cárnicos.

Vamos a considerar las variables "Gasto de frutas y verduras por semana de una familia colombiana" y "Gasto de azúcar, miel, chocolates y dulces de azúcar por semana de una familia colombiana".



De acuerdo con el diagrama de dispersión el gasto mensual de azúcares y derivados es (marca una, con una 'X') ❑ **dependiente** ❑ **independiente** del gasto mensual de frutas y verduras por semana para una familia colombiana. Esto quiere decir que _____

_____.

Vamos a considerar la variable aleatoria "gasto anual por concepto de hospitalización de una familia colombiana". Calculamos la media aritmética, correspondiente a

$$\bar{x} = 313005.6$$

Interpretación: En promedio una familia colombiana gasta anualmente por concepto de hospitalización aproximadamente COP 313000

DISCUSIÓN DE GRUPO: ¿Consideras alto o bajo tal gasto de hospitalización? ¿Cómo podría disminuirse tal gasto en caso de que lo consideres alto? _____

_____

_____.

Vamos a considerar la variable aleatoria "gasto trimestral en ropa y calzado de una familia colombiana". Calculamos la media aritmética, correspondiente a

$$\bar{y} = 244153.2$$

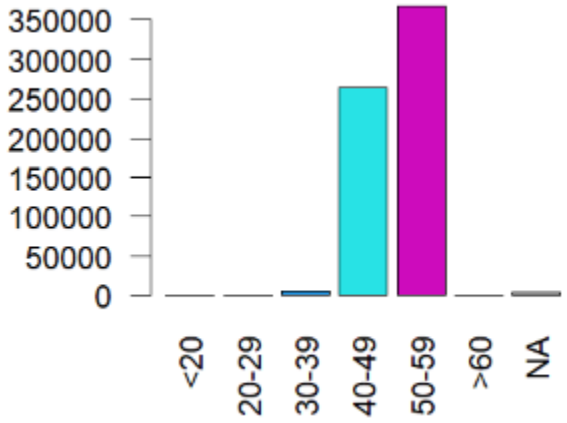
Interpretación: _____

_____.

Lección: Nota que la estadística nos permite responder preguntas sobre ***en qué*** gasta los recursos una familia colombiana y ***de qué*** depende se alimentación. Finalmente, con base en los gráficos estadísticos y las medidas descriptivas estamos explicando ***cómo nos comportamos*** o ***quienes somos***.

Respecto al **Censo de Recién Nacidos vivos en Colombia del año 2019** (Manual en [Da2]):

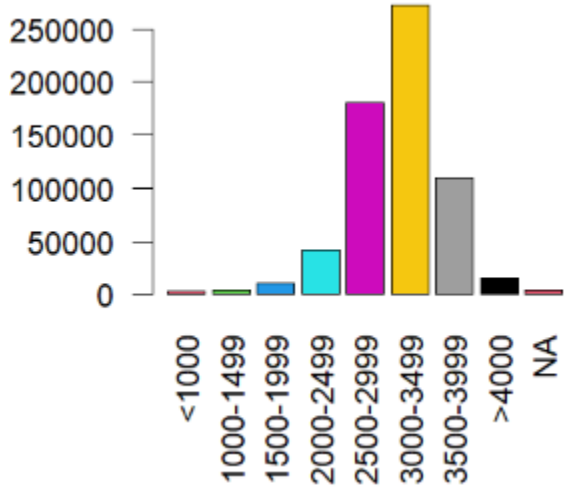
Vamos a considerar la variable "talla de un recién nacido (a) en Colombia". Observemos el *diagrama de barras*:



Teniendo en cuenta que, por ejemplo, la columna "30-39" corresponde entre 30 cm y 39 cm ¿cuáles consideras que son las *tallas normales* de un recién nacido? _____.
Si la columna NA se debe a valores sin información ¿De qué lugares crees que provienen estos datos?

_____.

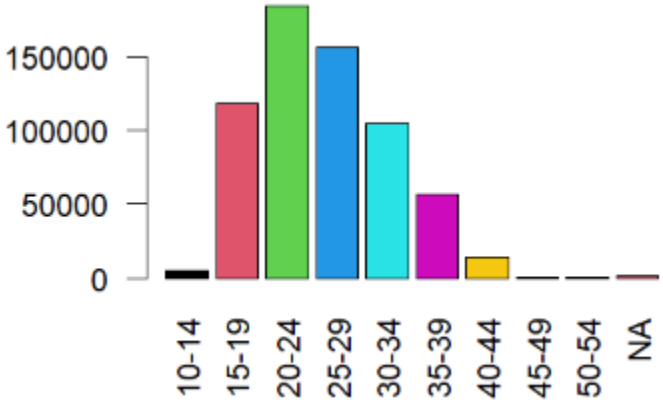
Vamos a considerar la variable "talla de un recién nacido (a) en Colombia". Observemos el *diagrama de barras*:



Teniendo en cuenta que, por ejemplo, la columna "1500-1999" corresponde entre 1500 gr y 1999 gr ¿cuáles consideras que son los *pesos normales* de un recién nacido? _____.

El porcentaje de datos faltantes corresponde aproximadamente al 0.7% de la información. Muchas veces al realizar análisis estadísticos se eliminan los datos faltantes o *datos no representativos*, ya que estos provocan ruido en las *estimaciones*. En otros casos se eliminan porque no determinan ninguna importancia para el análisis de nuestros datos.

Vamos a considerar la variable "edad de la madre de un recién nacido (a) en Colombia". Observemos el *diagrama de barras*:



¿En qué grupos de edades se observa que hay mayor cantidad de mujeres embarazadas? _____

En la siguiente tabla tenemos los conteos por categoría para cada grupo de edad de la madre del recién nacido. En esta tabla no consideramos los datos faltantes.

```
> tabla_madre[1:9]
 10-14  15-19  20-24  25-29  30-34  35-39  40-44  45-49  50-54
  4795 118163 184908 156451 105039  56421  13997    962    125
```

La proporción de madres entre 25 y 29 años corresponde a

$$\text{Proporción}_{25-29} = \frac{156451}{640861} \times 100\% = 24.4126261388975144\% \approx 24.4\%.$$

La proporción de madres mayores que 35 (considerado como alto riesgo de acuerdo con [Do]) corresponde a _____.

La proporción de madres adolescentes (es decir, hasta los 19 años) corresponde a _____.

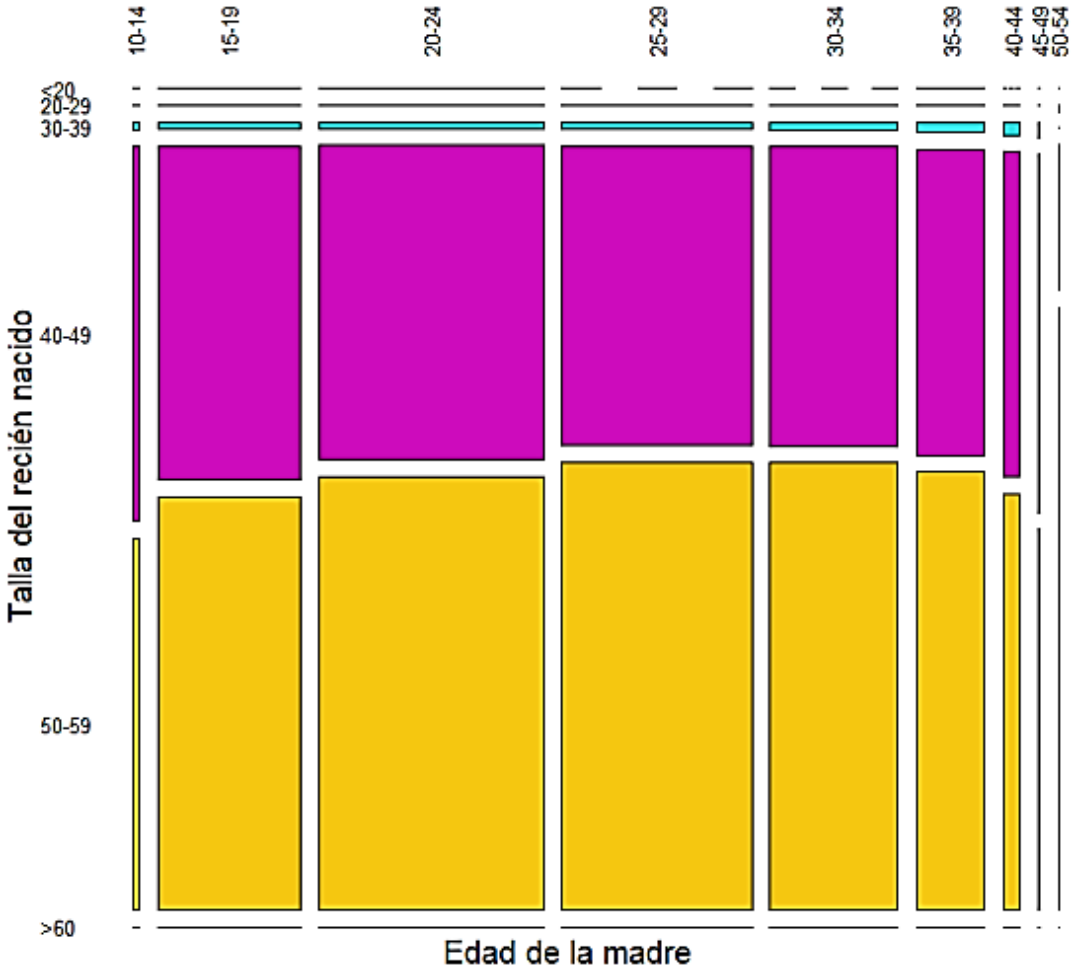
DISCUSIÓN EN GRUPO: De acuerdo con las proporciones ¿crees que es alto o bajo el porcentaje de madres en alto riesgo en Colombia? ¿crees que es alto o bajo el porcentaje de madres adolescentes en Colombia? ¿Cómo podría disminuirse tales proporciones en caso de que los consideres altos?

_____

_____

_____.

Ahora queremos establecer si el peso de un recién nacido depende de la edad de gestación de su madre. Para ello construimos un *diagrama de la tabla de contingencia* [Wn]:



Interpretación: En la gráfica no observamos un patrón definido, es decir, se observa que todos los rectángulos tienen un tamaño similar para la edad de la madre, dada la talla. De esta forma se concluye que la talla de un recién nacido no depende de la edad de la madre.

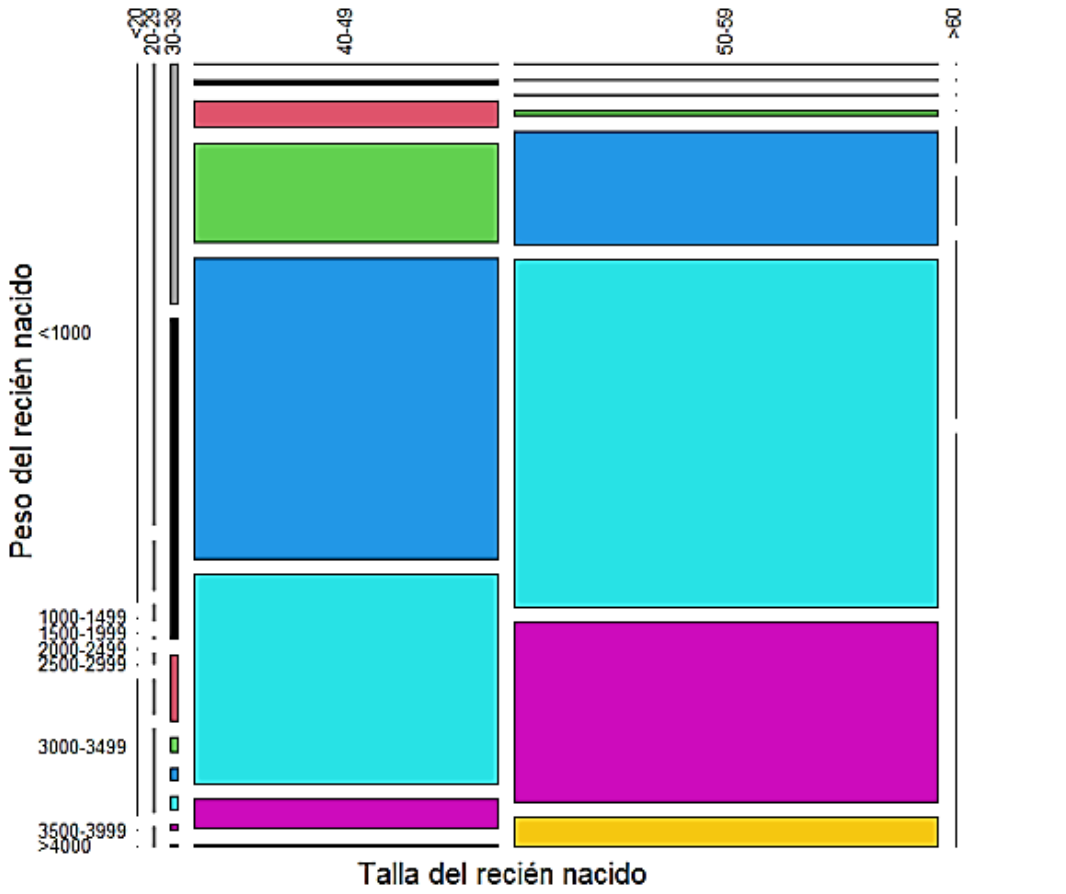
Interpretación: La variable Y ***depende*** de la variable X, cuando se establece que dada la ocurrencia de los valores de X estos implican la ocurrencia de determinados valores de Y (X implica Y). De otra parte, se dice que las variables están ***asociadas*** cuando no se puede establecer cual de las dos variables implica a la otra [Wn].

En el siguiente diagrama de la tabla de contingencia observamos la talla del recién nacido versus con el peso. En este caso se establece que las variables están asociadas[1] porque cuando la talla del recién nacido es menor a 20 cm, se observa una alta prevalencia de un peso menor a 1000 gr. Análogamente para los recién nacidos entre 20 y 29 cm. Respecto a los recién nacidos entre 30 y 39 cm se observa que hay mayor prevalencia de un peso entre 1500 y 1999 gr. _____

_____

_____

_____.



[1] No se puede decir que el peso implica la talla, ni la talla implica el peso; por tenerse dos eventos que ocurren simultáneamente cuando nace el (la) recién nacido (a).

DISCUSIÓN EN GRUPO: De acuerdo con los diagramas de tablas de contingencia ¿La edad de la madre representa un riesgo para que un bebé recién nacido presente una talla baja? ¿La talla y el peso de un recién nacido presentan alguna asociación del tipo 'mayor talla, mayor peso'? _____

_____

_____

_____.

Aspectos para resaltar:

- La encuesta de la última encuesta longitudinal de Protección Social fue realizada a 14407 hogares escogidos *heterogéneamente* en todo el país. El censo de recién nacidos del 2019 corresponde a 642660 bebés nacidos en ese año. ¿Te imaginas hacer los gráficos y calcular las medidas descriptivas a mano? ¡Nunca acabamos! Es por ello que se utilizan *softwares estadísticos* como `STATA`, `R`, `Python`, `Julia`, `SPSS`, etc. En la actualidad los más utilizados por las empresas (¡y que son gratuitos!) son `R` y `Python`.
- La redacción de conclusiones implica un manejo adecuado de competencias propias de la lectura y de la escritura. Nuestros resultados deben ser entendidos para todos nuestros supervisores. Adicionalmente, los gráficos estadísticos deben presentarse bonitos, para generar una armonía visual.
- Laboralmente un matemático, estadístico y/o científico de datos tiene muchas oportunidades como *Analista de Datos*. Las competencias a desarrollar corresponden a: (1) programación en `R` y `Python`, (2) conocimientos de algoritmos estadísticos básicos y (3) manejo de matemáticas para el entendimiento de la estimación y la predicción mediante el uso de técnicas estadísticas. Dentro del programa de matemáticas se ofrecen los cursos de ESTADÍSTICA I, ESTADÍSTICA II y TÓPICOS EN ESTADÍSTICA para desarrollar tales competencias ¡Te esperamos!

Referencias:

[Da1] Dirección de Metodología y Producción Estadística – DIMPE del Departamento Administrativo Nacional de Estadística (DANE) (2014). Reporte técnico de la Encuesta longitudinal de Protección Social - ELPS 2012. Disponible en https://microdatos.dane.gov.co//catalog/194/get_microdata.

[Da2] Dirección de Metodología y Producción Estadística – DIMPE del Departamento Administrativo Nacional de Estadística (DANE) (2014). Reporte técnico de Estadísticas Vitales - EEVV - 2019. Disponible en https://www.dane.gov.co/index.php/estadisticas-por-tema?id=34&phpMyAdmin=3om27vamm65hhkhrtgc8rrn2g4

[Do] Donoso, E., Carvajal, J. A., Vera, C., & Poblete, J. A. (2014). La edad de la mujer como factor de riesgo de mortalidad materna, fetal, neonatal e infantil. *Revista médica de Chile*, *142*(2), 168-174.

[Wn] Wnuk, A., Kozak, M., & Rochalska, M. (2009). Mosaic plots help visualize contingency tables. Example for a questionnaire survey on knowledge of and attitude towards GMO. In *Colloquium Biometricum* (Vol. 39).