Libro 5: Análisis exploratorio de datos

Sitio:	Agencia de Habilidades para el Futuro	Imprimido por:	Eduardo Moreno
Curso:	Estadística y probabilidades para el desarrollo del soft 1° D	Día:	domingo, 1 de junio de 2025, 14:26
Libro:	Libro 5: Análisis exploratorio de datos		

Tabla de contenidos

¿Qué es la Estadística?

¿Qué es la Estadística?

Si bien en la introducción del bloque algo mencionamos, consideramos conveniente definir qué es la Estadística.

Definición:

Etimológicamente la palabra “estadística” deriva del latín status que significa estado o situación. La estadística consiste en un conjunto de técnicas y procedimientos que permiten recoger datos, presentarlos, ordenarlos y analizarlos, de manera que, a partir de ellos, se puedan establecer conclusiones.

También, como mencionamos antes, vamos a enfocarnos en el estadística descriptiva. En el bloque de probabilidad hemos dado algunas ideas referidas a la estadística inferencial.

¿Qué es la estadística descriptiva?

¿A qué te suena la palabra “descriptiva”? Exacto, a la idea de mostrar, comentar algo. Los métodos de la estadística descriptiva ayudan a presentar los datos de modo tal que sea más fácil su interpretación. Hay varias formas simples e interesantes de organizar los datos, por ejemplo, en gráficos, que permiten detectar tanto las características sobresalientes como las características inesperadas. El otro modo de describir los datos es resumirlos en medidas que pretenden caracterizar el conjunto con la menor distorsión o pérdida de información posible.

En ocasiones un científico solo desea obtener alguna clase de resumen de un conjunto de datos representados en la muestra. En otras palabras, no requiere estadística inferencial. En cambio, le sería útil un conjunto de estadísticos o la estadística descriptiva. Tales números ofrecen un sentido de la ubicación del centro de los datos, de la variabilidad en los datos y de la naturaleza general de la distribución de observaciones en la muestra.

La investigación estadística

En este último libro les proponemos que tomen el rol de un investigador. Así que las explicaciones de los saberes teóricos los iremos contando a medida que vamos elaborando juntos una simulación de investigación.

Obviamente somos consientes de las complejidades que tiene elaborar una investigación, aquí solo pretendemos que ustedes se lleven herramientas que les sirvan en su futura profesión.

Vamos a seguir el siguiente esquema:

1. Formulación del problema.
2. Diseño de la investigación.
3. Recolección y organización de la información.
4. Análisis de la información.
5. Descripción de lo observado.

Podrán observar que varias ideas y conceptos ya fueron construidos en postas anteriores. De igual manera aquí se vuelven a explicar y comentar para quienes necesiten reforzarlos.

Comencemos...

Formulación del problema

Primera parada: Formulación del problema

Cuando una persona emprende el camino de una investigación, es porque tiene "algo" que lo motiva a hacerlo, es decir, el investigador tiene que tener un tema o una problemática que le llame la atención, la cual desconoce, pero maneja ciertas hipótesis o conjeturas, que intentará validarlas o refutarlas con la recolección y el análisis de la información que logre recabar.

En nuestro caso vamos a suponer la siguiente situación. Obviamente bastante simple para poder arrancar, afianzar ideas ya construidas y construir otras nuevas.

Comenzamos nuestra investigación juntos...

El equipo directivo de una institución educativa del nivel secundario, quiere realizar un estudio sobre la conectividad de sus estudiantes en la vida diaria.

Aquí tenemos claramente delimitada la problemática, la cual es la conectividad de los estudiantes de esta institución educativa.

A raíz de la delimitación del tema, surgen una serie de preguntas que iremos tratando de responderlas juntos:

¿Cuáles serán nuestras limitaciones?

Como estamos estudiando una escuela en particular, todas las conclusiones que elaboremos corresponderán a ella. De ninguna manera podemos generalizar a todos los estudiantes de la ciudad, provincia o país.

¿Sobre qué aspectos vamos a indagar?

Si bien ya tenemos definida la problemática, claro está que son varias las características que se pueden abordar cuando hablamos de "conectividad". Por nombrar solamente algunas tenemos:

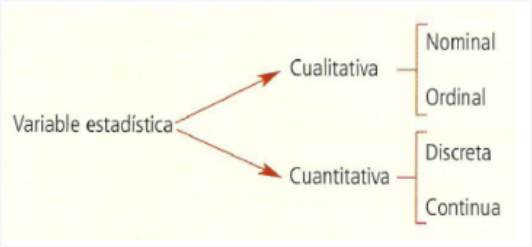
- Uso que le dan a sus celulares (comunicación, entretenimiento, educación, etc.).
- Si cuentan o no con acceso a internet.
- Cantidad y tipo de dispositivos con acceso a internet que hay en todo el grupo familiar (celulares, notebook, tablet, etc.).
- Tiempo diario de pantalla en sus celulares o computadoras.

Las características que analizamos se llaman **variables** . En la próxima sección las definiremos y clasificaremos.

Tipos de variables estadísticas

¿Qué son las variables estadísticas y cómo se clasifican?

Cada tema sobre el que se estudia una muestra y por consiguiente una población, se denomina variable. Las variables pueden ser cualitativas o cuantitativas, las cuales a su vez se subdividen en otras dos:



- **Variables cualitativas o categóricas:** son aquellas que pueden ser representadas a través de símbolos, letras, palabras, etc. Los valores que toman se denominan categorías y los elementos que pertenecen a estas se consideran idénticos respecto a la característica que se está midiendo. Por ejemplo, si la variable es "profesión", los valores que puede tomar son: programador, analista en sistemas, técnico en control de alimentos, etc.

A su vez, dentro de las categóricas encontramos a las nominales (que corresponden a aquellas en las cuales no existe una relación de orden, como el estado civil, el sexo de un individuo, la profesión, etc.), y las ordinales (que son aquellas en las cuales existe un orden implícito, como el nivel educacional, situación económica, etc.).

- **Variables cuantitativas o numéricas:** son aquellas que se pueden medir numéricamente, es decir, los valores que toman este tipo de variables son números. Una variable cuantitativa puede ser discreta o continua:

- Las **variables discretas** son aquellas que toman sus valores en un conjunto finito o infinito numerable. Ejemplo: cantidad de hijos, número de páginas de un libro, etc.
- Las **variables continuas** son aquellas que toman sus valores en un subconjunto de los números reales, es decir, en un intervalo. Ejemplos: la estatura de una persona, el ingreso mensual, etc.

Sigamos pensando juntos nuestra investigación...

En nuestro caso, podemos clasificar las variables antes mencionadas de la siguiente manera:

- Uso que le dan a sus celulares (comunicación, entretenimiento, educación, etc.) → CUALITATIVA NOMINAL.
- Si cuentan o no con acceso a internet → CUALITATIVA NOMINAL.
- Cantidad de dispositivos con acceso a internet que hay en todo el grupo familiar → CUANTITATIVA DISCRETA.
- Tipo de dispositivos con acceso a internet → CUALITATIVA NOMINAL.
- Tiempo diario de pantalla en sus celulares o computadoras. → CUANTITATIVA CONTINUA.

Para hacerlo más simple, trabajaremos con una sola variable la cual será: "cantidad de dispositivos con acceso a internet que hay en todo el grupo familiar".

Diseño de la investigación

Segunda parada: Diseño de la investigación

Siguiendo con las preguntas que surgen al momento de emprender una investigación, nos encontramos con una de las más importantes:

¿Se va a poder indagar a todos los estudiantes?

Como la escuela es una institución muy grande que cuenta con más de 600 estudiantes distribuidos en los diferentes niveles, llegamos a la conclusión de que analizarlos a todos sería muy trabajoso y llevaría más tiempo del que tenemos disponible. Por lo que decidimos seleccionar una parte de estos estudiantes formada por 50 estudiantes.

Aclaración: No vamos a debatir si la cantidad es acorde o no, ya que lo haremos para que sea más fácil manipular los datos en el libro.

Por lo tanto, nuestra **población** de estudio van a ser los 600 estudiantes de la institución educativa y la **muestra** serán los 50 estudiantes. Ya veremos que existen diferentes formas de elegirlos.

¿Qué limitaciones nos trae esto?

Que si estamos en la estadística descriptiva, todas las observaciones que hagamos será siempre en base a la muestra tomada y no a toda la población, ya que para ello necesitaríamos de la inferencia estadística. Sin embargo, para el tipo de trabajo que nos han encomendado, consideramos que una descripción de la muestra es más que suficiente.

¡OK! No podemos indagar a todos los estudiantes, ¿cómo elegimos a esos 50?

Primero, es importante mencionar que una muestra es representativa en la medida que es una imagen de la población.

En general, podemos decir que el tamaño de una muestra dependerá principalmente de:

- Nivel de precisión deseado.
- Recursos disponibles.
- Tiempo involucrado en la investigación.

Existen una gran cantidad de tipos de muestreo. En la práctica los más utilizados son los que explicaremos a continuación.

Tipos de muestreo

Tipos de muestreo:

1. Muestreo aleatorio simple:

Es un método de selección de n unidades extraídas de N, de tal manera que cada una de las posibles muestras tiene la misma probabilidad de ser escogida.

En la práctica, se enumeran las unidades de 1 a N, y a continuación se seleccionan n números aleatorios entre 1 y N, ya sea mediante tablas, utilizando algún programa como Excel o con alguna urna con fichas numeradas.

2. Muestreo aleatorio estratificado:

Se clasifican distintas partes o secciones existentes de la población según alguna característica propia de cada parte (como nivel económico o profesional, por ejemplo), y se toma la muestra en cada una de manera totalmente aleatoria. En este caso hablamos de afijación para referirnos a la importancia relativa que damos a cada sección estudiada por separado:

- si todas las partes tienen el mismo número de muestra será afijación simple,
- si cada parte tiene muestras proporcionales al tamaño de esa parte, afijación proporcional y
- será afijación óptima si el tamaño de la muestra de cada parte responde al análisis de la importancia que tendrá para el dato a estudiar.

3. Muestreo por conglomerado:

Se emplea cuando la población está dividida en grupos o conglomerados pequeños. Es muy similar al anterior, con la diferencia de que aquí las secciones o grupos en que dividimos la población ya existen naturalmente, como los vecinos de un barrio o los trabajadores de un centro comercial.

4. Muestreo sistemático:

Se utiliza cuando las unidades de la población están de alguna manera totalmente ordenadas.

Para seleccionar una muestra de n unidades, se divide a la población en n subpoblaciones de tamaño $K = \frac{N}{n}$ y se toma al azar una unidad de las K primeras (n_0) y de ahí en adelante cada K-ésima unidad, es decir, los elementos de la muestra serán:

$$\{n_0, n_0 + K, n_0 + 2K, \ldots, n_0 + (n-1)K\}$$

En la realidad es posible encontrarse con situaciones en las cuales no es posible aplicar libremente un tipo de muestreo, incluso estaremos obligados a mezclarlas en ocasiones.

Volvamos a nuestra investigación...

Llegó nuestro turno, debemos ver cómo elegir a esos 50 estudiantes que serán nuestra muestra. Como vimos, existen diferentes métodos para garantizar que la muestra sea lo más representativa posible.

Nosotros vamos a considerar que el aspecto de la edad o el nivel educativo, no es un factor determinante sobre la cantidad de dispositivos con acceso a internet que hay en el grupo familiar. Es por esto, que optamos por un muestreo aleatorio simple. A cada uno de los 600 estudiantes les asignamos un número (del 1 al 600) y elegimos de manera aleatoria (mediante un sorteo o un software) a 50 números, y por consiguiente, a los 50 estudiantes de nuestra muestra. Esta forma de elegir la muestra es uno de los más básicos y prácticos.

Ejemplos de muestreos

Veamos ahora otros ejemplos para comprender algunos tipos de muestreos:

Ejemplo de muestreo aleatorio simple:

Consideremos la producción de celulares de una compañía en un determinado turno de trabajo, la cual es de $N=35$. Para efectos de control de calidad de una de sus partes, se desea extraer una muestra aleatoria simple de tamaño $n=5$. Si los 35 celulares producidos son enumerados del 1 al 35, una posible muestra podrían ser los celulares $4, 7, 15, 24$ y 30.

Ejemplo de muestreo por conglomerado:

Si queremos analizar el rendimiento de los estudiantes de una escuela secundaria en Matemática, una muestra utilizando esta técnica podría consistir en elegir de manera aleatoria el 30% de los estudiantes de cada año.

Ejemplo de muestreo sistemático:

En una empresa, los empleados están ordenados por su número de legajo. Hay 500 empleados y vamos a seleccionar una muestra de 10. O sea, cada subpoblación será de 50 empleados. Elegimos al azar un número entre el 1 y el 10. Supongamos que es el 7. Por lo tanto, los elementos muestrales son:

$$\lbrace 7, 57, 107, 157, 207, 257, 307, 357, 407, 457 \rbrace$$

Observen que el azar está fundamentalmente en la determinación del n_0.

Recolección y agrupamiento de la información

Tercera parada: Recolección y agrupamiento de la información

¿Seguimos con nuestras preguntas orientadoras?

Ya definimos la muestra y la característica que vamos a analizar. Ahora debemos determinar cómo recolectar esa información, para lo cual existen diferentes maneras de hacerlo, las más usuales son las encuestas y entrevistas. No entraremos en detalles en esto pero recomendamos indagar más sobre este punto que es muy importante, ya que en base a lo que recabemos acá seguirá todo lo demás.

Supongamos que contactamos a los 50 estudiantes de la muestra y les hicimos una encuesta, la cual contenía la siguiente pregunta: ¿cuántos dispositivos con acceso a internet hay en tu grupo familiar? De cada uno obtuvimos una respuesta y las 50 se muestran en el siguiente cuadro:

4	7	4	7	1
2	1	4	8	3
0	1	7	1	0
1	2	4	1	3
2	4	1	2	2
2	3	1	4	1
4	3	4	3	3
4	1	3	3	4
4	2	3	3	9
4	0	5	2	7

Acá tenemos la información "en crudo" podríamos decirlo, pero dispuesta de esta manera no podemos observar demasiado. Así que ahora viene el momento de empezar a organizar la información.

Agrupamiento de la información

Como verán, los datos así distribuidos no nos dicen demasiado, por lo que es conveniente ordenarlos de una mejor manera para poder analizarlos de forma más sencilla. Una forma es armando lo que se conoce como "tabla de frecuencias". Ya hemos hablado de varios tipos de frecuencias pero vamos a volver a mencionarlas por si necesitan recordarlas:

- **Frecuencia absoluta (f_a):** es la cantidad de veces que se repite un dato de la variable.
- **Frecuencia relativa (f_r):** es el cociente entre la frecuencia absoluta y el tamaño de la muestra. Corresponde a la parte que representa ese dato en toda la muestra. Muchas veces a este valor se lo multiplica por 100 para obtener la frecuencia relativa porcentual.
- **Frecuencia absoluta acumulada (F_a):** es la suma de las frecuencias absolutas que se van acumulando hasta ese dato.
- **Frecuencia relativa acumulada (F_r):** es la suma de las frecuencias relativas que se van acumulando hasta ese dato.

En nuestro caso, la tabla de frecuencias completa nos queda así:

Cantidad de dispositivos	f_a	f_r	F_a	F_r
0	3	$0,06$	3	$0,06$

Cantidad de dispositivos	f_a	f_r	F_a	F_r
1	10	$0,2$	13	$0,26$
2	8	$0,16$	21	$0,42$
3	10	$0,2$	31	$0,62$
4	12	$0,24$	43	$0,86$
5	1	$0,02$	44	$0,88$
6	0	0	44	$0,88$
7	4	$0,08$	48	$0,96$
8	1	$0,02$	49	$0,98$
9	1	$0,02$	50	1
Totales	50	1		

¿Cómo se hizo para completar, por ejemplo, el renglón del valor 2?

- Frecuencia absoluta: se contó todas las veces que se repitió el dos.
- Frecuencia relativa: el valor de la frecuencia absoluta (8) se lo dividió por la cantidad total de datos (50).
- Frecuencia absoluta acumulada: se suman todas las frecuencias absolutas que son menores o iguales que la del valor 2.
- Frecuencia relativa acumulada: se suman todas las frecuencias relativas que son menores o iguales que la del valor 2.

Análisis de la información

Cuarta parada: Análisis de la información

Para analizar los datos recolectados, podemos calcular medidas que nos permitirán resumir la información y establecer conclusiones.

En particular, vamos a indagar sobre las medidas de tendencia central, de dispersión o variabilidad y de posición. Cada una de ellas nos aportarán un dato diferente y la integración de todas, una descripción más detallada y completa.

Además, visualizaremos la información mediante gráficos estadísticos exploratorios que permitan la identificación de patrones, de casos atípicos y la elaboración de conjeturas sobre lo observado.

Medidas de tendencia central

Las medidas de tendencia central son parámetros estadísticos que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: media, mediana y moda.

Para explicar la manera de calcularlas, vamos a suponer que las observaciones en una muestra son $x_1, x_2, \dots, x_n$ y que n es el tamaño muestral.

Media o promedio muestral:

La media o promedio muestral, es un punto de equilibrio entre todos los datos recolectados, representa una distribución equitativa entre todos ellos.

Entonces, para calcular la media de la muestra, que se denota $\bar{x}$, hacemos:

$$\bar{x} = \displaystyle\sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1+x_2+...+x_n}{n}$$

O bien,

$$\bar{x} = \displaystyle\sum_{i=1}^{n} \frac{x_i \cdot f_a(x_i)}{n}$$

Como el cálculo de la media toma todos los datos de la muestra, a veces los valores extremos pueden condicionar su valor. Por ejemplo, si las notas de dos estudiantes son $5,7$ y $2,10$, ambos tienen promedio 6. Esto conduce a definir otra medida de tendencia central que es la mediana y más adelante, las medidas de dispersión.

Mediana de la muestra:

Como ya dijimos anteriormente, el propósito de la mediana es reflejar la tendencia central de la muestra de manera que no sea influida por los valores extremos.

La mediana, que se denota $\tilde{x}$, es el valor que se encuentra en el medio de un conjunto de datos ordenados de manera creciente o decreciente. La mediana divide la muestra de forma tal que deja igual cantidad de datos a su izquierda que a su derecha, es decir, el 50%.

Para calcular la mediana de la muestra, primero debemos ordenar las observaciones de manera creciente o decreciente. Luego, hacemos lo siguiente:

$$\tilde{x} = \begin{cases} x_{(n+1)/2} \text{~~~~~~~si n es impar,} \\ \frac{1}{2}(x_{n/2}+x_{n/2+1}) \text{~~si n es par} \end{cases}$$

Lo anterior significa que, si el tamaño de la muestra es impar (por ejemplo $n=25$), la mediana va a ser la observación x_{13} ($(25+1)/2$). Mientras que si el tamaño muestral es par, (por ejemplo $n=30$), acá no vamos a encontrar un solo número que esté justo en el medio, por lo que se debe hacer un promedio de los dos centrales. En nuestro ejemplo sería $\frac{1}{2}(x_{15}+x_{16})$.

Moda o modo:

La moda (o también llamado modo) corresponde a aquel valor de la variable que se repite la mayor cantidad de veces, es decir, el valor que tiene mayor frecuencia absoluta. Por lo tanto, para obtener esta medida, no hace falta realizar ningún cálculo.

En ocasiones, puede ser que dos o más valores tengan la misma frecuencia absoluta y sea la mayor de todas.

Vamos a calcular las medidas de tendencia central para nuestra investigación...

1) Media:

$$\bar{x}= \frac{0 \cdot 3 + 1 \cdot 10 + 2 \cdot 8 + 3 \cdot 10 + 4 \cdot 12 + 5 \cdot 1 + 6 \cdot 0 + 7 \cdot 4 + 8 \cdot 1 + 9 \cdot 1 }{50} = \frac{154}{50} = 3,08$$

Esto significa que en promedio, los estudiantes de nuestra muestra tienen 3 dispositivos con conexión a internet.

2) Mediana:

Como el tamaño muestral es par ($n=50$), debemos calcular el promedio de los datos que ocupan las posiciones 25 y 26. Mirando la frecuencia acumulada, vemos que los valores 2 ocupan hasta la posición 21, y los 3 van hasta la posición 31. Esto implica que los valores x_{25} y x_{26} son ambos 3. Por lo tanto, el cálculo de la mediana nos queda:

$$\tilde{x} = \frac{3+3}{2}=3$$

Esto nos indica que el valor 3 deja por debajo y por encima al 50% de los datos. El hecho de que su valor esté cercano a la media, hace más representativo al valor promedio, indicando que los valores extremos no influenciaron demasiado en su cálculo.

3) Moda:

Para esta última medida de tendencia central, solo debemos buscar en la tabla el valor que tenga mayor frecuencia absoluta, ya que esto indica que es el dato que más veces se repitió. Es por esto que la moda es el valor 4.

Observen que si bien nos da que el centro de los datos está en el valor 3, éste no fue el que más veces se repitió. De todas maneras, es un valor muy cercano, siendo más importante los dos primeros resultados.

Medidas de dispersión

Ya hemos hablado mucho sobre las medidas de variabilidad. Estudiamos que hay muestras que pueden tener la misma media, pero que los datos pueden estar mayor o menor agrupados a su alrededor. Vamos a volver a definir las siguientes medidas de variabilidad: rango, varianza, desviación estándar y coeficiente de variación.

Nuevamente denotamos con $x_1,x_2,...,x_n$ los valores de la muestra.

Rango muestral:

Es quizás la medida de dispersión más simple, ya que consiste en la diferencia entre la observación máxima de la muestra y la observación mínima. En símbolos:

$$R= x_{máx}-x_{mín}$$

Podemos entender que si el rango es un valor alto, hay una mayor distancia entre el valor máximo y mínimo de la muestra y viceversa.

Varianza muestral:

Es un promedio de los desvíos cuadráticos de cada valor de la variable respecto a la media o promedio, o dicho de otra manera, es una medida de la desviación cuadrática promedio de la media $\bar{x}$.

Por lo tanto, la varianza muestral, que se denota con s^2, está dada por:

$$s^2= \displaystyle\sum_{i=1}^{n} \frac{(x_i- \bar{x})^2}{n-1}$$

O bien,

$$s^2= \displaystyle\sum_{i=1}^{n} \frac{(x_i- \bar{x})^2 \cdot f_a(x_i) }{n-1}$$

Una variabilidad grande en un conjunto de datos produce valores relativamente grandes de $(x- \bar{x})^2$ y, por consiguiente, una varianza muestral grande. La cantidad $n - 1$ a menudo se denomina **grados de libertad asociados con la varianza** estimada.

Observen que empleamos el término desviación cuadrática promedio aun cuando la definición utilice una división entre $n - 1$ grados de libertad en vez de n. Desde luego, si n es grande, la diferencia en el denominador es inconsecuente.

La varianza de la muestra tiene unidades que son el cuadrado de las unidades en los datos observados.

Desviación estándar de la muestra:

Es el valor esperado de la separación de los valores de la variable (valores observados) con respecto a la media. Indica cuánto se alejan en promedio cada uno de los valores de la variable de la media.

La desviación estándar de la muestra, denotada con s, es la raíz cuadrada positiva de la varianza, es decir,

$$s= \sqrt[]{s^2}$$

Su valor expresa de modo absoluto y en las mismas unidades que aquellas que se utilizan para medir las observaciones individuales de la muestra.

Una desviación estándar de poco valor absoluto indica que la dispersión de la muestra alrededor de la media es pequeña, es decir, los valores están concentrados, y viceversa.

Coeficiente de variación:

Otra forma de evaluar la variación en una muestra es considerar la variación relativa mediante el cálculo del coeficiente de variabilidad, simbolizado por CV. El valor de dicho coeficiente se define como la relación entre la desviación estándar y la media.

$$CV= \frac{s}{\bar{x}}$$

Vamos a calcular las medidas de dispersión para nuestra investigación...

1) Rango:

$$R=9-0=9$$

Implica que la distancia entre los valores extremos es de 9 unidades.

2) Varianza:

$$s^2= \frac{(0-3,08)^2 \cdot 3 + (1-3,08)^2 \cdot 10 + (2-3,08)^2 \cdot 8 + (3-3,08)^2 \cdot 10 + (4-3,08)^2 \cdot 12 + (5-3,08)^2 \cdot 1 + (7-3,08)^2 \cdot 4 + (8-3,08)^2 \cdot 1 + (9-3,08)^2 \cdot 1 }{50-1} = \frac{215,68}{49} \approx 4,4016$$

3) Desviación estándar:

$$s \approx \sqrt[]{4,4016} \approx 2,098$$

Esto nos indica que en promedio cada valor de la variable difiere en $2,098$ unidades de la media, ya sea por encima o por debajo.

Este valor adquiere también relevancia cuando comparamos más de una muestra.

4) Coeficiente de variación:

$$CV= \frac{2,098}{3,08} \approx 0,6812$$

Otras medidas de posición

A continuación, explicaremos otras medidas de posición que son los cuartiles, deciles y percentiles. No nos abocaremos a su cálculo, solo a su interpretación, ya que esto lo haremos con el software.

Así como la mediana divide a la serie de datos en dos grupos, los **cuartiles** dividen a la serie de datos en cuatro grupos de igual cantidad de elementos, los cuales se denotan Q_1, Q_2 y Q_3. El segundo cuartil coincide con la mediana.

Por su parte, los **deciles** dividen a la serie de datos en diez grupos de igual cantidad de elementos ($D_1, D_2, \dots, D_9$). El D_5 coincide con la mediana.

Por último, los **percentiles** dividen a la serie de datos en cien grupos de igual cantidad de elementos ($P_1, P_2, \dots, P_{99}$). El P_{50} coincide con la mediana.

Gráficos estadísticos: de barras e histogramas

¿De qué otra manera se puede presentar la información?

Ya vimos que una manera de organizar y presentar la información es mediante las tablas de frecuencias. Sin embargo, estas son útiles para el trabajo del investigador pero no para comunicar las datos recolectados. Es ahí cuando aparecen los gráficos estadísticos.

Cuando se tienen datos sobre muchos casos, los gráficos son una manera resumida y atractiva de mostrar la información. Existen diferentes tipos de gráficos y estos pueden variar según el tipo de información que se quiere brindar. Veremos los siguientes: histogramas, gráficos de barras, tallo y hojas, caja y bigotes, de torta o circular y pictogramas.

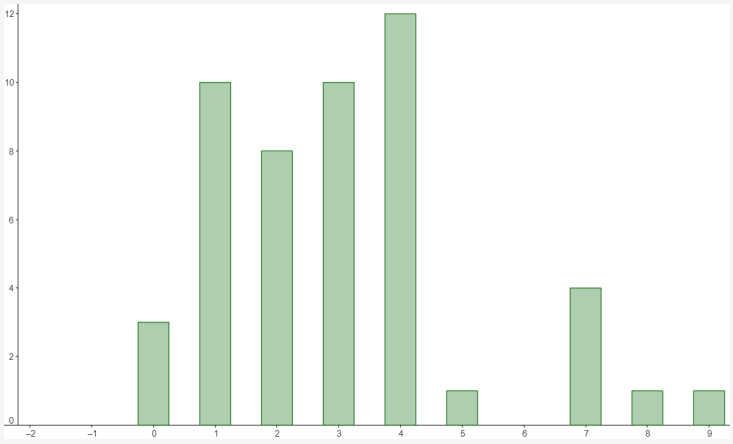
Si bien acá lo explicaremos y expondremos algunos ejemplos, recomendamos ingresar a la videoteca de esta posta para ampliar la explicación.

1) Histogramas y gráficos de barras:

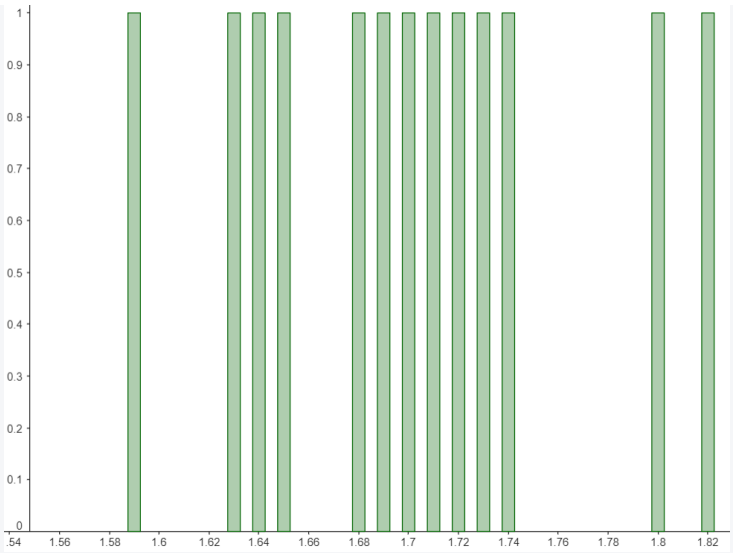
Ya hemos trabajado bastante con este tipo de gráficos. Su construcción es similar en ambos, con rectángulos de igual base y cuya altura corresponde con la frecuencia absoluta (o relativa). En lo que se diferencian es que los gráficos de barras se emplean para variables cuantitativas discretas y los histogramas para las continuas, ya que en estas últimas suele ser conveniente agrupar los datos en intervalos de clases.

Seguimos con nuestro ejemplo...

En nuestro caso, como la variable analizada es discreta, empleamos el gráfico de barras y nos queda así:



En cambio, si estuviéramos analizando la altura de los estudiantes, podemos generar grupos que abarquen a varias, ya que pueden ser todas diferentes. Observen que un gráfico de barras no nos permitiría realizar una observación acorde:



En cambio, con un histograma la situación cambia mucho:

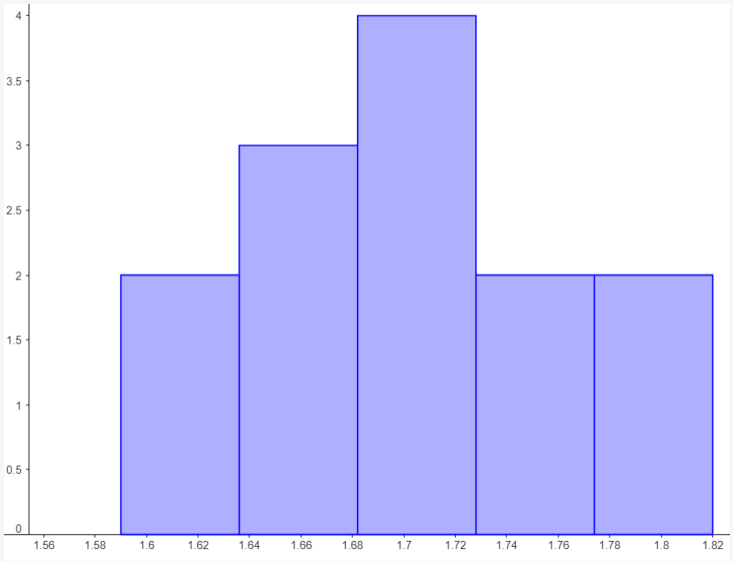


Diagrama de tallo y hojas

El diagrama de tallo y hojas es útil porque no solo que permite organizar los valores de la variable, sino que al mismo tiempo nos va generando un gráfico con aportes visuales.
Existen diferentes formas de organizar este diagrama y todo dependerá de los datos con los que se cuente. Pero en líneas generales, la idea es atribuir a una parte del número para que sea el tallo y a la otra parte, la hoja. Por ejemplo, si tenemos datos con dos cifras, el tallo puede ser la primera cifra y la hoja la segunda. O si tenemos números con decimales, el tallo puede ser la parte entera y la hoja, la parte decimal.

Para nuestra ejemplo, no tiene sentido realizar este diagrama ya que todos los números son de una cifra. Pero supongamos que analizamos otro aspecto de la conectividad como por ejemplo el tiempo promedio diario de pantalla en el celular. Para ilustrar mostramos 20 posibles valores:

3,8	4,6	3,9	4,5	7,3	9,1	8,6	2,4	3,6	3,7
4,9	5,3	6,6	6,2	4,8	5,6	5,3	9,2	3,8	2

Entonces, el diagrama de tallo y hojas nos queda:

2	0	4
3	6	7 8 8 9
4	5	6 8 9
5	3	3 6
6	2	6
7	3	
8	6	
9	1	2

La clave 3|1 significa 3.1

Lo que está al lado izquierdo de la línea vertical son los tallos y al lado derecho, las hojas. Entonces, todos los números que tienen parte entera 3 se ubican en ese tallo. En ese renglón tenemos los siguientes valores: $3,6$-$3,7$-$3,8$-$3,8$ y $3,9$.

También, este gráfico nos permite ver que la mayoría de los valores comienzan con las 3 horas, ya que es el renglón con más hojas.

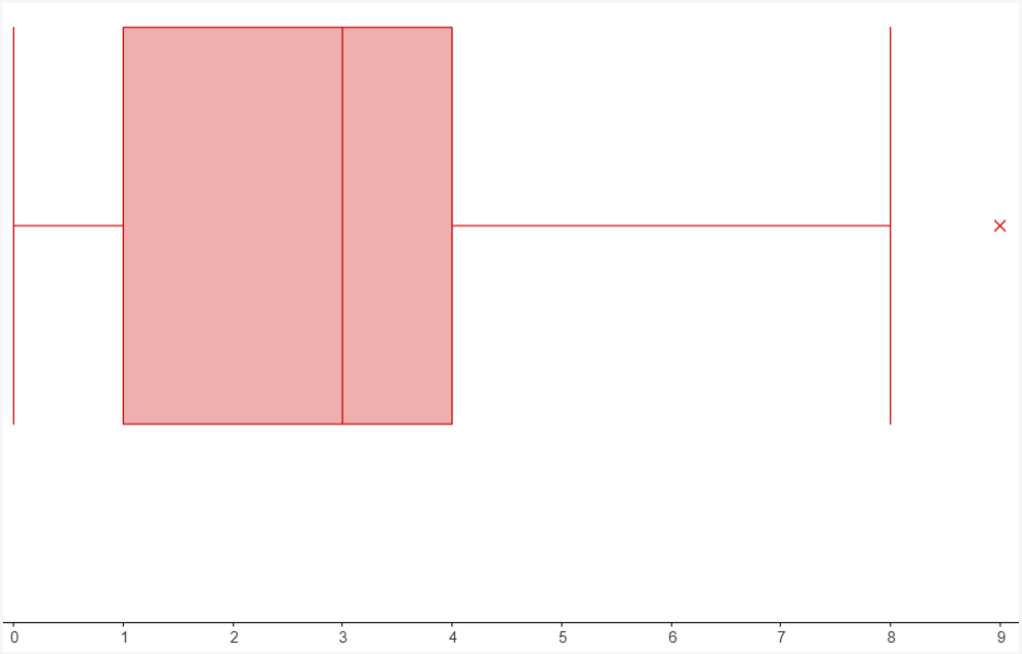
Caja y bigotes

Otra presentación que es útil para reflejar propiedades de una muestra es la gráfica de caja y bigote, la cual encierra el rango intercuartil de los datos en una caja que contiene la mediana representada. El rango intercuartil tiene como extremos el cuartil 3 (cuartil superior) y el cuartil 1 (cuartil inferior). Además de la caja se prolongan "bigotes", que indican las observaciones alejadas en la muestra. Para muestras razonablemente grandes la presentación indica el centro de localización, la variabilidad y el grado de asimetría.

Además, una variación de este tipo de gráficos puede ofrecer al observador información respecto de cuáles observaciones son valores atípicos. Los valores atípicos son observaciones que se consideran inusualmente alejadas de la masa de datos. Existen muchas pruebas estadísticas diseñadas para detectar este tipo de valores. Técnicamente se puede considerar que un valor atípico es una observación que representa un "evento raro" (existe una probabilidad pequeña de obtener un valor que esté lejos de la masa de datos).

Aunque la determinación de cuáles observaciones son valores atípicos varía de acuerdo con el tipo de software que se emplee, un procedimiento común para determinarlo consiste en utilizar un múltiplo del rango intercuartil. Por ejemplo, si la distancia desde la caja excede $1{,}5$ veces el rango intercuartil (en cualquier dirección), la observación se podría considerar un valor atípico.

¿Cómo es el diagrama de caja y bigotes para nuestra investigación?



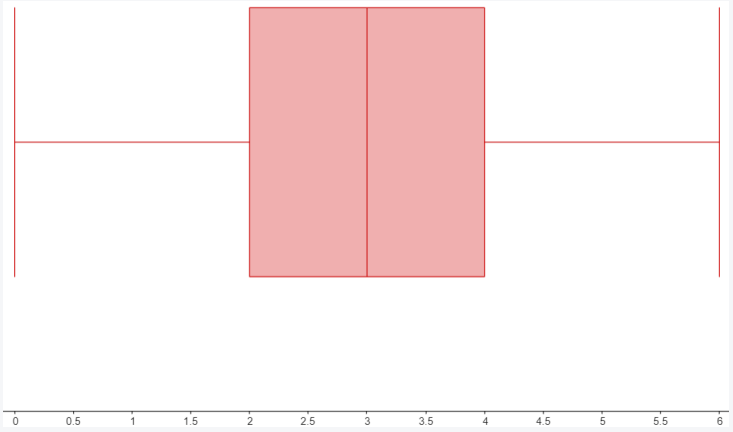
Estas son las partes básicas de un diagrama de caja y bigotes:

- La línea central de la caja indica la mediana de los datos. Una mitad de los datos está por debajo de este valor, y la otra por encima. Si los datos son simétricos, la mediana estará en el centro de la caja. Vemos que en nuestro caso, la mediana está un poco hacia la derecha.
- Los extremos de la caja indican los cuartiles 1 y 3. La longitud de la caja es la diferencia entre estos dos cuartiles y se conoce como rango intercuartílico. El tamaño de la caja es importante, ya que si la caja es grande, indica más dispersión de los datos y viceversa.
- Las líneas que se extienden desde la caja se llaman bigotes. Los bigotes representan la variación tolerada de los datos. Si los datos no llegan hasta el final de los bigotes, estos se ajustan a los valores mínimo y máximo de los datos. Observen que el bigote derecho es más extenso que el izquierdo, esto se debe a que

tenemos algunos valores grandes que fueron abarcados dentro del rango de tolerancia. El bigote izquierdo quedó más corto porque llegó hasta el valor mínimo y no hizo falta prolongarlo más.

- Si hay datos que quedan por encima o por debajo de los extremos de los bigotes, se los representa con puntos o cruces. Estos puntos se conocen como valores atípicos. En nuestro caso tenemos uno solo que es el valor \(9 \), lo cual nos da a entender que sería poco usual que los estudiantes tengan esta cantidad de dispositivos con acceso a internet.

Observen cómo los datos de esta otra muestra son más simétricos y no presentan valores atípicos:



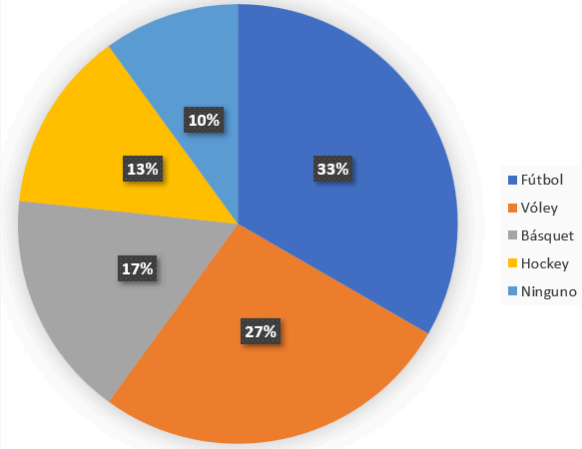
Otros tipos de gráficos

Para completar nuestro apartado de gráficos estadísticos, no podíamos dejar de mencionar dos gráficos que son muy utilizados en los medios de comunicación y redes sociales: el circular, de sectores o de torta y los pictogramas.

Gráfico circula, de sectores o de torta:

En este gráfico los datos son representados mediante sectores de un círculo. Cada uno indica diferentes categorías de la variable y el tamaño del sector es proporcional a la cantidad de veces que ese valor se repite, así a mayor cantidad de datos en una característica, mayor será el sector en el círculo y viceversa.
Su uso se limita más a las variables cualitativas.

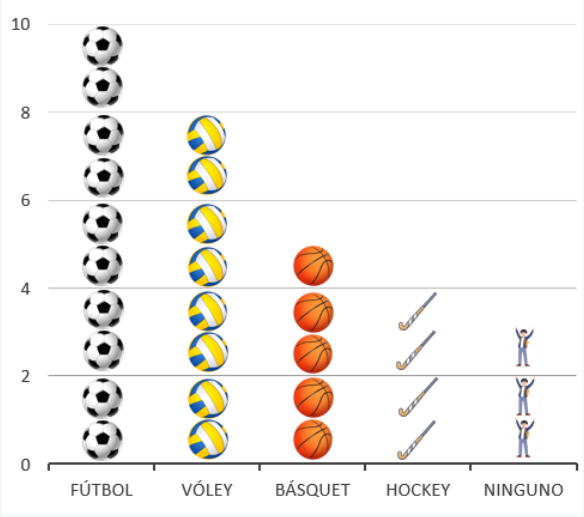
Por ejemplo, el siguiente gráfico muestra los deportes que realizan los estudiantes de un curso del nivel secundario:



Pictogramas:

El pictograma es un gráfico estadístico que se suele utilizar para caracteres cualitativos y que en lugar de barras para representar las frecuencias, utiliza dibujos alusivos a cada atributo, cuya dimensión es proporcional a la frecuencia absoluta.

La misma información que el gráfico circular se muestra en el siguiente pictograma:


10
8
6
4
2
0
FÚTBOL
VÓLEY
BÁSQUET
HOCKEY
NINGUNO

Descripción de lo observado

¿Cómo finalizamos nuestra investigación?

Hemos recolectado la información, la agrupamos en tabla, calculamos estadísticos para analizar el centro y la variabilidad de los datos, construimos gráficos para poder visualizar mejor estos aspectos y otros como la simetría y valores atípicos. Entonces ahora sí estamos en condiciones de poder establecer un informe describiendo todo lo observado y estableciendo algunas conclusiones, siempre teniendo en cuenta el alcance de nuestro estudio.

Quinta y última parada: Descripción de lo observado

Pudimos ver que en promedio los estudiantes de nuestra muestra cuentan con tres dispositivos con acceso a internet, teniendo en cuenta todo el grupo familiar. Este dato luego se afianzó con la mediana que mostró también que el 50% de los estudiantes se encuentran por encima y por debajo de este valor. Debemos remarcar de igual manera que el valor que más se repitió fue el 4 (su moda). Todo esto pudo apreciarse con el diagrama de barras, donde en estos valores encontramos los rectángulos de mayor altura.
Por otra parte, también estudiamos la variación de los datos. El desvió estándar fue de aproximadamente 2 celulares en este caso. ¿Qué nos muestra esto? Que en promedio cada valor de la variable dista de la media 2 unidades, es decir, esperaríamos que los valores estén o dos unidades por encima o dos por debajo del 3 (su media).
Finalmente, el diagrama de cajas y bigotes nos aportó que nuestra muestra no es simétrica, sino que la mediana está un poco hacia la derecha. Pero lo más importante es que vimos un valor atípico, lo cual nos indica que es poco común que los estudiantes de esta muestra tengan 9 celulares.

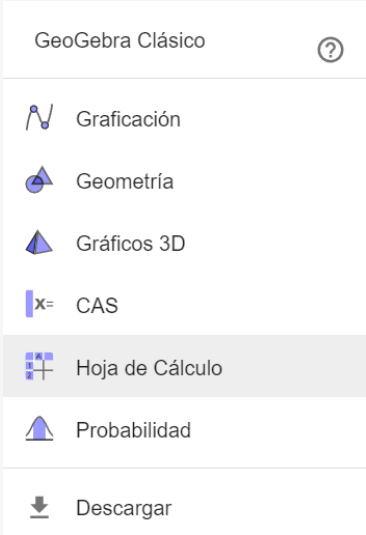
Claro está que ésta es una descripción posible, la cual seguramente pueda ser complementada con otras observaciones. Pero en líneas generales esto es lo básico que esperaríamos leer en un informe descriptivo.

En GeoGebra

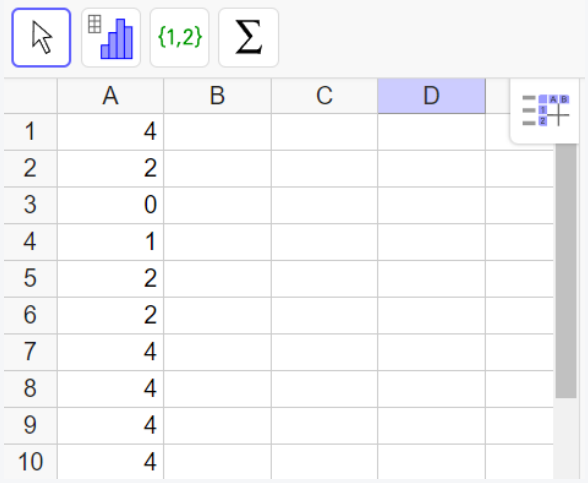
Todo el análisis de la información que explicamos anteriormente y mostramos el proceso "manual", se puede obtener de una manera simple y rápida empleando el GeoGebra.

1. Primero lo que debemos hacer es abrir el GeoGebra haciendo clic en el siguiente enlace: **https://www.geogebra.org/classic**

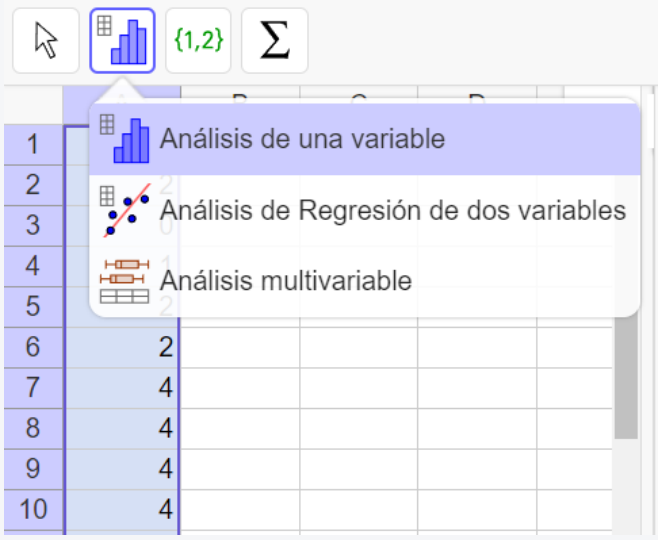
2. Una vez dentro, en el menú derecho elegir la opción "Hoja de Cálculo".



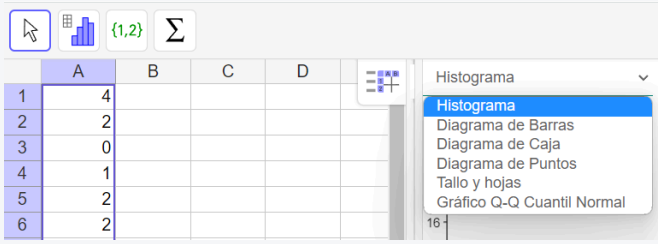
3. Se nos abrirá una grilla similar a la de un Excel. De manera vertical y uno debajo del otro, anotamos los valores de la variable estadística analizada.



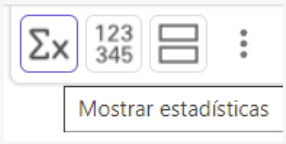
4. Seleccionar todos los datos (en nuestro caso los 50). Controlar que estén todos de color azul. Luego, seleccionar "análisis de una variable".

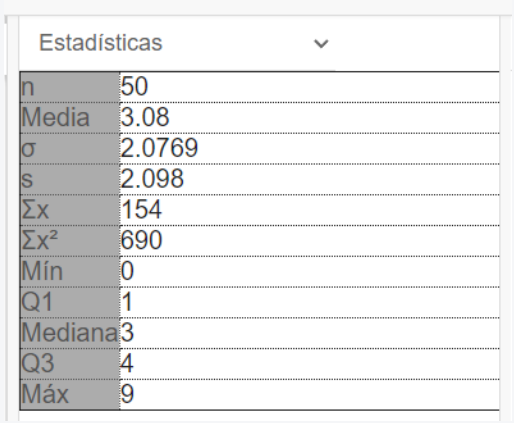


5. Aquí encontraremos un menú para los diferentes gráficos disponibles:



Además, si hacemos clic en la sumatoria, nos desplegará (entre otras cosas) el cálculo de las medidas de tendencia central, dispersión y posición:





y como podrán observar, estos resultados coinciden con los que nosotros calculamos "a mano".

Si bien la varianza muestral no aparece, podemos calcularla elevando al cuadrado el desvío estándar.

De esta manera vemos como una vez más los software nos facilitan el trabajo de realizar los cálculos y construcciones de gráficas, y nos podemos abocar al trabajo más importante que es la interpretación de los resultados y la descripción de lo observado.

Comparación de muestras

Comparemos más de una muestra de una misma población.

Supongamos que queremos analizar el peso de un perro salchicha adulto. Entonces decidimos tomar cinco muestras de 10 perros salchichas adultos (diferentes).

Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5
4.95	5.6	5	4	5
5.03	5.32	5	4	4.5
5.25	4.98	5.9	4.65	5.5
4	5	5.3	4.26	5.2
4.57	4.99	5.4	5	4.8
5.3	4.32	5.78	4.69	5.3
5	4.7	5.25	4.78	4.7
4.58	5.2	4.99	4.99	4
4.5	5.3	5.67	5.2	5.6
4.03	5.8	5.48	4.6	4.4

De cada muestra calculamos la media y el desvío estándar:

Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5
$\bar{x_1}=4,721$	$\bar{x_2}=5,1205$	$\bar{x_3}=5,377$	$\bar{x_4}=4,617$	$\bar{x_5}=4,9$
$s_1=0,463$	$s_2=0,4269$	$s_3=0,3315$	$s_4=0,4147$	$s_5=0,5142$

Ahora si tomamos las cinco medias muestrales como un conjunto de datos propio, le podemos calcular su desviación estándar. Esto es lo que se conoce como error estándar de la media, que definimos a continuación:

Definición:
En pocas palabras, el **error estándar de la media** de la muestra es una estimación de qué tan lejos es probable que la media de la muestra esté de la media de la población.

Otra investigación más

Para finalizar, vamos a dejar un ejemplo de otra simulación de investigación, para que sigamos afianzando las ideas nuevas que construimos en este último libro y también, para reforzar las que ya han aparecido a lo largo de toda esta materia.

Situación:

El gobierno municipal de una mediana ciudad nos encomienda la tarea de indagar sobre la situación de los negocios de rotisería que hay en ella.

1. Formulación del problema:

Situación de los negocios de rotisería de la ciudad, en lo que va del año 2023.

2. Diseño de la investigación:

Logramos acceder a la base de datos donde se registran los comercios y observamos que existen 200 negocios de rotisería distribuidos por diferentes barrios de la ciudad. Ante esto, pensamos que si analizamos a todos sería muy costoso y demandaría mucho tiempo, por lo que decidimos discriminar por sectores la cantidad de rotiserías existentes y tomar (de manera aleatoria) un porcentaje como muestra. Como la intensión es tener un panorama macro, juntamos todos los datos (aunque los dejamos diferenciados por si luego se quiere hacer alguna gestión específica en algún barrio). El tamaño muestral nos arrojó en total a 35 locales de rotiserías.

Nos resta por definir las características o temas que vamos a indagar de este tipo de negocios. Claro está que se nos abre un abanico de posibilidades, por nombrar algunas tenemos:

- Tipo de comidas que ofrecen (comida rápida, mexicana, pastas, café, etc.).
- Momento del día que están abiertos al público (mañana, mediodía, tarde, noche).
- ¿Cuentan con delivery y/o take away?
- Día de la semana con mayor y menor ventas.
- Principales dificultades que enfrentan a diario.
- Ingreso promedio diario del último mes.

Claro está que analizaríamos todas, pero para simplificar nuestro trabajo en el libro nos quedaremos con la última variable, que si la clasificamos es cuantitativa continua.

3. Recolección y agrupamiento de la información:

Supongamos que recorrimos los 35 locales de rotisería que conforman nuestra muestra y les hicimos una encuesta. En la pregunta sobre cuál fue el ingreso promedio diario del último mes, obtuvimos las siguientes respuestas:

15384	14872	17230	16432	16500
15400	19000	18540	15784	16080
16380	13450	16235	15974	17542
14852	14789	13000	16752	18462
16000	16745	16784	16742	17652
15872	14500	9000	18450	15000
12496	15640	25430	13486	16325

Como podrán advertir, los valores son muy diferentes como para hacer una tabla de frecuencias como la que hicimos antes. Se podría armar una empleando lo que se conoce como "intervalos de clase". Sin embargo, como utilizaremos el software GeoGebra, directamente cargamos los datos en él.

4. Análisis de la información:

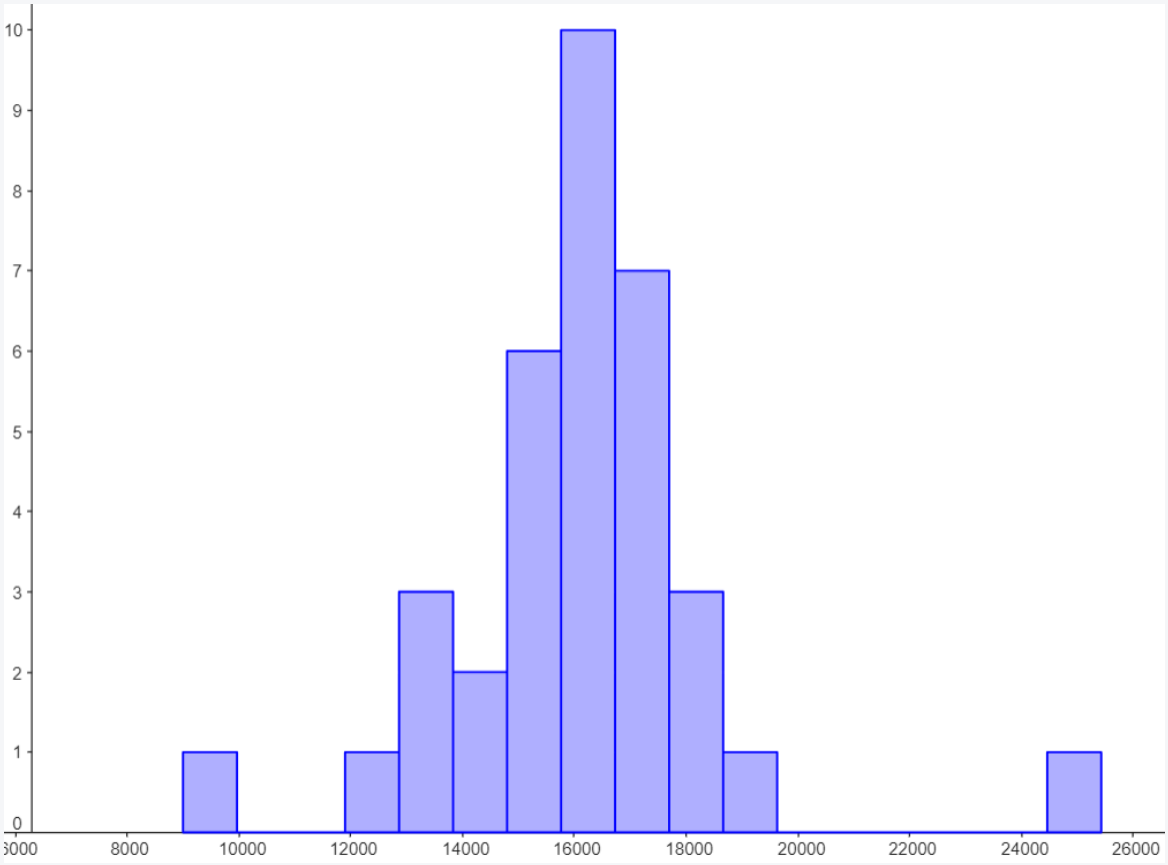
Lo primero que haremos será calcular las medidas de nuestro interés, fundamentalmente la media, mediana y el desvío estándar:

n	35
Media	16079.4286
σ	2491.2058
s	2527.5757
Σx	562780
Σx^2	9266394528
Mín	9000
Q1	14872
Mediana	16080
Q3	16784
Máx	25430

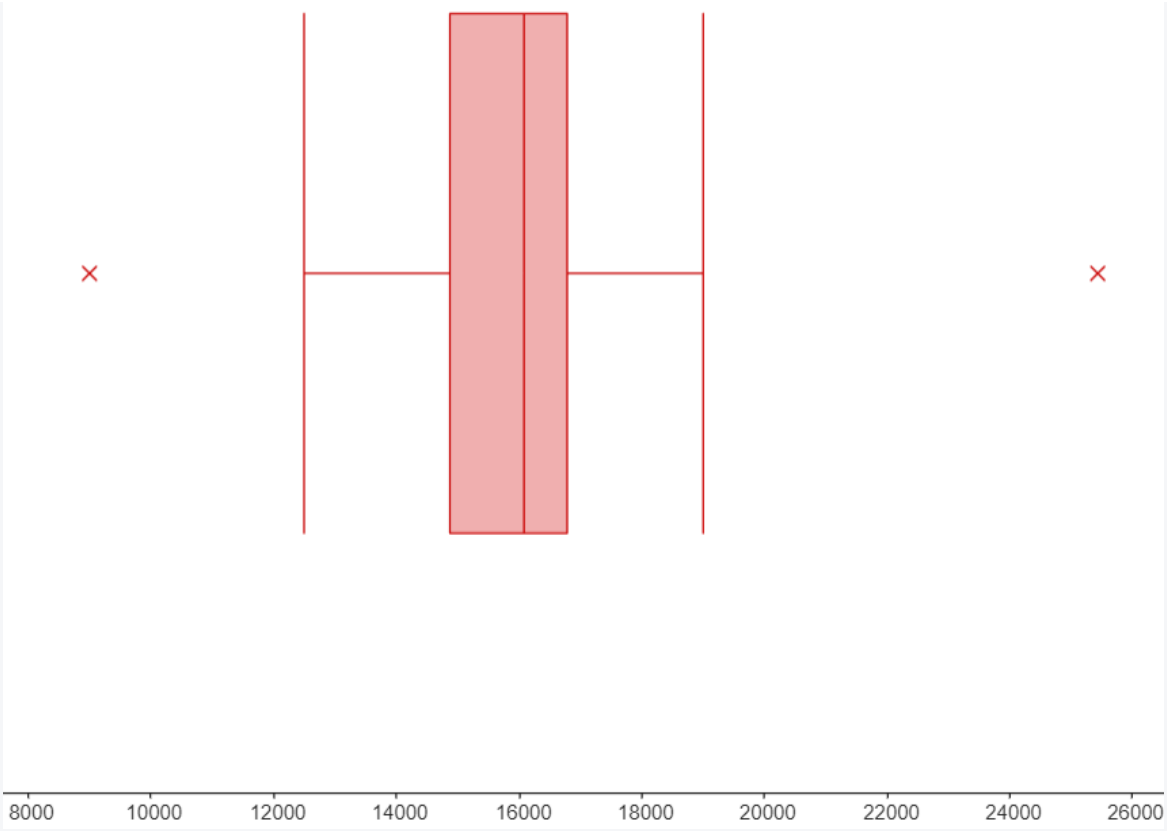
Con esta tabla vemos que la media muestral es aproximadamente \($16079 \), muy similar a la mediana que es de \($16080 \). En cuanto al desvió estándar observamos que es de aproximadamente \($2528 \). Además, tenemos el valor máximo y mínimo, lo cual nos permite obtener el rango que es de \($16430 \).

Ya haremos una interpretación de estos resultados, pero antes vamos a obtener dos gráficos:

- Histograma:



- Diagrama de caja y bigotes:



5. Descripción de lo observado:

En base a la información recolectada a partir de nuestra muestra tomada de manera representativa, observamos que los negocios de rotisería tienen un ingreso diario medio de \($16079 \), cuyo valor está muy cercano a la mediana (\($16080 \)), lo cual indica que ambos son buenos estadísticos para tomar como referencia sobre en donde se encuentra el centro de las observaciones. Estas afirmaciones son apoyadas por el histograma donde vemos que los mayores datos están concentrados alrededor de los \($16000 \).

Nos toca analizar la variabilidad de los datos. El desvió estándar nos muestra que los datos distan de la media en promedio \($2527 \). Si nos apoyamos en el diagrama de caja y bigotes, vemos que los datos no están tan dispersos de la mediana (tampoco de la media ya que son valores casi iguales en este caso). Podemos pensar que el desvío estándar hubiera dado un valor inferior si no tendríamos esos dos valores atípicos en nuestra muestra. Recordemos que la media suele estar influenciada por estos valores extremos.

Por último, podemos ver que el \(50\% \) de los valores por encima de la mediana están menos dispersos que los inferiores, esto lo vemos porque la raya vertical está más a la derecha, lo que hace que la parte de la caja en ese lugar sea más chica.

De esta manera hemos realizado una descripción bastante completa sobre nuestra variable, luego el gobierno municipal verá qué decisiones toma en base a ellas.