Capítulo 2

Estadística Descriptiva

2.1 Descripción Gráfica de Datos

Muchas veces hemos escuchado que "una imagen vale más que mil palabras". Esta observación es especialmente importante cuando se trata de resumir la información contenida en un conjunto de datos. En efecto, un buen resumen gráfico será mucho más fácil de interpretar que una larga lista de datos numéricos.

Para poder obtener representaciones gráficas que resuman en forma eficiente la información, necesitamos tener en cuenta que existen diferentes clases de datos, entre las cuales podemos mencionar las siguientes:

- Variables Cualitativas: Son aquellas variables que indican cualidades del objeto de estudio: sexo, nivel socioeconómico, etc.

- Variables Numéricas Discretas: En general, este tipo de variables indican cuántas veces ha sucedido un determinado fenómeno: número de automóviles vendidos, cantidad de estudiantes que aprueban un curso, etc.

- Variables Numéricas Continuas: En la mayoría de los casos, se obtienen a partir de mediciones.

Ejemplo 2.1 *En cierto distrito de Guatemala, en el año 1969, se entrevistó a un cierto número de mujeres casadas nacidas entre los años 1935 y 1944 y se les preguntó a qué edad contrajeron matrimonio. En la tabla 2.1 se muestran los datos correspondientes a 50 de estas mujeres:*

15	17	25	15	16	11	15	13	12	10
15	14	16	14	17	13	14	20	29	19
16	18	10	18	12	11	20	34	13	22
19	14	17	16	16	15	12	24	25	9
21	15	13	23	24	10	10	16	14	18

Tabla 2.1: Edades de matrimonio de 50 mujeres en cierto distrito de Guatemala, año 1969 (Ejemplo 2.1)

Describamos este conjunto de datos haciendo uso de algunas herramientas gráficas:

a **Diagrama de Puntos:**
Este tipo de diagrama resulta muy útil cuando el número de datos es pequeño (digamos, menor que 50). Se construye poniendo un punto sobre un valor en un eje horizontal cada vez que dicho valor aparece en el conjunto de datos.
Para el ejemplo que estudiamos, el diagrama de puntos tiene la forma que se indica en la figura 2.1.

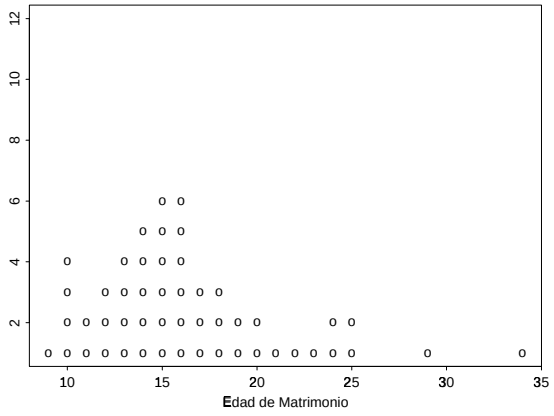

Figura 2.1: Diagrama de puntos para los datos de edad de matrimonio en Guatemala.

Este diagrama de puntos nos permite observar que de las 50 mujeres entrevistadas, más de la mitad se casaron antes de cumplir los 17 años, y sólo 2 de ellas eran mayores

de 25 años en el momento de casarse. Observamos además matrimonios realizados a los 9 o 10 años de edad.

b **Diagrama de tallo y hojas:**
Esta es una representación semigráfica que, si bien es semejante al diagrama de puntos, conserva mayor información. Es útil cuando los datos que se consideran tienen dos o tres cifras significativas. Se construye trazando una línea vertical a la izquierda de la cual se coloca la primera cifra significativa (tallo), y a la derecha de esa cifra se colocan todas las posibles últimas cifras significativas (hojas), tantas veces como aparezcan en los datos.

Para los datos del ejemplo 2.1 el diagrama de tallo y hojas es el siguiente:

0	9
1	000011222333344444555555666666778889
2	0012344559
3	4

Sin embargo, como hay pocos valores para el tallo, el diagrama proporciona pocos detalles. Una manera de solucionar este problema consiste en dividir cada tallo en dos o más clases. Por ejemplo:

0	9
1	000011222333344444
1	55555666666677788899
2	0012344
2	559
3	4

Para cada valor del tallo (decena), la primera clase contiene las observaciones con hojas entre 0 y 4 (unidades), mientras que la segunda contiene las hojas entre 5 y 9.

En este diagrama vuelve a observarse la tendencia a contraer matrimonio a una edad muy temprana.

c **Distribución de frecuencias (Histograma):**
Consiste en agrupar los datos por clases y mostrar el número de observaciones (frecuencia absoluta) o el porcentaje de éstas (frecuencia relativa) que hay en cada clase.

Puede ser presentada en forma tabular o en forma gráfica. A esta última se le suele llamar *histograma*.

Los pasos principales para la construcción del histograma son los siguientes:

- Establecer las clases: Para ello es necesario decidir el número de clases que se van a emplear y la longitud de las mismas.
  - Número de clases: Usualmente se toman entre 5 y 20 clases, pues menos de cinco proporcionan muy poco detalle, y más de 20 proporcionan un detalle excesivo que resulta de poca utilidad.
    Una regla empírica para la selección del número de clases consiste en tomar aproximadamente $\sqrt{n}$ clases, donde n es el número de datos.
  - Longitud de las clases: Habitualmente se toman clases de igual tamaño, es decir, si decidimos usar k clases, la longitud de cada una de ellas se obtiene como $\frac{(X_{max}-X_{min})}{k}$, donde X_{max} y X_{min} corresponden a los valores máximo y mínimo de los datos.
- Ordenar los datos en cada clase.
- Contar el número de datos en cada clase.
- Presentar los resultados.

En el ejemplo 2.1 tenemos 50 datos, y $\sqrt{n} = 7.071$. Además, $X_{max} - X_{min} = 34 - 9 = 25$. Como el múltiplo de 7 más cercano a 25 es 28, tomaremos 7 clases de longitud 4. En la tabla 2.2, se describen las clases, así como sus frecuencias absolutas y relativas.

El 64% de los datos se agrupa en la segunda y tercera clase, mientras que las tres últimas (mujeres de 23 o más años) apenas totalizan un 12%. Una vez mas, observamos que la mayor parte de estas mujeres se casaron muy jóvenes, en su mayoría entre los 11 y 19 años, lo cual corrobora las observaciones que hemos hecho anteriomente.

La representación gráfica correspondiente puede verse en la figura 2.2.

Si se desea, es posible usar clases de diferente longitud. En ese caso es el área de la barra, y no su altura, la que debe ser proporcional a la frecuencia. De lo contrario, la impresión visual producirá conclusiones incorrectas.

d Otros diagramas posibles incluyen diagramas de barra, gráficos XY, etc.

Límites de Clase	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada
$7 < X \leq 11$	7	0.14	0.14
$11 < X \leq 15$	18	0.36	0.50
$15 < X \leq 19$	14	0.28	0.78
$19 < X \leq 23$	5	0.10	0.88
$23 < X \leq 27$	4	0.08	0.96
$27 < X \leq 31$	1	0.02	0.98
$31 < X \leq 35$	1	0.02	1

Tabla 2.2: Distribución de frecuencias para los datos de edad de matrimonio en Guatemala (Ejemplo 2.1)

2.2 Medidas Numéricas Descriptivas

Así como los gráficos anteriores nos permiten resumir información sobre el conjunto de datos estudiados, también podemos obtener información a partir de ciertas funciones de los datos.

Consideraremos en particular dos tipos de medidas: *medidas de localización*, las cuales nos proporcionan una idea de la magnitud de los datos y *medidas de dispersión*, las cuales, como su nombre indica, permiten conocer en forma aproximada qué tan dispersos se encuentran los datos alrededor de dicha localización.

2.2.1 Medidas de Localización

Sean $X_1, X_2, \ldots, X_n$ los datos con los cuales estamos trabajando. En base a estos datos, podemos obtener las siguiente medidas:

a **Media Muestral (Promedio):**
La media muestral (o promedio) se define como:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2.1}$$

La media actúa como el "centro de masa" de los datos: si imaginamos los datos como los lugares en los cuales se colocan pesos en una barra horizontal, el promedio será el punto del cual es necesario colgar la barra para que se mantenga en equilibrio.

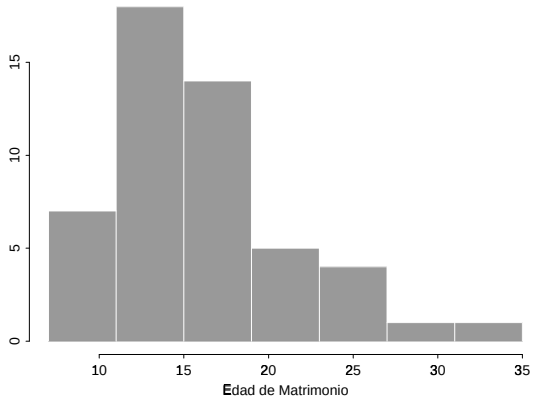

Figura 2.2: Histograma para los datos de edad de matrimonio en Guatemala.

Para los datos del ejemplo 2.1, $\bar{X} = 16.5$, es decir, la edad promedio de matrimonio entre el grupo estudiado es de 16 años y medio.

b **Mediana:**
Es el valor tal que el 50% de las observaciones está por encima y el otro 50% por debajo. Para obtener la mediana, ordenamos los datos en forma creciente. Si el número de datos es impar entonces la mediana es la observación central; si no, se toma como mediana el promedio de las dos observaciones centrales.

Ejemplo 2.2 *Consideremos el siguiente conjunto de observaciones:*

2 4 2 3 5 8 7

Al ordenarlas, obtenemos:

2 2 3 4 5 7 8

En este caso, la mediana es la observación central, es decir, 4.

Ejemplo 2.3 *Para el conjunto de datos*

2 2 3 4 5 7 8 10

La mediana se obtiene como el promedio de los dos datos centrales, 4 y 5. Es decir, la mediana para este conjunto de datos vale 4.5.

Para los datos del ejemplo 2.1, la mediana es 15.5 años.

Observemos que la media y la mediana para este caso son distintas, siendo la mediana menor que la media. Para discutir el significado de esta diferencia consideremos la siguiente situación

Ejemplo 2.4 *Para los datos del ejemplo 2.3 el promedio es 5.125 y la mediana es 4.5.*

Sustituyamos el último dato por 20. El nuevo conjunto de datos es

2 2 3 4 5 7 8 20

Para este nuevo conjunto de datos, la mediana permanece constante (4.5) pero la media cambia a 6.375.

Del ejemplo anterior se puede inferir que la media muestral es muy sensible a las observaciones alejadas (extremas), a diferencia de la mediana. Decimos entonces que la mediana es una medida más "robusta". Esta robustez de la mediana puede explicarse porque la mediana sólo emplea la información del *orden* de los datos, y no su tamaño.

En general, diferencias marcadas entre la media y la mediana indican la presencia de asimetría o de valores extremos en los datos. En el caso de los datos del ejemplo 2.1, la diferencia entre la media y la mediana viene dada por la presencia de dos valores extremos altos (29 y 34).

c **Media truncada:**
Cuando existe asimetría en los datos o se presentan valores extremos, una manera de reducir la sensibilidad de la media es eliminar un cierto porcentaje de datos en ambos extremos del conjunto de datos. A la medida así obtenida suele llamarsele *media truncada.* Como para su cálculo se eliminan los valores extremos, resulta menos afectada por la presencia de éstos que la media.

Para los datos del ejemplo 2.1, la media truncada obtenida eliminando el 5% de los datos es 16.152, mientras que si eliminamos el 10% obtenemos 15.975. Nótese que mientras mayor es el porcentaje de datos que se elimina, más se acerca la media truncada a la mediana.

d **Moda:**
Es la observación que más se repite. Toma sentido cuando se tiene una cantidad significativa de datos.

En el ejemplo 2.1 podemos identificar dos valores modales, que son 15 y 16 años. Cada una de estas edades aparece 6 veces en el conjunto de datos.

2.2.2 Medidas de Dispersión

a **Rango:**
El rango se calcula como la diferencia entre el máximo valor y el mínimo valor presentes en el conjunto de datos:

$$\text{Rango} = X_{max} - X_{min} \tag{2.2}$$

Se trata de una medida muy robusta, pues usa muy poca información del conjunto de datos; igualmente, proporciona poca información acerca del conjunto de datos.

En el ejemplo 2.1, el rango es 34-9=25.

b **Rango intercuartil:**
Los cuartiles q_1, q_2 y q_3 de un conjunto de datos son los números tales que dividen en cuatro partes iguales dicho conjunto de datos. Observemos entonces que el segundo cuartil, q_2, es la mediana.

Una manera de calcular los cuartiles (no es la única posible) es la siguiente: ordenemos los datos de menor a mayor, y llamemos $X_{(i)}$ al dato que está en la posición i de la muestra ordenada. Entonces,

$$q_1 = X_{([\frac{n+1}{4}])} + \left(\frac{n+1}{4} - \left[\frac{n+1}{4}\right]\right)\left(X_{([\frac{n+1}{4}]+1)} - X_{([\frac{n+1}{4}])}\right) \tag{2.3}$$

$$q_3 = X_{([\frac{3(n+1)}{4}])} + \left(\frac{3(n+1)}{4} - \left[\frac{3(n+1)}{4}\right]\right)\left(X_{([\frac{3(n+1)}{4}]+1)} - X_{([\frac{3(n+1)}{4}])}\right), \tag{2.4}$$

donde $[a]$ representa la parte entera de a.

Por ejemplo, si tenemos 20 datos, $\frac{20+1}{4} = 5.25$, de manera que

$$q_1 = X_{(5)} + (5.25 - 5)(X_{(6)} - X_{(5)}).$$

Como $\frac{3(20+1)}{4} = 15.75$, el tercel cuartil se calculará como

$$q_3 = X_{(15)} + (15.75 - 15)(X_{(16)} - X_{(15)})$$

.

El Rango Intercuartil es la longitud del intervalo donde está contenido el 50% central de los datos:

$$\text{Rango Intercuartil} = q_3 - q_1 \tag{2.5}$$

Para los datos del ejemplo 2.1, $n = 50$. Usando las ecuaciones (2.3) y (2.4), obtenemos $\frac{50+1}{4} = 12.75$, de modo que

$$q_1 = X_{(12)} + (12.75 - 12)(X_{(13)} - X_{(12)}) = 13 + .75(13 - 13) = 13$$

$$q_3 = X_{(38)} + (38.25 - 38)(X_{(39)} - X_{(38)}) = 19 + .75(19 - 19) = 19$$

El rango intercuartil es de $19 - 13 = 6$. Si comparamos este resultado con el rango (25) vemos que aunque el intervalo total de los datos abarca un rango de edades de 25 años, el 50% central se concentra en un período de 6 años, es decir, aproximadamente la cuarta parte del intervalo total.

c **Varianza muestral:**
La varianza muestral se define como:

$$\text{Varianza} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \tag{2.6}$$

Es decir, se trata del promedio de las desviaciones cuadráticas entre las observaciones y su media. Su equivalente físico es el momento de inercia: si nuevamente consideramos que los datos son pesos colocados en una barra, la varianza está relacionada con la resistencia que ofrecerá esta barra a girar cuando se la cuelga de su centro de masa (media): a mayor varianza, mayor será la resistencia.

Un posible inconveniente para la interpretación de la varianza es que, por el efecto del cuadrado en la definición, no está expresada en las mismas unidades que los datos, sino en su cuadrado (por ejemplo, si los datos se toman en metros, la varianza se expresará en metros cuadrados). Como una manera de eliminar este inconveniente, definimos la *desviación standard* del conjunto de datos como:

$$\text{Desviación Standard} = \sqrt{\text{Varianza}}$$

Esta medida tiene las mismas unidades que los datos en estudio.

En el caso del ejemplo 2.1, la varianza es de 26.5 años^2, lo cual implica una desviación standard de 5.15 años.

2.3 Diagramas de Caja.

El diagrama de caja (*Boxplot*) es una representación gráfica de los datos que permite analizar conjuntamente una serie de medidas numéricas, tales como el mínimo, el máximo, la mediana y los cuartiles. En este gráfico es posible observar características de los datos como simetría y posibles observaciones atípicas,

Los pasos a seguir para la construcción del diagrama de caja son los siguientes:

1. Ordenar los datos y obtener X_{min}, X_{max}, q_1, q_2, q_3.

2. Dibujar un rectángulo cuyos extremos sean q_1 y q_3, e indicar q_2 mediante una línea.

3. Calcular los "límites admisibles" superior e inferior:

$$LI = q_1 - f * (q_3 - q_1)$$

$$LS = q_3 + f * (q_3 - q_1)$$

   Se consideran posibles valores atípicos los situados fuera del intervalo (LI, LS).

   El factor f puede variar entre diferentes textos o paquetes estadísticos. Algunos de los valores más usados de f son $f = 0.75$ y $f = 1.5$.

4. Dibujar una línea que vaya desde cada extremo del rectángulo al valor más alejado no atípico.

5. Indicar todos los datos que están fuera del intervalo admisible marcándolos como atípicos.

La figura 2.3 muestra el diagrama de caja correspondiente a los datos del ejemplo 2.1, usando $f = 1.5$. En este gráfico podemos observar que los datos parecen distribuirse en forma simétrica alrededor de su mediana, pero se presentan dos observaciones mayores que el resto, las cuales podrían considerarse atípicas.

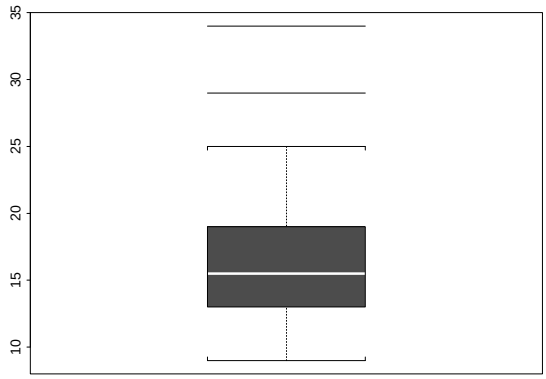

Figura 2.3: Diagrama de caja para los datos de edad de matrimonio en Guatemala.