1. Estadística

Definición: La estadística es un ciencia inductiva que permite inferir características cualitativas y cuantitativas de un conjunto mediante los datos contenidos en un subconjunto del mismo.

Definición: La población objetivo es el conjunto total de individuos u objetos con alguna característica que es de interés estudiar.

Definición: La muestra es un subconjunto de la población y contiene elementos en los cuales debe estudiarse la característica de interés para la población.

Definición: La observación o dato es cada uno de los valores obtenidos para los elementos incluidos en la muestra.

Al tener la muestra, usualmente se utilizan dos tipos de estadísticas para estudiar dichos datos. Primero se realiza una estadística descriptiva, con la cual se recopilan, organizan, procesan y presentan los datos obtenidos en la muestra, lo que permite tener una visión más amplia y certera acerca del problema en estudio. Luego se procede a realizar estadística inferencial, para obtener resultados basados en la información contenida en la muestra.

2. Estadística descriptiva

El primer paso para realizar un estudio estadístico es tener un conjunto de datos, estos pueden provenir de fuentes establecidas o a partir de una recolección realizada por el investigador. Los datos que se tienen pueden ser de dos tipos:

a Cualitativos: corresponden a respuestas categóricas (género de la persona).
b Cuantitativos: corresponden a respuestas numéricas (estatura de la persona). Además estos pueden dividirse en:
   1. Discretos (edad de una persona).
   2. Continuos (tiempo de realización de un proceso).

Cuando se obtiene la muestra hay varias técnicas de importancia para realizar la estadística descriptiva. Un primer paso es encontrar los estadísticos de orden. Si recordamos, $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$ representan los n datos de una muestra en forma ordenada, desde el valor más pequeño, hasta el más grande.

Ejemplo. Sea la muestra de tamaño 10, $(7, 5, 3, 11, 3, 1, 8, 5, 9, 6)$. Halle los estadísticos de orden.

Solución.
Es claro que $X_{(1)} = 1$, $X_{(2)} = 3$, $X_{(3)} = 3$, $X_{(4)} = 5$, $X_{(5)} = 5$, $X_{(6)} = 6$, $X_{(7)} = 7$, $X_{(8)} = 8$, $X_{(9)} = 9$, $X_{(10)} = 11$. Es importante notar que de existir dos observaciones con la misma medición, entonces habrán dos estadísticos de orden iguales.

Los datos que se obtienen pueden representarse de varias formas distintas, en tablas, en graficas o mediante números que caracterizan al grupo de datos.

2.1. Tablas de frecuencias. Las tablas de frecuencias es una técnica que se utiliza para agrupar los datos de forma ordenada. Supongamos que X es una muestra de tamaño n, entonces para construir la tabla de frecuencia seguiremos los siguientes pasos:

1. Calcular el rango de los datos, que no es más que la distancia entre el mayor y el menor de los valores de los datos. Es decir
$$R = X_{(n)} - X_{(1)}.$$

2. Seleccionar el número de clases k, para agrupar los datos. Como sugerencia para elegir el k

Otros metodos para calcular el número de clases incluye:

n	k
Menos de 50	4 a 7
Entre 50 y 100	6 a 10
Entre 100 y 200	7 a 12
Mas de 250	10 a 20

i $k = \sqrt{n}$.
ii $k \geq \frac{\log n}{\log 2}$.
iii $k = 1 + 3{,}322 \log n$.
iv $k = 1 + \log_2(n)$.
v $k = (2n)^{1/3}$.

3. Obtener la longitud de las clases como $L = R/k$. Usualmente se puede redefinir la longitud, el número de clases y los extremos de cada clase para que todas tengan la misma longitud y los intervalos de cada clase incluyan a todos los datos, sean excluyentes y los valores en los extremos de cada clase sean simples.
4. Realizar el conteo de datos para obtener la frecuencia en cada clase.
5. Usualmente se sigue la siguiente notación

n número de datos.
k número de clases.
m_i marca de la clase i, que es el punto medio del intervalo correspondiente.
f_i frecuencia de la clase i.
f_i/n frecuencia relativa de la clase i.
F_i frecuencia acumulada de la clase i.
F_i/n frecuencia acumulada relativa de la clase i.

EJEMPLO. Se realiza la medición de los tiempos que tardan N personas en utilizar el mismo servicio bancario, y se obtienen los siguientes datos, para una muestra de 28 de ellas:

9.16	11.95	11.47	11.35
13.29	15.15	9.54	8.73
12.07	14.75	11.26	10.00
11.97	14.79	13.66	9.75
13.31	15.48	11.18	11.71
12.32	13.47	15.03	12.45
11.78	13.06	14.86	12.38

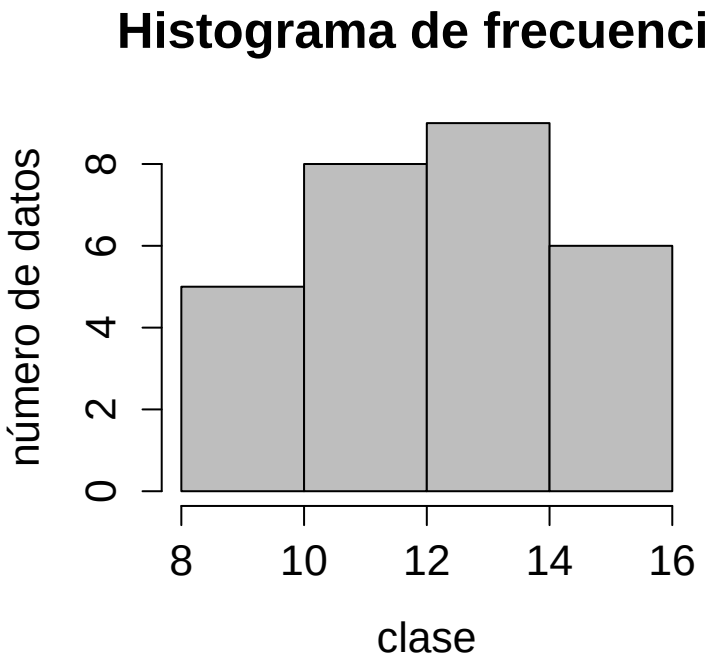
Obtenga la tabla de frecuencia para estos datos.

SOLUCIÓN.
El rango es $R = 15{,}48 - 8{,}73 = 6{,}75$. En este caso utilizaremos 4 clases. La longitud $L = 6{,}75/4 = 1{,}6875$, pero por facilidad utilizaremos intervalos de longitud 2, con extremos enteros. Así la tabla será

i	Clase k	m_i	f_i	f_i/n	F_i	F_i/n
1	[8, 10]	9	5	0.1786	5	0.1786
2	(10,12]	11	8	0.2857	13	0.4643
3	(12,14]	13	9	0.3214	22	0.7857
4	(14,16]	15	6	0.2143	28	1.0000

2.2. Graficas. La forma más común de representar graficamente un conjunto de datos es por medio de un histograma de frecuencia. Este regularmente nos muestra la frecuencia absoluta o relativa de un conjunto de datos.

EJEMPLO. Para los datos anteriores, realizar el histograma de frecuencias.



2.3. Medidas de Tendencia Central. DEFINICIÓN: La media muestral no es más que el promedio aritmético de los datos de una muestra. Si x_1, x_2, ..., x_n son los datos en la muestra, entonces

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

EJEMPLO. Calcular la media muestral de los datos anterior.

SOLUCIÓN.
$\bar{X} = 12{,}35$.

DEFINICIÓN: La moda muestral, Mo, es el valor que se repite mayor cantidad de veces.
Por ejemplo, en el caso de los datos que hemos venido utilizando, no existe la moda, ya que ningún valor se repite.

DEFINICIÓN: La mediana es el valor ubicado en el centro de los datos ordenados. Para una muestra de tamaño n

$$\widetilde{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{si } n \text{ es impar} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right), & \text{si } n \text{ es par.} \end{cases}$$

EJEMPLO. Calcular la mediana de los datos usados.

SOLUCIÓN.
$\widetilde{X} = \frac{1}{2}\left(X_{(14)} + X_{(15)}\right) = \frac{1}{2}(12{,}07 + 12{,}32) = 12{,}195.$

2.4. Medidas de dispersión. El rango, que ya fue definido anteriormente es una de las medidas de dispersión que suelen usarse.

DEFINICIÓN: La varianza muestral se define como

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$$

Más adelante se explicará con más detalle porque dividir por $n-1$ en vez de n.

EJEMPLO. Calcular la varianza de los datos que se tienen anteriormente.

SOLUCIÓN.
$S^2 = 3{,}5938$.

DEFINICIÓN: La desviación estándar muestral no es mas que la raíz cuadrada de la varianza muestral.

EJEMPLO. Calcular la desviación estándar de los datos anteriores.

SOLUCIÓN.
$S = 1{,}8957$.

2.5. Medidas de posición. DEFINICIÓN: Los cuartiles son números que dividen a los datos de la muestra en grupos de aproximadamente 25 %. El primer cuartil (Q_1) es el que tiene a su izquierda el 25 % de los datos. El segundo cuartil (Q_2) es igual a la mediana, y divide al grupo de datos en dos partes con un 50 % de los datos. El tercer cuartil (Q_3) es el que tiene a su izquierda el 75 % de los datos.

EJEMPLO. Calcular los cuartiles de los datos anteriores.

SOLUCIÓN.
Como hay 28 datos, entonces el 25 % de los datos están alrededor del séptimo dato, el 50 % alrededor del decimo cuarto dato y el 75 % alrededor del vigesimo primer dato. Luego

$$Q_1 = \left(X_{(7)} + X_{(8)}\right)/2 = (11{,}26 + 11{,}35)/2 = 11{,}305.$$

$$Q_3 = \left(X_{(21)} + X_{(22)}\right)/2 = (13{,}47 + 13{,}66)/2 = 13{,}565.$$

Como calculamos antes $Q_2 = \widetilde{X} = 12{,}195$.

DEFINICIÓN: Los percentiles son números que dividen los datos de la muestra en grupos de tamaño aproximado de 1 %. Por ejemplo, el percentil 1, P_1 es el que incluye a su izquierda aproximadamente el 1 % de los datos. Así sucesivamente hasta P_99.

DEFINICIÓN: El coeficiente de variación es una medida adimensional que se usa para comparar la variabilidad de los datos de diferentes grupos, y se calcula como $V = \dfrac{S}{\bar{X}}$.

2.6. Datos agrupados. Ahora supongamos que no se tienen los datos individuales, sino que se tiene la tabla de frecuencia entonces podemos calcular la media y la varianza de los datos.

DEFINICIÓN: La media de los datos agrupados se puede calcular como

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} m_i f_i.$$

La mediana de los datos agrupados se calcula como

$$\widetilde{X} = l_i + L\frac{\frac{n}{2} - F_a}{f_{\text{med}}}$$

donde el primer paso es calcular $n/2$ y ubicar el resultado en F_i. Si el resultado no esta, se toma el siguiente mas grande. Se toma l_i como la clase donde esta la mediana, que es donde se encuentra el valor escogido de F_i. F_a es la frecuencia acumulada de la clase anterior, f_{med} es la frecuencia absoluta de la clase en la que esta la mediana y L es la longitud de los intervalos.

En general si se quiere calcular cualquier cuartil

$$Q_k = l_i + L\frac{\frac{kn}{4} - f_{i-1}}{f_i}$$

donde $k = 1, 2, 3$.

La varianza de los datos agrupados se calcula

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(m_i - \bar{X})^2.$$

EJEMPLO. Para la tabla de frecuencias ya calculada

i	Clase k	m_i	f_i	f_i/n	F_i	F_i/n
1	[8, 10]	9	5	0.1786	5	0.1786
2	(10,12]	11	8	0.2857	13	0.4643
3	(12,14]	13	9	0.3214	22	0.7857
4	(14,16]	15	6	0.2143	28	1.0000

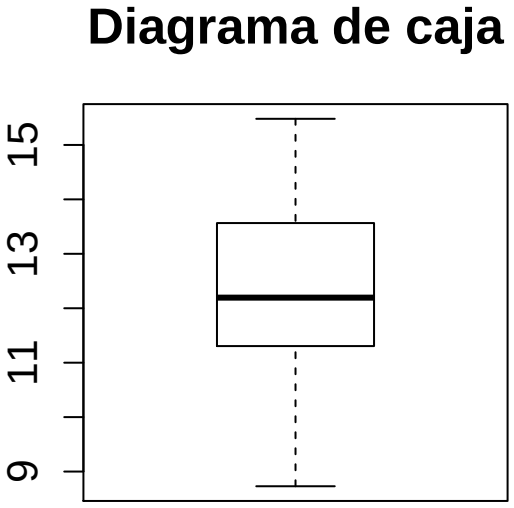
Calcula la media y la varianza.

SOLUCIÓN.

$$\bar{X} = \frac{1}{28}[(9)(5) + (11)(8) + (13)(9) + (15)(6)] = 12{,}14.$$

$$\widetilde{X} = 12 + 2\frac{14 - 13}{9} = 12{,}22.$$

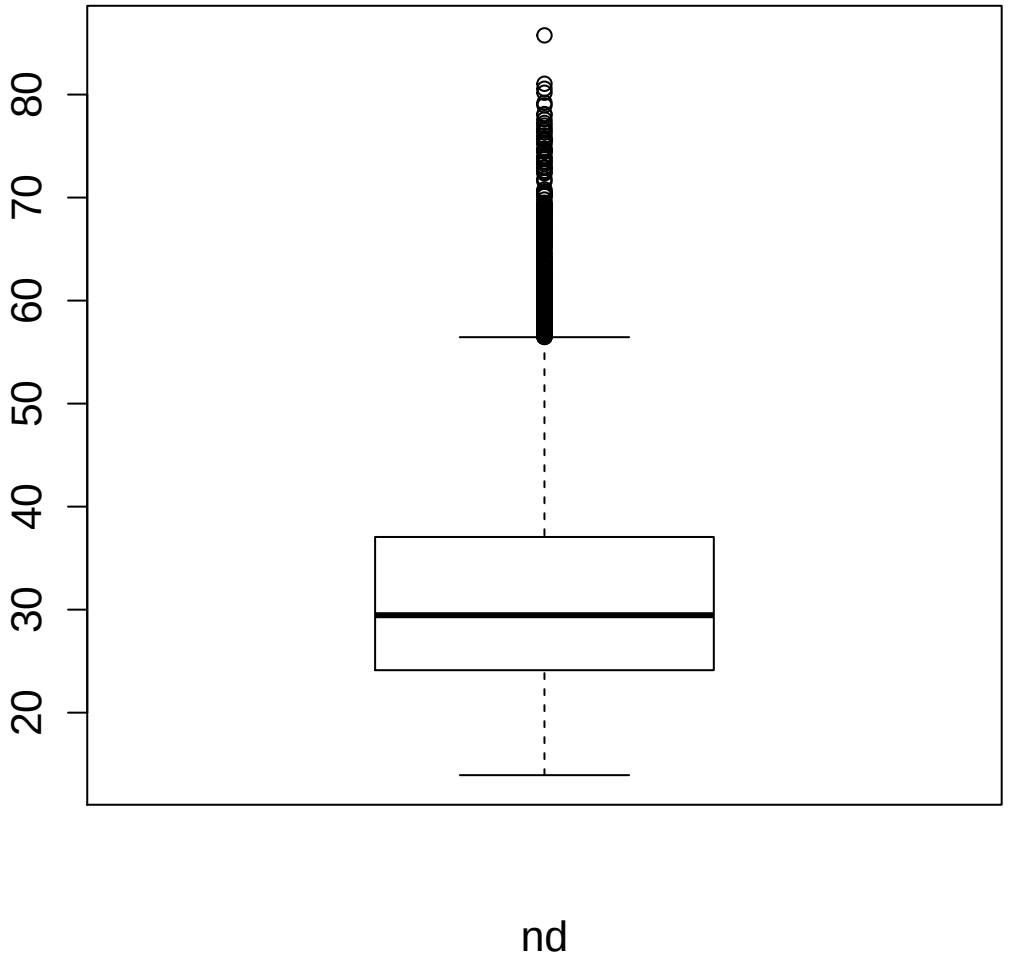
$$S^2 = \frac{1}{27}[5(9 - 12{,}14)^2 + 8(11 - 12{,}14)^2 + 9(13 - 12{,}14)^2 + 6(15 - 12{,}14)^2] = 4{,}2751.$$



2.7. Diagrama de caja. El diagrama de caja es una forma grafica de resumir las medidas de posición de los datos.

Como se puede ver en la figura la linea negra dentro de la caja representa la mediana de los datos, los límites de la caja son Q_1 y Q_3, mientras que las lineas punteadas son llamadas bigotes, y delimitan los límites normales dentro de los cuales deben estar los datos de ese conjunto. La distancia de la caja al final del bigote suele calcularse como 1.5 por el rango intercuartil, RQ, que no es más que $RQ = Q_3 - Q_1$. Los datos que quedan por encima o debajo de los bigotes son llamados datos atípicos.

En el segundo caso si se tiene un conjunto de datos con gran cantidad de datos atípicos.


20
30
40
50
60
70
80
nd