Informática y comunicaciones

ies azarquiel

SISTEMAS BIG DATA

ESTADÍSTICA DESCRIPTIVA

Sebastián Rubio Valero

Sptiembre 2025

1 ¿QUÉ ES LA ESTADÍSTICA DESCRIPTIVA?

La **estadística descriptiva** está orientada a la presentación de datos mediante tablas y gráficas que permiten resumir o describir el comportamiento de los mismos, sin realizar inferencias sobre ellos debido a que son obtenidos de una parte de la población

La **estadística inferencial**, en cambio, se deriva de la muestra, de mediciones u observaciones que se han hecho a una parte representativa de la población, por lo cual busca establecer generalidades para la población a partir de los datos obtenidos. En consecuencia, la estadística inferencial investiga o analiza una población a partir de la muestra seleccionada, utilizando técnicas adecuadas del muestreo.

En este curso sólo vamos a "estudiar" estadística descriptiva, ya que nos hará falta conocer algunas medidas de mayor uso en los cuadros de mandos y en el análisis de datos.Un científico de datos también realiza análisis por medio de la estadística inferencial.

2 CONCEPTOS GENERALES DE LA ESTADÍSTICA

2.1 Población

La población o universo es un conjunto de elementos a los cuales se le estudian algunas características comunes; por ejemplo, los docentes de una institución educativa, las empresas de un sector productivo, los barrios de una ciudad, los artículos vendidos en un supermercado, las calificaciones de una prueba de aptitud, entre otros.

La población puede ser finita o infinita. Se estima que una población es finita cuando el número de los elementos que la integran es conocido por el investigador; tal es el caso de los barrios de una ciudad, los docentes de una universidad, los operarios de una industria, etc., mientras que para la población infinita no se conoce el número de elementos, ya sea porque es muy grande o porque se sabe que existe pero no se conoce el tamaño, por ejemplo: los lanzamientos de un dado, el número de veces que una persona puede pasar por un sitio y demás.

2.2 Muestra

La muestra se define como un conjunto de elementos seleccionados adecuadamente, que pertenecen a una población determinada, o sea que es una parte de la población o universo.La muestra se define como un conjunto de elementos seleccionados ade-

cuadamente, que pertenecen a una población determinada, o sea que es una parte de la población o universo.

En la mayoría de los estudios se procura que el número de elementos de la muestra sea cercano al número de la población para evitar errores generados por el muestreo; sin embargo, con el fin de optimizar recursos de tiempo, dinero, etc., se asumen los errores generados por la diferencia en el número de unidades entre la muestra y la población y se acude al muestreo. En los casos en los cuales el número de elementos de la muestra es igual al de la población, el estudio se denomina censo.

2.3 Parámetros

Los parámetros son medidas cuantitativas que describen una característica de la población, entre ellas están: media aritmética, varianza y coeficiente de variación.

2.4 Variables

Una variable es una propiedad o característica a la que se le puede asignar un valor. Existe dos tipos de variables:

- **Variables cualitativas**. Las variables cualitativas son aquellas que representan atributos delos elementos y no permiten una representación numérica definida. Entre las variables cualitativas están: el estrato socioeconómico, el estado civil, la profesión, el color de una flor, entre otras.
- **Variables cuantitativas**. Estas variables permiten una escala numérica y las características de los elementos son observados cuantitativamente a través de una medida y una escala definidas. Entre las variables cuantitativas se encuentran: el salario de los empleados, la talla de una persona, el peso, el número de hijos en una familia, el número de artículos vendidos en un almacén, entre otros En las variable cuantitativas existen dos grupos, las variables **cuantitativas continuas** y las variables **cuantitativas discretas**. La cuantitivas contínuas son aquellas en las que un valor y el siguiente no existe saltos, por ejemplo la altura, el peso, un caso de una variable discreta sería el número de artículos vendidos.

Ejercicio

Dada la tabla de la figura 1, indicar de que tipo son las variables de la columna izquierda

Variable	Codificación	Tipo de variable
Edad	En años cumplidos	
Tiempo de antigüedad	En años cumplidos	
Estrato socioeconómico	Bajo, medio, alto	
Estado civil	Soltero, casado, viudo, separado, otro	
Escolaridad	Ninguna, primaria, secundaria, universitario, posgrado, otra	
Tipo de lesión presentada	Múltiples opciones	
Área del cuerpo afectada	Extremidades, pecho, cara, otra	
Grado de la lesión	Leve, moderada, grave	
Requirió atención médica	Sí, no	
Tiempo de incapacidad	En días	
Ha presentado previamente accidentes laborales	Sí, no	
Utiliza implementos de bioseguridad en su trabajo	Sí, no	
Fecha del accidente	Ubicación en el calendario (dd/mm/aaaa)	
En qué jornada ocurrió el accidente	Mañana, tarde, noche	
Temperatura del sitio del accidente	Temperatura en °C	
Fuma	Sí, no	
Número de cigarrillos consumidos diariamente	N° de cigarrillos	
Consumo de alcohol	Sí, no	
Frecuencia de consumo de alcohol	Diario, semanal, quincenal, mensual	
Personas con quienes consume licor	Amigos, familiares, pareja, compañeros de trabajo, otros	
Cociente intelectual (CI)	Medido en escala de CI	
Capacidad para el estudio	Puntuación en un test	
Barrio de residencia	Múltiples opciones	

Figure 1: Ejercicio 1

3 TABULACIÓN DE DATOS ESTADÍSTICOS

3.1 Rango

El rango o recorrido se defina como la variación numéria de la variable, es el recorrido que toma la variable desde el valor más pequeño hasta el valor más alto $R = [L_s - L_i]$, donde L_i : es el límite inferio y L_s es el límite superior.

En la siguiente tabla se presenta el tiempo en minutos requerido por un grupo de personas para realizar una actividad

Persona	Tiempo (min)	Persona	Tiempo (min)	Persona	Tiempo (min)	Persona	Tiempo (min)	Persona	Tiempo (min)
1	70	11	47	21	57	31	52	41	51
2	71	12	68	22	55	32	63	42	50
3	62	13	60	23	55	33	65	43	60
4	63	14	54	24	57	34	50	44	56
5	67	15	63	25	59	35	53	45	67
6	65	16	60	26	74	36	59	46	59
7	75	17	69	27	56	37	45	47	68
8	62	18	54	28	59	38	72	48	61
9	65	19	73	29	71	39	64	49	51
10	56	20	55	30	50	40	69	50	64

Figure 2: Enter Caption

En esta tabla observamos que el valor más alto L_s=75, el más bajo L_i=45

3.2 Número de intervalos o clases

Para calcular el número de intervalos en los que se puede dividir la muestra, se suele utilizar la fórmula propuesta por Sturges

$$m = 1 + 3,3 * log(n)$$

Donde n es el número total de datos, para el ejemplo de tabla anterior

$$m = 1 + 3,3 * log(50)$$

Por lo tanto $m = 6,60637$, en este caso es posible construir 6 0 7 intervalos, pero debemos de calcular la amplitud del intervalo (el logaritmo es en base 10)

3.3 Amplitud del intervalo de clase

La amplitud de los intervalos C no es necesaria que sea igual para todos, pero para nuestro estudios vamos considerarlos todos iguales. La forma de calcularlo es la siguiente

$$C = AR/m$$

Donde AR es el rango de la amplitud es el número de intervalos (calculado en el apartado anterior). Siguiendo con el ejemplo $AR = 75 - 45 = 30$, el número de intervalos calculados esta entre 6 o 7, vamos estudiarlos en los dos casos.

Si $m = 6$, la amplitud del intervalo es $C = 30/6$, esto es $C = 5,0$ minutos

Si $m = 7$, la amplitud del intervalo es $C = 30/7$, por lo tanto $C = 4,285714...$ minutos

Con los resultados obtenidos, se recomienda usar 6 intervalos ($m = 6$), con una amplitud de 5 minutos ($C = 5$).

Si la amplitud no es exacta por ninguno de los valores arrojado por fórmula de Sturges, el número de intervalos se puede incrementar hasta hacer la división exacta.

3.4 Límites de los intervalo

Para construir los intervalos, cada uno de ellos queda determinado por dos extremos: límite inferior (l_i) y el límite superior (l_s). En la siguiente tabla podemos ver como se obtienen los intervalos

Nº. de Intervalo	Límites de Clase l_i - l_s	Intervalos o Clase l_i - l_s
1	45 - 45 + 5 = 50	[45 - 50]
2	50 - 50 + 5 = 55	(50 - 55]
3	55 - 55 + 5 = 60	(55 - 60]
4	60 - 60 + 5 = 65	(60 - 65]
5	65 - 65 + 5 = 70	(65 - 70]
6	70 - 70 + 5 = 75	(70 - 75]

Figure 3: Enter Caption

3.5 Tabulación

Una vez que hemos construido los intervalos, se procede con el conteo o frecuencia de la información, como podemos ver en la siguiente tabla

Nº. de Intervalo	Intervalo (minutos)	Tabulación	Frecuencia (Nº. de personas)
1	[45 - 50]	II	2
2	(50 - 55]	IIII IIII	9
3	(55 - 60]	IIII IIII II	12
4	(60 - 65]	IIII IIII I	11
5	(65 - 70]	IIII IIII	9
6	(70 - 75]	IIII II	7

Figure 4: Enter Caption

3.6 Marca de clase

La marca de clase, conocida también como punto medio, es el valor representativo para cada intervalo. Se representa por x_i y se calcula promediando los límites del intervalo.

$$x_i = (l_s + l_i)/2$$

Es importante resaltar que el incremento entre marcas de clase es igual a la cantidad

Nº. de Intervalo	Intervalo (minutos)	Marca de Clase (x_i)
1	[45 - 50]	47,5
2	(50 - 55]	52,5
3	(55 - 60]	57,5
4	(60 - 65]	62,5
5	(65 - 70]	67,5
6	(70 - 75]	72,5

Figure 5: Enter Caption

del intervalo C

Ejercicio

En la siguiente tabla tenemos los datos del las calificaciones obtenidas de 100 aspirantes al concurso de oratoria

- Obten los límites superior e inferior, rango, número de intervalos, amplitud del intervalo y marca de clase
- Construye la tabulación y las frecuencias para cada intervalo
- ¿Entre qué puntuaciones está la mayor cantidad de aspirantes?
- ¿Qué porcentaje de aspirantes obtuvo los puntos más altos?
- ¿Qué porcentaje de aspirantes obtuvo los puntos más bajos?
- ¿Cuántos aspirantes obtuvieron los puntos más bajos?
- ¿Cuántos aspirantes obtuvieron los puntos más altos?

38	51	32	65	25	28	34	12	29	43
71	62	50	37	8	24	19	47	81	53
16	62	50	37	4	17	75	94	6	25
55	38	46	16	72	64	61	33	59	21
13	92	37	43	58	52	88	27	74	66
63	28	36	19	56	84	38	6	42	50
94	51	62	3	17	43	47	54	58	26
12	42	34	68	77	45	60	31	72	23
18	22	70	34	5	59	20	68	55	49
33	52	14	40	38	54	50	11	41	76

Figure 6: Enter Caption

4 DISTRIBUCIÓN DE FRECUENCIAS

4.1 Frecuencia absoluta

La frecuencia absoluta (n_i) es la cantidad de veces que se repite el valor x_i de la variable X en la muestra o la población. Cuando sumamos todas la frecuencias absolutas tenemos que obtener el total de la muestra.

Nº. de Intervalo	Intervalo (minutos)	Frecuencia absoluta (n_i)
1	[45 - 50]	2
2	(50 - 55]	9
3	(55 - 60]	12
4	(60 - 65]	11
5	(65 - 70]	9
6	(70 - 75]	7
Total		50

Figure 7: Enter Caption

4.2 Frecuencia relativa

La frecuencia relativa (h_i) se define como el porcentaje de frecuencia absoluta en relación con el total de datos de la muestra (n). Para obtener la frecuencia relativa aplicamos la siguiente fórmula

$$h_i = (n_i/n) * 100$$

Siendo n el total de datos de la muestra. En la siguiente tabla podemos ver las frecuencias relativas.

Nº. de Intervalo	Intervalo (minutos)	Frecuencia relativa (h_i)
1	[45 - 50]	(2/50)*100 = 4%
2	(50 - 55]	(9/50)*100 = 18%
3	(55 - 60]	(12/50)*100 = 24%
4	(60 - 65]	(11/50)*100 = 22%
5	(65 - 70]	(9/50)*100 = 18%
6	(70 - 75]	(7/50)*100 = 14%
Total		100%

Figure 8: Enter Caption

4.3 Frecuencia absoluta acumulada

La frecuencia absoluta acumulada (N_i) para un valor x_i de una variable X es la adición de las frecuencias absolutas n_i

$$N_i = \sum_{k=1}^{i} n_k \tag{1}$$

Nº. de Intervalo	Intervalo (minutos)	Frecuencia relativa (h_i)	Frecuencia absoluta acumulada (H_i)
1	[45 - 50]	4%	4%
2	(50 - 55]	18%	4% + 18% = 22%
3	(55 - 60]	24%	4% + 18% + 24% = 46%
4	(60 - 65]	22%	4% + 18% + 24% + 22% = 68%
5	(65 - 70]	18%	4% + 18% + 24% + 22% + 18% = 86%
6	(70 - 75]	14%	4% + 18% + 24% + 22% + 18% + 14% = 100%

Figure 9: Enter Caption

La siguiente tabla tenemos una relación de todos los cálculos realizados.

Nº. de Intervalo	Intervalo (Tiempo en minutos)	n_i	h_i	N_i	H_i	$\dot{x}_i$
1	[45 - 50]	2	4%	2	4%	47,5
2	(50 - 55]	9	18%	11	22%	52,5
3	(55 - 60]	12	24%	23	46%	57,5
4	(60 - 65]	11	22%	34	68%	62,5
5	(65 - 70]	9	18%	43	86%	67,5
6	(70 - 75]	7	14%	50	100%	72,5

Figure 10: Enter Caption

Ejercicio

La siguiente tabla representa las horas estudiadas (clase + horas de estudio) por semana (6 días) de un conjunto de 50 estudiantes del doble grado de matemáticas y físicas de segundo curso. Se pide realizar un tabla con los intervalos, las frecuencias y las marcas de clase. Discutir los resultados obtenidos con vuestro compañero de trabajo, sacar conclusiones sobre el grado y los estudiantes resultados.

Table 1: Muestra de 50 datos para ejercicios de estadística descriptiva

38	42	29	35	48	33	45	41	36	52
27	44	39	31	50	34	46	40	37	49
32	43	28	47	30	51	25	53	26	54
55	24	56	23	57	22	58	21	59	20
60	19	61	18	62	17	63	16	64	15

5 MEDIDAS DE TENDENCIA CENTRAL

En la mayoría de los casos, el conjunto de datos obtenidos, ya sea de una muestra o de una población, tienden a reunirse alrededor de un valor central. De esta manera, es posible obtener un valor típico o representativo de todo el conjunto de datos, el cual se llama medida de tendencia central. Las medidas de tendencia central más representativas son: media aritmética, mediana y moda.

5.1 Media aritmética

Representa el promedio del conjunto de datos de la muestra, su cálculo se realiza aplicando la siguiente fórmula

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

$\bar{X}$: Es la media aritmética de la muestra
n: Total de datos de la muestra
x_i: dato de la variable

Por ejemplo las horas dormidas por 5 estudiantes de 4^{o} de la ESO son: 6,8,5,8,7.

La media aritmética es:

$$\bar{X} = \frac{6+8+5+8+7}{5} = 6,8 \tag{3}$$

Cuando se agrupan los datos en una tabla de frecuencia, sin construir intervalos, la media se calcula mediante la siguiente formula:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i * n_i}{n} \tag{4}$$

Número de hijos x_i	Frecuencia n_i	$x_i * n_i$
0	1	0 x 1 = 0
1	2	1 x 2 = 2
2	4	2 x 4 = 8
3	2	3 x 2 = 6
4	1	4 x 1 = 4
$n = \sum n_i = 10$	$\sum x_i * n_i = 20$	

Figure 11: Enter Caption

n_i es la frecuencia absoluta para cada valor x_i

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i * n_i}{n} = 20/10 = 2 \tag{5}$$

Esto significa que el promedio de hijos es de 2 hijos por empleado.

Retomando la tabla de tiempos, tenemos los siguientes resultados

Nº. de Intervalo	Minutos	$\dot{x}$	n_i	$x_i * n_i$
1	[45 - 50]	47,5	2	95
2	(50 - 55]	52,5	9	472,5
3	(55 - 60]	57,5	12	690
4	(60 - 65]	62,5	11	687,5
5	(65 - 70]	67,5	9	607,5
6	(70 - 75]	72,5	7	507,5
			$n = \sum n_i = 50$	$\sum \dot{x}_i * n_i = 3060$

Figure 12: Enter Caption

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i * n_i}{n} = 3060/50 = 61,2 minutos \tag{6}$$

5.2 Mediana

La mediana en un conjunto de datos es el valor que ocupa el lugar central, de tal forma que se deja en cada extremo el 50% . Para la ubicación de la posición de la mediana se deben ordenar los datos de forma ascendente.

Si el conjunto de datos no se han agrupado, la posición de la mediana se ubica según los siguientes criterios:

- Cuando el total de datos (n) es impar, la posición de la mediana estará determinada por la fórmula:
$$i = X_{\frac{n+1}{2}} \tag{7}$$
- Mientras que si el total de datos (n) es par, la posición de la mediana estaría determinada por:
$$i = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} \tag{8}$$

En la siguiente,X es el número de errores por página cometidos por un grupo de operarios Inicialmente se deben ordenar los datos en forma ascendente, esto es: $3, 4, 5, 6, 8,$

Digitador	A	B	C	D	E
Nº de errores	3	6	4	5	8

Figure 13: Enter Caption

el total de datos $n = 5$ y la posición para el estimador será

$$i = X_{\frac{n+1}{2}} = X_{\frac{5+1}{2}} = X_3 \tag{9}$$

Según el orden que hemos establecido la mediana es 5

En el siguiente ejemplo la muestra de los errores cometidos por los operarios es $5, 5, 7, 9, 11, 13, 13, 15$

$$i = \frac{X_{\frac{8}{2}} + X_{\frac{8}{2}+1}}{2} = \frac{X_4 + X_5}{2} \tag{10}$$

Lo que se corresponde $i_4 = 9$ y $i_5 = 11$, luego el valor de la mediana $Me = (9+11)/2 = 10$, como puedes observar el valor de la mediana no se corresponde con ningún valor de la muestra.

El 50% de los operarios cometen menos de 10 errores, y el otro 50% cometen 10 o más errores.

En caso de que el conjunto de datos esté agrupado en intervalos, el cálculo de la mediana se realiza mediante el siguiente procedimiento:

- Hallar $N/2$
- Ubicar el intervalo cuya frecuencia absoluta acumulada N, contiene a $N/2$
- Calcular la mediana mediante la fórmula

$$Me = l_i + \left(\frac{\frac{N}{2} - N_{i-1}}{n_i} \right) * c \tag{11}$$

Donde:
l_i: limite inferior del intervalo que contiene a $N/2$
N: número total de datos de la población N_{i-1}: frecuencia absoluta acumulada anterior al intervalo que contiene a $N/2$
n_i: frecuencia absoluta del intervalo que contiene a $N/2$
c: amplitud del intervalo que contiene a $N/2$
Siguiendo con nuestro ejemplo, en la tabla siguiente tenemos los cálculos necesarios para obtener la mediana Los paso seguidos son los siguientes:

Nº. de Intervalo	Minutos	n_i	f_i	N_i	F_i	$\dot{x}$
1	[45 - 50]	2	4%	2	4%	47,5
2	(50 - 55]	9	18%	11	22%	52,5
3	(55 - 60]	12	24%	23	46%	57,5
4	(60 - 65]	11	22%	34	68%	62,5
5	(65 - 70]	9	18%	43	86%	67,5
6	(70 - 75]	7	14%	50	100%	72,5

Figure 14: Enter Caption

- El total de personas que realizaron la actividad es 50, entonces $N/2 = 50/2 = 25$

- Al analizar la frecuencia absoluta acumulada, se encuentra que 25 se ubica en el 4º intervalo (no es posible ubicar el valor 25 en el tercer intervalo, debido a que sólo acumula 23 personas)

- Los datos para el cálculo de la median serán: $l_i = 60$
  $n/2 = 25$
  $N_{i-1} = 23$
  $n_i = 11$
  $c = 65 - 60 = 5$

$$Me = l_i + \left(\frac{\frac{N}{2} - N_{i-1}}{n_i}\right) * c = 60 + (\frac{25-23}{11}) * 5 = 60 + 0.9 = 60.9 \text{ minutos}$$

Esto significa que el 50% de las personas realizaron la actividad en 60,9 minutos o menos y el otro 50% tardaron más de 60,9 minutos.Cabe preguntarse cuando utilizar la media y cuando la mediana, la respuesta se encuentra en los valores extremos. Si en una muestra algunos valores son muy extremos entonces interesa utilizar la mediana. Con el siguiente ejemplo va a quedar más claro. Sea X la edad de un grupo de personas pertenecientes a un club de actividades lúdicas, estas son:17,16,17,18,17,16,17,18,35 La edad de 35 hace subir la media, a 19 años, pero al calcular la mediana el resultado es de 17 años, que se ajusta más a la mayoría de los datos.

5.3 Moda

Se denomina **moda** de un conjunto de datos al valor que más se presenta, es decir el valor o atributo de mayor frecuencia, se puede aplicar tanto a las variables cuantitativas como a las cualitativas tanto discretas como continuas.
Para obtener la moda de un conjunto de datos sin agrupar se construyen las frecuencias y se ubica el valor o la característica que corresponde a la frecuencia mayor. En la siguiente tabla tenemos del color favorito de un grupo de 10 personas. En la tabla

Color	n_i
Blanco	2
Azul	4
Rosado	1
Negro	2
Morado	1

Table 2: Preferencia de colores

anterior el color de mayor frecuencia es el azul por lo tanto este es la moda.
En el ejemplo anterior sólo hay una moda por lo que estas distribuciones se les llaman *unimodal*. Cuando existen varias modas se les llama *multimodal*, y en el caso de que no tenga ninguna moda las distribución es *amodal*. Para poner un ejemplo de bimodal, supongamos que un grupo de estudiantes y un semestre se matriculan en el siguiente número de asignaturas, $6, 5, 6, 4, 6, 7, 7, 9, 4, 7$, como se puede comprobar que 6 asignaturas y 7 asignaturas tienen la misma frecuencia.

Cuando los datos han sido agrupados en clases o intervalos, la moda se calcula utilizando la ponderación en el intervalo, con el siguiente procedimiento

- Obtener el intervarlo (o los intervalos) con mayor frecuencia absoluta n_i
- Calcular la moda (o las modas) con la fórmula

$$Mo = l_i + (\frac{\Delta_1}{\Delta_2 + \Delta_1}) * c$$

Donde:
l_i:límite inferior del intervalo con mayor frecuencia absoluta
Δ_1:diferencia entre la mayor frecuencia absoluta y la anterior
Δ_2:diferencia entre la mayor frecuencia absoluta y la siguiente
c: amplitud del intervalo con mayor frecuencia absoluta

Continuando con nuestro ejemplo, vamos a calcular la moda partiendo de la tabla

Nº. de Intervalo	Minutos	n_i	f_i	N_i	F_i	$\dot{x}$
1	[45 - 50]	2	4%	2	4%	47,5
2	(50 - 55]	9	18%	11	22%	52,5
3	(55 - 60]	12	24%	23	46%	57,5
4	(60 - 65]	11	22%	34	68%	62,5
5	(65 - 70]	9	18%	43	86%	67,5
6	(70 - 75]	7	14%	50	100%	72,5

Figure 15: Enter Caption

- El intervalo de mayor frecuencia absoluta es el 3

- Los valores para el cálculo de la moda son:
  - $l_i = 55$
  - $\Delta_1 = 12 - 9 = 3$
  - $\Delta_2 = 12 - 11 = 1$
  - $c = 60 - 55 = 5$

  Por lo tanto la moda sería
  $Mo = 55 + (\frac{3}{4}) * 5 = 58,75$ minutos

Es decir, el tiempo que la mayoría de personas invierte para realizar la actividad es de 58,75 minutos

Ejercicio

Ejercicios: Continuando con nuestro ejercicio de las horas estudiadas por los alumnos. Calcula la media aritmética, mediana y moda

6 MEDIDAS DE POSICIÓN

Las medidas de posición, también llamadas cuantiles, son aquellas que permiten calcular valores en la distribución de los datos y que la dividen en partes iguales, de tal forma que los intervalos generados por los cuantiles contienen el mismo número de datos. Los cuantiles más usados son los cuartiles, deciles y percentiles.

6.1 Cuartiles

Los cuartiles (Q_k) son valores que fraccionan la distribución de los datos en cuatro partes iguales. Existen tres cuartiles y cada una de las partes representan un 25% de los datos.

El primer cuartil Q_1 deja por debajo el 25% de la distribución de los datos o el 75% por encima de él. El segundo cuartil (Q_2) acumula el 50% de los datos por debajo y el otro 50% por encima de él; y el tercer cuartil (Q_3) deja por debajo el 75% de los datos y por encima el 25%.

El cálculo de los cuartiles se realiza mediante el siguiente procedimiento:

- Ordenar los datos de forma ascendente.
- Calcular la posición i con la ecuación. $i = (k/4).n$.Donde k es el número de cuartil ($k = 1, 2, 3$) y n el número total de datos.
- Si i no es un númro entero, se debe redondear al entero siguiente y el valor que ocupa esta posición será el cuartil requerido. Si i es un número entero, el cuartil es el promedio de los valores de $i, i + 1$

Veamos un ejemplo de este algoritmo: Consultado un grupo de alumnos de Bachillerato por el número de horas que a la semana matemáticas, se obtiene el siguiente resultado: $3, 5, 2, 7, 6, 4, 9$ horas. Para calcular los cuartiles lo hacemos de la siguiente forma:

1. Ordenamos los datos de forma ascendente: $2, 3, 4, 5, 6, 7, 9$
2. Para el cuartil Q_1 la posición i sería: $i = \frac{1}{4} * 7 = 1, 75$
3. Dado que que i no es un entero, se redondea al entero siguiente, es decir a 2. En este caso, el cuartil Q_1 corresponde al valor ubicado en la posición 2, el cual es de 3 horas. Su interpretación significa que el 25% de los estudiantes dedican máximo 3 horas semanales

De forma similar, para el cuartil Q_2 la posición de i sería $i = (\frac{2}{4}) * 7 = 3, 5$. Como i no es un entero, se redondea al entero siguiente, es decir a 4. Por tanto, el cuartil Q_2 será el valor correspondiente a la posición 4, el cual es 5 horas. Es decir, el 50% de los estudiantes dedican máximo de 5 horas semanales. Este también se corresponde con la mediana.

Para el cuartil Q_3 la posición i sería: $i = (\frac{3}{4}) * 7 = 5, 25$, tomamos la posición 6 cuyo valor es de 7 horas. Indica que el 75% de los estudiantes dedican máximo 7 horas semanales.

Ejercicio

La altura de una muestra de alumnos de un curso de bachillerato es: 175,177,184,187,175,170,191,182,171 cm. Calcular los cuartiles siguiendo el ejemplo anterior. Interpreta los resultados.

Si los datos se han agrupado en clases o intervalos, los cuartiles se calculan mediante la siguiente ecuación:

$$Q_k = l_i + \left[\frac{k\left(\frac{n}{4}\right) - N_{i-1}}{n_i} \right] * C \tag{12}$$

Donde: k: número del cuartil, $k = 1, 2, 3$ n: número total de datos l_i: límite inferior del intervalo que contiene a $k(n/4)$ N_{i-1}: frecuencia absoluta acumulada anterior al intervalo que contiene a $k(n/4)$ C: amplitud del intervalo

Nº. de Intervalo	Intervalo (Estatura en cm)	n_i	f_i	N_i	F_i	$\dot{x}$
1	[150 - 155]	1	3%	1	3%	152.5
2	(155 - 160]	11	31%	12	34%	157.5
3	(160 - 165]	13	37%	25	71%	162.5
4	(165 - 170]	6	17%	31	89%	167.5
5	(170 - 175]	4	11%	35	100%	172.5

Figure 16: Enter Caption

EJEMPLO

En la tabla se dan los datos correspondiente a la estatura de un grupo de mujeres que asisten al gimnasio:
El Q_1 se calcula mediante el siguiente procedimiento:

1. Se halla $k(n/4), (1 * 35/4) = 8,75$
2. Se ubica el intervalo que contiene a contiene a $k(n/4)$ en la frecuencia absoluta acumulada N_i. (El segundo intervalo contine a 8,75 en la frecuencia absoluta acumulada).
3. El primer cuartil se obtiene mediante la fórmula.

$$Q_1 = l_i + \left\lfloor \frac{1\left(\frac{n}{4}\right) - N_{i-1}}{n_i} \right\rfloor * C$$

$$Q_1 = 155 + \left\lfloor \frac{1\left(\frac{35}{4}\right) - 1}{11} \right\rfloor * 5 = 159,4$$

De forma similar se calculan los cuartiles dos y tres. Se deja como ejercicio la resolución, sabiendo que el resultado $Q_2 = 162,1cm$ y $Q_3 = 166cm$

6.2 Deciles

Los deciles (D_k) son valores que fraccionan la distribución de los datos en diez partes iguales, las fórmulas son:

- Datos no agrupados y ordenados $i = (\frac{k}{10}) * n$
- Para datos agrupados en intervalos:
  $D_k = l_i + \left\lfloor \frac{k\left(\frac{n}{10}\right) - N_{i-1}}{n_i} \right\rfloor * C$

6.3 Percentiles

Los percentiles (P_k) son valores que fraccionan la distribución de los datos en cien partes iguales. En la distribución se presentan 99 percentiles: el primer percentil P_1 acumula el 1% del conjunto de datos, el percentil P_2 deja el 2%, y de forma similar los demás percentiles hasta llegar al percentil $P_9 9$ que acumula el 99% de los datos.

Para el cálculo de los percentiles se usa un procedimiento similar al empleado para los cuartiles y deciles:

- Datos no agrupados y ordenados $i = (\frac{k}{100}) * n$
- Para datos agrupados en intervalos:
  $P_k = l_i + \left\lfloor \frac{k\left(\frac{n}{100}\right) - N_{i-1}}{n_i} \right\rfloor * C$

7 MEDIDAS DE DISPERSIÓN

En el análisis de datos, es necesario conocer la variabilidad o la dispersión que los datos pueden tener en relación a una medida central. Las medidas de dispersión más representativas son: rango, rango intercuanrtil, varianza, desviación estándar y coeficiente de variación.

7.1 Rango

El rango es considerado como la medida de dispersión más simple para el análisis de los datos. No ofrece mucha información sobre la variabilidad de los datos por estar basada sólo en los valores extremos, razón por la cual debe ser usada como complemento de otras medidas de dispersión. Para el cálculo se utiliza la siguiente expresión.

Rango = valor máximo - valor mínimo

EJEMPLO

En una clase de bachillerato alumnos/as de mayor estatura miden 185 y alumnos/as de menor estatura es de 160.

Rango = 185 - 160=25 cm

7.2 Rango intercuartil

El rango intertcuartil (RIC) se denomina de esta manera porque es una medida de dispersión que evita que los valores extremos influyan en el conjunto de datos.Se calcula mediante la diferencia entre el cuartil tres (Q_3) y el cuartil uno (Q_1). Es decir, el rango intercuartil corresponde al rango del 50% ubicado en el centro de los datos. El RIC se calcula por medio de la siguiente expresión

$$\text{RIC} = Q_3\text{- } Q_1$$

EJEMPLO

En la tabla de la estatura de las mujeres, los cuartiles son $Q_1 = 159,4\text{cm}$ y $Q_3 = 166\text{cm}$ Así, el RIC resulta ser:

$$\text{RIC} = 166\text{- } 159,4 = 6,6\text{cm}$$

El intervalo dado por los Q_1 y Q_2 se le llama **mitad central**, es decir el 50% que contiene la información; y $6,6\text{cm}$ representa la dispersión media o rango intercuartil

7.3 Varianza

La varianza es una medida de dispersión basada en la diferencia de cada dato con la media aritmética. Hay que distinguir entre la *varianza poblacional* y la *varianza de la muestra*.
En la mayoría de los análisis estadísticos se emplea la varianza como una medida que permite comparar la dispersión entre dos o más variables, identificando la de mayor varianza como aquella que posee mayor dispersión o variabilidad. La importancia de la varianza está en que es una medida transitoria para el cálculo de la desviación típica o estándar de un conjunto de datos.
La *varianza poblacional*(σ^2) de una población de N datos y promedio μ, se obtiene con la siguiente expresión

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \tag{13}$$

La *varianza de la muestra*(s^2) tiene como objetivo convertirse en un estimador de la variación de la población

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{N - 1} \tag{14}$$

Donde:
$\bar{x}$: media aritmética de la muestra n: total de los datos de la muestra x_i: cada dato u observación de la variable X

Empleado	Calificación (x_i)	Media de la muestra ($\bar{x}$)	Desviación ($x_i - \bar{x}$)	Desviación al cuadrado (σ^2)
1	3,5	3,6	-0,1	0,01
2	4,5	3,6	0,9	0,81
3	4,2	3,6	0,6	0,36
4	3,0	3,6	-0,6	0,36
5	2,7	3,6	-0,9	0,81
6	3,3	3,6	-0,3	0,09
7	4,0	3,6	0,4	0,16
			$\sum(x_i - \bar{x}) = 0$	$\sum(x_i - \bar{x})^2 = 2,6$

Figure 17: Enter Caption

EJEMPLO

En la tabla la calificación del 1 al 5 por los comerciales de una empresa de productos lácteos. Con los resultados de la tabla la varianza será:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{2,6}{6} = 0,43 \tag{15}$$

En el ejemplo anterior el resultado obtenido indica la variación de la calificación entre los comerciales de 0,43.

Si los datos se agruparon en frecuencias o intervalos, la varianza puede calcularse mediante las siguiente fórmulas:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} - \mu^2$$
$$s^2 = \frac{\sum(x_i - \mu)^2}{N-1} - \bar{x}^2$$

Donde: $\bar{x}$:media aritmética
n: total de datos de la muestra
N: total de datos de la población
x_i: cada dato de la variable o marca de clase si es intervalo
n_i: frecuencia absoluta

Ejercicio

Se propone como ejercicio el cálculo de varianza del ejemplo que hemos ido haciendo a lo largo del tema

7.4 Desviación estándar

La desviación estándar es considerada la medida de dispersión con mayor representatividad para un conjunto de datos. La varianza para una muestra es (s) y para la población es (σ), su calculo es:

$$\sigma = \sqrt{\sigma^2}$$
$$s = \sqrt{s^2}$$

La desviación estándar indica la distribución de los datos alrededor de la media aritmética o promedio. Cuando la distribución de los datos se aproxima a una forma de campana o es simétrica, como se ilustra en la gráfica, la desviación estándar puede interpretarse mediante la regla empírica, esta es: el 68% de los datos se agrupan alrededor de la media, entre el intervalo $(\bar{x} - 1s)$ y $(\bar{x} + 1s$, el 95% entre$(\bar{x} - 2s)$ y $(\bar{x}+2s)$ Siguiendo con el ejemplo de las estaturas de las mujeres la desviación estándar

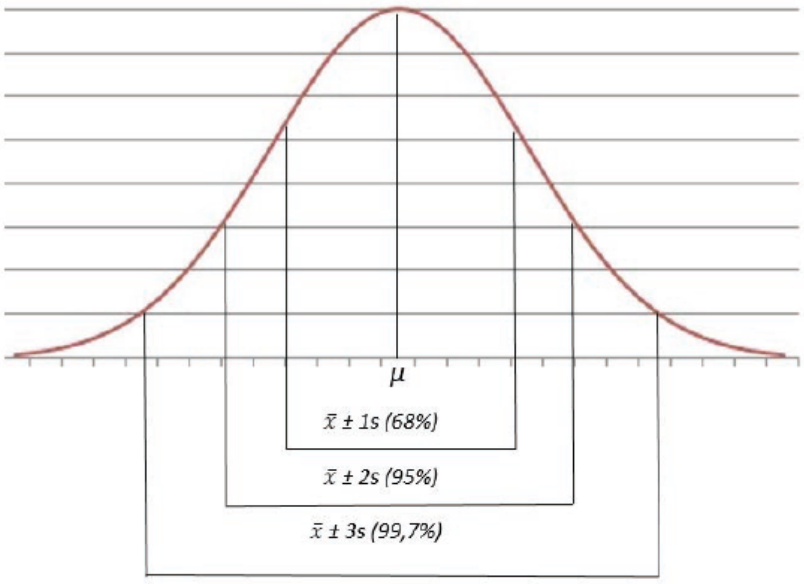

Figure 18: Enter Caption

sería:

$s = \sqrt{s^2} = \sqrt{819,1} = 28,6cm$

Se interpreta como que la estatura varía 28, 6 cm alrededor de la media $(162, 6cm)$.Por la regla empírica puede decirse que el 68% de las estaturas está dentro de una desviación estándar de la media.

7.5 Coeficiente de variación

El coeficiente de variación (CV) es una medida que relaciona la desviación estándar con la media aritmética para determinar qué tan homogénea o dispersa es la información. Expresa el porcentaje que representa la desviación con relación a la media aritmética y se calcula por medio de la siguiente ecuación:

$$CV = \frac{S}{\overline{X}} * 100 \tag{16}$$

Cuando se tiene una muestra, el coeficiente de variación puede ser utilizado para calificar estadísticamente la calidad de las estimaciones, un valor bajo indica que los datos están más cerca de la media, lo que significa menor dispersión. Para ello se consideran los siguientes criterios

- CV menor o igual al 7%, las estimaciones se consideran precisas.
- CV entre el 8% y el 14%, las estimaciones tienen precisión aceptable
- CV entre 15% y 20%, la precisión es regular
- CV mayor del 20%, la estimación es muy poco precisa

En el ejemplo de las estaturas de las mujeres $CV = \frac{S}{\overline{X}} * 100 = \frac{28,6}{162,6} * 100 = 17,6\%$

8 BIBLIOGRAFÍA

Elementos básicos de estadística descriptiva para el análisis de datos; Gabriel Jaime Posada Hernández, Editorial Luis Amigo