HISTOGRAMA Y POLIGONO DE FRECUENCIAS

El análisis detallado de histogramas es esencial para comprender la distribución de datos.
Esta herramienta clave permite visualizar la frecuencia de ocurrencia de diferentes valores en un conjunto de datos, lo que facilita la identificación de patrones y tendencias.
En esta presentación, exploraremos la importancia y el uso efectivo de los histogramas en el análisis de datos.

Un histograma es un gráfico que representa la distribución de frecuencias de un conjunto de datos.
Se compone de barras verticales donde la altura de cada barra indica la frecuencia de los valores que caen en un intervalo específico.
Este tipo de visualización proporciona una representación visual clara de la distribución de los datos.

La tabla de datos que figura a continuación corresponde a una muestra, tomada aleatoriamente durante 20 días, del peso en gramos de cierto embutido que puede ser elaborado por dos máquinas distintas (1 y 2), que a su vez son atendidas indistintamente por dos operarios (A y B).

Las especificaciones del peso son 220 ± 10 g, y últimamente se han detectado ciertos problemas a este respecto.

DÍA	OPERAR.	MÁQUINA 1				MÁQUINA 2			
1	A	220.3	215.5	219.1	219.2	220.3	208.0	214.4	219.2
2	B	215.8	222.0	218.9	213.6	216.9	213.4	217.7	217.7
3	B	220.4	218.7	218.6	219.6	222.9	219.7	209.4	221.6
4	B	221.5	227.0	219.5	222.5	223.1	215.3	220.4	215.6
5	A	215.7	225.3	223.0	218.0	216.0	210.9	221.4	210.9
6	A	222.7	215.1	219.6	217.3	212.1	213.0	218.0	216.5
7	A	216.0	218.8	217.9	213.0	216.9	216.0	213.5	219.2
8	B	219.4	218.3	216.7	224.1	216.2	218.4	216.6	214.9
9	B	219.8	222.6	219.1	217.7	216.2	212.2	216.9	214.9
10	A	220.2	219.5	222.4	219.9	222.9	214.3	219.1	216.7
11	B	218.0	223.9	219.6	221.9	214.9	212.6	219.4	212.3
12	B	219.3	219.6	218.8	219.9	219.0	216.7	216.4	213.5
13	B	220.0	214.1	224.3	217.4	218.0	219.5	219.5	222.3
14	A	223.9	220.6	219.5	219.6	211.8	218.2	218.3	217.4
15	A	218.1	218.8	218.4	217.9	214.6	215.7	218.0	216.4
16	B	216.9	221.6	220.6	222.6	215.6	220.4	217.3	216.2
17	B	217.9	225.7	222.2	216.1	212.5	214.6	209.7	211.3
18	A	224.2	216.2	219.9	220.4	215.8	219.9	216.5	211.9
19	A	214.1	219.7	222.4	224.5	213.7	209.7	216.9	213.1
20	A	221.1	225.0	222.7	222.2	212.5	217.5	217.4	215.7

Cuando se trata, como en este caso, de analizar la dispersión que presentan unos datos, la representación gráfica más adecuada es el histograma. Para realizar un histograma se marcan una serie de intervalos sobre un eje horizontal, y sobre cada intervalo se coloca un rectángulo de altura proporcional al número de observaciones (frecuencia absoluta) que caen dentro de dicho intervalo.

CONSTRUCCION DE UN HISTOGRAMA

Al construir los histogramas a mano conviene seguir una sistemática adecuada como la siguiente:

1. Colocar los datos a representar en filas.

2. Identificar y señalar el máximo y el mínimo de cada fila.

3. A partir del máximo y el mínimo de cada fila, localizar el máximo y el mínimo globales.

4. Calcular el rango (R) de los datos.

$$R = \text{Valor máximo} - \text{Valor mínimo}$$

5. Optar por un número de intervalos (k), en primera aproximación, utilizando la siguiente tabla o haciendo la raíz cuadrada del total de datos

NÚM. DE DATOS	NÚM. DE INTERVALOS
<50	5 - 7
50 - 100	6 - 10
100 - 250	7 - 12
>250	10 - 20

La selección del número de clases afecta su ancho y junto con la elección del punto de partida para el primer intervalo, afecta la forma del histograma. Hay algunas "reglas generales", como las antes indicadas que pueden ayudar a elegir un ancho apropiado, pero tenga en cuenta que ninguna de las reglas es exacta.

6. Determinar la amplitud (h) de los intervalos, haciendo:

$$h = \frac{R}{k}$$

y redondeando el valor obtenido a un múltiplo exacto, en base a la precisión de los datos presentados.

7. Fijar los límites de los intervalos. Para evitar el problema que se presenta al asignar un valor a un intervalo cuando dicho valor coincide con el extremo superior de un intervalo y el extremo inferior del otro, conviene fijar dichos extremos con una precisión igual a la mitad de la precisión de los valores. Así, si los datos se presentan con un solo decimal y los extremos de los intervalos son de la forma 2,15 - 2,35, está claro que los valores 2,2 y 2,3 deberán situarse en este intervalo; 2,4 en el intervalo siguiente, etc.

8. Rellenar la tabla de frecuencias, indicando el número de veces que aparecen datos dentro de cada uno de los intervalos definidos.

9. Construir el histograma.

INTERPRETACION DE HISTOGRAMAS

En las figuras debajo se presentan varias formas de histograma que responden a patrones de comportamiento típico.

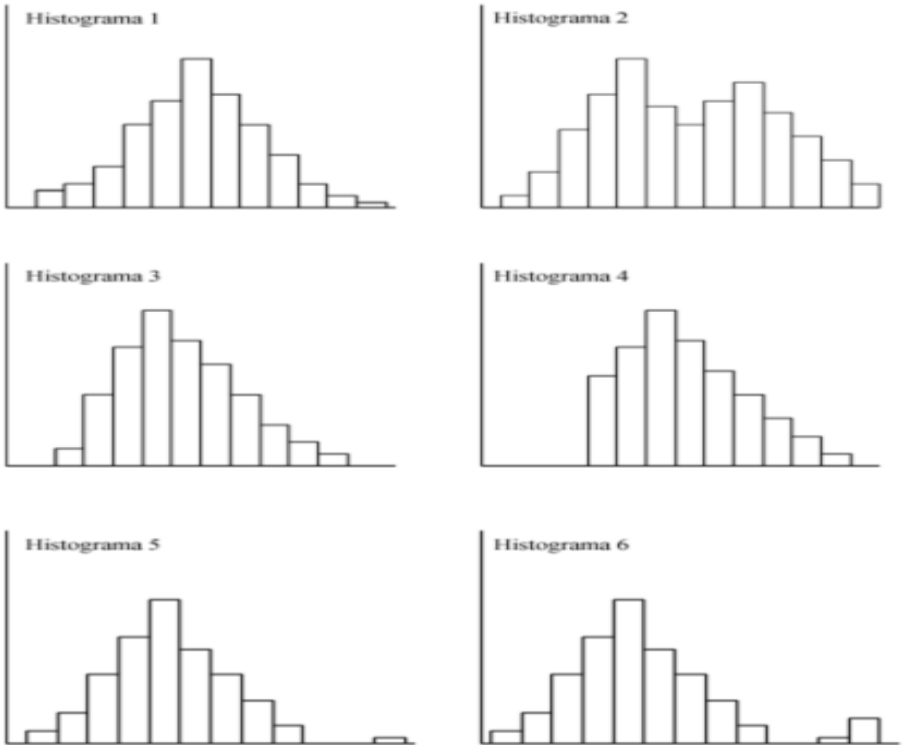
El histograma 1 corresponde a la forma de campana habitual que representa la variabilidad debida a causas aleatorias.

El histograma 2, con dos máximos diferenciados, responde a una distribución denominada bimodal y se presenta cuando están mezclados datos de distinto origen centrados en valores distintos.

El histograma 3 se denomina, por su forma, sesgado a la derecha, y responde a la variabilidad que presentan ciertas variables que no siguen una ley normal, como los tiempos de vida. También puede representar una magnitud con un "cero natural", como la tolerancia entre eje y cojinete.

Al histograma 4 parece faltarle una parte y por ello se le llama censurado (en este caso, a la izquierda). No representa una variabilidad natural y por tanto hay que sospechar que se han eliminado algunos valores. Esto ocurre si después de la producción se realiza una inspección al 100 % para separar las unidades fuera de tolerancias.

En los histogramas 5 y 6 aparecen datos que no siguen el patrón de comportamiento general (anomalías, errores, etc.). Su variabilidad puede atribuirse a alguna causa asignable que deberá ser identificada y eliminada.



ERRORES COMUNES DE INTERPRETACION

a) Interpretar el histograma como un gráfico de barras, donde las barras representan puntos de datos individuales, en lugar de agrupados;

b)

(b) Interpretar el histograma como un gráfico con dos variables, ya sea como un gráfico de dispersión o una serie de tiempo;

(c) Observar el eje vertical y comparar las diferencias en la altura de las barras para comparar la variación en dos histogramas;

(d) Una tendencia a pensar de manera determinista al interpretar una distribución en un contexto del mundo real.

Caso Práctico 1:

Construya un histograma y un polígono de frecuencia usando frecuencias relativas para distribución de kilómetros que corrieron 20 corredores seleccionados al azar durante una semana dada

Clases	Frecuencia	Frecuencia Acumulada
5.5 – 10	1	1
10.5 – 15	2	3
15.5 – 20	3	6
20.5 – 25	5	11
25.5 – 30	4	15
30.5 – 35	3	18
35.5 - 40	2	20

Convierta cada frecuencia en una proporción o frecuencia relativa dividiendo el frecuencia para cada clase por el número total de observaciones.
Para la clase 5.5–10, la frecuencia relativa es 1/20= 0.05; para la clase 10.5–15, la frecuencia relativa es 2/20 = 0.10; para la clase 15.5–20, la frecuencia relativa es 3/20= 0,15; y así.
Coloque estos valores en la columna etiquetada Frecuencia relativa.

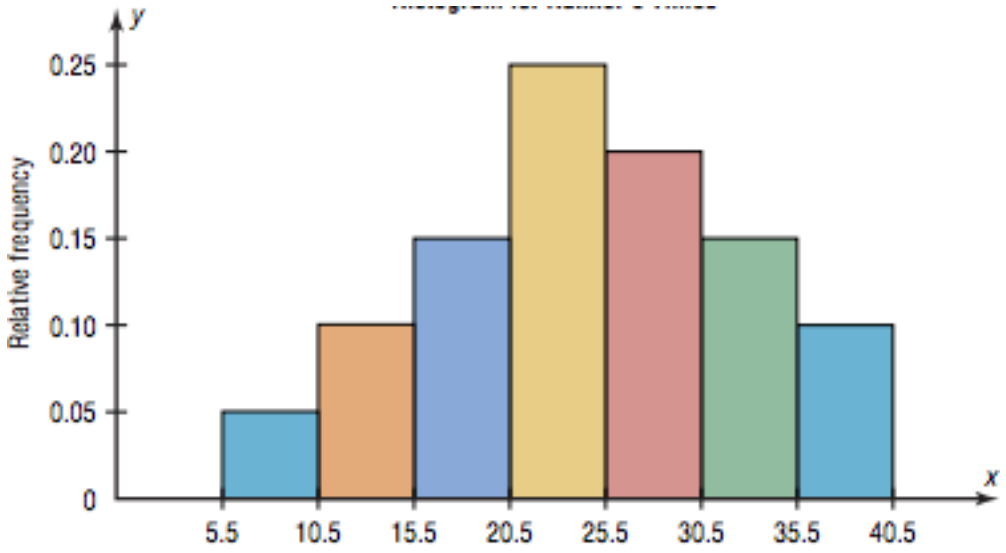
Encuentra las frecuencias relativas acumuladas. Para hacer esto, agregue la frecuencia en cada clase a la frecuencia total de la clase anterior. En este caso, 0 +0.05 =0.05, 0.05+0.10= 0.15, 0.15+ 0.15= 0.30, 0.30+ 0.25= 0.55, etc.
Colocar estos valores en la columna denominada Frecuencia relativa acumulada.

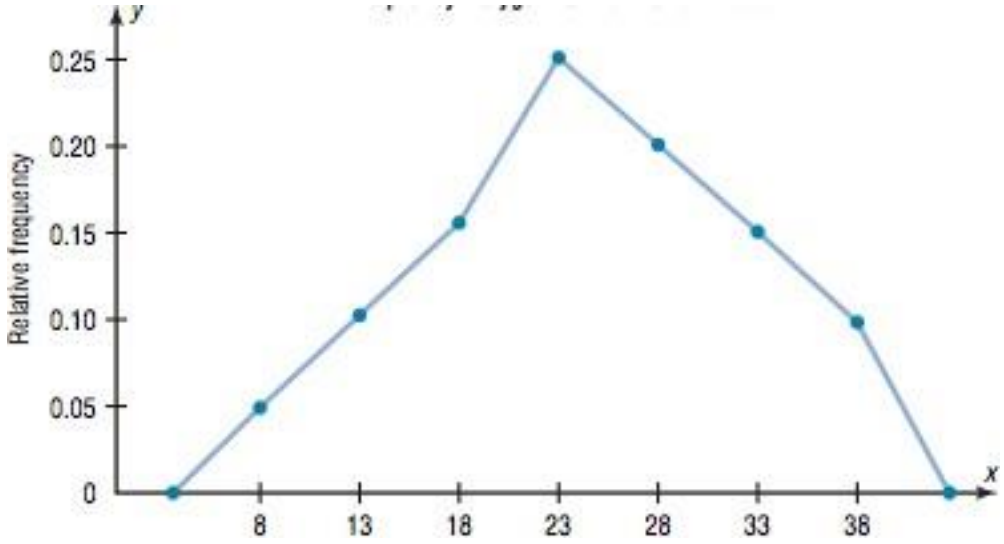
Usando el mismo procedimiento, encuentre las frecuencias relativas para la columna frecuencia acumulada. Las frecuencias relativas se muestran aquí.

Clases	Frecuencia Relativa	Frecuencia Relativa Acumulada
5.5 – 10	0.05	0.05
10.5 – 15	0.10	0.15
15.5 – 20	0.15	0.30
20.5 – 25	0.25	0.55
25.5 – 30	0.20	0.75
30.5 – 35	0.15	0.90
35.5 - 40	0.10	1.00

Dibuje cada gráfico como se muestra.

Para el histograma usar los límites de la clase a lo largo del eje *x*. Para el polígono de frecuencia, usar los puntos medios aproximados en el eje *x*. La escala en el eje *y* usa dimensiones.





Los histogramas pueden basarse en frecuencias relativas en lugar de frecuencias reales.

Los histogramas basados en frecuencias relativas muestran la proporción de repeticiones en cada intervalo en lugar del número de repeticiones.

En este caso, el eje Y va de 0 a 1 (o en algún punto intermedio si no hay proporciones extremas).

Puede cambiar un histograma basado en frecuencias a uno basado en frecuencias relativas dividiendo cada frecuencia de clase por el número total de observaciones, y luego utilizar los cocientes en el eje Y (etiquetados como porcentajes).

Caso Práctico 2:

Elaborar un histograma de frecuencias para el siguiente conjunto de datos:

271 236 294 252 254 263 266 220 262 278 288
262 237 247 282 224 263 267 254 271 278 263
262 288 247 252 264 263 247 225 281 279 238
252 242 248 263 255 294 268 255 272 271 291
263 242 288 252 226 263 269 227 273 281 267
263 244 249 252 256 263 252 261 245 252 294
288 245 251 269 256 264 252 232 275 284 252
263 274 252 252 256 254 269 234 285 275 263
246 263 294 252 231 265 269 235 275 288 294
263 247 252 269 261 266 269 236 276 248 298

Paso 1: Determinar el número de intervalos o clases

Podemos emplear cualquiera de los métodos mostrados en la explicación anterior; sin embargo, para nuestro ejemplo usaremos el método de la raíz cuadrada. De esta forma, para el conjunto de 110 datos, nuestro número de intervalos o clases es:

$$K = \sqrt{110} \cong 10$$

Paso 2: Calcular la amplitud del intervalo

En nuestro conjunto de datos, el valor mínimo de nuestro conjunto de datos es 220 y el valor máximo es 298; por lo tanto, la amplitud de la clase es:

$$Amplitud\ de\ Clase = \frac{(298 - 220)}{10} \cong 8$$

Paso 3: Calcular el número total de ocurrencias para intervalo / rango

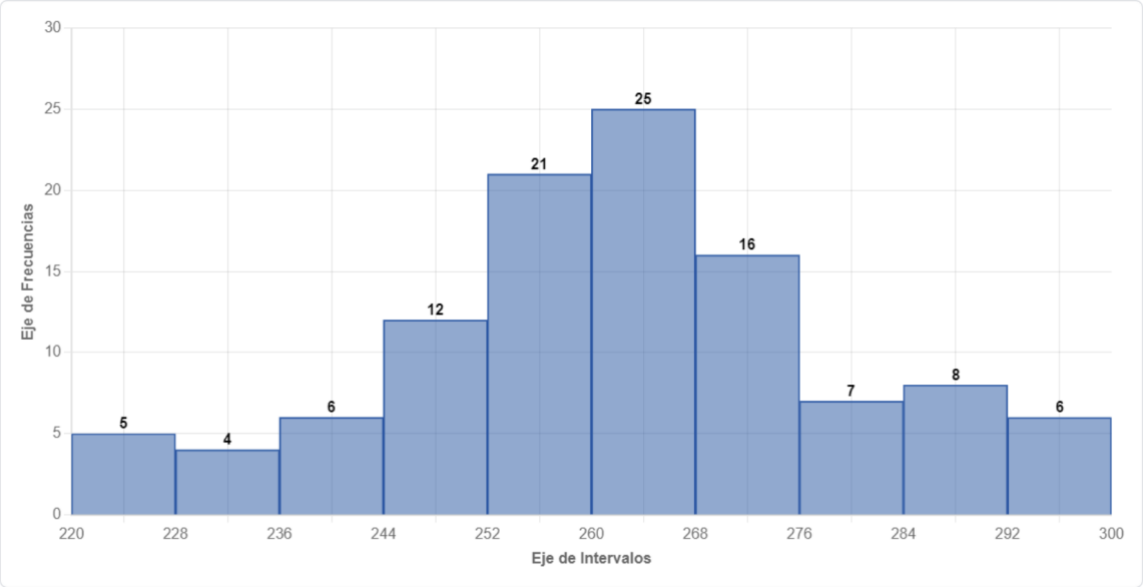
Realizamos el conteo de las ocurrencias de cada observación en su respectivo intervalo. Este conteo se conoce como frecuencias absolutas.
Nuestra tabla de frecuencias quedaría de la siguiente forma:

Clases	Punto Medio	Frecuencia Absoluta
[220-228)	224	5
[228-236)	232	4
[236-244)	240	6
[244-252)	248	12
[252-260)	256	21
[260-268)	264	25
[268-276)	272	16
[276-284)	280	7
[284-292)	288	8
[292-300)	296	6

En este caso se tomó intervalos con iguales límites pero una forma de nomenclatura que es posible usar es manejo de corchetes para inidcar intervalo cerrado y abierto, el corchete recto es cerrado y el corchete curvo abierto , por ejejmplo para la primer clasr el 228 pertenece a la clase uno pero no a la siguiente por estar el corchete cerrado.

En el curso la forma de trabajarlo es como se inidcó, tratando que solo coincidan las clases si no hay valores en ese punto o tmando valores límites fraccionarios para evitar tener coincidencias de dos valores en una misma clase.

Finalmente, graficamos las barras según las observaciones de cada intervalo



Caso Práctico 3:

Los siguientes datos se corresponden con los retrasos (en minutos) de una muestra de 30 vuelos de cierta compañía área por semana y 35 durante los fines de semana.

Por semana
12 25 3 16 11 7 21 54 19 6 11 23 21 7 13
14 187 15 21 7 5 17 21 13 26 21 19 8 24 16

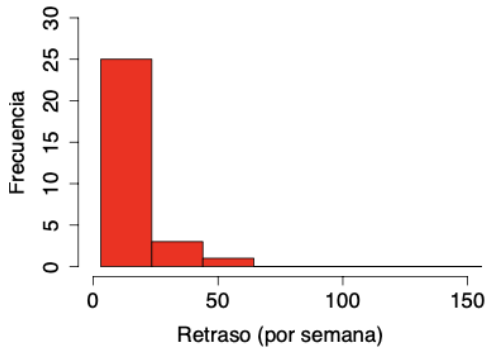
Fines de semana
16 11 3 12 21 13 17 12 24 21 14 18 7. 11 62
13 17 19 48 11 9 13 21 126 11 17 8 23 15 52
11 14 17 71 12

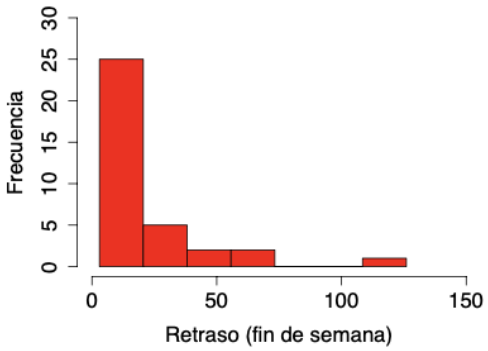
¿Se parecen las distribuciones de retrasos por semana y en fin de semana?

El objetivo es determinar si las distribuciones de retrasos por semana y en fin de semana se parecen a grandes rasgos
Planteamiento: el experimento consiste en seleccionar vuelos (individuos) y observar su retraso en minutos (variable). Se distinguen dos poblaciones: los vuelos por semana y los vuelos de fin de semana. Se tiene una muestra de 30 vuelos por semana y 35 de fin de semana. Los posibles valores de la variable son {9, 15, 18, . . . }.

Método y justificación: como se trata de comparar la distribución a grandes rasgos lo mejor es hacer un gráfico. Como la variable es continua, el gráfico más adecuado es el histograma.

Es necesario tener cuidado con los valores atípicos.





Si retiramos los valores atípicos