Contenido

1. Introducción

1.1. Definición de estadística

La estadística es la parte de las matemáticas que se encarga de estudiar sucesos no deterministas. Este carácter no determinista implica que no se pueden deducir el resultado de un experimento previamente a su realización.

Un ejemplo de una ciencia determinista es la física. Si un vehículo se desplaza a 50 km/h durante 5 minutos, puede determinarse con exactitud que distancia recorrerá.

A su vez la estadística trata con grupos grandes lo que genera la necesidad de organizar una cantidad de información más o menos grande.

1.2. Ramas de la estadística y objetivos

A lo largo de este curso se verán 3 bloques principales: estadística descriptiva, probabilidad en inferencia estadística.

La estadística descriptiva es la parte de la estadística que analiza las características de un grupo que se esté estudiando, pero no intenta predecir esas características. Ejemplo, ordenar todas las notas de una clase y calcular la nota media de la clase.

Por su parte la probabilidad se puede definir como la certeza de que ocurra un determinado evento en un experimento aleatorio

La estadística inferencial intenta predecir algunas características de un grupo de personas a partir del resultado de una muestra. Por ejemplo, si durante el trascurso de los años, una asignatura A tiene más nota media que una asignatura B, se puede considerar que las y los estudiantes consideran la asignatura A es más fácil y es esperable que en años posteriores el número de aprobados de la asignatura A sea mayor que el número de aprobados de la asignatura B.

2. Población y muestra

2.1. Población

Población es el conjunto de individuos sobre los cuales se desea analizar una característica. Ejemplos:

- Ejemplo 1. Se desea conocer el número de ciudadanos de un concejo que aprueba la gestión de su alcalde. La población son todos los ciudadanos
- Ejemplo 2. En una fábrica de refrescos se desea realizar un análisis de calidad y conocer si las latas finales están en buenas condiciones. La población son todas las latas de refrescos

La característica de la población que se desea estudiar se le llama parámetro. Si lo que se está estudiando, esto es, con lo que se está trabajando es con el conjunto de la población, al estudio se le llama censo.

2.2. Muestra

En algunas ocasiones no resulta práctico realizar el estudio a toda la población por lo que se coge una parte de ésta de manera que se pueda considerar que el resultado que se obtenga en la muestra representa de manera fiel al conjunto de la población. Determinar cuan fiel es esa representación forma parte de la estadística inferencial.

A la característica de la muestra que se desea estudiar se le llama estadístico. Si se está trabajando con muestras, al estudio se le llama muestreo.

3. Clasificación de los parámetros o los estadísticos

Los parámetros o los estadísticos se clasifican en atributos o variables.

- Atributos: son aquellos parámetros o estadísticos que representan características cualitativas, es decir, que no se pueden representar por números. Por ejemplo, el color de ojos de una población. Este grupo a su vez se puede dividir en atributos nominales o atributos ordinales. Los atributos nominales no se pueden ordenar, por ejemplo, las profesiones; mientras que los ordinales sí se pueden ordenar, por ejemplo, la carácter leve, grave o muy grave de un conjunto de faltas.
- Variables: son aquellos parámetros o estadísticos que representan una característica cuantitativa, es decir, se puede representar por números. Por ejemplo, el peso de una población. Las variables se suelen representar por letras mayúsculas: X, Y…. Estos a su vez se dividen en variables continuas o variables discretas. Las variables continuas son aquellas que pueden tomar infinitos valores, por ejemplo, el peso; sin embargo, las discretas solo pueden tomar unos valores concretos, por ejemplo, el número de años solo pueden tomar los valores 0, 1, 2, 3…

4. Distribuciones de frecuencias

A cada una de los posibles valores que puede tomar la variable se le designa por X_i, siendo i: 1, 2, 3…

4.1. Frecuencia absoluta, n_i

Representa el número de veces que se repite cada valor.

4.2. Frecuencia absoluta acumulada, N_i

La frecuencia absoluta acumulada de un valor representa el número de valores que son iguales o más pequeños que el valor indicado. Es decir, es como la frecuencia absoluta (n_i), pero sumando todas las frecuencias absolutas de todos los valores más pequeños al dado. Por eso se llama acumulada, porque según va aumentando el valor va sumando las frecuencias anteriores.

$$N_i = \sum_{k=1}^{k=i} n_i$$

4.3. Frecuencia relativa, f_i

Representa el número de veces que se repite cada valor en tanto por uno respecto al tamaño muestral N. Es decir, se calcula con la siguiente fórmula:

$$f_i = \frac{n_i}{N}$$

4.4. Frecuencia relativa acumulada, F_i

La frecuencia relativa acumulada de un valor representa el número de valores que son iguales o más pequeños que el valor indicado expresado en tanto por uno respecto al tamaño muestran N. Es decir, es como la frecuencia relativa (f_i), pero sumando todas las frecuencias relativas de todos los valores más

pequeños al dado. Por eso se llama acumulada, porque según va aumentando el valor va sumando las frecuencias anteriores.

$$F_i = \sum_{k=1}^{k=i} f_i = \sum_{k=1}^{k=i} \frac{n_i}{N}$$

4.5. Porcentaje, p_i

Es lo mismo que la frecuencia relativa (f_i) del apartado 4.3 pero expresado en %. Es decir:

$$p_i = f_i \cdot 100$$

4.6. Porcentaje acumulado, P_i

Es lo mismo que la frecuenta relativa acumulada (F_i) pero expresado en porcentaje. Es decir:

$$P_i = F_i \cdot 100$$

4.7. Ejemplo de tabla con todas las frecuencias

Se tiene un bombo con 20 bolas que tienen los siguientes números:

- 4 bolas con el número 1.
- 1 bolas con el número 2.
- 2 bolas con el número 3.
- 1 bolas con el número 4.
- 3 bolas con el número 5.
- 1 bolas con el número 6.
- 5 bolas con el número 7.
- 1 bolas con el número 8.
- 2 bolas con el número 9.

X_i	n_i	N_i	f_i	F_i	p_i	P_i
1	4	4	0,2	0,2	20 %	20 %
2	1	5	0,05	0,25	5 %	25 %
3	2	7	0,1	0,35	10 %	35 %
4	1	8	0,05	0,4	5 %	40 %
5	3	11	0,15	0,55	15 %	55 %
6	1	12	0,05	0,6	5 %	60 %
7	5	17	0,25	0,85	25 %	85 %
8	1	18	0,05	0,9	5 %	90 %
9	2	20 (1)	0,1	1 (2)	10 %	100 % (3)

(1) Este valor debe coincidir con el número total de bolas, N=20.
(2) Este valor debe coincidir siempre con 1.
(3) Este valor debe coincidir siempre con 100 %.

4.8. Marca de clase

En algunas situaciones se pueden agrupar varios valores X_i en intervalos. En esos casos se define la marca de clase como el valor medio del intervalo y se toma a este valor como representativo del intervalo pasando esta marca de clase a representarse como X_i y los 2 extremos del intervalo se denotarán como L_{i-1} y L_i.

Para ejemplificar esto usaremos como ejemplo la estura de 400 personas que serán tabuladas en intervalos que van de 10 cm en 10 cm salvo en el primer intervalo que tiene una amplitud de 20 cm.

L_{i-1} - L_i	X_i	n_i	N_i	f_i	F_i	p_i	P_i
130 – 150	140	38	38	0 ,095	0,095	9,5 %	9,5 %
150 – 160	155	80	118	0,2	0,295	20 %	29,5 %
160 – 170	165	186	304	0,465	0,76	46,5 %	76 %
170 – 180	175	60	364	0,15	0,91	15 %	91 %
180 – 190	185	24	388	0,06	0,97	6 %	97 %
190 - 200	195	12	400	0,03	1	3 %	100 %

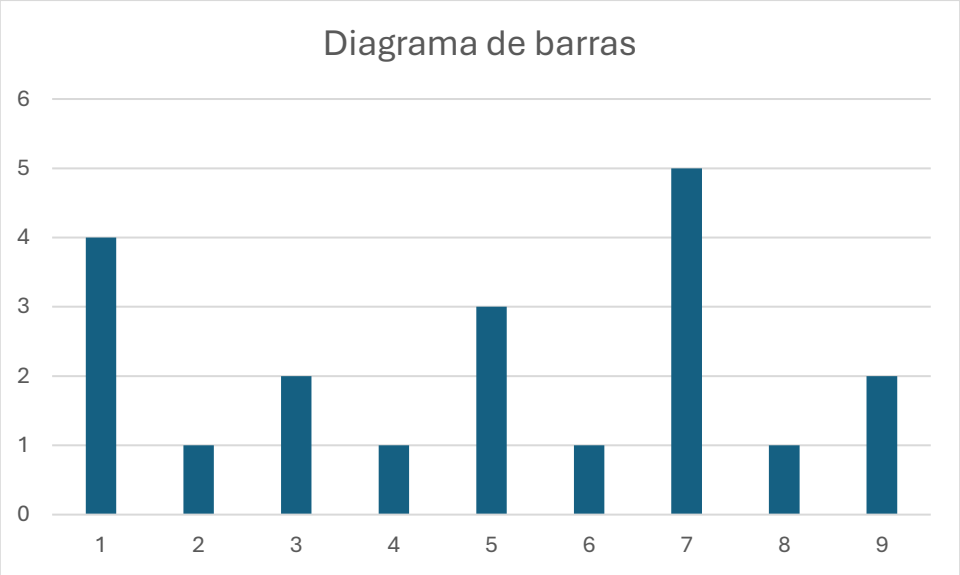
5. Representaciones gráficas

Los diagramas suelen representarse horizontalmente, es decir, los valores de la variable se anotan en el eje x y las frecuencias en el eje y. Si bien se puede hacer al revés, todos los ejemplos de este documento estarán hechos en horizontal.

5.1. Diagrama de barras

En el diagrama de barras se representa un rectángulo asociado a cada valor de la variable. La altura del rectángulo representa la frecuencia de cada valor. En este caso particular el área del rectángulo no es relevante, de hecho, puede representarse como una simple línea.

Para el ejemplo de las bolas del apartado 4.7 el diagrama de barras sería:



Hay que darse cuenta de que este diagrama de barras está realizado con frecuencias absolutas. También podría plantearse un diagrama de barras con frecuencias relativas o porcentajes. En cualquier caso, no se deben usar nunca los datos acumulados.

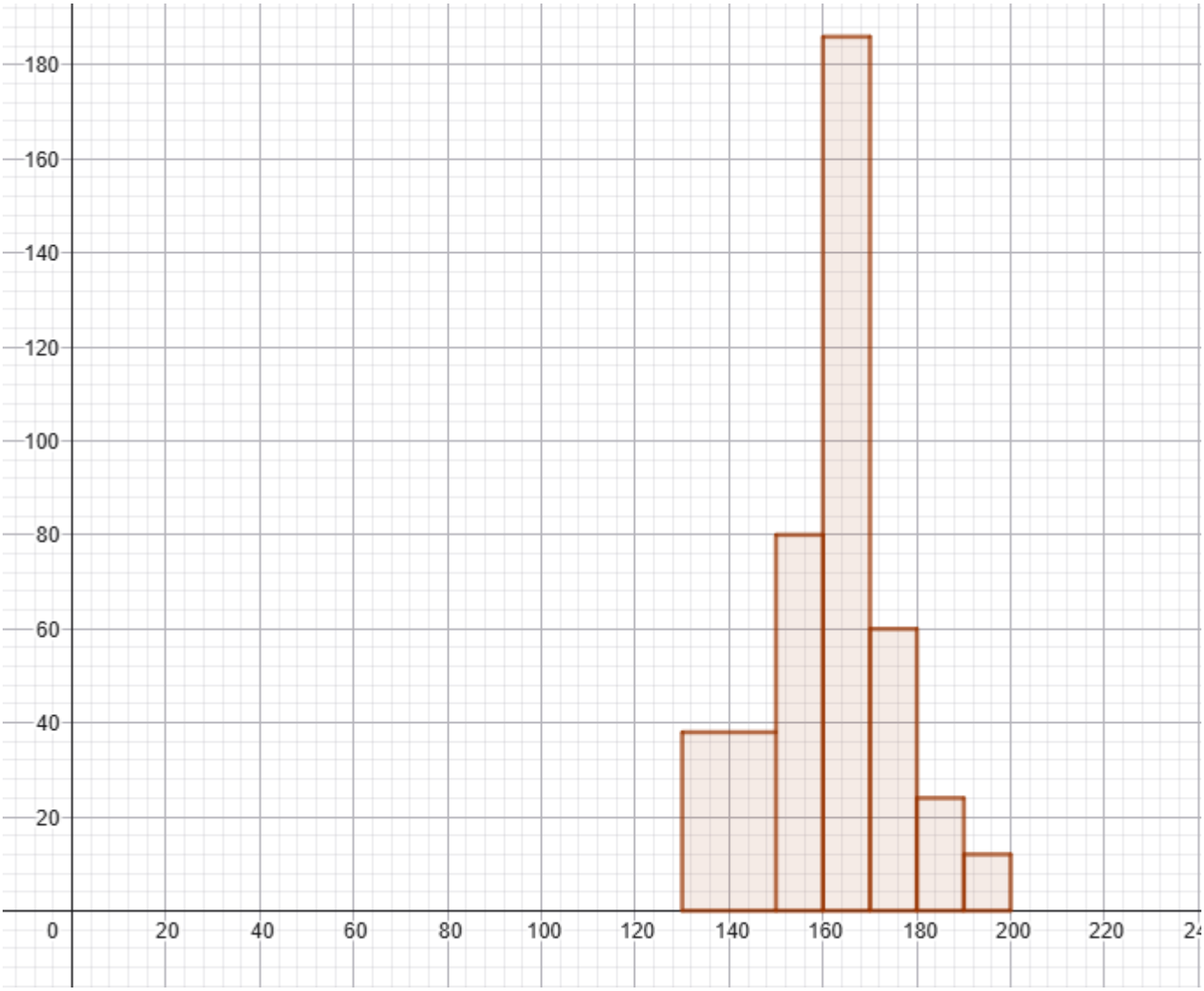
5.2. Histograma

El histograma se usa para valores agrupados como los vistos en el apartado 4.8. Para la representación de cada intervalo se usan rectángulos. En estos casos importa tanto la altura de la barra, como la base del rectángulo.

El histograma puede ser de frecuencias absolutas, frecuencias absolutas acumuladas, frecuencias relativas, frecuencias relativas acumuladas, porcentajes y porcentajes acumulados.

5.2.1. Altura del intervalo = frecuencia

En su versión más simple, la altura del intervalo es una de las frecuencias vistas y la base del rectángulo es la amplitud del intervalo. A continuación, se muestra como ejemplo el histograma las alturas del apartado 4.8



5.2.2. Altura del intervalo = densidad de frecuencia

- La base del rectángulo representa la longitud del intervalo.
- La altura del intervalo se llama densidad de frecuencia o densidad de porcentaje y se representa con h_i.
- El área del rectángulo representa la frecuencia o el porcentaje.

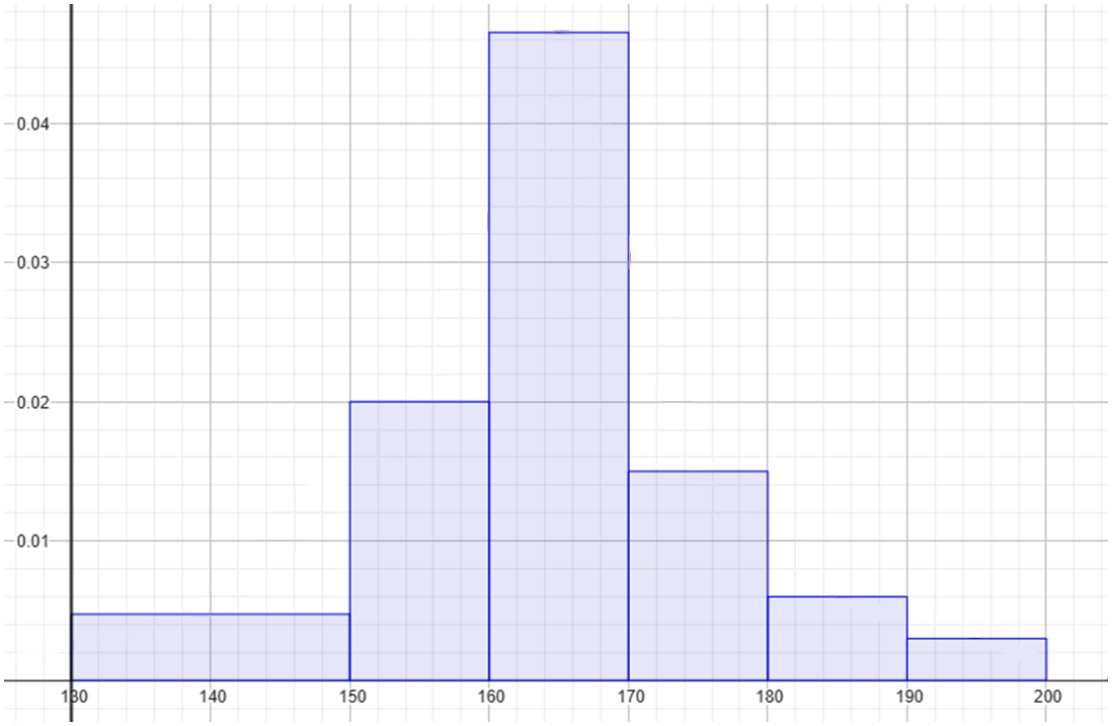
$$f_i = (L_i - L_{i-1}) \cdot h_i$$

Para ejemplificarlo se verá el mismo ejemplo de la tabla del apartado 4.8. Previamente se escribirá la tabla nuevamente, pero añadiendo una columna nueva para indicar las densidades de frecuencia h_i.

En la fórmula anterior, se puede despejar h_i:

$$h_i = \frac{f_i}{L_i - L_{i-1}}$$

L_{i-1} - L_i	X_i	n_i	N_i	f_i	h_i	F_i	p_i	P_i
130 – 150	140	38	38	0 ,095	$4{,}75 \cdot 10^{-3}$	0,095	9,5 %	9,5 %
150 – 160	155	80	118	0,2	0,02	0,295	20 %	29,5 %
160 – 170	165	186	304	0,465	0,0465	0,76	46,5 %	76 %
170 – 180	175	60	364	0,15	0,015	0,91	15 %	91 %
180 – 190	185	24	388	0,06	$6 \cdot 10^{-3}$	0,97	6 %	97 %
190 - 200	195	12	400	0,03	$3 \cdot 10^{-3}$	1	3 %	100 %



5.3. Diagrama de cajas

En el diagrama de cajas se dibuja un rectángulo en posición vertical sobre sistema de ejes. La base superior del rectángulo es el Q_3, la base inferior es el Q_1. Atravesando el rectángulo señalamos la mediana.

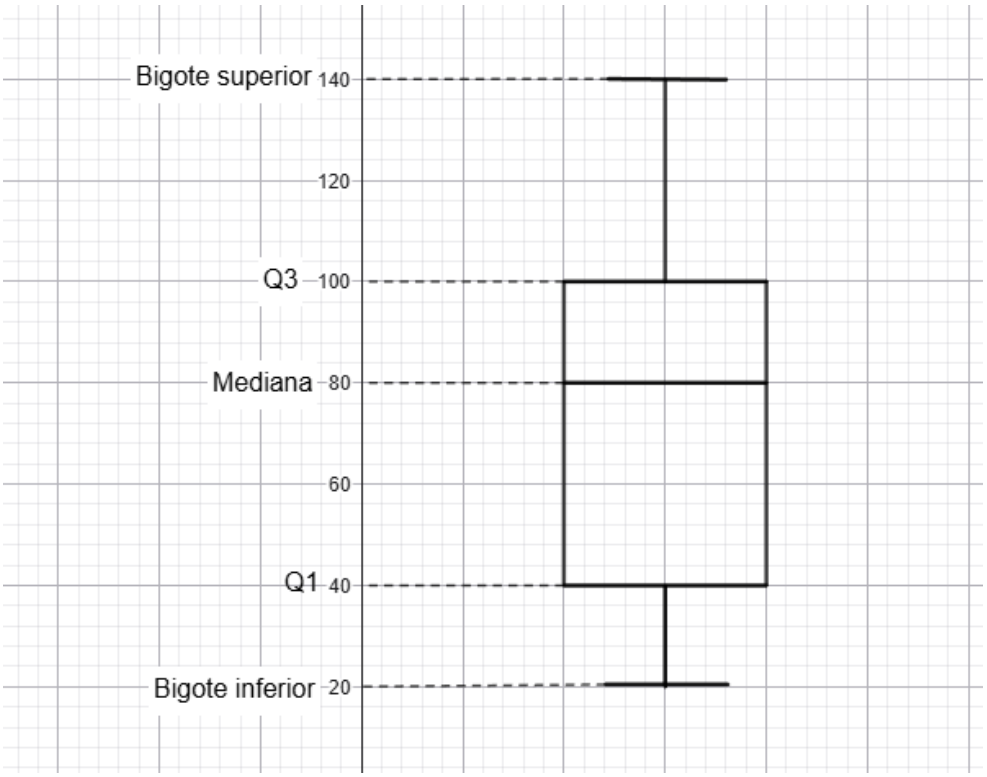
Por encima y por debajo del triángulo se pueden ver los bigotes superior e inferior. Los bigotes señalan las siguientes marcas:

- Bigote superior

$$Mínimo\ \{x_{máx}, Q_3 + 1{,}5 \cdot RI\}$$

- Bigote inferior

$$Máximo\ \{x_{mín}, Q_1 - 1{,}5 \cdot RI\}$$



Los valores de la distribución que estén por encima del bigote superior o por debajo del bigote inferior se consideran valores atípicos.

6. Medidas de posición central

6.1. Media aritmética

6.1.1. Datos no agrupados en frecuencias

En la mayoría de los casos cuando hablamos de tendencias centrales lo primero que se viene a la cabeza es la media aritmética que todas las personas conocen como “la suma de todos los valores dividido entre el número total de valores”. Esto matemáticamente se podría describir como:

$$\bar{x} = \frac{\sum x_i}{N}$$

6.1.2. Datos agrupados en frecuencias

Cuando los datos se dan agrupados en una tabla de frecuencias, se usa la siguiente fórmula:

$\bar{x} = \frac{\sum x_i \cdot n_i}{N} = \sum x_i \cdot f_i$	Donde: • $\bar{x}$, media aritmética. • n_i, frecuencia absoluta de cada valor • N, número total de valores. • f_i, frecuencia relativa de cada valor. • x_i, cada uno de los valores concretos o cada una de las marcas de clase para valores agrupados en intervalos

Ejemplo: las edades de una clase de universidad formada por 10 personas son:

- 5 personas tienen 18 años.
- 3 personas tienen 20 años.
- 2 personas tienen 23 años.

La edad media sería:

$$\bar{x} = \frac{\sum x_i}{N} = \frac{18 + 18 + 18 + 18 + 18 + 20 + 20 + 20 + 23 + 23}{10} = 19{,}6\ años$$

También se podría calcular realizando previamente la tabla de frecuencia absolutas del siguiente modo:

X_i	n_i	N_i
18	5	5
20	3	8
23	2	10

$$\bar{x} = \frac{\sum x_i \cdot f_i}{N} = \frac{18 \cdot 5 + 20 \cdot 3 + 23 \cdot 2}{10} = 19{,}6$$

6.2. Mediana, Me

La mediana es el valor x_i que divide al conjunto de los datos ordenados de manera que al menos el 50% de los elementos es menor o igual que la mediana. También se puede definir como el valor x_i que divide al conjunto de los datos ordenados de manera que al menos el 50% de los elementos es mayor o igual que la mediana

6.2.1. Determinación del valor de la mediana

Con frecuencias absolutas acumuladas

1. Hacer la tabla de frecuencias con al menos X_i y N_i.
2. Dividir $\frac{N_{máx}}{2}$
3. Si el resultado anterior es justo un valor N_i:
$$Me = \frac{X_i + X_{i+1}}{2}$$
4. En caso contrario, la mediana es el primer valor que sobrepase al resultado de la división del paso 2.

Con porcentajes acumulados

1. Hacer la tabla de frecuencias con al menos X_i y P_i.
2. Si hay un valor X_i con P_i=50 %:
$$Me = \frac{X_i + X_{i+1}}{2}$$
3. En caso contrario, la mediana es el primer valor X_i cuya P_i sea mayor que 50 %

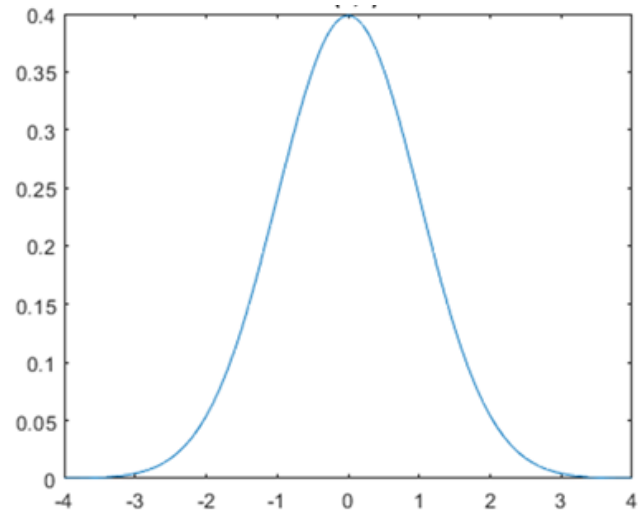
Nota: se podría hacer también con F_i

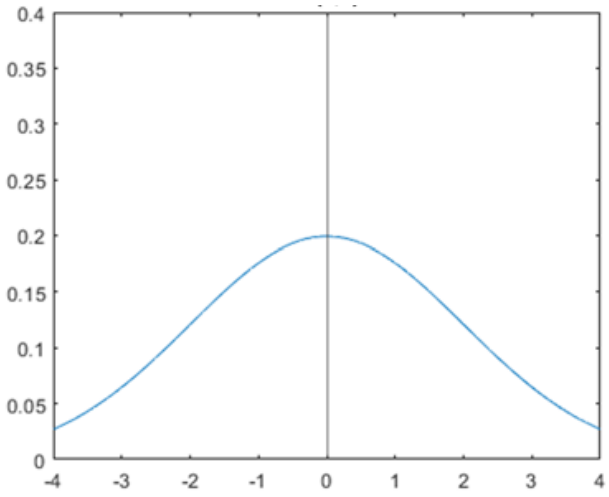
6.3. Moda, Mo

Es el valor que más veces se repite, es decir el que mayor frecuencia tenga. Si hay más de un valor con la misma frecuencia y esta es máxima, entonces, entonces todos esos valores son modas; la distribución tiene más de una moda.

7. Medidas de dispersión

Las medidas de tendencia central por si solas pueden no ser suficientes para caracterizar una distribución. Como se ve en las imágenes, ambas distribuciones tienen la misma tendencia central, que es 0; sin embargo, visualmente es fácil darse cuenta de que las distribuciones son diferentes, que la de la derecha está más dispersa respecto a la tendencia central lo que puede hacer que una distribución no sea válida





Las medidas de dispersión sirven para cuantificar esa dispersión respecto a las tendencias centrales.

9.1. Medidas de dispersión absoluta

Cuantifican la dispersión que hay entre los valores y la tendencia central en las mismas unidades en las que está dada la tendencia central.

9.1.1. Rango

El rango es la distancia que hay entre el valor máximo y el valor mínimo.

9.1.2. Percentiles y cuartiles

Antes de comenzar con las medidas de dispersión es necesario conocer estos dos términos.

Lo habitual es dividir la distribución en 100 trozos y se consideran 100 percentiles uno, por cada trozo, de manera que los percentiles son todos números enteros.

Los percentiles se asocian con porcentajes. Se van a ver como ejemplo los percentiles 25, 50 y 75.

- El percentil 25, $P_{25,}$ es el valor que divide al conjunto de datos ordenados de manera que al menos el 25 % de los elementos es menor o igual que P_{25}.
- El percentil 50, $P_{50,}$ es el valor que divide al conjunto de datos ordenados de manera que al menos el 50 % de los elementos es menor o igual que P_{25}. Es lo mismo que la mediana
- El percentil 75, $P_{75,}$ es el valor que divide al conjunto de datos ordenados de manera que al menos el 75 % de los elementos es menor o igual que P_{25}.

Los cuartiles son algunos percentiles concretos que reciben nombres propios:

- Cuartil 1, Q_1, es el percentil 25.
- Cuartil 2, Q_2, es el percentil 50, es decir, la mediana.
- Cuartil 3, Q_3, es el percentil 75.

Para buscar un percentil cualquier P_k seguimos el mismo método que el empleado para la mediana, pero cambiando el 50 % por el percentil de que se trate.

9.1.3. Rango intercuartílico

Es la diferencia entre cuartil 3 y el cuartil 1:

$$Ri = Q_3 - Q_1$$

9.1.4. Varianza, S^2

La varianza es una desviación cuadrática media que usa a la media aritmética como medida central

Para datos no agrupados: $S_X^2 = \frac{\sum(x_i - \bar{x})^2 \cdot n_i}{N} = \overline{(x_i)^2} - (\bar{x})^2$	Donde: • S_X^2, es la varianza. • x_i, es cada uno de los valores o de las marcas de clase. • $\bar{x}$, es la media artimética. • N, es el número total de valores. • n_i, es la frecuencia relativa de cada valor. • $\overline{(x_i)^2}$, es la media de los valores x_i elevado al cuadrado $\overline{(x_i)^2} = \frac{\sum {x_i}^2 \cdot n_i}{N}$

Las unidades de la varianza son las de los valores de la muestra elevados al cuadrado.

9.1.1. Cuasi-varianza, Sc^2

Para datos no agrupados: $Sc_X^2 = \frac{\sum(x_i - \bar{x})^2 \cdot n_i}{N - 1}$	Donde: • Sc_X^2, es la cuasivarianza. • x_i, es cada uno de los valores o de las marcas de clase. • $\bar{x}$, es la media artimética. • N, es el número total de valores. • n_i, es la frecuencia relativa de cada valor.

Las unidades de la cuasi-varianza tambiénson las de los valores de la muestra elevados al cuadrado.

9.1.2. Desviación típica, S

La necesidad de la desviación típica surge de dar una idea de dispersión con la media, pero en las mismas unidades en las que se dan los valores, en lugar de estar elevadas al cuadrado como ocurre con la varianza o la cuasi varianza. Por ello la fórmula de la desviación típica es:

$S_X = \left\lvert \sqrt{S_X^2} \right\rvert = \left\lvert \sqrt{\frac{\sum(x_i - \bar{x})^2 \cdot n_i}{N}} \right\rvert$	Donde: • S_X^2, es la varianza. • x_i, es cada uno de los valores. • $\bar{x}$, es la media artimética. • N, es el número total de valores. • n_i, es la frecuencia relativa de cada valor.

La desviación típica también se podría hacer con la cuasi-varianza.

9.2. Medidas de dispersión relativas

9.2.1. Coeficiente de variación de Pearson, CV

La desviación típica representa la variación con respecto a la media. Pero esta medida presenta un inconveniente: sus resultados se centran en la media, por lo que puede dar valores que no sean útiles para comparar 2 distribuciones con medias diferentes. Por ello se hace necesario dividir entre la media y expresarlo en % o ‰ (tanto por 1) para así homogeneizar esta medida de dispersión y poder comparar 2 muestras diferentes.

$$CV(X) = \frac{S}{\bar{x}} \quad o\ bien, \quad CV(X) = \frac{S}{\bar{x}} \cdot 100$$

- CV(X) grande → la media es poco representativa. Los datos no son homogéneos.
- CV(X) pequeña → la media representativa. Los datos son homogéneos.

10. Efectos de aumentos lineales y proporcionales

10.1. Aumento lineal

Un aumento lineal es cuando se obtiene una nueva distribución a partir de multiplicar y sumar un número a todos los elementos de otra distribución.

$$X' = a \cdot X + b$$

10.1.1. Consecuencias de un aumento lineal

$\boldsymbol{X' = a \cdot X + b}$	
Media	$\overline{X'} = a \cdot \overline{X} + b$
Mediana	$Me' = a \cdot Me + b$
Moda	$Mo' = a \cdot Mo + b$
Percentiles	$P' = a \cdot P + b$
Varianza	$S_X{}'^2 = a^2 \cdot S_X{}^2$
Desviación típica	$S_X' = \lvert a \rvert \cdot S_X$