AWAKELAB

BASECAMP

Ciencia de Datos

Análisis Exploratorio y Programación Estadística

Objetivo de la jornada

- Utilizar los conceptos básicos de estadística descriptiva para la caracterización de un conjunto de datos de una población.

Estadística descriptiva

La **estadística** descriptiva es una disciplina que se encarga de recoger, almacenar, ordenar, realizar tablas o gráficos y calcular parámetros básicos sobre el conjunto de datos.

La estadística descriptiva es, junto con la **inferencia estadística** o estadística inferencial, una de las dos grandes ramas de la estadística. Su propio nombre lo indica, trata de describir algo. Pero no describirlo de cualquiera forma, sino de manera cuantitativa. Pensemos en el peso de una caja de verduras, en la altura de una persona o en la cantidad de dinero que gana una empresa. De estas variables podríamos decir muchas cosas. Por ejemplo, podríamos indicar que esta o aquella caja de tomates pesan mucho o pesan menos que otras. Siguiendo con otro ejemplo, podríamos decir que el ingreso de una empresa varía mucho a lo largo del tiempo o que una persona tiene una altura promedio.

Para dictar las afirmaciones anteriores, sobre mucho, poco, alto, bajo, muy variable o poco variable necesitamos variables de medidas. Esto es, necesitamos cuantificarlas, ofrecer un número. Con esto en mente, podríamos utilizar los gramos o los kilogramos como unidad de medida para saber el peso de tantas cajas de tomates como consideremos. Una vez pesemos treinta cajas, sabremos cuales pesan más, cuales pesan menos, que cuantía es la que más se repite o si existe mucha disparidad entre los pesos de las diferentes cajas.

Con esta idea nace la estadística descriptiva, con la de recoger datos, almacenarlos, realizar tablas o incluso gráficos que nos ofrezcan

información sobre un determinado asunto. Adicionalmente, nos ofrecen medidas que resumen la información de una gran cantidad de datos.

Un ejemplo de estadística descriptiva sería cuando queremos calcular la media de goles por partido de un futbolista. Se trata de estadística descriptiva, ya que tratamos de describir una variable (número de goles). En este caso, mediante el cálculo de una métrica.

Así pues, decir que Ronaldo metió 1,05 goles por partido durante los últimos 30 partidos, es una frase propia de estadística descriptiva.

También podríamos decir, por ejemplo, que el 30% de los compañeros de clase de Juan tienen los ojos azules, el 60% castaños y el 10% restante negros. Se trataría de una variable cualitativa (color de ojos), pero estamos describiendo la frecuencia con la que aparece.

Tipos de Variables estadísticas

Una variable estadística es una característica de una muestra o población de datos que puede adoptar diferentes valores.

Cuando hablamos de variable estadística estamos hablando de una cualidad que, generalmente adopta forma numérica. Por ejemplo, la altura de Juan es de 180 centímetros. La variable estadística es la altura y está medida en centímetros. También podríamos, por ejemplo, decir que el beneficio de una empresa ha sido de 22.300 dólares el último año. En este caso, la variable sería el beneficio y estaría medido en dólares. Ambas variables son del tipo cuantitativo (se expresan con un número).

Claro que no todas las variables estadísticas son iguales y, por supuesto, no todas se pueden expresar, en principio, en forma de número. Así, otra variable que podríamos encontrarnos es el color de ojos de una persona. Por ejemplo, Juan tiene los ojos verdes y Andrés los tiene azules. La variable sería el color de ojos y sería una variable cualitativa. Es decir, no se expresa con número.

Aunque hay decenas de tipos de variables estadísticas, por norma general podemos encontrarnos dos tipos de variables:

- **Variable cuantitativa:** Son variables que se expresan numéricamente.
    - *Variable continua:* Toman un valor infinito de valores entre un intervalo de datos. El tiempo que tarda un corredor en completar los 100 metros lisos.
    - *Variable discreta*: Toman un valor finito de valores entre un intervalo de datos. Número de helados vendidos.
- **Variable cualitativa:** Son variables que se expresan, por norma general, en palabras.
    - *Variable ordinal*: Expresa diferentes niveles y orden. Por ejemplo, primero, segundo, tercero, etc.
    - *Variable nominal*: Expresa un nombre claramente diferenciado. Por ejemplo el color de ojos puede ser azul, negro, castaño, verde, etc.

Además cada una de estas variables podría tener más subtipos, ya que tenemos variables de tipo económico, categóricas, dicotómicas, dependientes, independientes. Es decir, como ya hemos dicho, hay muchos tipos de variables estadísticas. Por ejemplo, podríamos tener una variable estadística de tipo cuantitativo, discreta y dependiente.

Adicionalmente, también debemos aclarar que el hecho que las variables cualitativas se expresan con nombre no quiere decir que no puedan ser parte de un modelo matemático. Así pues, podríamos crear una variable cuantitativa a partir de una variable cualitativa. Por ejemplo, para el color de ojos podríamos asignar un 1 si tiene los ojos azules, un 2 si tiene los ojos verdes y un 3 si tiene los ojos marrones. O, en otros casos, podríamos también convertir variables dicotómicas que indica SI o NO, en 1 o 0.

Matrices de Datos y tablas de frecuencia.

Una cosa es escribir un diario en una libreta y otra cosa es recopilar datos que sirvan para un posterior análisis. Esos datos se pueden recoger de muchas maneras y las que normalmente son buenas para el campo (fichas para cada salida de campo u objeto de estudio) o quedan bonitas en un documento (con celdas unidas, por ejemplo) no son buenas para el posterior análisis de los datos.

Si ya están metidos en el computador, en el formato que sea: csv, excel, …, la mejor manera es como están. Nunca (o casi nunca) es buena idea retocar los datos originales ya que este proceso puede (y suele) producir errores que no son subsanables posteriormente.

Es decir, si los datos ya están en algún formato digital, lo mejor es que las modificaciones se hagan ya directamente en el script de importación de datos, de forma que esas modificaciones entren dentro de lo que se llama «Ciencia reproducible» (Rodríguez-Sánchez et al. 2016).

La matriz de datos

La matriz de datos ideal debe ser:

- Única (no siempre es posible, pero debe intentarse).
- Consistente
- Rectangular (todas las casillas deberían estar llenas)
- Las columnas son Variables.
- Las filas son observaciones.
- No debe tener variables obtenidas a partir de cálculos con otras variables de la matriz. Si no puedes evitar la tentación, haz los cálculos en otra hoja de cálculo diferente. Recuerda poner cuáles y cómo las has calculado en la leyenda de los datos.
- No debe tener fórmulas, ni cálculos, ni pre-análisis, en la matriz de datos.
- No debe haber información en los formatos. Los colores o negritas de las filas o columnas no se exportan.
- Formato csv, o en su defecto xlsx u ods.

Una cosa importante, NUNCA juntes celdas, ni de la misma fila, ni de la misma columna, en una tabla de datos.

Tablas de frecuencia

Las tablas de frecuencias son cuadros en los que se registran los datos estadísticos en forma organizada con la frecuencia de cada uno de los valores que puede tomar la variable estudiada. Se presentan en columnas y filas con la finalidad de analizar, sintetizar e ilustrar la información producida por los datos recopilados de una investigación o estudio determinado

Tipos de tablas de Frecuencia.

Las frecuencias son el número de veces que se repite un valor determinado de una variable. Por ejemplo; si se mide la edad de 6 personas y se obtiene 22, 21, 30, 24, 22 y 30 se tiene que la frecuencia de 22 es 2, es decir, dos personas de veintidós años. 21 tendrá frecuencia 1, 24 frecuencia 1 y 30 frecuencia 2.

Los tipos de frecuencia pueden ser, Frecuencia Absoluta, Frecuencia Acumulada, Frecuencia Relativa y Frecuencia Relativa Acumulada, el siguiente cuadro explica cada una de ellas.

Cuadro Nº 1

Definición de Frecuencias Estadísticas

	TIPO	**CARACTERÍSTICAS**
Tablas	**Frecuencia Absoluta**	Se define como el número de observaciones que cumple una característica determinada y se denotará como fi
	Frecuencia Acumulada	Es la suma o acumulación de las frecuencias Absoluta y se denotará por Fa; así $F_1 = f_1, F_2 = f_1 + f_2, F_3 = f_1 + f_2 + f_3$, etc
	Frecuencia Relativa	Es la proporción que representa fi en el total de datos observados, se denotará por ahí y se calcula de la siguiente manera: $hi = (fi/N)$donde Nes el total de observaciones realizadas.
	Frecuencia relativa acumulada	Es la acumulación o suma de las frecuencias relativas; $H_1 = h_1, H_2 = h_1 + h_2, H_3 = h_1 + h_2 + h_3$, etc.

Cuadro Nº 2

Componentes de una Distribución de Frecuencias

Variable X_1	f_1	F_a	h_i	H_a
X_1	f_1	f_1	h_1	h_1
X_2	f_2	$f_1 + f_2$	h_2	$h_1 + h_2$
X_3	f_3	$f_1 + f_2 + f_3$	h_3	$h_1 + h_2 + h_3$
- - -				
X_n	f_n	$(f_1 + f_2 + f_3 + \ldots + f_n) = N$	h_n	$(h_1 + h_2 + h_3 + \ldots + h_n) = N$
	$\sum f_i = N$		$\sum h_i = N$	

Ejemplo 1

Se visitaron varias viviendas de una comunidad ubicada al sur de Maracay, en las cercanías del lago de Valencia y se observó la presencia de algunas enfermedades en su población infantil, producto del deterioro ambiental. Los resultados de la observación fueron los siguientes:

Respiratoria	Eruptiva	Respiratoria	Infección	Respiratoria	Eruptiva
Dengue	Diarrea	Respiratoria	Eruptiva	Eruptiva	Diarrea
Dengue	Respiratoria	Respiratoria	Respiratoria	Eruptiva	Respiratoria
Eruptiva	Eruptiva	Diarrea	Infección	Eruptiva	Diarrea
Dengue	Respiratoria	Diarrea	Eruptiva	Respiratoria	Respiratoria

Tipo de Enfermedad

X_i	f_i	F_a	h_i	H_a
Respiratoria	11	11	$11/30 = 0{,}36666$	$11/30 = 0{,}36666$
Eruptiva	9	20	$9/30 = 0{,}3$	$20/30 = 0{,}6666$
Infección	2	22	$2/30 = 0{,}0666$	$22/30 = 0{,}7333$
Diarrea	5	27	$5/30 = 0{,}1666$	$27/30 = 0{,}90$
Dengue	3	30	$3/30 = 0{,}1$	$30/30 = 1{,}00$

$$\sum f_i = 30 \qquad \sum h_i = 0{,}9998 \approx 1{,}00$$

De los datos ordenados en la tabla de frecuencias se puede obtener la siguiente información:

- Se observó una muestra de 30 niños de la comunidad.
- De ellos 11 presentan enfermedades Respiratorias, 9 Eruptivas, 2 algún tipo de Infección, 5 con cuadro Diarreico y 3 casos de Dengue.
- La mayoría de los niños enfermos se deben a enfermedades respiratoria o erupciones en la piel con 36,66 % y 30,00% respectivamente
- Entre las enfermedades Respiratorias y las Eruptivas acumulan 20 de los 30 casos observados.

Construcción de Tablas de Frecuencias para Datos Cuantitativos Discretos

Se llaman Datos No Agrupados porque no han sido categorizados en grupos, se presentan tal como se recogen del campo. Una vez agrupados, estos se convierten en Datos Agrupados.

Ejemplo 2

En vista de la alta incidencia de enfermedades en los niños de la comunidad se está diseñando un plan de vacunación para infantes con edad comprendidas entre 0 y 5 años, se pasó casa por casa y se preguntó la edad de los niños que habitan en la vivienda obteniéndose los siguientes resultados:

1	1	1	0	1	0
1	2	0	1	1	2
0	3	2	1	2	3
1	0	2	4	0	1
5	0	1	4	5	3
1	1	1	3	1	0
0	0	2	0	0	2
2	2	1	0	5	1
5	2	1	3	1	0
4	2	1	1	0	5

Edad de los niños

X_i	f_i	F_a	h_i	H_a
0	15	15	$15/60 = 0{,}25$	$15/60 = 0{,}25$
1	21	36	$21/60 = 0{,}35$	$36/60 = 0{,}60$
2	11	47	$11/60 = 0{,}1833$	$47/60 = 0{,}7833$
3	5	52	$5/60 = 0{,}0833$	$52/60 = 0{,}8666$
4	3	55	$3/60 = 0{,}05$	$55/60 = 0{,}9166$
5	5	60	$5/60 = 0{,}0833$	$60/60 = 1$

$\sum f_i = 60$ $\quad$ $\sum h_i = 0{,}9999 \approx 1{,}00$

Se encontraron 15 niños de cero años, 21 de un año, 11 de dos años, 5 de tres años, 3 de cuatro años y 5 de cinco años.

- De ellos 47 tienen entre cero y dos años
- Los niños de dos años representan 18,33% de los observados
- Los niños entre 0 años y 2 años representan 78,33% de los niños observados.
- Se observó un total de 60 niños en la comunidad.

Construcción de Tablas de Frecuencias para Datos Cuantitativos Continuos

Cuando los datos tienen una variación muy alta entre el valor más pequeño de la serie y el valor más alto, la variable puede tomar muchos valores distintos, como ocurriría por ejemplo si se quiere saber la edad de todos los habitantes de la comunidad estudiada en los ejemplos anteriores, de allí que el valor menor es cero (0) y el valor mayor podría inclusive ser mayor que cien (100) y no sería muy práctico diseñar una tabla de frecuencia como la que se construyó para los niños.

Cuando esto ocurre o la variable que se mide se considera como continua se agrupan los datos en intervalos de frecuencias y a las tablas así construidas se les dice tablas de frecuencias para Datos Agrupados y la metodología de construcción de los intervalos es lo único diferente a las tablas de frecuencias para Datos No Agrupados

Construcción de una Tabla de Frecuencias para Datos Agrupados

A continuación se presentan los pasos para la construcción de una tabla de frecuencias para datos agrupados, para posteriormente ilustrar su construcción con un ejemplo.

1. Primero se elige el número de intervalos que se desea tener en su tabla, se aconseja que sea como mínimo de 4 intervalos y como máximo 7, aunque hay autores que consideran hasta 14 intervalos.

   A ese número de intervalos lo llamaremos k, es decir k es el número de intervalos que se van a tener en la tabla y un método más efectivo para obtenerlo es de acuerdo a la fórmula de Stuges

$$k = 1 + 3{,}322 log 10(n)$$

2. Una vez escogido el valor k de intervalos se determina la longitud Ic que debe tener cada intervalo, todos los intervalos serán de igual longitud. Y esta longitud se obtendrá utilizando la fórmula.

$$Ic = \frac{Máximo(X_i) - Máximo(X_i)}{k}$$

   El número resultante de restar Máximo menos Mínimo se llama longitud total o rango de la variable y se denota como Rang(X) = Máx(X) – Min(X)

3. El primer intervalo debe contener al dato más pequeño que se observó y de allí debe comenzar ese intervalo de clases y el último debe contener al valor máximo de los datos recogidos. Se debe procurar que todos los intervalos tengan la misma longitud.

4. Se comienza a construir la tabla, iniciando el primer intervalo en el valor mínimo de la variable, aumentando el valor Ic para fijar su límite superior.

5. Se colocan las frecuencias respectivas como se hizo en los ejemplos estudiados anteriormente.


AWAKELAB

Ejemplo 3

Construcción de una Tabla de Frecuencias para Datos Agrupados

Una micro empresa está interesada en participar en la licitación para la fabricación de uniformes (bragas mecánicas) para el personal de mantenimiento de la Fuerza Aérea Venezolana (FAV) y para conocer las tallas decide medir la estatura de una muestra de 50 trabajadores obteniendo los datos que se presentan a continuación:

1,75	1,78	1,70	1,69	1,78	1,80
1,66	1,66	1,72	1,84	1,72	1,75
1,73	1,66	1,66	1,70	1,75	1,69
1,68	1,72	1,74	1,79	1,80	1,81
1,75	1,74	1,82	1,74	1,76	1,66

Se debe construir la tabla de frecuencias para datos agrupados porque los datos son continuos.

1. Se determina el número de intervalos más conveniente mediante la fórmula de Stuges.

   k = 1 + 3,322log(N) como son 30 datos entonces N = 30

   k = 1 + 3,322log(30) = 1 + 3,322x1,4771 = 1 + 4,9069 = 5,9069 ≅ 6 intervalos.

   Se hallarán el valor mínimo y máximo de los datos

   Mín(X) = 1,66 y Máx(X) = 1,854

   Rang(X) = 1,84 – 1,66 = 0,20

2. Ic = 0,18/6 = 0,03
3. Se comienza a construir la tabla de frecuencia partiendo del valor mínimo de los datos que es 1,66 en este caso

$L_i - L_s$	f_i	F_a	h_i	H_a
1,66 - 1,69	8	8	8/30=0,2666	8/30=0,2666
1,69 - 1,72	5	13	5/30=0,1666	13/30?0,4333
1,72 - 1,75	8	21	8/30=0,2666	21/30=0,70
1,75 - 1,78	3	24	3/30=0,1	24/30=0,80
1,78 - 1,81	4	28	4/30=0,1333	28/30=0,9333
1,81 - 1,84	2	30	2/30 = 0,0666	30/30=1,0000
	30			

- Ocho personas observadas miden entre 1,66 y 1,69 m
- Las personas que miden entre 1,66 y 1,75 m representan 70% del total de observadas
- Los dos grupos de estatura más frecuentes son de 1,66 a 169 y de 1,72 a 1,75 con frecuencia de 8 cada una
- Esos dos intervalos de clase representan 52,23% del total de datos, lo que viene a ser un poco más de la mitad de los observados.

Gráficos y formas de distribución.

Aunque las tablas y los gráficos no son algo exclusivo de la estadística descriptiva, sí que la caracterizan. En informes, en estudios e investigaciones es muy frecuente el uso de gráficos. Nos ayudan a mostrar la información de manera más sencilla y acotada.

Eso sí, dentro de las tablas y gráficos existe una cantidad de tipos inmenso. A continuación se definirán algunos.

Histograma

Un histograma es una representación gráfica de una variable en forma de barras.

Se utilizan para variables continuas o para variables discretas, con un gran número de datos, y que se han agrupado en clases.

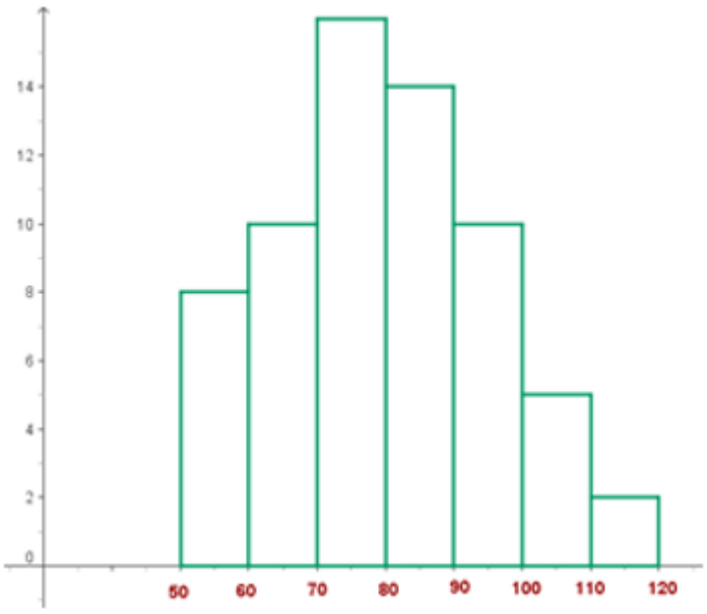
En el eje abscisas se construyen unos rectángulos que tienen por base la amplitud del intervalo, y por altura, la frecuencia absoluta de cada intervalo.

La superficie de cada barra es proporcional a la frecuencia de los valores representados.

Ejemplo:

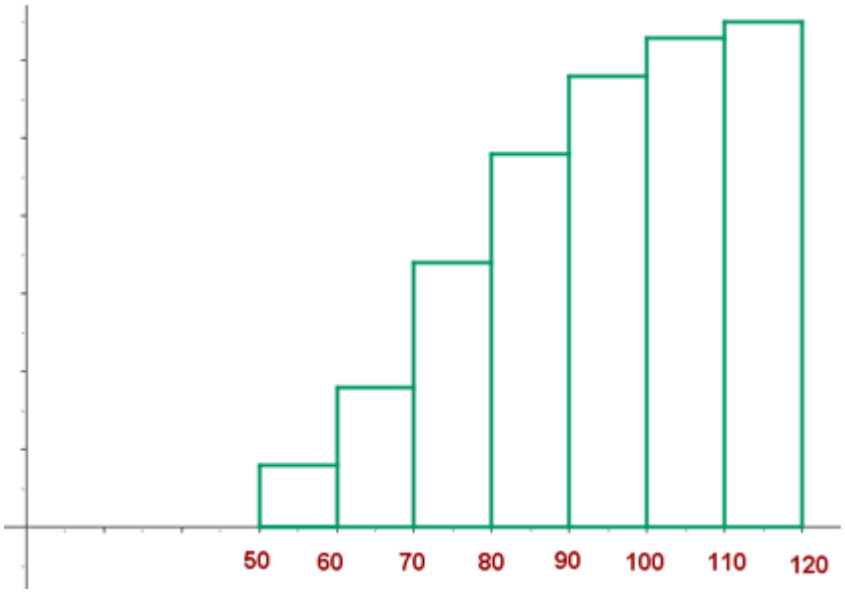
El peso de 65 personas adultas viene dado por la siguiente tabla:

	C_i	f_i	F_i
[50,60]	55	8	8
[60,70]	65	10	18
[70,80]	75	16	34
[80,90]	85	14	48
[90,100]	95	10	58
[100,110]	110	5	63
[110,120]	115	2	65
		65	



Histograma de frecuencias acumuladas

Si se representan las frecuencias acumuladas de una tabla de datos agrupados se obtiene el histograma de frecuencias acumuladas.



Histogramas con intervalos de amplitud diferente

Para construir un histograma con intervalo de amplitud diferente tenemos que calcular las alturas de los rectángulos del histograma.

$$h_i = \frac{f_i}{a_i}$$

$\mathbf{h_i}$ es la altura del intervalo

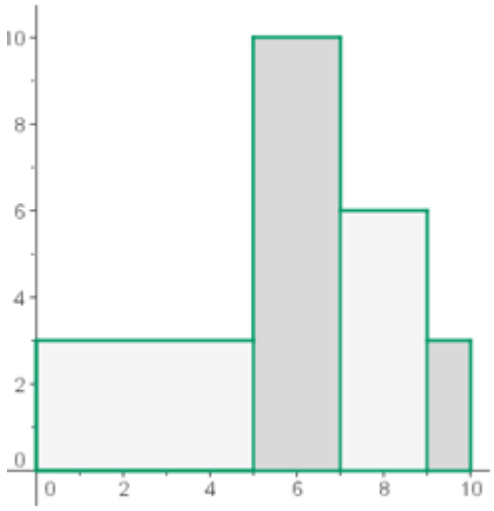
$\mathbf{f_i}$ es la frecuencia del intervalo

$\mathbf{a_i}$ es la amplitud del intervalo

Ejemplo:

En la siguiente tabla se muestran las calificaciones (suspenso, aprobado, notable y sobresaliente) obtenidas por un grupo de 50 alumnos.

	f_i	h_i
[0, 5)	15	3
[5, 7)	20	10
[7, 9)	12	6
[9, 10)	3	3
	50	



Barras

Un diagrama de barras se utiliza para presentar datos cualitativos o datos cuantitativos de tipo discreto.

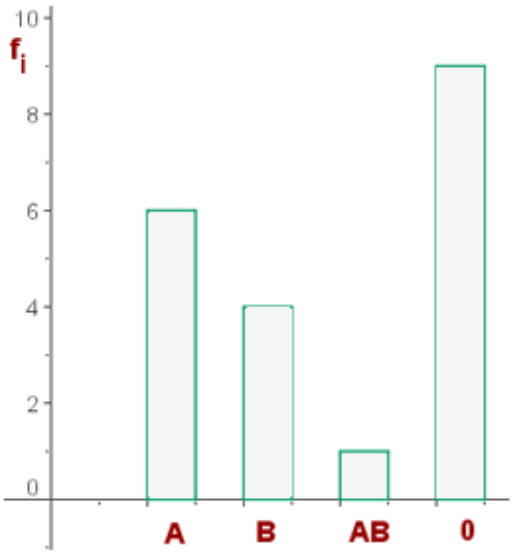
Se representan sobre unos ejes de coordenadas, en el eje de abscisas se colocan los valores de la variable, y sobre el eje de ordenadas las frecuencias absolutas o relativas o acumuladas.

Los datos se representan mediante barras de una altura proporcional a la frecuencia.

Ejemplo

Un estudio hecho al conjunto de los 20 alumnos de una clase para determinar su grupo sanguíneo ha dado el siguiente resultado:

Grupo sanguíneo	f_i
A	6
B	4
AB	1
0	9
	20



Torta

Un diagrama de sectores se puede utilizar para todo tipo de variables, pero se usa frecuentemente para las variables cualitativas.

Los datos se representan en un círculo, de modo que el ángulo de cada sector es proporcional a la frecuencia absoluta correspondiente.

$$\alpha = \frac{360°}{N} \cdot f_i$$

El diagrama circular se construye con la ayuda de un transportador de ángulos.

Ejemplo:

En una clase de 30 alumnos, 12 juegan baloncesto, 3 practican natación, 4 juegan fútbol y el resto no practica ningún deporte.

$$\alpha_1 = \frac{360°}{30} \cdot 12 = 144° \qquad \alpha_2 = \frac{360°}{30} \cdot 3 = 36°$$

$$\alpha_3 = \frac{360°}{30} \cdot 9 = 108° \qquad \alpha_4 = \frac{360°}{30} \cdot 6 = 72°$$

	Alumnos	Ángulo
Baloncesto	12	144°
Natación	3	36°
Fútbol	9	108°
Sin deporte	6	72°
Total	30	360°

imagen:fuente propia

Medidas de Tendencia central: Moda, Media, Mediana

Las medidas de tendencia central son parámetros estadísticos que informan sobre el centro de la distribución de la muestra o población estadística.

A veces, tratamos con una gran cantidad de información, como variables que presentan muchos datos o son muy dispares. Datos con muchos decimales, de diferente signo o longitud. En estos casos, siempre es preferible calcular medidas que nos ofrezcan información resumida sobre dicha variable. Por ejemplo, medidas que nos indiquen cuál es el valor que más se repite.

Sin perjuicio de lo anterior, no hay que irse tan lejos. Si miramos la siguiente tabla que muestra el salario que cobra cada uno de los trabajadores de una empresa que fabrica cajas de cartón, tendremos lo siguiente:

Trabajador	Salario
1	€ 1.235
2	€ 1.002
3	€ 859
4	€ 486
5	€ 1.536
6	€ 1.248
7	€ 1.621
8	€ 978
9	€ 1.125
10	€ 768

Alguien podría preguntarse, ¿cuánto gana el trabajador promedio de esta empresa? En ese caso las medidas de tendencia central nos podrían ayudar. Concretamente, la media. Sin embargo, a priori, lo único que sabemos es que el número estará entre el mínimo y el máximo.

Entre las medidas de tendencia central podemos encontrarnos con las siguientes:

Media

La media es el valor promedio de un conjunto de datos numéricos, calculada como la suma del conjunto de valores dividida entre el número total de valores. A continuación se muestra la fórmula de la media aritmética:

$$Media\ aritmética = \frac{\sum_1^N x_i}{N} = \frac{x_1 + x_2 + x_3 \ldots . + x_n}{N}$$

Formula media aritmetica

Donde x es el valor de la observación i, y N el número total de observaciones.

Ejemplo:

Supongamos que nuestras calificaciones en la escuela son:

Asignatura	Nota
Matemáticas	*7*
Educación Física	*8*
Biología	*5*
Economía	*10*

N = número total de asignaturas = 4

Entonces aplicando la fórmula que acabamos de exponer, el resultado sería:

$$Media = \frac{Nota\ de\ Mates + Nota\ de\ E.F. + Nota\ de\ Biología + Nota\ de\ Economía}{4}$$

$$Media\ aritmética = \frac{7+8+5+10}{4} = 7{,}5$$

Nuestra nota media será de un 7,5.

Mediana.

La mediana es un estadístico de posición central que parte la distribución en dos, es decir, deja la misma cantidad de valores a un lado que a otro. Las fórmulas propuestas no nos darán el valor de la mediana, lo que nos darán será la posición en la que está dentro del conjunto de datos. Las fórmulas que indica la posición de la mediana en la serie son las siguientes:

- Cuando el número de observaciones es par:

Mediana = (n+1) / 2 → Media de las posiciones observaciones

- Cuando el número de observaciones es impar:

Mediana = (n+1) / 2 → Valor de la observación

Ejemplo:

Imaginemos que tenemos los siguientes datos:

2,4,12,6,8,14,16,10,18.

En primer lugar los ordenamos de menor a mayor con lo que tendríamos lo siguiente:

2,4,6,8,**10**,12,14,16,18.

Pues bien, el valor de la mediana, como indica la fórmula, es aquel que deje la misma cantidad de valores tanto a un lado como a otro. ¿Cuántas observaciones tenemos? 9 observaciones. Calculamos la posición con la fórmula de la mediana correspondiente.

Mediana = 9+1 / 2 = 5

¿Qué quiere decir este 5? Nos dice que el valor de la mediana, se encuentra en la observación cuya posición es la quinta.

Por lo tanto la mediana de esta sería de datos sería el número 10, ya que está en la quinta posición. Además, podemos comprobar como tanto a la izquierda del 5 hay 4 valores (2, 4, 6 y 8) y a la derecha del 10 hay otros 4 valores (12, 14, 16 y 18).

Moda

La moda es el valor que más se repite en una muestra estadística o población. No tiene fórmula en sí mismo. Lo que habría que realizar es la suma de las repeticiones de cada valor. Por ejemplo, ¿cuál es la moda de la siguiente tabla de salarios?

Trabajador	Salario
1	€ 1.236
2	€ 1.236
3	€ 859
4	€ 486
5	€ 1.536
6	€ 1.536
7	€ 1.621
8	€ 978
9	€ 1.236
10	€ 768

La moda sería 1.236€. Si vemos los salarios de los 10 trabajadores, veríamos que 1.236€ se repite en tres ocasiones

Medidas de Dispersión

Las medidas de dispersión tratan, a través del cálculo de diferentes fórmulas, de arrojar un valor numérico que ofrezca información sobre el grado de variabilidad de una variable.

En otras palabras, las medidas de dispersión son números que indican si una variable se mueve mucho, poco, más o menos que otra. La razón de ser de este tipo de medidas es conocer de manera resumida una característica de la variable estudiada. En este sentido, deben acompañar a las medidas de tendencia central. Juntas, ofrecen información de un sólo vistazo que luego podremos utilizar para comparar y, si fuera preciso, tomar decisiones.

Las medidas de dispersión más conocidas son: el rango, la varianza, la desviación típica y el coeficiente de variación (no confundir con coeficiente de determinación). A continuación veremos estas cuatro medidas.

Rango

El rango es un valor numérico que indica la diferencia entre el valor máximo y el mínimo de una población o muestra estadística. Su fórmula es:

$$R = Máx_x - Mín_x$$

Donde:

- **R** → Es el rango.
- **Máx** → Es el valor máximo de la muestra o población.
- **Mín** → Es el valor mínimo de la muestra o población estadística.
- **x** → Es la variable sobre la que se pretende calcular esta medida.

Ejemplo:

Supongamos que tenemos una empresa que produce microchips para luego venderlos a las principales marcas de computadoras. Esta empresa encarga a un economista que realice un estudio sobre la evolución de las ventas (últimos 4 años) para, posteriormente, ofrecer consejos que mejoren los resultados empresariales. Entre otras muchas métricas, se pide que se calcule el rango de producción de microchips. A continuación se muestra la siguiente tabla de datos:

Mes 1	44.347
Mes 2	12.445
Mes 3	26.880
Mes 4	23.366
Mes 5	42.464
Mes 6	15.480
Mes 7	21.562
Mes 8	11.625
Mes 9	39.496
Mes 10	39.402
Mes 11	47.699
Mes 12	44.315
Mes 13	29.581
Mes 14	44.320
Mes 15	35.264
Mes 16	10.124
Mes 17	43.520
Mes 18	26.360
Mes 19	19.534
Mes 20	30.755
Mes 21	37.327
Mes 22	15.832

Mes 23	33.919
Mes 24	29.498
Mes 25	46.136
Mes 26	18.007
Mes 27	36.339
Mes 28	27.696
Mes 29	47.413
Mes 30	47.636
Mes 31	20.978
Mes 32	49.079
Mes 33	40.668
Mes 34	45.932
Mes 35	40.454
Mes 36	46.132
Mes 37	35.054
Mes 38	11.906
Mes 39	22.532
Mes 40	43.045
Mes 41	45.074
Mes 42	16.505
Mes 43	27.336
Mes 44	37.831

Mes 45	29.757
Mes 46	37.765
Mes 47	22.237
Mes 48	38.601
MÁXIMO	**49.079**
MÍNIMO	**10.124**
RANGO	**38.955**

El mes que más microchips produjo la empresa (MÁXIMO) fue el mes 32 con 49.079 microchips producidos. Por su parte, el momento que menos microchips produjo tuvo lugar en el mes 16 con 10.124 microchips producidos. Por tanto, el rango estadístico que es la diferencia (49.079-10.124) se sitúa en 38.955.

¿Cómo se interpreta esto? Esto quiere decir, que durante los últimos 4 años la variación máxima que ha habido ha sido de 38.955 microchips producidos. Gráficamente podemos verlo del siguiente modo:

imagen: https://economipedia.com/

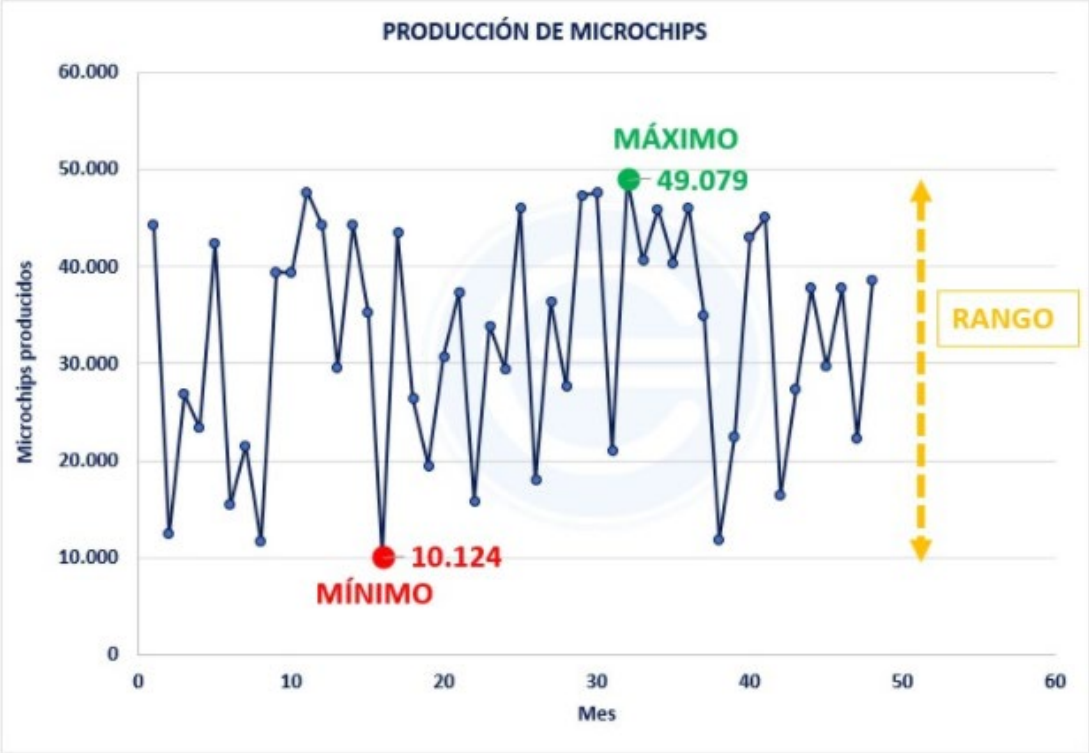


imagen: https://economipedia.com/

El punto verde es el máximo, el punto rojo el mínimo y la línea discontinua amarilla situada a la derecha es la diferencia. Esto es, el rango.

Rango intercuartílico

El rango intercuartílico es una medida de dispersión de un conjunto de datos que expresa la diferencia o la distancia entre el primer y el tercer cuartil.

En otras palabras, el rango intercuartílico es la diferencia entre el penúltimo y el primer cuartil de una distribución utilizada en el diagrama de caja. Generalmente utilizado en el diagrama de caja que utiliza la mediana como medida central.

La forma abreviada de nombrar al rango intercuartílico es RIC o RQ.

El rango intercuartil utiliza la mediana como medida central. Entonces, el resultado del rango intercuartil será próximo a la mediana o segundo cuartil (Q2) si hay pocos valores extremos.

El rango intercuartil está considerado un estadístico robusto por su baja exposición a valores extremos. Esto es debido a que solo se consideran las observaciones entre el tercer cuartil y el primer cuartil. Todas las observaciones fuera de ese rango quedan excluidas del cálculo y, por tanto, solo se tienen en cuenta las observaciones más cercanas a la mediana, es decir, al segundo cuartil.

La presencia de varios valores extremos entre el primer y el tercer cuartil hará aumentar mucho el rango intercuartílico y también la mediana, pero a una tasa menor. Esta situación es poco probable dado que los datos muy extremos tienden a ser poco comunes.

Sabiendo que el rango intercuartil es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1), entonces, simplemente tenemos que hacer la diferencia entre ambos valores.

$$\mathbf{RIC = Q3 - Q1}$$

Para recordar esta medida estadística de forma sencilla y rápida, tenemos que pensar en intercuartílico. Intercuartílico significa entre cuartiles y rango lo entendemos como distancia entre dos puntos. Entonces, podemos entender rango intercuartílico como distancia o diferencia entre dos cuartiles. Estos dos cuartiles son el tercer cuartil (Q3) y el primer cuartil (Q1)


AWAKELAB

Ejemplo:

Suponemos que queremos calcular el rango intercuartílico y la desviación del número de ciclistas que pasan por delante de nuestra casa durante el año.

1. Primero, contamos los ciclistas y recogemos la información en una tabla.

Número de ciclistas

Mes	Ciclistas
Enero	200
Febrero	140
Marzo	200
Abril	300
Mayo	370
Junio	400
Julio	600
Agosto	700
Septiembre	760
Octubre	500
Noviembre	300
Diciembre	200

2. Segundo, calculamos los cuartiles que necesitamos para calcular el rango intercuartil.

   Q3 = 525

   Q1 = 200

   RIC = Q3 – Q1 = 525 – 200 = 325

El rango intercuartílico de este conjunto de datos es 325. Cuanto mayor es el rango intercuartílico, mayor la dispersión entre los datos.

Varianza

La varianza es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media. Formalmente se calcula como la suma de los residuos al cuadrado divididos entre el total de observaciones. Su fórmula es la siguiente:

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$$

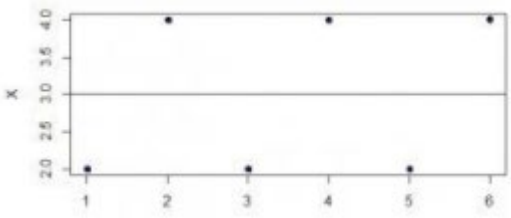
- **X →** Variable sobre la que se pretenden calcular la varianza
- **x_i →** Observación número i de la variable X. i puede tomará valores entre 1 y n.
- **N →** Número de observaciones.
- **$\bar{x}$** → Es la media de la variable X.

O lo que es lo mismo:

$$Var\,(X) = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \cdots + (x_n - \bar{X})^2}{n}$$

Ejemplo:

Si tuviésemos datos sobre los salarios de un conjunto de personas en euros, el dato que arroja la varianza sería en euros cuadrados. Para que tenga sentido la interpretación calcularíamos la desviación típica y pasaríamos el dato a euros.



1. Desviación -> (2-3) = -1
2. Desviación -> (4-3) = 1
3. Desviación -> (2-3) = -1
4. Desviación -> (4-3) = 1
5. Desviación -> (2-3) = -1
6. Desviación -> (4-3) = 1

Si sumamos todas las desviaciones el resultado es cero.

Desviación Estándar o típica.

La **desviación típica** es otra medida que ofrece información de la dispersión respecto a la media. Su cálculo es exactamente el mismo que la varianza, pero realizando la raíz cuadrada de su resultado. Es decir, la desviación típica es la raíz cuadrada de la varianza.

$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{N}}$$

- **X →** Variable sobre la que se pretenden calcular la varianza
- **x_i →** Observación número i de la variable X. i puede tomar valores entre 1 y n.
- **N →** Número de observaciones.
- **$\bar{x}$ →** Es la media de la variable X.

Ejemplo:

Vamos a comprobar como, con cualquiera de las dos fórmulas expuestas, el resultado de la desviación típica o desviación media es el mismo.

Según la fórmula de la **varianza** (raíz cuadrada):

$$a = \sqrt{\frac{\sum_i^N \quad (X_i - \underline{X})^2}{N}}$$

$$= \sqrt{\frac{(2-3)^2+(4-3)^2 + (2-3)^2 + (4-3)^2(2-3)^2 + (4-3)^2}{6}}$$

$$= \sqrt{\frac{1+1+1+1+1+1}{6}} = \sqrt{\frac{6}{6}} = \sqrt{1} = 1$$

Según la fórmula del valor absoluto:

$$\sigma = \frac{\sum_i^N \blacksquare |X_i - \underline{X}|}{N}$$

$$= \frac{|2-3| + |4-3| + |2-3| + |4-3| + |2-3| + |4-3|}{6}$$

$$= \frac{|-1| + |1| + |-1| + |1| + |-1| + |1|}{6}$$

$$= \frac{1 + 1 + 1 + 1 + 1 + 1}{6} = 1$$

Tal como dictaba el cálculo intuitivo. La desviación media es de 1. Pero, ¿no habíamos dicho que la fórmula del valor absoluto y de la desviación típica daban valores diferentes? Así es, pero hay una excepción. El único caso en que la desviación estándar y la desviación respecto de la media ofrecen el mismo resultado es el caso en que todas las desviaciones son igual a 1.

Coeficiente de Variación

El coeficiente de variación, también denominado como coeficiente de variación de Pearson, es una medida estadística que nos informa acerca de la dispersión relativa de un conjunto de datos.

Su cálculo se obtiene de dividir la desviación típica entre el valor absoluto de la **media** del conjunto y por lo general se expresa en porcentaje para su mejor comprensión.

$$CV = \frac{\sigma_x}{|\bar{X}|}$$

- **X** → Variable sobre la que se pretenden calcular la varianza
- σ_x → Desviación típica de la variable X.
- $|\bar{x}|$ → Es la media de la variable X en valor absoluto con $\bar{x} \neq 0$

El coeficiente de variación se utiliza para comparar conjuntos de datos pertenecientes a poblaciones distintas. Si atendemos a su fórmula, vemos que este tiene en cuenta el valor de la media. Por lo tanto, el coeficiente

de variación nos permite tener una medida de dispersión que elimine las posibles distorsiones de las medias de dos o más poblaciones.

Ejemplo:

Pensemos en una población de elefantes y otra de ratones. La población de elefantes tiene un peso medio de 5.000 kilogramos y una desviación típica de 400 kilogramos. La población de ratones tiene un peso medio de 15 gramos y una desviación típica de 5 gramos. Si comparáramos la dispersión de ambas poblaciones mediante la desviación típica podríamos pensar que hay mayor dispersión para la población de elefantes que para la de los ratones.

Sin embargo al calcular el coeficiente de variación para ambas poblaciones, nos daríamos cuenta que es justo al contrario.

Elefantes: 400/5000=0,08
Ratones: 5/15=0,33

Si multiplicamos ambos datos por 100, tenemos que el coeficiente de variación para los elefantes es de apenas un 8%, mientras que el de las ratones es de un 33%. Como consecuencia de la diferencia entre las poblaciones y su peso medio, vemos que la población con mayor dispersión, no es la que tiene una mayor desviación típica.

Estandarización de Datos

La normalización estadística es la transformación de escala de la distribución de una variable con el objetivo de poder hacer comparaciones respecto a conjuntos de elementos y a la media mediante la eliminación de los efectos de influencias.

En otras palabras, la normalización son proporciones sin unidades de medida (adimensionales o invariantes de escala) que nos permiten poder

comparar elementos de distintas variables y distintas unidades de medida.

En **estadística** y **econometría** se emplean tablas de distribución de probabilidad tipificadas para encontrar la probabilidad que tome una observación dada la función distribución que siga la variable.

Es importante no limitar el término de normalización sólo a conjuntos de elementos donde la **distribución normal** sea una buena aproximación a su frecuencia

En la siguiente tabla se detallan las normalizaciones más comunes en estadística aplicada a las finanzas y economía.

Normalización	Fórmula	Tablas (Distribuciones)
Puntuación tipificada o estándar	$\frac{X-\mu}{\sigma}$	Normal estándar
T de Student	$\frac{X-\bar{X}}{s}$	T-Student
Coeficiente de variación	$\frac{\mu}{\sigma}$	Poisson
Momento estandarizado	$\frac{\mu_k}{\sigma_k}$	-

- La puntuación tipificada o estándar normaliza los errores cuando podemos calcular los parámetros muestrales.
- La normalización en la **distribución T de Student** normaliza los residuos cuando los parámetros son desconocidos y hacemos una estimación para obtenerlos.
- El **coeficiente de variación** utiliza la media como medida de escala a diferencia de la puntuación tipificada y la T de Student que utilizan la desviación estándar. La distribución se normaliza para las distribuciones de Poisson y exponencial.

- El momento estandarizado puede aplicarse a cualquier distribución de probabilidad que tenga una función generadora de momentos. En otras palabras, que las integrales de los momentos sean convergentes.

¿Cuántas veces hemos leído que la distribución de probabilidad normal parece una buena aproximación a la frecuencia de las observaciones y nos piden encontrar la probabilidad de que la variable X tome un valor concreto?

En otras palabras, establecemos **$X \sim N(\mu,\sigma^2)$**, y nos piden encontrar **$P(X \leq x_i)$**

Sabemos que para encontrar **$P(X \leq x_i)$**, necesitamos buscar la probabilidad en las tablas de distribución de probabilidad. En este caso, en las tablas de la distribución de la distribución normal. Las tablas de distribución de probabilidad más usadas en econometría y finanzas cuantitativas son: ji-cuadrado, t de Student, F de Fisher-Snedecor, Poisson, exponencial, cauchy y la normal estándar.

Las probabilidades calculadas en las tablas de distribución cumplen la propiedad:

$$X \sim N(\mu, \sigma^2), \quad Z = \frac{X - \mu}{\sigma} \rightarrow N(0,1)$$

Es decir, las probabilidades (los números dentro de la tabla) están tipificadas. Entonces, tendremos que tipificar también nuestra variable según los parámetros de la función de distribución si queremos encontrar la probabilidad de **$P(X \leq x_i)$**.

Ejemplo:

Queremos saber la probabilidad de que el número de esquiadores que vayan a esquiar un viernes por la mañana sea de 288.

La estación de esquí nos dice que la frecuencia de la variable esquiadores puede aproximarse a una distribución normal de media 280 y varianza 16.

Entonces, tenemos:

$$X \sim N(\mu, \sigma^2)$$

Donde X la definimos como la variable 'esquiadores'

Nos piden, la probabilidad de que el número de esquiadores que van a esquiar un viernes sea menor o igual que 288. Es decir:

$$P(X \leq 288)$$

Procedimientos:

Para buscar la probabilidad de que el número de esquiadores sea igual a 288, primero tenemos que tipificar la variable.

$$X \sim N(280,16), \quad Z = \frac{288 - 280}{4} = 2 \rightarrow N(0,1)$$

$$P(X \leq 288) = P\left(Z \leq \frac{288 - 280}{4}\right) = P(Z \leq 2)$$

Luego miramos la tabla de distribución de la normal estándar continua:

Z	0	1	2	3
2,0	**0,9772**	0,9778	0,9783	0,9788

$$P(X \leq 288) = P\left(Z \leq \frac{288 - 280}{4}\right) = P(Z \leq 2) = 0{,}9772$$

$$P(X \leq 288) = P(Z \leq 2) = 0{,}9772$$

La probabilidad de que un viernes por la mañana 288 esquiadores vayan a esquiar es de 97,72% dados los parámetros media y varianza.

Referencias

[1] Estadística Descriptiva.

https://www.dm.uba.ar/materias/estadistica_Q/2011/1/modulo%20descriptiva.pdf

[2] Tipos de Gráficos

http://soy-staff.blogspot.com/2015/10/estadistica-descriptiva-graficas.html

[3]Medidas de Tendencia central

https://economipedia.com/definiciones/media.html

[4] Medidas de dispersión

https://www.youtube.com/watch?v=AbN977Xd96k

[5] Teorema de Cherbyshev

https://www.teorema.top/teorema-de-chebyshev/