Departamento de matemática y física
Estadística I – 0834405T

Tema 2

Profa. Mayle leal

Previamente al estudio de este manual deben revisar los siguientes conceptos de los capítulos 3y 4 del libro: Estadística aplicada a los negocios y la economía

MEDIA POBLACIONAL $$\mu = \frac{\Sigma X}{N} \quad [3.1]$$

en la cual:

- μ representa la media poblacional; se trata de la letra minúscula griega *mu;*
- N es el número de valores en la población;
- X representa cualquier valor particular;
- Σ es la letra mayúscula griega *sigma* e indica la operación de suma;
- ΣX es la suma de X valores en la población.

Cualquier característica medible de una población recibe el nombre de **parámetro**. La media de una población es un parámetro.

PARÁMETRO Característica de una población.

Media de una muestra

Como se explicó en el capítulo 1, con frecuencia se selecciona una muestra de la población para encontrar algo sobre una característica específica de la población. Por ejemplo, el departamento de control de calidad necesita asegurarse de que los rodamientos de balas fabricados tengan un diámetro externo aceptable. Resultaría muy costoso y consumiría demasiado tiempo verificar el diámetro externo de todos los rodamientos producidos. Por consiguiente, se selecciona una muestra de cinco rodamientos y se calcula el diámetro externo de cinco rodamientos para aproximar el diámetro medio de todos.

En el caso de los datos en bruto, de los datos no agrupados, *la media es la suma de los valores de la muestra, divididos entre el número total de valores de la muestra.* La media de una muestra se determina de la siguiente manera:

$$\text{Media de la muestra} = \frac{\text{Suma de todos los valores de la muestra}}{\text{Número de valores de la muestra}}$$

La media muestral y la media poblacional se calculan en la misma manera, pero la notación abreviada que se emplea es diferente. La fórmula de la media muestral es:

MEDIA DE UNA MUESTRA $$\bar{X} = \frac{\Sigma X}{n} \quad [3.2]$$

en la cual:

- $\bar{X}$ es la media de la muestra; se lee: X barra;
- n es el número de valores de la muestra.

La media de una muestra o cualquier otra medición basada en una muestra de datos recibe el nombre de **estadístico.** Si el diámetro promedio externo de una muestra de cinco rodamientos de bala es de 0.625 pulgadas, se trata de un ejemplo de estadístico.

ESTADÍSTICO Característica de una muestra.

Media ponderada

La media ponderada constituye un caso especial de la media aritmética y se presenta cuando hay varias observaciones con el mismo valor. Para explicar esto, suponga que el Wendy´s Restaurant vende refrescos medianos, grandes y gigantes a $0.90, $1.25 y $1.50. De las 10 últimas bebidas vendidas 3 eran medianas, 4 grandes y 3 gigantes. Para determinar el precio promedio de las últimas 10 bebidas vendidas recurra a la fórmula 3.2.

$$\bar{X} = \frac{\$.90 + \$.90 + \$.90 + \$1.25 + \$1.25 + \$1.25 + \$1.25 + \$1.50 + \$1.50 + \$1.50}{10}$$

$$\bar{X} = \frac{\$12.20}{10} = \$1.22$$

el precio promedio de venta de las últimas 10 bebidas es de $1. 22.

Una manera fácil para determinar el precio promedio de venta consiste en determinar la media ponderada; multiplique cada observación por el número de veces que aparece. La media ponderada se representa como $\bar{X}_w$, que se lee: "X subíndice w".

$$\bar{X}_w = \frac{3(\$0.90) + 4(\$1.25) + 3(\$1.50)}{10} = \frac{\$12.20}{10} = \$1.22$$

En este caso las ponderaciones son conteos de frecuencias. Sin embargo, cualquier medida de importancia podría utilizarse como una ponderación. En general, la media ponderada del conjunto de números representados como $X_1, X_2, X_3, \ldots, X_n$ con las ponderaciones correspondientes $w_1, w_2, w_3, \ldots, w_n$, se calcula de la siguiente manera:

MEDIA PONDERADA

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n}{w_1 + w_2 + w_2 + \cdots + w_n} \qquad [3.3]$$

La cual se abrevia de la siguiente manera:

$$\bar{X}_w = \frac{\Sigma(wX)}{\Sigma w}$$

Mediana

Ya se ha insistido en que si los datos contienen uno o dos valores muy grandes o muy pequeños, la media aritmética no resulta representativa. Es posible describir el centro de dichos datos a partir de una medida de ubicación denominada **mediana**.

Para ilustrar la necesidad de una medida de ubicación diferente de la media aritmética, suponga que busca un condominio en Palm Aire. Su agente de bienes raíces le dice que el precio típico de las unidades disponibles en este momento es de $110 000. ¿Aún insiste en seguir buscando? Si usted se ha fijado un presupuesto máximo de $75 000, podría pensar que los condominios se encuentran fuera de su presupuesto. Sin embargo, la verificación de los precios de las unidades individuales podría hacerle cambiar de parecer. Los costos son de $60 000, $65 000, $70 000, $80 000 y de $275 000 en el caso de un lujoso penthouse. El importe promedio aritmético es de $110 000, como le informó el agente de bienes raíces, pero un precio ($275 000) eleva la media aritmética y lo convierte en un promedio no representativo. Parece que un precio de poco más o menos $70 000 es un promedio más típico o representativo, y así es. En casos como éste, la mediana proporciona una medida de ubicación más válida.

MEDIANA Punto medio de los valores una vez que se han ordenado de menor a mayor o de mayor a menor.

El precio mediano de las unidades disponibles es de $70 000 Para determinarlo, ordene los precios de menor ($60 000) a mayor ($275 000) y seleccione el valor medio

($70 000). En el caso de la mediana los datos deben ser por lo menos de un nivel ordinal de medición.

Precios ordenados de menor a mayor		Precios ordenados de mayor a menor
$ 60 000		$275 000
65 000		80 000
70 000	← Mediana →	70 000
80 000		65 000
275 000		60 000

Observe que existe el mismo número de precios bajo la mediana de $70 000 que sobre ella. Por consiguiente, a la mediana no le afectan precios bajos o altos. Si el precio más alto fuera de $90 000 o de $300 000, incluso de $1 000 000, el precio mediano aún sería de $70 000. Asimismo, si el precio más bajo fuera de $20 000 o $50 000, el precio mediano todavía sería de $70 000.

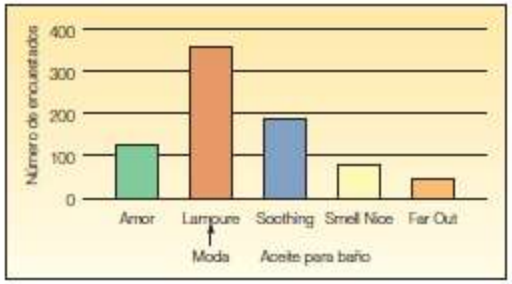
En el ejemplo anterior hay un número *impar* de observaciones (cinco). ¿Cómo se determina la mediana en el caso de un número *par* de observaciones? Como antes, se ordenan las observaciones. Enseguida, con el fin de obtener un único valor por convención, calcule la media de las dos observaciones medias. Así, en el caso de un número par de observaciones, la mediana quizá no sea uno de los valores dados.

Moda

La **moda** es otra medida de ubicación.

MODA Valor de la observación que aparece con mayor frecuencia.

La moda es de especial utilidad para resumir datos de nivel nominal. Un ejemplo de esta aplicación en datos de nivel nominal: una compañía creó cinco aceites para baño. La gráfica de barras 3.1 muestra los resultados de una encuesta de mercado diseñada para determinar qué aceite para baño prefieren los consumidores. La mayoría de los encuestados se inclinó por Lamoure, según lo evidencia la barra más grande. Por consiguiente, Lamoure representa la moda.



GRÁFICA 3.1 Número de encuestados que prefieren ciertos aceites para baño

Los salarios anuales de los gerentes de control de calidad en algunos estados seleccionados aparecen enseguida.

Estado	Salario	Estado	Salario	Estado	Salario
Arizona	$35 000	Illinois	$58 000	Ohio	$50 000
California	49 100	Louisiana	60 000	Tennessee	60 000
Colorado	60 000	Maryland	60 000	Texas	71 400
Florida	60 000	Massachusetts	40 000	Virginia Oeste	60 000
Idaho	40 000	Nueva Jersey	65 000	Wyoming	55 000

Un examen de los salarios revela que el salario anual de $60 000 se presenta con mayor frecuencia (seis veces) que otros salarios. Por tanto, la moda es $60 000.

En resumen, es posible determinar la moda para todos los niveles de datos, nominal, ordinal, de intervalo y de razón. La moda también tiene la ventaja de que no influyen en ella valores extremadamente grandes o pequeños.

No obstante, la moda tiene sus desventajas, por las cuales se le utiliza con menor frecuencia que a la media o a la mediana. En el caso de muchos conjuntos de datos no existe la moda, porque ningún valor se presenta más de una vez. Por ejemplo, no hay moda en el siguiente conjunto de datos de precios: $19, $21, $23, $20 y $18. Sin embargo, como cada valor es diferente, podría argumentar que cada valor es la moda. Por lo contrario, en el caso de algunos conjuntos de datos hay más de una moda. Suponga que las edades de los miembros de un club de inversionistas son 22, 26, 27, 27, 31, 35 y 35. Ambas edades, 27 y 35 son modas. Así, este agrupamiento de edades se denomina *bimodal* (tiene dos modas). Alguien podría cuestionar la utilización de dos modas para representar la ubicación de este conjunto de datos de edades.

Medidas de dispersión

Consideraremos diversas medidas de dispersión. El rango se sustenta en los valores máximo y mínimo del conjunto de datos. La desviación media, la varianza y la desviación estándar se basan en desviaciones de la media aritmética.

Rango

La medida más simple de dispersión es el **rango**. Representa la diferencia entre los valores máximo y mínimo de un conjunto de datos. En forma de ecuación:

RANGO Rango = Valor máximo − valor mínimo **[3.6]**

El rango se emplea mucho en aplicaciones de control de procesos estadísticos (CPE) como consecuencia de que resulta fácil de calcular y entender.

Consulte la gráfica 3.6. Determine el rango del número de monitores de computadora producidos por hora en las plantas de Baton Rouge y Tucson. Interprete los dos rangos.

El rango de la producción por hora de monitores de computadora en la planta de Baton Rouge es de 4, el cual se determina por la diferencia entre la producción máxima por hora de 52 y la mínima de 48. El rango de la producción por hora en la planta de Tucson es de 20 monitores de computadora, obtenido con el cálculo 60 − 40. Por tanto: 1. Existe menos dispersión en la producción por hora en la planta de Baton Rouge que en la planta de Tucson, porque el rango de 4 monitores de computadora es menor que el rango de 20 monitores; 2. La producción se acumula más alrededor de la media de 50 en la planta de Baton Rouge que en la planta de Tucson (ya que un rango de 4 es menor que un rango de 20). Así, la producción media en la planta de Baton Rouge (50 monitores de computadora) resulta una medida de ubicación más representativa que la media de 50 monitores de computadora en la planta de Tucson.

Varianza y desviación estándar

La **varianza** y la **desviación estándar** también se fundamentan en las desviaciones de la media. Sin embargo, en lugar de trabajar con el valor absoluto de las desviaciones, la varianza y la desviación estándar lo hacen con el cuadrado de las desviaciones.

VARIANZA Media aritmética de las desviaciones de la media elevadas al cuadrado.

La varianza es no negativa y es cero sólo si todas las observaciones son las mismas.

DESVIACIÓN ESTÁNDAR Raíz cuadrada de la varianza.

Varianza de la población Las fórmulas de la varianza poblacional y la varianza de la muestra son ligeramente diferentes. La varianza de la población se estudia primero. (Recuerde que una población es la totalidad de las observaciones estudiadas.) La **varianza de la población** se determina de la siguiente manera:

VARIANZA DE LA POBLACIÓN

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \qquad [3.8]$$

En esta fórmula:

σ^2 es la varianza de la población (σ es la letra minúscula griega sigma); se lee *sigma al cuadrado;*
X es el valor de una observación de la población;
μ es la media aritmética de la población;
N es el número de observaciones de la población.

Observe el proceso de cálculo de la varianza:

- Comience determinando la media;
- En seguida calcule la diferencia entre cada observación y la media, y eleve al cuadrado dicha diferencia;
- Entonces sume todas las diferencias elevadas al cuadrado;
- Por último divida la suma de las diferencias elevadas al cuadrado entre el número de elementos de la población.

Varianza muestral La fórmula para la media poblacional es $\mu = \Sigma X/N$. Sencillamente cambie los símbolos para la media de la muestra; es decir, $\bar{X} = \Sigma X/n$. Por desgracia, la conversión de una varianza poblacional en una varianza muestral no es tan directa.

Requiere un cambio en el denominador. En lugar de sustituir n (el número en la muestra) por N (el número en la población), el denominador es $n-1$. Así, la fórmula de la **varianza muestral** es:

VARIANZA MUESTRAL

$$s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} \qquad [3.10]$$

en la cual:

s^2 es la varianza muestral;
X es el valor de cada observación de la muestra;
$\bar{X}$ es la media de la muestra;
n es el número de observaciones en la muestra.

El teorema de Chebyshev establece lo siguiente:

TEOREMA DE CHEBYSHEV En cualquier conjunto de observaciones (muestra o población), la proporción de valores que se encuentran a k desviaciones estándares de la media es de por lo menos $1 - 1/k^2$, siendo k cualquier constante mayor que 1.

La media aritmética de la suma quincenal que aportan los empleados de Dupree Saint para el plan de reparto de utilidades de la compañía es de $51.54 y la desviación estándar, de $7.51. ¿Por lo menos qué porcentaje de las aportaciones se encuentra en más 3.5 desviaciones estándares y menos 3.5 desviaciones de la media?

Alrededor de 92%, que se determina de la siguiente manera:

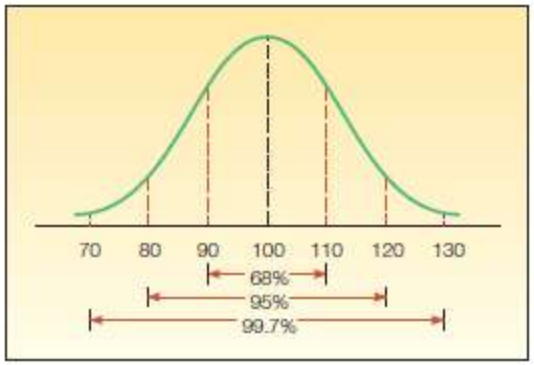
$$1-\frac{1}{k^2}=1-\frac{1}{(3.5)^2}=1-\frac{1}{12.25}=0.92$$

La regla empírica

El teorema de Chebyshev tiene que ver con cualquier conjunto de valores; es decir, que la distribución de valores puede tener cierta forma. Sin embargo, en cualquier distribución simétrica con forma de campana, como muestra la gráfica 3.7, es posible ser más precisos en la explicación de la dispersión en torno a la media. Estas relaciones que implican la desviación estándar y la media se encuentran descritas en la **regla empírica**, a veces denominada **regla normal**.

REGLA EMPÍRICA En cualquier distribución de frecuencias simétrica con forma de campana, aproximadamente 68% de las observaciones se encontrarán entre más y menos una desviación estándar de la media; cerca de 95% de las observaciones se encontrarán entre más y menos dos desviaciones estándares de la media y, de hecho todas (99.7%), estarán entre más y menos tres desviaciones estándares de la media.

Estas relaciones se representan en la gráfica 3.7 en el caso de una distribución con forma de campana con una media de 100 y una desviación estándar de 10.



Otras medidas de dispersión

La desviación estándar es la medida de dispersión más generalmente utilizada. No obstante, existen otras formas de describir la variación o dispersión de un conjunto de datos. Un método consiste en determinar la *ubicación* de los valores que dividen un conjunto de observaciones en partes iguales. Estas medidas incluyen los **cuartiles**, **deciles** y **percentiles**.

Los cuartiles dividen a un conjunto de observaciones en cuatro partes iguales. Para explicarlo mejor, piense en un conjunto de valores ordenados de menor a mayor. En el capítulo 3 denominamos *mediana* al valor intermedio de un conjunto de datos ordenados de menor a mayor. Es decir, que 50% de las observaciones son mayores que la mediana y 50% son menores. La mediana constituye una medida de ubicación, ya que señala el centro de los datos. De igual manera, los **cuartiles** dividen a un conjunto de observaciones en cuatro partes iguales. El primer cuartil, representado mediante Q_1, es el valor debajo del cual se presenta 25% de las observaciones, y el tercer cuartil, representado como Q_3, es el valor debajo del cual se presenta 75% de las observaciones. Es lógico, Q_2 es la mediana. Q_1 puede considerarse como la *mediana* de la mitad inferior de los datos y Q_3 como la *mediana* de la parte superior de los datos.

Asimismo, los **deciles** dividen a un conjunto de observaciones en 10 partes iguales y los **percentiles** en 100 partes iguales. Por tanto, si su promedio general en la universidad se encuentra en el octavo decil, usted podría concluir que 80% de los estudiantes tuvieron un promedio general inferior al de usted y que 20%, un promedio superior. Un promedio general ubicado en el trigésimo tercer percentil significa que 33% de los estudiantes tienen un promedio general más bajo y 67% tienen un promedio general más alto. Las calificaciones expresadas en percentiles se utilizan a menudo para dar a conocer resultados relacionados con pruebas estandarizadas en Estados Unidos, como SAT, ACT, GMAT (empleado para determinar el ingreso en algunas maestrías de administración de empresas) y LSAT (empleado para determinar el ingreso a la escuela de leyes).

Cuartiles, deciles y percentiles

Para formalizar el proceso de cálculo, suponga que L_p representa la ubicación de cierto percentil que se busca. De esta manera, si quiere encontrar el trigésimo tercer percentil, utilizaría L_{33}, y si buscara la mediana, el percentil 50°, entonces L_{50}. El número de observaciones es n; así que, si desea localizar la mediana, su posición se encuentra en $(n + 1)/2$, o podría escribir esta expresión como $(n + 1)(P/100)$, en la que P representa el percentil que busca.

LOCALIZACIÓN DE UN PERCENTIL

$$L_p = (n+1)\frac{P}{100} \qquad [4.1]$$

Diagramas de caja

Un **diagrama de caja** es la representación gráfica, basada en cuartiles, que ayuda a exhibir un conjunto de datos. Para construir un diagrama de caja, sólo necesita cinco estadísticos: el valor mínimo, Q_1 (primer cuartil), la mediana, Q_3 (tercer cuartil) y el valor máximo. Un ejemplo ayudará a explicarlo.

Alexander's Pizza ofrece entregas gratuitas de pizza a 15 millas a la redonda. Alex, el propietario, desea información relacionada con el tiempo de entrega. ¿Cuánto tiempo tarda una entrega típica? ¿En qué margen de tiempos deben completarse la mayoría de las entregas? En el caso de una muestra de 20 entregas, Alex recopiló la siguiente información:

Valor mínimo = 13 minutos

Q_1 = 15 minutos

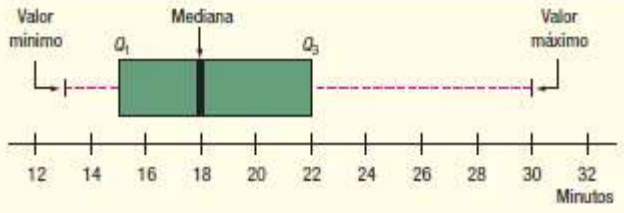
Mediana = 18 minutos

Q_3 = 22 minutos

Valor máximo = 30 minutos

Elabore un diagrama de caja para los tiempos de entrega. ¿Qué conclusiones deduce sobre los tiempos de entrega?

El primer paso para elaborar un diagrama de caja consiste en crear una escala adecuada a lo largo del eje horizontal. Enseguida, dibujamos una caja que inicie en Q_1 (15 minutos) y termine en Q_3 (22 minutos). Dentro de la caja trazamos una línea vertical para representar a la mediana (18 minutos). Por último, prolongamos líneas horizontales a partir de la caja dirigidas al valor mínimo (13 minutos) y al valor máximo (30 minutos). Estas líneas horizontales que salen de la caja, a veces reciben el nombre de *bigotes*, en virtud de que se asemejan a los bigotes de un gato.

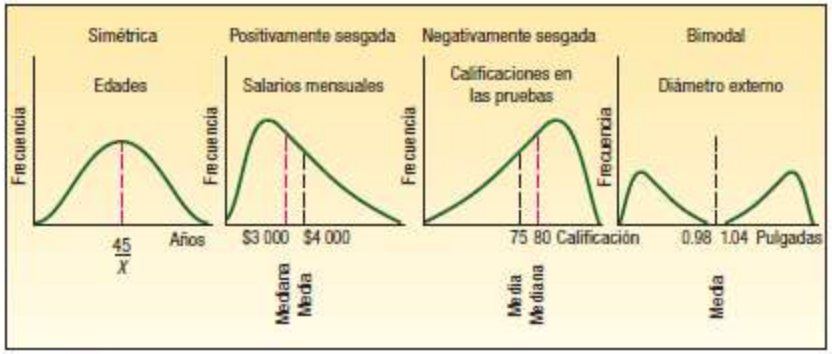


Sesgo

En el capítulo 3 se trataron las medidas de ubicación central para un conjunto de observaciones por medio de la presentación de un informe sobre la media, la mediana y la moda. También se describieron medidas que muestran el grado de propagación o variación de un conjunto de datos, como el rango y la desviación estándar.

Otra característica de un conjunto de datos es la forma. Hay cuatro formas: simétrica, con sesgo positivo, con sesgo negativo y bimodal. En un conjunto **simétrico** de observaciones la media y la mediana son iguales, y los valores de datos se dispersan uniformemente en torno a estos valores. Los valores de datos debajo de la media y de la mediana constituyen una imagen especular de los datos arriba de estas medidas. Un conjunto de valores se encuentra **sesgado a la derecha** o **positivamente sesgado** si existe un solo pico y los valores se extienden mucho más allá a la derecha del pico que a la izquierda de éste. En este caso la media es más grande que la mediana. En una distribución **negativamente sesgada** existe un solo pico, pero las observaciones se extienden más a la izquierda, en la dirección negativa, que a la derecha. En una distribución negativamente sesgada, la media es menor que la mediana. Las distribuciones positivamente sesgadas son más comunes. Los salarios con frecuencia obedecen este patrón. Piense en los salarios de los empleados de una pequeña compañía con aproximadamente 100 personas. El presidente y unos cuantos altos ejecutivos tendrían salarios muy altos respecto de los demás trabajadores, y de ahí que la distribución de salarios mostraría un sesgo positivo. Una **distribución bimodal** tendrá dos o más picos.

Con frecuencia éste es el caso cuando los valores provienen de dos o más poblaciones. Esta información se resume en la gráfica 4.1.



GRÁFICA 4.1 Formas de los polígonos de frecuencias


ESTADÍSTICAS EN EXCEL

Análisis de datos estadísticos con excel

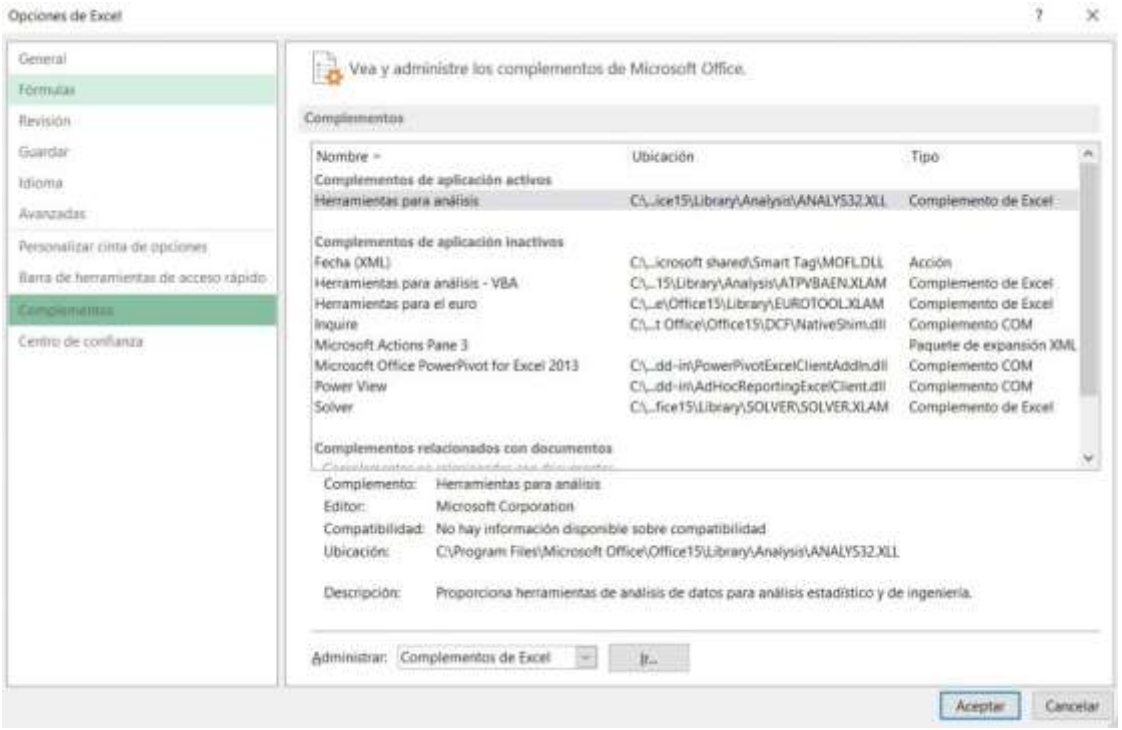
Contenido

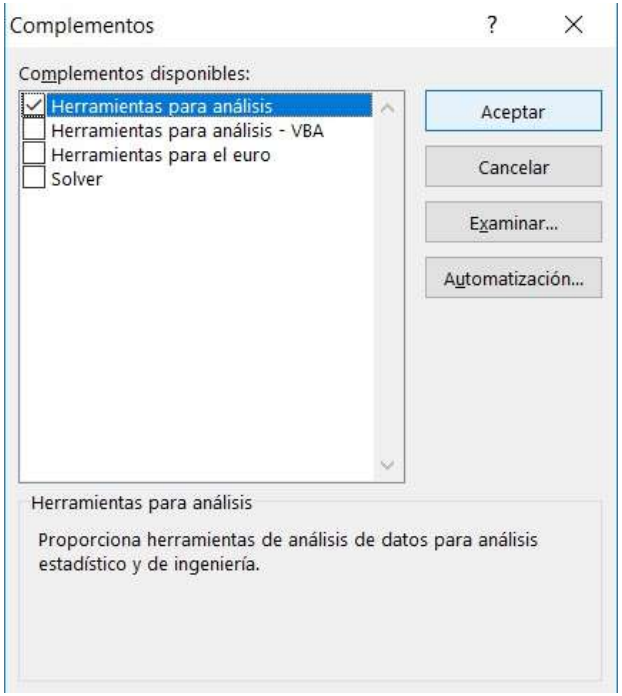
1. ¿Cómo cargar el paquete análisis de datos?

Antes de empezar deben cargar el paquete herramientas para análisis que tiene Excel.
En la web oficial de microsoft office o aquí los pasos a seguir:

1. Archivo >> opciones >>
2. Ves a complementos >> ir a…
3. Marca la opción herramienta para el análisis y aceptar

Esta imagen muestra la ventana de opciones >> complementos
La siguiente imagen refleja la opción para activar las herramientas de análisis de datos.





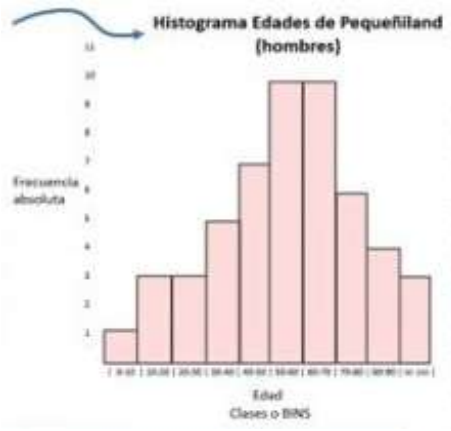
En la pestaña datos aparecerá una nueva opción Análisis de datos

2. ¿Cómo construir una tabla de frecuencias y el histograma con la herramienta para el análisis de excel?

En este apartado pueden crear una tabla de frecuencias con los intervalos escogidos de manera eficaz y hacer el histograma con la herramienta para el análisis que has cargado antes.

Histograma y la tabla de frecuencias

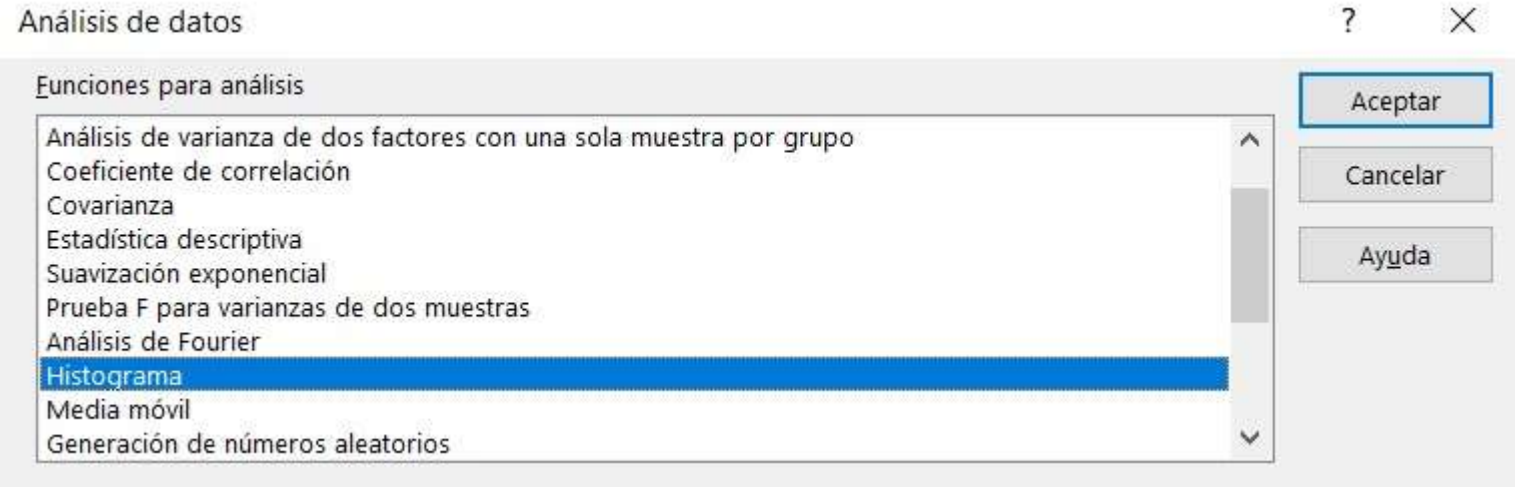
- Un histograma es un diagrama de barras. La altura de cada barra refleja la cantidad de individuos de un clase o intervalo.
- La tabla de frecuencias es una manera de resumir los datos. Imagínate un montón de números en fila. Puedes agrupar por intervalos y contar cuántos individuos tienen de un mismo intervalo o clase.



Clases	Frecuencia absoluta	Frecuencia absoluta acumulada	Frecuencia relativa	Frecuencia relativa acumulada
0-10 años	1	1	1.92%	1.92%
10-20 años	3	4	7.69%	7.69%
20-30 años	3	7	13.46%	13.46%
30-40 años	5	12	23.08%	23.08%
40-50 años	7	19	36.54%	36.54%
50-60 años	10	29	55.77%	55.77%
60-70 años	10	39	75.00%	75.00%
70-80 años	6	45	86.54%	86.54%
80-90 años	4	49	94.23%	94.23%
90-100 años	3	52	100.00%	100.00%

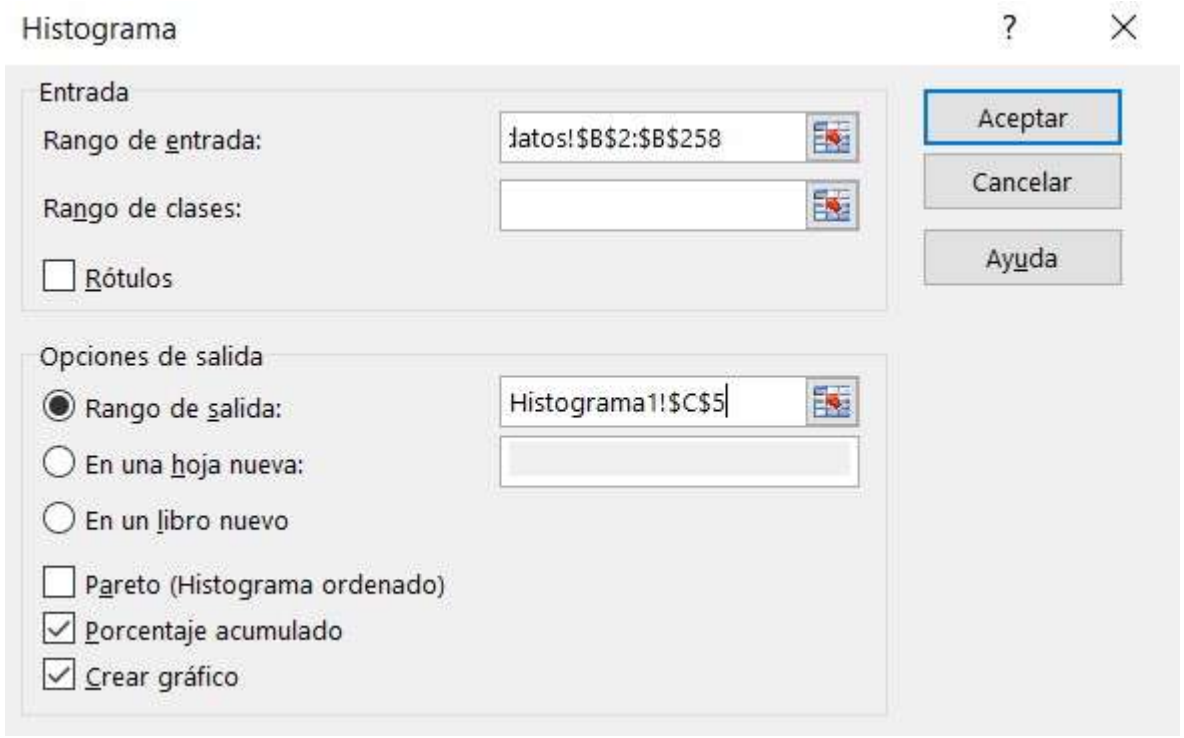
Los datos por si solos no dicen nada pero se pueden agrupar las edades en grupos de 10 años en 10 años. Estos grupos son las clases o intervalos. La frecuencia será la cantidad de personas de cada grupo. Puedes crear una tabla muy sencilla con las frecuencias por cada rango de edad, esto es la tabla de frecuencias. Solo hace falta dibujar el histograma pintando las barras con la altura de cada frecuencia.

En la pestaña datos y al final a la derecha tienes la herramienta que has cargado antes análisis de datos. Selecciona histograma. Con esta opción vas a crear la tabla de frecuencias que mejor se ajusta a tus datos y el histograma.



En la ventana siguiente:

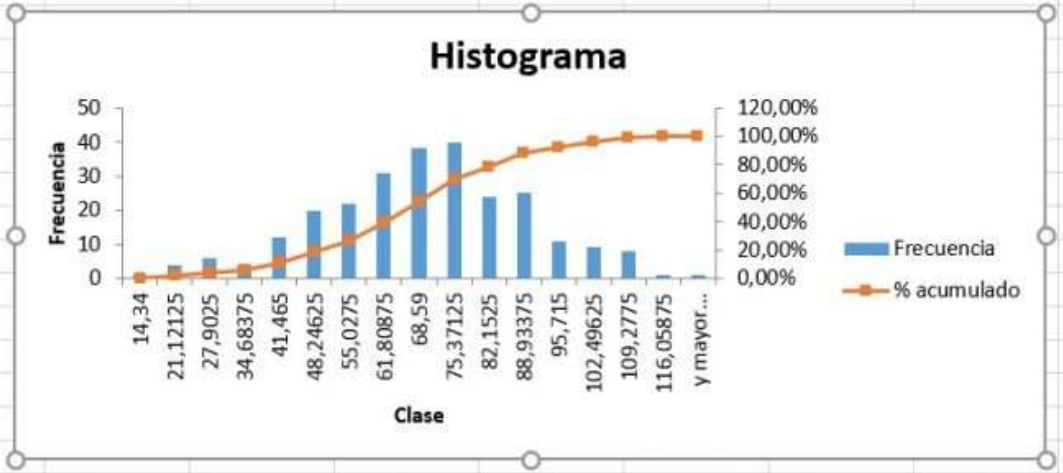
- Seleccionar el rango de datos de entrada
- Seleccionar dónde quieres que copie la tabla de frecuencias y el histograma
- Activar la opción crear gráfico y si quieres % acumulado



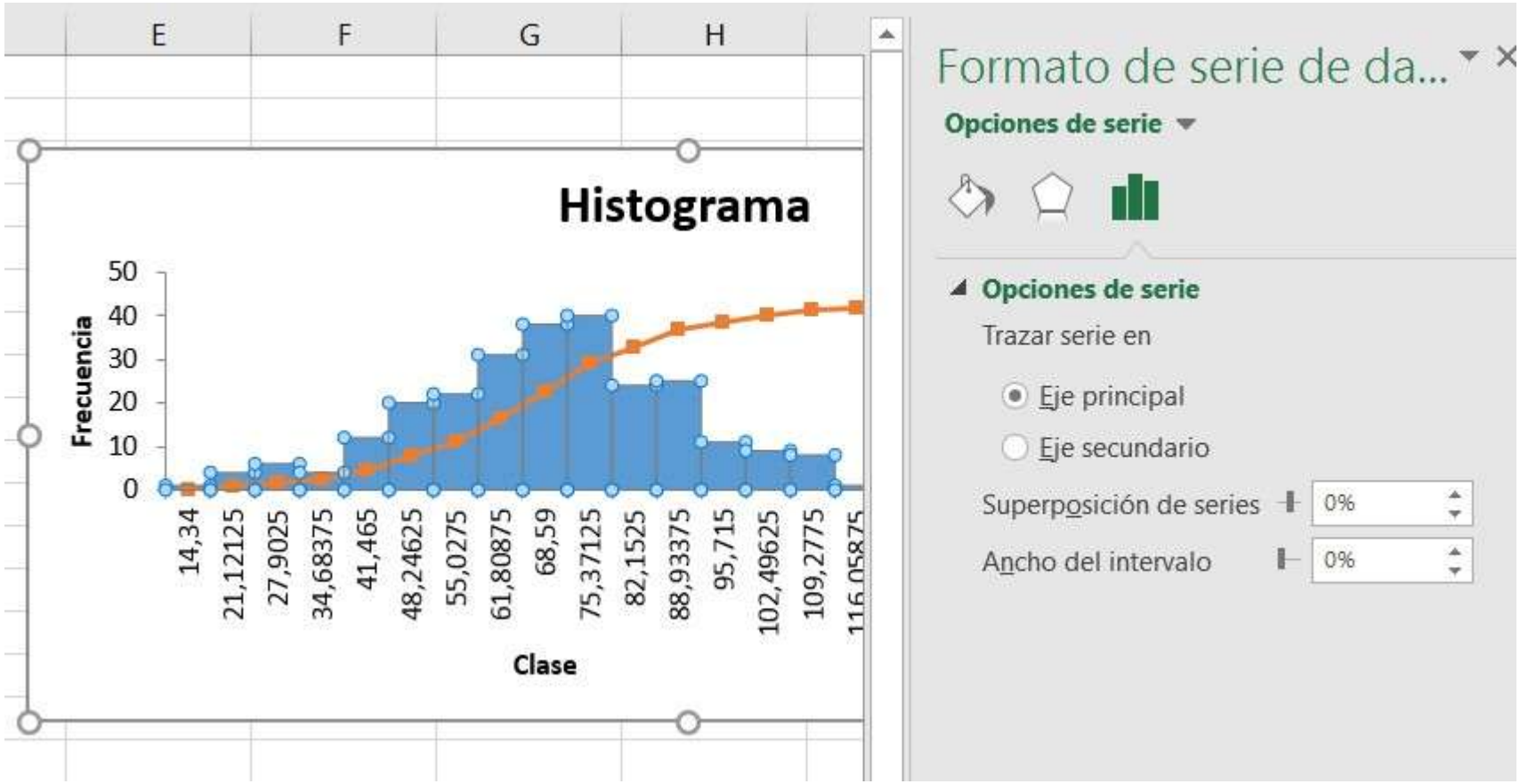
Excel se ha encargado de seleccionar el rango de clases

Este es el resultado:

	A	B	C
1	*Clase*	*Frecuencia*	*% acumulado*
2	14,34	1	0,39%
3	21,12125	4	1,95%
4	27,9025	6	4,28%
5	34,68375	4	5,84%
6	41,465	12	10,51%
7	48,24625	20	18,29%
8	55,0275	22	26,85%
9	61,80875	31	38,91%
10	68,59	38	53,70%
11	75,37125	40	69,26%
12	82,1525	24	78,60%
13	88,93375	25	88,33%
14	95,715	11	92,61%
15	102,49625	9	96,11%
16	109,2775	8	99,22%
17	116,05875	1	99,61%
18	y mayor...	1	100,00%
19			



Para juntar las barras del histograma de manera muy sencilla. Selecciona las barras >> botón derecho y ancho de intervalo 0.

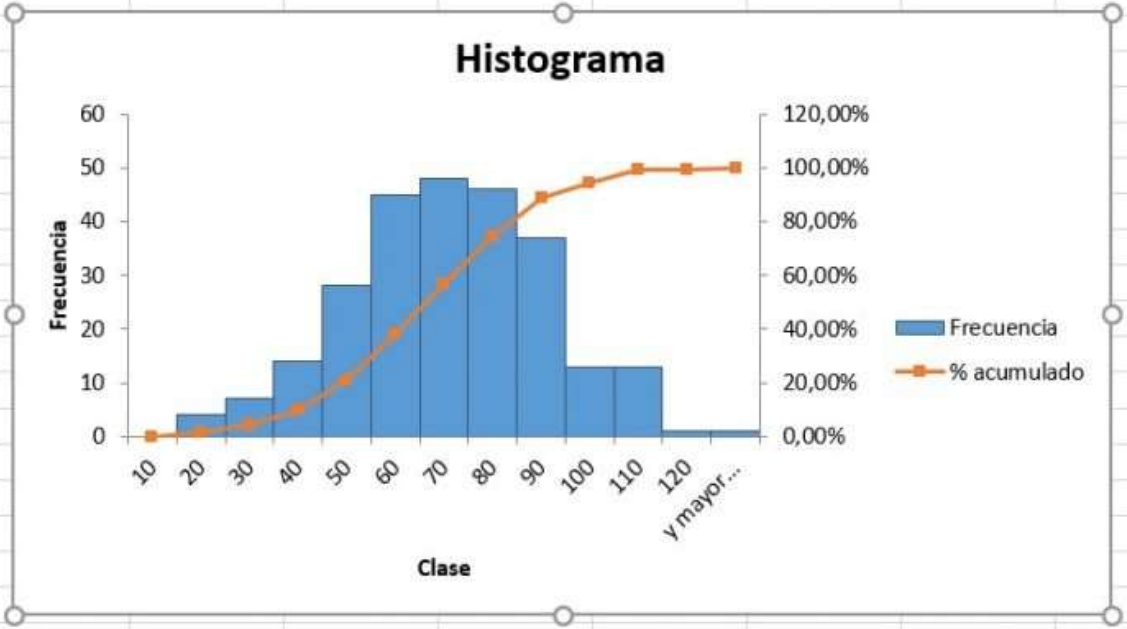


También se puede crear el intervalo de clases manualmente. Por ejemplo: usar de 10 en 10 kg. Se crea una nueva columna con las clases personalizadas.

	A	B	C
1	Sexo	Peso [kg]	clases [kg]
2	HOMBRE	72,72	10
3	HOMBRE	69,12	20
4	HOMBRE	65,26	30
5	MUJER	49,35	40
6	MUJER	61,84	50
7	HOMBRE	64,82	60
8	MUJER	72,07	70
9	HOMBRE	71,48	80
10	HOMBRE	82,92	90
11	MUJER	51,07	100
12	HOMBRE	90,47	110
13	HOMBRE	82,24	120
14	MUJER	47,70	
15	MUJER	64,32	
16	MUJER	38,07	
17	MUJER	59,48	
18	MUJER	67,19	
19	MUJER	42,94	

Para crear este nuevo histograma con los intervalos de clases de 10 en 10 se introducen estos intervalos de clase en la opción rango de clases > seleccionas el rango. Este es el resultado:

	A	B	C
1	*Clase*	*Frecuencia*	*% acumulado*
2	10	0	0,00%
3	20	4	1,56%
4	30	7	4,28%
5	40	14	9,73%
6	50	28	20,62%
7	60	45	38,13%
8	70	48	56,81%
9	80	46	74,71%
10	90	37	89,11%
11	100	13	94,16%
12	110	13	99,22%
13	120	1	99,61%
14	y mayor...	1	100,00%

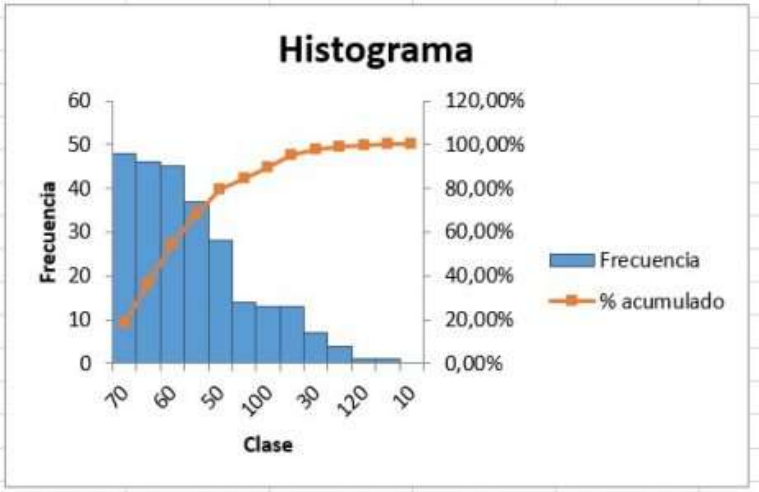


Ahora el histograma tiene los intervalos que yo he puesto.

Diagrama de Pareto

En el diagrama de Pareto se ordenan las alturas de las barras del histograma de mayor a menor. Sencillamente selecciona la opción Pareto (histograma ordenado), Este grafico es muy usado en control de calidad ya que identifica cuales son las causas que generan la mayor parte de un problema.

	A	B	C	D	E	F
1	*Clase*	*Frecuencia*	*% acumulado*	*Clase*	*Frecuencia*	*% acumulado*
2	10	0	0,00%	70	48	18,68%
3	20	4	1,56%	80	46	36,58%
4	30	7	4,28%	60	45	54,09%
5	40	14	9,73%	90	37	68,48%
6	50	28	20,62%	50	28	79,38%
7	60	45	38,13%	40	14	84,82%
8	70	48	56,81%	100	13	89,88%
9	80	46	74,71%	110	13	94,94%
10	90	37	89,11%	30	7	97,67%
11	100	13	94,16%	20	4	99,22%
12	110	13	99,22%	120	1	99,61%
13	120	1	99,61%	y mayor...	1	100,00%
14	y mayor...	1	100,00%	10	0	100,00%
15						
16						



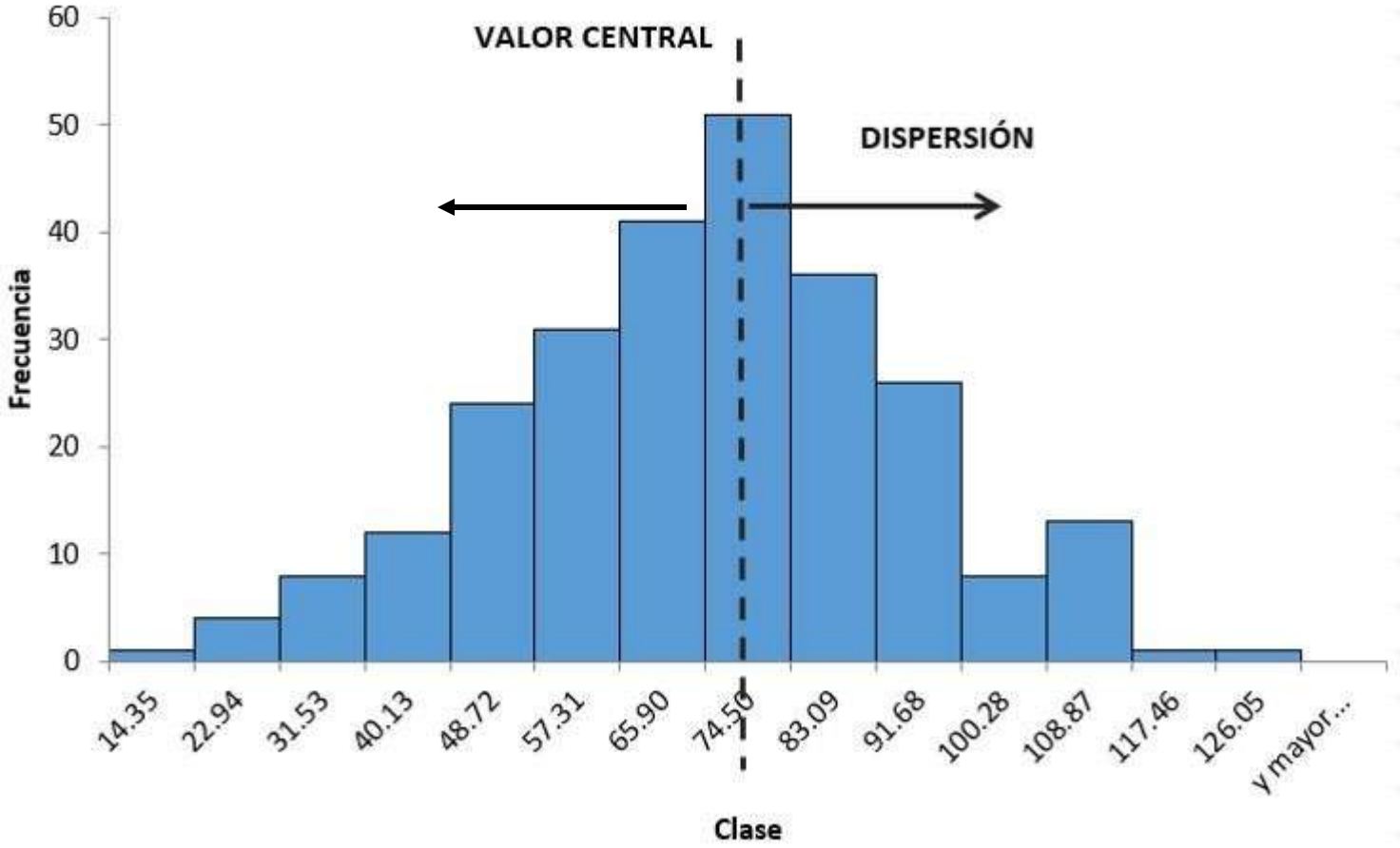
Ahora tienes las barras ordenadas de mayor altura a menor.

3. ¿Cómo crear el resumen numérico de los datos en Excel?

Se trata de calcular los números que pueden resumir las propiedades del histograma:

- Centralidad
- Y dispersión

En la siguiente figura puedes ver la idea de centralidad y dispersión en un histograma:



El valor central es algo así como el valor medio (hacia donde tienden los datos) y la dispersión es la distancia de los datos al valor central. Cuanta más dispersión más alargado es el histograma.

Los valores más interesantes son:

1. Para la centralidad >> el promedio o media y la mediana o cuartil 2.
2. Para la dispersión >> la desviación típica o estándar

La media es la suma de todos los datos dividido por el total de individuos. La desviación típica o estándar es la media de las distancias al cuadrado de los datos a la media. Los cuartiles se calculan ordenando los datos de menor a mayor. Y agrupando en 4 grupos iguales en número.

- La mediana es el cuartil 2.
- El rango intercuartílico es la diferencia entre el cuartil 1y 3.

Cómo calcular las medidas de tendencia central y la dispersión en Excel

Hay varias formas de encontrar estas medidas como se muestra a continuación

1. Con funciones:

   - Media >> = promedio(datos)
   - Mediana >> =mediana(datos)

   cuartil 2 o mediana >> =cuartil.exc(datos;2)

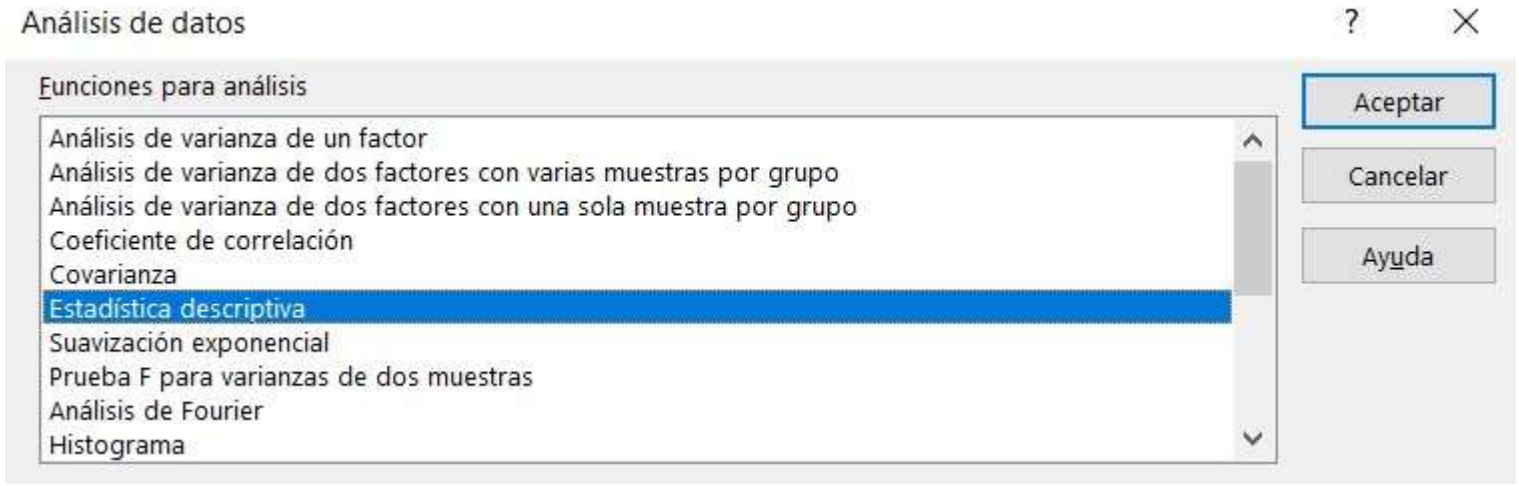
   - Desviación típica >> =devesta(datos;2)
   - Rango intercuartílico >> =cuartil.exc(datos;3)-cuartil.exc(datos;1)
   - Número de observaciones o “n” >> =contar(datos)

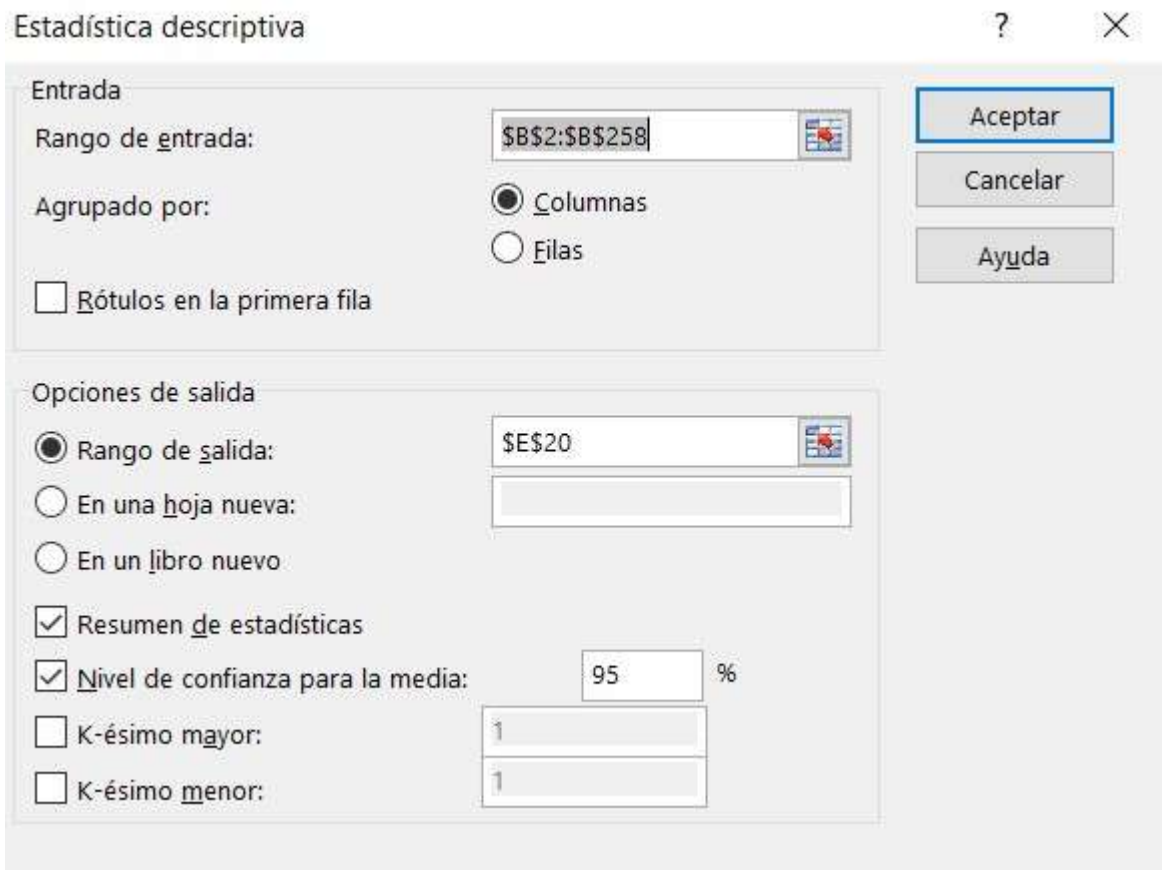
Aquí tienes los resultados

Mediana	67,140000	=MEDIANA(B2:B258)
Q2	67,14	=CUARTIL.EXC(B2:B258;2)
Q1	53,185	=CUARTIL.EXC(B2:B258;1)
Q3	80,505	=CUARTIL.EXC(B2:B258;3)
Q3-Q1	27,32	=CUARTIL.EXC(B2:B258;3)-CUARTIL.EXC(B2:B258;1)
n	257	=CONTAR(B2:B258)
Desviación típica	19,995737	=DESVESTA(B2:B258)

2. Con la Herramienta Análisis de Datos

Para hacerlo: en la pestaña datos >> análisis de datos y después selecciona la opción estadística descriptiva.
Selecciona el rango de entrada de datos como siempre y el el rango de salida. Dónde quieres pintar la tabla del resumen numérico:





Esta opción sirve para obtener una tabla con distintas características numéricas:

Variable Peso		
Media	66,2063813	Promedio
Error típico	1,2472998	
Mediana	67,14	Cuartil2 o mediana
Moda	58,28	Es el valor que más veces se repite
Desviación estándar	19,995737	
Varianza de la muestra	399,829498	Es el cuadrado de la Desviación estándar
Curtosis	-0,03871721	
Coeficiente de asimetrí	-0,08258888	
Rango	108,5	La diferencia entre el máximo y el mínimo
Mínimo	14,34	El valor mínimo
Máximo	122,84	El valor máximo
Suma	17015,04	La suma de todos los individuos
Cuenta	257	El número de individuos N
Nivel de confianza(95,0	2,45627494	Con un probabilidad del 95% la media de la población se encuentra entre 66.206 +- 2.45 kg

3. Con las tablas dinámicas: para esto primero revisemos primero la siguiente información

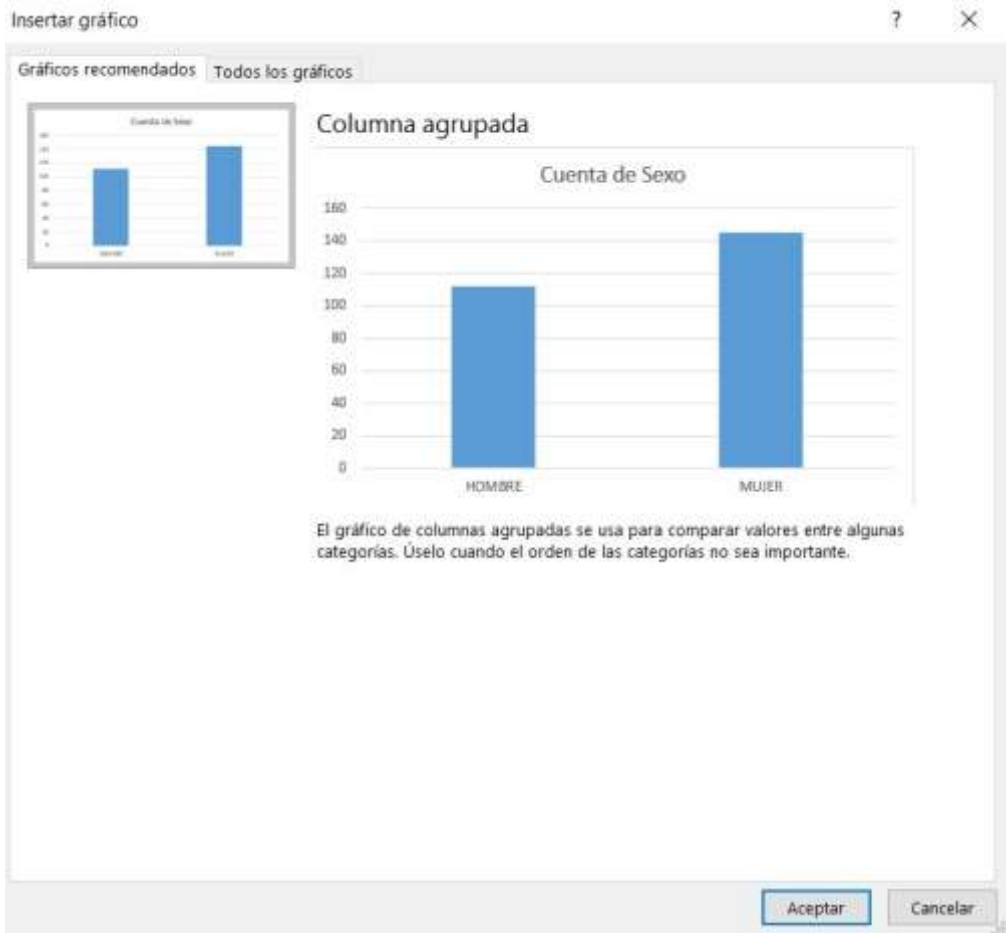
¿Cómo construir un diagrama de barras con variables categóricas?

Hasta ahora he mostrado las opciones con variables numéricas. Pero a veces aparecen variables categóricas. En este caso la variable sexo, es categórica. Tienes las categorías hombre y mujer. El concepto del diagrama de barras de

variables categóricas es bastante sencillo. La altura de cada barra será el número de personas de cada grupo o categoría. Para hacerlo en Excel: Selecciona la columna de variables categóricas con el título de la columna:

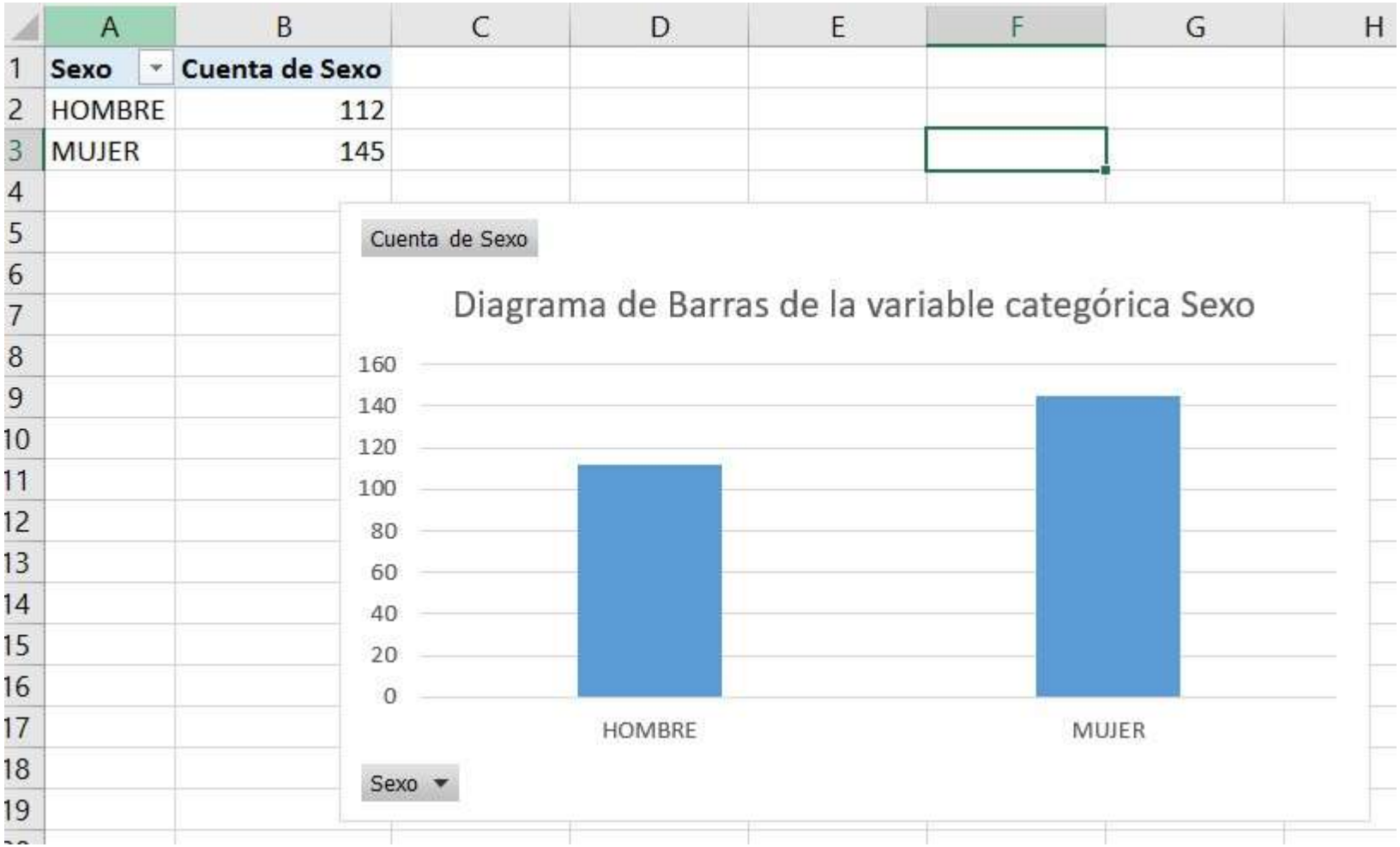
	A	B
1	Sexo	Peso [kg]
2	HOMBRE	72,72
3	HOMBRE	69,12
4	HOMBRE	65,26
5	MUJER	49,35
6	MUJER	61,84
7	HOMBRE	64,82
8	MUJER	72,07
9	HOMBRE	71,48
10	HOMBRE	82,92
11	MUJER	51,07
12	HOMBRE	90,47
13	HOMBRE	82,24
14	MUJER	47,70
15	MUJER	64,32
16	MUJER	38,07

Vas a insertar gráfico > gráficos recomendados



Excel crea automáticamente un gráfico con tablas dinámicas. Recuerda mirar el video de tablas dinámicas

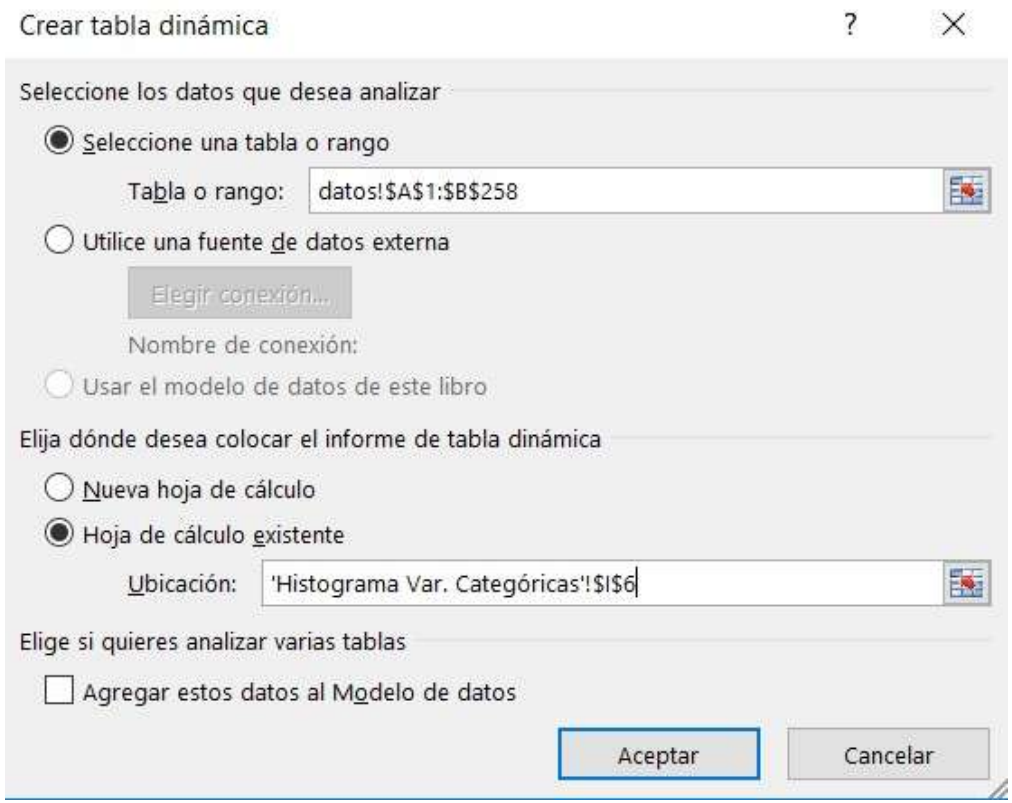
Este caso la tabla dinámica consiste en contar el número de personas de cada grupo.



También lo puedes hacer manualmente con la función contar y obtener la misma tabla y hacer el gráfico de barras sencillamente. Otra opción es hacer la tabla dinámica como tu quieras. Sólo tienes que seleccionar los datos así:

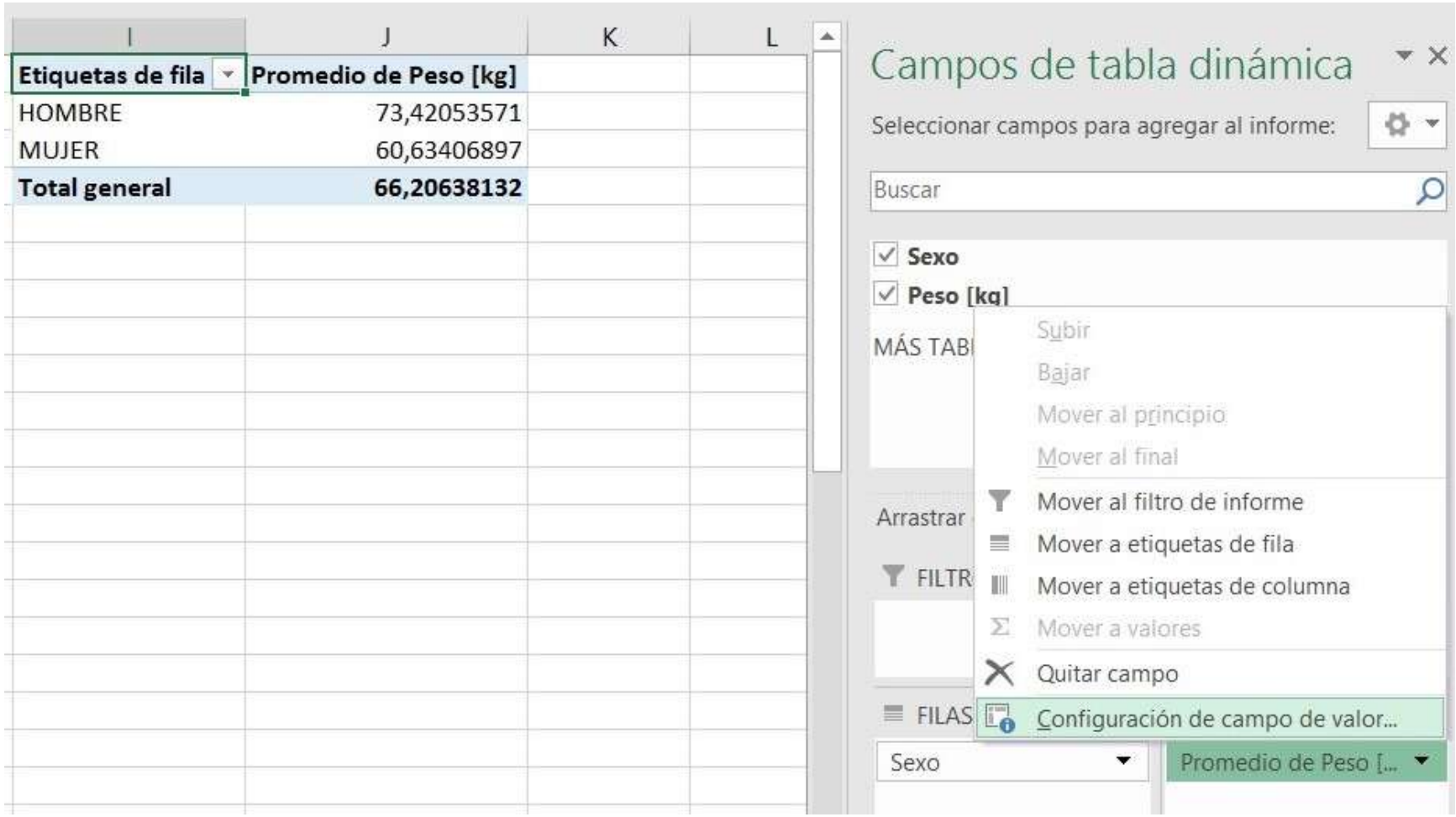
	A	B
1	Sexo	Peso [kg]
2	HOMBRE	72,72
3	HOMBRE	69,12
4	HOMBRE	65,26
5	MUJER	49,35
6	MUJER	61,84
7	HOMBRE	64,82
8	MUJER	72,07
9	HOMBRE	71,48
10	HOMBRE	82,92
11	MUJER	51,07
12	HOMBRE	90,47
13	HOMBRE	82,24
14	MUJER	47,70

Después insertar >> tabla dinámica y seleccionar la celda donde quieres que se calcule la tabla:



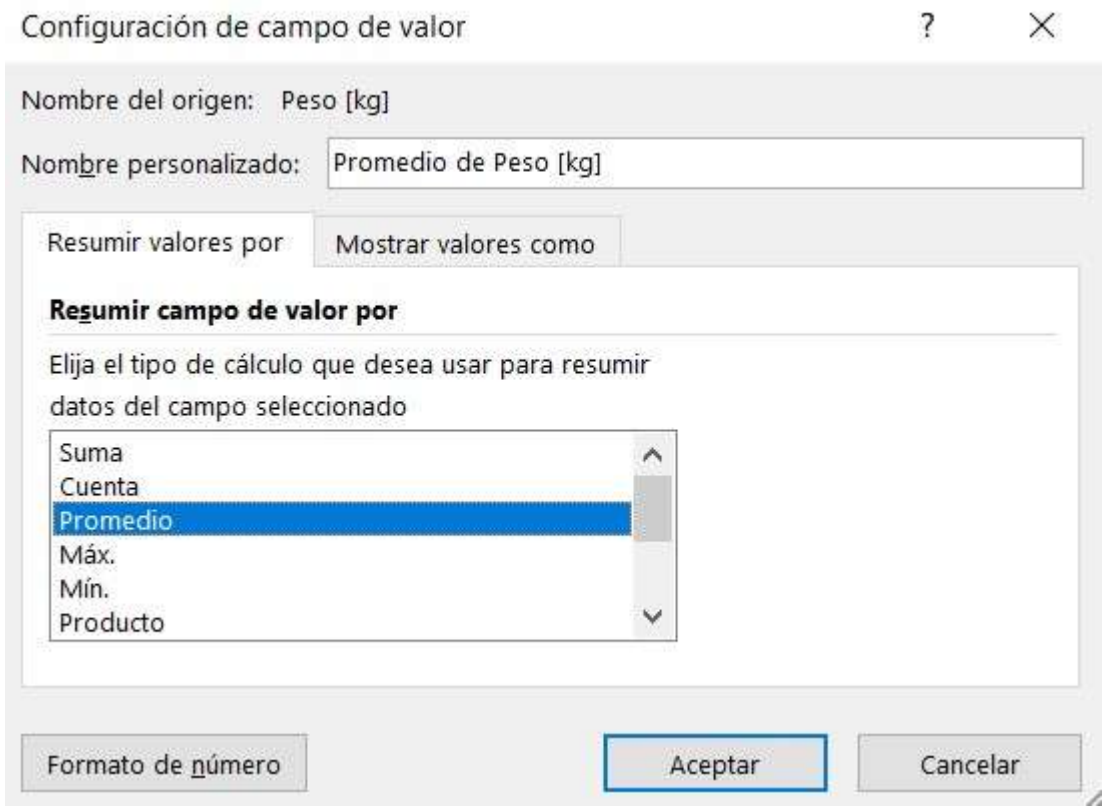
Ahora teniendo seleccionada la tabla dinámica que acabas de crear puedes cambiar el tipo de cálculo:

- Puedes calcular el peso total por cada categoría
- O puedes calcular el promedio o media por cada categoría



Si buscas el tipo de filtro (abajo a la derecha) y le das a configuración de campo de valor… >> puede cambiar el tipo de cálculo.

He seleccionado el promedio:



Y puedes crear el gráfico de barras para comparar los dos grupos…

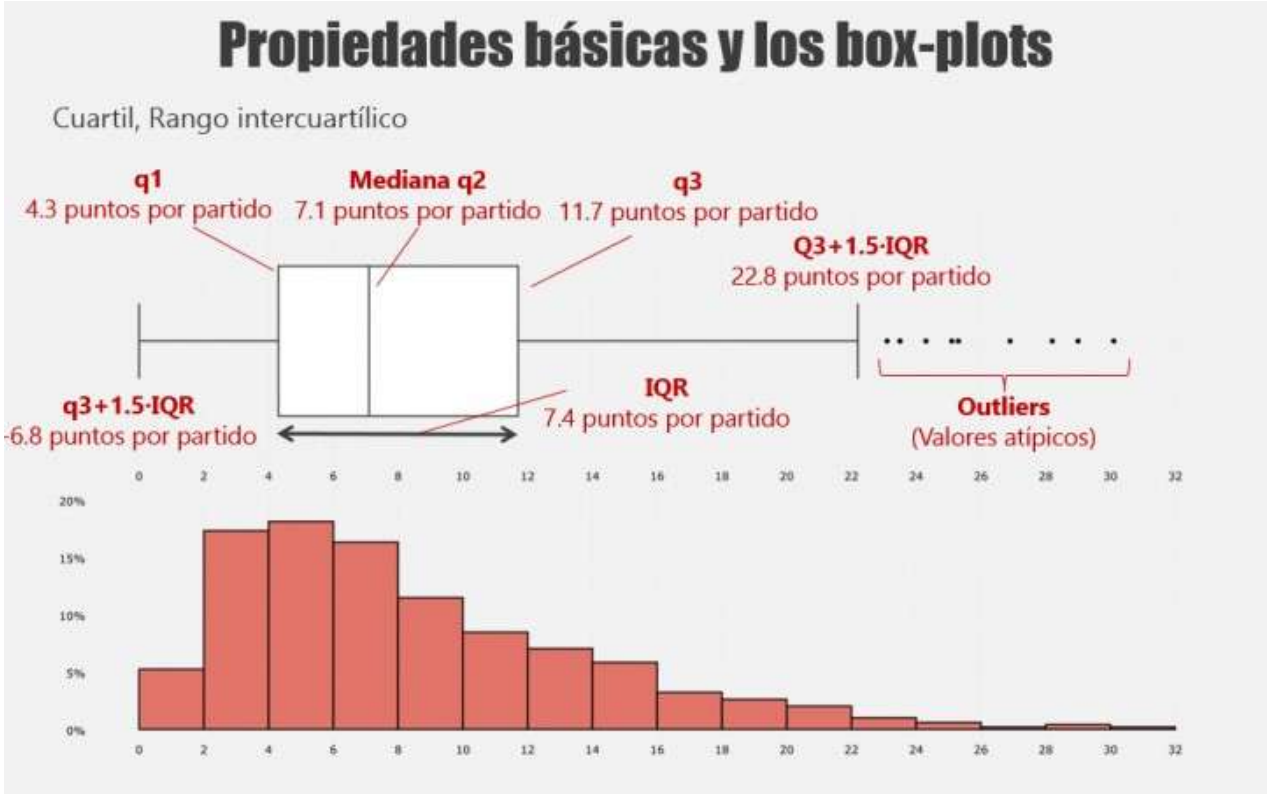
¡En un gráfico puedes comparar la media de las las categorías fácilmente.

¿Cómo construir el boxplot con excel?

El boxplot es una herramienta que te permite comparar grupos muy fácilmente. En el ejemplo tienes dos grupos: hombres y mujeres. Puedes comparar los dos grupos de un vistazo en un solo gráfico.

¿qué es un boxplot, whisker o gráfico de bigotes?

el boxplot se basa en el uso de los cuartiles. El cuartil 1, el cuartil 2 y el cuartil 3. La siguiente imagen ilustra muy bien qué es un boxplot.



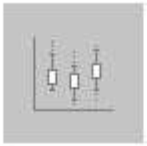
	A	B
1	Sexo	Peso [kg]
2	HOMBRE	72,72
3	HOMBRE	69,12
4	HOMBRE	65,26
5	MUJER	49,35
6	MUJER	61,84
7	HOMBRE	64,82
8	MUJER	72,07
9	HOMBRE	71,48
10	HOMBRE	82,92
11	MUJER	51,07
12	HOMBRE	90,47
13	HOMBRE	82,24
14	MUJER	47,70

De esta manera tienes seleccionada la variable categórica sexo (la primera columna) que corresponde al grupo. Y el variable numérico peso (la segunda columna). Que es la variable que quieres comparar. Sólo tienes que insertar >> gráfico >> ver todos los gráficos y te saldrá la ventana siguiente:

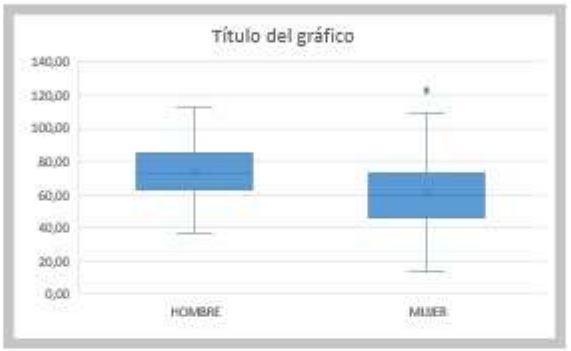
Insertar gráfico

Gráficos recomendados | Todos los gráficos

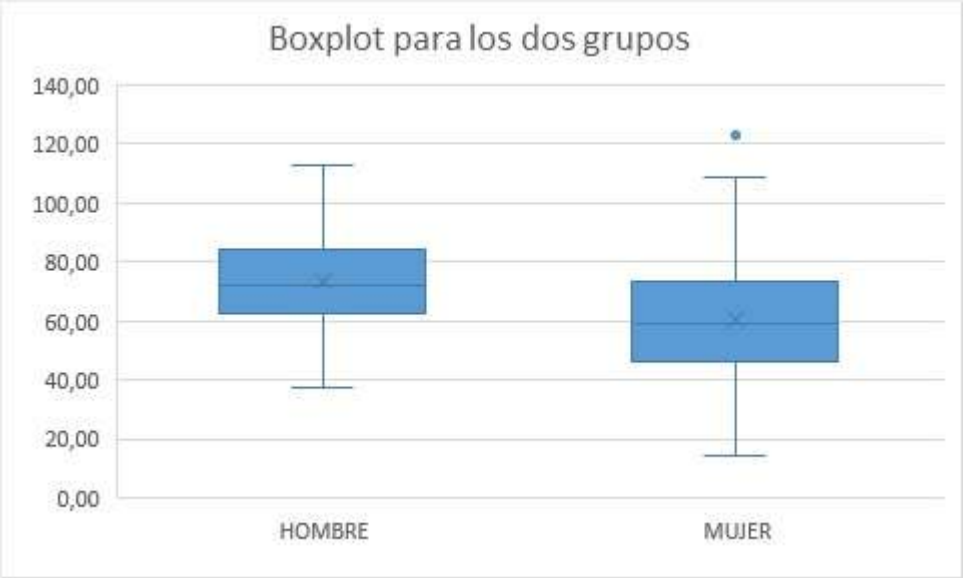
- Reciente
- Plantillas
- Columna
- Línea
- Circular
- Barra
- Área
- X Y (Dispersión)
- Cotizaciones
- Superficie
- Radial
- Gráfico de rectángulos
- Proyección solar
- Histograma
- Cajas y bigotes
- Cascada
- Cuadro combinado



Cajas y bigotes



Vas a la opción cajas y bigotes. Dando en aceptar tienes creado el gráfico de boxplot para los dos grupos:



Este gráfico es muy útil para comparar grupos muy rápidamente.

Condensa la dispersión y el valor central en un caja. En este caso puedes ver como el grupo de mujeres tiene un peso menor que el de hombres, pero no muy significativamente diferente.