ESTATÍSTICA

1. Estatística

A estatística é a ciencia que utiliza os números para o estudo das leis que dependen do azar, é dicir, os fenómenos aleatorios. Na estatística distinguimos dúas partes perfectamente diferenciadas. A primeira delas dedícase a recoller datos, ordenalos, simplificalos, clasificalos, obter deles un conxunto de valores que os identifican, facer comparación con outros conxuntos de datos e estudar relacións entre eles, esta parte coñécese como Estatística Descritiva, que é a que abordaremos ao principio da unidade. A outra parte da estatística denomínase Inferencia Estatística e ademais de resumir os datos preténdese predicir ou estimar o comportamento no futuro ou en xeral dunha poboación, esta abordarase na seguinte unidade.

2. Distribucións unidimensionais

2.1 Variables estatísticas

A estatística estuda unha característica ou carácter dun conxunto de individuos chamado poboación. Cando a poboación e moi grande recórrese a un subconxunto denominado mostra. A elección dunha mostra representativa da poboación é un proceso que pode ser complicado ás veces. Existen diferentes técnicas da mostraxe para asegurar a representatividade dos datos recollidos, por exemplo: mostraxe aleatoria simple, na que os individuos se numeran para ser elixido aleatoriamente, mostraxe estratificada, na que se divide a poboación en estratos e elíxese a mostra de modo que conteña individuos de cada un dos estratos,... Un individuo é cada un dos elementos da poboación ou mostra.

Nun estudio estatístico, temos que elixir a característica a estudar e a poboación ou mostra, que é un subconxunto representativo da poboación. Tendo en conta , que a elección dunha mostra representativa é un proceso que pode ser complicado, ás veces. Existen distintas técnicas de mostraxe para asegurar a representatividade da mostra: mostraxe simple, na que se enumeran os individuos para logo ser elixidos ao azar; mostraxe estratificada, na que se divide a poboación en estratos e elíxese a mostra de modo que conteña os mesmos individuos de cada estrato,... este contido o profundizaremos no seguinte curso. Unha vez elixida variable e mostra, o estudo o podemos resumir en catro fases de actuación:

- Recolección de datos.
- Ordenar ou organizar os datos.
- Analizar os datos.
- Interpretar os resultados.

Unha variable estatística unidimensional é a característica que se vai a estudar dos individuos dunha poboación. A variable X queda determinada polos datos $x_1, x_2, x_3, \ldots$

As variables pódense clasificar en :

- Cualitativas ou atributos: cando a característica que se vai estudar non se pode describir numericamente. Por exemplo, cor de ollos, partido político o que se vota, profesión,... Cada

unha das posibilidades que pode tomar a variable cualitativa se lle chama modalidade, no exemplo cor de ollos, as modalidades serían: verde, azul, negro e marrón.

- Cuantitativas: cando a característica que se estuda é de tipo numérico e poden ser:
  - Discretas: soamente toma valores enteiros ou valores illados en cada intervalo. Por exemplo, número de fillos dunha familia, número de alumnos dun centro escolar...
  - Continuas: a variable pode tomar tantos valores somo queiramos por pequeno que sexa o intervalo. Por exemplo, altura das árbores dun bosque, temperatura...

Cando o estudo faise dunha única variable da poboación, fálase de estatística unidimensional. Se interesa a relación existente entre dúas variables da mesma poboación utilizarase o que se chama estatística bidimensional.

Exemplo

Queremos estudar mediante unha enquisa, a intención de voto nunhas eleccións estatais.

Poboación: Españois con dereito a voto

Mostra: Eliximos unha mostra por exemplo 2000 persoas ao chou ou ben poderíamos facer unha estratificación por comunidades e elixir de todas un número fixo.

Variable estatística: intención de voto é unha variable cualitativa.

Táboas de distribución de frecuencias

Unha vez que recollemos os datos, a través dun cuestionario, observación de individuos, recollida automática mediante dispositivos informáticos coma un móbil ou un PC... deberemos organizalos. Hoxe en día adoitase usar o ordenador para organizar os datos, existen multitude de programas especializados no tratamento de datos, entre eles os máis sinxelos son as follas de cálculo. Nesta fase adoitase a "limpar os datos", é dicir, a eliminar posibles erros.

Para organizar os datos dunha variable estatística utilizamos as táboas de frecuencias, nelas ordenamos os datos na primeira columna e as frecuencias nas seguintes columnas.

- Frecuencia absoluta, f_i: número de veces que se repite un certo valor (ou conxunto de valores). Se as acumulamos obtemos o que se chama frecuencia absoluta acumulada, F_i.
- Frecuencia relativa, h_i: Calcúlase dividindo a frecuencia absoluta entre o número total de datos, se o multiplicamos por 100 transfórmase nunha porcentaxe, que nos da unha información moi fácil de valorar.

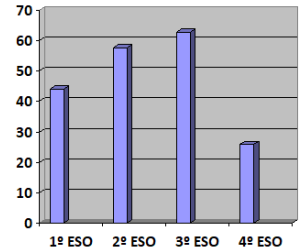
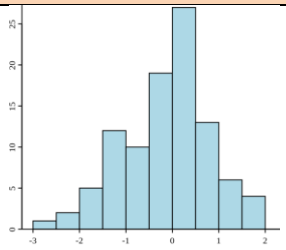
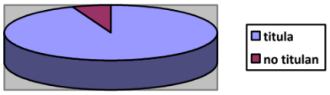
Cando o número de datos do estudo é moi grande ou a variable é continua, organizamos ditos datos en intervalos que denominaremos clases, a amplitude dos intervalos será en todos a mesma. O punto medio da cada intervalo será o valor que representa a todos, e se lle chama marca da clase. Este valor nun intervalo de extremos $[x_1, x_2]$ calcúlase como a semisuma dos extremos $\frac{x_1+x_2}{2}$.

A hora de elixir o número de intervalos nos que agruparemos os datos dunha mostra existen varias formas como:

- Fórmula de Sturgers: nº clases= 1+3,32 log N, sendo N o número de datos.
- Calculando a raíz cadrada do número de datos.
- De forma intuitiva.

2.2 Gráficos estatísticos

Para organizar e interpretar os datos obtidos nun estudo estatístico utilizamos, principalmente os seguintes gráficos estatísticos.

Diagrama de Barras	Histograma	Diagrama de Sectores
 *Ilustración 1: Diagrama de barras*	 *Ilustración 2: Histrograma*	 *Ilustración 3: Sectores*
No eixe de abscisas escríbense os datos da variable e no de ordenadas as frecuencias. As barras deben estar separadas.	Divídese o eixe de abscisas en intervalos e levántase un rectángulo en cada tramo con altura a súa frecuencia. Úsanse para variables agrupadas en intervalos.	O círculo divídese en tantos sectores como datos teña a variable, sendo a amplitude proporcional a frecuencia, e calcúlase cunha simple regra de tres.

Nos histogramas e diagramas de barras se unimos os puntos medios dos lados superiores dos rectángulos obtemos un novo gráfico que se chama poligonal de frecuencias.

O gráfico máis axeitado para representar os datos obtidos nun estudo estatístico dependerá da variable estudada:

- Cualitativa: diagrama de barras ou diagrama de sectores.
- Cuantitativa discreta: diagrama de barras, polígono de frecuencias ou diagrama de sectores.
- Cuantitativa continua: Histograma.

2.3 Parámetros estatísticos

Para resumir a información referente a os datos dunha variable estatística cuantitativa pódese calcular valores representativos do conxunto de datos denominados parámetros, vexamos os máis usuais.

Medidas de centralización

A media aritmética, $\overline{x}$, é o cociente da suma de todos os datos multiplicados pola súa frecuencia entre o número de datos. É a medida de centralización máis utilizada.

Media se os datos non están agrupados: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{N}$

Media para datos agrupados: $\overline{x} = \frac{\sum_{i=1}^{n} f_i \cdot x_i}{N}$

A moda, Mo, é o valor dos datos que aparece con maior frecuencia. Se a variable atópase por intervalos fálase de intervalo modal.

A mediana, Me, é o valor que ocupa a posición central dos datos, despois de ordenalos ou a media do datos centrais, se o numero é par.

Medidas de dispersión

As medidas de dispersión permiten coñecer o grao de agrupamento dos datos en torno as medidas de centralización.

Medida	Cálculo
Rango ou recorrido: é a diferenza entre o maior e o menor valor da variable. A información que proporciona é imprecisa, pois soamente ten en conta os valores extremos.	$R = x_{Máx - x_{Min}}$
Varianza: é a media do cadrados das desviacións. A varianza é sempre positiva, pero ten un pequeno inconveniente, non se expresa nas mesmas unidades que os datos (Ex: se os datos son cm a varianza é cm^2)	$\sigma^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \overline{x})^2}{N} = \frac{\sum_{i=1}^{n} f_i {x_i}^2}{N} - \overline{x}^2$
Desviación típica: é a raíz cadrada positiva da varianza. É o parámetro máis utilizado, ten as mesmas unidades que os datos, e estuda a desviación dos datos entorno a media.	$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \overline{x})^2}{N}}$
Coeficiente de variación: é o cociente da desviación típica e a media, adoitase a expresar en %. Mide a homoxeneidade do conxunto.	$CV = \frac{\sigma}{\overline{x}}$

Ao comparar a dispersións de dúas distribucións pode acontecer que teñan desviación moi similares pero que a dispersión sexa totalmente diferentes, non é o mesmo unha desviación de 2kg nunha poboación de polos que de vacas, para elo usamos o coeficiente de variación. O coeficiente de variación mide, en tantos por cen, a desviación dunha dispersión con respecto á media. Canto máis pequeno sexa o coeficiente de variación, os datos estarán máis concentrados arredor da media, e esta resultará máis significativa, canto maior sexa o coeficiente de variación o conxunto estará moito máis disperso.

2.4 Análise estatístico

As medidas estatísticas proporcionan maior información cando se analizan conxuntamente. Vexámolo a continuación.

Exemplo

As cualificacións de 85 alumnos nun exame son as da táboa seguinte, realiza un pequeno estudo:

Nota x_i	1	2	3	4	5	6	7	8	9	10
Nº alumnos f_i	2	3	6	10	22	15	12	7	5	3

Calculamos unha táboa cos datos precisos para calcular os parámetros estatísticos:

Nota x_i	1	2	3	4	5	6	7	8	9	10	sumas
Nº alumnos f_i	2	3	6	10	22	15	12	7	5	3	85
$x_i \cdot f_i$	2	6	18	40	110	90	84	56	45	30	481
${x_i}^2$	1	4	9	16	25	36	49	64	81	100	-
${x_i}^2 \cdot f_i$	2	12	54	160	550	540	588	448	405	300	3059

Media, datos agrupados: $\overline{x} = \frac{\sum_{i=1}^{n} f_i \cdot x_i}{N} = \frac{481}{85} \approx 5{,}66$

Mediana: 5

Moda: 5

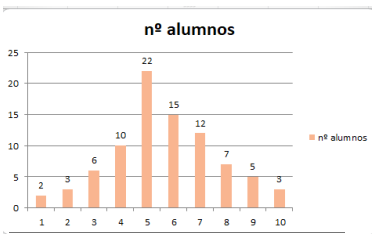


Ilustración 4: Diagrama de barras

R= $x_{Máx-x_{Min}} = 10 - 1 = 9$

Varianza: $\sigma^2 = \frac{\sum_{i=0}^{n} f_i {x_i}^2}{N} - \bar{x}^2 = \frac{3059}{85} - 5{,}66^2 = 3{,}95$

Desviación Típica: $\sigma = \sqrt{3{,}95} \approx 1{,}99$

Coeficiente de Variación: $CV = \frac{\sigma}{\bar{x}} = \frac{1{,}99}{5{,}66} = 0{,}352$; 35,2%

Podemos deducir que os datos presentan unha agrupación entorno a media, xa que os valores da desviación típica e o coeficiente non son relativamente elevados.

3. Estatística bidimensional

Unha variable estatística bidimensional é a que resulta ao estudar dúas características diferentes dos individuos dunha poboación, está formada por dúas variables estatísticas unidimensionais. A variable bidimensional (X,Y) queda determinada polos pares de datos $(x_1, y_1), (x_2, y_2)$ ...
Por exemplo, notas de Matemáticas e Física dos alumnos e alumnas de primeiro de bacharelato, metros cadrados da superficie dun local comercial e volume de vendas anual do mesmo... Para organizar os datos dunha variable estatística bidimensional, faise usando diferentes táboas: táboas simples e táboas de dobre entrada.
Pódese utilizar unha táboa simple se a frecuencia absoluta de cada par é un, é dicir, se todos os pares de números que forman a variable son únicos.

X	x_1	x_2	...	x_n
Y	y_1	y_2	...	x_n

Se algún dos pares se repite, pódese construír a táboa engadino outra fila ou columna, que indicaría a frecuencia absoluta de cada par, é dicir, cantas veces se repite cada par de números.

X	x_1	x_2	...	x_n
Y	y_1	y_2	...	x_n
f_i	f_1	f_2	...	f_n

Adoitase tamén a utilizar as táboas de dobre entrada, xa que resulta a forma máis clara de representar a frecuencia de cada par. Nas columnas colócanse os datos da variable X e, nas filas, os da variable Y, no interior da táboa anótanse as frecuencias absolutas de cada para (x,y).
A partires da táboa estatística da distribución bidimensional, pódese obter a súa representación gráfica. A forma máis usual de representar este tipo de distribucións e a que se chama diagrama de dispersión ou nube de puntos

Exemplo

Dada a seguinte distribución bidimensional, representa táboa de dobre entrada e o seu diagrama de de dispersión:

Matemáticas	2	3	3	4	4	6	6	7
Lingua	2	6	8	4	7	2	5	6
f_i	1	1	1	2	3	1	2	2

Táboa dobre entrada e diagrama de dispersión:

MAT / LING	2	3	4	6	7	Total
2	1	0	0	1	0	2
4	0	0	2	0	0	2
5	0	0	0	2	0	2
6	0	1	0	0	2	3
7	0	0	3	0	0	3
8	0	1	0	0	0	1
Total	1	2	5	3	2	13

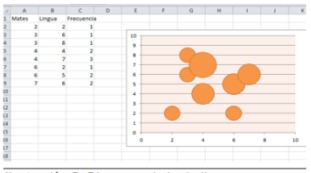


Ilustración 5: Diagrama de burbullas

O diagrama de dispersión está feito cunha folla de cálculo que veremos ao final da unidade.

Cando se estuda por separado as variables X e Y que forman unha variable bidimensional fálase de distribucións marxinais. Neste caso poderemos falar de táboas de frecuencias marxinais que se obteñen ao estudar por separado cada unha das variables e tamén de táboas de frecuencias condicionadas que se obteñen ao estudar unha variable bidimensional e impoñerlle unha condición sobre os valores da outra variable. Vexamos un exemplo.

Exemplo

O número de persoas por fogar, *X*, e o numero de coches de cada un *Y*, ven reflectido na seguinte táboa de dobre entrada.

X / Y	1	2	3	4	Frecuencias X
1	7	10	11	16	44
2	0	2	6	7	15
3	0	0	1	5	6
Frecuencias Y	7	12	18	28	65

a) Cantos fogares están formados por tres membros?
b) En cantos fogares hai 2 coches?
c) Cantos fogares están formados por 4 membros e teñen 2 coches?
d) Escribe a táboa por separado da variable X(persoas), isto é a táboa de frecuencias marxinais?
e) Escribe a marxinal de Y?
f) Determina a táboa de frecuencias para a variable Y condicionada a que o número de persoas no fogar sexan dúas.

a) Dos entrevistados, 18 fogares están formados por tres membros.
b) Hai 15 fogares nos que teñen 2 coches
c) Dos fogares de 4 membros en 7 deles teñen dous coches.
d) Táboa de frecuencias marxinais de X é:

X	1	2	3	4	Total
Frecuencia	7	12	18	28	65

e) Táboa de frecuencias marxinais de Y:

Y	1	2	3	Total
Frecuencia	44	15	16	65

f) Táboa de frecuencias condicionada:

Y/X=2	1	2	3	Total
Frecuencia	10	2	0	12

Parámetros estatísticos dunha distribución marxinal son os parámetros vistos no apartado de distribución unidimensionais.

- As medias marxinais son as medias das variables de *X e Y:*

  $\overline{x} = \frac{\sum f_i \cdot x_i}{N}$ $\quad \bar{y} = \frac{\sum f_i \cdot y_i}{N}$

- El centro de gravidade é o par de valores das medias marxinais: G$(\overline{x}, \overline{y})$

- As desviacións marxinais son as desviacións típicas das variables X e Y e denotamos coma:

  $$\sigma_x = \sqrt{\frac{\sum f_i \cdot {x_i}^2}{N} - \overline{x}^2} \sigma_y = \sqrt{\frac{\sum f_i \cdot {y_i}^2}{N} - \overline{y}^2}$$

- A covarianza duna variable bidimensional (X,Y) é :

  $$\sigma_{xy} = \frac{\sum f_{ij} x_i y_i}{N} - \overline{xy}$$

  A covarianza é un valor que indica o grao de variación conxunta de dúas variables aleatorias. O signo da covarianza apórtanos a seguinte información:
  - Covarianza positiva: ao aumentar a variable X aumenta os valores da variable Y. A nube de puntos orientase a dereita e cara arriba.
  - Covarianza negativa: ao aumentar os valores da variable X diminúen os valores da variable Y. A nube de puntos orientase a dereita e cara abaixo.

Exemplo

Calcula a covarianza destes datos:

X / Y	1	2	3	4	Total
1	7	10	11	16	44
2	0	2	6	7	15
3	0	0	1	5	6
Total	7	12	18	28	65

Primeiro: Calculamos as medias marxinais $\bar{x} = \frac{\sum f_i \cdot x_i}{N} = 3{,}031$; $\bar{y} = \frac{\sum f_i \cdot y_i}{N} = 1{,}415$

Segundo: Calcúlase a covarianza cos datos de cada variable e as súas frecuencias:

$$\sigma_{xy} = \frac{\sum f_{ij} x_i y_i}{N} - \bar{x}\bar{y} = \frac{293}{65} - 3{,}032 \cdot 1{,}415 = 0{,}219$$

Como Covarianza é positiva quérenos dicir que ao aumentar a variable x aumenta a variable y, a nube de puntos oriéntase cara arriba.

3.1 Táboas de contixencia

Nos apartados anteriores centráronse en variables cuantitativas. Vexamos que ocorre cando as variables son cualitativas ou atributos, xa que os datos que temos xa non son cantidades. Os atributos non son susceptibles dunha valoración numérica, a análise da mesma baséase no reconto que se realiza de cada unha das súas modalidades. Deste xeito elabóranse táboas nas que a cada modalidade asígnaselle a súa frecuencia absoluta.

Se consideramos unha soa variable colócanse dúas columnas unha coas modalidades e outra coas frecuencias, a este tipo de táboas se lle chama táboas de contixencia. No caso de dúas variables recolleremos os valores nunha táboa de dobre entrada.

Exemplo 1

Preguntamos a 10 persoas a súa cor favorita:

Cor	Frecuencia
vermello	4
verde	3
azul	3

Exemplo 2

Preguntamos a 10 persoas cal é a súa cor favorita e súa comida:

Comida / Cor	Tortilla	Pizza	Churrasco	Total
vermello	2	1	1	4
verde	0	2	1	3
azul	1	0	2	3
Total	3	3	4	10

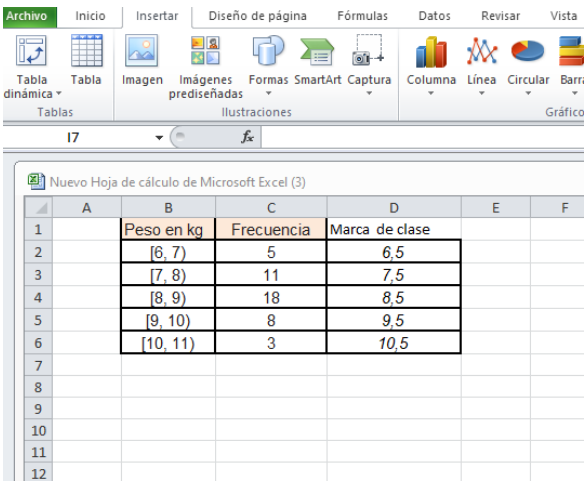
4. Recursos informáticos

Como xa mencionamos anteriormente existen diferentes recursos que se poden utilizar para facer menos laboriosa as tarefas de cálculo de parámetros e confeccións de gráficos.

Para o cálculo de parámetros pódese utilizar a calculadora científica que ten unha serie de teclas que nos axudan a calcular a media e a a desviación típica, deberemos poñer a calculadora en ”modo estatístico” e isto proceso dependerá da calculadora coa que traballemos.

Vexamos con máis precisión como se constrúen gráficos coa axuda da folla de cálculo Excel:

1. Deberemos introducir a táboa dos datos que queremos representar. (imaxe 6)
2. Para facer o histograma deberemos ir a inserir gráfico de columnas e logo inserir datos: Pinchar en Agregar: poñeremos “Histograma de frecuencias” e en valores inserimos os nosos datos, isto é as nosas frecuencias. (imaxe 7 e 8)
3. Aceptar e temos un diagrama de barras. Editamos datos para que aparezan as marcas da clases. (imaxe 9 e 10)
4. Volvemos a agregar agora “Polígono de frecuencias” e de novo as frecuencias. Aparecen dúas columnas.(imaxe 11)
5. Seleccionando nas últimas columnas inserimos liña poligonal. (imaxe 12)
6. Por último observamos que ao estar nunha variable continua os intervalos son seguidos e non deberá existir espazo entre as barras o eliminamos, pinchando sobre as barras desprégase “opcións de serie” “ancho de intervalo” e o eliminamos. (imaxe 13 e 14)



	A	B	C	D	E	F
1		Peso en kg	Frecuencia	Marca de clase		
2		[6, 7)	5	6,5		
3		[7, 8)	11	7,5		
4		[8, 9)	18	8,5		
5		[9, 10)	8	9,5		
6		[10, 11)	3	10,5		
7						
8						
9						
10						
11						
12						

Ilustración 6: Táboas de frecuencias

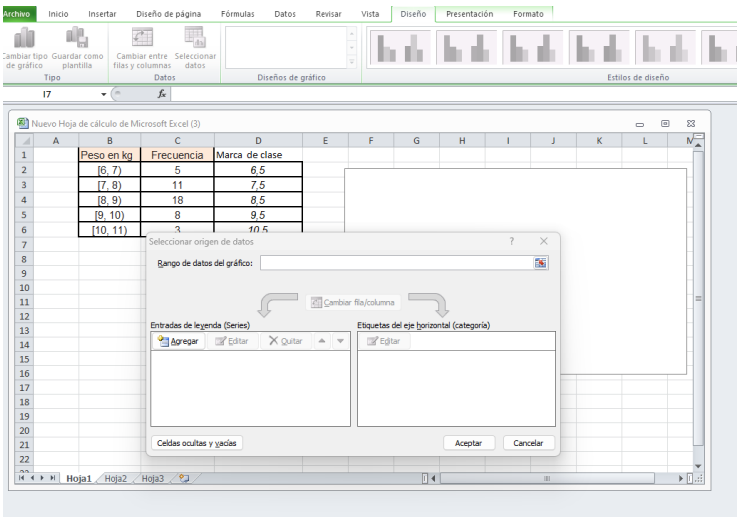


Ilustración 7:Agregar datos

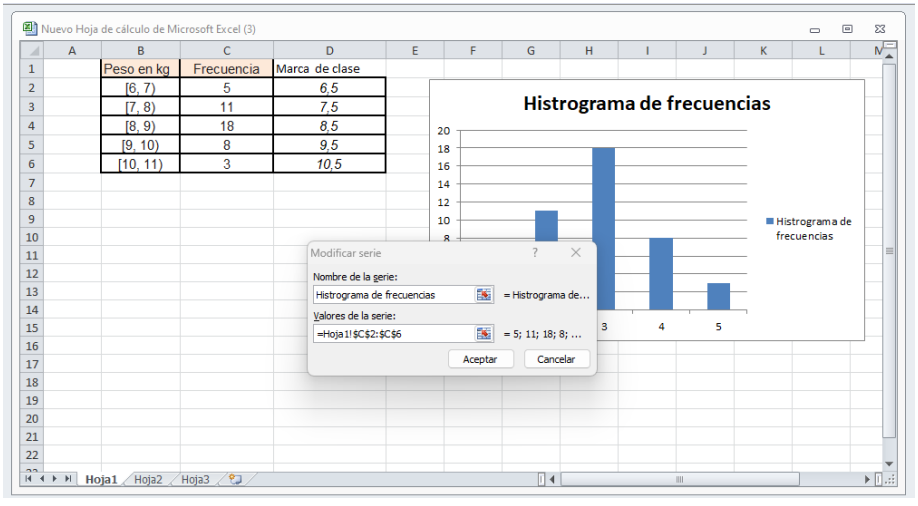


Ilustración 8: Agregar frecuencias

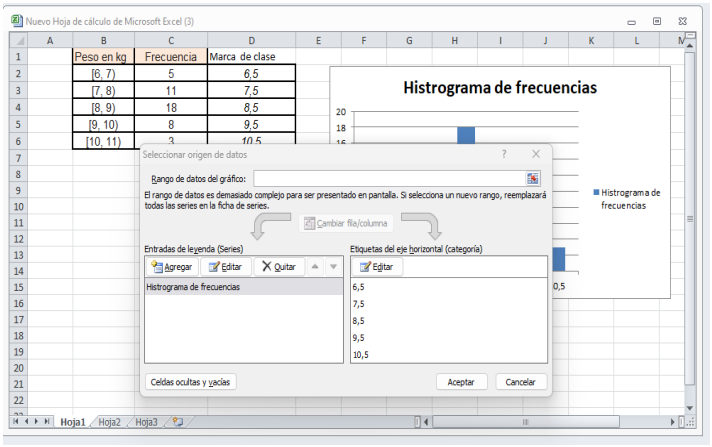


Ilustración 9: Editar marcas

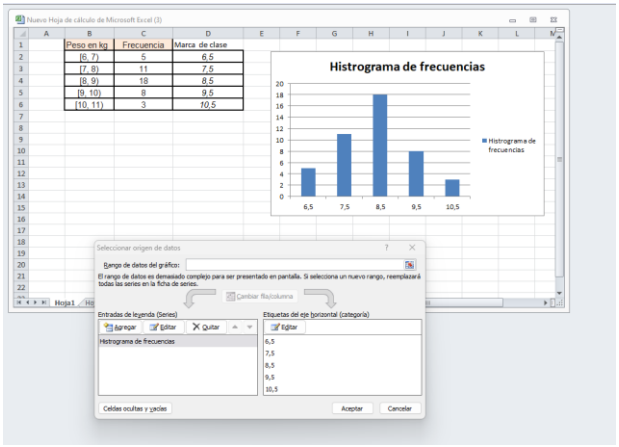


Ilustración 10: Agregar Datos 2

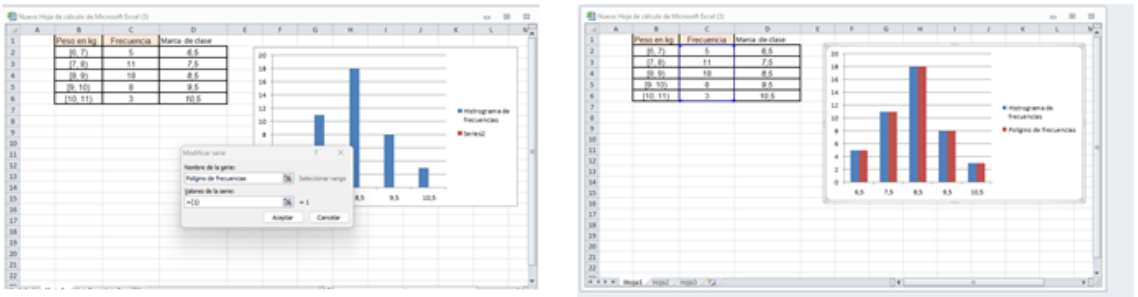


Ilustración 11: Polígono de frecuencias

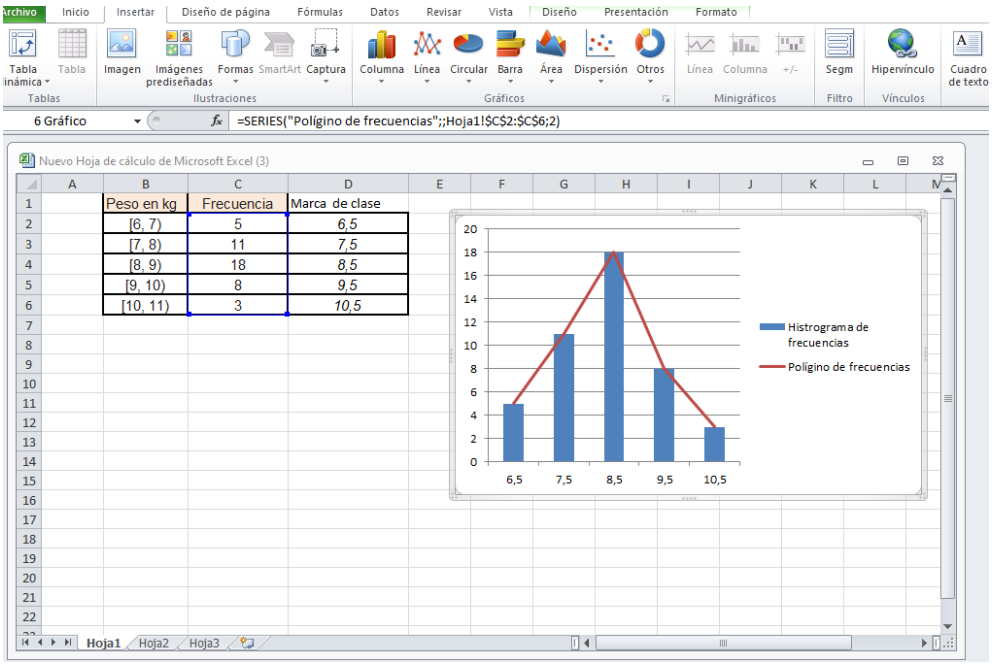


Ilustración 12: Polígono de frecuencias 2

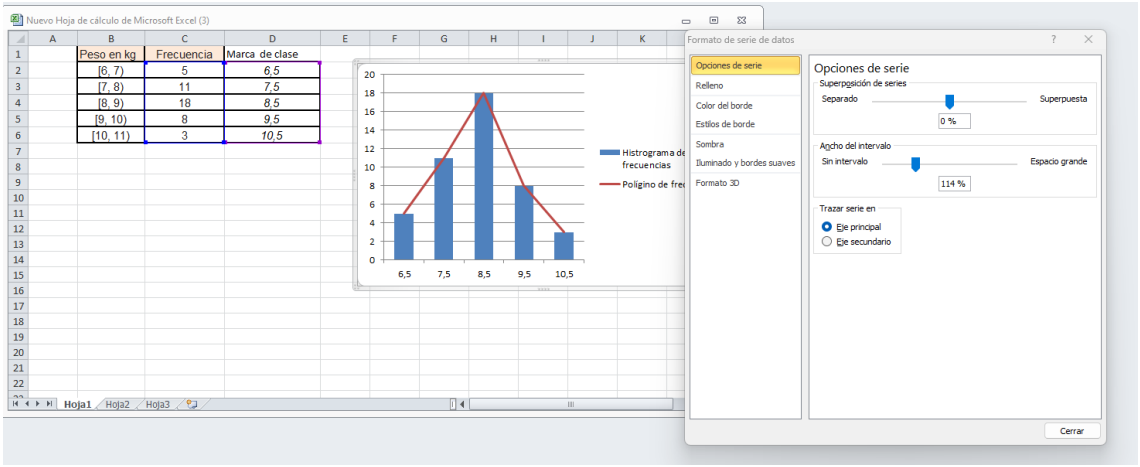


Ilustración 13: Eliminar os espazos

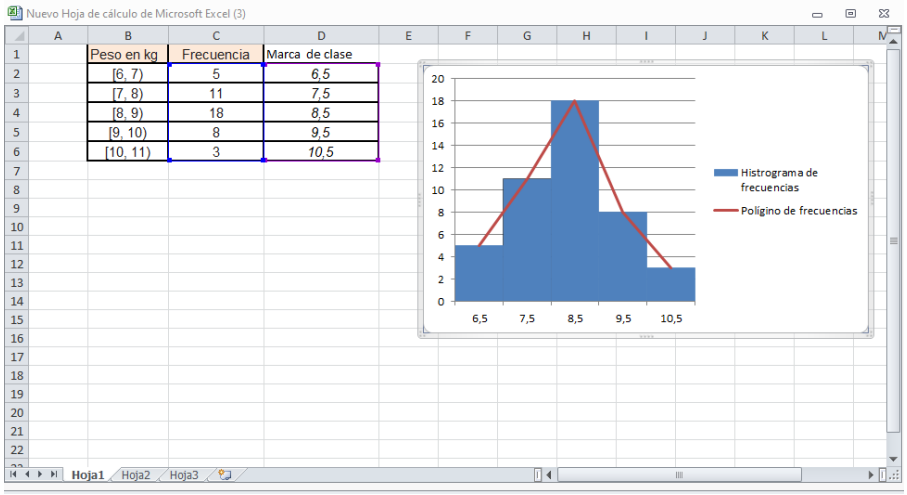


Ilustración 14: Historgrama

Licenzas das ilustracións

Ilustración	Recurso
Ilustración 1. Diagrama de barras	Autoría: Elaboración Propia
Ilustración 2. Histograma	Autoría: Elaboración Propia
Ilustración 3. Sectores	Autoría: Elaboración Propia
Ilustración 4. Diagrama de barras	Autoría: Elaboración Propia
Ilustración 5. Diagrama de burbullas	Autoría: Elaboración Propia
Ilustración 6: Táboas de frecuencias	Autoría: Elaboración Propia
Ilustración 7: Agregar datos	Autoría: Elaboración Propia
Ilustración 8: Agregar frecuencias	Autoría: Elaboración Propia
Ilustración 9: Editar marcas	Autoría: Elaboración Propia
Ilustración 10: Agregar Datos 2	Autoría: Elaboración Propia
Ilustración 11: Polígono de frecuencias	Autoría: Elaboración Propia
Ilustración 12: Polígono de frecuencias 2	Autoría: Elaboración Propia
Ilustración 13: Eliminar os espazos	Autoría: Elaboración Propia
Ilustración 14: Historgrama	Autoría: Elaboración Propia