MATEMÁTICA Y ESTADÍSTICA

Unidad 3

Estadística descriptiva

Variables cualitativas y cuantitativas. Población, muestra, aleatoriedad, inferencia.
interpretación de datos. Frecuencias absoluta, relativa y acumulada.
.

Roberto Fiadone

UNIDAD 3. ESTADISTICA

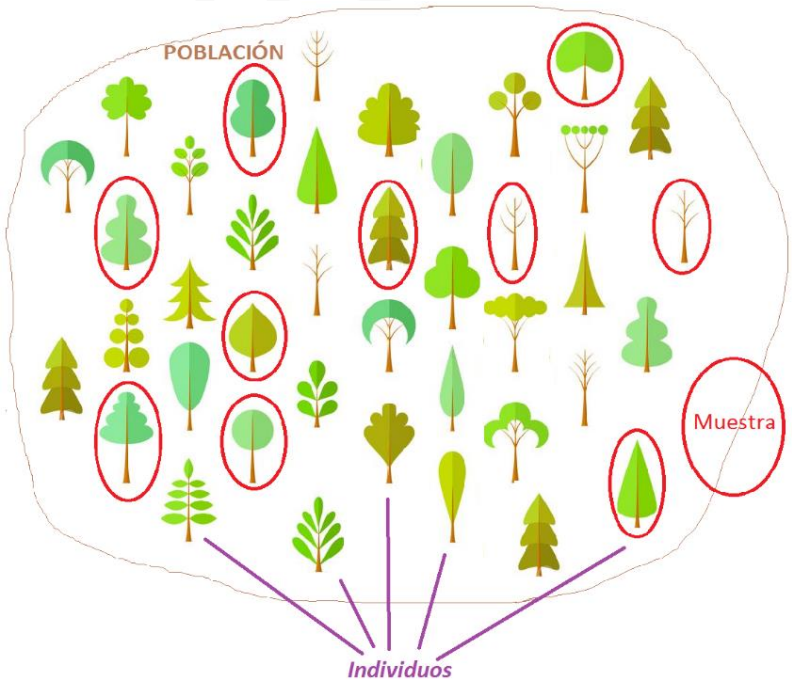
Por estadística entendemos el conjunto de métodos por medio de los cuales podemos recolectar, organizar, resumir y presentar datos relativos a un conjunto de observaciones para así obtener y comunicar conclusiones. Todo ello forma parte de la *estadística descriptiva*.

Además, la estadística colabora en la toma de decisiones y la predicción de situaciones. Es lo que se llama *estadística inferencial.*

Los datos que se analizan se corresponden a ciertas características de una ***población*** o conjunto de ***individuos*** que se está investigando y que **no** necesariamente son personas (pueden ser animales, plantas, tornillos, publicaciones, etc.). Si, como es frecuente, es imposible obtener los datos de toda la población en estudio, entonces se toma un subconjunto de ésta (generalmente de forma aleatoria). A este subconjunto se lo denomina ***muestra***.

El ***tamaño*** (o sea, el número de individuos) que debe tener una muestra para que sea representativa de la población, cómo seleccionar la muestra, etc.; son contenidos que exceden los contenidos de este curso. En lo que sigue, supondremos que las muestras son efectivamente representativas de la población, es decir, que podemos suponer que cualquier conclusión que hagamos sobre las muestras es extrapolable a la población en estudio.

El siguiente gráfico describe en forma sencilla lo dicho anteriormente para una muestra de árboles de un bosque en estudio.



Individuo: Es la unidad en estudio, y no necesariamente son personas (pueden ser animales, familias, autos, casas, etc).

Población: Conjunto de individuos sobre el que interesa obtener conclusiones, inferir algo. Generalmente es demasiado grande como para acceder a todos sus elementos.

Muestra: Es el subconjunto de individuos de la población que se seleccionaron para el análisis. A la cantidad de individuos de la muestra se lo llama el *tamaño* de la muestra, y lo simbolizaremos con la letra N.

Cuando los individuos son elegidos al azar se dice que la muestra es *aleatoria*.

VARIABLES

Las características o propiedades de interés relacionadas con cada individuo que se estudia se denominan ***variables,*** ya que dichas características varían de un individuo a otro. La variable puede tomar distintos *valores* (dónde, como veremos, estos valores no tienen por qué ser necesariamente un número) al efectuarse mediciones en cada individuo. Por ejemplo:

- edad de los habitantes de una ciudad.
- color de los peces de un lago
- preferencia por un club de futbol de los alumnos de una escuela
- altura de los edificios de una ciudad

son ejemplos variables que podemos proponernos estudiar.

A efectos de su posterior análisis, se consideran distintos tipos de variables, según puedan tomar valores numéricos o no. De acuerdo con ello, las variables pueden ser:

- *Cualitativas:* representan cualidades. Los valores que toman no pueden expresarse mediante números (sexo, nacionalidad, preferencia por un equipo de futbol, etc.).
- *Cuantitativas:* Los valores que toman pueden expresarse mediante números (temperatura, salario, número de hijos).

A su vez, las variables ***cuantitativas*** se clasifican en dos clases:

***Variables discretas*:** Sólo toman **valores aislados**, es decir, no pueden tomar ningún valor entre dos consecutivos. Generalmente son aquellas variables que consisten en **contar** algo

Ejemplos de variables discretas: Cantidad de frutos en un árbol (0,1,2,3,4…), cantidad de hijos de una familia (0,1,2,3 etc.), resultados al lanzar un dado (1, 2, 3, 4, 5, 6).
O sea, comúnmente es cualquier variable que solo toma valores aislados, generalmente enteros. Nunca vamos a obtener valores intermedios absurdos del tipo "familia con 2,4 hijos", "árbol con 7,82 peras", etc.

***Variables continuas*:** Pueden tomar **cualquier valor perteneciente al conjunto de los números reales dentro de un intervalo** continuo. Suelen ser aquellas variables que **miden** algo.

Ejemplo de variables continuas: Medida en metros de la altura de las personas, peso en kg de bolsas de arena, temperatura en grados Celsius, etc. Estas variables, a diferencia de las discretas, sí pueden tomar un rango infinito de valores dentro de un intervalo, es decir, los valores que pueden tomar pueden ser **cualquier número real** y, por lo tanto, pueden incluir todo tipo de fracciones o decimales. Dicho de otra manera, no toman valores aislados, entre dos valores cualquiera (por ejemplo, altura de 1,71 m y de 1,72 m) puede haber otros infinitos valores más (altura de 1,715 m, de 1,718 m, etc).

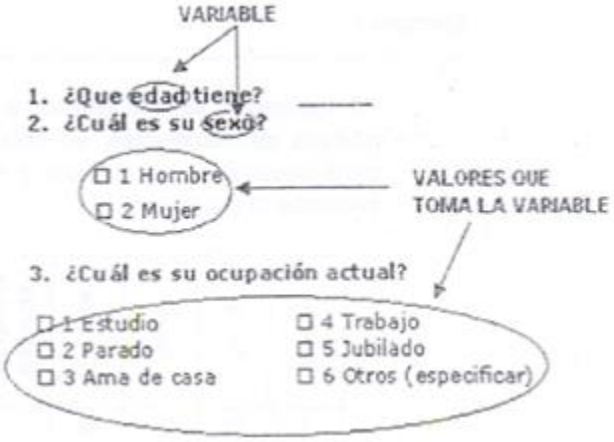
$$Variable \begin{cases} Cualitativa\ (no\ numérica) \\ Cuantitativa\ (numérica) \begin{cases} Discreta\ (valores\ aislados) \\ Continua\ (intervalos) \end{cases} \end{cases}$$

DATOS

RECOLECCIÓN: Para cada individuo se registra el valor que toma la característica observada, y este valor observado representa un ***dato***.
Si, por ejemplo, mi variable es "estatura" y una persona mide 1,70 m, este es un dato que difiere del de otra que mida 1,68 m. El conjunto de datos recogidos de cada individuo constituirá la base para el análisis estadístico de la variable estatura.

En una población o muestra se puede estudiar más de una variable.

Muchas veces los datos para un estudio estadístico se recogen a través de encuestas o cuestionarios, en las que el investigador traduce en preguntas las características variables que le interesa estudiar.



Cuando se realiza un ***censo***, se está encuestando a **toda** la población. En cambio, una encuesta **solo abarca un subconjunto de la población**, no se puede decir que se trate de un censo, es solo una muestra.

ORGANIZACIÓN Y RESUMEN: Una vez terminada la etapa de recolección de datos, comienza la etapa de organización de éstos para poder resumirlos e interpretarlos. Una forma posible es la de organizarlos en una tabla o matriz de datos, en donde se vuelca toda la información.

Por ejemplo, si se está llevando a cabo un estudio sobre una muestra de estudiantes en donde se les pregunta, entre otras cuestiones, edad, sexo y si trabaja o no, se puede presentar la siguiente lista:

Orden	**Edad**	**Sexo**	**Trabaja**	**Materias aprobadas**	**Especialidad**
1	23	F	SI	4	Humanidades
2	40	M	SI	8	Exactas
3	35	M	SI	10	Económicas
4	26	F	NO	5	Humanidades
5	24	F	SI	6	Económicas
6	33	F	NO	5	Económicas
7	35	M	SI	1	Económicas
8	39	M	SI	3	Exactas
9	25	M	NO	4	Económicas
10	22	M	NO	6	Humanidades
11	23	M	SI	10	Económicas
12	26	F	SI	8	Económicas
13	28	F	NO	2	Exactas
14	26	M	NO	7	Económicas
15	24	M	NO	4	Económicas

Cada columna de la lista recoge los datos de cada uno de los individuos observados

Variables en estudio

Orden	**Edad**	**Sexo**	**Trabaja**	**Materias aprobadas**	**Especialidad**
1	23	F	SI	4	Humanidades
2	40	M	SI	8	Exactas
3	35	M	SI	10	Económicas
4	26	F	NO	5	Humanidades
5	24	F	SI	6	Económicas

Respuestas de cada individuo (datos)

Aun habiendo organizado la información en una lista como la anterior, su interpretación puede ser dificultosa. Vamos a reorganizarla de modo tal que se facilite su lectura y análisis.

El primer paso consiste en separar cada variable y elaborar una tabla de dos columnas. En la primera se colocan la categorías o valores que toma cada variable. En la segunda se contabilizan el número de veces que se repite dicho valor de variable.

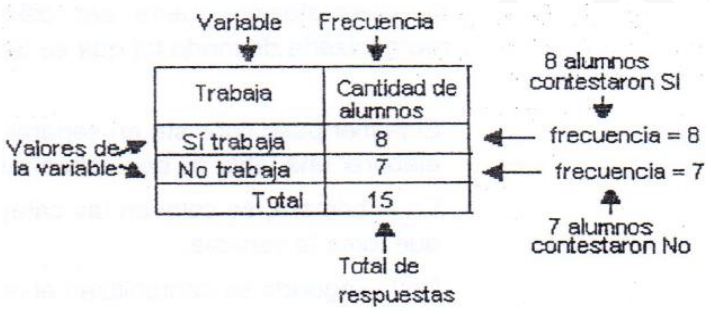
En el ejemplo, para la variable "*trabaja*", queda una tabla como la siguiente:

Trabaja	Cantidad de alumnos
Sí	8
No	7
Total	**15**

Una tabla como la anterior se denomina *tabla de frecuencias*.

> Una **tabla de frecuencias** es un cuadro que asocia a cada valor de la variable el número de veces (frecuencia) que se repite dicho valor.

Componentes de la tabla



De acuerdo con lo que hemos visto, los datos obtenidos a partir de un relevamiento de datos se ordenan en forma de tablas llamadas tablas de distribución de frecuencias.
La primera columna de una tabla está formada por los valores que toma la variable.
La segunda columna por la cantidad de veces que se registró cada uno de los datos obtenidos. A esta cantidad de veces se la denomina **frecuencia absoluta** (**f**) correspondiente a dicho valor.

Diario	Cantidad de lectores
Clarín	1.711.770
La Nación	312.670
Crónica	352.140
Olé	288.190
Popular	264.310
Total	**2.929.080**

También podemos expresar esas frecuencias como una proporción respecto al total de observaciones; utilizando las **frecuencias relativas** (**fr**) que se obtienen dividiendo la frecuencia absoluta por el total de observaciones. O mediante las **frecuencias porcentuales** (**f%**), que se obtienen multiplicando las frecuencias relativas por 1OO.

$$f_r = frecuencia\ relativa = \frac{frecuencia\ absoluta}{total\ de\ observaciones} = \frac{f}{N}$$

$$f_{\%} = frecuencia\ porcentual = frecuencia\ relativa \cdot 100 = f_r \cdot 100$$

Como ejemplo, mostramos la tabla anterior con las distintas frecuencias correspondientes.

Diario	Frecuencia Absoluta (f)	Frecuencia Relativa (f_r)	Frecuencia Porcentual ($f_{\%}$)
Clarín	1.711.770	0,58	58%
La Nación	312.670	0,11	11%
Crónica	352.140	0,12	12%
Olé	288.190	0,10	10%
Popular	264.310	0,09	9%
Total	**2.929.080**	**1,00**	**100%**

Por último, cuando la variable en estudio es numérica, puede aportar mucho ordenar los valores que pueda tomar de menor a mayor y calcular la llamada **frecuencia acumulada**. Esta va sumando todas las frecuencias absolutas hasta el valor en consideración, es decir, es un subtotal formado por la suma de todas las frecuencias absolutas menores a un determinado valor. Por ejemplo, en el siguiente cuadro se anotó la información sobre los ingresos de 110 trabajadores de una empresa.

Ingresos	Frecuencias
1800	10
1900	23
2000	25
2100	31
2200	21
Total	**110**

Las **frecuencias acumuladas** (F) se obtienen sumando todas las frecuencias absolutas **hasta** el valor de la variable en consideración.

Ingresos	f	F
1800	10	**10**
1900	23	(10+23=) **33**
2000	25	(10+23+25 = 33+25=) **58**
2100	31	(10+23+25+31 = 58+31=) **89**
2200	21	(10+23+25+31+21 = 89+21=) **110**
Total	**110**	

Así, por ejemplo, el “33” en la fila del 1900 nos dice que hay 33 sueldos cuyos valores son de 1800 o de 1900. Y el “89” en la fila del 2100 nos dice que hay 89 trabajadores cuyos sueldos son menores o a lo sumo iguales a 2100.

Es útil también agregar la **frecuencia acumulada relativa**,

$$F_r = \frac{F}{N}$$

o sea, el cociente de cada frecuencia absoluta por la cantidad total N de individuos.
O mejor aún, la **frecuencia acumulada relativa porcentual**

$$F_{\%} = F_r \cdot 100\%$$

En nuestro ejemplo, ($N = 110$)

Ingresos	f	F	$F_r = \frac{F}{110}$	$F_{\%} = F_r \cdot 100\%$
1800	10	10	$\frac{10}{110} = 0,09$	$0,09 \cdot 100 = 9\%$
1900	23	33	$0,3$	30%
2000	25	58	$0,52$	52%
2100	31	89	$0,89$	89%
2200	21	110	1	100%
Total	**110**			

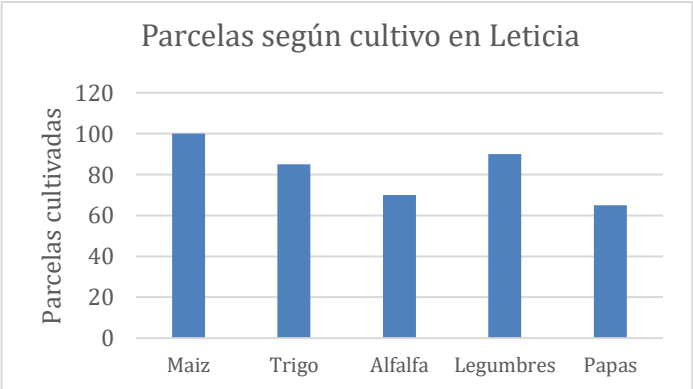
Y así, en la última columna, podemos ver que el 9 % de los trabajadores ganan $1800, el 30% ganan 1900 o menos, el 52% ganan $2000 o menos, etc.

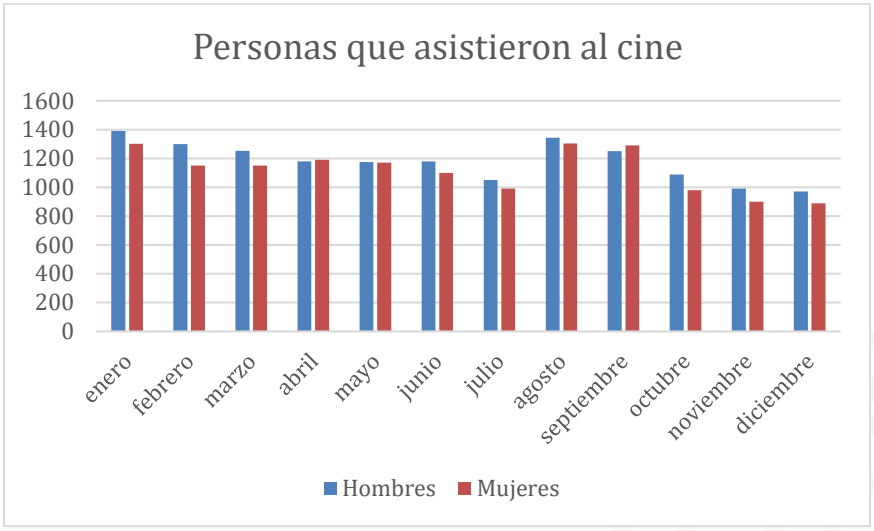
Representación gráfica de los datos

Para facilitar el análisis de un conjunto de datos se pueden utilizar distintos tipos de representaciones gráficas que describan las características o propiedades de las variables en estudio. Veremos algunas de la más clásicas.

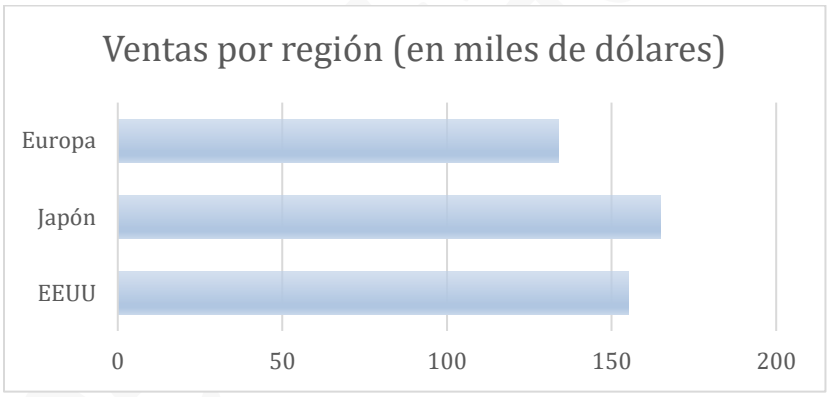
- Gráficos de barras verticales: nos permiten comparar variables cuyos valores son representados por rectángulos de igual base y altura proporcional al valor que representan.

Parcelas por cultivo en Leticia	
Cultivos	Parcelas
Maíz	100
Trigo	85
Alfalfa	70
Legumbres	90
Papas	65
	410





- Gráfico de barras horizontales: similar al anterior, la longitud de cada barra es proporcional a la cantidad a representar y pero las barras van de izquierda a derecha.



- Gráficos circulares (también llamados "de torta" ó "pastel"): permiten ver la distribución interna de los datos, generalmente en forma de porcentaje sobre un total.

Consideremos el ejemplo visto antes, donde están representadas la cantidad de parcelas que, en la localidad de "Leticia", le asignan a cada cultivo:

Cultivos	Parcelas	f_r	$f_{\%}$	f_r x 360
Maíz	100	0,24	24%	88
Trigo	85	0,21	21%	75
Alfalfa	70	0,17	17%	61
Legumbres	90	0,22	22%	79
Papas	65	0,16	16%	57
	410	**1,00**	**100%**	**360**

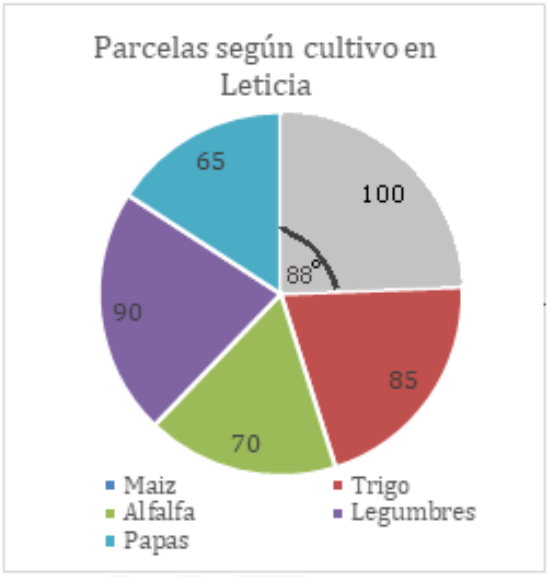
¿De qué manera se construye el gráfico circular correspondiente?

Como, por ejemplo, hay 100 parcelas de maíz, y esas 100 parcelas representan el 24% de las 440 parcelas que hay en total en Leticia, entonces a "maíz" le corresponderá un 24% del total del "pastel".

Y como en total una circunferencia mide 360°, y el 24% de 360° es igual a

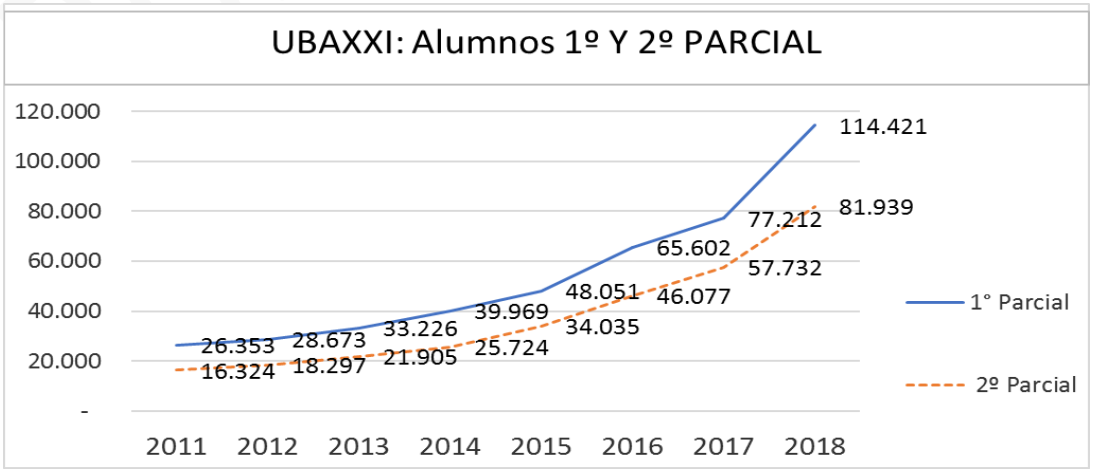
$$\frac{24}{100}360° = 0{,}24 \cdot 360° = 88°$$

entonces le asignamos a las 100 parcelas de maíz el trozo de pastel cuyo ángulo mide 88°. Y luego repetimos este procedimiento para cada uno de los cultivos, para así saber que porción de la "torta" le corresponde a cada cultivo.



Observemos en este ejemplo que si no aparecieran en cada porción del pastel los rótulos indicando la cantidad de parcelas correspondientes a cada sector, sería difícil, por ejemplo, darse idea de si hay más parcelas cultivadas con maíz que de legumbres, pues las porciones de pastel asignadas son casi iguales. Por eso, este tipo de gráficas es desaconsejable: para el ojo no es sencillo comparar dos áreas bidimensionales o dos ángulos, por lo que podríamos decir que, en la mayoría de las ocasiones, la información mostrada con ellos se representaría más adecuadamente con gráficas más simples de barras o columnas, o incluso simplemente con una serie de datos tabulados. La única ventaja que podemos encontrar en esta representación es que cuando alguien la observa, inmediatamente entiende que está viendo partes de **un todo**.

- <u>Gráfico de líneas:</u> En este tipo de gráficos se representan los valores de los datos en dos ejes perpendiculares entre sí, donde el eje horizontal suele representar el tiempo. Los distintos valores se unen mediante una recta para dar una idea de que los datos van variando en el tiempo de manera continua.



Ejemplo: Un estudio de mercado

Supongamos que una empresa textil, dedicada a la fabricación para mujeres de pantalones, camisas y camperas de "jean", decide renovar la presentación de sus productos. Para ello, realiza un estudio de mercado con el fin de conocer determinadas características de las consumidoras.
Hasta el momento, los productos que comercializa están dirigidos a mujeres de entre 18 y 28 años, y se presentan en una escasa variedad de modelos.
Las innovaciones consisten en ofrecer una línea de productos con diseños para distintas edades, y en confeccionar los pantalones en tres tonalidades de color diferentes para cada talle.
Para conocer las preferencias de los consumidores, se realizó una encuesta a 30 mujeres que usan este tipo de prendas de vestir, de entre 18 y 28 años.
Se analizan los siguientes aspectos: tonalidades preferidas de "jean", cantidad de prendas de "jean" compradas por cada consumidor el último año, y estatura.

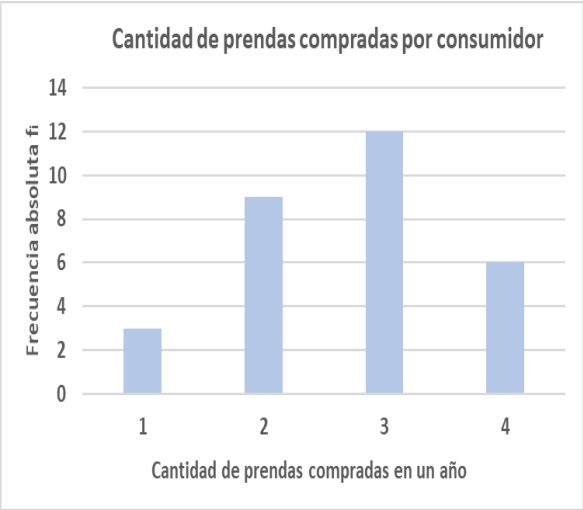
a) Variable "*tonalidades preferidas de jean*": Se efectúa un análisis de la preferencia de la tonalidad de las prendas, con el propósito de rediseñar modelos, teniendo en cuenta el gusto del público al cual va dirigido el producto. Los datos obtenidos se ordenan en tablas de distribución de frecuencias, como se muestra a continuación:

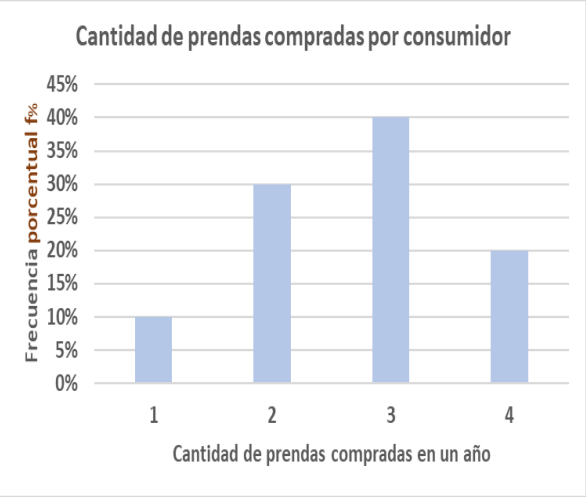
Mujeres			
Tonalidades de jean	f	f_r	$f_{\%}$
Claro	3	0,10	10
Mediano	17	0,57	57
Oscuro	10	0,33	33
Total	**30**	**1**	**100**

b) Variable "*cantidad de prendas de jean compradas por cada consumidora el último año*": Con la finalidad de poder estimar el valor de compras de esta clase de prendas se les preguntó a las **30** encuestadas cuantas prendas de este tipo habían adquirido en el último año.

Cantidad prendas compradas por consumidor en un año (x_i)	f	f_r	$f_{\%}$
1	3	0,10	10
2	9	0,30	30
3	12	0,40	42
4	6	0,20	18
Totales	**30**	**1**	**100**

Los resultados obtenidos se presentan en los siguientes diagramas de barras (versus frecuencia absoluta y versus frecuencia porcentual):





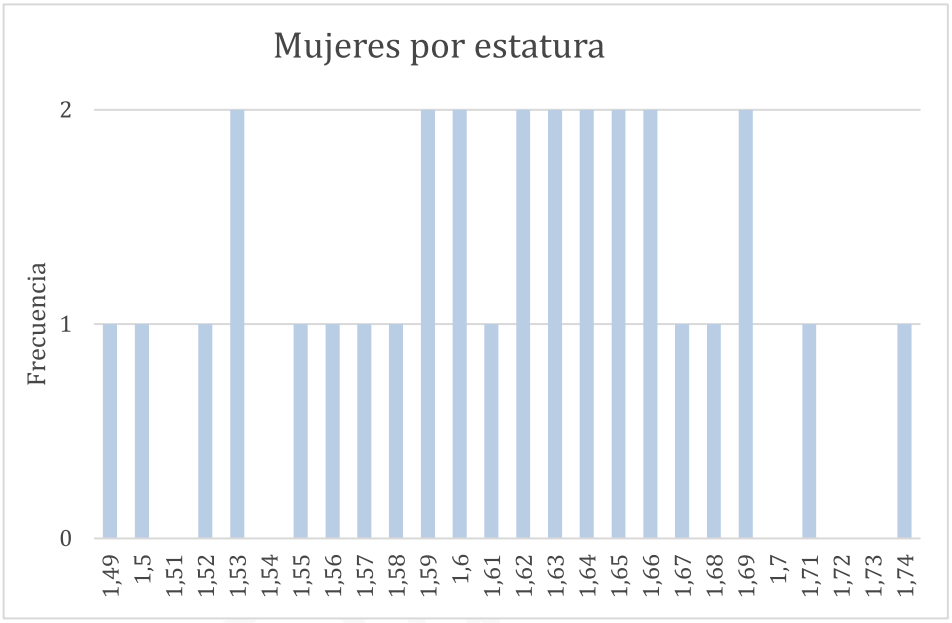
c) Variable "*estatura de las mujeres*": Se analizó esta característica con el fin de poder determinar tres largos diferentes de pantalones, para cada talle, en la línea mujer.

Cuando el número de observaciones es grande, es conveniente agrupar los datos por categorías denominadas ***intervalos de clase*****.** Para explicar de que se tratan estos intervalos y como se procede, vamos a utilizar los **30 datos** de estatura (en metros) de las mujeres obtenidos en la encuesta:

1,53	1,5	1,69	1,55	1,66	1,67	1,56	1,57	1,58	1,59
1,59	1,60	1,60	1,61	1,62	1,74	1,63	1,63	1,64	1,64
1,65	1,65	1,66	1,49	1,68	1,52	1,53	1,62	1,69	1,71

Supongamos que contásemos cuantas veces se repite cada estatura (por ejemplo, el 1,49 está una vez, el 1,50 está una vez, etc.) Hacemos una tabla de frecuencias y creamos un diagrama de barras:

Estatura	Cantidad
1,49	1
1,5	1
1,51	0
1,52	1
1,53	2
1,54	0
1,55	1
1,56	1
1,57	1
1,58	1
1,59	2
1,6	2
1,61	1
1,62	2
1,63	2
1,64	2
1,65	2
1,66	2
1,67	1
1,68	1
1,69	2
1,7	0
1,71	1
1,72	0
1,73	0
1,74	1



No parece que este diagrama de barras sirva para arribar a alguna descripción útil…
Esto suele ocurrir cuando tenemos **muchos datos muy distintos en valor entre sí**. Y esto a su vez suele ocurrir cuando trabajamos con variables cuantitativas continuas, como lo es la altura, en que hay una gran variedad de posibles respuestas, valores muy similares, pero distintos por muy poco. En estos casos, conviene agrupar a los datos en lo que se llama "intervalos de clase", y realizar lo que se llama un "histograma". Veamos como…

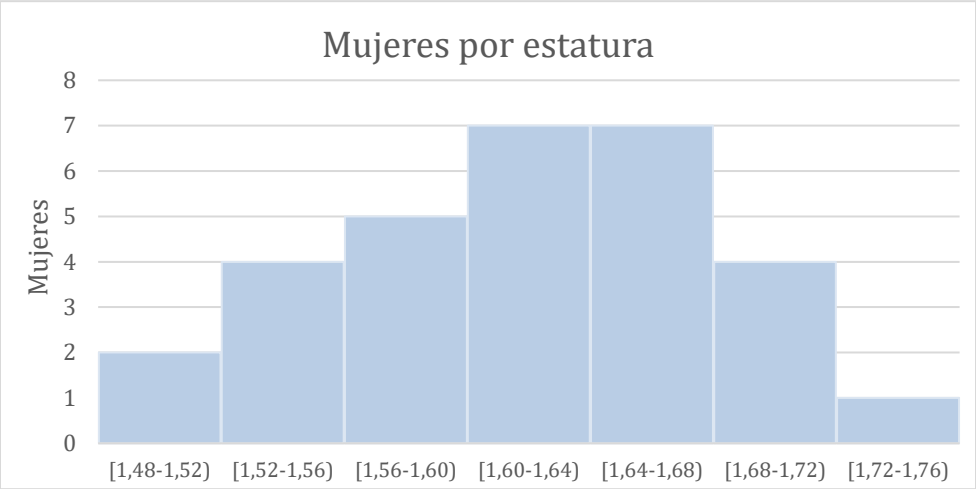
Histograma

Supongamos que agrupamos primeramente los 30 datos en 7 intervalos (a los que llamaremos "**intervalos de clase**"), y contamos cuantos datos pertenecen a cada intervalo, confeccionando entonces la siguiente tabla de frecuencias absolutas de cada intervalo.
Es como si convirtiéramos en discreta una variable que era continua: ahora puede haber mediciones con distintos valores, pero que pertenecen a una misma clase o categoría de intervalo.

Estatura	Cantidad
[1,48-1,52)	2
[1,52-1,56)	4
[1,56-1,60)	5
[1,60-1,64)	7
[1,64-1,68)	7
[1,68-1,72)	4
[1,72-1,76)	1

Veamos entonces como construir el gráfico denominado "**histograma**":

En el eje horizontal marcamos los intervalos de clase y se toma cada uno de ellos como la base de un rectángulo cuya altura es la frecuencia de cada clase.



Vendría a ser como un diagrama de barras, pero con la diferencia de que acá las barras se hallan pegadas, **una al lado de la otra**, y que lo que se representa en el **eje horizontal** son los **intervalos** numéricos.

Ahora sí nos es posible llegar a algunas conclusiones con la simple observación: por ejemplo, la altura de casi la mitad de las mujeres seleccionadas es de 1,60 a 1,68. También, entre otras cosas, observamos que es poco frecuente que una mujer mida más de 1,72.

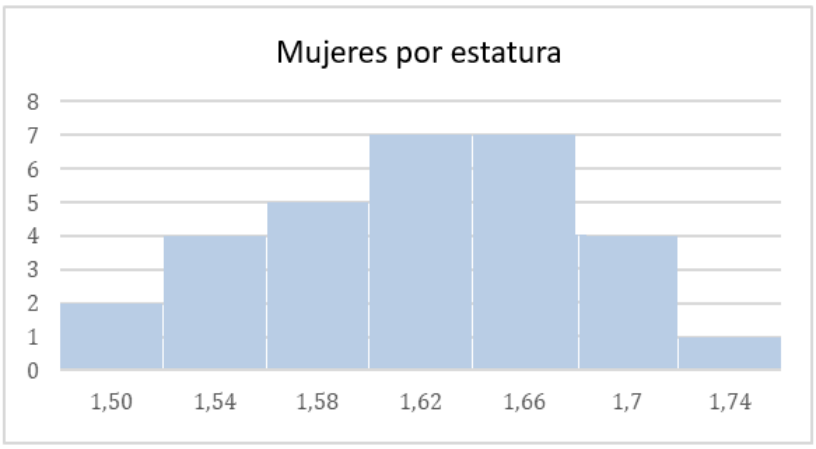
A veces, en vez de exhibir en el eje x los intervalos de clase, se prefiere representar a cada uno mediante su "**marca de clase**". Se le dice así al valor central, al punto medio, que representa a cada intervalo de clase y se obtiene al sumar los límites del intervalo y dividir el resultado por dos:

Marca de clase del intervalo "i" $= X_i = \dfrac{extremo\ inferior + extremo\ superior}{2}$

Para nuestro último ejemplo, procederíamos del siguiente modo:

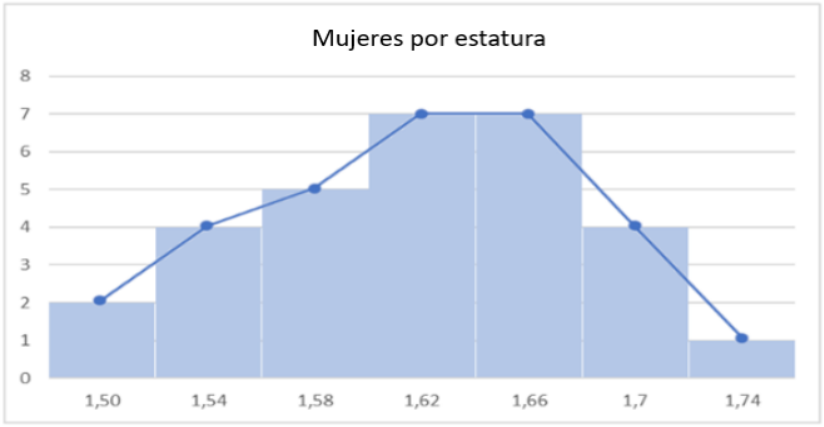
Estatura	X_i
[1,48-1,52)	$\frac{1{,}48 + 1{,}52}{2} = 1{,}50$
[1,52-1,56)	$\frac{1{,}52 + 1{,}56}{2} = 1{,}54$
[1,56-1,60)	1,58
[1,60-1,64)	1,62
[1,64-1,68)	1,66
[1,68-1,72)	1,70
[1,72-1,76)	1,74

Y ahora en el histograma sustituimos los intervalos por las marcas de clase.



Es como si asumiéramos que hay dos personas que miden 1,50; cuatro que miden 1,54, etc. Si bien sacrificamos la precisión del dato (pues representamos a todos los de un mismo intervalo con un único valor) ganamos en cuanto a la claridad del gráfico.

Además, a veces se traza sobre el histograma un **polígono de frecuencias**, que resulta de conectar con segmentos la parte superior de cada columna, justo en su mitad para así dar una mejor idea de la forma de distribución de la variable.



A veces solo se deja representado dicho polígono de frecuencias.