INSTITUTO de ENSEÑANZAS a DISTANCIA de ANDALUCÍA

PAU

Mayores de 25 años

Contenidos

Matemáticas Aplicadas a las Ciencias Sociales
Sucesiones y estadística: Estadística unidimensional

1. Estadística: ¿Una nueva ciencia?

Seguro que estarás de acuerdo con nosotros en que las estadísticas forman parte sustancial de nuestras vidas. No solo porque todos estemos incluidos en multitud de estadísticas, sino porque los medios de comunicación cada vez están más saturados con datos estadísticos: el número de desplazamientos en unas determinadas fechas vacacionales, el número de usuarios de Internet, el desarrollo de un jugador a lo largo de un partido de baloncesto, las preferencias en el tipo de películas que vemos, los hábitos de lectura o de salud, y multitud de temas más. Muchos de ellos acompañados de gráficas que permiten ver de un solo vistazo mucha más información comparada que de cualquier otra manera.

Además, en muchos aspectos sociales y económicos se utilizan las estadísticas para deducir que puede pasar en el futuro, como puedes ver en el siguiente vídeo:

De aquí en adelante, queremos que aprendas a: manejar datos, reconocer la forma en que pueden expresarse, estudiar las posibles relaciones que hay entre unos resultados y sacar conclusiones. Es decir, vamos a seguir las etapas de un proceso estadístico:

Todo ello debe tener un objetivo claro, formarnos como consumidores críticos de los medios de comunicación con el fin de no dejarnos manipular por la información matemática expresada de forma tendenciosa.

1.1. Conceptos básicos

Ya te habrás hecho la idea de que la **Estadística** es la rama de las matemáticas que se encarga de recolectar y organizar datos con el objeto de inferir conclusiones sobre ellos.

La Estadística se divide en dos partes:

- **Estadística descriptiva**:es la parte que se encarga de recoger, organizar, expresar gráficamente y resumir los datos que se han recogido. En ella construiremos tablas de frecuencias, agrupando los datos, los representaremos gráficamente y calcularemos parámetros que nos indicarán claramente cómo se han distribuido los datos recogidos.

- **Estadística inferencial o inferencia estadística**: es la parte que estudia regularidades en los datos recogidos para elaborar conclusiones futuras, permitiendo una toma de decisiones más efectiva. La bondad de esas deducciones se mide de forma probabilística, es decir, estudiaremos cómo es la probabilidad de acertar eligiendo una opción u otra. También se pretende estudiar si se puede generalizar el estudio hecho de unos datos a toda la población. Es lo que se hace cuando se realiza una estadística de predicción de votos antes de unas elecciones.

Por ahora nos centraremos en estudios estadísticos descriptivos. Como ya has visto en la introducción, un estudio estadístico puede dividirse en etapas, y sin duda, la primera es la elaboración de un plan de actuación:

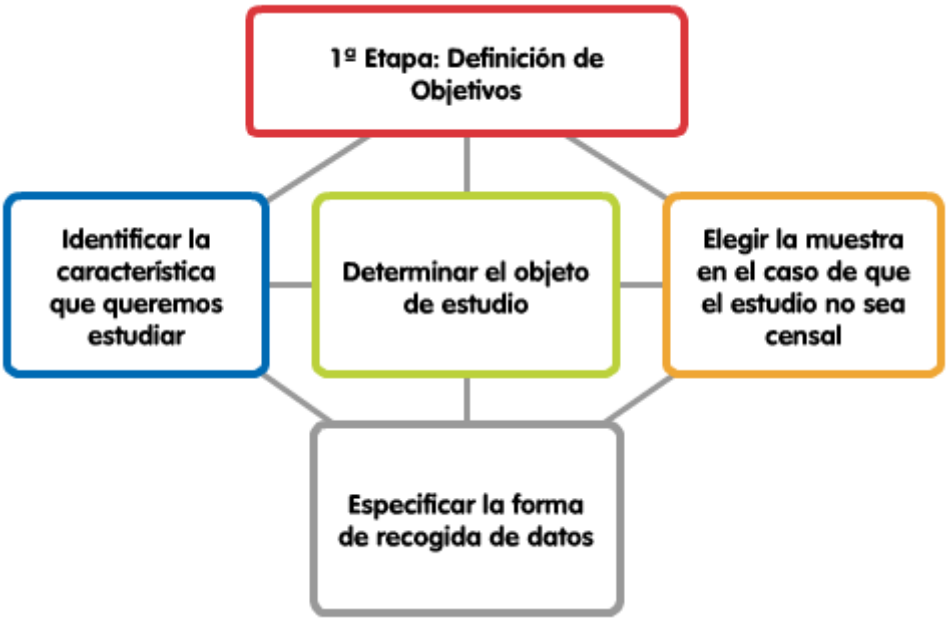


Imagen de elaboración propia

El objeto de estudio

Llamamos **población** al conjunto de elementos que son objeto de estudio estadístico, e **individuo** a cada uno de los elementos de la población.

Aunque tengan estos nombres, esos elementos pueden referirse a cualquier cosa y no necesariamente personas. Por ejemplo podemos estudiar los televisores que se montan en una determinada fábrica, la cantidad de vehículos que se desplazan por carretera un fin de semana de agosto, o los programas de televisión que se prefieren en una determinada franja horaria. Cada televisor, vehículo o programa televisivo sería un individuo de ese estudio.

A veces es necesario estudiar a todos los elementos de la población (investigación censal), por ejemplo cuando se realiza el censo de población de una determinada ciudad. Pero en general, es muy costoso, en tiempo y dinero, entrevistar a todos los elementos objeto del estudio, por ello se selecciona solo una parte, a la que llamamos **muestra**.

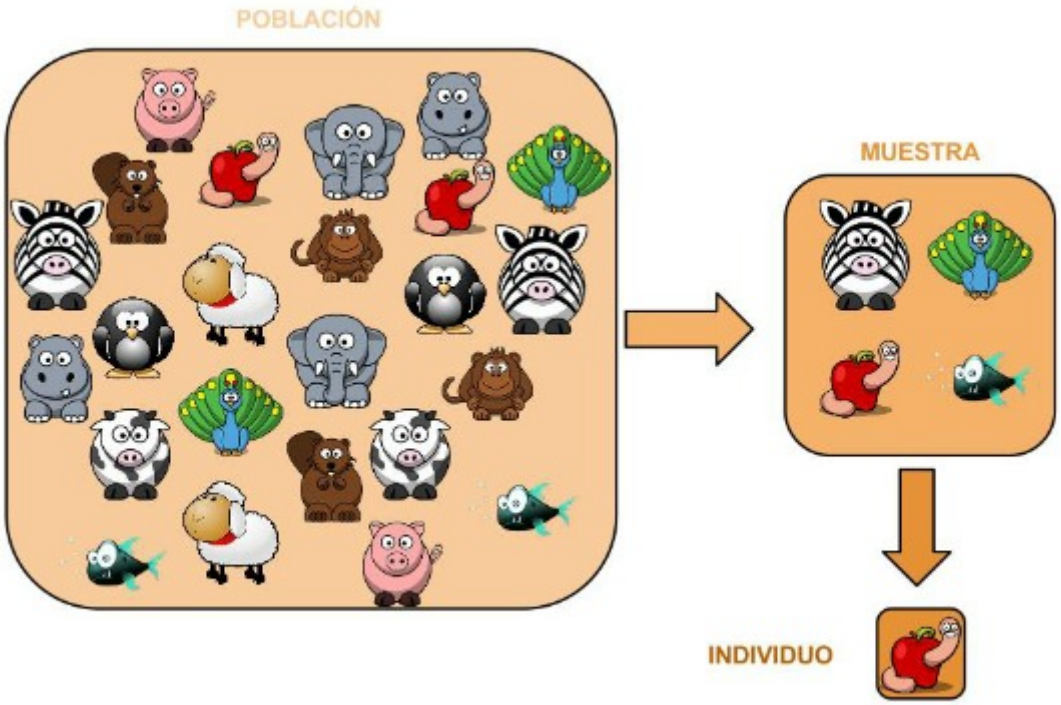


Imagen de elaboración propia

Importante

La elección de la muestra es muy importante para que los resultados que se extraigan de ella se puedan generalizar a toda la población. Debe haber pocos individuos para que no sea muy costosa su realización, pero elegidos de forma que aparezcan todos los estratos diferentes que forman la población. Si se cumplen ambas condiciones decimos que la **muestra** es **representativa**.

¿Qué característica queremos estudiar de la población?

Se llama **variable estadística** o carácter a cada una de las características que pueden estudiarse de la población.

Las variables estadísticas pueden ser de dos tipos:

- **Cualitativas**: son aquellas en la que los resultados posibles no son valores numéricos. Por ejemplo, color del pelo, tipo de ropa preferida, lugar de veraneo, etc.
- **Cuantitativas**: aquellas cuyo resultado es un número. A su vez las hay de dos tipos:

    1. **Cuantitativas discretas**: cuando toman valores aislados. Por ejemplo: número de amigos de tu pandilla, número de veces que vas al cine al mes, número de coches que tiene tu familia.

    2. **Cuantitativas continuas**: cuando entre dos valores cualesquiera pueden haber valores intermedios, es decir, se toman todos los valores de un determinado intervalo. Por ejemplo: peso de las personas, nivel sobre el mar en que se encuentra tu ciudad, medida del perímetro torácico.

A cada una de las posibles respuestas de una variable estadística se les llama **modalidad**. Estas modalidades tendrán que estar definidas sin ambigüedad, de manera que cada individuo pueda representar una y sólo una de las modalidades de cada variable.

Comprueba lo aprendido

A continuación te mostramos tres variables estadísticas, dos corresponden a sendas encuestas que ha realizado la edición digital de el diario El País, y la tercera es una estadística que se incluye en la memoria del año 2009 sobre Seguridad Vial, publicada por la Dirección General de Tráfico (DGT).

Completa con **sí** o **no** los espacios en blanco que aparecen en la tabla en función de que sea cierto o falso lo que se afirma.

Encuesta/Estadística	Variable/Respuesta	Cualitativa	Cuantitativa	Discreta	Continua
¿Qué presupuesto (en euros), tienes para esta Semana Santa?	0 Menos de 100 De 100 a 300 Más de 300	☐	☐	☐	☐
¿Qué harías para reducir las muertes por violencia de género?	Publicar las listas de maltratadores Promover la eduación en igualdad Aumentar el presupuesto Introducir cambios en la ley	☐	☐	☐	☐
¿Cuántos puntos del Carnet de conducir has perdido en los últimos 3 años?	0 2 3 4 6	☐	☐	☐	☐

Enviar

Seguro que has completado todos los espacios en blanco bien. Si no es así, repasa las definiciones anteriores. Recuerda que debes escribir sí o no.

2. Organización de datos

Ya tenemos claro qué vamos a estudiar, a quiénes les vamos a hacer el estudio, y cuál será el medio por el que obtendremos la información. El siguiente paso es obtener los datos y ordenarlos.

Imagen de elaboración propia

El documento donde se van anotando las características o las respuestas aportadas por los individuos es lo que se llama el **cuestionario**, y al igual que la elección de la muestra, su diseño es otra de las características más importantes que hay que tener en cuenta cuando planificamos una encuesta, pues a partir de él tenemos que ser capaces de sacar toda la información necesaria para realizar la investigación deseada.

Imagen en Flickr de @pach bajo CC

Pueden plantearse distintos tipos de preguntas. Las preguntas de respuesta cerrada exigen más trabajo en su preparación, pues tiene que cubrir todas las posibles respuestas que pueda pensar el encuestado, no debe ofrecer alternativas inadecuadas que incomoden al entrevistado, deben ser excluyentes las alternativas..., pero a cambio, el procesamiento de los datos es mucho más fácil que en las preguntas de respuestas abiertas. En general, ningún tipo de pregunta es mejor que otro, todo depende de lo pensado y trabajado que tengamos el cuestionario, así que, lo ideal es una mezcla entre un tipo y otro.

Otros tipos de preguntas que se pueden plantear en un cuestionario son las preguntas de ordenación en una escala, las preguntas con más de una elección (no demasiado aconsejable pues puede que las respuestas no se traten de manera correcta) y preguntas en las que hay que ordenar según las preferencias del encuestado los ítems ofrecidos como respuesta.

Cuando se diseña un cuestionario no se tiene claro al cien por cien si va a funcionar bien o no. Por eso, muchas veces se hace un estudio previo o encuesta piloto a una muestra bastante más pequeña para ver si el cuestionario se ajusta a lo que los estadísticos quieren.

2.1. Frencuencia y tablas

Las personas a veces tendemos a acumular objetos, ropa... y para organizarnos las ordenamos en armarios, estanterías, etc. De esta forma, sabemos qué tenemos y dónde lo tenemos. Pues bien, cuando realizamos un estudio estadístico muchos son los datos que obtenemos sin ningún tipo de orden.

Pero de esta cantidad de datos, ¿podemos obtener alguna información?

Parece lógico pensar que ordenándolos y posteriormente encontrando semejanzas podamos agruparlos y obtener, de esta forma, una información que pueda describirnos la población escogida.

Imagen en Flickr de Andy.Schultz bajo CC

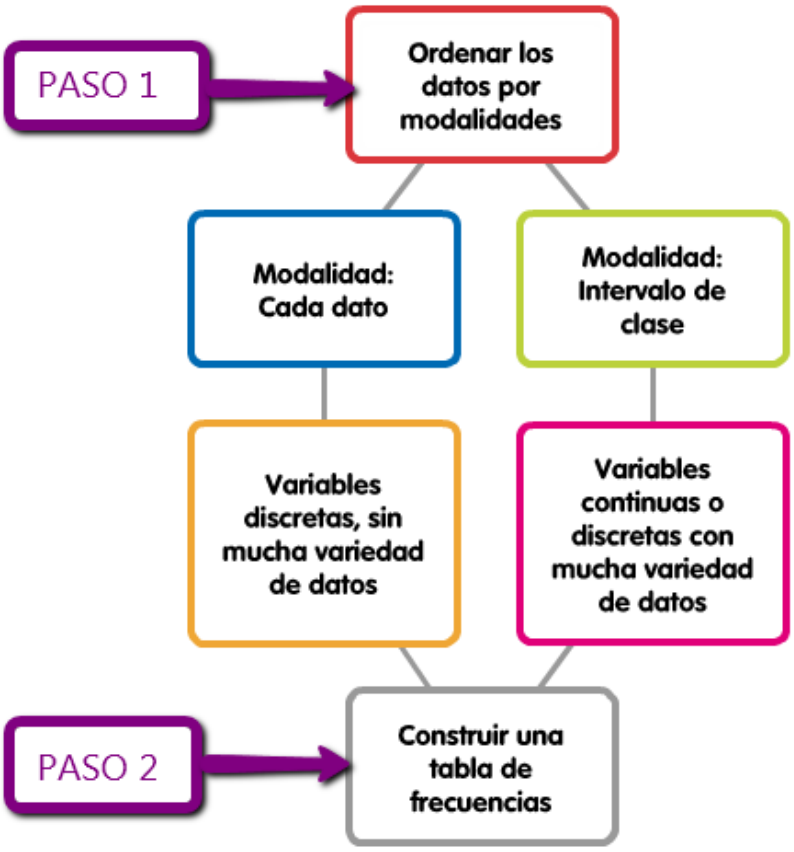


Imagen de elaboración propia

El primer paso para ordenar estos datos es definir las distintas modalidades según los resultados obtenidos.

Esto podemos hacerlo de dos formas:

- **Ordenarlos por elementos**. En general, optamos por esta técnica en variables discretas, en los que la variedad de datos no es muy amplia (variedad no cantidad). Las modalidades serían todas las posibles repuestas.

- **Ordenarlos por intervalos.** En general, optamos por esta técnica en variables cuantitativas continuas o en variables discretas en las que hemos obtenido datos muy variados. Estos intervalos se reciben el nombre de intervalos de clase, en los cuales se agruparán los datos observados en una muestra. Así, cada intervalo es considerado como una modalidad. Una vez construidos los intervalos de clase, se elige un representante de cada uno de ellos, llamado **marca de clase**, que normalmente es el **punto medio del intervalo**.

El segundo paso es colocar los datos en tablas que indiquen el número de veces que se repite cada modalidad, estas tablas se llaman **tablas de frecuencias**.

Frecuencia absoluta y frecuencia relativa

Ahora bien no es lo mismo que una modalidad se repita 5 veces en un total de 50 datos que en un total de 500. Por tanto, hablaremos de frecuencias absoluta y relativa.

- La **frecuencia absoluta** es el número de veces que se repite una modalidad de una variable en un estudio estadístico. Se suele representar por n_i. En el caso de que las modalidades sean intervalos tomamos como frecuencia absoluta de cada modalidad el número de observaciones agrupadas en el intervalo correspondiente.

- La **frecuencia relativa** es el cociente entre la frecuencia absoluta (n_i) y el número total de datos (N). Se representa por f_i.

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

La suma de todas las frecuencias absolutas es igual al número total de datos (N). Y por tanto, la suma de todas las frecuencias relativas es 1.

Por último, añadir que multiplicando las frecuencias relativas por 100, obtenemos el **tanto por ciento** de esa modalidad en la muestra.

Frecuencias acumuladas

En multitud de ocasiones nos interesa agrupar los datos, de manera que, sin perder información, nos sean más útiles.

Por ejemplo, es interesante conocer en cuántas provincias se puede dormir en verano, es decir, cuántas están por debajo del umbral del sueño.

Tomando los siguientes datos, que se refieren a las temperaturas de una sola noche, obtenemos la tabla:

Temperaturas	n_i	N_i
15ºC	1	1
16ºC	3	4
17ºC	4	8
18ºC	2	10
19ºC	3	13
20ºC	1	14
21ºC	5	19
22ºC	6	25
23ºC	2	27
24ºC	2	29

Sabiendo que para poder conciliar el sueño necesitamos una temperatura inferior a 22ºC, de la tabla anterior se deduce que solo en 19 de las provincias estudiadas han dormido a pierna suelta.

La tercera columna de la tabla hace referencia al concepto de **frecuencia absoluta acumulada**.

Importante

La **frecuencia absoluta acumulada** es el resultado de sumar la frecuencia absoluta de una modalidad de la variable con todas las frecuencias de las modalidades anteriores.

$$N_i = n_1 + n_2 + \ldots + n_i$$

En consecuencia, las **frecuencias relativas acumuladas** representan la proporción de individuos de una muestra que presentan alguna de las i primeras modalidades.

$$F_i = f_1 + f_2 + \ldots + f_i$$

La *última* frecuencia absoluta acumulada es igual al número total de datos.

Y por tanto, la *última* frecuencia relativa acumulada es igual a 1.

Ejercicio resuelto

En una biblioteca se ha realizado una encuesta entre los usuarios sobre los libros que se han leído en el último mes:

4,1,3,10,5,2,2,5,1,19,8,3,5,15,2,1,1,1,6,3,2,12,3,7,6,3,4,1,10,7,11,6,7,12,4,2,8,

5,9,3,6,8,2,1,12,9,8,5,2,3,3,4,3,7,9,1,4,9,5,8,6,12,17,3,9,6,7,5,5,3,9,7,8,11,2.

Elaboraremos una tabla de frecuencias agrupando los valores en intervalos.

Mostrar retroalimentación

Elegiremos los intervalos de clases:

La modalidad de interés son los libros que leen los usuarios de esta biblioteca que van entre 1 libro y 19 libros. Podemos realizar intervalos de 5 en 5. Quedando:

Intervalos de clase Libros leídos
[0,5)
[5,10)
[10,15)
[15, 20)

Calcularemos las marcas de clases: recordamos que son el punto medio del intervalo:

Intervalos de clase	Marca de clase
[0,5)	2,5
[5,10)	7,5
[10,15)	12,5
[15, 20)	17,5

Construiremos el resto de la tabla de frecuencias:

Intervalos de clase	Marca de clase	n_i	f_i	N_i	F_i
[0,5)	2,5	32	0,427	32	0,427
[5,10)	7,5	32	0,427	64	0,854
[10,15)	12,5	8	0,106	72	0,960
[15, 20)	17,5	3	0,040	75	1
TOTAL		75	1		

Ejercicio resuelto

Curso 2010/2011

En la corrección de errores tipográficos de un texto se han encontrado 22 páginas con un solo error en cada una, 9 páginas con 2 errores en cada una, 6 páginas con 3 errores en cada una, 3 páginas con 4 errores en cada una, 2 páginas con 5 errores en cada una y ningún error en las 58 páginas restantes.

Construya las tablas de frecuencias absolutas y relativas de la distribución del número de errores por página de ese texto.

Mostrar retroalimentación

Vamos a estudiar el número de errores por página, por lo tanto la modalidad de la variable es el número de errores, y su frecuencia absoluta es el número de páginas que tienen esos errores.

Número de errores	Número de páginas Frecuencia absoluta n_i	Frecuencia relativa f_i	Frec. Abs. Acumulada N_i	Frec. Rel. Acumulada F_i
0	58	$\frac{58}{100}=0,58$	58	0,58
1	22	$\frac{22}{100}=0,22$	80	0,8
2	9	$\frac{9}{100}=0,09$	89	0,89
3	6	$\frac{6}{100}=0,06$	95	0,95
4	3	$\frac{3}{100}=0,03$	98	0,98
5	2	$\frac{2}{100}=0,02$	100	1
TOTAL	100	1		

Gráficos estadísticos

Siempre se ha dicho: "*Vale más una imagen que mil palabras*". Y en estadística también es cierto. Los datos que hemos ordenado antes en tablas también los podemos representar gráficamente.

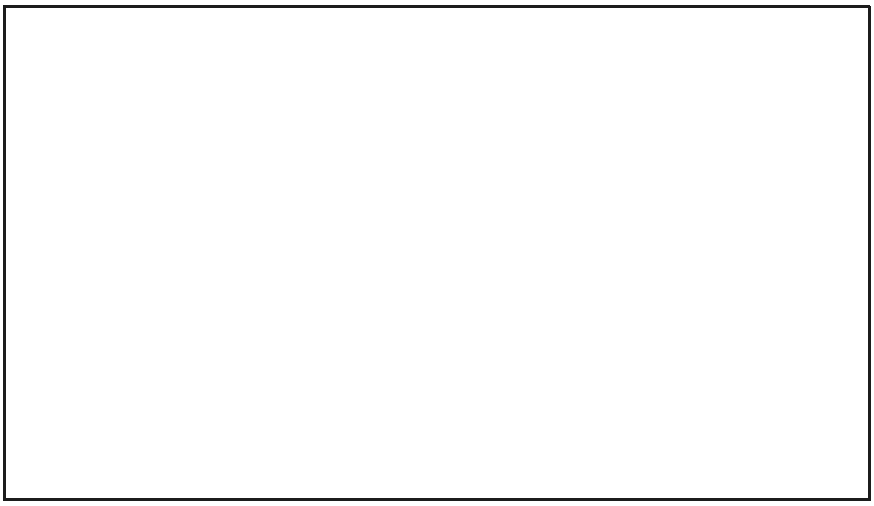
Un gráfico es una representación visual mediante elementos geométricos (líneas, círculos...) de una serie de datos estadísticos.

La ventaja de utilizar gráficos estadísticos es que nos permite comprender fácilmente el fenómeno que estamos estudiando, pudiendo saber de esta forma cómo evoluciona, hasta qué niveles llega...

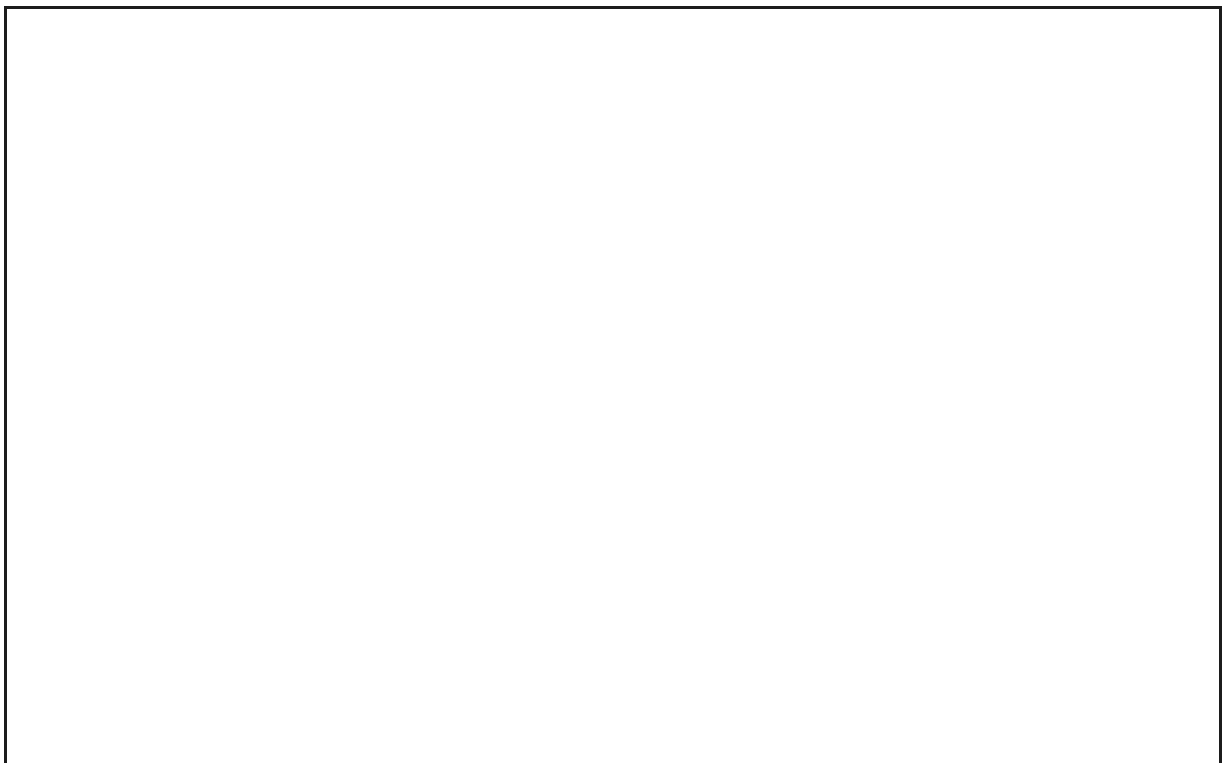
Según el objetivo de la investigación, las variables que queremos estudiar, sus características..., optaremos por un tipo de representación o por otra.

Tipos de Gráficos

En España el INE (Instituto Nacional de Estadística) es el encargado de la coordinación general de los servicios estadísticos. Pero para que puedas hacerte una idea de la importancia y la magnitud de este órgano, puedes ver el siguiente vídeo:



Por lo tanto, quién mejor que ellos para describirte los distintos tipos de gráficos estadísticos existentes y sus principales funciones, en la siguiente escena de Flash:

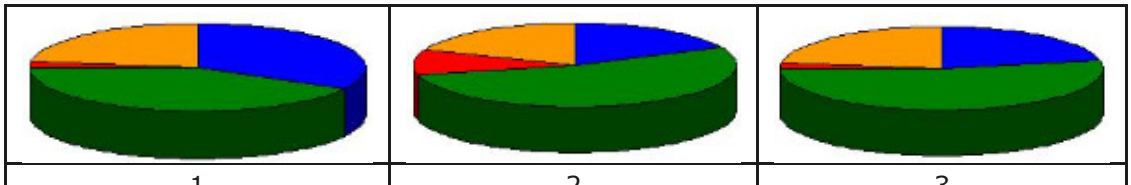


Comprueba lo aprendido

En la siguiente tabla aparecen las frecuencias relativas en tanto por ciento, de las respuestas que dieron los lectores, donde se preguntaba a los lectores sobre qué harían para reducir las muertes por violencia de género.

Respuestas	**Frecuencia relativa en tanto por ciento**
Publicar las listas de maltratadores	22
Promover la educación en igualdad	53
Aumentar el presupuesto	2
Introducir cambios en la ley	23

Indica qué **número** de los siguientes diagramas de sectores pertenece a la tabla anterior: ☐



1	2	3

Además, escribe el **color** que corresponde a las siguientes opciones: a "Promover la educación en igualdad" le pertenece el ☐ y a "Aumentar el presupuesto" el ☐

.

Enviar

3. Análisis de datos

Para analizar la gran cantidad de datos que se utilizan al realizar un estudio estadístico, se hace imprescindible resumirlos de manera que se conserve la mayor información posible y que representen el comportamiento global de la población.

Esto lo vamos a hacer a través de una serie de medidas que complementarán unas a otras:

- **Medidas de centralización**, buscan las características del centro de la distribución, y son la **media, moda y mediana**.
- **Medidas de posición**, indican, una vez ordenados, cuántos elementos quedan a la izquierda o derecha de uno dado: **cuartiles, deciles, centiles o percentiles**.
- **Medidas de dispersión**, proporcionan una idea sobre la separación de los datos: **rango o recorrido, desviación media, varianza, desviación típica y coeficiente de variación**.
- **Medidas de forma**, proporcionan una idea de la simetría de la distribución: **coeficiente de asimetría**.

Imagen en Flickr de HeavyWeightGeek bajo CC

Importante

Si recuerdas en la unidad 1, hablábamos de la importancia del lenguaje algebraico a la hora de "escribir" fórmulas. La mayoría de las medidas estadísticas, vienen dadas por estas expresiones algebraicas, en las que por lo general intervienen todos los datos. Este es el motivo que para facilitar la mnemotecnia, busquemos siempre la concisión y brevedad de estas fórmulas, intentando simplificarlas al máximo posible. Es por ello por lo que utilizamos el signo $\sum$ (sumatorio), para indicar de forma abreviada la suma de varios números.

Por ejemplo, para acortar la siguiente suma $x_1+x_2+x_3+x_4$, escribimos: $\sum_{i=1}^{4} x_i$.

3.1. Medidas de centralización

Las medidas de centralización son datos que representan de forma global a toda la población, y en torno a los cuales están agrupados todos los valores.

Estas medidas de centralización no son conceptos desconocidos, muy al contrario, aparecen continuamente en nuestra vida. Por ejemplo:

- A los niños pequeños, uno de los primeros conceptos que se les enseña son los tamaños. A muy corta edad aprenden la diferencia entre grande, **mediano** y pequeño.

- Hoy en día ir a la **moda** es una de las preocupaciones de mucha gente. Para no quedarte desfasado tienes que estar a la moda, es decir, en este caso, vestirte de la misma forma que mucha otra gente.

Imagen en Flickr de laverrue bajo CC

- ¿Sabías que van a situar dos radares seguidos y a cierta distancia, para de esa forma calcular la velocidad **media** a la que circulamos, y así poder comprobar si infringimos la ley?

Importante

Notación

Llamaremos $x_1, x_2, \cdots, x_n$ a los datos si la distribución está ordenada por elementos, o a las marcas de clase si está ordenada por intervalos. N al número total de datos, que coincide con la suma de las frecuencias absolutas, n_i. Recuerda que i va desde 1 hasta n, que es el número de datos distintos.

Media aritmética

Se calcula al sumar todos los elementos y dividir por el número total de elementos de la población, o lo que es lo mismo, si los datos vienen en una tabla de frecuencias, se multiplica cada dato por su frecuencia absoluta y se suman los resultados obtenidos, y este resultado se divide por el número total de datos. Se denota por $\overline{x}$ o por μ.

Si los datos vienen agrupados en intervalos, se multiplica la marca de clase por su frecuencia, se suman los resultados obtenidos y este total se divide por el número de datos.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \cdot n_i$$

Propiedades

- Solo se puede obtener con datos cuantitativos.

- Puede no coincidir con ninguno de los datos, y se encuentra entre el menor y el mayor de los datos de la distribución.

- Cuando aparecen valores extremos y poco significativos la media puede que no sea representativa.

- Es única. Solo existe una por cada distribución.

Imagen de elaboración propia

Moda

En el caso de una variable no agrupada, se define como el valor que más se repite entre los datos de que disponemos, es decir, es el dato que tiene mayor frecuencia absoluta. Vamos a representarla por M_o . En el caso de una variable agrupada en intervalos de igual amplitud se busca el intervalo con mayor frecuencia, al que llamaremos, **intervalo modal**. Dentro del intervalo modal se suele tomar como moda el punto medio del intervalo (marca de clase).

Propiedades

- La moda es el único estadístico que puede utilizarse con cualquier tipo de variable. En concreto es el único parámetro que tiene sentido calcular en las variables cualitativas.

- Tiene además la particularidad de ser el único parámetro estadístico que puede tomar más de un valor. Por ejemplo, una familia muy numerosa, con ascendencia de partos múltiples tiene 6 hijos cuyas edades son 3, 3, 6, 10, 10, 14. Entre esos datos, la moda correspondería a los valores 3 y 10 ya que ambos se repiten dos veces.

Mediana

Es el valor de la variable que divide a la serie de datos ordenados en dos partes iguales. Si los datos no están agrupados por intervalos, tenemos dos casos:

- El número de datos de los que disponemos es impar, en cuyo caso se ordenan en orden creciente y la mediana es el término que ocupa el lugar central.

$$\underbrace{1,\ 1,\ 1,\ 1,\ 1,\ 1,}_{Mitad\ inferior}\ \underbrace{2,}_{Mediana}\ \underbrace{2,\ 2,\ 2,\ 3,\ 3,\ 4}_{Mitad\ superior}$$

Mediana datos impares

$$\underbrace{1,\ 1,\ 1,\ 1,\ 1,}_{Valores\ inferiores}\ \underbrace{1,\ 2,}_{Valores\ intermedios}\ \underbrace{2,\ 2,\ 3,\ 3,\ 4}_{Valores\ superiores}$$

- Si el número de datos es par, la mediana es la media aritmética de los dos valores centrales.

Se representa por M_e .

$$1,5 = \frac{1+2}{2}$$

Mediana datos pares

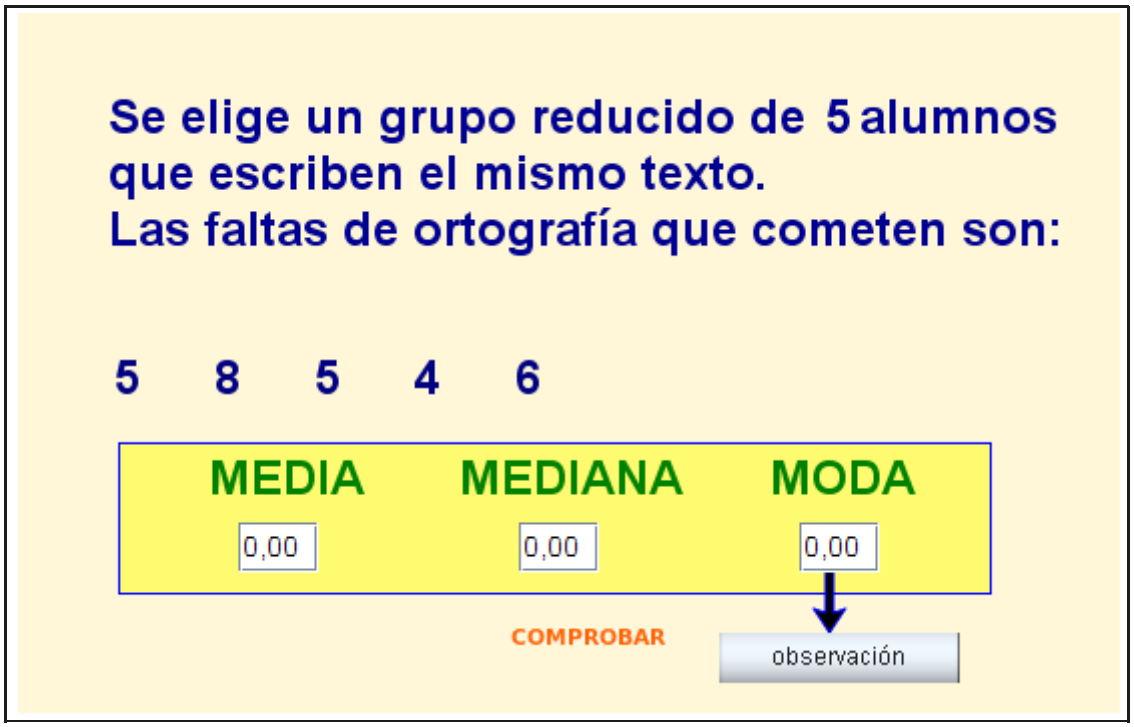
Propiedades

- Solo se puede obtener para datos cuantitativos.
- Es única, y puede no ser un dato de la distribución.

Si los datos están agrupados por intervalos, hablamos del intervalo mediano o central que es aquel donde se encuentre el o los valores centrales. Es posible utilizar una fórmula para precisar un valor para la mediana dentro del intervalo mediano. En el siguiente enlace puedes encontrarla:

Mediana para datos agrupados en intervalos

Si haces clic en la siguiente imagen puedes practicar con el cálculo de algunos parámetros de centralización:



Ejercicio resuelto

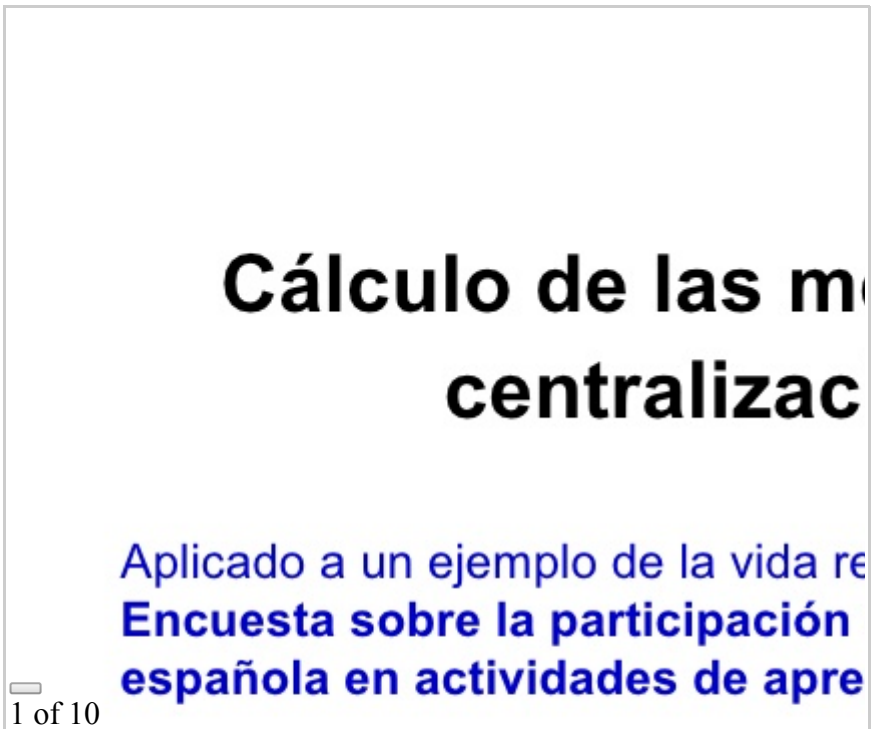
La siguiente distribución agrupada por intervalos, es fruto de una encuesta sobre la participación de la población adulta española en actividades de aprendizaje (EADA 2007):

Tramos de	Marca de	Frecuencia

edad	clase, x_i	absoluta, n_i
[25, 35)	30	3018026
[35, 45)	40	2433843
[45, 55)	50	1624953
[55, 65)	60	805375
[65, 75)	70	298697

Calcula las medidas de centralización asociadas.

Mostrar retroalimentación



Presentación en Slideshare por Jesús Fernández

Ejercicio resuelto

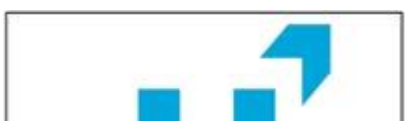
Curso 2011/2012

En una urbanización se ha realizado un estudio sobre el número de personas que habitan en cada piso y se obtienen los siguientes datos:

Personas	1	2	3	4	5
Pisos	20	60	52	35	18

a) ¿Cuántos pisos hay en la urbanización?

b) Determine la media y la moda de la distribución.



Universidades Públicas de Andalucía

Mostrar retroalimentación

a) El número de pisos se corresponde con la suma de todas las frecuencias absolutas de la distribución, por lo tanto, el total de pisos es 20+60+52+35+18=185.

b) La **moda** de la distribución es aquel valor que tiene mayor frecuencia absoluta, por lo que, la moda de nuestra distribución es 2.

La **media**:

$$\overline{x} = \frac{1 \cdot 20 + 2 \cdot 60 + 3 \cdot 52 + 4 \cdot 35 + 5 \cdot 18}{185} = \frac{526}{185} = 2{,}84$$

3.2. Medidas de dispersión

Al igual que pasa en el mundo, las variables estadísticas no siempre están bien repartidas, de tal forma que la media puede no ser representativa si hay valores dispares que le afecten. Para eso tenemos las medidas de dispersión, que nos informan de la concentración de esos datos dentro de la variable.

Las medidas de dispersión nos informan de hasta qué punto las medidas de centralización son representativas como síntesis de la información.

Las medidas de dispersión cuantifican la separación, la dispersión, la variabilidad de los valores de la distribución respecto al valor central.

Estudiaremos a continuación el recorrido, la desviación media, la varianza, la desviación típica y el coeficiente de variación.

Imagen en Flickr de Håkan Dahlström bajo CC

Recorrido

El recorrido de una variable es la diferencia entre el mayor y menor valor que toma esa variable. Se representa por R . En el caso de tener los datos agrupados por intervalos, el recorrido sería la diferencia entre el extremo superior del mayor intervalo y el extremo inferior del menor intervalo (no intervienen las marcas de clase).

Desviación media

Se define el parámetro desviación media como la suma de las diferencias entre los valores y la media, en valor absoluto, dividido por el número total de valores.

$$D_m = \frac{1}{N}\sum_{i=1}^{n} |x_i - \bar{x}| \cdot n_i$$

Imaginemos que estamos interesados en comprar en bolsa acciones de una determinada empresa. Para saber si es una inversión segura, investigamos la evolución del precio de las acciones en el último año. Por supuesto, hemos calculado previamente cuál ha sido el precio medio de las acciones y mirando el resto de los valores podremos decir que si muchos días el precio ha estado alejado de la media (por debajo o por encima), diremos que la inversión es arriesgada o volátil. Y por el contrario, si la mayoría de los días el precio ha estado cerca de la media diremos que la inversión es segura.

La medida estadística que nos indica esta variación de los datos respecto de la media es la varianza.

Varianza y desviación típica

La varianza se define como la media aritmética de los cuadrados de las desviaciones de cada valor respecto a la media. Se representa por S^2 , y se calcula mediante la fórmula:

$$S^2 = \frac{1}{N}\sum_{i=1}^{n} (x_i^2 - \bar{x})^2 \cdot n_i$$

La desviación típica es la raíz cuadrada positiva de la varianza. Se representa por $S = +\sqrt{S^2}$.

Aunque tanto la varianza como la desviación típica son medidas de la variación con respecto a la media y tienen un significado semejante, la desviación típica se utiliza con más frecuencia porque se expresa

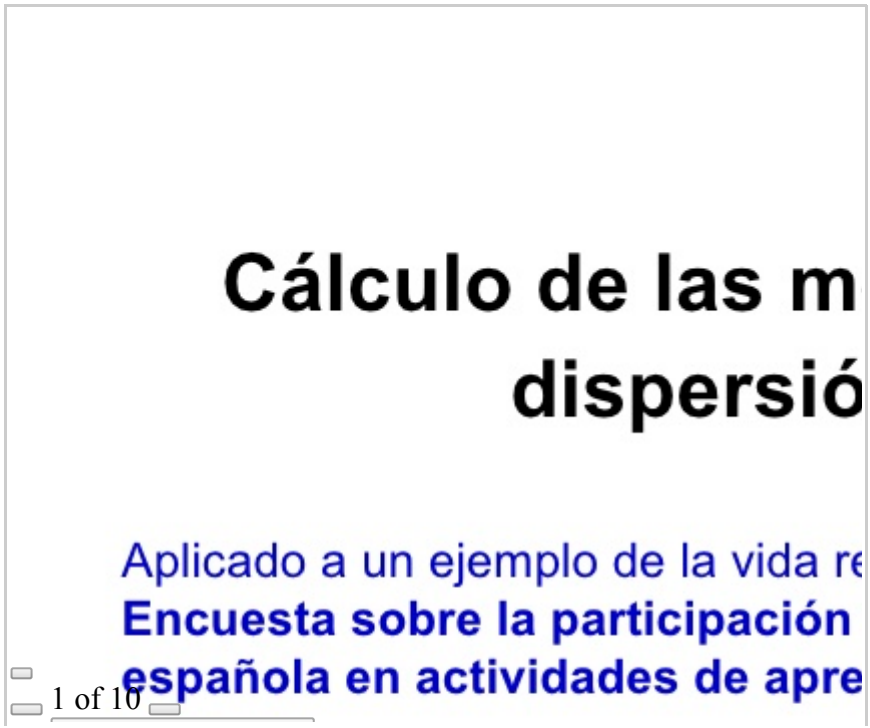
en las mismas unidades que los valores de las variables. Al contrario que la varianza, cuyos valores se expresan en unidades al cuadrado.

Haciendo algunos cambios y realizando algunos cálculos podemos obtener una fórmula más operativa para la varianza:

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot n_i}{N} - \bar{x}^2$$

La varianza y la desviación típica también podemos encontrarla representadas como σ^2 y σ respectivamente.

A continuación un ejemplo en un contexto real:



Presentación en Slideshare por **Jesús Fernández**

Cuando queremos comparar la dispersión entre dos poblaciones distintas, es muy útil la siguiente medida:

Coeficiente de variación

El coeficiente de variación se define como el cociente entre la desviación típica y la media. Se representa por CV.

$$CV = \frac{\sigma}{\bar{x}}$$

El coeficiente de variación se suele expresar en porcentajes: $CV = \frac{\sigma}{\bar{x}} \cdot 100\ \%$

Cuanto mayor sea el coeficiente de variación, más dispersión hay en la variable y menos representativa es la media.

El coeficiente de variación es una medida independiente de las unidades de medición, ya que es el cociente de la desviación típica y la media, y ambas se miden en las mismas unidades.

Ejercicio resuelto

Curso 2010/2011 (Continuación)

En la corrección de errores tipográficos de un texto se han encontrado 22 páginas con un solo error en cada una, 9 páginas con dos errores en cada una, 6 páginas con 3 errores en cada una, 3 páginas con 4 errores en cada una, 2 páginas con 5 errores en cada una y ningún error en las 58 páginas restantes.

a) Construya las tablas de frecuencias absolutas y relativas de la distribución del número de errores por página de ese texto

b) Halle la media y la desviación típica del número de errores por página de dicho texto.

Mostrar retroalimentación

a) Resuelto en el apartado 2.1.

b)

Número de errores	Número de páginas Frecuencia absoluta n_i
0	58
1	22
2	9
3	6
4	3
5	2
TOTAL	100

Vamos a calcular la media:

$$\overline{x} = \frac{0 \cdot 58 + 1 \cdot 22 + 2 \cdot 9 + 3 \cdot 6 + 4 \cdot 3 + 5 \cdot 2}{100} = 0{,}8$$

Para calcular la desviación típica, primero calculamos la varianza:

$$s^2 = \frac{0^2 \cdot 58 + 1^2 \cdot 22 + 2^2 \cdot 9 + 3^2 \cdot 6 + 4^2 \cdot 3 + 5^2 \cdot 2}{100} - 0{,}8^2 = \frac{210}{100} - 0{,}64 = 1{,}46$$

Por último haciendo la raíz cuadrada a la varianza, obtenemos la desviación típica:

$$\sigma = \sqrt{1{,}46} = 1{,}21$$

Ejercicio resuelto

Curso 2011/2012 (Continuación)

En una urbanización se ha realizado un estudio sobre el número de personas que habitan en cada piso y se obtienen los siguientes datos:

Personas	1	2	3	4	5
Pisos	20	60	52	35	18

a) ¿Cuántos pisos hay en la urbanización?

b) Determine la media y la moda de la distribución.

c) Determine la varianza y la desviación típica de la misma.

Mostrar retroalimentación

a) y b) Resueltos en el apartado 3.1.

c) Para calcular la varianza y la desviación típica tenemos que partir de la media que calculamos en el apartado anterior, $\overline{x}=2,84$.

Varianza

$$s^2=\frac{1^2\cdot 20+2^2\cdot 60+3^2\cdot 52+4^2\cdot 35+5^2\cdot 18}{185}-2,84^2=\frac{1738}{185}-8,0656=1,33$$

Desviación típica

$$\sigma=\sqrt{1,33}=1,15$$

Medidas de posición

Imagen en Flickr de tomypelluz bajo CC

Seguro que muchas veces te habrás encontrado realizando grandes colas: para sacar entradas de un concierto o espectáculo muy requerido, para renovar o entregar alguna documentación, o en situaciones parecidas. En algunos sitios, donde las colas de espera están bien organizadas, nosotros al menos, hemos encontrado de pronto un letrero que indicaba "a partir de aquí 1 hora de espera", unos metros más adelante te encuentras otro de "desde aquí 45 minutos" y así sucesivamente. Es una forma de compartimentar la cantidad de personas que esperan y el tiempo que se tardará en llegar a la entrada o ventanilla. Algo similar hacen los parámetros que vamos a ver en este subapartado.

Se llaman **parámetros de posición** aquellos que dividen a los datos obtenidos en partes proporcionales, de forma que cada parte tenga el mismo número de elementos. Para poder hacerlo necesitamos que los datos estén ordenados de menor a mayor. A veces se les llama con el nombre genérico de **cuantiles**. Los hay de tres tipos: **cuartiles, deciles y percentiles**, aunque vamos a desarrollar el primero y el último.

Importante

Se definen los **cuartiles** como los valores que dividen a la distribución de valores ordenados en cuatro partes iguales. Son los siguientes:

- Q_1 : primer cuartil, tiene el 25 % de los datos delante de él y el 75 % detrás.
- M_e : segundo cuartil, que coincide con la mediana. Tiene el 50 % de los datos delante y el otro 50 % detrás de él.
- Q_3 : Deja delante de él el 75 % de la distribución y detrás el 25 %.

Para calcular los cuartiles basta generalizar el cálculo de la mediana que ya habíamos visto. Se halla $\frac{N}{4}$ y el primer valor cuya frecuencia absoluta acumulada supera ese valor es Q_1 . Para Q_3 debemos hallar $\frac{3 \cdot N}{4}$ y seleccionar aquel valor cuya frecuencia acumulada supera esa cantidad. Hay que recordar, como en la mediana, que si un valor tiene como frecuencia acumulada exactamente ese valor se halla la media aritmética con el valor siguiente.

Se define el **recorrido intercuartílico** como la diferencia entre el tercer y el primer cuartil. Dentro de este intervalo se encuentra el 50 % de la distribución.

$$R_q = Q_3 - Q_1$$

Un estudio conjunto del recorrido y del recorrido intercuartílico nos da información sobre la dispersión de la muestra. Si el recorrido general es grande pero el intercuartílico pequeño, eso indica que hay valores extremos. Si ambos son grandes los datos son dispersos y si ambos son pequeños los datos están muy agrupados respecto a los valores centrales.

Ejercicio resuelto

Para realizar un estudio sobre el gasto farmacéutico en la sanidad pública, nos encargan que hagamos un estudio sobre el número de medicamentos por paciente que se receta en una determinada consulta a lo largo de una semana. Se obtiene la siguiente tabla:

Nº de medicamentos (x_i)	1	2	3	4	5	6	7	8	9	10
Nº de pacientes (n_i)	12	24	15	13	9	6	2	1	1	1

Calcula los cuartiles de esa distribución.

Mostrar retroalimentación

Necesitamos las frecuencias acumuladas y dividir N en cuartos.

x_i	n_i	N_i
1	12	12
2	24	36
3	15	51
4	13	64
5	9	73
6	6	79
7	2	81
8	1	82
9	1	83
10	1	84

- $\frac{N}{4}=\frac{84}{4}=21$. La primera frecuencia absoluta acumulada que supera a 21 es 36, que corresponde al valor 2.
  Por tanto $Q_1=2$.
- $\frac{N}{2}=\frac{84}{2}=42$. La primera frecuencia absoluta acumulada que supera a 42 es 51, que corresponde al valor 3.
  Por tanto $M_e=3$.
- $\frac{3N}{4}=\frac{3\cdot 84}{4}=63$. La primera frecuencia absoluta acumulada que supera a 63 es 64, que corresponde al valor 4.
  Por tanto $Q_3=4$.

Importante

Se definen los **percentiles** como aquellos parámetros que dividen el conjunto ordenado de valores en 100 partes iguales. De esta manera, el percentil 34, por ejemplo, es aquel que tiene delante el 34% y detrás el 66% restante.

De forma análoga a los cuartiles, para hallar los percentiles dividimos el número total de datos (N) entre 100 y multiplicarlo por el orden del percentil que se busca y después hallar qué valor iguala o sobrepasa a esa cantidad.

Ejercicio resuelto

En el ejemplo anterior de las recetas calcula los percentiles 12, 50 y 67.

Mostrar retroalimentación

Teniendo la tabla de las frecuencias acumuladas que hemos usado en el ejercicio anterior, hallamos los percentiles pedidos.

x_i	n_i	N_i
1	12	12
2	24	36
3	15	51
4	13	64
5	9	73
6	6	79
7	2	81
8	1	82
9	1	83
10	1	84

- $\frac{12N}{100} = \frac{12 \cdot 84}{100} \approx 10 \Rightarrow P_{12} = 1$
- $\frac{50N}{100} = \frac{50 \cdot 84}{100} \approx 42 \Rightarrow P_{50} = 3$
- $\frac{67N}{100} = \frac{67 \cdot 84}{100} \approx 56 \Rightarrow P_{67} = 4$

Como quizás te hayas dado cuenta en el ejercicio anterior, el percentil 50 coincide con la mediana y, de forma análoga, $P_{25} = Q_1$ y $P_{75} = Q_3$.

Igual que hay unos parámetros que dividen el conjunto de los valores en 100 partes iguales, existen los **deciles** que la dividen en 10 partes iguales, pero en lugar de los deciles se suelen utilizar más los percentiles.

Cuando los datos vienen agrupados en intervalos, además de encontrar el intervalo que contiene al percentil correspondiente, es posible precisar el valor del mismo utilizando la fórmula que puedes encontrar en el siguiente enlace:

Percentiles para datos agrupados en intervalos

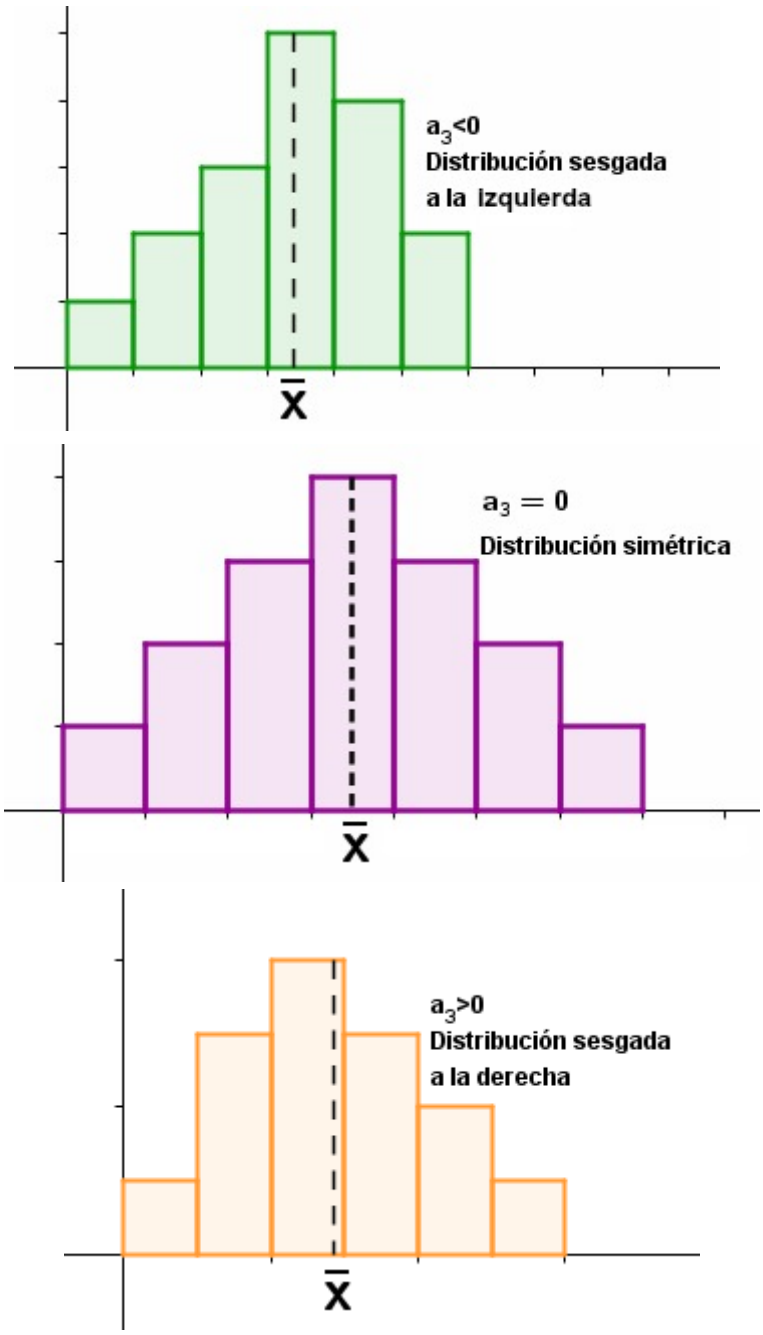
Medidas de simetría

Existen otras medidas que nos permiten caracterizar la forma de la distribución.

El **coeficiente de asimetría de Fisher** se define como:

$$a_3 = \frac{1}{\sigma^3} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3 \cdot n_i}{N}$$

Como $(x_i - \bar{x})^3$ puede ser positivo o negativo, este coeficiente puede ser positivo o negativo. Atendiendo al signo tenemos los siguientes casos:



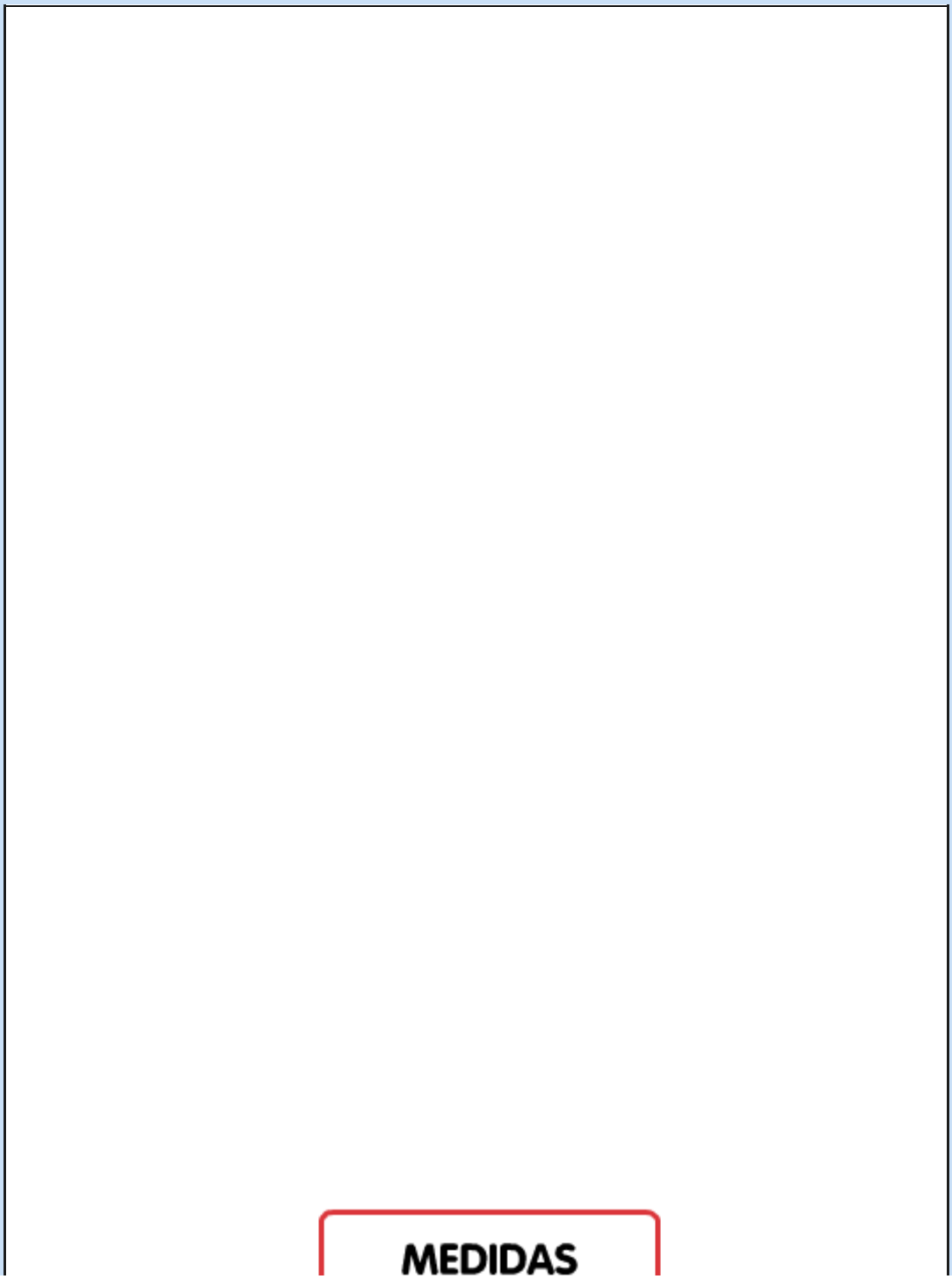
Un coeficiente de asimetría positivo corresponde a distribuciones cuya parte a la derecha de la media es más larga, en un sentido intuitivo, que al otro lado. La asimetría es negativa en el caso opuesto.

4. Apéndice

Si te has fijado en las actividades que te hemos propuesto de exámenes anteriores, habrás descubierto que el cálculo de las medidas estadísticas, es una pregunta que aparece con cierta frecuencia, por eso es imprescindible que adquieras soltura y desarrolles destrezas en este campo y así evitar errores. Como en unidades anteriores en este apartado te ofrecemos, unos recursos que te ayudarán a que ciertos cálculos se hagan menos tediosos. Además, también puedes ampliar conocimientos con las "Curiosidades" y "Para saber más".

Importante

Sin duda lo más complicado de este tema de estadística es memorizar las distintas fórmulas que aparecen, para poder luego aplicarlas con soltura. Es por ello por lo que te adjuntamos la siguiente imagen recopilación de todas:



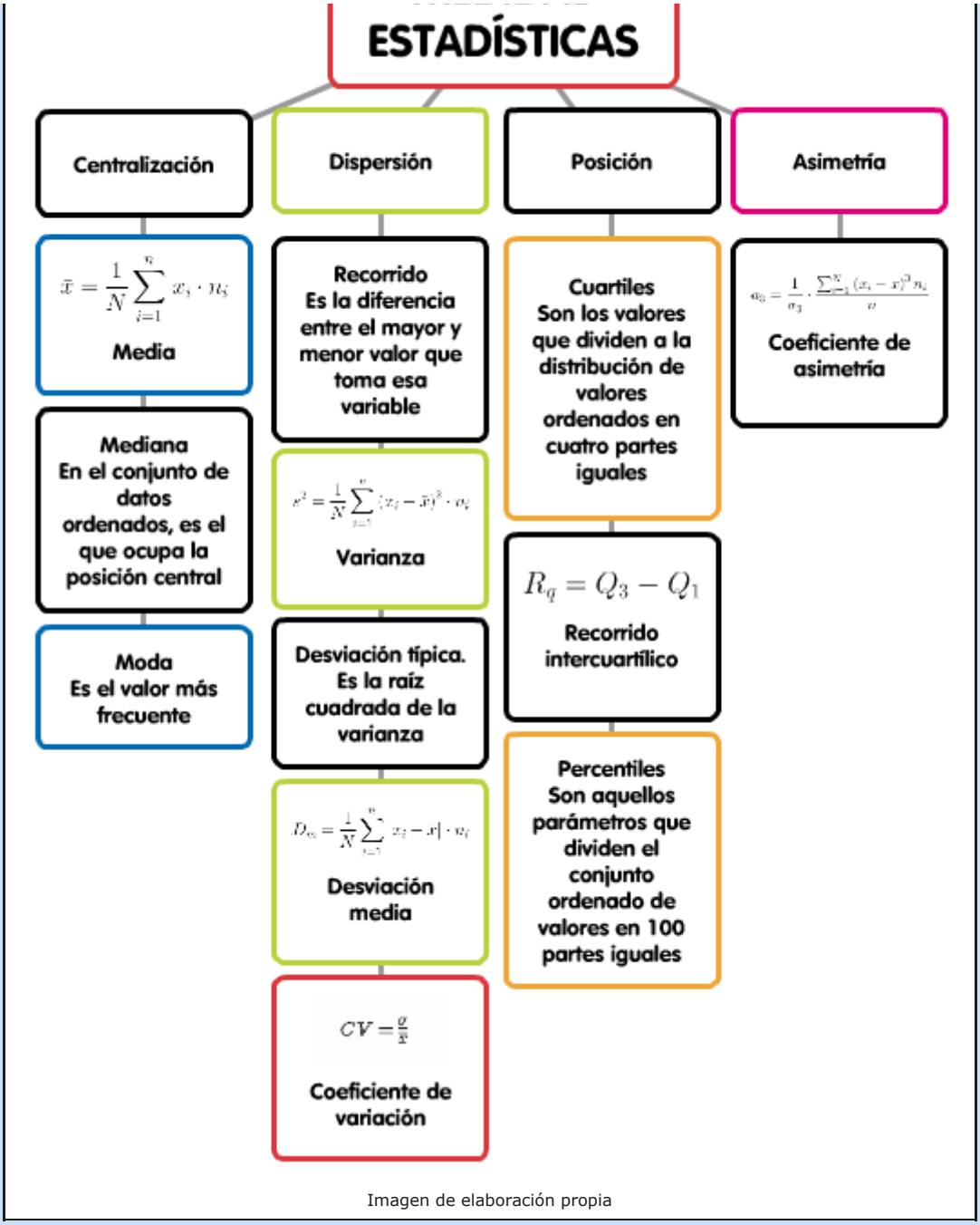


Imagen de elaboración propia

Haz clic en la imagen para ampliar

Calculadora

Aunque probablemente te sea más sencillo, conocer las fórmulas y utilizar la calculadora simplemente para hacer cálculos, te dejamos una lista de vídeos donde te explican como utilizar las funciones de estadísticas disponibles en algunos modelos CASIO:

Como ya sabes no dispondremos de otras herramientas en la prueba, pero si en algún momento decides practicar con ejercicios de los que no tengas la solución, esto te ayudará a corregirlos.

Open Calc

Haz clic en la imagen y descubrirás una página en la que te explican cómo trabajar con Estadística descriptiva tanto con la hoja de Cálculo de OpenOffice.

Estadística con:

Curiosidad

Historia

Estamos acostumbrados a hablar de las distintas ramas de las matemáticas empezando por culturas ancestrales como la griega, egipcia, china... Sin embargo, la Estadística es una disciplina relativamente reciente. Por eso te recomendamos el siguiente vídeo:

Curiosidad

Varianza y póquer

La varianza es el nombre técnico que dan los jugadores de póquer a una racha de mala suerte. La varianza justifica que aunque un jugador esté jugando bien vaya perdiendo: todas sus acciones tienen EV (valor esperado) positivo, pero su banca registra pérdidas.

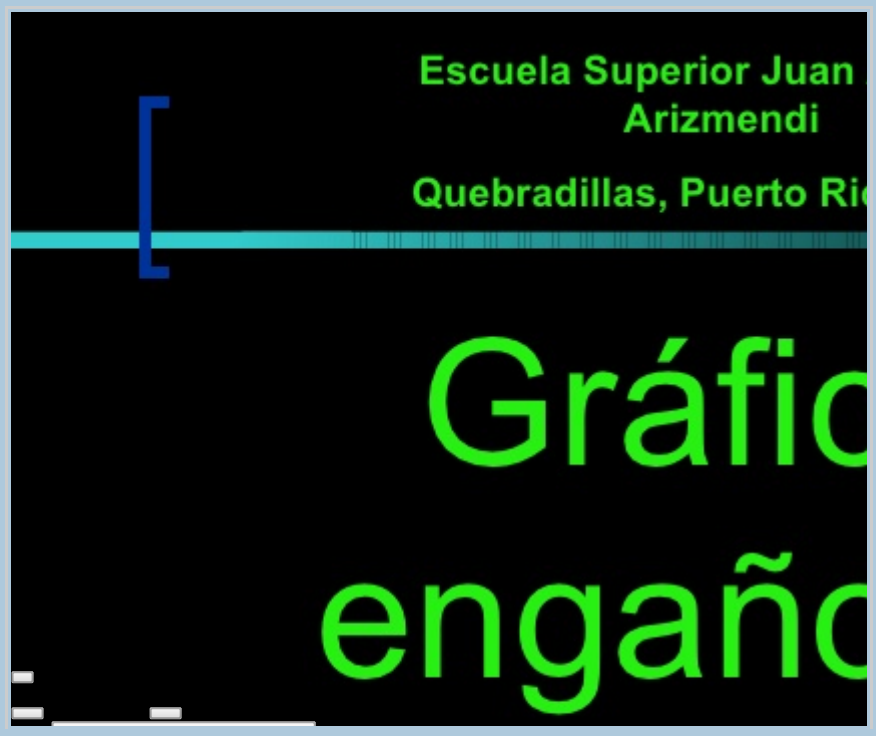
Los jugadores dicen que no importa perder porque "ha sido culpa de la varianza". El buen jugador de póquer parece que solo tiene un enemigo: la varianza.

Para defenderse de la varianza lo más común es mejorar la gestión de los recursos propios (banca), para que "un zarpazo de la varianza" no termine en "bancarrota". Es decir, que una racha negativa no acabe con toda nuestra banca y podamos seguir jugando hasta que llegue una racha positiva.

Gráficas engañosas

Para terminar el tema, no podemos olvidar que es fundamental estar atentos y ser críticos con la información que recibimos. Y ser mucho más cautos si esta información nos llega en formato gráfico, ya que son muchas las ocasiones en que se intentan tergiversar los datos representados.

Lo puedes comprobar en esta presentación:



Presentación en Slideshare por jaa_math

Para saber más

Si haces clic en la siguiente imagen del Proyecto EDAD, descubrirás una serie de actividades relacionadas con los conceptos trabajados a lo largo del tema:

Para saber más

A continuación, en la siguiente presentación puedes repasar y ampliar los conceptos vistos anteriormente. Si te animas, te percatarás que no hemos estudiado todas las herramientas disponibles para hacer un análisis estadístico, pero sí las fundamentales: