Índice

¿Te lo sabes?

Continúa

¡A estudiar un suceso!

Introducción

En la primera parte de este apartado hemos aprendido algunas nociones sobre cómo calcular probabilidades, así que ¡vamos a ello! Vamos a estudiar un suceso aleatorio, comenzando por su observación. Veamos tres ejemplos:

1. En general los aparatos eléctricos se venden con cierta garantía de que van a durar un cierto tiempo, desde las bombillas hasta los frigoríficos. El tiempo que dura un aparato eléctrico sin estropearse es un fenómeno aleatorio.

Vamos a hacer un estudio sobre la calidad de las bombillas que fabrica una factoría. Tenemos que coger las bombillas y ver cuánto duran encendidas. Para medir eso no hay más remedio que tener la bombilla encendida hasta que se funda, y entonces anotamos las horas que ha durado encendida. Así que cogemos todas las bombillas fabricadas, las encendemos y esperamos a que se fundan. Los resultados son excelentes, todas han durado muchísimo encendidas... el problema es que ahora están todas rotas, y no podemos vender ninguna... Está claro que el procedimiento que hemos seguido no es el adecuado... ¿qué hacemos?

Introducción

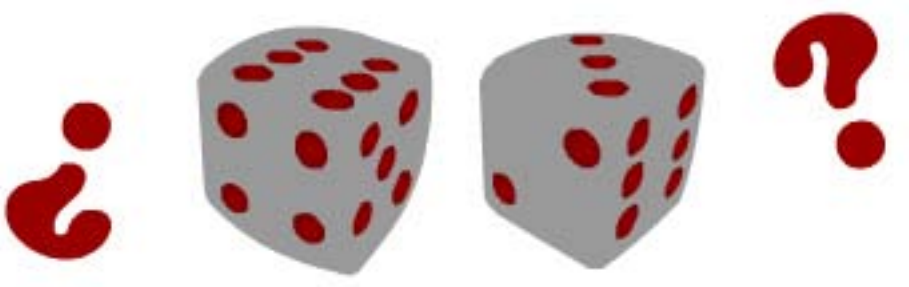
2. Decidimos estudiar el peso y la estatura de todos los habitantes de España. En España hay casi 50 millones de habitantes, así que vamos a tener que emplearnos a fondo. Cada día logramos medir a 10 personas, por lo que vamos a tardar 5.000.000 días. Aunque trabajásemos todos los días del año, sin fines de semana ni vacaciones, tardaríamos aproximadamente 13.700 años en medirlas a todas.

Como no parece muy práctico tardar tanto, decidimos que tenemos que hacer el trabajo más rápido, así que contratamos a gente para que nos ayude. Si contratáramos a 10.000 personas lograríamos medir 100.000 personas al día. Con esto tardaríamos tan solo 50 días, que son casi 2 meses de trabajo. Si tengo que pagar 1.000 euros al mes a cada persona, el estudio cuesta 20 millones de euros... igual es un poco caro... Si no me ayuda nadie sale muy barato pero tardo mucho, y si contrato a mucha gente tardo muy poco pero me sale muy caro... ¿qué hacemos?

Introducción

3. Como nuestra experiencia con las bombillas y la estatura de la gente no ha sido muy buena, decidimos estudiar algo más sencillo.

Vamos a lanzar un dado todas las veces y a anotar el número que sale, y luego vamos a contar cuántas veces ha salido el 1, el 2 etc. Así que nos ponemos a lanzar un dado hasta... ¡un momento, un momento! ¿Cuántas veces se puede lanzar un dado? En teoría se podría no parar nunca. Podemos ponernos a lanzar el dado y cuando nos cansemos que siga otra persona (si es que logramos convencer a alguien), y no tiene por qué acabarse nunca. Y si no se acaba nunca no puedo estudiarlo por completo... ¿qué hacemos?



¡A estudiar un suceso!

Muestreo

Parece que en muchas de las ocasiones no podemos observar todas las veces que el suceso ocurre.

- A veces la toma de información es destructiva.
- Otras veces tomar toda la información es muy caro, en tiempo y dinero. Además llegar a toda la población no es fácil (a veces no sabemos dónde vive alguien y no podemos llegar a su casa, por ejemplo).
- En otras ocasiones el suceso puede ocurrir infinitas veces, por lo que nunca vamos a poder observarlo *del todo*.

¿Y si estudiamos solo unos pocos casos?

La mayoría de las veces con estudiar únicamente ciertos casos es suficiente. Este conjunto de unos pocos casos se llama **muestra**.

Muestreo

¿Se han cocido los macarrones?

Cuando se ponen a cocer macarrones hay que estar pendiente de que queden como a uno le gusten. Para saber si están listos hay dos opciones:

1- Comerse la olla entera. Es posible que el resto de los comensales no quede muy satisfecho si hacéis eso, porque se quedarán sin comer (este es un caso clarísimo de toma de información destructiva... para el resto de los invitados, claro).

2- Probar 2 o 3 macarrones, y si esos pocos están hechos... pues lo más seguro es que todos los macarrones estén ya listos. Naturalmente suponer que como unos pocos están hechos lo estarán todos (extender el resultado de unos pocos a toda la población) tiene sus riesgos. Siempre habrá alguien que encuentre en su plato un macarrón que esté un poco más duro, o un poco más blando que el resto de los macarrones. Es decir, que **al generalizar cometemos un error**.

Muestreo

¿Por qué se comete este error?. En primer lugar porque no estudiamos todos los casos. En segundo lugar porque podemos tener mala suerte y coger una muestra que no represente bien a la totalidad. Puede que ningún macarrón estuviera en su punto excepto precisamente ese que has ido a probar, y como has probado el único que está bien hecho pienses que todos lo están cuando en realidad no es así.

Es muy importante que las **muestras** sean un **buen representante de lo que queremos estudiar**, que recojan **toda la variedad**.

Por ejemplo, si queremos estudiar la estatura de la población no debemos coger una muestra que se componga únicamente de jugadores de baloncesto, que suelen ser gente muy alta; ¿verdad? Porque la estatura de estas personas no es la habitual; lo frecuente es que la gente sea más baja; y además, si cogemos solo a jugadores de baloncesto estamos cogiendo solo a gente alta cuando en la población hay muchas alturas diferentes. Cuanto mejor sea la muestra, menor será el error cometido por la generalización.

Muestreo

La rama de la Estadística que se encarga de estudiar cómo seleccionar buenas muestras se llama **Muestreo**. El Muestreo establece tanto la forma de selección como el tamaño de la muestra. Además, muy importante, el Muestreo se encarga también de **estudiar el error cometido** por generalizar los resultados obtenidos para la muestra.

¿Qué acabo de recoger? La Estadística Descriptiva

Introducción

Muy bien, ya tenemos el suceso que queremos estudiar y dónde lo vamos a observar (la muestra). Ahora recogemos los datos y...

¿Qué hago con los datos?

Para empezar, asegurarnos de que los hemos escrito bien (este proceso se llama **depuración**, por ejemplo si estamos estudiando las alturas de adultos y nos encontramos que hemos anotado que un adulto mide 180 metros seguramente nos hayamos equivocado).

El análisis de los datos depende lógicamente de lo que queramos hacer con ellos, y lo habitual es comenzar por **observar** cómo son los datos: ¿Cuántos casos hay de cada clase? ¿Qué características tienen?... La rama de la Estadística que se encarga de estudiar los datos recogidos se denomina **Estadística Descriptiva**.

¿Qué acabo de recoger? La Estadística Descriptiva

Frecuencias absolutas y relativas

A continuación vamos a explicar brevemente los indicadores que más se utilizan al estudiar los datos:

Las frecuencias absolutas
La frecuencia absoluta de un suceso es el número de veces que se da ese suceso.

Las frecuencias relativas
Es la proporción con respecto al total de la frecuencia absoluta. Normalmente la frecuencia relativa se da en porcentaje.

Veamos un ejemplo (los datos son ficticios):

¿Qué acabo de recoger? La Estadística Descriptiva

	Sexo	Tipo	Región
1	Hombre	Niño	Este
2	Hombre	Niño	Norte
3	Hombre	Adulto	Este
4	Hombre	Adulto	Oeste
5	Hombre	Adulto	Norte
6	Hombre	Adulto	Norte
7	Hombre	Adulto	Sur
8	Hombre	Adulto	Sur
9	Hombre	Anciano	Oeste
10	Hombre	Anciano	Oeste
11	Hombre	Anciano	Este
12	Hombre	Anciano	Sur
13	Mujer	Niña	Este
14	Mujer	Adulta	Norte
15	Mujer	Adulta	Sur
16	Mujer	Adulta	Oeste
17	Mujer	Adulta	Sur
18	Mujer	Adulta	Este
19	Mujer	Anciana	Norte
20	Mujer	Anciana	Norte

Frecuencias absolutas y relativas

Estamos haciendo un estudio sobre la incidencia de una enfermedad, y tenemos los siguientes datos sobre 20 personas enfermas.

¿Qué acabo de recoger? La Estadística Descriptiva

Frecuencias absolutas y relativas

Casos por sexo	Hombres	Mujeres	Total (suma)
Frecuencias absolutas	12	8	20

Casos por tipo	Niños	Adultos	Ancianos	Total (suma)
Frecuencias absolutas	3	11	6	20

Casos por región	Norte	Sur	Este	Oeste	Total (suma)
Frecuencias absolutas	6	5	5	4	20

¿Qué acabo de recoger? La Estadística Descriptiva

Frecuencias absolutas y relativas

$$\text{Frecuencia relativa (en \%)} = \frac{\text{Frecuencia absoluta}}{\text{Total}} * 100$$

Casos por sexo	Hombres	Mujeres	Total (suma)
Frecuencia absoluta	**12**	**8**	**20**
Frecuencia relativa	$\frac{12}{20} * 100 = 60\%$	$\frac{8}{20} * 100 = 40\%$	**100%**

¿Qué acabo de recoger? La Estadística Descriptiva

Frecuencias absolutas y relativas

Casos por tipo	Niños	Adultos	Ancianos	Total (suma)
Frecuencias absolutas	**3**	**11**	**6**	**20**
Frecuencia relativa	$\frac{3}{20} * 100 =$ **15%**	$\frac{11}{20} * 100 =$ **55%**	$\frac{6}{20} * 100 =$ **30%**	**100%**

Medidas de centralización

Manejar una tabla de datos puede ser muy complicado. Suele ser muy útil disponer de una cifra única que **resuma** todos los datos, puesto que da información sobre la muestra y además es más fácil operar con un solo número. Se conocen como **medidas de centralización**, y las más importantes son:

La media: consiste en sumar todos los datos y dividir entre el total.

La moda: es el valor que más se repite en la muestra.

La mediana: es el valor que queda *en medio* al ordenar los datos.

Veámoslo con un ejemplo.

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de centralización

Ejemplo: estamos haciendo un estudio sobre la estatura media de los miembros de la **familia García**, y los resultados son estos:

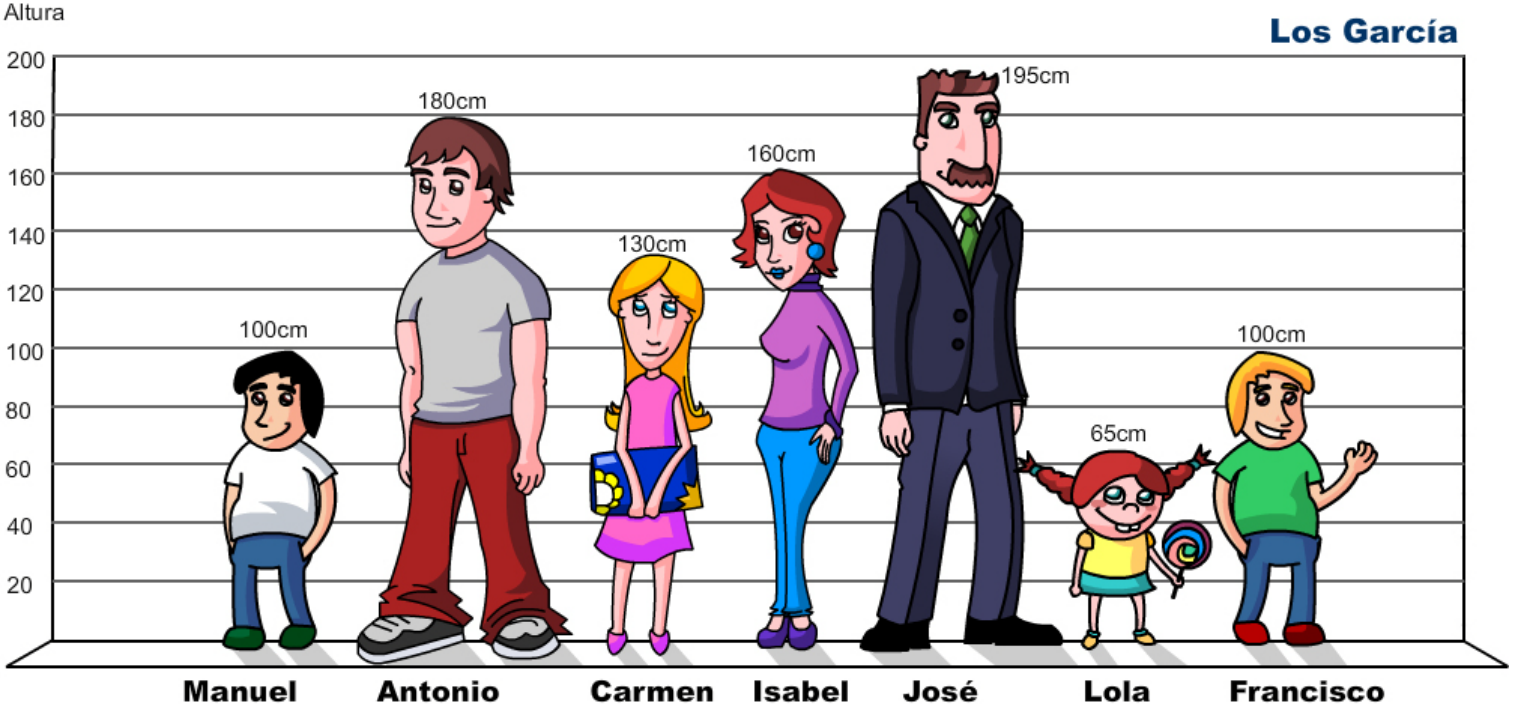
	Altura (centímetros)
Antonio	180
Carmen	130
Francisco	100
Lola	65
Manuel	100
Isabel	160
José	195

Los nombres no son casuales. Son los nombres más frecuentes en España a 1 de enero de 2009, igual que el apellido García.

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de centralización

La familia García

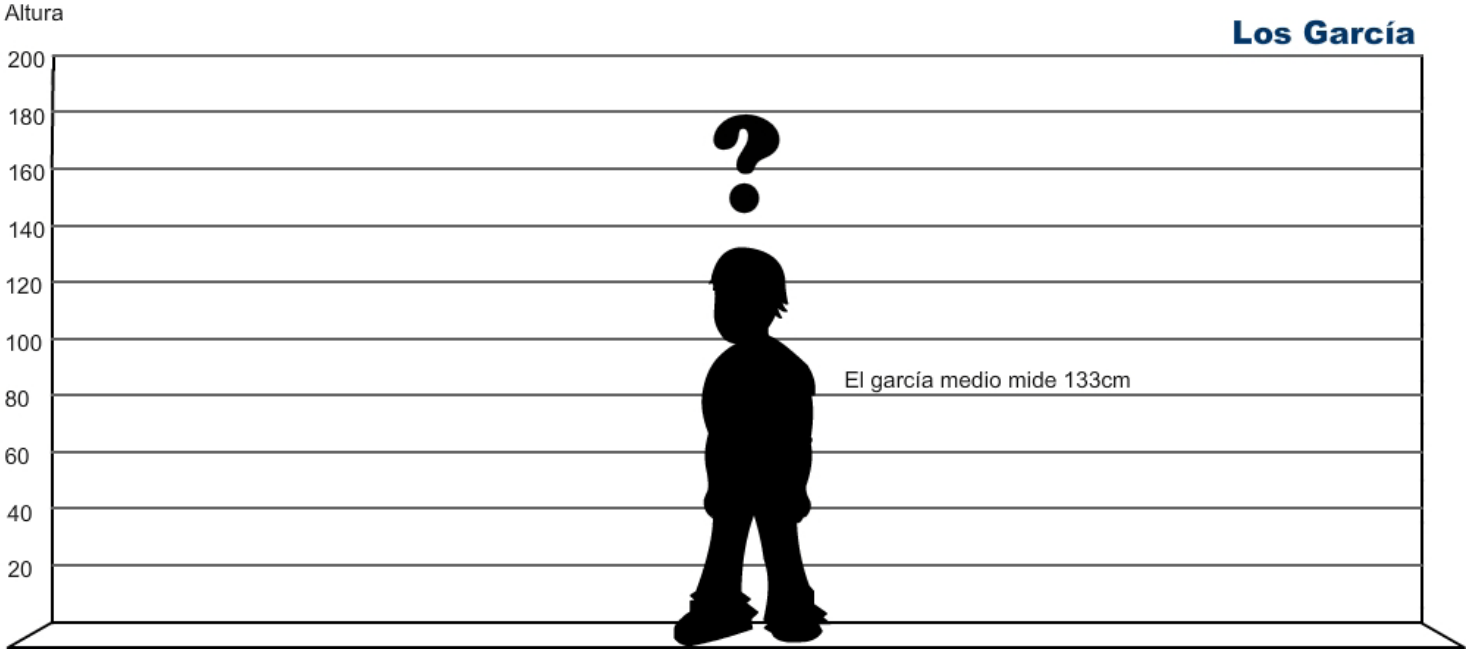


Media de altura = $\frac{100+180+130+160+195+65+100}{7} \approx$ **133** cm

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de centralización

¡Ojo a esto!
La media de estatura en esta familia es 133 cm, pero nadie mide exactamente esa altura. A veces la media toma un valor que no existe en la muestra.



¿Qué acabo de recoger? La Estadística Descriptiva

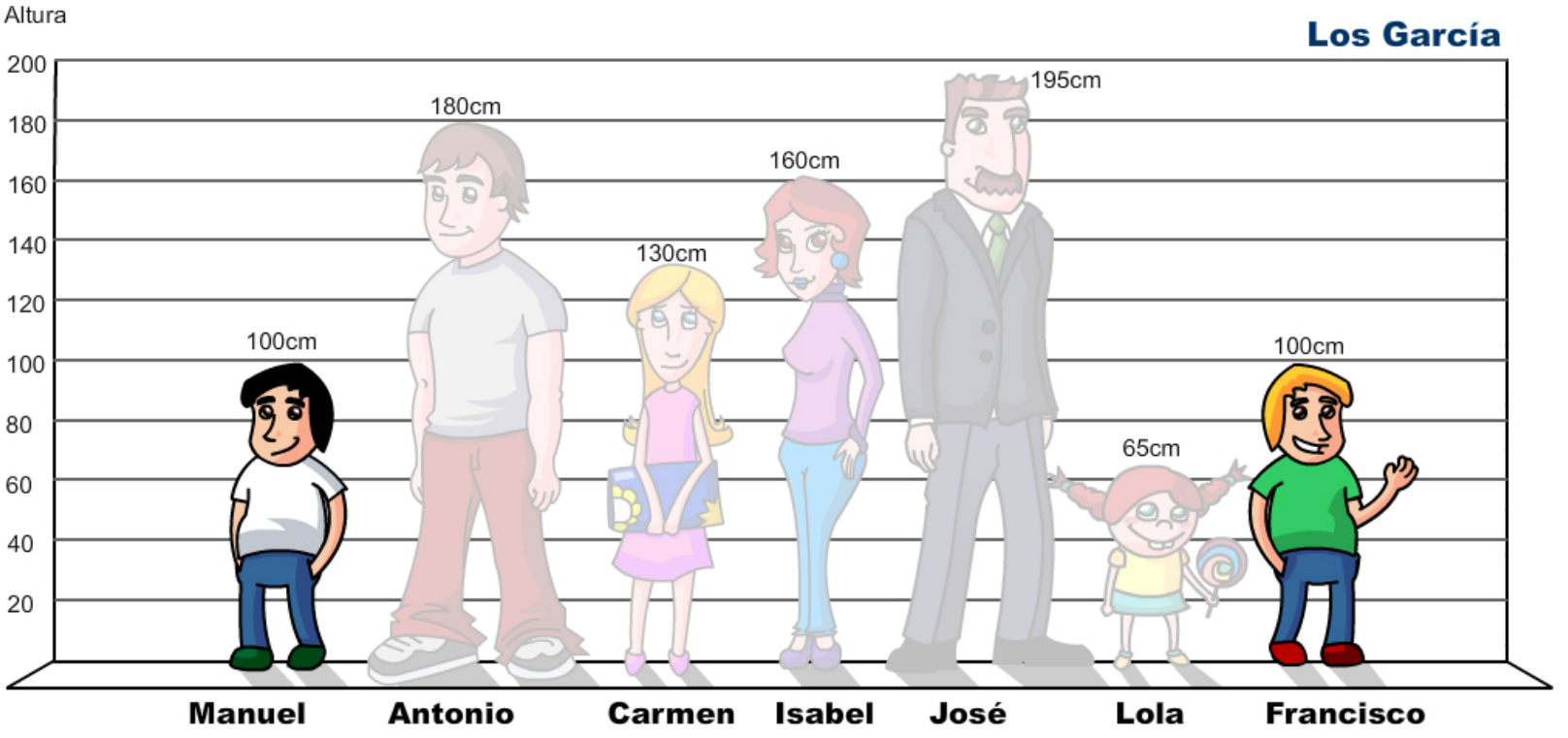
Medidas de centralización

Moda: es el valor que más se repite. En este caso es **100**; puesto que es el valor que más aparece (**2** veces).

Valor	Veces que aparece
65	1
100	**2**
130	1
160	1
180	1
195	1

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de centralización



Manuel y Francisco tienen una estatura que está *de moda*.

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de centralización

Mediana: para calcular la mediana necesitamos ordenar los datos (da lo mismo si de mayor a menor o de menor a mayor):

		Alturas ordenadas (cm)
1	Lola	65
2	Manuel	100
3	Francisco	100
4	Carmen	130
5	Isabel	160
6	Antonio	180
7	José	195

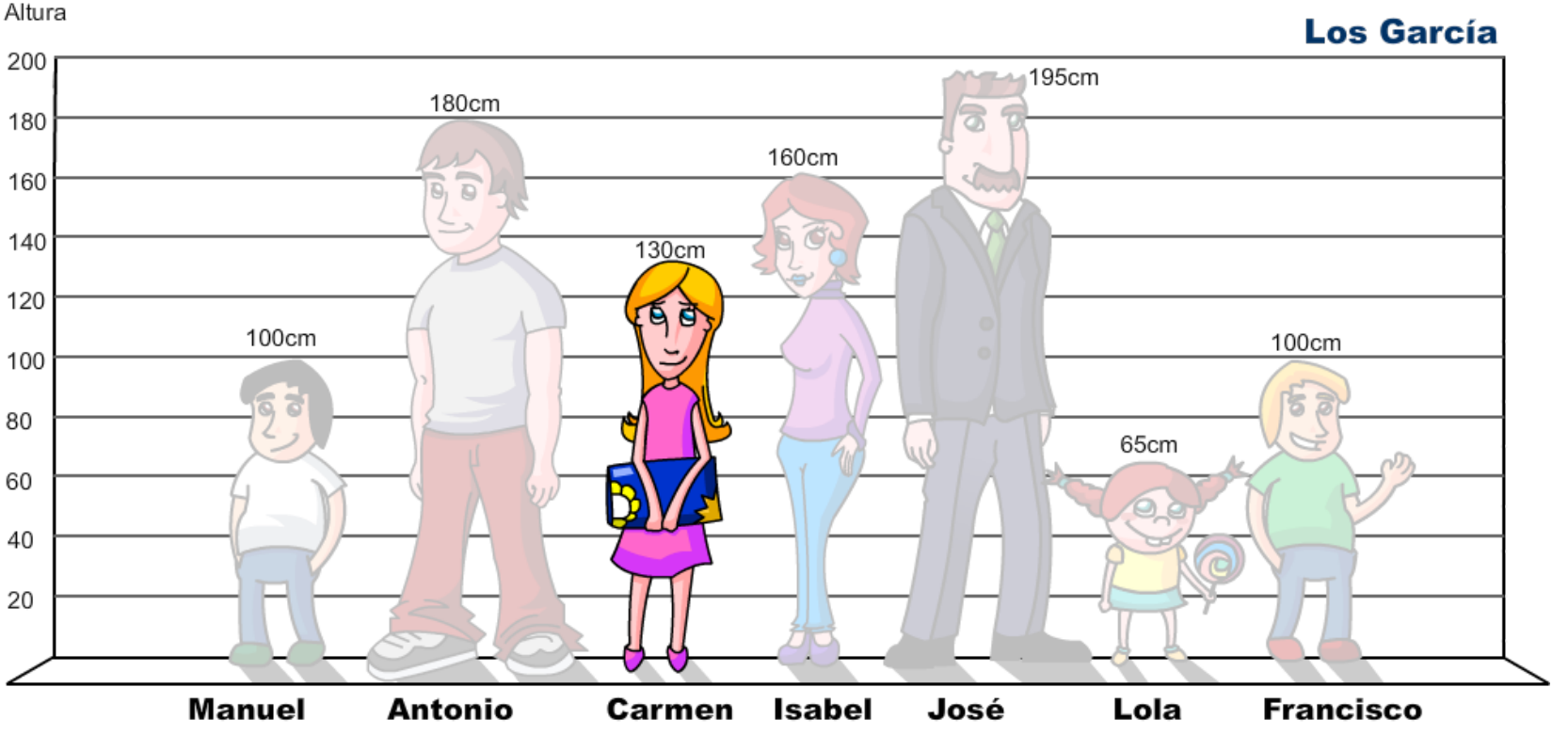
El **4º** valor es el que queda en medio, deja tres valores por encima y tres por debajo, luego la mediana es **130**.

En caso de que hubiera un número par de observaciones, la mediana se calcularía como la media de las dos observaciones que quedaran en medio.

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de centralización

La familia García



Al ser Carmen de estatura **mediana** sabe que hay tantas personas más altas que ella como personas más bajas.

Continúa Inicio

Problemas de las medidas de centralización

Pues yo no recuerdo haber comido pollo...

Lo que tienen las medidas resumen es eso, que son un resumen. Sintetizar varios datos en uno solo tiene muchas ventajas, básicamente porque manejar y almacenar un solo dato requiere mucho menos esfuerzo y recursos que manejar una tabla con un millón de números. Pero también hay inconvenientes, tal vez el principal sea que al reducir varios datos a uno sólo **es inevitable perder información**.

Si una persona se come cuatro pollos y otra ninguno, en media se han comido dos cada una. Seguramente para la persona que se haya quedado con hambre no le sea de mucho consuelo saber que *en media* él también ha comido pollo, dos en concreto. Este es un ejemplo muy instructivo de cómo la media a veces *falla*. ¿Qué ha sucedido? Que en este caso la media no es un buen resumen de la realidad de la población. Hay información importante, como el hecho de que la diferencia entre 4 y 0 es *muy grande*, que al hacer la media se pierde. Afortunadamente tiene arreglo, vamos a ver cómo con otro ejemplo.

Problemas de las medidas de centralización

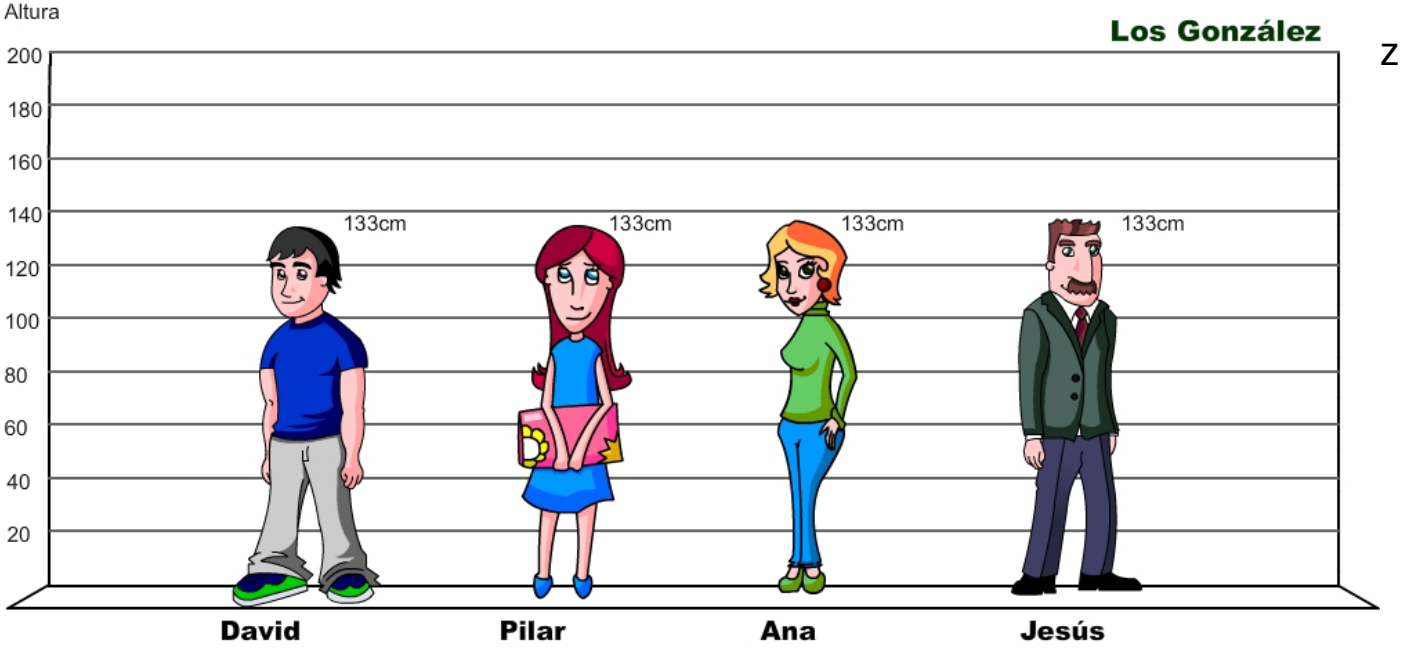
Vamos a estudiar las **alturas medias** de los miembros de dos familias distintas.

La primera es la familia García, de la que ya hemos estudiado la media, la moda y la mediana de sus alturas, y ya sabemos que la media de su altura es 133 cm.

Ahora vamos a estudiar a la familia González.

¿Qué acabo de recoger? La Estadística Descriptiva

Problemas de las medidas de centralización



z

$$\textbf{Media de altura} = \frac{133+133+133+133}{4} = \textbf{133} \text{ cm}$$

González es el segundo apellido más común a 1 de enero de 2009. Los nombres también son de los más comunes en esa fecha.

Problemas de las medidas de centralización

También podemos calcular la **moda** y la **mediana** de la familia González.

Moda: el valor que más se repite es el 133.

Mediana: como todos los miembros miden lo mismo ni nos hace falta ordenarlos; la mediana es también 133 cm.

Pero no queremos centrarnos en eso. Queremos estudiar qué información se pierde al hacer la media, y cómo solucionarlo.

Medidas de dispersión

Ambas familias tienen la misma media, pero son claramente distintas. La familia García tiene miembros de distintas alturas, mientras que en la familia González todos miden lo mismo. Aun así la media sale la misma para ambas familias. ¿Cómo podemos reflejar que las alturas varían mucho en la familia García, mientras que en la familia González no varían nada? Con las **medidas de dispersión** de los datos.

Las medidas de dispersión, como su nombre indica, dan una idea de lo que varían los datos de la muestra. Son muy útiles para evaluar la fiabilidad de las medidas de centralización como la media. Cuanto más alta sea una medida de dispersión, menos representativa será la medida de centralización.

Las medidas de dispersión más conocidas son el **rango** y la **varianza.**

El **rango** (R) se define como el valor máximo de las observaciones menos el mínimo. Cuanta más dispersión haya en los datos, mayor será el rango.

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de dispersión

Para calcular el rango es necesario ordenar los datos. Para la **familia García** el rango es:

		Alturas ordenadas (cm)
1	Lola	65
2	Manuel	100
3	Francisco	100
4	Carmen	130
5	Isabel	160
6	Antonio	180
7	José	195

En este caso el valor máximo es 195 y el mínimo, 65; de modo que el rango de los datos es: R = valor máximo – valor mínimo = 195 – 65 = **130 cm**

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de dispersión

La **varianza** (Var, o también S^2): media de cuánto se alejan los datos de la media

$$S^2 = \frac{(dato_1 - media)^2 + ... + (dato_n - media)^2}{n}$$

Con n el tamaño de la muestra.
Calculamos la varianza para la **familia García**. Primero calculamos el numerador

	Altura (cm)	Altura menos la media (133 cm)	Elevamos al cuadrado la altura - media
Antonio	180	47	2209
Carmen	130	-3	9
Francisco	100	-33	1089
Lola	65	-68	4624
Manuel	100	-33	1089
Isabel	160	27	729
José	195	62	3844
		SUMA	**13593**

Medidas de dispersión

Y con el numerador ya calculado, hacemos la división y obtenemos la varianza para las estaturas de la familia García. Esta familia tiene 7 miembros, así que n = 7. Por lo tanto:

$$S^2 = \frac{13593}{7} = \mathbf{1941{,}85}\ cm^2$$

Cuanto mayor sea la varianza, mayor será la dispersión de los datos. Y cuanto mayor sea la dispersión de los datos, menor será la representatividad de la media como resumen de la información de la muestra.

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de dispersión

Sin embargo para la **familia González** el rango es:

		Alturas ordenadas (cm)
1	David	133
2	Pilar	133
3	Jesús	133
4	Ana	133

En este caso el valor máximo coincide con el mínimo porque todos los valores son iguales.
R = valor máximo – valor mínimo = 133 – 133 = **0**

¿Qué acabo de recoger? La Estadística Descriptiva

Medidas de dispersión

La **varianza** (Var, o también S^2) es la media de la distancia entre los datos y la media.

$$S^2 = \frac{(dato_1 - media)^2 + ... + (dato_n - media)^2}{n}$$

Con n el tamaño de la muestra.

Calculamos la varianza para la **familia González**. Primero calculamos el numerador:

		Altura (cm)	Altura menos la media (133)	Elevamos al cuadrado la altura-media
1	David	133	0	0
2	Pilar	133	0	0
3	Jesús	133	0	0
4	Ana	133	0	0
			SUMA	0

Medidas de dispersión

Y con el numerador ya calculado, hacemos la división y obtenemos la varianza para las estaturas de la familia González. Esta familia tiene 4 miembros, así que n = 4. Por lo tanto:

$$S^2 = \frac{0}{4} = \mathbf{0}\ cm^2$$

Cuando las medidas de dispersión de una muestra son muy pequeñas se dice que la muestra es muy **homogénea**. Como puede verse, las alturas de la familia García varían mucho, y las de la familia González nada, y eso se refleja en los rangos (130 frente a 0) y en las varianzas (1941,85 frente a 0).

¿Qué conclusión sacamos entonces? Que las medidas de centralización son muy cómodas y útiles, pero deben venir siempre **acompañadas de una medida de dispersión**, que nos indica si realmente la medida de centralización resume bien la muestra (cuando la medida de dispersión es pequeña) o si por el contrario no recoge bien toda la información (medida de dispersión alta).

Relación medidas de centralización, dispersión y probabilidades

¡Un momento, un momento! En el primer apartado hemos dicho que para estudiar fenómenos aleatorios calculamos las probabilidades de sus resultados. Después hemos pasado a estudiar algún fenómeno aleatorio, hemos hablado de tomar muestras, de repente han aparecido las medidas de dispersión y centralización y de calcular probabilidades nunca más se supo...

Es cierto, de repente parece que al tomar una muestra nos hemos olvidado de calcular las probabilidades, por ejemplo, de que al tirar un dado 1000 veces el 6 aparezca sólo 3 veces, la probabilidad de que una bombilla dure encendida más de 512 horas seguidas, o la probabilidad de que la estatura de una persona de las que hemos ido a medir esté entre 180 y 197 cm. ¿Ya no son importantes? ¿Qué tienen que ver las medidas de centralización y dispersión con el cálculo de probabilidades?

Relación medidas de centralización, dispersión y probabilidades

Como ya hemos dicho en el primer apartado calcular la probabilidad de que un suceso tenga tal resultado no es fácil. Lo habitual es que vengan dadas por una fórmula, y no sea necesario ir resultado por resultado calculando su probabilidad. En ocasiones esa fórmula viene dada por una función de distribución concreta, como los ejemplos que hemos visto en el primer apartado de la binomial, geométrica, normal etc. En otras ocasiones directamente no conocemos la fórmula.

Por un lado, si no conocemos la fórmula para calcular probabilidades, indicadores como la media, la mediana, la varianza... nos pueden ayudar a aproximar su cálculo (para aquellos que ya sepan del tema, que miren por ejemplo los teoremas de Markov y Chebyshev). Por otro lado, las fórmulas están llenas de letras (**parámetros**) cuyo valor muchas veces es desconocido, y la media y la varianza nos pueden servir para estimar su valor. Y por último, pese a lo que hemos dicho, lo cierto es que no siempre queremos estudiar las probabilidades de los resultados. A veces sólo queremos conocer las principales características de la población o de los resultados de los experimentos porque con eso nos basta, y para eso las medidas de centralización y dispersión son de mucha ayuda.

Dato curioso: la sinopsis de los libros

Nos apetece comprarnos un libro, pero no tenemos nada decidido. Una solución es ir a la tienda y echar un vistazo. El procedimiento habitual es coger aquel libro que nos llama la atención y leernos el resumen, y en base a si nos gusta o no la sinopsis, decidimos comprar.

Sabemos que en el resumen no nos cuentan el libro entero. A veces las sinopsis sintetizan muy bien de qué va el libro y a veces no. Este resumen puede verse como una especie de media del libro. Si en el libro pasan muchas cosas (mucha variabilidad), la sinopsis no puede recogerlas todas, porque la sinopsis tiene que entrar en media contraportada y no se puede contar todo. Si por el contrario el libro consistiera en una sola frase repetida muchas veces (poca variabilidad) y la sinopsis fuese esa frase, que sí cabría en la contraportada, sería una sinopsis perfecta.

Dato curioso: ¿por qué al calcular la varianza se elevan las restas al cuadrado?

El concepto y cálculo de la varianza puede parecer complicado la primera vez que se ve. Cojo cada dato. Le resto la media para ver si el dato está cerca o lejos de la media. Lo elevo al cuadrado. Lo sumo todo. Divido entre el número de observaciones.

Entender que hacemos una media de lo que se aleja cada dato de la media es más o menos sencillo. Lo que no está tan claro es por qué de pronto elevamos al cuadrado las restas. ¿Qué tiene eso que ver con calcular esa media un poco extraña que es la varianza?

Que quede claro que nadie eleva las cosas al cuadrado porque sí, si se hace es porque hay una buena razón. En este caso el motivo suele estar relacionado con evitar que cantidades positivas y negativas se compensen, y nos salga una varianza que no refleja bien la dispersión.

Por ejemplo, medimos la temperatura de nuestra ciudad en varios momentos del año y obtenemos los siguientes resultados:

¿Qué acabo de recoger? La Estadística Descriptiva

Dato curioso: ¿por qué al calcular la varianza se elevan las restas al cuadrado?

	Temperatura (ºC)
Medida 1	+25
Medida 2	+15
Medida 3	+20
Medida 4	-15
Medida 5	-25
Medida 6	-20

Calculamos la media:

$$\textbf{Media de temperatura} = \frac{+25 + 15 + 20 - 15 - 25 - 20}{6} = \textbf{0}\ ^{o}C$$

¿Qué acabo de recoger? La Estadística Descriptiva

Dato curioso: ¿por qué al calcular la varianza se elevan las restas al cuadrado?

Y ahora calculamos la varianza. Parece una ciudad con temperaturas que varían mucho de los meses de frío a los meses de calor, así que es de esperar que la varianza sea grande. Vamos a ver qué pasa si no elevamos al cuadrado y nos limitamos a restar la media a cada valor y sumarlo todo

	Temperatura (ºC)	Temperatura menos la media (0)
Medida 1	+25	+25
Medida 2	+15	+15
Medida 3	+20	+20
Medida 4	-15	-15
Medida 5	-25	-25
Medida 6	-20	-20
	SUMA	**0**

$$\textbf{Varianza} = \frac{0}{6} = \textbf{0 ºC}$$

¿Qué acabo de recoger? La Estadística Descriptiva

Dato curioso: ¿por qué al calcular la varianza se elevan las restas al cuadrado?

La “varianza” sale 0. Al no elevar al cuadrado, las cantidades positivas se han compensado con las negativas y nos sale un resultado raro: ¡una varianza 0 significa que a lo largo de todo el año ha habido la misma temperatura! Y los datos nos dicen que eso es mentira: no se repite ni uno. Por esto se elevan las restas al cuadrado. Al hacerlo todos los datos se vuelven positivos, y ya no hay esas compensaciones que nos llevan a resultados engañosos:

	Temperatura (ºC)	Temperatura menos la media (0)	Elevamos la temperatura – media al cuadrado
Medida 1	+25	+25	625
Medida 2	+15	+15	225
Medida 3	+20	+20	400
Medida 4	-15	-15	225
Medida 5	-25	-25	625
Medida 6	-20	-20	400
		SUMA	**2500**

¿Qué acabo de recoger? La Estadística Descriptiva

Dato curioso: ¿por qué al calcular la varianza se elevan las restas al cuadrado?

La varianza correctamente calculada sale:

$$\textbf{Varianza} = \frac{2500}{6} \approx \textbf{416,67 ºC}^2$$

Y este es el motivo por el que se elevan al cuadrado cada una de las restas “dato – media”, antes de sumarlas todas y hacer la media.

Y recuerda: una temperatura media anual de 15ºC puede significar dos cosas: que o bien en esas ciudad todos los días hay 15ºC, o bien que en esa ciudad a veces hay 0ºC y a veces 30ºC. Y la pista sobre si es una situación u otra nos la dan las medidas de dispersión.

Introducción

Todo esto que hemos contado hasta ahora está muy bien, pero en el primer apartado hemos dicho que la Estadística es una herramienta que utilizan otras ciencias, y aún no hemos visto ninguna aplicación práctica, ningún ejemplo de para qué se usa la Estadística.

Cuando observamos un suceso al final lo que queremos es sacar conclusiones. La mayoría de las veces ni siquiera podemos observar todas las veces que el suceso ocurre, pero aún así queremos ser capaces de explicarlo y poder responder a preguntas como:

- ¿Qué variables influyen sobre el suceso? *(¿Influye la lluvia sobre el crecimiento de los cultivos? ¿Influye tu edad sobre la velocidad del viento?).*
- Las variables que influyen, *¿cómo lo hacen?* (mucho o poco, positiva o negativamente...).
- ¿Podemos elaborar un modelo que explique el suceso y permita **predecir** su comportamiento? (*como las presiones están descendiendo, lo más seguro es que llueva*).

Introducción

La **Inferencia Estadística** es la rama de la Estadística que recoge las técnicas que permiten responder a las preguntas anteriores. Se basa en el **uso de las observaciones** y de los indicadores descriptivos del apartado anterior (la media, la varianza, la mediana...). La Inferencia Estadística también establece cómo hay que planificar los experimentos para estudiar los sucesos. No vamos a explicar estos procedimientos en este apartado de conceptos básicos, puesto que requieren un conocimiento avanzado de Matemáticas y Probabilidad, pero sí vamos a dar algunos ejemplos.

Es muy importante indicar que la Inferencia Estadística arroja resultados sobre la población total en base a lo observado en una muestra (rara vez se puede estudiar la población entera) y muchas veces lo que estudia la Estadística son fenómenos aleatorios, por lo que no son resultados precisos, sino que siempre se dan con un **grado de incertidumbre**. La Inferencia Estadística dispone de herramientas para medir ese grado de incertidumbre. Veamos unos ejemplos.

La Inferencia Paramétrica

Ejemplo 1: en el primer apartado, al referirnos a la distribución de Poisson, ¿cuánto vale λ?

En la distribución de Poisson que hemos visto en el primer apartado, aparece una letra griega, **λ** (lambda), y de hecho hemos llamado a la distribución Poisson de **parámetro λ**. Además en muchas de las otras distribuciones que hemos visto salen las letras p y q como las probabilidades de un resultado u otro. Si no conocemos el valor de λ, o de p o q, no podemos hacer nada con las fórmulas ¿De dónde se obtiene el valor de estas letras (**parámetros**)*?*

El valor de los parámetros de las distribuciones normalmente es desconocido, pero la Inferencia Paramétrica nos da técnicas para **estimarlos** en base a lo observado en una muestra de la población.

La Inferencia no Paramétrica

Ejemplo 2: ¿Cómo sabes que es una binomial?

Ya hemos dado ejemplos de varias situaciones que podían explicarse según una distribución geométrica, hipergeométrica, binomial negativa... pero las cosas no siempre son sencillas, y dado un suceso aleatorio, en general no se conoce de antemano qué distribución de probabilidad sigue. No obstante la inferencia no paramétrica dispone entre otras cosas de técnicas que pueden ayudar a decidir si un suceso aleatorio sigue una distribución en particular.

La dependencia de variables

Ejemplo 3: La hipertensión y el consumo de sal

Hemos oído muchas veces que el consumo excesivo de sal eleva la tensión arterial. ¿Cómo lo saben los médicos? ¿Qué han visto para poder llegar a esa conclusión, que la sal influye sobre la presión arterial?

La inferencia estadística tiene una serie de indicadores (como la **covarianza** y la **correlación** entre variables, las **curvas de regresión** etc.) que permiten decidir si una de ellas influye sobre otra positiva o negativamente, así como estimar en qué medida influye una variable en otra.

Los modelos predictivos

Ejemplo 4: ¡Pero si no iba a llover!

La Meteorología es la ciencia de los sucesos atmosféricos. Entre otras cosas estudia qué pasa en la atmósfera -cómo se mueven las capas de aire, las variaciones de presión, humedad, temperatura- por qué pasa (para lo cuál se nutre de leyes físicas) y cómo algunas variables influyen sobre otras.

Una de las aplicaciones que tiene la Meteorología es que permite predecir el tiempo que va a hacer en el futuro (sea dentro de dos horas o en cuatro días), lo que resulta fundamental para los aeropuertos, transporte marítimo, agricultura... y también para la vida en general. Pero ¿quién no ha tenido alguna vez la sensación de que el hombre del tiempo a veces se equivoca? Predice lluvias y no llueve. Predice sol y llueve. ¿Qué pasa?

Los modelos predictivos

La predicción meteorológica no se hace al azar, ni a ojo. Se hace estudiando la evolución de muchas variables: presión, temperatura, humedad, velocidad y dirección del viento, radiación solar, movimiento de las masas frías / calientes de aire y un largo etcétera, e introduciéndolas todas en un modelo estadístico.

La predicción meteorológica tiene un componente aleatorio y eso hace que predecir el tiempo con exactitud no sea posible; el modelo no es perfecto y a veces *se equivoca*. Y además, cuanto mayor es el tiempo de previsión, más inexactitud. No es lo mismo predecir para dentro de dos horas que para dentro de una semana.

La Inferencia Estadística cuenta con métodos que permite predecir, con cierto margen de error, cómo va a comportarse una variable en función de los valores de otras, por ejemplo con las **curvas de regresión**.

El contraste de hipótesis

Ejemplo 5: ¡No puede ser casualidad!

Estamos en un laboratorio farmacéutico, y vamos a probar un nuevo medicamento para aliviar el dolor. Así que le pedimos a una persona a la que le duele la rodilla que se lo tome y espere media hora. La persona se lo toma y se le pasa el dolor. ¿Podemos concluir que el medicamento funciona?

Todavía no. A la persona se le ha pasado el dolor, sí, pero ¿y si ha sido casualidad? Los dolores normalmente se pasan, ¿y si justo le hemos dado la pastilla cuando el dolor ya se le estaba pasando? Y también hay otra cosa, muchas veces da igual que la pastilla funcione o no, con que nos creamos que va a quitar el dolor nos basta para que se nos quite (esto se conoce como *efecto placebo*), así que ¿y si se le ha pasado el dolor por efecto placebo? Necesitamos más datos.

El contraste de hipótesis

Muy bien, repetimos el experimento, y le pedimos a una persona a la que le duele la cabeza que se tome nuestra pastilla y espere media hora. No se le pasa el dolor. ¿Podemos concluir que nuestra pastilla no funciona?

Tampoco. Puede ser que la pastilla no funcione, sí, pero también puede ser que tarde más en actuar, que la persona a la que se la hayamos dado tenga más tolerancia a los analgésicos (y que por lo tanto necesite más cantidad para conseguir el mismo efecto), o puede que nuestra pastilla no funcione para los dolores de cabeza, pero sí sea útil para los dolores articulares. Así que... de nuevo, necesitamos más datos.

El contraste de hipótesis

Repetimos el experimento por tercera vez. En esta ocasión tenemos a 1.000 personas a las que les duele una articulación (rodilla, tobillos, muñecas...). Se toman la pastilla y de las 1.000 a 950 se les pasa el dolor. Que a una persona se le pase el dolor por casualidad y la pastilla no tenga nada que ver, nos lo creemos. Que a dos personas se les pase el dolor por casualidad y la pastilla no tenga nada que ver... también. Que a 950 personas se les pase el dolor por casualidad y la pastilla no tenga nada que ver... es más difícil creérselo. Ha podido ser por casualidad, sí, pero es **poco probable**. Si tuviéramos que apostar por algo, sería porque el medicamento funciona.

El razonamiento anterior tiene un nombre en Estadística: **contraste de hipótesis**. Contado muy rápido, un contraste de hipótesis es una técnica estadística que permite **decidir** si una propiedad se da o no en una muestra en base a lo observado. Lo que se estudia con el contraste de hipótesis, dicho de forma muy simplificada, es *Respecto a lo que ha pasado, ¿qué es más probable: que haya sido por casualidad o que realmente se dé la propiedad que estoy estudiando?*

Y por último...

Con esto concluye este apartado. Si has llegado hasta aquí enhorabuena, ya tienes las nociones básicas sobre qué esta ciencia, para qué puede usarse y las herramientas que emplea. Esperamos que te haya sido útil y que tengas ganas de ampliar conocimientos sobre este tema.

También puedes comprobar lo que has aprendido en nuestro test.

¿Te lo sabes?

Inicio