Manipulaciones y utilización de la estadística

por

Arantza Urkaregi Etxepare, Universidad del País Vasco-Euskal Herriko Unibertsitatea

Existen mentiras, grandes mentiras, y ... estadísticas.

Benjamín Disraeli

Se atribuye al político británico del siglo XIX Benjamín Disraeli esta famosa clasificación de las mentiras en tres tipos: mentiras, grandes mentiras, y ... estadísticas. ¿Mienten las estadísticas? ¿No es una ciencia la Estadística?
Los números no engañan, pero las personas que nos los presentan, a menudo, sí lo hacen.
Seguramente, sería más correcta formular la frase como:
"Existen medio mentirosos, mentirosos y estadísticos embaucadores".

Como apunta el estadístico Stephen K. Campbell, esta continua perversión de la estadística hace que la ciudadanía "en un principio, acepte las conclusiones estadísticas sin ejercer crítica alguna, por suponer que las cifras no mienten. A veces nos desalentamos con el solo hecho de que se nos ofrezcan afirmaciones que empiecen "según las estadísticas..." o "las estadísticas demuestran que ...". Pero, a veces, nos pasamos al extremo opuesto. Tendemos a creer que con las estadísticas se puede probar cualquier cosa, y por lo tanto no prueban nada. Mientras que en un momento creímos que las cifras no podían mentir, ahora se deduce que lo único que pueden hacer es engañar.

Una de las formas de engañar es **omitir una parte importante de la información**.

Comparemos estas dos noticias informativas:

- "El Estatuto de Autonomía de Andalucía ha sido aprobado con el 87 % de los votos. La abstención superó el 63 %".
- "Casi un 77 % de los españoles" ha ratificado la Constitución europea.

Esta segunda frase fue la afirmación que realizó el presentador del Telediario al día después del referéndum europeo. "Olvidó" añadir que se trataba del porcentaje sobre el número de votantes, y no sobre el total de la población, mayoritariamente abstencionista (58.23 %). La Constitución europea no fue aprobada por el 77 % de la ciudadanía española, sino por el 77 % de las personas que acudieron a votar. En definitiva, sólo el 32 % del censo ratificó la Constitución europea.

Cuando un medio de comunicación quiere impresionar a su audiencia con la gravedad de una situación, suele emplear números absolutos en vez de porcentajes: "50 muertos en el último puente de cuatro días".

Si la noticia se acompañara de las estadísticas de muertes por accidente de tráfico, observaríamos que se trata de aproximadamente el mismo valor que el número de víctimas por accidente de tráfico en cualquier periodo de cuatro días.

Pero, entonces, el titular no sería noticia... Se trata de lo que se llama falacia de "base extensa".

Esta argucia se suele emplear también en forma inversa. Por ejemplo, es frecuente leer en los periódicos afirmaciones como: "El número de asesinatos en la ciudad ha aumentado un 60 % respecto al pasado año."

Si el periodista o el político nos dijeran que el año pasado se cometieron 5 homicidios y éste 8, probablemente el dato no nos impactaría de la misma manera.

Y, sin embargo, en ninguno de los dos ejemplos podemos decir que el periodista o el político hayan mentido, simplemente han dicho una media verdad, la que les interesaba.

Lo mismo ocurre con noticias como la siguiente:

DOMINGO, 21 ENERO 2007 POLÍTICA LA VANGUARDIA 15

SONDEO DEL INSTITUTO NOXA PARA 'LA VANGUARDIA' ▸▸ *El proceso de paz en el País Vasco*

Más de la mitad culpa al PP de la desunión frente al terrorismo y el 82% apuesta por un pacto de todos los partidos

La mayoría avala el intento de diálogo con ETA

BATASUNA

El 77% rechaza que los abertzales puedan concurrir a las elecciones

Estos porcentajes, ¿se refieren a la población o a la muestra? ¿Cuántas personas han sido entrevistadas?

1. Población vs. muestra

Cuando se dan datos estadísticos, es necesario señalar si los datos recogidos se refieren a la población o constituyen una muestra (subconjunto de la población).

En el caso más habitual de que se trate de una muestra, ésta debe haber sido seleccionada de forma que sea representativa de la población. Cuanto más representativa sea la muestra, obtendremos mejores estimaciones para los parámetros de la población.

La primera condición para que una muestra sea representativa, sus elementos deben ser elegidos al azar. Pero eso es mucho más difícil de lo que parece. Por ejemplo, si la encuesta es telefónica, dejaremos fuera a aquella personas que no tienen teléfono o que no están en casa cuando llamamos. Si pedimos a las y los lectores de una publicación, ya sea en papel o electrónica, que den su opinión, estaremos construyendo una muestra autoseleccionada. Sólo las personas que leen esa publicación, y entre ellas las verdaderamente comprometidas con la cuestión, perderán su tiempo en dar su opinión. ¿Qué validez tiene entonces una información tan sesgada?

Para conseguir una muestra representativa de la población es necesario aplicar la técnica de muestreo adecuada, que se seleccionará en base a los objetivos de la investigación que se quiere realizar. Analicemos el ejemplo de las estimaciones de los resultados electorales:

1) El recuento de papeletas realizado en una muestra de mesas electorales, proporcionan una estimación muy precisa del verdadero resultado. Esto es así, porque la muestra seleccionada de mesas electorales, se obtiene en base a anteriores resultados electorales y se logra así que sea muy representativa.

2) Cuando la estimación de dichos resultados se basa en entrevistas realizadas en las puertas de los Colegios Electorales, la precisión de las estimaciones varía muy notablemente. Es lógico que la muestra sea menos representativa, ya que, por ejemplo, está condicionada por el hecho de que las personas preguntadas quieran responder o no.

3) El recuento final de los votos nos da el resultado real, dado que, en este caso, no se trata de una muestra, sino de la población.

Una mala selección de la muestra da origen a sesgos en la información estadística:

- Sesgos producidos por errores en la selección de las mesas electorales y al modo de selección de las personas que salgan de ejercer su derecho al voto. Estos errores son medibles por procedimientos estadísticos y se pueden solucionar.

- Sesgos producidos por ocultamiento de información o falseamiento de la misma.

En el primer caso, cuando se trata de entrevistas a personas que acaban de ejercer su derecho a voto, el error se puede reducir mediante un riguroso procedimiento de sustitución de quienes se han negado a declarar su voto.

Sin embargo, el falseamiento de la respuesta tiene una solución difícil y sólo sobre la base de experiencias anteriores resulta posible aplicar índices de corrección a los datos agregados.

Las ocultaciones de información no son novedosas. En la Revista Estudios de Economía Aplicada, José Aranda Aznar[1] señala algunos ejemplos del siglo XVIII:

- En el Censo de Floridablanca de 1787 existe una advertencia en la que el redactor duda sobre el aumento de población entre 1768 y 1787, estimando que debería ser un 45 por ciento superior.

  Seguidamente da como razón de esta subestimación de población "el cuidado con que los pueblos y sus vecinos procuran disminuir el número de sus habitantes, temerosos de que tales numeraciones se dirijan a aumentar las cargas de los servicios personales o de los tributos".

- Lo mismo ocurre con el Censo de Frutos y Manufacturas de 1799, que pretendía nada menos que dar "una razón de los frutos y de las manufacturas que ha producido cada Provincia en dicho año; sus precios corrientes, la cantidad que ha consumido y sobrado; la proporción que hay entre los productos y la población, y entre esta y la extensión territorial; y asimismo los lugares donde se hallan establecidos los artículos principales de la industria".

  Tan ambicioso proyecto no debió ser un éxito por el comentario que figura en la citada publicación: "La poca exactitud que se encuentra en muchos de los estados remitidos por los Intendentes; las faltas que se han notado en algunos, y la obscuridad que han presentado otros, hicieron demasiadamente trabajosa la redacción de este Censo, el cual carece para ello de la certeza que desearán los que la leyeran".

2. Error de estimación

La Estadística no proporciona la "verdad", nos acerca a la realidad con un cierto nivel de confianza o un cierto margen de error.

[1] Estudios de Economía Aplicada, 1995, vol. 3, pag. 6-11

Una vez seleccionada adecuadamente la muestra, hay que **señalar el error que se está cometiendo al generalizar los datos obtenidos en la muestra a la población**.

Volvamos al ejemplo de La Vanguardia 21 - 01 - 2007: "Un 76 % de los consultados por el Instituto Noxa avala el diálogo con ETA".

Efectivamente, se trata de una muestra y, por tanto, se debe indicar el error de la estimación. Sin embargo, no hemos encontrado a lo largo del artículo ninguna mención sobre el número de personas encuestadas, es decir, sobre el tamaño de la muestra. Y el error de estimación está relacionado con el tamaño muestral: a mayor tamaño muestral, menor error de estimación.

Así, si se han realizado 2.000 encuestas, con un 95 % de confianza, podemos afirmar que el porcentaje que avalaría, en el conjunto de la población española, el diálogo con ETA estaría entre: $76 \pm 2{,}2 = (73{,}8, 78{,}2)$.

Con 1.000 encuestas y el mismo nivel de confianza, este porcentaje se situaría entre el 72.9 % y el 79.1 % ($76 \pm 3{,}1$).

3. El problema de la no-respuesta

Otro elemento que contribuye a falsear la información estadística son los **datos ausentes**.

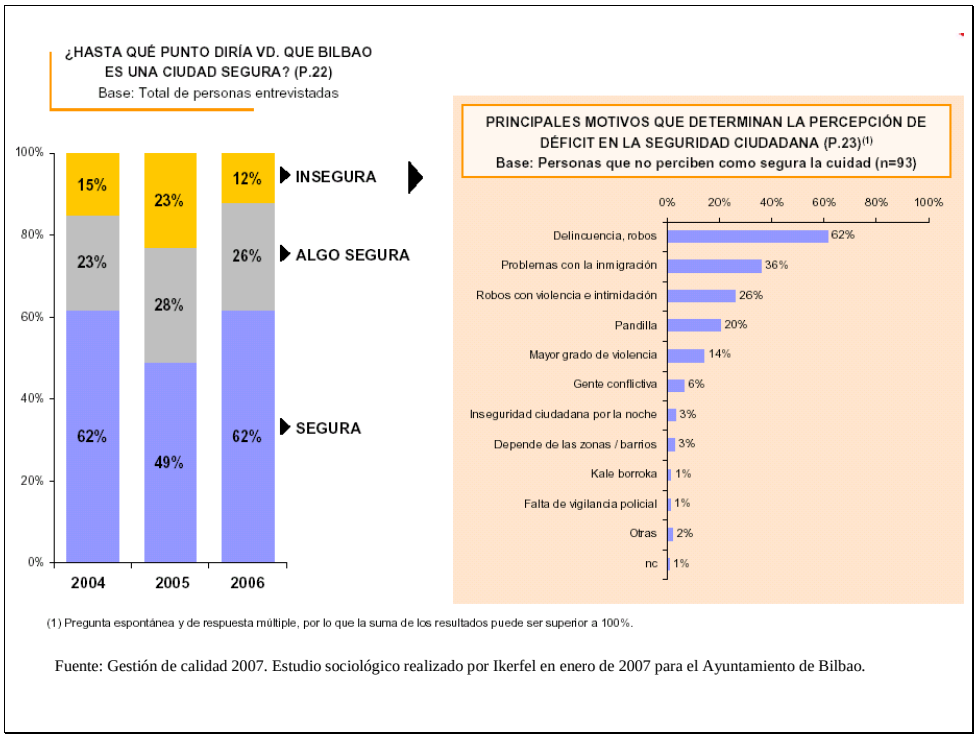
En un estudio realizado por Ikerfel para el Ayuntamiento de Bilbao, se indica que se han realizado 800 entrevistas telefónicas, lo que da lugar, con un nivel de confianza del 95 %, a un error de estimación de 3,5 %.

Sin embargo, no todas las personas han respondido a todas las preguntas y, en consecuencia, el error cometido para cada pregunta será diferente.

Si consideramos la pregunta sobre seguridad ciudadana (P22), observamos que el 12 % de personas entrevistadas considera que Bilbao es una ciudad insegura. Por lo tanto, con un 95 % de confianza, la proporción correspondiente para la población estaría en el intervalo $12 \pm 3{,}5 = (8{,}5, 15{,}5)$. Es decir, con un 95 % de confianza, podemos afirmar que entre el 8.5 % y el 15.5 % de la ciudadanía bilbaína considera que Bilbao es una ciudad insegura.

Sin embargo, si observamos la P23 (Principales motivos que determinan la percepción de déficit de la seguridad ciudadana), se indica que el tamaño muestral es 93, el número de personas que no consideran a Bilbao como ciudada segura. La proporción señalada en el gráfico, de un 36 % que relaciona la inseguridad con la inmigración se refiere a un 36 % de esas 93 personas, por lo que, si quisiéramos generalizar esa proporción para el conjunto de la ciudadanía bilbaína, el error de estimación no sería del 3.5 %, sino del 10.2 %, por lo que esta estimación sería mucho menos exacta; es decir, con un 95 % de confianza, la propoción de la ciudadanía bilbaína que relaciona la inseguridad con la inmigración estaría entre un 25.8 % y un 46.2 %, lo que nos da un error de estimación excesivamente grande,

dado el reducido tamaño muestral.



En definitiva, cuando trabajamos con datos estadísticos relativos a una muestra de la población, debemos tener en cuenta que:

- La muestra debe ser representativa de la población.
- Cuanto más pequeña es la muestra mayor es el error cometido.
- Es necesario reducir al máximo la falta de respuesta.

4. Forma de resumir la información: media vs. mediana

En un estudio que se está realizando en el Ayuntamiento de Bilbao, se quiere analizar la proporción media de mujeres entre sus trabajadores. Se han seleccionado 4 áreas con los siguientes datos:

Area	% **mujeres**
Obras y Servicios	40
Mujer y Cooperación al Desarrollo	92
Acción Social	90
Urbanismo	50

El Ayuntamiento nos indica que la proporción media es del 68 %, es decir, han calculado la media aritmética:

$$\overline{x} = \frac{40 + 92 + 90 + 50}{4} = \frac{272}{4} = 68.$$

¿Es cierta esta afirmación?

Primera cuestión: ¿Hay el mismo número de trabajadores/as en estas áreas? Nos proporcionan la siguiente información:

Area	**nº mujeres**	**nº total trabaj.**
Obras y Servicios	60	150
Mujer y Coop. al Desarrollo	46	50
Acción Social	63	70
Urbanismo	40	80
Total	**209**	**350**

¿Cuál es la proporción media de mujeres en estas 4 áreas?

$$\overline{x} = \frac{209}{350} = 59,71.$$

Es decir, hay que calcular la media ponderada de las proporciones de cada área, siendo los pesos el número de trabajadores/as de cada área.

Pero no siempre es adecuado resumir un conjunto de datos mediante la media, sea ésta aritmética o ponderada.

La media aritmética

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{k} x_i f_i}{n},$$

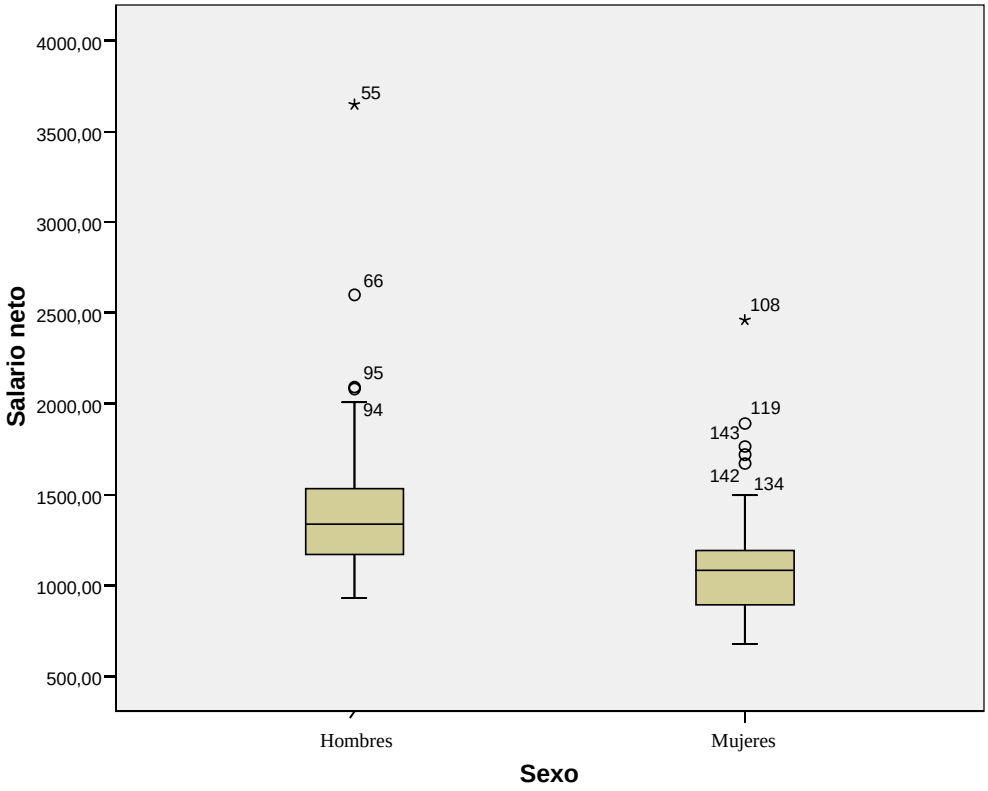
es representativa cuando los datos no están muy dispersos, pero no lo es cuando la dispersión es grande, debido a que los valores extremos tienen mucha influencia en la media.

Cuando tenemos un grupo de datos dispersos, el estadístico que mejor representa a ese conjunto de datos es la mediana: el valor que divide a ese conjunto de datos ordenado en dos partes iguales.

$$F(Me) = f(X \leq Me) = \frac{n}{2}.$$

El análisis de la estructura salarial es un ejemplo en el que el salario medio no es representativo de dicha estructura salarial.

La representación del salario neto en base a los datos de la Encuesta de estructura salarial realizada por el INE en 2002 nos proporciona el siguiente gráfico, en el que observamos las diferencias por sexo, así como la existencia de unos pocos salarios muy altos respecto al conjunto (mayores en los hombres que en las mujeres). Estos salarios altos van a tener una gran influencia en la media, por lo que, a la hora de representar la estructura salarial, el estadístico de tendencia central más adecuado es la mediana y no la media:



Cuando describimos un conjunto de datos mediante la media, el estadístico de dispersión asociado es la desviación estandar:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}.$$

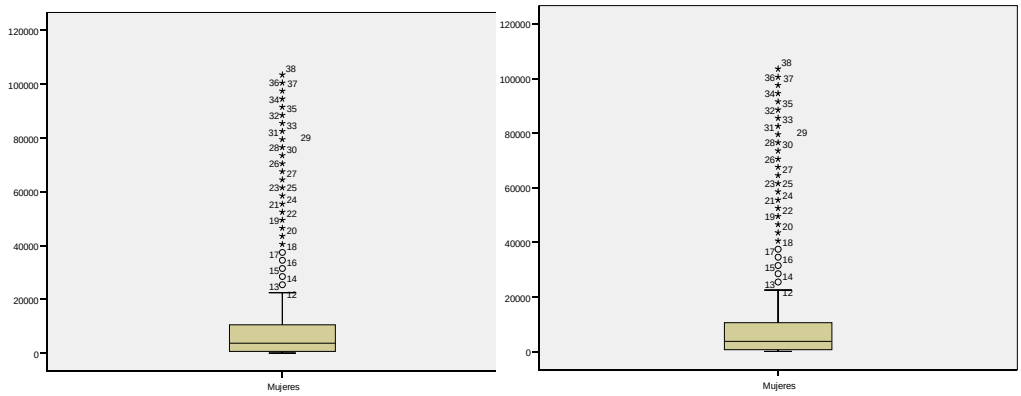
Mientras que cuando utilizamos la mediana, el estadístico de dispersión asociado es el rango intercuartílico:

$$RI = Q_3 - Q_1.$$

Si comparamos estos estadísticos para el salario neto diferenciados para hombres y mujeres, obtenemos la siguiente tabla:

Salario neto		
Media	Hombres	1442,52
	Mujeres	1112,06
Desv. Stand.	Hombres	461,04
	Mujeres	332,93
Mediana	Hombres	1336,06
	Mujeres	1082,03
Rango Interc.	Hombres	366,65
	Mujeres	341,65

Si analizamos la distribución de la renta personal de la población de la CAV con 18 y más años en 2001 obtenemos también una distribución con altos valores extremos, como se observa en el gráfico:



Los estadísticos correspondientes a la renta anual en la CAV se recogen en la siguiente tabla:

Renta	Media	Desv. Stand	Mediana	Rango Intercuartílico
Global	12.526,19	14.166,41	9.520,47	18.750,00
Hombres	17.467,75	15.525,63	16.500,00	15.508,00
Mujeres	7.882,16	10.852,92	3.750,00	9.750,00

Según los datos de la tabla, el 50 % de los hombres empleados de la EAE tiene una renta menor de 16.500 euros, mientras que el 50 % de las mujeres empleadas de la EAE tiene una renta menor de 3.750 euros. Sin embargo, si utilizamos la media, se obtiene que la renta media de las mujeres sería de 7.882 euros, frente a los 17.467,75 euros de los hombres. Las desviaciones típicas son muy grandes, incluso mayores que las medias correspondientes, lo que nos indica una gran dispersión, por lo que deberíamos utilizar la mediana de la renta como estadístico más representativo del conjunto de las rentas, tanto a nivel global, como en función del sexo.

5. A modo de conclusión

La aplicación de la estadística en diferentes campos de la ciencia se ha ido incrementando a medida que se ha extendido la utilización de los ordenadores y se han popularizado los paquetes estadísticos.

Esta amplia utilización de la Estadística tiene elementos positivos, pero también negativos: no es suficiente introducir los datos y tratarlos con un paquete estadístico para realizar un buen análisis estadístico.

Consejos para una buena utilización de la Estadística:

- Definir claramente los objetivos de la investigación.
- Seleccionar la muestra de forma aleatoria, asegurando la representatividad de la misma y su adecuación a los objetivos.
- Depurar los datos y reducir al máximo los valores ausentes.
- Estudiar los datos y seleccionar los estadísticos más adecuados para resumir la información de los datos.
- Presentar los datos indicando la metodología seguida, el nivel de confianza y el error que se puede cometer.

La siguiente noticia aparecida en El Correo, incluyendo la ficha técnica, es un buen ejemplo de cómo se deben presentar los resultados estadísticos:

EL CORREO, 21 – 01 – 2007

El PP aventaja al PSOE en 2,4 puntos en intención de voto tras el atentado de ETA

Los populares obtendrían entre 163 y 169 escaños frente a los 148 actuales, mientras los socialistas pasarían de 164 a entre 142 y 152

El Partido Popular, con una expectativa de voto del 40,7%, aventaja en 2,4 puntos al PSOE (38,3%), según los resultados del sondeo elaborado por el Instituto Metra-Seis, en exclusiva para la agencia Colpisa, sobre una muestra de 2.000 entrevistas realizadas entre los días 7 y 15 del presente mes de enero. La traducción en escaños de estos porcentajes de votos permitiría al PP obtener entre 163 y 169 escaños frente a los 148 actuales, mientras que el PSOE pasaría de 164 diputados a 142-152. Los populares parecen así romper, de manera coyuntural, el recurrente empate técnico que le viene atribuyendo la mayoría de las encuestas publicadas durante los últimos meses.

El resto de partidos con representación parlamentaria apenas sufriría alteración, aunque CiU e IU se muestran al alza y podrían obtener 1 ó 2 diputados más de los 10 y 5, respectivamente, que tienen ahora; el PNV mantendría sus 7 escaños, mientras ERC figura a la baja, con la posible pérdida de uno de sus 8 escaños.

La encuesta pone de manifiesto que el PP mantiene unos índices de fidelidad entre sus votantes más elevados de los que es capaz de conservar el PSOE. O, dicho de otro modo, la proporción de votantes fieles al PP -haga lo que haga- es significativamente más alta que la de los fieles del PSOE.

FICHA TÉCNICA

Universo y ámbito: Personas de 18 y más años, censadas en cualquier municipio del territorio nacional.

Muestra: 2.000 personas, con afijación proporcional a la población residente en las distintas comunidades y niveles de hábitat.

Error estadístico de los datos obtenidos: +- 2,2%, en el caso más desfavorable, con una probabilidad del 95,5% (2 sigma).

Tipo de encuesta: Telefónica, asistida por ordenador (CATI).

Sistema de selección: Aleatoria de teléfonos/hogares y cumplimentación de cuotas cruzadas de género y edad en los hogares seleccionados.

La Estadística nos permite acercarnos a la realidad, pero siempre que actuemos en base a unas determinadas reglas que posibiliten la aplicación de las técnicas estadísticas, sin falsear esa realidad.

Yale y Kendal (1954) definen la Estadística como la Ciencia que trata de la recolección, clasificación y presentación de los hechos sujetos a una apreciación numérica como base a la explicación, descripción y comparación de los fenómenos. Espero que esta charla os haya servido para no creer a ciegas en afirmaciones que dicen basarse en la Estadística, pero al mismo tiempo os haya animado a profundizar en la estadística, para utilizarla y hacerlo de una manera adecuada.

Arantza Urkaregi Etxepare
Universidad del País Vasco-
Euskal Herriko Unibertsitatea
Zientzia eta Teknologia Fakultatea
Matematika Aplikatua, Estatistika eta I.O. Saila
Sarriena auzoa, z/g, 48940 Leioa
e-mail: ***arantza.urkaregi@ehu.es***