Macro versus micro: La población desde el individuo, pasando por los datos.

«Existen tres tipos de engaños: las mentiras, las grandes mentiras y las estadísticas»

Popularizada por ***Mark Twain***

«Si queremos que la gente crea algo, lo único que hay que hacer es organizar una encuesta que diga tal cosa y después darle publicidad, preferiblemente por televisión»

Daniel Estulin

Introducción.

En el proceso de información y toma de decisiones que sigue una persona, empresa o incluso algoritmo de inteligencia artificial, es innegable que la Estadística juega un papel crucial. No en vano, se dice ya frecuentemente que *los datos* son la materia prima del s. XXI (comparándolo incluso con el petróleo) y que éste es *el siglo de los datos*. Por tanto, la alfabetización estadística de los individuos es crucial para evitar manipulaciones por parte de sus congéneres a través de informaciones sesgadas mediante el uso interesado de ciertos datos o interpretaciones de gráficos. En esta Unidad Didáctica proponemos actividades para acercarnos a diferentes formas de hacer estadística y comprender la información que está detrás de ésta.

Tradicionalmente, la Estadísticia se puede subdividir en dos grandes bloques, según los objetivos que se persigan con los correspondientes procesos, a saber: la estadística descriptiva, que únicamente pretende recolectar, organizar e interpretar los datos recabados en una muestra; y la estadística inferencial, que aspira a obtener conclusiones globales (macro) sobre la población a través de los datos particulares (micro) de las muestras o los individuos. Finalmente hay un retorno de las conclusiones sobre la población para obtener predicciones sobre el individuo o un suceso puntual concreto, que se realiza a través de las distribuciones de probabilidad.

Sin embargo, con el auge de los ordenadores y la programación informática, existen hoy día muchos procedimientos para recabar datos y obtener conclusiones que nada tienen que ver con las tradicionales encuestas y su representación en gráficos sobre ejes coordenados, ni con el cálculo de parámetros estadísticos. En particular, tras un recorrido por nociones rudimentarias que nos permitirán comprender los fundamentos de procesos más complejos, hacemos una breve aproximación al sistema de 'Cookies' al navegar por Internet que nos servirá para atisbar qué significa el archimencionado 'Big Data'.

Interpretando datos.

Los estudios estadísticos tradicionales se componen, en esencia, de tres grandes fases:

1. Muestreo y recogida de datos.
    - Tipo de variable y población de estudio.
    - Selección de la muestra: aleatoria, por estratos, etc...
    - Proceso de recabado: encuesta, entrevista, etc...
2. Organización y representación de la información.
    - Parámetros estadísticos:
        - De posición: mediana, media, cuartiles, ...
        - De dispersión: varianza, desviación típica, ...
        - Dependencia entre varias variables: covarianza, correlación, ...
    - Gráficos: diagramas de sectores, barras, cajas y bigotes, etc...
    - Varias variables: rectas y otras curvas de regresión.
3. Interpretación de los resultados.
    - Obtención de conclusiones a partir de los puntos anteriores.

Con este breve resumen en mente, **visitia alguna web con estudios estadísticos** como por ejemplo la web del INE (Instituto Nacional de Estadística) o la web del CIS (Centro de Investigaciones Sociológicas), elige alguno de sus estudios y trata de determinar sus elementos más importantes a partir del **análisis de la ficha técnica.** Como ejemplo, proponemos la encuesta realizada por el CIS en 2021 llamada "Infancia y Juventud ante la pandemia de Covid-19".

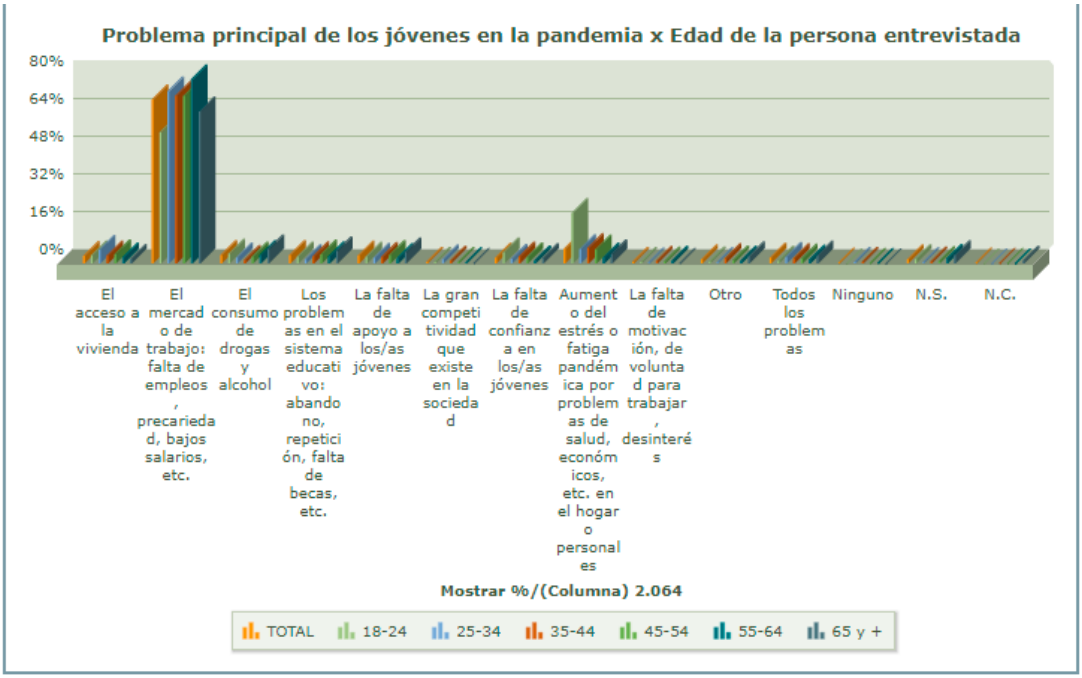

Imagen 1. Representación en diagrama de barras de las respuestas obtenidas en la encuesta del CIS a la pregunta sobre el problema principal que se considera sufrirán los jóvenes, segmentando por la edad del individuo encuestado.

Debes tener en cuenta que, a medida que el tamaño de la muestra crece, los datos y la información aportada se parecen más a los reales de la población pero por contra su tratamiento se vuelve más inoperativo. Sin embargo, esta idea es la que permite pasar de los hechos concretos de un individuo (que es un ente discreto) al comportamiento general teórico de toda la población (que se trata como un continuo).

En concreto, gracias a la ***Ley de los Grandes Números*** podemos asegurar que el valor esperado de un hecho coincide con el valor observado a través de una muestra. **Busca información sobre este importante teorema,** trata de comprender su enunciado y explícalo con tus propias palabras. También es de extremada importancia el ***Teorema del Límite Central,*** del cual un caso particular es la aproximación de la distribución Normal a través de la Binomial y que puede visualizarse mediante un experimento atribuido al inglés Sir Francis Galton.

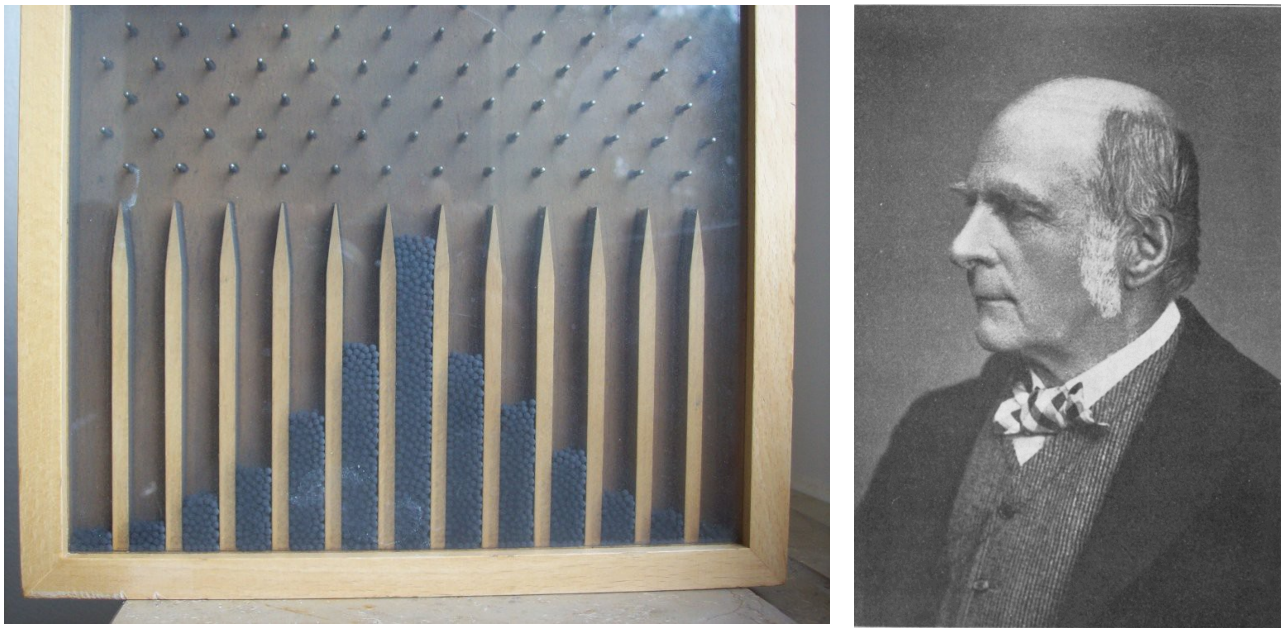

Imagen 2. A la izquierda, se ilustra cómo la repetición de un fenómeno binomial da lugar a la aproximación de la distribución normal. A la derecha, imagen de Sir Francis Galton.

Quizá uno de los ejemplos prácticos más claros en el que se aplica este salto del discreto al continuo sean las predicciones de resultados electorales que se realizan tras las encuestas sobre intención de voto, como por ejemplo esta realizada en la Comunidad de Madrid para las elecciones de abril de 2021.

Sin embargo, los resultados oficiales finalmente fueron muy diferentes de los esperados... **Realiza un análisis comparativo** entre los resultados esperados y los que efectivamente se dieron en las elecciones. ¿A qué crees que se debe esta disonancia? Vuelve a la ficha técnica de la encuesta y presta atención a los datos que hablan del nivel de confianza del estudio y el margen de error que se acepta para considerar acertada la previsión dada por la encuesta.

Una (buena) imagen vale más que mil datos.

La mejor forma de hacer llegar las conclusiones de un estudio estadístico es, sin duda, la que entra por los ojos. La ruta visual es potente y por tanto capta e interpreta la información mostrada con menor esfuerzo que la que nos llega a través del canal auditivo. En ocasiones, la información plasmada en gráficos resulta un bastón en el que apoyar un eslógan publicitario, un titular informativo o un mantra electoral y no siempre le es fiel a los datos en los que dice soportarse.

Ya sea de forma (mal) intencionada, por desconocimiento estadístico o por un simple error, el individuo que recibe la información debe permanecer alerta para detectar, con un análisis crítico, las ocasiones en que vemos representaciones gráficas que exageran o distorsionan alguna característica del estudio. **Analiza los dos ejemplos siguientes** y discute, argumentando tu postura, si transmiten una información visual adecuada:

Imagen 3. Dos gráficos estadísticos desajustados. A la izquierda, comparativa presupuestaria aparecida en Telemadrid el 05/10/2010 y, a la derecha, datos de desempleo aparecidos en TVE el 21/01/2015.

En cualquier caso, no siempre la estadística va asociada a parámetros o gráficos estadísticos en el sentido usual que les atribuimos. Uno de los primeros casos documentados del tratamiento de datos y obtención de conclusiones a partir de la representación de frecuencias es el que permitió resolver al médico John Snow un brote de cólera en Londres en 1854.

Observa el siguiente plano de los alrededores de Broad Street, en el que Snow fue señalando el domicilio de los difuntos aquejados de cólera. ¿Qué conclusiones puedes extraer a partir de la representación? ¿Cómo crees que finalmente se resolvió el brote? Tras realizar algunas conjeturas, busca la información del caso real.

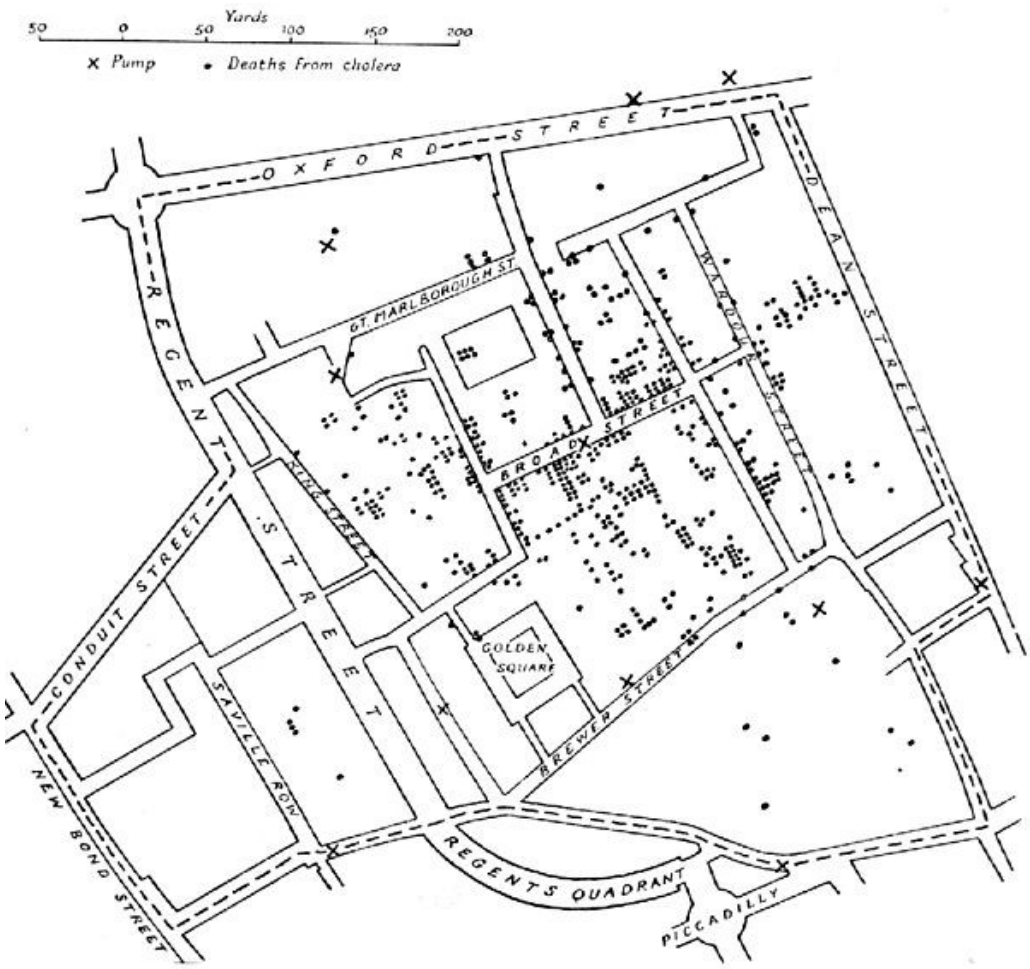

Imagen 4. Réplica del plano de los alrededores de Broad Street que elaboró J. Snow con las defunciones producidas por cólera en 1854.

Los datos en la era de Internet.

Como puedes imaginar, el desarrollo de los ordenadores y, en particular, de la tecnología que permite la navegación por internet, ha hecho que la recogida de datos y su tratamiento evolucione de una forma muy significativa en las últimas décadas. Cualquier diseñador de contenidos web (blogs, vídeos, etc...) está muy interesado en conocer cómo un usuario particular se comporta al visitar su dominio.

Esta recogida de información se realiza a través de las 'cookies', que son pequeños archivos que se instalan en el dispositivo, ya sea un ordenador o un teléfono móvil, y que permite recordar las preferencias del usuario y sus costumbres de navegación. La información que se recaba a través de las cookies puede utilizarse con diversas finalidades y,entre otras, podemos agruparlas en dos grandes secciones:

- Cookies estadísticas (analíticas): tienen por objetivo que el dueño del dominio conozca las ostumbres de navegación de los usuarios, qué partes de su web se visitan con más frecuencia, qué pestañas se despliegan, la procedencia de las visitas, el tiempo que pasan navegando en la web, etc... Esto le sirve para entender si el portal está bien construido o si hay algunos enlaces que llaman poco la atención o no se visitan.

- Cookies de seguimiento (tracking): los motores de búsqueda pueden posicionar los resultados que se muestran según el número de visitas que reciben y los enlaces que llegan a esas páginas desde otras. Además, las empresas comerciales muestran diferentes anuncios en las banners de publicidad según las preferencias detectadas (dos personas visitando la misma página visualizarán diferentes productos en las ventanas destinadas a mostrar publicidad) y almacenan información sobre las costumbres de navegación del usuario.

Todo el conglomerado de datos del cual vamos dejando rastro cada vez que accedemos a un servicio a través de Internet, suele nutrir a los algoritmos de 'Machine Learning' que son un caso concreto de 'Big Data'. Este último es en general cualquier tipo de proceso que requiera del manejo de una cantidad enorme de datos y, en particular, el Machine Learning es cualquier tipo de algoritmo que en su programación haga uso de los datos sin la intervención de un humano y que evolucione con el tiempo mejorando sus decisiones conforme aumentan su experiencia y los datos recabados, encontrando patrones en ellos. **¿Podrías buscar algunos ejemplos de uso de Big Data?** ¿Son Machine Learning o no?

En conclusión, como puedes intuir tras el trabajo en esta Unidad, es importante ser consciente de la existencia de las técnicas de tipo estadístico para evitar una posible manipulación y no ver influenciados de forma externa nuestros intereses en la dirección que pretendan otros individuos, o las empresas (especialmente de marcas comerciales). Para acabar, organizad por grupos un **debate sobre los límites legales y/o éticos** que deberían existir en torno al desarrollo de algoritmos Big Data.