

Técnicas de análisis de datos para la empresa

Ivan Soler-Ramos

Joan Torrent-Sellens

25 horas

Ivan Soler-Ramos

Licenciado en Sociología (UB). Posgrado en Minería de datos (UOC). Investigador del Grupo de investigación interdisciplinaria sobre las TIC, i2TIC y experto en análisis de datos, estadística y econometría. Diez años de experiencia en investigación de mercados en el sector privado.

Joan Torrent-Sellens

Doctor en Sociedad de la Información y el Conocimiento (IN3-UOC). Máster en Análisis de economía aplicada (UAB). Licenciado en Ciencias Económicas y Empresariales (UAB). Profesor de los Estudios de Economía y Empresa, y director de la UOC Business School. Director del Grupo de investigación interdisciplinaria sobre las TIC, i2TIC, e investigador sobre la economía internacional, española y catalana.

Índice

Introducción	5
Objetivos	7
1. Planteamiento de una investigación	8
2. Los datos en las metodologías cuantitativas	9
2.1. El origen de los datos.....	9
2.2. Datos primarios y datos secundarios.....	10
2.3. Técnicas de recogida de información.....	11
3. Las variables	12
3.1. Concepto de variable.....	12
3.2. Variables cuantitativas y variables cualitativas.....	12
3.3. Variables independientes y variables dependientes.....	13
4. Población y muestra	14
5. Análisis exploratorio de datos	15
5.1. Análisis exploratorio univariante.....	15
5.1.1. Análisis exploratorio univariante con variables cuantitativas....	15
5.1.2. Análisis exploratorio univariante con variables cualitativas.....	24
5.2. Análisis exploratorio bivariante.....	30
5.2.1. Transformación de variables cuantitativas en cualitativas.....	31
5.2.2. Análisis exploratorio bivariante con variables cualitativas.....	32
Resumen	38
Glosario	39
Bibliografía	41

Introducción: la necesidad de los datos

Desde la empresa más pequeña hasta la más grande, todas precisan datos y todas generan datos. La generación de datos y su correcta interpretación y análisis son esenciales para producir el valor de la empresa. Conocer el entorno y los mercados en los que se ubica la actividad empresarial, interpretar sus cambios y prever el futuro, y analizar el proceso interno de generación de valor de la empresa son algunos de los ejemplos en los que el análisis de datos puede ser muy útil.

De hecho, para que una empresa funcione correctamente en el mercado es esencial que efectúe una buena toma de decisiones. Esta depende de múltiples cuestiones, pero en este curso nos centraremos en una de ellas realmente importante: el papel de los datos en la toma de decisiones.

Las decisiones, para ser buenas, se deben guiar por un conocimiento lo más completo posible de aquello que se ha de decidir. Y la información, la buena información, es esencial para lograr el conocimiento que nos conduce hacia una correcta toma de decisiones. A la vez, una de las mejores maneras de obtener una buena información es realizar un análisis de datos cuidadoso.

Hasta ahora hemos utilizado tres conceptos: datos, información y conocimiento, que demasiado habitualmente se emplean de manera indistinta y apropiada.

Los datos son las unidades mínimas y, por sí mismas, son irrelevantes para la toma de decisiones. Mediante el procesamiento de estos datos les damos significado, y este significado lo denominamos información. Es decir, la información se da cuando los datos adoptan un significado.

Este significado lo logramos procesando los datos, pero no solamente mediante cálculos matemáticos o estadísticos. Los datos pueden acabar convirtiéndose en información si los resumimos, los contextualizamos o los categorizamos.

Estos datos con significado finalmente se convierten en conocimiento cuando se combinan con las otras informaciones de las que disponemos, con nuestra experiencia o con nuestro razonamiento. El proceso de conversión de datos/información/conocimiento es imprescindible para una correcta toma de decisiones en la empresa.

Con todo, muy a menudo la información necesaria para la toma de decisiones puede no estar estructurada ni grabada ni disponible. Muchos pequeños

empresarios conocen de memoria a sus clientes, saben qué quieren, cómo lo quieren, cuándo lo quieren, y también conocen a sus competidores. Saben cómo funciona su sector y cuáles son las características de sus mercados. Sin embargo, esta información valiosa no está disponible para otras personas y otros procesos, lo que puede ser perjudicial para la empresa. Una estructuración de esta información mediante datos siempre será útil para optimizar los procesos de decisión y facilitará el crecimiento de la empresa. Y, todavía más, una empresa que se dedique a recopilar los datos y no haga ningún tipo de análisis está derrochando un activo esencial para su correcto funcionamiento en el mercado. Así, por ejemplo, puede perder la oportunidad de identificar nuevos nichos de mercado, nuevas maneras de diversificar sus productos, nuevas vías de innovación o de internacionalización y así mejorar su posición relativa en el mercado.

Acabamos de constatar que el análisis de datos es primordial en el proceso de conversión de datos en información y de información en conocimiento. Este proceso es imprescindible para una correcta toma de decisiones que, al final, redundará en un mejor funcionamiento de la empresa. Precisamente, el curso que ahora empezamos pretende realizar una primera aproximación al análisis de datos en la empresa y dar al estudiante unos primeros instrumentos para realizar este análisis. No se trata de un curso de estadística ni de un curso de hojas de cálculo. Pretendemos realizar una introducción rigurosa pero práctica al análisis de datos para la empresa utilizando ejemplos sobre la información de mercado.

Objetivos

Los objetivos de este curso son los siguientes:

1. Distinguir entre datos, información y conocimiento, y captar cómo el análisis de datos es imprescindible para una correcta toma de decisiones en la empresa.
2. Entender qué es una investigación y cómo se puede plantear.
3. Interpretar y distinguir los diferentes tipos de variables, y aprender a diseñar un proceso de investigación.
4. Comprender los fundamentos prácticos del análisis de datos exploratorio univariante y bivariante.
5. Hacer una primera diagnosis sobre las implicaciones del análisis de datos en la toma de decisiones de la empresa.

1. Planteamiento de una investigación

El primer paso para obtener la información necesaria y efectuar un análisis de datos es **plantear una investigación**. Desde la más sencilla hasta la más compleja, toda investigación debe responder a las siguientes cuestiones.

Debemos tener claro qué queremos investigar. Es decir, ¿qué tipo de fenómeno queremos investigar y por qué lo queremos investigar? ¿Qué haremos con la información que obtengamos?

También hay que diseñar cómo conseguir esta información. Para ello, hemos de saber qué datos serán necesarios para poder recopilarlos y transformarlos en información y, finalmente, en conocimiento.

Por lo tanto, debemos establecer el método, la metodología y, finalmente, las herramientas para poder recopilar los datos que nos interesan.

En resumen, las principales preguntas que nos deberemos formular a la hora de plantear una investigación son:

- ¿Cuál es el objeto de la investigación?
- ¿Cuál es el objetivo?
- ¿Cuáles son los datos que necesitaremos para darle respuesta?
- ¿Dónde están estos datos?
- ¿Han sido recopilados con anterioridad?
- ¿Están disponibles?

La decisión sobre la metodología de investigación...

... está condicionada por las posibilidades que nos ofrecen los datos y por los recursos que disponemos para analizarlos.

2. Los datos en las metodologías cuantitativas

2.1. Datos primarios y datos secundarios

Básicamente, los datos pueden ser de dos tipos: los **datos primarios** y los **datos secundarios**. Los datos primarios son aquellos que son tomados en una primera instancia, es decir, con los instrumentos que nosotros mismos hemos definido. Los datos secundarios son datos tomados (y procesados) previamente por otras personas o instituciones.

La diferencia esencial entre los datos primarios y los secundarios no responde al tipo de información que se puede extraer de ellos. Si nos fijamos, todo dato secundario fue en el momento de su recopilación un dato primario, y todo dato primario será una vez recopilado y procesada un dato secundario para los otros. La diferencia entre los datos primarios y secundarios tiene lugar por medio de la ensambladura entre la información que necesitamos y la que tenemos disponible. Cuando queremos una información que no tenemos disponible, entonces hay que realizar una investigación mediante los datos primarios.

Por lo tanto, a la hora de plantear una investigación debemos responder a la siguiente pregunta:

- ¿La información que necesito la tengo disponible en alguna fuente fiable?

Si la respuesta es afirmativa, el sentido común nos llevará a utilizar el dato ya procesado, el dato secundario, puesto que usualmente el dato primario se encuentra en una fase inicial y necesita más esfuerzo para ser tomado. Con todo, a menudo sucede que la información que precisamos no está disponible o, por la razón que sea, no tenemos acceso a ella porque nadie la ha recogido antes.

Algunos breves apuntes sobre los datos secundarios

El análisis de datos secundarios es muy importante en los pasos iniciales de una investigación como modo de contextualizar y aumentar nuestro conocimiento previo a la recopilación primaria de información. El análisis de datos secundarios nos ayuda a identificar variables clave, seleccionar muestras, formular diseños de investigación o sugerir nuevos enfoques.

La principal ventaja de los datos secundarios es su accesibilidad e inmediatez, dado que ya han sido procesados y reunidos.

Con todo, también tienen algunos inconvenientes: pueden no encajar bien con el problema que se quiere estudiar, no tener la precisión deseada o estar referenciados en unidades inadecuadas o en periodos diferentes. Por lo tanto, es esencial verificarlos.

Por ejemplo...

... los datos del IPC son secundarios porque los elabora el Instituto Nacional de Estadística. En cambio, si hiciéramos una investigación sobre la dinámica de los precios en nuestra empresa, estos datos serían primarios.

Otro ejemplo...

... la información referente a las opiniones y preferencias de nuestros clientes no tiene demasiado sentido que alguien la tome por nosotros y, si lo hace, seguramente no nos la entregará así como así.

2.2. El origen de los datos

Según su origen, los datos se pueden clasificar en:

1. **Datos internos:** cuando la fuente de datos se encuentra dentro de nuestra organización. Estos tipos de datos pueden ser muy interesantes para realizar análisis que optimicen el funcionamiento de nuestra empresa. Algunas de las ventajas que presentan este tipo de datos son su fácil accesibilidad y el bajo coste que representan.
2. **Datos externos:** los datos externos son datos que podemos obtener de otra fuente. Aquí se incluyen todos los datos primarios que reuniremos para resolver nuestras necesidades de información. No obstante, conviene realizar un breve apunte sobre los datos externos secundarios. Existen varios tipos de fuentes de datos externos. Si las dividimos en función de su accesibilidad, obtenemos:
 - 2.1. **Fuentes abiertas:** son datos externos que están disponibles de manera abierta, en bibliotecas o en Internet. Estos datos responden a publicaciones generadas por las administraciones públicas, libros o los que se pueden obtener en cualquier Instituto de Estadística.
 - 2.2. **Fuentes semiabiertas:** son datos que generan asociaciones empresariales, gremios o federaciones de empresas y que, generalmente, se ofrecen a sus integrantes o asociados. En función de si somos integrantes o no de la organización, los datos serán una fuente abierta (para nosotros) o cerrada (para los otros).
 - 2.3. **Fuentes cerradas:** las fuentes cerradas son habitualmente datos estandarizados que una empresa ha ido reuniendo y que procesa y, finalmente, vende a quien quiera disponer de esta información. Habitualmente la adquisición de estos datos es muy costosa. Algunos de los datos con fuente cerrada más habituales hacen referencia a información de mercado relativa a datos de consumidores, al estado de diferentes sectores de actividad, a datos de comunicación y audiencia, y a otros elementos de valor de la empresa.
 - 2.4. Las **fuentes sindicadas:** son aquellas que son de pago, bien de pago por cada uso, bien de suscripción. Están procesadas y recopiladas por empresas que por iniciativa propia realizan estudios de mercado y publican los informes correspondientes. Esta información es de pago y cualquier persona puede acceder a ella. Por ejemplo, los estudios que lleva a cabo AC Nielsen son algunos de los más conocidos.

Algunas posibles fuentes de datos internos...

... son aquellas que se generan con la actividad comercial (distribución de ventas por productos, zonas o clientes), con la actividad productiva (costes, existencias) o aquellas resultantes de la información contable y financiera, por nombrar algunas de las más importantes.

Visita las siguientes webs:

www.idescat.cat

www.ine.es

<http://epp.eurostat.ec.europa.eu/>

Visita la web:

<http://es.nielsen.com>

2.3. Técnicas de recogida de información

Las técnicas de recogida de información dependen de los objetivos que queremos lograr con la investigación; de la realidad que queremos estudiar; del tipo de investigación que en consecuencia queremos emprender, y del tipo de análisis de datos que realizaremos una vez obtenida la información.

En función de si son datos primarios o secundarios, algunas de las técnicas de recopilación de información más frecuentes son:

1. Datos primarios:

1.1. La **recogida de información por observación** es aquel procedimiento mediante el cual se obtiene la información por observación directa de los fenómenos de interés. Se trata de un procedimiento en el que la información se graba por las anotaciones que realiza el propio investigador. La recopilación de datos por observación es una técnica de recogida de información propiamente dicha si se orienta a un objetivo de investigación, está planificada y busca obtener respuestas a proposiciones generales.

1.2. La **encuesta** es la técnica de recopilación de información más habitual en la metodología cuantitativa. La encuesta utiliza procedimientos estandarizados de interrogación con el fin de conseguir medidas cuantitativas sobre las características subjetivas y/o objetivas de una población.

2. Datos secundarios:

2.1. **Estadísticas**. Corresponden a la información proporcionada por los Institutos de Estadística Oficial.

3. Las variables

3.1. Concepto de variable

Llegados a este punto, a continuación daremos un breve repaso de los conceptos más básicos relacionados con **la estadística descriptiva**.

La estadística descriptiva o análisis exploratorio de datos es la forma más básica de la estadística y tiene un notable contenido instrumental.

No es objeto de este curso convertirse en un manual de estadística, pero en cuanto que herramienta que necesitaremos para analizar datos primarios sí que es de interés que hagamos un breve repaso muy orientado a la práctica de los principales conceptos que necesitaremos trabajar.

Lo primero que debemos tener en cuenta es que para obtener la información que deseamos, hemos de hacer operativos los conceptos. Es decir, hemos de ser capaces de transformar los conceptos, las preguntas esenciales que nos formulamos, en forma de variables. Es decir, debemos hacer operativas las dimensiones de la realidad para que se puedan trabajar de manera empírica.

Así pues, hay que definir el concepto de variable.

Una **variable** no es más que una de las diferentes dimensiones de la realidad que queremos conocer. Hablaremos de **valor de la variable** cuando nos refiramos a la medida que la variable presenta. Hablaremos de **variación de la variable** cuando nos refiramos a la diferencia que existe entre los valores de una variable.

Trataremos la estadística de modo muy intuitivo...

... sin profundizar en las matemáticas que hay detrás, aplicando de una manera práctica conceptos básicos de análisis.

3.2. Variables cuantitativas y variables cualitativas

Desde el contenido de una variable, es decir, desde el valor que representa, y de un modo simple, es posible agrupar dos grandes tipos de variables:

1. **Variable cuantitativa:** aquella en la que el valor representa una cifra.
2. **Variable cualitativa:** aquella en la que el valor representa una categoría.

A su vez, las variables cuantitativas se pueden dividir en:

1. **Variable cuantitativa continua:** cuando el valor no representa ninguna interrupción más allá de las que el instrumento de medición nos plantea.
2. **Variable cuantitativa discreta:** aquella que presenta una interrupción entre sus valores porque los intermedios no existen.

Las variables cualitativas se pueden dividir en:

1. **Variable cualitativa ordinal:** es aquella en la que los valores presentan un orden establecido dentro de una escala. Por ejemplo, la típica valoración de opinión de 1 a 10. En la práctica, y por el tema que nos ocupa, no existe diferencia entre las variables cualitativas ordinales y las variables cuantitativas discretas.
2. **Variable cualitativa nominal:** en este caso, los valores de las variables representan categorías que no tienen relación de medida entre ellas.

Un caso específico y muy habitual de variable cualitativa nominal es la llamada **variable dicotómica**. Esta variable solamente toma dos valores. Habitualmente 1 o 2, o 0 o 1. Se utiliza para señalar la presencia o ausencia de una característica, o la afirmación o negación a una pregunta de opinión de un cuestionario.

3.3. Variables independientes y variables dependientes

Finalmente, hay una última diferencia que cabe abordar cuando tratamos variables. Se trata de la función que pueden representar en nuestro análisis. Según esta aproximación, las variables pueden ser dependientes e independientes.

1. Las **variables dependientes** son aquellas variables que suponemos que varían en función de otra variable.
2. Las **variables independientes** son aquellas que suponemos que no evolucionan hacia ninguna otra variable, a pesar de que su variación sí que incide en la evolución de otras variables dependientes.

Por ejemplo, imaginemos que medimos la altura de los niños de un colegio y queremos averiguar cómo varían estas alturas. Para hacerlo, disponemos de dos variables: la edad del niño y su altura. La variable independiente sería la edad y la variable dependiente sería la altura. Es decir, la secuencia lógica del pensamiento nos lleva a interpretar directamente que la altura del niño varía con su edad, no al contrario.

Por ejemplo, la altura...

... si medimos en centímetros la altura de un individuo, evidentemente esta presenta interrupciones porque no es infinita.

Como ejemplo...

... el número de hijos de una persona.

Por ejemplo, los colores de un coche...

... si asignamos 1 al color verde y 2 al color rojo, evidentemente no podemos interpretar que el color rojo es el doble del color verde.

Hay que añadir, sin embargo, que hay muchos casos en los que esta relación de dependencia no está nada clara y hay que realizar análisis más sofisticados para averiguar su relación. Estos tipos de análisis quedan fuera del alcance de este curso.

4. Población y muestra

Una vez tenemos decididos los datos, y su tipología, que necesitaremos, a continuación hemos de ver dónde los debemos encontrar. Existen varios tipos de lugares donde encontrarlos, pero en este curso nos centraremos en los datos obtenidos según la declaración de un entrevistado. Para ello, utilizaremos **encuestas a muestras representativas y fiables de una población**.

Se entiende por **población** el conjunto de individuos que disponen de las características que precisamos observar.

Se entiende por **muestra** un subconjunto de esta población que representa a escala las características principales de la población.

Cuando el tamaño de la población es grande y no podemos llegar a todo el mundo, entonces debemos definir una muestra. Para esta definición de la muestra hay diferentes técnicas. Con todo, nosotros nos centraremos en dos condiciones: la representatividad y la fiabilidad de la muestra.

Una **muestra es representativa** cuando representa las características de la población objeto de estudio. Es decir, cuando obtenemos un conjunto de individuos inferior a la población y que la representan. Además, con el diseño de la muestra hemos de suponer que cometemos un error. Este error lo denominamos **error muestral**.

Una **muestra es fiable** cuando el error muestral que supone trabajar con una muestra es asumible. El hecho de que una muestra sea o no fiable depende esencialmente del criterio del investigador y del tipo de investigación.

Metodológicamente, para poder trabajar con una muestra existen varias técnicas. A continuación se citan dos.

- **La muestra aleatoria o probabilística simple (MAS)**. En esta técnica todos los individuos que conforman la población tienen las mismas probabilidades de ser elegidos como unidades muestrales.
- **La muestra estratificada**. En esta técnica la población se segmenta en diferentes estratos o clases y en cada uno de ellos se elige una muestra aleatoria. Los elementos de cada estrato deben ser más similares entre sí que respecto a la población. Por lo tanto, se trata de buscar la máxima homogeneidad en cada estrato y la máxima heterogeneidad entre los diferentes estratos. A cada estrato le corresponde una **submuestra**, y el conjunto de las submuestras constituye el total de la muestra poblacional.

Por ejemplo...

... podemos obtener datos midiendo alguna de las características de un producto o servicio que producimos. Pero, para conocer la opinión que tienen nuestros clientes de nuestro producto o servicio, se lo debemos preguntar. Y esto se debe hacer de manera estructurada.

5. Análisis exploratorio de datos

Una vez están identificados y recopilados los datos que necesitamos para nuestra investigación, ya estaremos en disposición de iniciar el análisis de datos. Para el análisis de datos emplearemos la estadística descriptiva, también denominada análisis exploratorio de datos.

El análisis exploratorio de datos es un conjunto de técnicas que tienen como finalidad describir, presentar y, finalmente, analizar los datos que queremos observar en una población.

Para lograr los objetivos...

... del análisis exploratorio de datos utilizaremos las tablas, los gráficos y resúmenes de estadísticas.

El análisis exploratorio de datos se denomina **univariante** cuando mide una de las características de los individuos de la población. Y se denomina **bivariante** cuando mide una de las características de la población en función de otra característica.

Para trabajar con los datos, estos se organizan siguiendo una estructura particular. Habitualmente, la empleada para las hojas de cálculo y los programas de análisis estadístico. En general, cada unidad muestral (registro/individuo) ocupa una fila diferente. Y las diferentes variables, es decir, las dimensiones de la realidad que queremos observar, se presentan en columnas. Cada variable ocupa una columna diferente.

Además, en el análisis de datos por medio de encuesta debemos tener en cuenta que:

- La cantidad y la calidad de las respuestas que queremos obtener no está necesariamente relacionada con la dimensión de la muestra.
- Los datos serán más útiles a medida que los podamos comparar con otros similares o con otras poblaciones de referencia.

5.1. Análisis exploratorio univariante

La estadística exploratoria univariante mide una característica del individuo que puede presentar varias modalidades. Según la información que contenga el tipo de variable que queremos analizar, el análisis descriptivo que llevaremos a cabo será de un tipo u otro. A lo largo de este curso trataremos los análisis más habituales en estadística exploratoria.

5.1.1. Análisis exploratorio univariante con variables cuantitativas

Antes de adentrarnos en el análisis exploratorio univariante, primero revisaremos brevemente dos tipos de medidas relevantes para las variables cuantitativas. Se trata de las medidas de la posición central y las de dispersión.

Las **medidas de posición central** son aquellas que hacen referencia a la posición del parámetro dentro de la distribución. En nuestro caso, abordaremos la **media aritmética**, la **mediana** y la **moda**.

La **media aritmética** es el valor resultante de dividir el sumatorio de un conjunto de datos por el número total de datos. La media aritmética es una medida muy utilizada en el análisis de datos. Con todo, puede ser muy sensible a valores extremos. Es decir, si uno de los valores del conjunto de datos que tratamos es mucho mayor que el resto, la media aritmética no será una medida demasiado precisa para extraer conclusiones. Para solucionarlo, la media **truncada** (que calcula la media entre el 95% de los casos centrales) o la desviación **estándar** (o típica) son medidas que nos ayudarán a no equivocarnos en el momento de extraer conclusiones del análisis efectuado.

A continuación, y con el objetivo práctico de explicar cómo se va construyendo este conjunto de estadísticas, haremos un análisis exploratorio utilizando la **hoja de cálculo Microsoft Excel**. En Excel, la función **PROMEDIO** [PROMEDIO (número 1, número 2, etc.)] permite calcular la media aritmética de un conjunto de datos. Basándonos en unos datos hipotéticos, la tabla 1 muestra el procedimiento de cálculo de la media aritmética.

Tabla 1. Procedimiento de cálculo de la media aritmética en Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	A6	A1	A2	A3	A4	A5					
2	46.847	1	2	0	0	4	1					
3	46.626	1	9	0	0	2	1					
4	46.715	1	3	0	0	3	1					
5	46.757	1	7	0	0	3	1					
6	46.665	2	4	0	0	5	2					
7	46.636	2	0	0	0	4	1					
8	47.259	1	1	0	0	5	1					
9	46.754	1	10	12	2	2	1					
10	46.507	2	1	0	0	4	2					
11	46.861	2	8	0	0	3	1			PROMEDIO	=PROMEDIO(C2:C37)	
12	46.849	1	13	0	2	2	1					
13	46.788	1	12	0	0	3	1					
14	46.929	2	1	0	0	4	2					
15	46.808	2	1	0	0	5	2					
16	46.959	1	8	0	0	5	2					
17	46.655	2	5	0	0	4	1					
18	46.645	1	10	1	0	2	1					
19	46.536	2	0	0	0	3	2					
20	46.532	1	10	0	0	3	1					
21	46.714	2	1	0	0	4	1					
22	46.686	1	5	1	0	4	1					
23	46.747	2	8	0	0	4	1					
24	46.726	2	3	0	0	5	1					
25	46.967	1	9	0	0	3	1					
26	46.914	2	2	0	0	5	2					
27	46.958	1	20	0	0	4	1					
28	46.698	2	3	0	0	4	2					
29	46.743	1	3	0	0	5	1					
30	46.608	1	4	0	0	4	1					
31	46.821	2	2	0	0	4	2					
32	46.699	1	3	0	0	4	1					
33	46.738	2	1	0	0	5	2					
34	46.518	2	0	0	0	4	2					
35	46.974	1	4	0	0	3	1					
36	46.658	1	1	0	0	4	1					
37	46.748	1	10	0	0	4	1					
38												

Una vez efectuada la operación, el resultado se hace visible en la celda correspondiente, donde se ha realizado la operación estadística (tabla 2):

La **mediana** es el valor que divide toda una serie de datos en dos partes exactamente iguales. Es decir, donde la cantidad de datos inferiores a la mediana es igual a la cantidad de datos superiores a esta. En Excel, la función MEDIANA [MEDIANA (número 1, número 2, etc.)] permite calcular la mediana de un conjunto de datos (tabla 4).

Tabla 4. Procedimiento de cálculo de la mediana en Excel

1	A	B	C	D	E	F	G	H	I	J	K	L
2	46.847	14	2	0	0	4	1	1				
3	46.626	14	9	0	0	2	1	1				
4	46.715	14	3	0	0	3	1	1				
5	46.757	14	7	0	0	3	1	1				
6	46.665	14	4	0	0	5	99	2				
7	46.636	14	0	0	0	4	1	2				
8	47.259	14	1	0	0	5	1	1				
9	46.754	14	10	12	2	2	1	1			PROMEDIO	5,11
10	46.507	14	1	0	0	4	2	2			MODA	1
11	46.861	14	8	0	0	3	1	2			MEDIANA	=Mediana(C2:C37)
12	46.849	14	13	0	2	2	1	1				
13	46.788	14	12	0	0	3	1	1				
14	46.929	14	1	0	0	4	2	2				
15	46.808	14	1	0	0	5	2	2				
16	46.959	14	8	0	0	5	99	1				
17	46.655	14	5	0	0	4	1	2				
18	46.645	14	10	1	0	2	1	1				
19	46.536	14	0	0	0	3	2	2				
20	46.532	14	10	0	0	3	1	1				
21	46.714	14	1	0	0	4	1	2				
22	46.886	14	5	1	0	4	1	99				
23	46.747	14	8	0	0	4	1	2				
24	46.726	14	3	0	0	5	1	2				
25	46.967	14	9	0	0	3	1	1				
26	46.914	14	2	0	0	5	2	2				
27	46.958	14	20	0	0	4	1	1				
28	46.698	14	3	0	0	4	2	2				
29	46.743	14	3	0	0	5	1	1				
30	46.608	14	4	0	0	4	1	1				
31	46.821	14	2	0	0	4	2	2				
32	46.699	14	3	0	0	4	1	1				
33	46.738	14	1	0	0	5	2	2				
34	46.518	14	0	0	0	4	2	2				
35	46.974	14	4	0	0	3	1	1				
36	46.658	14	1	0	0	4	1	1				
37	46.748	14	10	0	0	4	1	1				
38												

Las medidas de dispersión son profusamente utilizadas en el análisis de datos y, a menudo, sirven para completar la información que nos aportan las medidas de tendencia central.

Las **medidas de dispersión** nos indican la distancia promedio de los datos respecto a la media aritmética. En este curso trataremos el **rango**, la **desviación mediana**, la **varianza** y la **desviación estándar o típica**.

El **rango** es la medida de desviación que tiene un cálculo más sencillo. Simplemente, se trata de la diferencia entre el mayor y el menor valor de la distribución. Se suele denominar con la letra R.

La **desviación mediana** es la media de las diferencias entre los valores de la variable y la media aritmética en valor absoluto. En Excel, la función DESVPROM [DESVPROM (número 1, número 2, etc.)] permite calcular la desviación mediana de un conjunto de datos (tabla 5).

Tabla 5. Procedimiento de cálculo de la desviación mediana en Excel

C	D	E	F	G	H	I	J	K	L
A1	A2	A3	A4	A5	A6				
2	0	0	4	1	1				
9	0	0	2	1	1				
3	0	0	3	1	1				
7	0	0	3	1	1				
4	0	0	5	99	2				
0	0	0	4	1	2				
1	0	0	5	1	1				
10	12	2	2	1	1				
1	0	0	4	2	2		PROMEDIO	5,11	
8	0	0	3	1	2		MODA	1	
13	0	2	2	1	1		MEDIANA	4	
12	0	0	3	1	1		DESVIACIÓN MEDIA	=DESVPROM(C2:C37)	
1	0	0	4	2	2				
1	0	0	5	2	2				
8	0	0	5	99	1				
5	0	0	4	1	2				
10	1	0	2	1	1				
0	0	0	3	2	2				
10	0	0	3	1	1				
1	0	0	4	1	2				
5	1	0	4	1	99				
8	0	0	4	1	2				
3	0	0	5	1	2				
9	0	0	3	1	1				
2	0	0	5	2	2				
20	0	0	4	1	1				
3	0	0	4	2	2				
3	0	0	5	1	1				
4	0	0	4	1	1				
2	0	0	4	2	2				
3	0	0	4	1	1				
1	0	0	5	2	2				
0	0	0	4	2	2				
4	0	0	3	1	1				
1	0	0	4	1	1				
10	0	0	4	1	1				

La **varianza** (S^2). Para asegurar que las diferencias entre la media y los puntos de un valor sean positivos, es posible elevarlas al cuadrado. La varianza es el resultado de la división de la sumatoria de la distancia entre cada dato y la media aritmética del conjunto de datos elevados al cuadrado. En Excel, la función VAR [VAR (número 1, número 2, etc.)] permite calcular la varianza de un conjunto de datos (tabla 6).

Tabla 6. Procedimiento de cálculo de la varianza en Excel

C	D	E	F	G	H	I	J	K	L
A1	A2	A3	A4	A5	A6				
2	0	0	4	1	1				
9	0	0	2	1	1				
3	0	0	3	1	1				
7	0	0	3	1	1				
4	0	0	5	99	2				
0	0	0	4	1	2				
1	0	0	5	1	1				
10	12	2	2	1	1				
1	0	0	4	2	2		PROMEDIO	5,11	
8	0	0	3	1	2		MODA	1	
13	0	2	2	1	1		MEDIANA	4	
12	0	0	3	1	1		DESVIACIÓN MEDIA	3,25	
1	0	0	4	2	2		VARIANZA	=VAR(C2:C37)	
1	0	0	5	2	2				
8	0	0	5	99	1				
5	0	0	4	1	2				
10	1	0	2	1	1				
0	0	0	3	2	2				
10	0	0	3	1	1				
1	0	0	4	1	2				
5	1	0	4	1	99				
8	0	0	4	1	2				
3	0	0	5	1	2				
9	0	0	3	1	1				
2	0	0	5	2	2				
20	0	0	4	1	1				
3	0	0	4	2	2				
3	0	0	5	1	1				
4	0	0	4	1	1				
2	0	0	4	2	2				
3	0	0	4	1	1				
1	0	0	5	2	2				
0	0	0	4	2	2				
4	0	0	3	1	1				
1	0	0	4	1	1				
10	0	0	4	1	1				

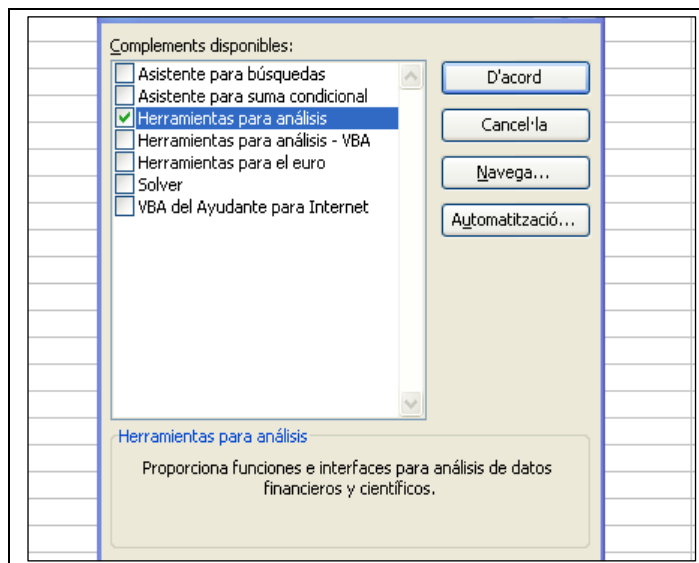
La varianza tiene el problema de que eleva también al cuadrado las unidades de los datos originales. Es decir, si estamos midiendo, por ejemplo, segundos o minutos, obtenemos la varianza en segundos o minutos. Para evitarlo, a menudo se utiliza la **desviación estándar o típica**. La desviación estándar es la raíz cuadrada de la varianza. En Excel, la función DEVEST [DEVEST (número 1, número 2, etc.)] permite calcular la desviación estándar de un conjunto de datos (tabla 7).

Tabla 7. Procedimiento de cálculo de la desviación estándar en Excel

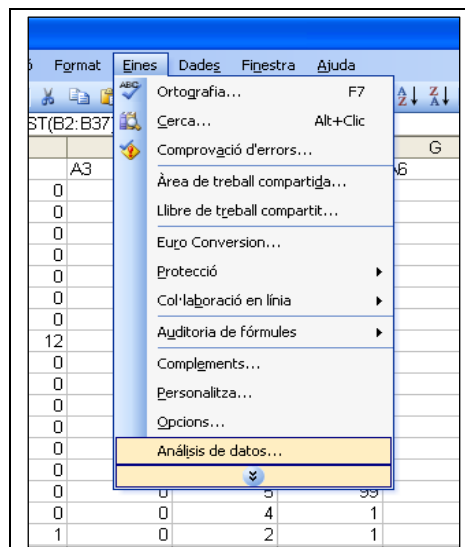
C	D	E	F	G	H	I	J	K
0	0	4	1	1				
0	0	2	1	1				
0	0	3	1	1				
0	0	3	1	1				
0	0	5	99	2				
0	0	4	1	2				
0	0	5	1	1				
12	2	2	1	1	PROMEDIO		5,11	
0	0	4	2	2	MODA		1	
0	0	3	1	2	MEDIANA		4	
0	2	2	1	1	DESVIACIÓN MEDIA		3,75	
0	0	3	1	1	VARIANZA		21	
0	0	4	2	2	DESVIACIÓN ESTANDAR		=DEVEST(B2:B37)	
0	0	5	2	2				
0	0	5	99	1				
0	0	4	1	2				
1	0	2	1	1				
0	0	3	2	2				
0	0	3	1	1				
0	0	4	1	2				
1	0	4	1	99				
0	0	4	1	2				
0	0	5	1	2				
0	0	3	1	1				
0	0	5	2	2				
0	0	4	1	1				
0	0	4	2	2				
0	0	5	1	1				
0	0	4	1	1				
0	0	4	2	2				
0	0	4	1	1				
0	0	5	2	2				
0	0	4	2	2				
0	0	3	1	1				
0	0	4	1	1				
0	0	4	1	1				

Esta información y alguna más, sin embargo, la podemos obtener de una manera más fácil y directa que si utilizamos la herramienta [ANÁLISIS DE DATOS] que nos ofrece Microsoft Excel. Con esta herramienta podemos obtener, simultáneamente, todas estas estadísticas descriptivas que hemos visto hasta ahora.

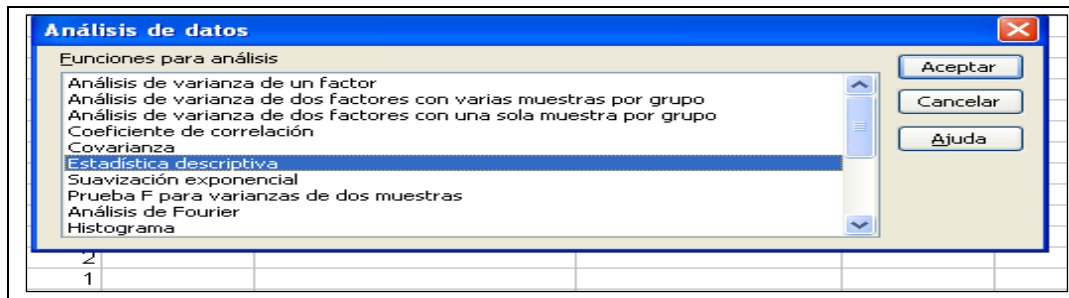
Para poder utilizar esta herramienta, lo primero que hemos de hacer es verificar si ha sido habilitada en el ordenador. En primer lugar, debemos entrar en el menú [HERRAMIENTAS] y observar si nos aparece la opción [ANÁLISIS DE DATOS]. Si esta no está presente, la hemos de activar. Para hacerlo, hay que seleccionar la opción [COMPLEMENTOS] dentro del menú [HERRAMIENTAS] y seleccionar [HERRAMIENTAS PARA ANÁLISIS].



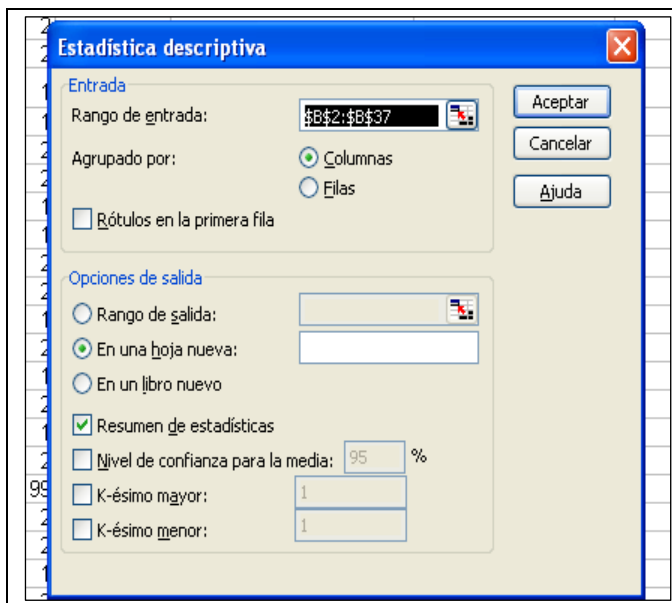
Posteriormente, debemos hacer clic en [DE ACUERDO] y verificar que en el menú [HERRAMIENTAS] esta opción esté presente.



Una vez esta herramienta esté habilitada, haremos un clic encima y se desplegará un menú en el que seleccionaremos la opción [ESTADÍSTICA DESCRIPTIVA].



Después nos aparecerá la siguiente pantalla:



En esta pantalla nos aparecen toda una serie de opciones que debemos tener en cuenta. Hay opciones de entrada, es decir, las opciones que se refieren a cómo los datos entran en el ordenador, y opciones de salida, es decir, de qué manera nos devolverá el ordenador los datos ya procesados.

En cuanto a las opciones de entrada, tenemos:

- [RANGO DE ENTRADA]. Hemos de indicar qué datos queremos analizar. Los podremos seleccionar con el ratón.
- [AGRUPADO POR]. Aquí deberemos decir si los datos están colocados en columnas o en filas. En este caso la variable A1 está situada en una columna.

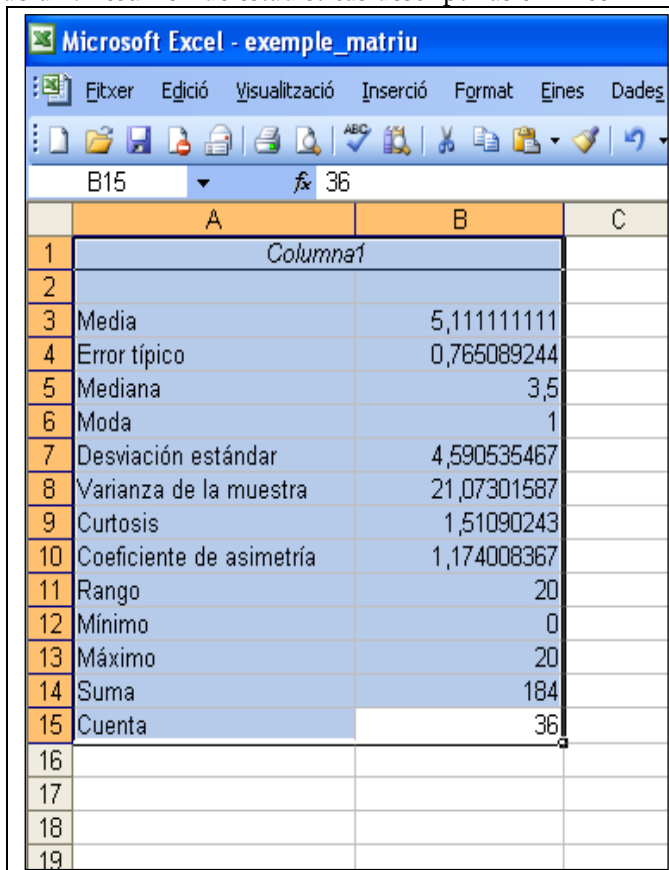
En cuanto a las opciones de salida, tenemos:

- [RANGO DE SALIDA], [EN UNA HOJA NUEVA], [EN UN LIBRO NUEVO] son opciones que se refieren al lugar donde queremos que Excel sitúe los datos una vez estos estén procesados. En este caso, le pedimos que los abra en una nueva hoja.

- [RESUMEN DE ESTADÍSTICAS]. Esta opción la deberemos marcar, puesto que nos dará un resumen de las estadísticas que hemos visto hasta ahora (y algunas otras).

Una vez seleccionadas las opciones, haremos clic en [DE ACUERDO] y nos aparecerá en una nueva hoja el siguiente resumen:

Tabla 7. Resumen de estadísticas descriptivas en Excel



	A	B	C
1	Columna1		
2			
3	Media	5,111111111	
4	Error típico	0,765089244	
5	Mediana	3,5	
6	Moda	1	
7	Desviación estándar	4,590535467	
8	Varianza de la muestra	21,07301587	
9	Curtosis	1,51090243	
10	Coeficiente de asimetría	1,174008367	
11	Rango	20	
12	Mínimo	0	
13	Máximo	20	
14	Suma	184	
15	Cuenta	36	
16			
17			
18			
19			

Así pues, hemos visto dos maneras de aproximarnos al análisis cuantitativo exploratorio univariante utilizando Excel. Usar la herramienta [ANÁLISIS DE DATOS] es más completo, rápido y fácil que la utilización de cada una de las funciones una a una.

Con todo, hay que tener en cuenta que habrá ocasiones en las que preferiremos utilizar una única función. Debemos tener en cuenta que si realizamos alguna modificación en los datos de entrada o en la salida procesada, si utilizamos la herramienta [ANÁLISIS DE DATOS], estos cambios no se verán reflejados y, por lo tanto, deberemos repetir el análisis. Esto no sucede si utilizamos la función individual en una celda de la hoja de datos.

5.1.2. Análisis exploratorio univariante con variables cualitativas

El análisis exploratorio más habitual con variables cualitativas es la tabla de frecuencias.

Esta tabla nos puede presentar las siguientes estadísticas:

Frecuencia absoluta. Es el número de individuos que comparten la misma modalidad de la variable.

Frecuencia relativa. Es la proporción de individuos que comparte la misma modalidad de la variable.

Frecuencia absoluta acumulada. Es el número de individuos que comparten una modalidad o alguna menor.

Frecuencia relativa acumulada. Es la proporción de individuos que comparten una modalidad o alguna menor.

Las frecuencias acumuladas tienen sentido en el supuesto de que las modalidades de las variables sean ordenables. Por ejemplo, si tomamos una variable cuantitativa discreta, como es el número de hijos, puede tener sentido observar sus frecuencias acumuladas.

Es decir, si:

- El 20% de una población no tiene ningún hijo.
- El 35% tiene un hijo.
- El 20% tiene dos hijos.

Podría ser una información relevante conocer que $\frac{3}{4}$ partes ($20\% + 35\% + 20\% = 75\%$) de la población estudiada tiene dos hijos o menos.

En cambio, si la característica que se analizara fuera el nombre del hijo mayor, ninguna de las frecuencias acumuladas tendría sentido, puesto que no es posible determinar qué nombre es mayor o menor a otro.

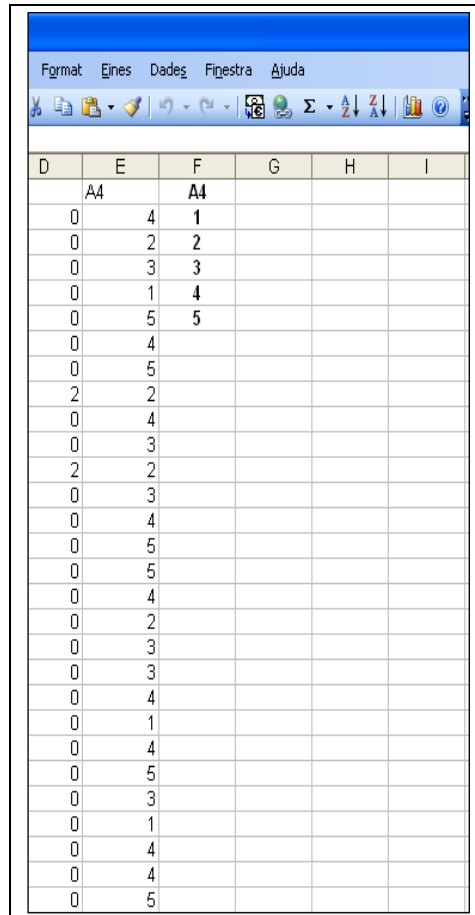
Es decir, si:

- El 20% de los niños se llaman Marta.
- El 35% se llama Jordi.
- El 20% se llama Joan.

Podríamos decir que el 75% de los niños de la población objeto de estudio se llaman Marta o Jordi o Joan, pero no habría a priori ninguna relación entre las modalidades de las variables.

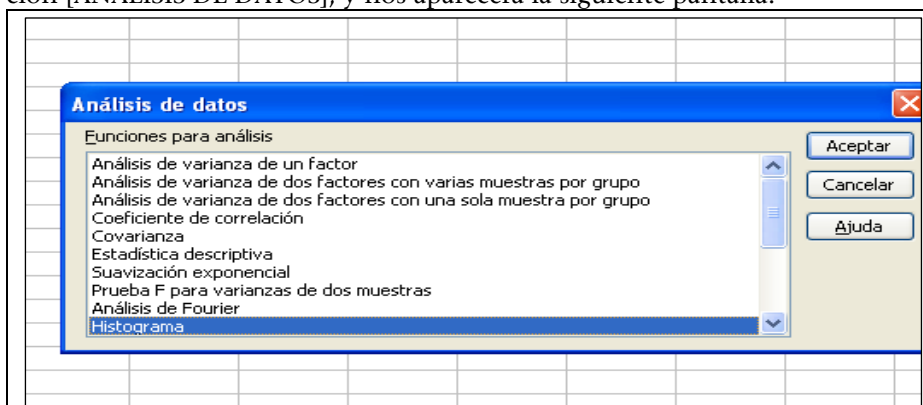
Veamos, a continuación, cómo se construye una tabla de frecuencias en Microsoft Excel.

Lo primero que haremos es escribir en una columna el nombre de la variable y, sucesivamente, las diferentes modalidades que aparecen en la variable. Como se ve en la imagen de ejemplo, la variable A4 en la columna E presenta cinco modalidades diferentes (1, 2, 3, 4 y 5).



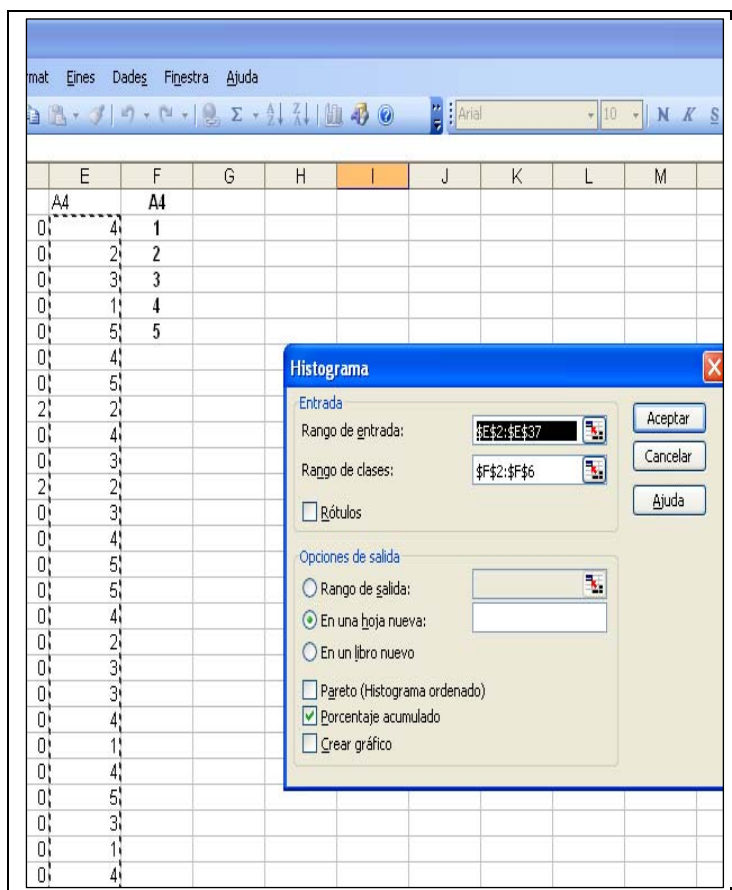
The image shows a spreadsheet window with a menu bar (Format, Eines, Dades, Finestra, Ajuda) and a toolbar. The spreadsheet has columns labeled D, E, F, G, H, and I. Column E is labeled 'A4' and contains the following values: 4, 2, 3, 1, 5, 4, 5, 2, 4, 3, 2, 3, 4, 5, 5, 4, 2, 3, 3, 4, 5, 5, 4, 2, 3, 3, 4, 1, 4, 5, 3, 1, 4, 4, 5. Column F is labeled 'A4' and contains the following values: 1, 2, 3, 4, 5. All other cells in the visible range are empty.

Ahora volveremos a seleccionar dentro del menú [HERRAMIENTAS] la opción [ANÁLISIS DE DATOS], y nos aparecerá la siguiente pantalla:



En esta pantalla deberemos seleccionar la opción [HISTOGRAMA] y [ACEPTAR].

Una vez dentro de la opción [HISTOGRAMA], nos aparecerá la siguiente pantalla:



En esta pantalla encontramos unas opciones muy parecidas a las que hemos visto con anterioridad. Así, vuelven a aparecer opciones de entrada, es decir, las opciones que se refieren a cómo los datos entran en el ordenador, y opciones de salida, es decir, de qué manera nos devolverá el ordenador los datos una vez procesados.

Opciones de entrada:

- [RANGO DE ENTRADA]. Debemos indicar qué datos queremos analizar. Los podremos seleccionar con el ratón.
- [RANGO DE CLASES]. Aquí hemos de indicar las diferentes modalidades que toma la variable A4. En este caso son aquellas situadas en la columna F (\$F\$2:\$F\$6).

Opciones de salida:

- [RANGO DE SALIDA], [EN UNA HOJA NUEVA], [EN UN LIBRO NUEVO]. Son opciones que se refieren al lugar en el que queremos que Excel sitúe los datos una vez están procesados. Nuevamente, le pediremos que los abra en una nueva hoja. En este caso marcaremos la opción [PORCENTAJE ACUMULADO].

Una vez aceptamos, nos aparecerá en una nueva hoja la siguiente tabla (tabla 8):

Tabla 8. Histograma en Excel

	A	B	C	D
1	Clase	Frecuencia	% acumulado	
2	1	3	8,33%	
3	2	4	19,44%	
4	3	7	38,89%	
5	4	15	80,56%	
6	5	7	100,00%	
7	y mayor...	0	100,00%	
8				
9				

Como Excel no nos ofrece la frecuencia absoluta acumulada ni la frecuencia relativa, las deberemos construir manualmente para completar la tabla de frecuencias. Así, lo primero que hay que hacer es abrir dos nuevas columnas mediante la opción [COLUMNAS] dentro del menú [INSERCIÓN] asegurando que estas se inserten entre la [FRECUENCIA] y el [% ACUMULADO]. En la columna C escribiremos en la primera fila [FRECUENCIA ACUMULADA] y en la columna D [%].

	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3			8,33%	
3	2	4			19,44%	
4	3	7			38,89%	
5	4	15			80,56%	
6	5	7			100,00%	
7	y mayor...	0			100,00%	
8						
9						

En la celda C2, copiaremos la frecuencia absoluta (anotaremos [=B2]) y en la celda C3, sumaremos la frecuencia absoluta de la segunda modalidad (anotaremos [=C2+B3]).

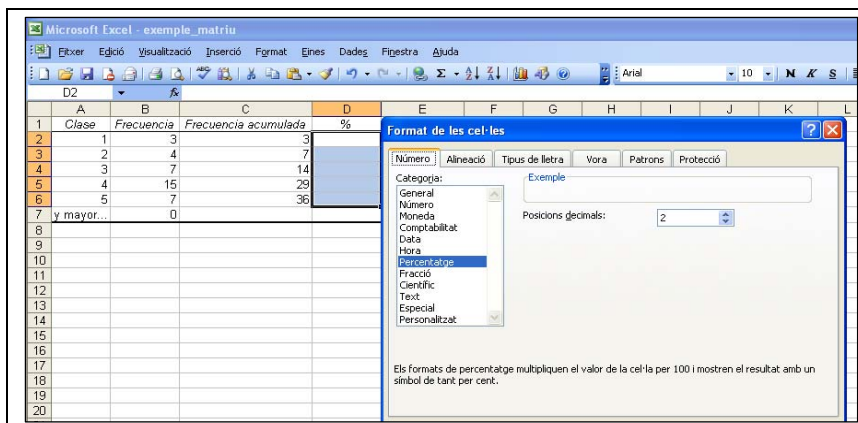
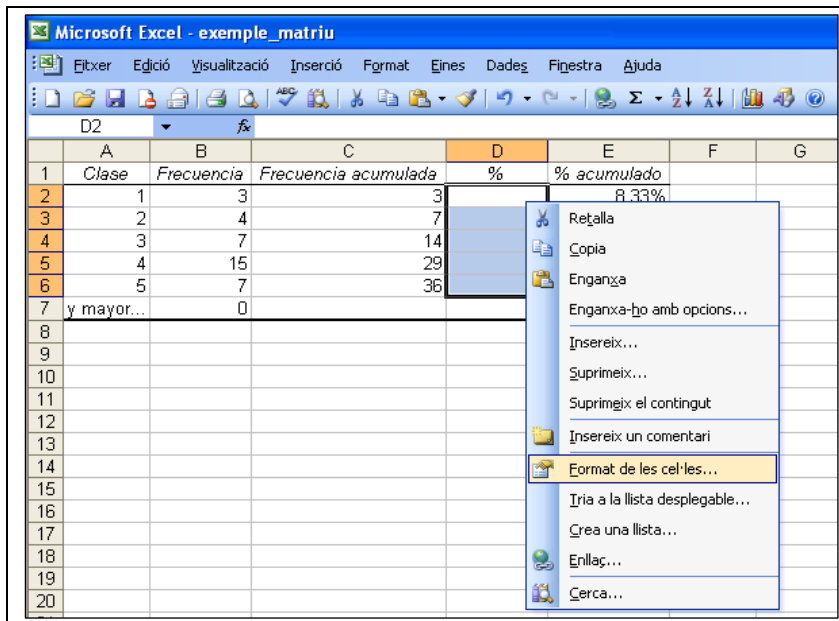
	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	=B2		8,33%	
3	2	4			19,44%	
4	3	7			38,89%	
5	4	15			80,56%	
6	5	7			100,00%	
7	y mayor...	0			100,00%	
8						
9						

	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	3		8,33%	
3	2	4	=C2+B3		19,44%	
4	3	7			38,89%	
5	4	15			80,56%	
6	5	7			100,00%	
7	y mayor...	0			100,00%	
8						
9						

Una vez efectuado este paso, tan solo habrá que arrastrar con el ratón la marca situada en la parte inferior a la derecha de la celda y ya tendremos las frecuencias acumuladas.

	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	3		8,33%	
3	2	4	7		19,44%	
4	3	7	14		38,89%	
5	4	15	29		80,56%	
6	5	7	36		100,00%	
7	y mayor...	0			100,00%	
8						
9						

Llegados a este punto, procederemos a construir la frecuencia relativa. Esta frecuencia se expresa en porcentaje. Por lo tanto, lo primero que haremos será definir el formato de las celdas como porcentaje. Esto lo podremos conseguir seleccionando las celdas que ocuparán estas frecuencias y abriendo el menú contextual haciendo clic con el botón derecho del ratón. Aquí seleccionaremos la opción [FORMATO DE LAS CELDAS] y, una vez en la pantalla siguiente, dentro de la pestaña [NÚMERO], la opción [PORCENTAJE]. Dejaremos las dos posiciones decimales que por defecto ofrece Excel.



Finalmente, procederemos a completar la columna (D) de la frecuencia relativa calculando la proporción de B2 respecto al total de los casos. Esto lo haremos mediante el cálculo $[=B2/SUMA(b\$2:b\$6)]$, en el que en B2 encontramos la frecuencia absoluta de la clase (o modalidad) 1 de la variable A4, que dividimos por la suma (utilizando la función [SUMA] de Excel) de las frecuencias de todas las modalidades de la variable. Es decir, el total de los casos. Hay que resaltar que añadiremos el signo [\$] después de las indicaciones de la columna B para fijar la operación [SUMA].

	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	3	=B2/ suma(b\$2:b\$6)		
3	2	4	7		19,44%	
4	3	7	14		38,89%	
5	4	15	29		80,56%	
6	5	7	36		100,00%	
7	y mayor...	0			100,00%	
8						
9						

En el último paso, y del mismo modo que lo hemos hecho con anterioridad, arrastraremos la marca de la parte inferior de la derecha de la celda para completar la columna. Así, ya tendremos la tabla de frecuencias completa (tabla 9):

Tabla 9. Tabla de frecuencias en Excel

	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	3	8,33%	8,33%	
3	2	4	7	11,11%	19,44%	
4	3	7	14	19,44%	38,89%	
5	4	15	29	41,67%	80,56%	
6	5	7	36	19,44%	100,00%	
7	y mayor...	0			100,00%	
8						
9						

5.2. Análisis exploratorio bivariante

El análisis exploratorio bivariante, igual que el univariante, se realiza tanto con variables cuantitativas como con variables cualitativas. Nos centraremos en el análisis bivariante con variables cualitativas. Específicamente, aprenderemos a hacer tablas de contingencia. Las razones son principalmente dos: en primer lugar, las tablas de contingencia son profusamente utilizadas en el análisis de datos; en segundo lugar, debemos saber que cualquier variable cuantitativa puede transformarse en una variable cualitativa.

Es importante recordar que...

... este curso es introductorio y, por lo tanto, solamente pretendemos lograr unos conocimientos básicos de análisis de datos. En este sentido, el análisis exploratorio bivariante con variables cuantitativas quedará fuera de nuestros objetivos.

5.2.1. Transformación de variables cuantitativas en cualitativas

La transformación de variables cualitativas en variables cuantitativas tiene como razón principal simplificar la interpretación de las variables, de tal modo que la clasificación en categorías facilite la toma de decisiones.

El proceso de conversión de una variable cuantitativa en una variable cualitativa se denomina categorización.

Esta categorización se puede llevar a cabo siguiendo criterios ya establecidos previamente, y que facilitan una lectura comparativa de los datos. La conversión también se puede efectuar por razones teóricas de cualquier tipo, o para facilitar la toma de decisiones de la persona que está realizando el propio análisis de datos.

Existen varias técnicas de categorización de datos cuantitativos. Sin embargo, en el presente curso solo veremos un ejemplo muy sencillo. Para llevarlo a cabo, retomaremos el ejemplo que hemos utilizado en el análisis exploratorio univariante, y continuaremos utilizando como instrumento la hoja de cálculo Excel.

Nuestro punto de partida lo establece la distribución de la variable A1, de la que hemos calculado las estadísticas descriptivas en los apartados anteriores. Una de ellas ha sido la media. Así, la media aritmética de la variable A1 es ($\bar{x} = 5,11$).

Podríamos pues categorizar la variable A1 en una nueva variable que denominaremos A1_COD en función de si los valores de A1 son menores o mayores que la media aritmética. En la tabla 10 se puede observar el procedimiento y el resultado en forma de nueva variable cualitativa de esta categorización.

Tabla 10. Categorización de una variable en Excel

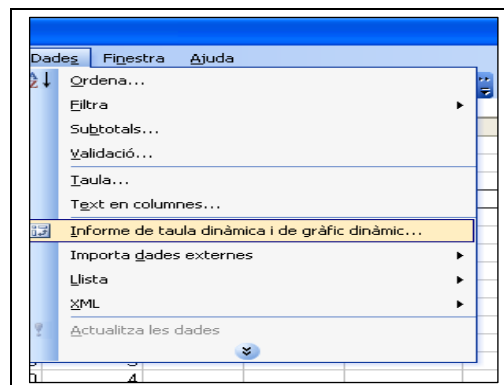
	A	B	C	D	E	F	G	H	I	J
1	ID	A1 cod	A2	A3	A4	A4				
2	46.847	2	Inferior	0	0	4	1			
3	46.626	9	Superior	0	0	2	2			
4	46.715	3	Inferior	0	0	3	3		PROMEDIO	5,11
5	46.757	7	Superior	0	0	1	4			
6	46.665	4	Inferior	0	0	5	5			
7	46.636	0	Inferior	0	0	4				
8	47.259	1	Inferior	0	0	5				
9	46.754	10	Superior	12	2	2				
10	46.507	1	Inferior	0	0	4				
11	46.861	8	Superior	0	0	3				
12	46.849	13	Superior	0	2	2				
13	46.788	12	Superior	0	0	3				
14	46.929	1	Inferior	0	0	4				
15	46.808	1	Inferior	0	0	5				
16	46.959	8	Superior	0	0	5				
17	46.655	5	Inferior	0	0	4				
18	46.645	10	Superior	1	0	2				
19	46.536	0	Inferior	0	0	3				
20	46.532	10	Superior	0	0	3				
21	46.714	1	Inferior	0	0	4				
22	46.886	5	Inferior	1	0	1				
23	46.747	8	Superior	0	0	4				
24	46.726	3	Inferior	0	0	5				
25	46.967	9	Superior	0	0	3				
26	46.914	2	Inferior	0	0	1				
27	46.958	20	Superior	0	0	4				
28	46.698	3	Inferior	0	0	4				
29	46.743	3	Inferior	0	0	5				
30	46.608	4	Inferior	0	0	4				
31	46.821	2	Inferior	0	0	4				
32	46.699	3	Inferior	0	0	4				
33	46.738	1	Inferior	0	0	5				
34	46.518	0	Inferior	0	0	4				
35	46.974	4	Inferior	0	0	3				
36	46.658	1	Inferior	0	0	4				
37	46.748	10	Superior	0	0	4				
38										

5.2.2. Análisis exploratorio bivalente con variables cualitativas

El análisis exploratorio bivalente con variables cualitativas más habitual se realiza mediante la construcción de una tabla de contingencia. Para poder trabajar con variables cualitativas en Excel necesitamos utilizar las tablas dinámicas.

Para construir una tabla dinámica seguiremos los siguientes pasos:

1. En el menú [DATOS] seleccionaremos la opción [INFORME DE TABLA DINÁMICA Y DE GRÁFICO DINÁMICO]. Nos aparecerá un cuadro de diálogo en el que indicaremos la localización de los datos que queremos analizar y el tipo de información que queremos crear.



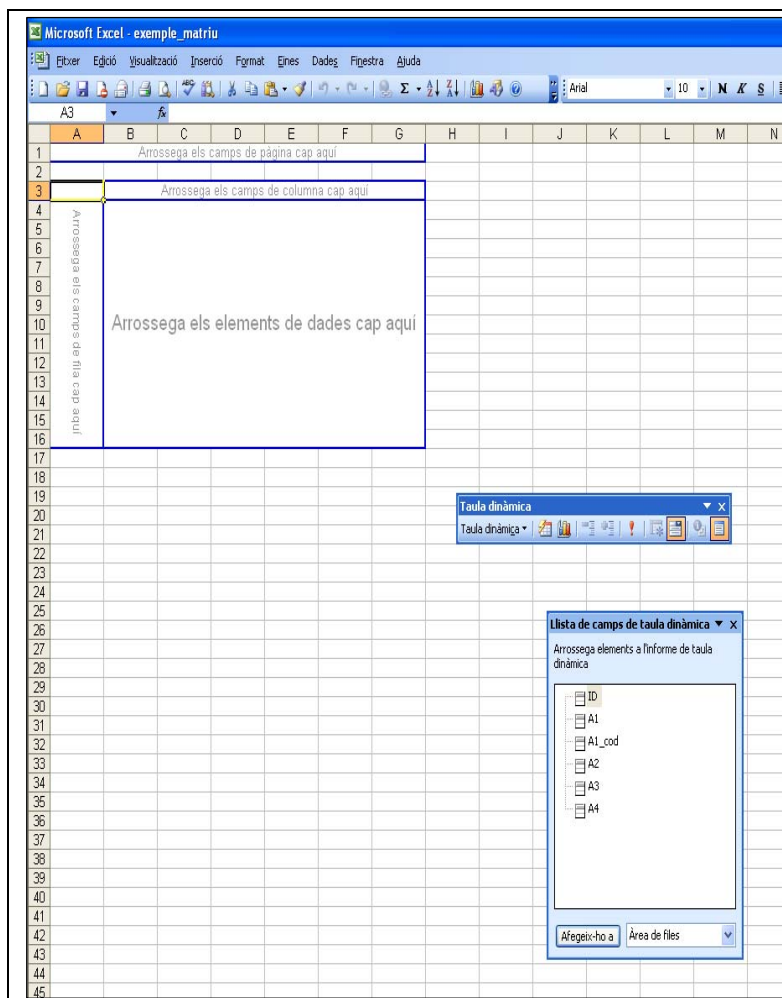
2. En nuestro caso, seleccionaremos las opciones por defecto. Es decir, el origen de los datos será una [LISTA O BASE DE DATOS DE MI-

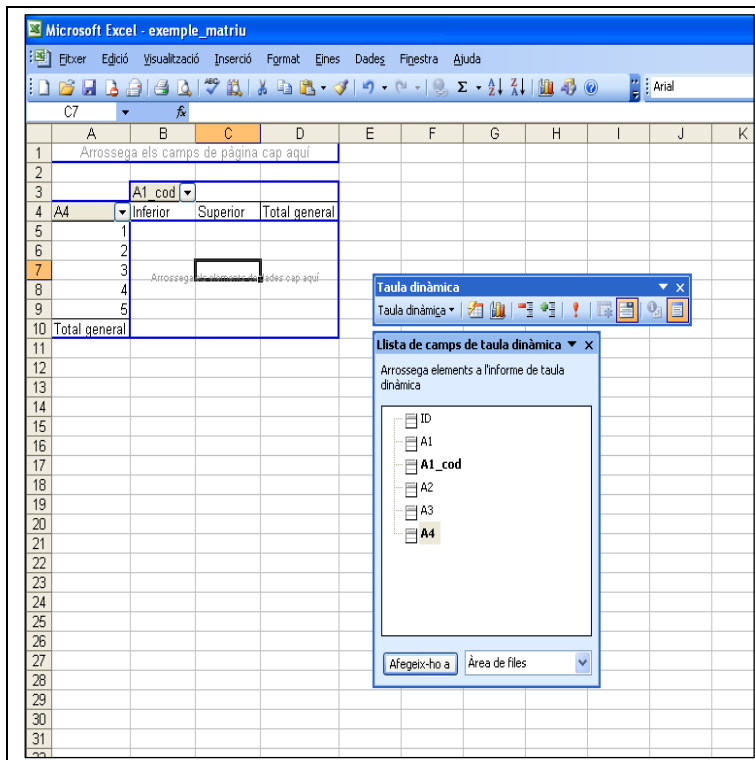
4. Llegados a este punto, nos aparecerán en una nueva hoja los cuadros de diálogo que nos guiarán para poder construir la tabla dinámica.
5. En primer lugar, debemos seleccionar las variables que la compondrán y situarlas en la tabla. Arrastraremos una de las variables y la situaremos en columna, y después arrastraremos la otra hacia las filas.

En nuestro ejemplo, situamos la variable A1_COD, la variable cualitativa que hemos categorizado previamente, en columna, y la variable A4, en fila. De este modo nos será fácil observar cómo varía A4 en función de las categorías establecidas (inferior y superior en la media) de A1_COD.

Para facilitar el análisis y la toma de decisiones es habitual...

... situar en columnas la teórica variable independiente o la variable en la que queremos observar la variación de otra según las categorías de la primera.

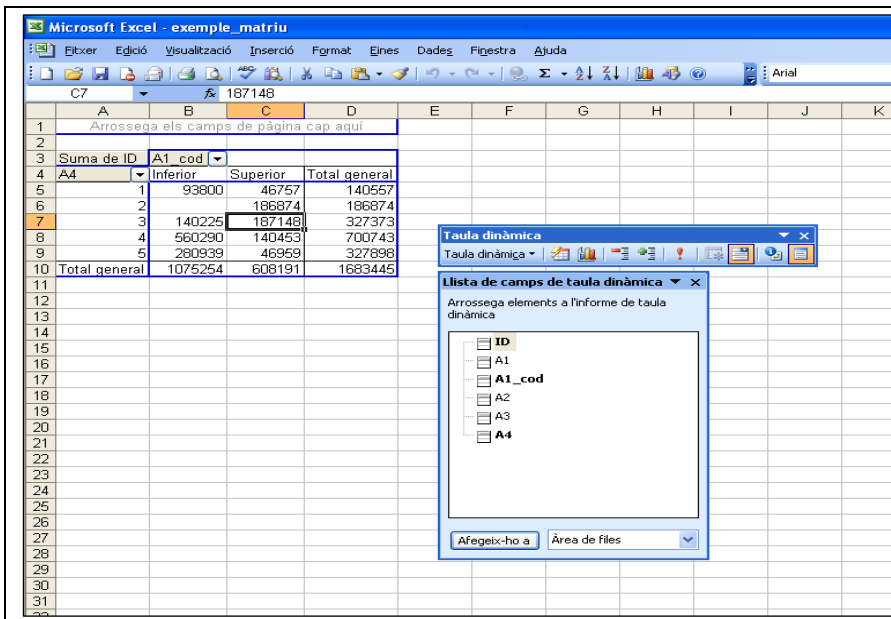




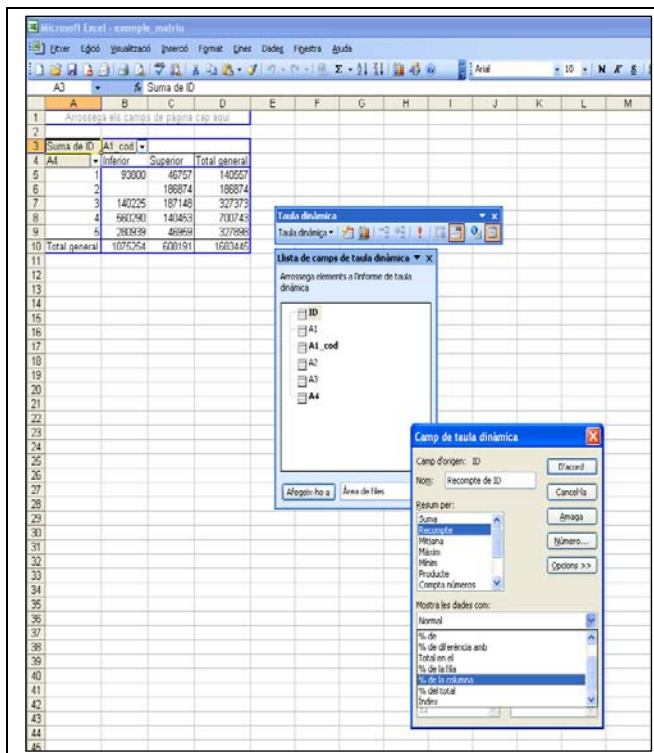
6. Para completar la tabla de contingencia, hemos de decir al ordenador qué es lo que queremos hacer con estas variables. Es decir, qué información queremos sacar de ellas. La lectura de datos que queremos obtener la hemos de indicar en la tabla dinámica.

En nuestro ejemplo, y como se puede observar en la figura siguiente, vemos cómo hemos arrastrado a la celda A3 la variable ID. Esta variable se denomina identificador y es una variable que identifica todos y cada uno de los diferentes registros (filas). Para poder identificar el registro, cada valor (modalidad) de la variable debe ser diferente. Operar con una variable identificadora también es muy útil para realizar diferentes operaciones. En la práctica, contar identificadores es lo mismo que contar individuos (registros).

Por defecto, Excel da [SUMA de ID] como resumen de la variable. Pero en este caso, lo que queremos es ver la distribución de los individuos en una variable en función de las categorías de otra. Así, la suma de los valores de los identificadores de los individuos no es una ayuda, más bien al contrario. No los hemos de sumar, los hemos de contar.



7. Con el objetivo de contar los individuos, haremos doble clic en la celda A3 y se abrirá un cuadro de diálogo llamado [CAMPO DE TABLA DINÁMICA]. En este cuadro seleccionaremos un resumen por [RECUENTO]. Seguidamente haremos clic en [OPCIONES], que es donde podremos seleccionar el modo en el que queremos que nos sean mostrados los datos. Como lo que queremos es ver la distribución de A4 según A1_COD, lo que haremos es elegir el [% DE LA COLUMNA].



8. Finalmente, obtenemos nuestra tabla de contingencia (tabla 11). La lectura que debemos hacer es vertical, puesto que lo que aparece son los porcentajes de la columna.

Si observamos detenidamente los datos, podremos observar que la distribución de la variable A4 es muy diferente para los individuos con valores de A1 inferior a la media, que para los individuos que tienen valores de A1 superior a la media. Los individuos con valores pequeños en A1 muestran una mayor tendencia a presentar valores mayores en A4.

La información que hemos obtenido mediante el procesamiento de datos, y que está presente en la tabla de contingencia, nos puede llevar a pensar que existe una relación entre las respuestas de las variables tratadas.

Razonar esta información, completándola con análisis estadísticos de mayor profundidad, muy posiblemente optimizará cualquier toma de decisiones más allá del conocimiento que hayamos generado en el análisis de la tabla de contingencias.

A título práctico:

El trabajo con tablas (y gráficos) dinámicas con Excel es muy sencillo e intuitivo.

Las pruebas por ensayo y error, directamente relacionadas con las tablas, y el auxiliar de tablas dinámicas son una excelente manera de comprender rápidamente las posibilidades de esta herramienta.

Tabla 11. Tabla de contingencias en Excel

Microsoft Excel - exemple_matriu

Fitxer Edició Visualització Inserció Format Eines Dades Finestra A

D11 fx

	A	B	C	D	E	F
1						
2						
3	Recompte de ID	A1_cod				
4	A4	Inferior	Superior	Total general		
5	1	8,70%	7,69%	8,33%		
6	2	0,00%	30,77%	11,11%		
7	3	13,04%	30,77%	19,44%		
8	4	52,17%	23,08%	41,67%		
9	5	26,09%	7,69%	19,44%		
10	Total general	100,00%	100,00%	100,00%		
11						
12						
13						
14						

Taula dinàmica
Taula dinàmica ▾

Resumen

En este curso introductorio sobre el análisis de datos hemos trabajado conceptos estadísticos básicos que nos permiten tener una primera visión de cómo los datos pueden apoyar a la toma de decisiones en la empresa. El análisis de datos es un gran instrumento para la toma de decisiones, y esta es imprescindible para el buen funcionamiento de la empresa en el mercado.

En primer lugar, hemos repasado algunos aspectos estadísticos clave. Hemos diferenciado entre datos, información y conocimiento, conceptos que se refieren a realidades muy diferentes pero que, muy habitualmente, se utilizan de manera indistinta. Hemos conocido las preguntas básicas que nos debemos plantear en el momento de iniciar una investigación apoyada en datos. Y, finalmente, también hemos trabajado qué son los datos y cómo se pueden diferenciar.

A partir de aquí, y como este curso se basa en la metodología cuantitativa de análisis de datos, hemos conocido qué es una variable y cuáles son sus diferentes tipos. Hemos trabajado competencias sobre cómo diferenciar entre población y muestra y hemos aprendido a hacer los análisis univariantes y bivariantes sencillos más habituales, tanto en cuanto a las variables cuantitativas como a las variables cualitativas.

En conjunto, a lo largo de este curso hemos efectuado una revisión introductoria a la estadística descriptiva que debe permitir, una vez concluida la comprensión de los materiales didácticos y realizadas las actividades, tener una primera visión de cómo orientar las respuestas a las necesidades de información de la empresa, y conocer unos primeros instrumentos para realizar el análisis estadístico básico para la toma de decisiones.

Glosario

análisis univariante *m* Análisis que mide una de las características de los individuos de una población.

análisis bivariante *m* Análisis que mide una de las características de un conjunto de individuos en función de otra característica.

desviación estándar *f* Raíz cuadrada de la varianza.

desviación mediana *f* Media de las diferencias entre los valores de la variable y la media aritmética en valor absoluto.

frecuencia absoluta *f* Número de individuos que comparten la misma modalidad de la variable.

frecuencia absoluta acumulada *f* Número de individuos que comparten una modalidad o alguna otra menor.

frecuencia relativa *f* Proporción de individuos que comparte la misma modalidad de la variable.

frecuencia relativa acumulada *f* Proporción de individuos que comparten una modalidad o alguna otra menor.

mediana (Me) *f* Valor que divide toda una serie de datos en dos partes exactamente iguales.

media aritmética *f* Valor resultante de dividir el sumatorio de un conjunto de datos por el número total de datos. Cuando la media aritmética se refiere a una población, se utiliza el símbolo (μ) y cuando se refiere a una muestra se utiliza (\bar{x}).

moda (Mo) *f* Valor que más veces se repite, es decir, aquel que tiene la mayor frecuencia.

muestra *f* Subconjunto de una población que representa a escala las características principales de esta.

población *f* Conjunto de individuos que disponen de las características que precisamos observar.

rango (R) *m* Diferencia entre el mayor y el menor valor de una distribución.

tabla de contingencia *f* Tabla para grabar y analizar la relación entre dos o más variables, habitualmente de naturaleza cualitativa nominal u ordinal.

valor de una variable *m* Medida que una variable presenta.

variable *f* Reducción de una dimensión de una realidad.

variable dependiente *f* Aquella variable que suponemos que varía en función de otra variable.

variable dicotómica *f* Caso específico (y muy habitual) de variable cualitativa nominal. Esta variable solamente toma dos valores y se utiliza habitualmente para mostrar la presencia o ausencia de una característica.

variable independiente *f* Aquella variable que funciona como *input* del sistema y su variación incide en la variación otros variables dependientes.

variable cualitativa *f* Aquella en la que el valor de la variable representa una categoría.

variable cualitativa ordinal *f* Aquella variable cualitativa en la que los valores presentan un orden establecido dentro de una escala.

variable cualitativa nominal *f* Aquella variable cuantitativa en la que los valores de las variables representan categorías sin relación de medida entre ellas.

variable cuantitativa *f* Aquella en la que el valor de la variable representa una cifra.

variable cuantitativa continua *f* Cuando el valor no representa ninguna interrupción más allá de las que impone el instrumento de medida.

variable cuantitativa discreta *f* Aquella que presenta una interrupción entre los valores porque los valores intermedios no existen.

variación de una variable *f* La diferencia que existe entre los valores de una variable.

varianza (S^2) *f* Resultado de la división entre el sumatorio de las distancias entre cada dato y la media aritmética de los datos elevadas al cuadrado.

Bibliografía

Levin, J. (1997). *Fundamentos de estadística en la investigación social*. México DF: Oxford University Press.

Montero Lorenzo, J. M. (2009). *Problemas resueltos de estadística descriptiva para ciencias sociales*. Madrid: Paraninfo.

Otamendi, F. J.; Diaz Chao, A. (2011). *Estadística para emprendedores. Lecciones prácticas y casos con Minitab*. Madrid: Addlink Media.