

DOCENCIA EN ESTADÍSTICA CON MICROSOFT® EXCEL: ESTADÍSTICA DESCRIPTIVA

Úrsula Faura Martínez, Isabel Parra Frutos, Fuensanta Arnaldos García,
M^aTeresa Díaz Delfa, Lourdes Molera Peris
Universidad de Murcia

RESUMEN

Con la finalidad de que los alumnos comprendan mejor los conceptos teóricos desarrollados en cualquier curso de estadística y, además, dotar a los futuros profesionales de recursos que les ayuden a manejar y tratar datos del mundo real, es conveniente estudiar la estadística desde una perspectiva más práctica apoyándonos en herramientas informáticas como las hojas de cálculo, más versátiles que el software puramente estadístico (SPSS, Statgraphics, S-PLUS,...). En particular, centramos nuestra atención en Microsoft® Excel, dada su amplia implantación actual.

El primer problema que se plantea a la hora de analizar una gran cantidad de datos es cómo obtener una visión inicial de la información que contienen los mismos, que permita ofrecer una primera valoración del comportamiento de la característica que se está estudiando. Por ello es conveniente disponer de resúmenes como los proporcionados por las tablas de frecuencias, ya sean unidimensionales, bidimensionales o condicionadas, tanto para datos agrupados como no agrupados en intervalos. Para lograr este objetivo de una forma rápida y sencilla, utilizamos las *tablas dinámicas* de Excel.

Palabras clave: estadística descriptiva, tabulación, hoja de cálculo.

1. INTRODUCCIÓN

Las técnicas básicas de análisis de datos, objeto de la Estadística Descriptiva, son cada día más utilizadas como elemento de apoyo a la toma de decisiones en cualquier ámbito, pero más aún lo son dentro del mundo de la Economía y la Administración de Empresas. Son técnicas que requieren de procesos de cálculo que, en muchas ocasiones, son tediosos y necesitan de un cierto nivel de esfuerzo o dedicación. Tradicionalmente, y quizás por ello, los estadísticos han sido considerados calculistas, siendo su tarea principal la de sintetizar una gran masa de información y, así, reducir la incertidumbre asociada a toda toma de decisiones. Sin embargo, esta situación traía consigo un cierto alejamiento del usuario final de las técnicas, especialmente cuando se disponía de una gran cantidad de datos. Hoy día la situación ha cambiado, desde hace tiempo se dispone de herramientas informáticas adecuadas y de gran potencia que simplifican los cálculos, pero su manejo requiere, lógicamente, un entrenamiento.

La informática descentralizada es un hecho en el mundo empresarial que permite que cada usuario, cada responsable de un área o departamento de la empresa que requiera información, encuentre posibilidades para su tratamiento autónomo. Raro es ya encontrar una empresa, que disponga de un personal mínimamente cualificado, en la que no esté instalado el software más difundido actualmente en el mercado, el paquete Microsoft® Office, que incluye la hoja de cálculo Excel.

Para hacer más efectiva la enseñanza de las diferentes herramientas estadísticas, ya sean de descriptiva, de modelos de probabilidad o de inferencia, es aconsejable darle un enfoque práctico. Sin embargo, en la docencia en Estadística Descriptiva usualmente, y por motivos didácticos y de facilidad de cálculo, se han utilizado tradicionalmente ejemplos prácticos basados en conjuntos pequeños de valores o, si eran grandes, se presentaban tabulados, dejando al alumno desprovisto de herramientas que le permitieran afrontar el análisis estadístico de un número elevado de datos que no estuviesen organizados, situación que en la práctica puede ser más frecuente que las anteriores. El objetivo del presente trabajo es proponer diferentes prácticas a realizar con la hoja de cálculo Excel que incidan en este vacío docente.

2. LA DOCENCIA EN ESTADÍSTICA DESCRIPTIVA Y LAS HERRAMIENTAS INFORMÁTICAS

Una de las ramas de la Estadística más accesible por su bajo nivel de dificultad y por su utilidad es la Estadística Descriptiva. Además, constituye el paso previo en cualquier análisis de datos más sofisticado. Se dedica, fundamentalmente, al ordenamiento y tratamiento mecánico de la información para su presentación por medio de tablas y representaciones gráficas, así como a la obtención de algunos estadísticos útiles para la explicación de la(s) característica(s) poblacional(es) objeto de estudio.

En el mundo real se ha generado una dinámica de producción de datos enorme, datos que es necesario comprender para obtener la máxima información que nos permita conocer la realidad en la que se está inmerso, así como tomar las decisiones oportunas. Para ello se hace necesario una depuración y procesamiento de los datos. Aquí es donde la Estadística Descriptiva juega un papel fundamental. El análisis estadístico, por tanto, empieza con el objetivo de resumir la información disponible (datos) de tal forma que podamos obtener una primera impresión del comportamiento de la característica poblacional bajo estudio en forma de tablas y representaciones gráficas. En particular, los problemas prácticos de la tabulación son precisamente los que pretendemos abordar en este trabajo.

En los cursos tradicionales de Estadística Descriptiva (sin el complemento de herramientas informáticas) se impartían una serie de conocimientos muy útiles; sin embargo, en el mundo real el alumno encuentra un problema que desde una asignatura de Estadística Descriptiva no se le había solucionado, constituyendo dicho problema, en ocasiones, un obstáculo que le impedía la aplicación de las valiosas herramientas estadísticas que había estudiado. Nos referimos al hecho de que para la aplicación de buena parte de los conocimientos que les transmitimos necesitan disponer de tablas de frecuencias. Estas tablas se podrían elaborar manualmente si se dispone de pocos datos, con la consiguiente pérdida de tiempo y cierta probabilidad de error en la elaboración. Sin embargo, hoy día se produce una gran cantidad de datos que los alumnos podrían entender si les proporcionamos un abanico completo de herramientas de análisis, tanto informáticas como estadísticas. De ahí nuestro empeño desde hace unos años, puesto que la tecnología y disposición de medios en nuestra universidad lo permiten, en introducir prácticas en el aula de informática que les ayuden a enfrentarse a situaciones

reales, esto es, a una gran cantidad de datos correspondientes a variables multidimensionales.

Nuestro objetivo aquí, como ya avanzamos, es la tabulación. Las tablas de frecuencias tienen un doble objetivo: resumir la información disponible y simplificar el cálculo de estadísticos. También es cierto que cuando se recurre a herramientas informáticas el segundo objetivo pierde validez dado que los programas informáticos (hojas de cálculo o programas estadísticos) proporcionan funciones que calculan estadísticos directamente a partir de los datos sin organizar, esto es, sin tabular. Sin embargo, el primer objetivo señalado no pierde validez, dado que la tabla nos permitirá obtener una primera visión del comportamiento de la característica poblacional, además de ser la base para la construcción de gráficos cuando trabajamos con hojas de cálculo.

3. OBJETIVOS DE LAS PRÁCTICAS DE TABULACIÓN

Las prácticas de tabulación en el aula de informática son las primeras a las que se enfrenta un alumno. En ellas se le proporciona una realidad que debe analizar. Por ejemplo, en una de las prácticas llevadas a cabo se deseaba estudiar la actividad comercial de una cadena de tiendas de ropa infantil, en la que se proponía una gran cantidad de datos (en concreto $7 \times 178 = 1.246$ datos), para que observara las ventajas y la necesidad de un tratamiento informático. Los objetivos que perseguimos con este tipo de prácticas son varios y en todos ellos la herramienta de Excel a utilizar, aunque de forma diferente, es *Tablas dinámicas*.¹

Objetivo 1. Identificación de población, elemento poblacional y variables.

A partir de un fichero de datos como el de la siguiente figura, pretendemos que perciban cómo (en este caso, sobre la realidad comercial de una empresa) se pueden observar diferentes características poblacionales en los elementos de la población o muestra, que conforman una variable multidimensional. Esto conlleva, en primer lugar, a que el alumno identifique quién es el elemento poblacional que se observa (el cliente) y cómo sobre ese elemento se pueden distinguir diferentes características (tienda en la que ha comprado, zona en la que vive, valor de la compra que ha realizado, descuento que ha obtenido, marca de ropa que ha comprado, edad del cliente -niño/a- y talla comprada). En segundo lugar, a que identifique los caracteres cuantitativos y

¹ Es interesante hacerles resaltar que el término tabla dinámica es exclusivo de una herramienta de Excel y no es un término estadístico, dado que tienden a confundirlo.

cualitativos, con el problema añadido de una columna (CLIENTE) que representa a cada cliente y que no constituye ninguna característica poblacional.

	A	B	C	D	E	F	G	H	I
1	Cliente	Tienda	Zona	Ventas	Descuento	Marca	Edad	Talla	
2	1	T2	B	35,6	3,56	Marca 3	7	9	
3	2	T2	A	159,6	19,96	Marca 1	24	36	
4	3	T2	B	98,05	9,81	Marca 1	32	36	
5	4	T2	E	202,4	20,24	Marca 3	14	24	
6	5	T2	C	162,88	20,29	Marca 3	5	6	
7	6	T2	A	146,97	14,7	Marca 1	23	30	
8	7	T2	C	126,93	9,69	Marca 3	9	12	
9	8	T2	A	225,53	24,55	Marca 2	27	36	
10	9	T2	E	205,57	15,56	Marca 2	19	24	
11	10	T2	C	189,83	20,98	Marca 2	23	30	
12	11	T2	E	DEV	0	Marca 3	DEV	DEV	
13	12	T2	D	214,54	23,45	Marca 2	26	30	
14	13	T2	D	183,29	20,33	Marca 2	15	30	
15	14	T2	B	152,04	13,2	Marca 3	28	36	
16	15	T2	A	119,22	16,92	Marca 3	23	36	
17	16	T1	A	19,77	4,98	Marca 1	15	30	
18	17	T2	A	70,07	2,01	Marca 2	19	24	

Las características poblacionales en nuestro ejemplo son las siguientes:

- TIENDA: código de la tienda
- ZONA: zona de residencia del cliente
- VENTAS: volumen de venta al cliente, en euros
- DESCUENTO: descuento practicado al cliente
- MARCA: marca de los artículos vendidos
- EDAD: edad del cliente (bebe o niño), en meses
- TALLA: talla vendida

Además, para contemplar las situaciones en las que se producen devoluciones, algunos datos han sido sustituidos por DEV.

Objetivo 2. Tabulación de variables cualitativas (atributos).

En general, la tabulación de un atributo consiste en saber las distintas modalidades del atributo y la frecuencia con la que éstas han aparecido. Esta es la teoría que les damos en clase y que ahora deben aplicar. El problema práctico que soluciona Excel es conocer las distintas modalidades que tiene el atributo y después contar las repeticiones cuando disponemos de un número elevado (178) de observaciones.

Una vez obtenida la tabla de frecuencias es importante que entiendan su contenido. Así, para el atributo TIENDA la información que ofrece es el número de clientes que ha tenido cada una de las tres tiendas (en el periodo estudiado).

TIENDA	n _i
T1	61
T2	84
T3	33
	178

Objetivo 3. Tabulación de variables cuantitativas.

Este caso es más complejo. Por ejemplo, para la variable EDAD, el alumno se va a encontrar con dos problemas. En primer lugar, si realiza la tabulación siguiendo los pasos que ya ha visto con los atributos, el resultado que obtendría sería incorrecto dado que aparecería el dato DEV que no es un valor posible de la variable. Por lo tanto, se enfrenta a la depuración de los datos (mediante la opción *Autofiltro* de Excel).

El segundo problema es la gran cantidad de valores distintos que toma esta variable y que se refleja en una tabla de frecuencias excesivamente larga que no cumple el propósito buscado de resumir la información. Los alumnos entienden aquí la necesidad de agrupar los datos. Una vez planteada esta necesidad, hay que abordar como dar una solución. Aparecen nuevas cuestiones, ¿cuántos intervalos hay que hacer? ¿qué amplitud deben tener? El alumno ahora es consciente del problema práctico que se le va a presentar. En esta situación la solución no es única y depende en buena medida del uso final que se le vaya a dar a la tabla. En el caso de la variable EDAD quizás resulte interesante estudiar tramos de edades que correspondan con las diferentes tallas. En general, puede existir un condicionante para la agrupación o no, pudiendo elegir en esta última situación la agrupación que uno desee, observando unas reglas generales. Todo esto, además, muestra que la tabulación no es algo mecánico, que el alumno debe tener presente qué datos está estudiando, en qué unidades están medidos², etc., esto es, debe hacer un esfuerzo por involucrarse en la realidad que está estudiando.

La agrupación de datos en intervalos se puede hacer de diversas formas, bien de manera automática con la opción de Excel *Agrupar*³, bien determinando a priori el número de intervalos de igual amplitud que se desean construir, o bien la amplitud constante que se desea que tengan los intervalos. En estos dos últimos casos, hay que determinar el rango de la variable y hacer los cálculos correspondientes que permitan la construcción (no automatizada en Excel) de los intervalos. Aquí, una vez más, el alumno debe decidir si el extremo inferior del primer intervalo y el extremo superior del último intervalo van a coincidir o no con el dato más pequeño y más grande, respectivamente, de la muestra. Esta decisión puede depender de la clase de datos que tengamos (valores enteros o no), del tipo de intervalos que se quiera hacer (extremos

² Una problema frecuente de los alumnos es que cuando calculan la edad media del cliente (niño o bebé) la expresan en años, dando como resultado una edad media de 19 años.

enteros o con decimales), etc. En estos casos, además, para calcular las frecuencias absolutas se puede recurrir a la función matricial de Excel denominada *Frecuencia*.

EDAD	n _i	EDAD	n _i
0-3	6	0-3	6
3-6	12	3-6	12
6-9	13	6-9	13
9-12	9	9-12	9
12-15	22	12-18	41
15-18	19	18-24	32
18-21	18	24-36	53
21-24	14		166
24-27	15		
27-30	12		
30-33	12		
33-36	14		
	166		

Objetivo 4. Tabulación de variables bidimensionales.

Son diversas las variables unidimensionales que se pueden estudiar conjuntamente. A continuación describimos dos ejemplos.

* *Estudie conjuntamente ZONA y TIENDA.*

La tabla de frecuencias absolutas conjuntas de ambas variables, en este caso tabla de contingencia, es

TIENDA	ZONA					
	A	B	C	D	E	
T1	17	18	14	5	7	61
T2	27	16	21	9	11	84
T3	13	4	9	6	1	33
	57	38	44	20	19	178

en la que se pone de manifiesto, por ejemplo, que las tres tiendas son visitadas por clientes residentes en las cinco zonas, dado que no hay ninguna frecuencia absoluta conjunta igual a cero.

* *Estudie conjuntamente TALLA y EDAD. ¿Qué conclusiones extrae en cuanto a las tallas utilizadas? Elabore una tabla de edades para las diferentes tallas que sirva de ayuda, tanto a los vendedores como a los clientes que entran a la tienda, a la hora de solicitar alguna talla.*

³ Con esta opción se debe tener en cuenta que el tipo de intervalo que realiza Excel es cerrado por la izquierda y abierto por la derecha. El tipo de intervalo no tiene consecuencias si ninguno de los extremos

EDAD	TALLA								
	3	6	9	12	18	24	30	36	
0-3	1		5						6
3-6		2	8	2					12
6-9			3	4	6				13
9-12				3	4	2			9
12-18					7	26	8		41
18-24						9	17	6	32
24-30							9	18	27
30-36								26	26
	1	2	16	9	17	37	34	50	166

La cuestión propuesta en cuanto a la elaboración de una tabla de ayuda tanto a vendedores como a compradores que indique las tallas más adecuadas para cada edad del niño no tiene una solución única. La construcción de la tabla de doble entrada simplifica la tarea, ya que en ella pueden observar para qué edades se está vendiendo cada talla y a partir de aquí elegir un estadístico calculado sobre la edad (media, moda, máximo, etc.) que crean adecuado para elaborarla.

Objetivo 5. Tabulación de variables condicionadas.

En el contexto del ejemplo que estamos analizando se pueden estudiar múltiples variables condicionadas. Una de las propuestas puede ser la siguiente, en la que, además, añadimos cuestiones adicionales.

- * *Considere ahora el carácter poblacional TIENDA pero, por un lado, condicionado a que NO se ha producido devolución y, por otro, a que SÍ se ha producido devolución ¿Cuál es la tienda que más ventas realizó finalmente (en número de clientes)?, ¿cuál es la tienda que mayor número de devoluciones recibió?*

Es evidente que se necesitan dos distribuciones condicionadas, por un lado, la de la variable TIENDA condicionada a que se haya producido devolución y, por otro, a que no se haya producido. Estas distribuciones se consiguen en Excel realizando un filtrado de los datos que permite separar aquellos que corresponden a devolución de los que no corresponden a devolución, para después tabularlos por separado.

TIENDA/DEV	n _{i/j}	f _{i/j}
T1	2	16,67%
T2	9	75,00%
T3	1	8,33%
	12	100

TIENDA/No DEV	n _{i/j}	f _{i/j}
T1	59	35,54%
T2	75	45,18%
T3	32	19,28%
	166	100

En el caso de que existiera una variable que recogiera la situación de devolución o no devolución (que no es el caso en nuestro ejemplo) se podría optar por una tabla bidimensional para obtener a partir de ella las condicionadas.

- * *Obtenga la distribución de la ZONA para los clientes de cada tienda. ¿Cuál es la zona de procedencia más frecuente y menos frecuente en cada tienda?*

Para resolver esta cuestión no sería necesario el filtrado de los datos para no incluir aquellos clientes que devuelven su compra. Aquí lo más rápido es elaborar una tabla de contingencia que tabule ZONA y TIENDA, y a partir de ella extraer las distribuciones condicionadas.

ZONA/T1	$n_{i/j}$
A	17
B	18
C	14
D	5
E	7
	61

ZONA/T2	$n_{i/j}$
A	27
B	16
C	21
D	9
E	11
	84

ZONA/T3	$n_{i/j}$
A	13
B	4
C	9
D	6
E	1
	33

- * *Determine la marca más vendida en cada zona y en cada tienda. Explique estadísticamente cómo obtiene este resultado.*

El concepto estadístico que se está pidiendo a los alumnos es la moda de distintas distribuciones condicionadas calculadas a partir de las ventas (sin incluir las devoluciones). Sin embargo, tienen dificultad en identificar los conceptos estadísticos involucrados en la pregunta, aunque en general son capaces de resolver lo que se les pide sin aplicar estrictamente la teoría, esto es, recurriendo a la intuición.

El primer problema práctico que se plantea es el filtrado de los datos para eliminar aquellos clientes que han devuelto su compra. Una vez eliminados los datos que hacen referencia a devoluciones, se realizan dos tablas de doble entrada con las variables MARCA y ZONA y con las variables MARCA y TIENDA. A partir de estas tablas es posible extraer las distintas distribuciones condicionadas y obtener de forma sencilla lo que se pide.

Objetivo 6. Tablas de estadísticos de distribuciones condicionadas.

Se trata de tablas que se realizan en Excel de forma similar a las anteriores, pero que no son tablas de frecuencias. El alumno, en general, no se da cuenta de la diferencia y a menudo las trata igual. La realización de gráficos en estos casos, como el diagrama de sectores o el diagrama de rectángulos, es frecuente, incluso en la práctica profesional,

y no puede considerarse incorrecto. Aunque, una vez más, no es consciente de que no está aplicando la teoría estadística expuesta en clase. Veamos algunos ejemplos.

* *Estudie el volumen de ventas y las ventas medias por cliente de cada tienda.*

Para resolver esta cuestión en Excel es necesario insertar una *Tabla dinámica* para la variable TIENDA introduciendo en el campo *datos* la variable VENTAS, solicitando la *Suma* y el *Promedio* de la misma. De esta forma, para cada modalidad de TIENDA se obtiene la suma de las ventas y las ventas medias.

TIENDA	VENTAS totales	VENTAS medias/cliente
T1	8.130,28	137,801
T2	11.796,08	157,281
T3	5.300,92	165,654
Cadena de tiendas	25.227,28	151,9716

Esta información sobre la variable VENTAS también se podría pedir por zonas, por tallas, por marcas, etc.

Asimismo se les puede plantear una situación más compleja en la que deben recopilar diversos estadísticos de distribuciones condicionadas junto con alguna o algunas distribuciones unidimensionales con la intención de tomar una decisión. Veamos el siguiente ejemplo.

* *La cadena desea conocer el funcionamiento de cada tienda para estudiar su efectividad o rentabilidad, con el objetivo de realizar una mayor promoción de alguna de ellas o incluso dar cursos de formación en ventas a sus empleados si esto fuera necesario. Evalúe la situación y extraiga conclusiones calculando para cada tienda*

1. *total de clientes (con o sin devolución),*
2. *efectividad (porcentaje de clientes que no devuelven su compra),*
3. *volumen total de ventas,*
4. *volumen total de ventas netas (ventas menos descuento),*
5. *la rentabilidad por cliente (ventas medias y ventas netas medias por cliente),*
6. *riesgo (dispersión de las ventas).*

TIENDA	Clientes	DEV	Efectividad	Volumen de ventas	Volumen de ventas netas	Ventas medias	Ventas netas medias	Desv. típica de ventas
T1	61	2	96,721%	8.130,28	7.290,24	137,80	123,56	58,63
T2	84	9	89,286%	11.796,08	10.620,42	157,28	141,61	64,43
T3	33	1	96,970%	5.300,92	4.780,82	165,65	149,40	55,28
Cadena de tiendas	178	12	93,258%	25.227,28	22.691,48	151,97	136,70	61,70

Las decisiones a tomar se pueden justificar, en base a la tabla anterior, de diferentes formas, y en última instancia dependerá de los intereses de la empresa o del empresario. Por ejemplo, se podría señalar el mejor valor en cada estadístico (en negrita) y el peor valor (en cursiva). Y a partir de ahí intentar tomar una decisión en cuanto a la mejor y peor tienda o en cuanto a saber qué falla en cada una, provocando para ello una discusión en clase.

4. CONCLUSIONES

En las últimas décadas se ha ido intensificando notablemente la aplicación de la informática en las diferentes esferas del quehacer económico–social. Un papel relevante en este sentido le ha correspondido al perfeccionamiento de las hojas de cálculo como herramienta que permite manejar gran cantidad de datos y que además, son sencillas de utilizar. Entre estas, destaca la hoja de Microsoft Excel, con la ventaja de que los futuros egresados resultarán directamente beneficiados dondequiera que vayan a realizar su labor profesional al utilizar un software que resulta mucho más accesible que los dirigidos sólo al análisis de datos. Con esta herramienta, los alumnos pueden adquirir una mayor solidez en los conocimientos estudiados, mejorar su capacidad de análisis y síntesis, realizar prácticas más realistas y, fomentar el aprendizaje activo y la motivación. En cuanto a la tabulación, el uso de Excel permite obtener tablas de forma sencilla, totalmente dirigida y controlada por el usuario.

5. REFERENCIAS BIBLIOGRÁFICAS

- ARNALDOS, F.; DÍAZ, M.T.; FAURA, U.; MOLERA, L. y PARRA, I. (2003). Estadística descriptiva para economía y administración de empresas: cuestiones tipo test y ejercicios con Microsoft® Excel. Ed. AC- Thomson.
- BERK K.N. y CAREY P. (2003). Análisis de datos con MS Excel (actualizado para Office 2000). Ed. Thomson.
- CAO ABAD, R. y otros (2001). Introducción a la estadística y sus aplicaciones. Ed. Pirámide.
- FERNÁNDEZ, C. Y FUENTES, F. (1995): Curso de estadística descriptiva. Teoría y práctica. Ed. Ariel Economía.
- PÉREZ, C. (2002). Estadística Aplicada a través de Excel. Prentice-Hall, Universidad Autónoma de Madrid.