

Relaciones entre dos variables

Una visión de urgencia

Vicente Manzano-Arrondo, 2014

Índice de contenido

Dos apuntes antes de comenzar – 2
De qué va esto – 2
Hola, ¿qué puedo hacer con este cuestionario? – 2
Esquema para el estudio de una relación – 3
Traducir el objetivo o interés de investigación en una relación entre dos variables – 3
Identificar de qué tipo es cada una de las variables – 4
Visualizar la variable (tabla o gráfica) – 5
Identificar problemas y solucionarlos – 8
Generar, si procede, estadísticos específicos – 8
Visualizar la relación (tabla o gráfica) – 9
Tabla de contingencia – 9
Diagrama de dispersión – 10
Diagrama de medias – 12
Identificar problemas y solucionarlos – 13
Cuantificar la relación – 13
V de Cramer – 13
r de Pearson – 14
Eta cuadrado – 15
Realizar inferencia, si procede – 17
Estado de derecho – 18
Lógica y fases en una PSHN – 18
Identificar la compatibilidad – 19
PSHN para dos categóricas – 19
PSHN para dos cuantitativas – 20
PSHN para una categórica con una cuantitativa – 20
Vocabulario mínimo – 20

Dos apuntes antes de comenzar

De qué va esto

Si va a llevarse a cabo una tarea, mejor si se hace bien. Este principio general, como todos los principios, tiene excepciones y matizaciones. Si alguien pretende hacerme daño, desearía que su cometido no tuviera éxito. De cuantas situaciones cuentan con objetivos que pudiera catalogar como positivos, loables, felicitantes... algunos tienen consecuencias más trascendentes que otros, algunos son más robustos (admiten más problemas por el camino) que otros. Si necesitas que un cirujano opere en tu interior, exigirás que el quirófano se encuentre en condiciones óptimas de higiene, con buenos profesionales y todos los medios a su alcance. Imagina ahora que tienes un accidente caminando por la calle y hay que tapar rápidamente una herida para que no te desangres. Por allí pasa un cirujano. Tiene conocimientos suficientes como para salvarte, por ejemplo, arrancándote una parte del pantalón (bastante sucio, por cierto) y haciendo un torniquete. El cirujano tiene las manos llenas de una mezcla diversa de grasas, cereales y no sabemos qué más, porque en esos momentos estaba poniéndose perdido con una hamburguesa. Él puede salvarte la vida. Pero tú gritas "¡No! ¡Que me lleven a un quirófano y que este señor se duche!". Así que son tus últimas palabras.

En la teoría del muestreo decimos que cuanto mayor es el tamaño de una muestra aleatoria, mejores serán las inferencias posteriores. Pero somos conscientes de que conforme una muestra aumenta de tamaño, también hace el trabajo más difícil, cuesta más dinero, requiere más tiempo... Cuando se ponen todas las consecuencias sobre la mesa, las decisiones dejan de ser simples.

Todo este discurso pretende justificar este documento. Se inserta en las situaciones donde alguien se ha caído en medio de la calle o se requiere una muestra para hacerse realidad de algo, pero con muy bajo coste. Cuando se opera con pocos medios (incluyendo de conocimiento), no se puede pedir unos resultados impresionantes. Pero menos da una piedra. Si tienes que llevar a cabo una relación entre dos variables y no te encuentras en una situación ideal para resolver de forma pausada y excelente el reto, entonces quizá puedas tirar de las respuestas que encuentras aquí.

Este documento puede servir también para romper el hielo en el mundo de las relaciones, con el objetivo de profundizar y fundamentar en temas sucesivos, así como para recordar aprendizajes ya adquiridos pero algo oxidados.

Hay multitud de situaciones en las que interesa estudiar la relación entre variables. De esa inmensidad, vamos a considerar únicamente la existencia de dos variables, no más. Pueden ser de tipos diferentes, pero vamos a agruparlos solo en dos: variables categóricas o cuantitativas. Los objetivos también pueden diferir, así como las pruebas o comparaciones de relación. Pero aquí sólo vamos a considerar una posibilidad para cada una de las tres combinaciones posibles (categórica-categórica, categórica-cuantitativa y cuantitativa-cuantitativa).

El último apartado se encarga de la definición de un vocabulario mínimo imprescindible para entender correctamente estos asuntos. El primer apartado presenta el esquema general para estudiar la relación. Los demás se encargan de ir desarrollando ese esquema, punto a punto. He acudido a la explicación más breve que me ha parecido suficiente y a ejemplificarlo todo, si bien con un único caso y esperando que sea sencillo. Por último: salvo en algún caso excepcional, no vamos a realizar cálculos. Se entiende que ya sabes hacerlo, o encontrar cómo se hace, o acudir a un programa de ordenador que realiza las operaciones de cálculo.

Hola, ¿qué puedo hacer con este cuestionario?

Por el despacho de la universidad pasa una sucesión de investigadores, tanto con experiencia como en formación, estudiantes y miembros de organizaciones diversas, con las mismas inquietudes una y otra vez. Por lo general, han recogido datos a través de cuestionarios, observaciones u otros procedimientos y ahora se plantean ¿qué puedo hacer con toda esa información?

El modo en que se lleva a cabo una investigación científica a partir de datos es muy conocido. Todas las personas que pasan por el despacho lo conocen. Básicamente consiste en 1) definir con claridad los objetivos, 2) traducir los objetivos a un procedimiento para recoger información, 3) recogerla, 4) procesarla, 5) establecer conclusiones. Con unos términos u otros, con mayor generalidad o concreción, todas estas personas conocen el esquema. Pero no lo cumplen, ni son conscientes de la trascendencia que implica no cumplirlo. La fase 1 es “definir con claridad los objetivos”. Cuando alguien me pregunta “¿Qué puedo hacer con estos datos?”, yo le respondo “¿Cuáles eran tus objetivos?” y algunos rostros muestran sorpresa ante la pregunta. Cuando consigo conocer los objetivos del estudio, descubro que hay mucha más información de la necesaria y parte de la necesaria no se ha recogido. Algo pasó en el diseño y ejecución del estudio, pues se olvidaron de lo más relevante. No es raro. Pasa en la vida cotidiana.

Así pues, una primera tentación que hay que vencer a la hora de afrontar el análisis de los datos es buscar cuanto pueda obtenerse a partir de los datos obtenidos. El análisis no funciona así. El camino consiste en 1) traducir los objetivos de investigación a objetivos de análisis, 2) realizarlos y 3) volver a los objetivos con respuestas estadísticas. Por ejemplo, nos planteamos el objetivo de encontrar si ser hombre o mujer afecta a la importancia que se da a renovar el vestuario. La pregunta 2 del cuestionario es el sexo biológico (1.Hombre – 2.Mujer). La pregunta 23 es “Indica a continuación tu grado de acuerdo con la afirmación «Hoy en día resulta imprescindible ir renovando el vestuario continuamente» (Desde 1.Muy en desacuerdo; hasta 5. Muy de acuerdo). El objetivo de investigación puede traducirse en el objetivo de análisis “comparar las medias de acuerdo entre hombres y mujeres”. Tal vez el cuestionario pregunte por el nombre que pondría a un perro negro de caza, o cuántos electrodomésticos tiene en casa. Si esas respuestas no tienen interés para los objetivos del estudio, no tienen presencia en los objetivos de análisis, aunque queden preciosas en el instrumento de recogida de información.

Esquema para el estudio de una relación

1. Traducir el objetivo o interés de investigación en una relación entre dos variables.
2. Identificar de qué tipo es cada una.
3. Estudiar cada variable por separado.
 - a) Visualizar la variable (tabla o gráfica).
 - b) Identificar problemas y solucionarlos.
 - c) Generar, si procede, estadísticos específicos.
4. Estudiar específicamente la relación
 - a) Visualizar la relación (tabla o gráfica).
 - b) Identificar problemas y solucionarlos.
 - c) Cuantificar la relación.
 - d) Realizar inferencia, si procede.

Traducir el objetivo o interés de investigación en una relación entre dos variables

El objetivo de este documento se centra en el análisis de los datos, por lo que no voy a dedicar más que algún que otro consejo general para abordar esta fase. Es un asunto complejo que requiere sentido común y muy especialmente haber tomado decisiones conscientes antes de llegar a este punto.

Los análisis de datos se llevan a cabo con los datos, no con los objetivos. Estos deben ser expresados de tal forma que sea posible referirse a ellos a través de los datos. Así, por ejemplo, puedo tener como objetivo “establecer el grado en que el uso de whatsapp dificulta la memoria”. A

la hora del análisis, ¿dónde está el uso del whatsapp o la dificultad de memoria? Tal vez hemos utilizado un cuestionario para la investigación. En él, el ítem 23 es pregunta por la frecuencia de uso del whatsapp: “Nunca – Cada varios días – Diariamente – Varias veces a lo largo del día”. Los ítems 8 a 12 muestran diversas pruebas de memoria. A la persona que responde al cuestionario se le presenta una tarjeta con objetos durante 20 segundos. Cuando se retira la tarjeta, se le pide que marque con una cruz los elementos del ítem 8 que estaban presentes en la tarjeta. Se comprueba el número de aciertos (máximo 10). Lo mismo ocurre con los ítems restantes hasta el 12. Así que sumamos el número de aciertos de los cinco ítems y tendremos como resultado una variable, que vamos a llamar “memoria”, que va de 0 (ni un solo acierto) a 50 (acierto total). La variable “uso de whatsapp” se expresa directamente según el ítem 23. La variable “dificultades de memoria” es el inverso de la variable “memoria”, pues conforme menor sea el valor obtenido en “memoria”, mayor será la dificultad que expresa. Para no tener problemas de interpretación, calculamos:

$$\text{dificultad de memoria} = 50 - \text{memoria}$$

Con este recurso, la variable “dificultad de memoria” va de 0 (máxima memoria, mínima dificultad) a 50 (mínima memoria, máxima dificultad). Con ello, el objetivo “estudio de la relación entre el uso de whatsapp y las dificultades de memoria” se aborda estadísticamente operando con dos variables: la medición directa del ítem 23 y la variable construida “dificultad de memoria”. El modo concreto con que se analizaría esta relación lo vamos a ir desgranando en los siguientes apartados.

Este ejemplo pretende mostrar las reflexiones y operaciones previas que tal vez tengamos que poner en marcha antes de abordar un análisis de relación entre dos variables.

Si no tienes claro cómo resolver estas situaciones, de casuísticas muy diversas, va a ser difícil que abordes el siguiente paso. Así que echa el tiempo que haga falta para traducir todos tus objetivos a nivel de variables, tal y como hemos hecho en este apartado. Después, sigue con el texto.

Identificar de qué tipo es cada una de las variables

El análisis de datos opera con variables y sus valores. Las operaciones que se pueden realizar dependen de la escala de medida de las variables en juego. Vamos a partir de una clasificación con dos únicas categorías: variables categóricas y variables cuantitativas. Vamos a construir este par a partir de un trío: variable nominales, ordinales y cuantitativas.

Una variable nominal es aquella en la que sus valores son nombres. La provincia andaluza de nacimiento cuenta con ocho posibles valores. Cada uno de ellos es un nombre. Se trata de una variable nominal. En lugar de los nombres originales (Huelva, Córdoba, etc.), podemos utilizar números: 1. Cádiz; 2. Málaga; 3... Esta decisión facilita muchas gestiones. Entre otras ventajas, minimiza la probabilidad de errar cuando se introducen los datos. Es fácil que alguien escriba Cordoba en lugar de Córdoba, o se le vaya una tecla y escriba Códoba. Aunque una persona leyendo estas informaciones identificaría el mismo valor, para un programa de ordenador se trata de tres valores diferentes. Así que utilizar números hace esta tarea más fácil. No obstante no hay que olvidar que en las variables nominales los números siguen siendo nombres. 3 no es el triple de 1, del mismo modo que Sevilla no es el triple de Almería.

Las variables ordinales son también una colección de nombres, pero añaden una particularidad potente a las nominales: los nombres expresan una mayor o menor presencia de aquello a lo que se refieren. Pregunto “¿Cómo te encuentras de ánimo esta mañana?”. Las respuestas posibles son “1.Deprimido – 2.Algo triste – 3.Voy tirando – 4.Con buen ánimo –

5.Pletórico”. Cada estado de ánimo considerado tiene un nombre. El nombre no solo representa un estado de ánimo diferente, sino también *mayor* o *menor* respecto a los que tiene al lado. La respuesta “voy tirando” expresa mejor ánimo que “algo triste” y peor que “con buen ánimo”. Así pues, los números que representan a los estados de ánimo expresan no solo estados de ánimo diferentes, sino también estados mejores o peores. El orden de los números se corresponde con el orden de los fenómenos a los que apuntan los números.

Por último, las variables cuantitativas expresan directamente cantidades y no tiene sentido ponerles nombres. ¿Cuántas llamadas de teléfono has hecho en lo que va de semana? La respuesta 15 es directamente comprensible. No se refiere a otra cosa que a sí misma: 15 llamadas.

En este documento vamos a considerar únicamente dos tipos de variables. Las nominales serán denominadas *categóricas*, pues una variable nominal se refiere a un sistema de categorías: categoría hombre, categoría mujer; categoría Cádiz, categoría Jaén, etc. Las cuantitativas siguen siendo consideradas como cuantitativas. ¿Y las ordinales? Aquí viene un pequeño problema.

El análisis de datos está ideado principalmente para variables cuantitativas. Con ellas podemos llevar a cabo multitud de operaciones. La más famosa de todas es calcular la media aritmética: media de edad en la muestra, media de litros de alcohol consumidos en la feria, media de hijos por familia, etc. No tiene sentido calcular una media con una variable categórica. Imagina que lo haces con la variable *provincia de nacimiento*. ¿Qué significa una media de valor 3,7? Por muchas vueltas que le des, no vas a obtener ningún sentido. 3,7 no tiene significado en esa variable.

Si la variable es ordinal, tendremos que considerar si es tratada como categórica o como cuantitativa. En muchas ocasiones, las ordinales se abordan como cuantitativas para poder aplicar con ellas la potencia de la estadística. Son variables incómodas, pero abundantes. No es raro que la ordinal se trate como categórica en una representación gráfica, pero como cuantitativa en el estudio del efecto en una relación. No hay un listado preciso de criterios para tomar la decisión sobre cómo tratar a una variable ordinal. Lo que hay es un principio fundamental: haz lo que tengas que hacer, que tenga sentido. Si haces una representación gráfica de la relación entre una variable ordinal y una variable cuantitativa, puedes tratar a la primera como categórica o como cuantitativa. De las dos opciones, escoge aquella que te permite conseguir la gráfica más útil, más claramente interpretable. Con los programas de ordenador se hacen análisis con rapidez y poco esfuerzo. Así que ante una duda, prueba las dos opciones y escoge la que de ambas cumpla mejor con la función de permitirte averiguar qué está pasando con tus datos.

Antes de seguir con la lectura, repasa tus objetivos de análisis y clasifica las variables implicadas según su tipo o escala de medida. Todos los apartados que siguen te ofrecen procedimientos que dependen del tipo de variable, por lo que no puedes tomar decisiones concretas si no has resuelto esta etapa.

Visualizar la variable (tabla o gráfica)

Para hacer buenos análisis lo primero es conocer lo mejor posible la materia prima: los datos, organizados en variables. Existe un lamentable y extendido hábito: entrar de lleno en análisis de decisiones, sin haber ocupado un tiempo previo en conversar con las variables, observar su comportamiento y que lleguen a ser familiares y comprensibles. El primer acto en un estudio de relación entre dos variables es, por tanto, tomar a cada una de ellas por separado y realizar una tabla de frecuencias o una representación gráfica para observar qué ocurre.

Del abanico de posibilidades para variables categóricas, una buena opción es un diagrama de barras. Consiste en una gráfica donde el eje horizontal muestra las categorías, mientras que el eje vertical expresa la frecuencia con que se presenta cada categoría en el conjunto de datos. Observa

por ejemplo la representación gráfica de la figura 1. Proviene de un estudio en el que se preguntaba a una muestra de personas sobre el barrio en el que viven. Una de las preguntas se refiere a los aspectos que más le gustan del barrio. Las personas respondieron de forma abierta. Después, el equipo de investigación construyó un sistema de categorías, es decir, clasificó las respuestas anotando si en ellas se había mencionado o no las categorías que ves en la figura 1: la gente del barrio, el ambiente, las posibilidades de ocio, los establecimientos, aspectos relacionados con los transportes y la locomoción, servicios y otros. Las categorías se han ordenado en función de su frecuencia, mostrando primero la categoría con más elecciones. Lo que se expresa no son las frecuencias absolutas, sino los porcentajes. La mayoría de las personas manejamos bien los tantos por ciento, por lo que es una buena opción escogerlo como vehículo para comunicar resultados en la representación gráfica. Hay muchas posibilidades estéticas para expresar en una gráfica. El objetivo es que, sin necesidad de ser desagradable a la vista, no contenga añadidos que obliguen a un extra de esfuerzo de interpretación, y que no genere equívocos: escoge la gráfica más simple o sencilla que se te ocurra. Son las mejores. En el ejemplo, no hay líneas horizontales, sino que cada barra viene coronada por el porcentaje que le corresponde. Quien la mira tiene una primera impresión general intuitiva, gracias a la altura de las barras. Si quiere información más precisa, mirará las cantidades. Es importante que la altura comience en cero. Algunos programas de ordenador maximizan las diferencias. Si la barra de menor altura tiene el valor 30%, por ejemplo, entonces sitúa el origen vertical en 20% en lugar de 0%. Esta decisión facilita conclusiones erróneas, pues exagera la diferencia entre categorías.

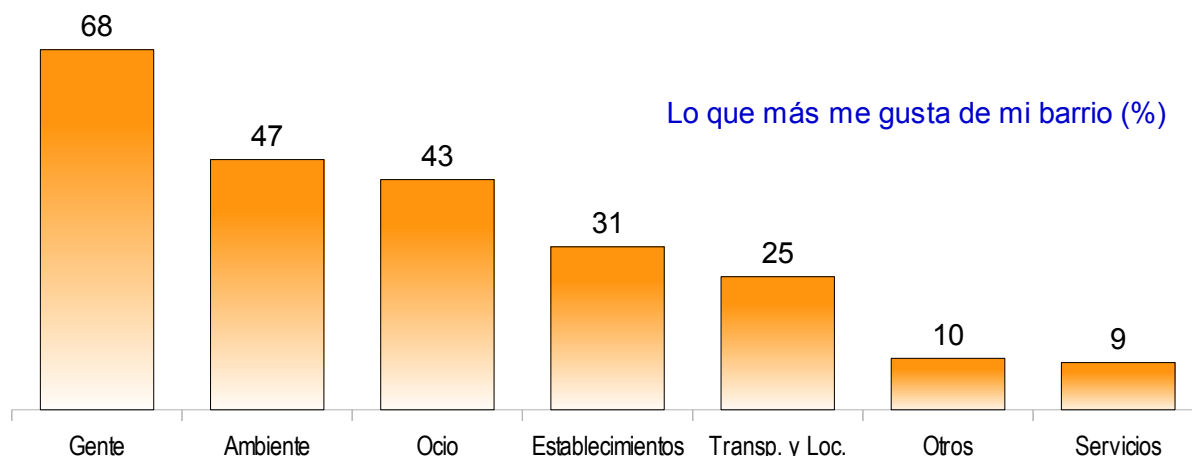


Figura 1. Ejemplo de diagrama de barras.

En la figura 2 tienes otro ejemplo. En este caso las barras se disponen en sentido horizontal. Es una buena opción cuando se representan muchas categorías. En esa investigación, se preguntó a una muestra de algo más de 300 niños y niñas por las cosas importantes de la vida. Se ofrecieron 9 valores como posibles respuestas y los participantes propusieron más. Las respuestas propuestas por los participantes se agruparon en otras nueve, marcándolas en la gráfica con un asterisco. Los valores se expresan en una escala de 1 a 12, donde 1 expresa mínima importancia, y 12 importancia máxima. Las barras se han ordenado en función de la importancia media recibida. Observa que la longitud de cada barra no es ya una frecuencia, sino una media aritmética: la importancia media otorgada por los niños y niñas de la muestra a ese valor o aspecto de la vida.

Además de las representaciones gráficas, una buena tabla de frecuencias es una opción a tener en cuenta. Presenta información de forma menos intuitiva y más precisa. Es una elección buena especialmente en los informes técnicos y para que el propio equipo de investigación observe qué pasa. Pero a la hora de exponer resultados a un público más diverso, las representaciones gráficas constituyen la mejor decisión. La tabla 1 expresa información de una variable ordinal. Si llevas a cabo una suma literal de la columna de porcentajes, el resultado será de 100,1 y no de

100,0. También existen algunos desfases del mismo tipo en la columna de porcentajes acumulados. Ocurre porque primero se han hecho los cálculos y después se han redondeado. Es normal que en los redondeos de las cantidades decimales se vaya acumulando imprecisión. Esto no es preocupante. Lo relevante es que contamos con información numérica de buena calidad para construir una idea sobre lo que está ocurriendo. En el ejemplo, se observa que la acumulación de las respuestas se encuentran entre los valores intermedios-superiores. La gente tiende a estar de buen ánimo antes que lo contrario.

Valor	Frecuencia	%	% acumulado
Deprimido	12	7,4	7,4
Algo triste	29	17,8	25,0
Voy tirando	58	35,6	60,7
Con buen ánimo	43	26,4	87,1
Pletórico	21	12,9	100,0
Total	163	100,0	

Tabla 1. Tabla de frecuencias para una variable ordinal.

Importancia media de los 18 valores

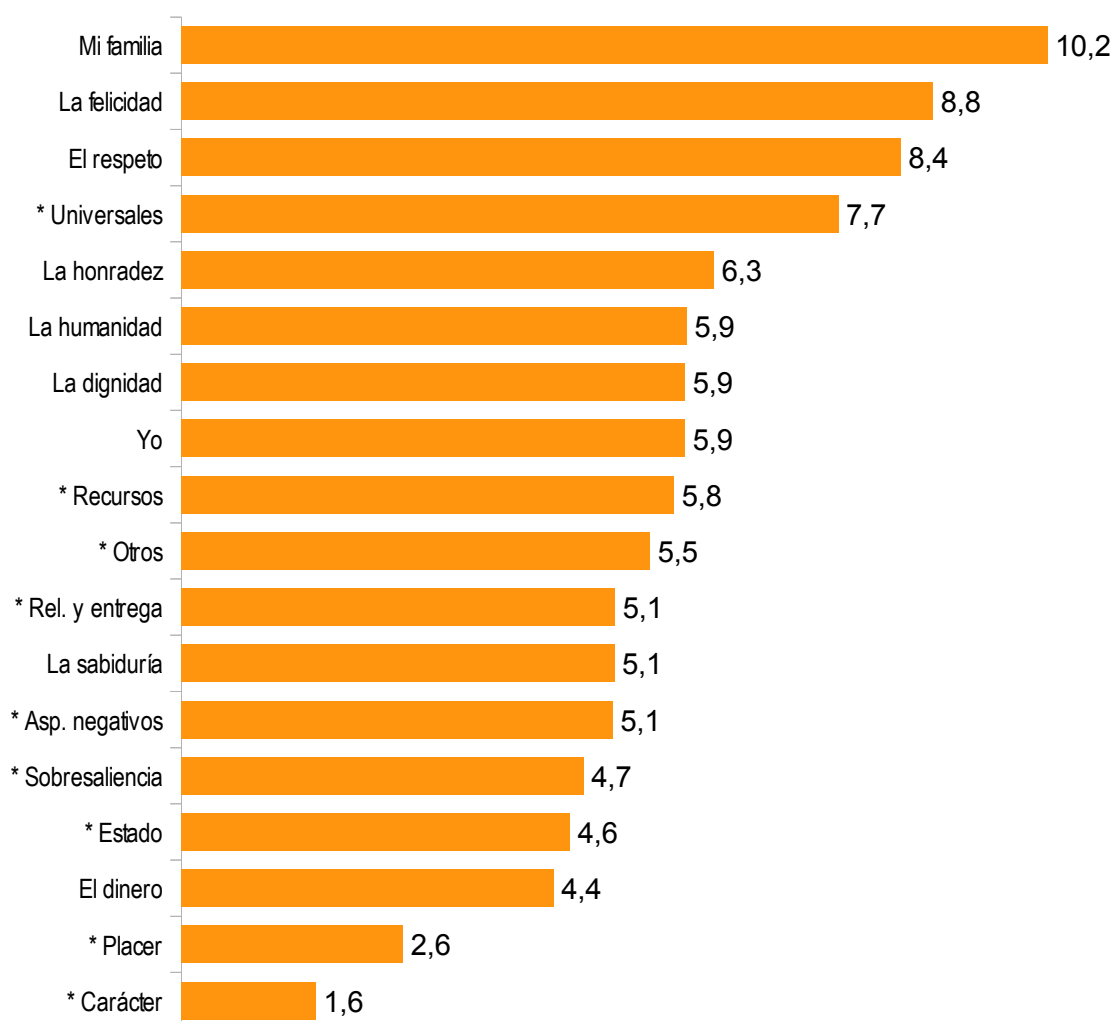


Figura 2. Ejemplo de diagrama vertical de medias.

Identificar problemas y solucionarlos

Las tablas y representaciones gráficas permiten no solo familiarizarse con las variables en juego, sino identificar problemas. Al observar una tabla, por ejemplo, podemos encontrar que hay valores que no deberían existir. Una variable como la de las tablas 1 y 2 cuenta con valores comprendidos entre 1 y 5. Al realizar la tabla se puede observar la existencia de un 0 o un 7. Son dos valores que no deberían estar. No tienen ningún significado específico. Se deben a un error posiblemente cometido durante la introducción de datos en el ordenador. Al observarlo en la tabla, hay que buscar el registro de ese caso y corregir el error.

Si no se lleva a cabo esta labor, esos errores participan de los cálculos posteriores y propician conclusiones que no deberían establecerse.

Generar, si procede, estadísticos específicos

Además de los recursos tabulares y gráficos, es posible ganar información calculando estadísticos específicos. Por lo general, contamos con dos tipos especialmente relevantes: medidas que expresan el valor más característico o que mejor representa al conjunto (estadísticos de tendencia central o de representación numérica) y medidas que expresan la dispersión o variabilidad de las respuestas (estadísticos de dispersión). Estos segundos, además, sirven para valorar a los primeros: si hay mucha variabilidad en las respuestas, escoger un valor concreto para representar al conjunto será una decisión poco acertada. Por este motivo, las medidas de dispersión también se consideran estadísticos de bondad de la representación numérica. A continuación observa la tabla 2, en la que se han insertado columnas de cálculos intermedios para facilitar la obtención de la media aritmética y de la desviación tipo, como estadísticos de representación numérica y de variabilidad.

X	f	Xf	X ² f
1	12	12	12
2	29	58	116
3	58	174	522
4	43	172	688
5	21	105	525
	163	521	1863

Tabla 2. Cálculos intermedios.

$$\bar{X} = \frac{\sum X}{n} = \frac{521}{163} = 3,20 \quad S = \sqrt{\frac{\sum X^2}{n} - \bar{X}^2} = \sqrt{\frac{1863}{163} - 3,20^2} = 1,10$$

La media aritmética tiene el valor 3,2. El valor 3 representa a la categoría “voy tirando”. Observa que se trata de una variable ordinal que estamos tratando como cuantitativa. La información de la media es útil, aunque no haya que tomarse demasiado literalmente el resultado, que abunda en que el ánimo general es intermedio-alto.

La desviación tipo suministra el resultado 1,1. ¿Qué significa? Así, de entrada, no lo sabemos. La media se puede interpretar bien porque se encuentra acotada entre el mínimo y el máximo (entre 1 y 5). Pero la desviación tipo es más difícil. Para solucionarlo, contamos con una transformación: la dispersión acotada. Lo que hace es re-expresar S comparándolo con S máxima. Al dividir S entre la máxima S que podemos encontrar, el resultado se encuentra entre cero (cuando no hay variabilidad ninguna, todos los datos tienen el mismo valor y, por tanto, S vale cero) y 1 (cuando S es máxima y por tanto se divide entre sí misma). Para hacer la interpretación más sencilla, se multiplica el

resultado por 100 y, con ello, contamos con un estadístico que va de 0 a 100. Cuanto más cercano a 0, menos variabilidad; cuanto más cercano a 100, más variabilidad. En el caso del ejemplo:

$$S^A = \frac{100 S}{\sqrt{(\bar{X} - \text{mín})(\text{máx} - \bar{X})}} = \frac{100 \cdot 1,10}{\sqrt{(3,2 - 1)(5 - 3,2)}} = 55$$

La desviación acotada tiene el valor 55%, es decir, la variabilidad del conjunto de datos es un 55% de la máxima que cabría esperar. Eso es mucha variabilidad como para contar solo la media aritmética. Habría que indicar que hay un poco de todo, una dispersión apreciable, si bien cierta coincidencia o agolpamiento de datos en torno a 3,2.

Los programas de ordenador para análisis de datos suministran sin problemas los valores de medias y de desviaciones tipo (también llamadas, desviación estándar y típica), pero no de la desviación acotada. No obstante, como has podido ver, no es algo difícil.

Llegados a este punto, ya sabes llevar a cabo un estudio de urgencia de cada variable por separado. No sigas leyendo hasta que no hayas completado esta fase y conozcas las variables que participan de tus relaciones, corrigiendo los errores si es que ocurren. Los cuatro apartados siguientes versan sobre la relación específicamente.

Visualizar la relación (tabla o gráfica)

Del mismo modo que hemos llevado a cabo un conocimiento previo sobre cada variable por separado, el siguiente paso consiste en visualizar la relación entre ambas, recurriendo también a tablas o a representaciones gráficas.

Para escoger una buena tabulación o gráfica, no basta con tener en cuenta el tipo de variable según su escala de medida. Es necesario contemplar el número de valores. La tabla 3 considera únicamente los tipos de variable. Después veremos cómo afecta el número de valores.

Recurso	Pensado para:
Tabla de contingencia	Dos variable categóricas.
Diagrama de dispersión	Dos variables cuantitativas
Diagrama de medias	Categórica x cuantitativa

Tabla 3. Recursos de visualización.

Tabla de contingencia

Es una tabla, es decir, una organización de filas y columnas, en cuyas casillas se expresa la frecuencia de ocasiones en que se presenta el par valor_fila x valor_columna. En la siguiente tabla se puede estudiar la relación entre el destino turístico preferido y el poder adquisitivo en una muestra de 186 personas.

		Nivel adquisitivo			Total
		Bajo	Medio	Alto	
Destino preferido	Campo	17	43	9	69
	Costa	38	19	11	68
	Ciudad	8	20	21	49
	Total	63	82	41	186

En el ejemplo, vemos que 17 personas de nivel adquisitivo bajo prefieren el campo, que 11 personas de nivel adquisitivo alto prefieren la costa y que 20 personas de nivel adquisitivo medio prefieren la ciudad, por mencionar solo tres pares de los nueve existentes. En los márgenes de la tabla se observa, por ejemplo, que hay 68 personas que prefieren la costa, o que hay 63 personas de nivel adquisitivo bajo.

Observando la tabla queda claro que las personas con bajo nivel adquisitivo prefieren la costa, puesto que hablamos de 38 de 63, frente a 17 para el campo y 8 para la ciudad. Con una reflexión similar se concluye que el campo es el destino preferido para el nivel medio y la ciudad lo es para el nivel alto. Es más fácil establecer estas conclusiones si se utilizan porcentajes. Los porcentajes se pueden calcular según las filas o según las columnas. Ninguna de ambas opciones es mejor que la otra, depende de cuál de ambas perspectivas nos parezca más fácil, cómoda o natural para concluir. En el caso del ejemplo, nos podemos centrar en ver los resultados desde el nivel adquisitivo, por lo que vamos a calcular los porcentajes según esa mirada:

		Nivel adquisitivo		
		Bajo	Medio	Alto
Destino preferido	Campo	27,0	52,4	22,0
	Costa	60,3	23,2	26,8
	Ciudad	12,7	24,4	51,2
	Total	100	100	100

Vemos que el 60% de las personas de nivel adquisitivo bajo prefieren costa, frente al 27% campo y 13% ciudad, por lo que la preferencia parece clara. Del mismo modo se opera con el resto, para llegar a la misma conclusión que en el párrafo anterior, si bien con más facilidad. Los programas de ordenador suelen realizar esta labor (calcular los porcentajes) con sencillez.

Diagrama de dispersión

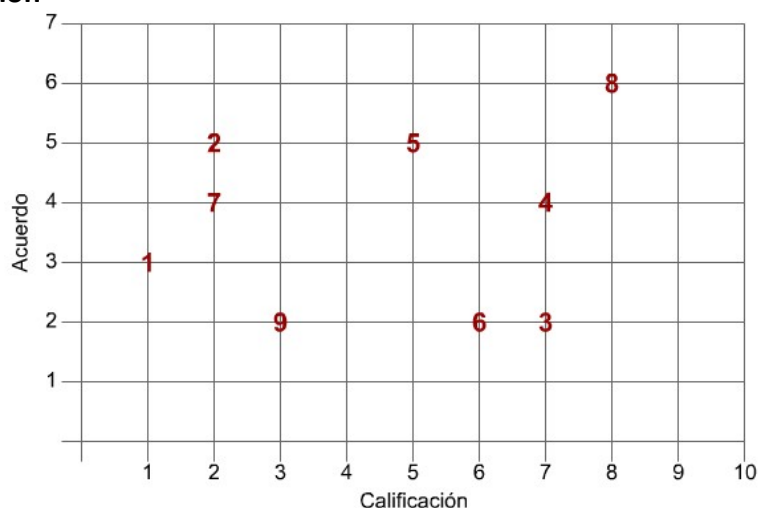


Figura 3. Diagrama de dispersión con identificación de casos.

Las variables cuantitativas no solo se diferencian de las categóricas en la escala de medida. Por lo general cuentan con una mayor amplitud de valores. Cuando dos variables cuentan con muchos valores, una tabla de contingencia genera un número excesivo de celdas o casillas, que hace inviable la interpretación. Ocurre además que la existencia de *cantidad* permite operaciones no posibles en una tabla, como es ordenar los datos según su valor.

Un diagrama de dispersión es una representación gráfica basada en dos ejes. Cada uno de ambos representa a una de las dos variables cuantitativas. Cada caso o unidad del archivo de datos será representado con un punto en la gráfica. En la figura 3 puedes observar un diagrama de dispersión donde se representan los valores de nueve estudiantes, de quienes se ha obtenido un par de datos para cada uno: su calificación de curso y el grado de acuerdo con ello.

El estudiante 1 ha obtenido un 1 y tiene un nivel de acuerdo 3. El estudiante 9 ha tiene un nivel de acuerdo de valor 2 por una calificación de valor 3. El estudiante 8 ha obtenido también un 8, con cuya nota tiene un acuerdo de valor 6. Si se tienen relativamente pocos casos (en el ejemplo, 9), procede identificarlos con un número. No es lo que suele ocurrir. De hecho, con pocos casos, es prescindible realizar una representación gráfica. Lo que hacemos es que cada caso está representado por un punto. La figura 4 muestra un ejemplo con más casos.

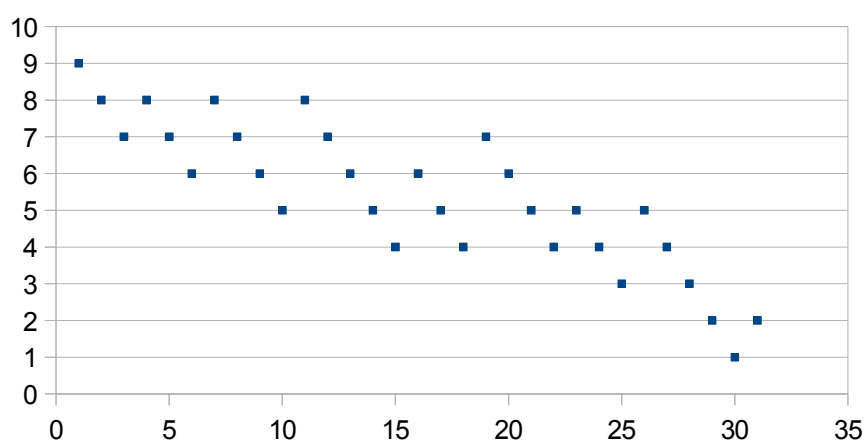


Figura 4. Diagrama de dispersión con puntos.

La interpretación de la figura 4 es relativamente sencilla: se observa que conforme aumenta el valor de la variable del eje horizontal, disminuye el valor de la variable del eje vertical. En cualquier caso, conforme una aumenta, la otra disminuye. Es lo que se denomina una relación negativa o descendente.

En algunas ocasiones, una o ambas de las variables cuantitativas cuentan con muy pocos valores. En tales casos no tiene sentido representar la relación con un diagrama de dispersión, pues no hay suficiente variabilidad. Cada punto representa uno o más casos. Si contamos con dos estudiantes, por ejemplo, con una calificación de 7 y un grado de acuerdo de valor 3, la coordenada (7,3) contará con dos puntos, aunque solo veremos uno. Si hay poca variabilidad y muchos datos, algunas coordenadas estarán saturadas con muchos puntos, mientras que otras contarán con uno solo. En cualquier caso, no podremos tener una idea sobre qué ocurre, puesto que falta información sobre la densidad de los cruces de coordenadas. Si ocurre así y solo con intención de visualizar la relación, habrá que cambiar de tabla o gráfica. Si la falta de variabilidad suficiente le ocurre a una sola variable, entonces la trataremos como categórica y llevaremos a cabo un diagrama de medias. Si le ocurre a las dos, habrá que acudir a una tabla de contingencia. Imagina que queremos estudiar la relación entre el número de hermanos que tiene cada persona de una muestra, y el número de teléfonos móviles que poseen entre todos. Se trata de dos variables cuantitativas. Observa la figura 5, que muestra el diagrama de dispersión que le corresponde a los mismos datos que puedes observar en la tabla 4, que muestra la tabla de contingencia. Aunque se trate de dos variables cuantitativas, al contar con tan pocos valores cada una, coincidirás conmigo en que nos enteramos mejor de lo que ocurre acudiendo a la tabla de contingencia.

		Número de hermanos				Total
		0	1	2	3	
Número de móviles	1	10	6	1	1	18
	2	5	17	8	2	32
	3	1	5	21	7	34
	Total	16	28	30	10	84

Tabla 4. Relación entre dos cuantitativas con una tabla de contingencia.

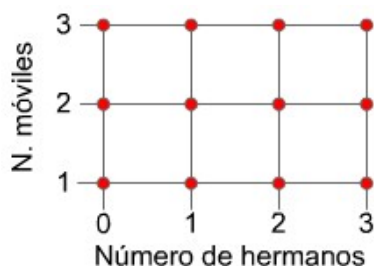


Figura 5. Diagrama de dispersión con pocos valores.

Diagrama de medias

Un diagrama de medias es como un diagrama de barras, pero donde la altura de las barras no está definida por la frecuencia con que se presenta cada valor o categoría de eje horizontal, sino por la media aritmética de otra variable. La figura 2 muestra un ejemplo de diagrama de medias. La vimos en el momento de estudiar las variables por separado. No nos dimos cuenta en ese momento, pero se trata de un estudio de relación entre dos variables: valores o aspectos importantes de la vida en relación con el grado de importancia que le dan los jóvenes de la muestra. En la figura puede observarse, por ejemplo, que el valor con mayor importancia es la familia.

Quien dice barras, dice puntos o líneas. Cualquier recurso gráfico sirve mientras respete dos principios: el origen es cero (el eje vertical comienza en cero) y la posición del punto, de la línea o del extremo superior de la barra es proporcional a la media que representa.

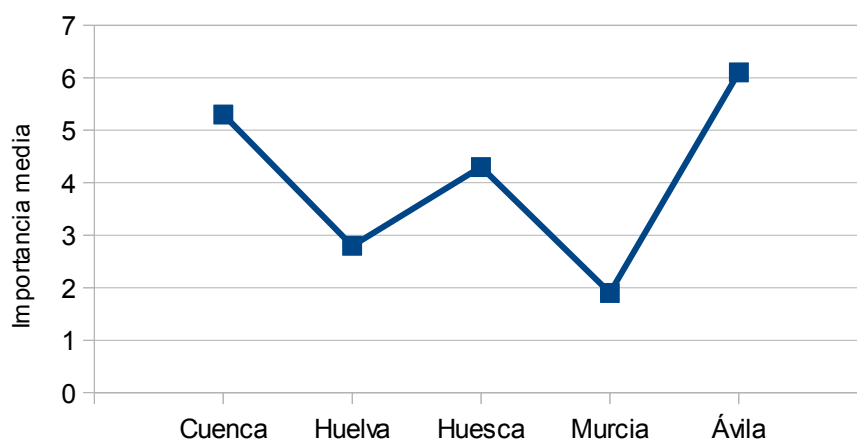


Figura 6. Diagrama de medias.

La figura 6 muestra un ejemplo, esta vez recurriendo a líneas en lugar de a barras. Hemos preguntado a un conjunto de personas provenientes de cinco provincias por la importancia que dan al festival de Eurovisión. La figura muestra la relación entre las variables *importancia otorgada al festival* y *provincia de residencia*. Se puede observar que la máxima importancia tiene lugar en Ávila, la mínima en Murcia.

Identificar problemas y solucionarlos

Del mismo modo que ocurrió con la visualización o estudio preliminar de variables por separado, así ocurre también con la visualización de relaciones: sirven también para identificar casos raros y solucionar el asunto. Los casos raros muestran dos aspectos de interés. Por un lado, pueden estar señalando errores. Cuando los errores se refieren a una variable por separado, ya habrán sido detectados y corregidos en la visualización de variables. Hay otro tipo de errores: las incoherencias.

Imagina, por ejemplo, que contamos con un cuestionario que pregunta si la persona está o no casada. Si lo está, se le pregunta desde cuándo. Si alguien no está casada, esta segunda pregunta no tiene sentido y debería aparecer vacía o con el valor 0 (u otro con ese significado). Al visualizar una tabla o una gráfica que combina ambas variables, habrá de observarse que todas las personas que no están casadas cuentan con el valor 0 en la segunda variable. En caso contrario, se trata de un error que aconsejará ir al archivo de datos y corregirlo.

El otro fenómeno de interés en la corrección de casos raros es observar que hay un caso o un pequeño grupo de casos que no se comporta del mismo modo que el grueso del conjunto de datos. Si ocurre así, está pasando algo digno de estudio, pues interesa conocer por qué ese pequeño grupo de casos se comporta de forma diferente. La mejor decisión es excluir a esos casos del análisis y estudiarlos por separado. Si se mezclan con el resto provocarán que las conclusiones finales no sirvan ni para ellos ni para el grueso de los datos.

Y ahora, antes de seguir, toma cada uno de tus objetivos de relación y realiza una visualización, redactando unas primeras conclusiones, separando casos raros, corrigiendo errores de coherencia... Termina estas tareas antes de implicarte en el siguiente paso.

Cuantificar la relación

La visualización permite conocer la relación y establecer unas primeras conclusiones. No obstante, suele ser insuficiente. Contamos con recursos estadísticos que permiten *cuantificar* la relación, es decir, expresar el grado de relación entre dos variables a través de una cantidad, de un número concreto. Al aplicar el procedimiento y obtener el número, lo siguiente es saber interpretarlo. De esta forma podemos concluir indicando que existe una relación de una cuantía determinada. El procedimiento que se aplica depende del tipo de variables implicadas. Siguiendo el mismo esquema de la tabla 3, vamos a plantear también tres procedimientos específicos de cuantificación.

V de Cramer

Cuando contamos con dos variables categóricas, la visualización se lleva a cabo mediante una tabla de frecuencias. La tabla indica cuántos casos se han contabilizado para cada una de las combinaciones posibles entre pares de categorías de ambas variables. Ya lo hemos visto en un apartado específico.

Karl Pearson ideó un procedimiento para cuantificar la relación entre dos variables categóricas a través de una tabla de contingencia. La lógica es comparar las frecuencias observadas con las que cabría esperar si ambas variables no tuvieran nada que ver entre sí, es decir, las frecuencias que cabría esperar si no hubiera ninguna relación entre ambas. Cuanto más se aleje lo observado respecto a esa situación, concluiremos que la relación es de mayor grado o cuantía. El procedimiento que ideó recibió el nombre de Ji cuadrado de Pearson. Ji por referirse a la letra griega, χ , del mismo nombre. Cuadrado, χ^2 , porque el procedimiento, cuyas tripas no voy a describir aquí, se expresa como distancias al cuadrado entre las frecuencias esperadas y observadas. Y de Pearson, por su autor.

Ocurre que χ^2 vale 0 cuando las frecuencias observadas y las esperadas coinciden y por tanto las dos variables que estamos estudiando son totalmente independientes entre sí y no existe un atisbo de relación. No obstante, χ^2 no tiene un valor máximo. Su cota superior depende del número de datos y del número de categorías de las variables en juego. Esto dificulta la interpretación, puesto que un valor por ejemplo de $\chi^2=12$ puede implicar un grado ridículo de relación o una relación máxima dependiendo de los aspectos mencionados.

Para solucionar este aspecto, Harald Cramer dedujo el valor máximo que podría alcanzar χ^2 e ideó un estadístico que se mueve entre 0 (ausencia de relación) y 1 (relación máxima). Los programas estadísticos de ordenador suelen ofrecer este valor. Por si no ocurriera así, vamos a ejemplificar el cálculo.

En el ejemplo que vimos de la tabla de contingencia entre destinos turísticos y nivel adquisitivo, $\chi^2=38,49$. No sabemos cómo masticar este resultado. Así que calculamos la V de Cramer del siguiente modo:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} = \sqrt{\frac{38,49}{186(2-1)}} = 0,45$$

El valor de $V = 0,45$ expresa una relación apreciable.

¿Cómo interpretar el grado de relación en el intervalo (0,1)?

Una buena conducta implica interpretar la cuantía literalmente: 0,45 implica una relación intermedia o de cuantía media (un 45% de lo máximo que cabría esperar). No obstante, la cosa no suele funcionar así. En muchas ocasiones se espera que concluyamos con una sentencia dicotómica: ¿existe o no existe relación? Este hábito exige transformar en un sistema de dos categorías lo que se expresa originalmente en un continuo. Una solución intermedia es basarse en algunas orientaciones establecidas por Jacob Cohen para medir lo que llamó el *tamaño del efecto*. Aunque la expresión “tamaño del efecto” suele aplicarse sobre un efecto estandarizado, aquí vamos a considerarlo sencillamente como efecto acotado en (0,1). Las orientaciones de Cohen las traduciré a:

- De 0 a 0,1: no hay relación.
- Hasta 0,3: relación pequeña.
- Hasta 0,5: relación media.
- Por encima de 0,5: relación alta.

Luego, la relación encontrada entre el nivel adquisitivo y el destino turístico preferido es de grado medio.

r de Pearson

Karl Pearson no se ciñó únicamente al caso de la Ji cuadrado. Colaboró en la idea y desarrollo del que ha pasado a la historia como el coeficiente de correlación lineal de Pearson, representado por la letra minúscula r.

La r de Pearson cuantifica la relación entre dos variables cuantitativas. Va de -1 a +1, con la siguiente interpretación:

- El valor absoluto del estadístico, es decir $|r|$, es una medida directa de tamaño de efecto: cuanto más cerca está de 0, menos relación lineal. Conforme más cerca está de 1: mayor relación.
- Conforme r se aleja de 0 y se acerca a -1, la relación es negativa o descendiente. Es el caso que vimos en la figura 4: al aumentar el valor de una variable, disminuye la otra.
- Conforme r se aleja de 0 y se acerca a 1, la relación es más positiva o ascendiente: al aumentar el valor de una variable, aumenta también la otra.

La interpretación de la cuantía de la relación sigue el mismo esquema que en el caso de la V de Cramer: relación nula ($|r| \leq 0,1$), pequeña ($0,1 < |r| \leq 0,3$), media ($0,3 < |r| \leq 0,5$) y alta ($|r| > 0,5$).

r mide “relación lineal”, es decir la que se adecúa a una línea recta. Si la relación entre ambas variables sigue otro patrón (curvilínea, exponencial...), r no es un buen estadístico. Lo que se suele hacer en tales casos es llevar a cabo transformaciones de las variables, pero tal estrategia excede los objetivos de este documento brevísimo. Como antes de calcular r ya hemos visualizado la relación a

través de un diagrama de dispersión, sabremos ante qué tipo de relación nos encontramos. Observa la figura 7. Espero que visualizar estos patrones te ayude a interpretar correctamente r .

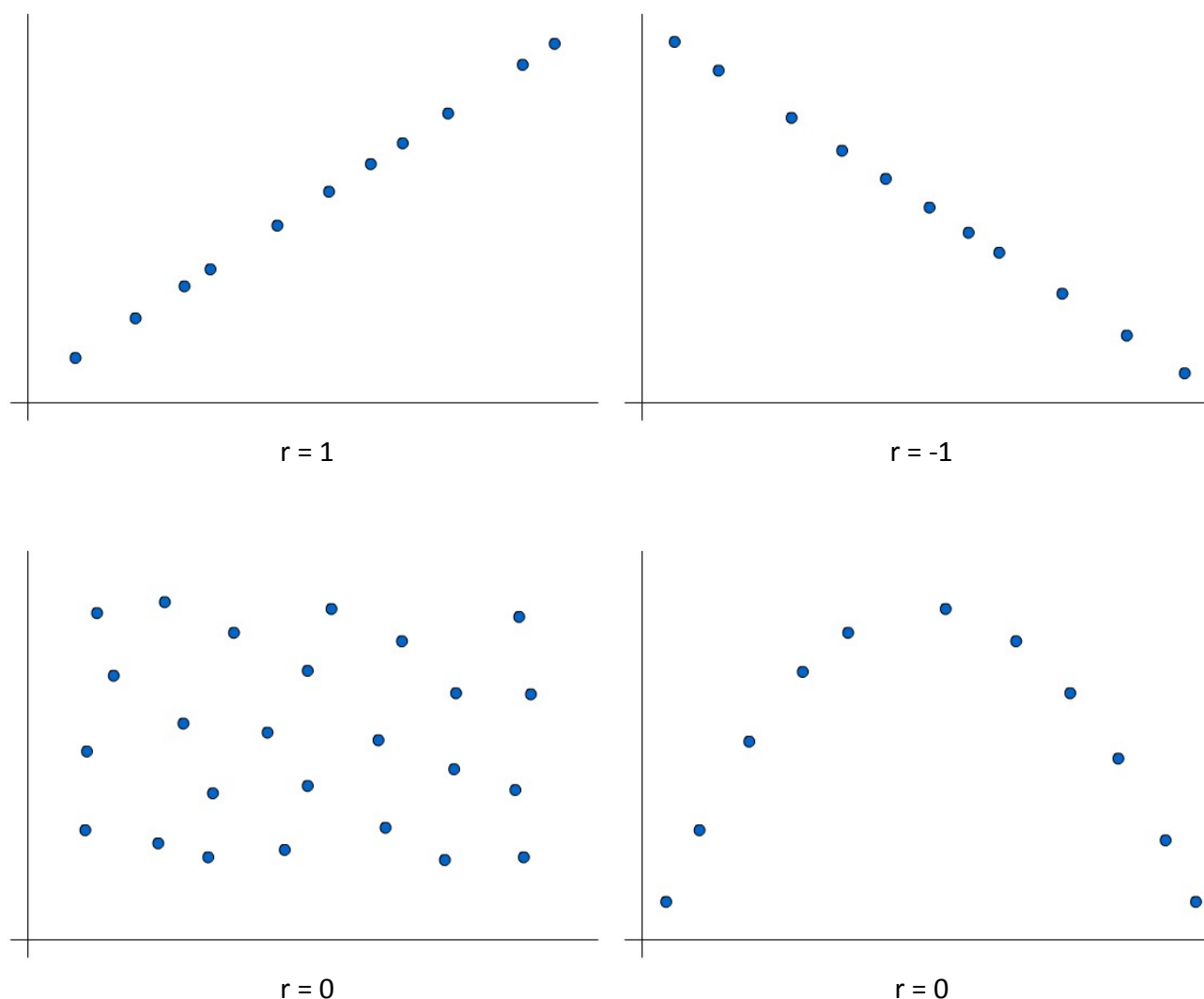


Figura 7. Ejemplos de relación con r de Pearson.

Eta cuadrado

La situación en la que contamos con una variable categórica y otra cuantitativa cuenta con un amplio abanico de posibilidades. De todas ellas vamos a recurrir a un único procedimiento, el que permite cubrir más objetivos. Fue ideado por Sir Ronald Fisher y se denomina *análisis de la varianza*.

La desviación tipo o cualquier otra medida de variabilidad para cuantías, permite expresar el grado en que una variable cuantitativa muestra dispersión entre sus valores. Medir la relación entre una variable categórica y una cuantitativa es tanto como observar el grado en que la variación de la cuantitativa disminuye al incluir la categórica.

En un estudio sobre actitudes frente a la inmigración, ideamos una variable cuantitativa a partir de la combinación de varias preguntas del cuestionario. Esta variable expresa actitud abierta ante la inmigración (interés por conocer otras culturas, religiones, hábitos, etc. y a las personas que las representan, además de considerar positiva la llegada de inmigrantes y el intercambio de perspectivas). Se expresa en el intervalo (0,10).

Al preguntar a un grupo extenso de personas, hemos observado respuestas muy diversas, una amplia variabilidad. ¿A qué es debido tanta dispersión entre las respuestas? En el cuestionario también hemos incluido algunas preguntas sobre el nivel de estudios, considerando tres

posibilidades: bajo, medio y alto. Se trata de una variable ordinal que a los efectos de estudiar la relación vamos a tratar como categórica. La representación gráfica (un diagrama de medias) consta en la figura 8.

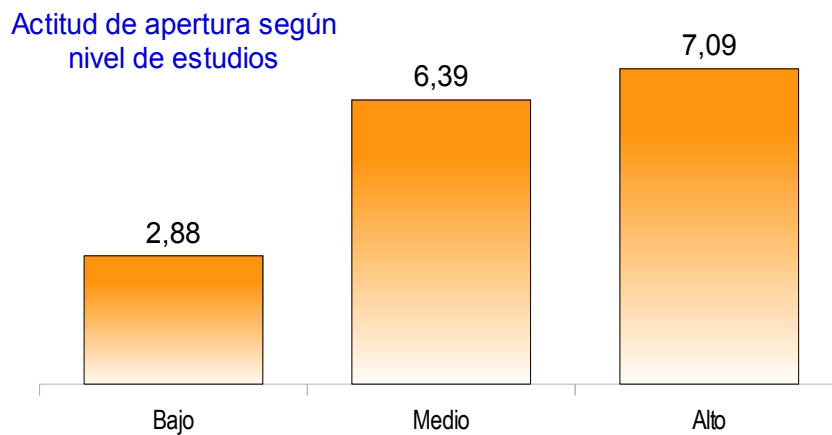


Figura 8. Diagrama de medias para relación entre actitud y estudios.

El diagrama de medias es muy ilustrativo. Puede observarse que el nivel bajo de estudios cuenta con una actitud menos abierta ante la inmigración, frente a valores altos de actitud en los otros dos casos. Se puede afirmar con claridad que existe *variabilidad* entre los tres grupos de personas respecto a la variable cuantitativa *actitud de apertura*. Esto implica que parte de la variabilidad de esta variable cuantitativa queda recogida por las diferencias entre los tres grupos.

La lógica del análisis de la varianza consiste en comparar la variabilidad que existe entre los grupos (variación *entre*) con la variabilidad que queda dentro de cada uno de los grupos (variación *dentro*). Existe tanto más relación entre ambas variables cuanto la variabilidad resida más *entre* que *dentro*. Es bastante lógico. Si no existiera relación, entonces no importaría qué grupo observar, pues todos contarían con el mismo valor medio y toda la varianza observada en los datos se encontraría dentro de los grupos. En el caso opuesto, de máxima relación, toda la variabilidad existiría entre los grupos (mostrando medias muy diferentes), mientras que dentro de cada grupo todos los datos coincidirían en un mismo valor.

El recurso que se utiliza para medir la variabilidad es la *suma de cuadrados*. Consiste en calcular las distancias al cuadrado que separa a cada dato respecto a la media aritmética del conjunto de datos. La tabla 5 muestra los cálculos intermedios para el ejemplo en el que nos encontramos.

La fuente de variación se refiere a las variaciones entre y dentro. La suma de ambas da lugar a la variación total. La suma de cuadrados es esa medida de dispersión entre los valores dentro de la fuente de variación, ya descrita. Los grados de libertad expresan las posibilidades de variación. Conforme más grados de libertad exista, es más fácil obtener grandes sumas de cuadrados. La media cuadrática expresa la suma de cuadrados en función de los grados de libertad. La F de Fisher es el cociente entre las medias cuadráticas. Si su valor es 1 implica que numerador y denominador coinciden, es decir, que la variabilidad media coinciden en las fuentes de variación *entre* y *dentro*. Si es inferior a 1, expresa que hay más variabilidad dentro de los grupos que entre ellos. Y si es superior a 1, tenemos que la variable categórica funciona para explicar la variabilidad, pues consigue ser mayor que la que todavía queda dentro de los grupos. Para concluir relación buscamos entonces encontrar valores de F que sean superiores a 1.

Con los datos del ejemplo, vemos que F es muy elevada (33,77). No obstante, se trata de una cuantía con cota inferior (0, que ocurre cuando las medias de todos los grupos coinciden), pero no con un máximo superior, del mismo modo que ocurre con la Ji cuadrado de Pearson. Por este motivo acudimos a otro estadístico: eta cuadrado, representado por su letra griega homónima. Consiste en expresar la variabilidad (suma de cuadrados) entre los grupos respecto a la total. $\eta^2=0$ indica que no

hay variabilidad entre los grupos y por tanto existe una ausencia absoluta de relación. En el caso opuesto, $\eta^2=1$ y representa la situación en que coinciden la variabilidad total con la entre grupos, es decir, ocurre cuando toda la variabilidad se encuentra entre los grupos y no existe ninguna variación dentro de ellos. Como puede suponerse, es una situación muy difícil de esperar en la práctica.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	F	η^2
Entre	182,73	2	$182,73 / 2 = 91,37$	$91,37 / 2,17 = 33,77$	$182,73 / 336,96 = 0,54$
Dentro	154,23	57	$154,23 / 57 = 2,17$		
Total	336,96	59			

Tabla 5. Cuadro del análisis de la varianza.

η^2 se interpreta del mismo modo que V y $|r|$. Con los datos del ejemplo, se observa un efecto grande o una relación alta.

Los programas estadísticos de ordenador realizan análisis de la varianza. No todos incluyen el cálculo de η^2 , pero sí los valores de sumas de cuadrados según la fuente de variación, por lo que el cálculo manual de η^2 es inmediato.

Antes de continuar con el último apartado, cubre todos tus objetivos de relación e interpreta el resultado en términos de cuantía de efecto. Cuando hayas finalizado esa tarea, ya puedes culminar este aprendizaje, aperitivo o repaso con el paso de la inferencia.

Realizar inferencia, si procede

Cuanto hemos visto hasta el momento se refiere al conjunto de datos con el que estamos trabajando. En muchas ocasiones, realizamos los cálculos con una muestra pero nos interesa concluir a nivel de la población. En tales casos, se realizan procesos de inferencia estadística.

La inferencia estadística cuenta con muchas posibilidades, recursos y procedimientos. Para el estudio de relaciones vamos a acudir a uno solo de ellos: la prueba de significación de la hipótesis nula (PSHN). Antes de ponerla en marcha es importante considerar dos condiciones. La primera es que la muestra ha debido de obtenerse según un procedimiento aleatorio. Si la muestra no es aleatoria, en sentido estricto no se puede aplicar inferencia estadística. Es así porque la inferencia estadística acude a modelos de probabilidad para definir las conclusiones. La otra condición es haber encontrado un tamaño de efecto o cuantía de relación suficiente. Vamos a considerar como suficiente una relación al menos pequeña (superior a 0,1). Personalmente defiendo que es una cuantía demasiado pequeña, muy ridícula, que no significa nada en la práctica. Pero el hábito extendido es el de no atender siquiera al tamaño de efecto, por lo que no voy a ponerme excesivamente escrupuloso con esto.

Ocurre que cuando se manejan muestras de mucho tamaño, la PSHN puede llevar a una situación paradójica, ayudando a concluir que hay relación cuando nuestros ojos están viendo que no la hay. Esto lo vamos a ver más despacio, pero no con exceso, ya que este documento no tiene por objetivo sumergirse en demasiadas profundidades.

La lógica de una PSHN, debida al mismo Ronald Fisher que ideó el análisis de la varianza, es un tanto enrevesada. No puede ser de otro modo, como ya veremos. Para comprenderla, voy a basarme en la misma lógica pero procedente de otro contexto: el manejo de la culpabilidad y la inocencia en un estado de derecho.

Estado de derecho

Sepas más o menos de leyes, seguro que has visto películas donde aparece un juicio, alguien sentado en el banquillo de los acusados, con fiscal, jueza, abogado, jurado... Con esa impronta, seguro que reconoces la lógica del proceso que voy a presentar de forma esquemática.

0. En la paz de las situaciones, se considera que ha pasado algo y que hay una persona que podría ser la causante. En las películas y series más extendidas, que provienen del mismo lugar del planeta, esta perturbación de la paz suele deberse a un asesinato. Y se sospecha de alguien, que vamos a denominar G, como asesino de otro alguien, que denominamos T.
1. Hipótesis o presunción de inocencia: G es inocente mientras no se demuestre lo contrario.
2. Datos o evidencias: se recogen pruebas sobre el móvil, dónde se encontraba la noche del crimen, testimonios de personas del entorno de T un vagabundo que pasaba por allí...
3. Decisión. Son posibles dos desenlaces:
 - a) La acumulación de pruebas puede ser suficiente como para que el jurado considere que la hipótesis de inocencia no se mantiene y rechaza la hipótesis, decidiendo por tanto que el acusado es culpable.
 - b) Las pruebas no son concluyentes, o hay evidencias para todos los gustos. En cualquier caso, los datos no tienen la fuerza necesaria como para rechazar la hipótesis de inocencia y, por tanto, se mantiene.
4. Conclusión. Respectivamente:
 - a) G es culpable del asesinato de T.
 - b) G es inocente del asesinato de T.

Lógica y fases en una PSHN

Es normal que las personas que acceden por vez primera a la lógica de una PSHN sientan que el proceso es muy enrevesado. No obstante es literalmente el mismo que el visto en el punto anterior. Vamos a reproducirlo:

0. En la paz de las situaciones, se tiene la sospecha de que dos variables están relacionadas, es decir, que cuando varía una parece variar también la otra y en un sentido concreto.
1. Hipótesis nula (H_0): las dos variables son independientes entre sí (no están relacionadas) mientras no se demuestre lo contrario.
2. Datos o evidencias: se obtiene una muestra aleatoria de la población de interés. Se realizan en ella los cálculos para medir la relación entre ambas variables, obteniendo resultados.
3. Decisión. Son posibles dos desenlaces:
 - a) Los resultados son incompatibles con H_0 . Si hay que escoger entre una hipótesis previa o los datos recogidos bajo control, creemos a los datos y se rechaza H_0 .
 - b) Los datos son compatibles con H_0 , por lo que se mantiene.
4. Conclusión. Respectivamente:
 - a) Existe relación entre ambas variables.
 - b) Las dos variables son independientes.

Además de comprender la lógica de la PSHN (repasa el esquema anterior hasta que lo comprendas), es necesario comprender correctamente qué es eso de la compatibilidad entre la hipótesis y los datos.

La hipótesis nula se refiere siempre a la población, no a la muestra. De una población donde dos variables no están relacionadas entre sí es posible obtener muestras aleatorias donde exista una relación no nula. Así que H_0 afirma que, encontremos lo que encontremos en la muestra, en la población las dos variables no están relacionadas entre sí.

Esta afirmación general debe concretarse en términos estadísticos. Para el caso de relación entre dos variables categóricas, diremos que $\chi^2 = 0$ en la población. Para el caso de dos variables cuantitativas, $\rho = 0$ (se utiliza la letra griega ρ para expresar la r a nivel poblacional). Para una cuantitativa con una categórica, $F \leq 1$.

Identificar la compatibilidad

Hay varios métodos para pronunciarse acerca de la compatibilidad entre la H_0 y los datos de la muestra. El que voy a utilizar aquí es el más extendido, presente en los programas de análisis de datos habituales. Se basa en probabilidades.

Podemos calcular la probabilidad de que, siendo cierta la H_0 , es decir siendo cierto que no hay relación en la población, obtengamos una muestra al azar con resultados como los que hemos obtenido en nuestro caso. Esa probabilidad recibe varios nombres: grado de significación, valor p , p -value,... Pues bien, si esa probabilidad es pequeña, entonces la decisión es rechazar la hipótesis nula, ya que es muy difícil encontrar la relación obtenida en la muestra suponiendo que no existe tal relación en la población. La conclusión será que existe relación. Pero si esa probabilidad es alta, entonces los resultados de la muestra son compatibles con la hipótesis nula y por tanto no puede ser rechazada (se mantiene). La conclusión será que no existe relación.

Bien, ya sabemos que H_0 y datos son compatibles o incompatibles si, respectivamente, el grado de significación es alto o bajo. Ahora tenemos otro problema: ¿cómo identificar si el grado de significación es alto o bajo?

Para identificarlo, utilizamos un umbral, un valor teórico que marca el límite entre lo bajo y lo alto. Si no se llega al umbral, entonces es bajo. Si se supera, entonces es alto. Ese umbral recibe el nombre de nivel de significación, error de primera especie o error tipo I, y se simboliza con la letra griega alfa, α . Así pues:

- Si $p < \alpha$, la probabilidad de que los resultados que hemos obtenidos provengan de una población donde no existe relación, es baja. Luego, datos e hipótesis nula son incompatibles entre sí: o los datos dicen lo cierto o lo cierto es lo que dice la hipótesis nula. Como creemos a nuestros datos, se rechaza la hipótesis nula y concluimos que existe relación. También se dice con frecuencia que existe una relación *estadísticamente significativa*.
- Si $p \geq \alpha$, la probabilidad de que los resultados que hemos obtenidos provengan de una población donde no existe relación, es alta. Luego, datos e hipótesis nula son compatibles entre sí: los datos no contradicen la hipótesis nula, por lo que se mantiene. Conclusión: no existe relación.

Y, por último, ¿qué valor tiene α ?

No tiene un valor fijo. Lo ideal sería que pensáramos en las consecuencias de nuestras decisiones operando con PSHN. Si rechazar la hipótesis nula tiene consecuencias graves, exigiremos un valor de α muy bajo. Si las consecuencias son despreciables, α podrá ser alto. Sin embargo, no suele hacerse esta reflexión.

Ronald Fisher jugó con su imaginación e inventó a una viejecita que decía saber si en una taza de té con leche se había vertido antes el té o la leche. Fisher ideó un experimento con tazas y órdenes diferentes de vertido e imaginó determinadas situaciones. Finalmente consideró que una combinación determinada de aciertos le daba suficiente seguridad para concluir que la viejecita decía la verdad. Esa combinación se traduce en un valor de probabilidad del 5%. Desde entonces, el hábito está en considerar $\alpha = 0,05$ (o, en tantos por ciento, $\alpha = 5\%$) como umbral para decidir si un grado de significación es bajo o alto, siguiendo el esquema anterior.

PSHN para dos categóricas

En el ejemplo que vimos para estudiar la relación entre el nivel adquisitivo y el destino turístico preferido, el valor de J_i cuadrado es 38,49. Según la hipótesis nula, $\chi^2 = 0$ en la población. Necesitamos traducir ese valor en un grado de significación para tomar la decisión. Al usar un programa de ordenador, ya suministra el valor p que, en este caso es 0,000 (0 redondeado a tres dígitos decimales). Dado que $p < \alpha$, se rechaza la hipótesis nula y concluimos que el nivel adquisitivo y el destino turístico están relacionados entre sí.

Si hemos hecho los cálculos a mano o con una hoja de cálculo, podemos consultar algunas tablas que traducen los valores de Ji cuadrado a grados de significación, o podemos acudir a ejecutar una función de la hoja de cálculo. Si utilizas LibreOffice Calc, la función es **1-chisqdist(ji;(f-1)*(c-1);1)**, donde *ji* es el valor de la ji cuadrado, *f* es el número de filas de la tabla (número de categorías de la variable en filas), y *c* es el número de columnas (número de categorías de la variable en columnas). Si utilizas Microsoft Excel, la función es **distr.chiquad**. En la situación del ejemplo, la función es **1-chisqdist(38,49;4;1)** o bien **1-distr.chiquad(38,49;4;1)**, cuyo resultado es 0,00000002229940732068.

PSHN para dos cuantitativas

Pongamos que estudiamos la relación entre el número de veces que una persona ha ido al cine el último año, con el grado de acuerdo con la afirmación “Me gusta mucho el cine”, medido en un intervalo de 0 a 10. El resultado es $r = 0,35$ al realizar ambas preguntas a una muestra aleatoria de $n = 27$ personas. ¿Existe relación entre ambas variables?

Hemos realizado un análisis de cada una de ellas por separado y obtenido un diagrama de dispersión. Al cuantificar obtenemos $r = 0,35$ que nos parece una cuantía aceptable. Dado que el tamaño del efecto es suficiente y que la muestra es aleatoria, ponemos en marcha una PSHN para ver si podemos concluir que existe relación en términos generales en la población.

El programa de ordenador indica un grado de significación de valor 0,074. Se trata de un riesgo alto, puesto que $p = 0,074 > 0,05 = \alpha$. Luego, se mantiene la hipótesis nula y se concluye que no existe relación entre el grado de acuerdo con la afirmación “Me gusta mucho el cine” y el número de veces que se ha ido al cine durante el año anterior.

En LibreOffice Calc, la función a utilizar es **distr.t(t;n-2;2)**, mientras que en Microsoft Excel es **distr.t.2c(t;n-2)**, donde $t = \sqrt{\frac{r^2 (n-2)}{1 - r^2}}$. En el ejemplo: **distr.t(1,8682;25;2) = 0,073503516729599**.

PSHN para una categórica con una cuantitativa

La hipótesis nula establece que incluir la variable nivel de estudios no reduce la variabilidad observada en la actitud de apertura, pues no existe relación entre ambas variables. Los cálculos indican 2 grados de libertad entre-grupos, 57 dentro-grupos y un valor de $F = 33,77$. El programa de ordenador suministra un grado de significación de valor $p = 0,000$ redondeado a tres cifras decimales. Luego, dado que $p < \alpha$, se rechaza la hipótesis nula y concluimos que el nivel de estudios y la actitud de apertura son dos variables relacionadas entre sí.

En LibreOffice Calc, la función a utilizar es **distr.f(F;gl entre;gl dentro)**, mientras que en Microsoft Excel es **distr.f.cd(F;gl entre;gl dentro)**. En el ejemplo, **distr.f(33,77;2;57) = 0,000000000211899941680096**.

Vocabulario mínimo

<i>Constante</i>	Característica que se mantiene sin variar en el contexto del estudio. Si la investigación versa sobre las preferencias turísticas de la población andaluza, las preferencias son variables, pero la población andaluza es constante, pues en el contexto del estudio no se aborda otro tipo de población.
<i>Escala de medida</i>	Tipo de variable según la operaciones que pueden realizarse con ella. Las escalas de medida de las variables se estudian con taxonomías diferentes. Una de ellas considera variables nominales, ordinales y cuantitativas.

- Estadístico* Es un tipo de operación matemática que genera un número. Ejemplo de estadísticos son media, proporción, porcentaje, varianza, desviación tipo, amplitud... En sentido estricto, se denomina estadístico a estas operaciones realizadas en una muestra, mientras que se denomina *parámetro* a lo mismo pero calculado en una población. Sin embargo, es habitual hablar de estadístico para referirse a ambos. Es una especie de comportamiento sexista aplicado a la estadística: muchas personas hablan de mujer para referirse a las mujeres, mientras que utilizan el término *hombre* para referirse tanto a los hombres como al conjunto de hombres y mujeres.
- Inferencia* Inferir es concluir algo en un nivel a partir de resultados obtenidos a otro nivel. La inferencia estadística consiste en establecer conclusiones basadas en probabilidades, referidas a una población pero realizando los cálculos en una muestra aleatoria de esa población.
- Muestra* Porción de una población, en la que se realizan los cálculos de estadísticos. La estadística se ocupa de las muestras aleatorias, es decir, de las porciones que han sido obtenidas al azar.
- Población* Contexto al que se refieren o desean referirse los resultados del estudio. Tal vez se haya trabajado con la población directamente o quizá se ha acudido a trabajar con una muestra, aunque los resultados se orientan para ser aplicados a la población.
- Relación* Dos variables están relacionadas entre sí cuando observamos que al variar una de ellas también varía la otra y en un sentido concreto.
- Valor* Expresión concreta de una variable. “Córdoba” es uno de los valores de la variable “Provincia de nacimiento”. “32” es uno de los valores de la variable “edad”.
- Variable* Característica que varía, es decir, que se expresa con un valor concreto de entre un conjunto de valores posibles. El color del cabello es una variable, cuyos valores posibles son moreno, rubio, castaño, pelirrojo, plateado...