Técnicas cuantitativas para la Administración pública

PID_00267225

Camilo Cristancho Mantilla

Tiempo mínimo de dedicación recomendado: 5 horas

Universitat Oberta de Catalunya

Camilo Cristancho Mantilla

Investigador post-doctoral en el departamento de Ciencias Políticas de la Universidad de Barcelona. Es miembro del grupo de investigación Calidad de la Democracia y su investigación más reciente se centra en desigualdades y procesos de influencia política. Su trabajo se basa en métodos de ciencia social computacional, experimentos y análisis estadístico.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Daniel Rajmil (2019)

Primera edición: septiembre 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción

La Administración pública se enfrenta a problemas y desafíos de enormes proporciones y complejidad, como los déficits de vivienda, la pobreza, la inseguridad o el paro, entre mucho otros. Esto implica la necesidad de comprender muy bien los problemas para poder aprovechar de la mejor forma los fondos públicos y buscar las mejores soluciones. Generalmente, la observación de estas situaciones implica comprender a nivel agregado el comportamiento o las circunstancias de los individuos, las familias o los grupos de interés, como las escuelas o los barrios. También comprende la necesidad de evaluar cómo cambian estos comportamientos y circunstancias ante determinadas intervenciones sociales, como los programas de gobierno, o ante los cambios institucionales propiciados por las leyes y las regulaciones.

El gestor público necesita contar con las herramientas adecuadas para afrontar estos retos en la medida en que es responsable de proveer las mejores soluciones y de garantizar además la eficiencia de políticas financiadas con impuestos. Afortunadamente, la estadística provee las herramientas para interpretar de una manera rigurosa estas situaciones complejas. Por una parte, el uso de estadísticas permite realizar la observación y el seguimiento de un gran número de características o atributos de los sujetos de estudio (individuos, familias, colectivos, etc.). Para ello, se crean variables a partir de los conceptos de interés que miden rigurosamente los atributos. Por otra parte, las estadísticas permiten interpretar las variables y evaluarlas a nivel agregado para comprender fenómenos colectivos y patrones temporales. De esta manera, es posible determinar las necesidades del gasto social, crear mediciones para definir criterios de asignación de recursos y calcular el desempeño de las inversiones sociales. Este tipo de tareas implica tener la capacidad de examinar una gran cantidad de variables simultáneamente para ordenar y dar sentido a situaciones sociales complejas.

Adicionalmente, el gestor público tiene la responsabilidad de seguir y evaluar la evidencia disponible sobre otras intervenciones públicas. Esto implica comprender la evidencia o las afirmaciones en la investigación académica o aplicada que se publica en informes oficiales, evaluaciones, comunicados de prensa y artículos académicos. Por ello es necesario desarrollar habilidades para evaluar las conclusiones empíricas y la validez de los métodos de investigación. Esto es especialmente relevante en las evaluaciones de impacto que proporcionan información sobre los efectos que un programa puede tener sobre la población beneficiaria. Los métodos para conocer si dichos efectos son atribuibles al programa suelen ser bastante complejos, pero esta guía proporciona

las bases para aproximarse a este tipo de trabajos estadísticos. Además, tiene como objetivo desarrollar las habilidades básicas para realizar análisis de gran utilidad, que son la base de estudios más sofisticados.

Esta guía pretende introducir técnicas de análisis estadístico de una forma aplicada a los problemas de la gestión pública y de la manera más práctica posible. Esto significa que se ha tratado de simplificar al máximo las nociones matemáticas y teóricas con el fin de presentar los métodos aplicados utilizando datos reales o ficticios para posibilitar cálculos simples.

La guía se divide en cinco secciones. En la primera, se introducen conceptos de medición, como los tipos de variables y la gestión de los datos. En la segunda sección, se presentan conceptos y técnicas de análisis descriptivo para una sola variable. En la tercera sección, se trata el tema de las relaciones entre variables categóricas; en este punto, se introduce el tema del diseño de investigación y la prueba de hipótesis. En la cuarta sección, se introducen las técnicas de análisis de relaciones entre variables continuas, y en la quinta y última sección, se expone el modelo de regresión. Varias técnicas de análisis para casos particulares se han omitido por cuestiones de simplicidad o espacio, pero se sugieren algunas fuentes complementarias para ampliar los temas.

1. Medición

Las actividades de planeación y control propias de la Administración pública requieren observar y monitorear los recursos y los resultados de la gestión. Esto implica comprender los fenómenos de interés mediante representaciones numéricas que permitan medirlos sistemáticamente. Una **medida** es una asignación de números (u operacionalización) a un fenómeno que nos interesa analizar.

La **medición** es la base del análisis estadístico en cuanto permite transformar conceptos tales como la salud pública, el logro educativo o la eficiencia recaudatoria en cantidades específicas que pueden evaluarse mediante la comparación.

1.1. Tipos de variables

Una **variable** es una característica o condición que puede cambiar o asumir diferentes valores. Un **valor** es una cantidad específica que es posible para una variable. Cuando observamos un fenómeno de interés nos centramos en una característica que puede variar de valor entre los sujetos o en el tiempo. Empezaremos por definir tres tipos de variables: cuantitativas, categóricas y ordinales.

Las **variables cuantitativas** toman valores numéricos. Hay aspectos que pueden medirse a partir de cantidades observables, como los presupuestos o los gastos, los cuales se representan en cantidades de dinero. Las variables cuantitativas pueden tener **valores discretos** o **continuos**. En el caso de los discretos, solo pueden tomar valores numéricos específicos (por ejemplo, edad, dinero, número de personas en un programa, etc.). En el caso de los valores continuos, la variable puede tomar cualquier valor numérico (por ejemplo, la altura, el peso, el porcentaje de ejecución de un proyecto, etc.).

Las **variables categóricas** se utilizan para describir fenómenos de interés que no pueden medirse en cantidades numéricas o en el caso de que nos interese tan solo comprenderlos en términos de categorías. Por ejemplo, para conocer las características de la población suele ser interesante medir la composición de grupos de acuerdo con atributos particulares, como el género, la raza o la religión. A pesar de que estas variables no tienen escalas de medición cuantitativas, es posible medirlas en términos de categorías; por ejemplo: femenino *versus* masculino; blanco *versus* no blanco; católicos, protestantes, judíos, musulmanes y otros. El gestor público necesita saber cómo medir, describir y analizar estos grupos estadísticamente en cuanto determinan cantidades de inte-

rés como, por ejemplo, el efecto diferencial de las políticas sobre grupos poblacionales. Estas variables también se conocen como **variables cualitativas**. Para definir las categorías se utilizan diferentes criterios:

- Criterio único, como usuario *versus* no usuario.
- Más de un criterio. Taxonomías o tipologías, es decir, conjuntos de categorías que pueden ser mutuamente excluyentes (por ejemplo, bajo la línea de pobreza, en riesgo de pobreza, fuera de riesgo) o colectivamente exhaustivas (por ejemplo, estudia, estudia y trabaja, trabaja, jubilada).

Las **variables ordinales** representan categorías que están ordenadas. Un ejemplo de esto serían los grados de educación (EGB, ESO, educación secundaria no obligatoria, formación profesional o ciclo formativo, diplomatura universitaria o 1.r ciclo, licenciatura o grado universitario, máster, posgrado o doctorado). La medición ordinal permite clasificar objetos o eventos por categorías y ordenarlos por grados. Se puede asociar un número a cada caso, y ese número no solo indica la categoría a la que pertenece, sino cómo se relaciona con las demás categorías.

1.2. Datos

Un elemento central del análisis estadístico es la **gestión de la información**. ¿Cómo debemos almacenar nuestras observaciones para poder analizar la información? ¿Cuál es la mejor forma de registrar las variables que hemos medido?

> Lo más importante es tener claro qué medimos y de dónde hemos tomado este dato. Para ello, debemos tener muy claro cuál es nuestra **unidad de análisis**.

Podemos hacer observaciones de individuos, parejas, familias, grupos, organizaciones, etc., y de cada una de estas unidades de análisis podemos medir diferentes atributos (variables). La forma más sencilla de organizar nuestras observaciones es mediante la construcción de una **tabla** o **matriz de datos**, que es un conjunto organizado de datos en una estructura tabular (similar a una hoja de cálculo) y constituye el elemento fundamental de una base de datos. Consiste en organizar la información en una cuadrícula en la cual las observaciones o registros son cada una de las filas en las que se divide la tabla, y las variables o campos son cada una de sus columnas. Con esta organización, es posible registrar cada valor de una variable para una observación en cada celda o intersección entre fila y columna.

Es importante tener en cuenta que cada observación siempre debe tener un **identificador único** (un número o código que represente cada individuo, pareja, familia, organización o unidad de análisis utilizada). Esto es indispensable

para identificarlo correctamente, dado que puede estar registrado de múltiples formas (por ejemplo, con los nombres escritos de manera diferente –mayúsculas, segundo nombre, abreviaciones, etc.– en la misma o en diferentes tablas.

En la tabla 1 se muestra un fragmento del estudio del CIS «3242. Macrobarómetro de marzo de 2019». (http://cis.es/cis/opencm/ES/1_encuestas/estudios/ver.jsp?estudio=1444). Se trata de una encuesta realizada a 16.194 personas que representan a la población con derecho a voto en las elecciones generales y que reside en España. Podemos observar la estructura de matriz de datos en la que cada fila corresponde a un individuo entrevistado y cada columna a una variable. La primera columna es una variable que identifica a cada individuo con un número único de cuestionario. Las siguientes columnas contienen información de interés sobre los atributos del individuo, su percepción sobre el principal problema en España y sobre su intención de voto en las elecciones generales de 2019. Sin embargo, podemos notar que esta información necesita una interpretación para conocer los nombres de las variables y las etiquetas de cada categoría. Para ello, es necesario contar con información adicional que nos permita interpretar los datos y transformarlos para su análisis. Más específicamente, necesitamos conocer los nombres, los tipos y las definiciones de las variables, las unidades con las que se miden o las escalas en las que se clasifican y los nombres o las etiquetas para cada categoría. Esta información se guarda en un documento adicional de **metadatos** (datos sobre los datos) que contiene información sobre las variables y las preguntas o fuentes utilizadas. Usando este documento (http://cis.es/cis/export/sites/default/-Archivos/Marginales/3240_3259/3242/cues3242.pdf podemos procesar la información de la tabla 1 y obtener los datos tal como se muestran en la tabla 2.

Tabla 1. Tabla de datos sin procesar.

CUES	P23	P22	P25	Estudios	P601	P10
1	81	1	2	6	13	94
2	74	2	3	2	1	1
3	43	2	1	6	1	97
4	58	2	1	5	19	2
5	65	2	2	5	1	94
6	57	1	1	6	1	2
7	47	1	1	6	8	12
8	64	2	2	6	8	1
9	68	1	2	2	1	11
10	53	2	4	4	1	11

Fragmento del estudio 3242 del CIS *Macrobarómetro*.

En la tabla 2 podemos ver claramente qué variables tenemos a nuestra disposición y los valores que toma cada variable para cada uno de los diez individuos del fragmento del estudio. Debemos observar que el hecho de que las categorías se representen mediante valores numéricos o mediante sus etiquetas o nombres no cambia el tipo de variable. El uso de las etiquetas facilita el análisis, mientras que el uso de códigos numéricos se utiliza para optimizar el tamaño de los archivos de datos. Los programas estadísticos tratan de diferentes maneras los datos almacenados, sean códigos o etiquetas o los dos tipos de información.

Tabla 2. Tabla de datos procesados.

Id	Edad	Sexo	Situacion_laboral	Estudios	Principal_problema	Intencion_voto
1	81	Hombre	Jubilado/a o pensionista (anteriormente ha trabajado)	Superiores	Los/as políticos/as en general, los partidos y la política	No lo tiene decidido aún
2	74	Mujer	Pensionista (anteriormente no ha trabajado, sus labores, etc.	Primaria	El paro	PP
3	43	Mujer	Trabaja	Superiores	El paro	No votará
4	58	Mujer	Trabaja	F.P.	La violencia contra la mujer	PSOE
5	65	Mujer	Jubilado/a o pensionista (anteriormente ha trabajado)	F.P.	El paro	No lo tiene decidido aún
6	57	Hombre	Trabaja	Superiores	El paro	PSOE
7	47	Hombre	Trabaja	Superiores	Los problemas de índole económica	EH Bildu
8	64	Mujer	Jubilado/a o pensionista (anteriormente ha trabajado)	Superiores	Los problemas de índole económica	PP
9	68	Hombre	Jubilado/a o pensionista (anteriormente ha trabajado)	Primaria	El paro	EAJ-PNV
10	53	Mujer	Parado/a y ha trabajado antes	Secundaria 2ª etapa	El paro	EAJ-PNV

Fragmento del estudio 3242 del CIS *Macrobarómetro*.

La «Encuesta continua de hogares» del Instituto Nacional de Estadística (INE) proporciona un ejemplo de documento de metodología con toda la información necesaria para interpretar los datos de las encuestas. Utilizaremos esta encuesta en el siguiente apartado, por lo que es recomendable descargar los datos y los documentos del estudio de 2018 en la página del INE (http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176952&menu=resultados&secc=1254736

195203&idp=125473557298). El fichero de datos en formato de valores separados por comas (.csv) puede importarse desde hojas de cálculo de Microsoft Excel, LibreOffice Calc o Google.

2. Análisis descriptivo univariado

2.1. La descripción estadística y la inferencia

La **estadística descriptiva** es un conjunto de métodos de organización y resumen de los datos. Su principal función es describir las propiedades agregadas de un conjunto de datos.

De esta manera, es posible comprender un fenómeno a partir de la agregación de observaciones individuales, resumiendo adecuadamente las diversas características de ese conjunto. En este punto, es crucial notar la diferencia entre la población que quiere estudiarse y el conjunto de observaciones que es posible hacer sobre una parte de esta población. Un valor descriptivo para una población se llama **parámetro**, y un valor descriptivo para una muestra es una **estadística**.

Las **estadísticas inferenciales** son métodos para utilizar la muestra con el fin de sacar conclusiones generales (inferencias) sobre la población. El objetivo es utilizar estadísticas de la muestra para hacer inferencias sobre parámetros poblacionales.

Dado que una muestra es típicamente una parte de la población, los datos de la muestra proporcionan información limitada sobre esa población. Por esta razón, las estadísticas de la muestra son generalmente representaciones imperfectas de los parámetros poblacionales correspondientes. Por ejemplo, un parámetro es el porcentaje de adultos que viven en pareja sin hijos en España, y la estadística es el porcentaje de 100.000 adultos que viven en pareja sin hijos de la «Encuesta continua de hogares» del INE.

Las diferencias que existen, por casualidad, entre la muestra estadística y el parámetro de población se conocen como **error de muestreo**. El error de muestreo de una estadística es igual a la diferencia que existe cuando usamos una estadística de muestra para predecir el valor de un parámetro de población. Definir y medir el error de muestreo es una parte importante de la estadística inferencial.

La diferencia entre la estadística inferencial y la estadística descriptiva es el uso de un **modelo de probabilidad**. Un modelo es una descripción (matemática) de las conexiones entre las variables de interés. Se trata de una simplificación

de la realidad, y por lo tanto nunca es «correcto» o «falso», pero puede ser más o menos útil. La inferencia estadística y algunas nociones básicas sobre la teoría de la probabilidad se presentan en los siguientes apartados.

La estadística descriptiva nos permite resumir los datos con tablas y gráficos (tanto para variables cuantitativas como categóricas). Las **descripciones numéricas** informan sobre las tendencias, la variabilidad y la posición (para variables cuantitativas). Las **descripciones bivariadas** nos informan sobre cómo se relacionan dos variables (para variables cuantitativas o categóricas).

En las siguientes secciones veremos cómo reconocer diferentes tipos de datos, producir una variedad de medidas estadísticas, interpretar estas medidas estadísticas básicas y saber qué medidas son adecuadas para qué tipos de datos.

2.2. Distribuciones de frecuencia

Después de recopilar datos, la primera tarea para un investigador es organizar y simplificar estos datos para que sea posible obtener una visión general de los resultados. Este es el objetivo de las técnicas estadísticas descriptivas. El primer método para simplificar y organizar los datos es construir una distribución de frecuencia.

> Una **distribución de frecuencia** es una tabulación organizada que muestra exactamente cuántos individuos están ubicados en cada categoría en la escala de medición. Presenta una imagen organizada de todo el conjunto de valores y muestra dónde se encuentra cada individuo en relación con otros en la distribución.

Una tabla de distribución de frecuencia consta de al menos dos columnas: una que enumera las categorías en la escala de medición (X) y otra para la frecuencia (f). En la columna X, los valores se enumeran de mayor a menor, sin omitir ninguno. Para la columna de frecuencia, se determinan las cuentas para cada valor (con qué frecuencia ocurre cada valor de X en el conjunto de datos). Estas cuentas son las frecuencias para cada valor de X. La suma de las frecuencias debe ser igual al número de observaciones (N).

La tabla 3 muestra veinte observaciones de la «Encuesta continua de hogares» tomadas al azar. Si nos interesa investigar cómo afecta el número de habitaciones de la vivienda a algún resultado de interés de política pública (por ejemplo, el rendimiento educativo), debemos empezar por describir cuántas habitaciones tienen las viviendas en nuestra población de interés. En concreto, cómo varía el número de habitaciones de la vivienda.

Tabla 3. Fragmento de la *Encuesta continua de hogares* 2018.

id_viv	ca	idq_pv	tipoviv	metrosvi	regvi	habvi	densidadvi
073443	05	35	1	0180	1	5	60
059321	13	28	5	0138	2	5	46
024757	10	46	5	0090	2	6	22.5
065139	19	52	5	0092	2	5	92
068892	01	18	2	0100	1	6	100
090857	11	10	5	0118	1	6	59
032499	13	28	5	0080	2	3	40
025713	02	50	5	0065	1	4	65
052257	02	50	5	0074	1	4	24.7
030525	01	18	5	0086	2	5	43
024614	10	46	4	0099	3	5	16.5
055340	11	10	5	0100	4	5	33.3
021115	14	30	2	0070	1	5	35
079361	01	11	5	0058	2	5	29
051262	10	46	5	0068	1	5	13.6
072956	03	33	1	0090	1	4	90
000874	07	05	5	0100	1	6	100
064281	16	48	1	0182	1	9	60.7
082938	13	28	5	0138	1	6	69
051712	16	48	5	0069	2	4	34.5

Fuente: http://www.ine.es

Además de las dos columnas para las categorías en la escala de medición y para la frecuencia (el número de observaciones en cada categoría) que acabamos de explicar, una tabla de distribución de frecuencia también puede incluir una tercera columna que indique el porcentaje acumulado. En la columna de las categorías, los valores pueden aparecer ordenados (en el caso de variables ordinales) o seguir un orden alfabético o de interés sustantivo, sin omitir ninguno. Para la columna de frecuencia, se determinan las cuentas para cada valor (con qué frecuencia ocurre cada valor de la variable de interés en el conjunto de datos); estas cuentas son las frecuencias para cada valor de la variable de interés. La suma de las frecuencias debe ser igual al número de observaciones.

La tabla 4 muestra la distribución de frecuencias de la variable «número de habitaciones» (Habvi).

Tabla 4. Distribución de frecuencias del número de habitaciones.

habvi	Frecuencia	Porcentaje	Porcentaje acumulado
1	98	0.1	0.1
2	1,009	1	1.1
3	3,336	3.32	4.42
4	14,296	14.22	18.64
5	41,455	41.23	59.87
6	23,733	23.61	83.47
7	9,566	9.51	92.99
8	4,039	4.02	97.01
9	1,857	1.85	98.85
10	782	0.78	99.63
11	202	0.2	99.83
12	80	0.08	99.91
13	50	0.05	99.96
14	17	0.02	99.98
15	10	0.01	99.99
16	6	0.01	99.99
17	1	0	100
18	2	0	100
19	1	0	100
20	1	0	100
22	1	0	100
Total	100,542	100	

Encuesta continua de hogares 2018.

Cuando una tabla de distribución de frecuencia enumera todas las categorías individuales se llama una **distribución de frecuencia regular**. Sin embargo, en ciertas ocasiones, un conjunto de categorías cubre una amplia gama de valores. En estas situaciones, una lista de todos los valores de la variable de interés sería bastante larga, demasiado larga para ser una presentación «simple» de los datos. Para corregir esta situación, se utiliza una tabla de **distribución de frecuencia agrupada**. Dado que existen pocos hogares con viviendas con más de diez habitaciones (menos del 2 %, tal como puede observarse en la columna de porcentaje acumulado), estos se agruparon en una categoría de 10 o más habitaciones (tabla 5).

Tabla 5. Distribución de frecuencias agrupada del número de habitaciones.

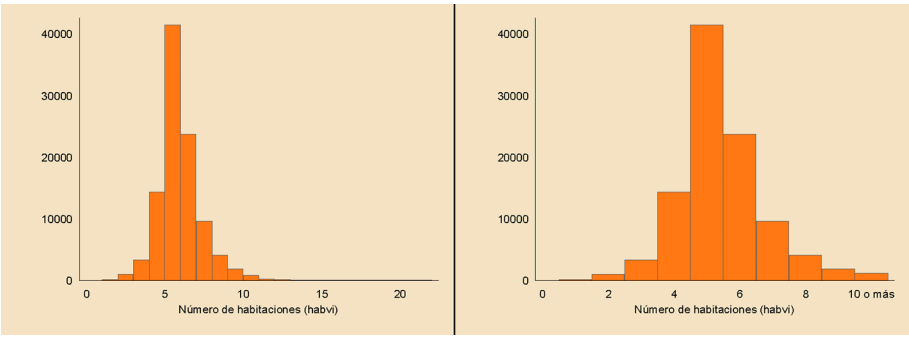
habvi	Frecuencia	Porcentaje	Porcentaje acumulado
1	98	0.1	0.1
2	1,009	1	1.1
3	3,336	3.32	4.42
4	14,296	14.22	18.64
5	41,455	41.23	59.87
6	23,733	23.61	83.47
7	9,566	9.51	92.99
8	4,039	4.02	97.01
9	1,857	1.85	98.85
10 o más	1,153	1.15	100
Total	100,542	100	

Encuesta Continua de Hogares 2018.

Una forma alternativa de representar la distribución de frecuencias es mediante un gráfico. La convención es representar los valores de la variable de interés en el eje *X* y las frecuencias en el eje *Y*. En caso de que estemos interesados en representar una variable categórica, es posible utilizar un diagrama de barras.

En un **histograma**, una barra está centrada sobre cada categoría (o intervalo de clase), de modo que la altura de la barra corresponde a la frecuencia y el ancho se extiende hasta los límites reales, así que las barras adyacentes se tocan. La figura 1 muestra los histogramas para las distribuciones de frecuencias y frecuencias agrupadas.

Gráfico 1. Histogramas del número de habitaciones.



Cuando las categorías se miden en una escala nominal u ordinal, la representación gráfica de la distribución de frecuencias debe ser un gráfico de barras. Un **gráfico de barras** es como un histograma, excepto porque se dejan espacios entre las barras adyacentes. La principal diferencia entre un histograma y un diagrama de barras es que un histograma solo se usa para trazar la frecuencia de ocurrencias de un valor en un conjunto de datos continuos que se ha

dividido en clases, llamadas *intervalos*. Los gráficos de barras, por otro lado, pueden usarse para muchos otros tipos de variables, incluidos los conjuntos de datos ordinales y nominales.

Los gráficos de distribución de frecuencia son útiles porque muestran el conjunto completo de valores. De un vistazo, es posible determinar los valores más altos, los valores más bajos y dónde se centran los valores. El gráfico también muestra si los valores están agrupados o dispersos en un rango amplio. Por esta razón, nos interesa la forma de la distribución. La forma de la distribución nos permite ver, por ejemplo, en qué medida hay simetría con respecto a un eje vertical. Una distribución es simétrica si el lado izquierdo del gráfico es (aproximadamente) una imagen espejo del lado derecho. Un ejemplo de una distribución simétrica es la distribución normal, en forma de campana. En cambio, las distribuciones están sesgadas cuando las puntuaciones se acumulan a un lado de la distribución, dejando una «cola» de unos pocos valores extremos en el otro lado. Analizar la forma de la distribución nos indica en qué medida nuestra población está sesgada hacia los extremos.

En nuestro caso, podemos ver que si no agrupamos los datos (panel izquierdo del gráfico 1) tenemos una cola hacia la derecha con muy pocas viviendas que tienen entre 10 y 22 habitaciones. Esto se llama una distribución sesgada positivamente, dado que los valores tienden a acumularse en el lado izquierdo de la distribución con la cola disminuyendo hacia la derecha. En caso de que nos interese agrupar los hogares con más de 10 habitaciones, podemos ver que la distribución es simétrica (panel de la derecha del gráfico 1).

2.3. Medidas de centralidad

Las estadísticas descriptivas más comunes son las medidas de tendencia central.

Tal como su nombre indica, las **medidas de tendencia central** intentan ubicar el punto central o medio en un grupo de datos. En términos generales, la tendencia central es una medida estadística que determina un valor único que describe con precisión el centro de la distribución y representa la distribución completa de los valores. El objetivo de la tendencia central es identificar el valor único que sea el mejor representante para todo el conjunto de datos.

Al identificar el «**valor promedio**», la tendencia central permite a los investigadores resumir o condensar un gran conjunto de datos en un solo valor. Por lo tanto, la tendencia central sirve como una estadística descriptiva porque permite a los investigadores describir o presentar un conjunto de datos de

forma concisa y muy simplificada. Además, es posible comparar dos (o más) conjuntos de datos simplemente comparando el valor promedio (tendencia central) para un conjunto frente al valor promedio para otro conjunto.

Es esencial que la tendencia central esté determinada por un procedimiento objetivo y bien definido para poder expresar exactamente cómo se obtuvo el valor promedio y poder duplicar el proceso. Ningún procedimiento único produce siempre un valor representativo. Hay tres técnicas comúnmente utilizadas para medir la tendencia central:

- la **media**, que es el promedio aritmético de las observaciones;
- la **mediana**, que es la observación que cae exactamente en la mitad del grupo, y
- la **moda**, que es el valor que se produce con mayor frecuencia.

Estas medidas de resumen se denominan *estadísticas*, entendiéndose por *estadística* cualquier cantidad numérica cuyo valor está determinado por los datos.

La **media** es la medida de tendencia central más utilizada. El cálculo de la media se hace sobre variables cuantitativas (con valores numéricos o de intervalos). La media se obtiene calculando la suma, o el total, para el conjunto completo de datos y luego dividiendo esta suma por el número de observaciones. Supongamos que tenemos una muestra de *n* observaciones cuyos valores designamos por *x1*, *x2*, … *xn*. El valor promedio de la muestra es:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Conceptualmente, la media también puede definirse como la cantidad que recibe cada individuo cuando el total se divide por igual entre todos; o bien, como el punto de equilibrio de la distribución, porque la suma de las distancias por debajo de la media es exactamente igual a la suma de las distancias por encima de la media.

Para ilustrar el cálculo de la media, la tabla 6 muestra los Presupuestos Generales del Estado para 2018. Para conocer el valor medio del presupuesto para cada una de las políticas, debemos sumar el presupuesto total y dividirlo por el número de políticas. Esto es, 368,323 miles de millones de euros divididos en 27 políticas, que dan como resultado una media de 13,642.59 euros por partida.

Tabla 6. Presupuestos Generales del Estado para 2018.

Políticas	Millones de euros
1. Acceso a la Vivienda y Fomento de la Edificación	481

Fuente: Estadísticas de los Presupuestos Generales del Estado (www.sepg.pap.hacienda.gob.es).

Políticas	Millones de euros
2. Otras actuaciones de carácter económico	639
3. Investigación militar	679
4. Órganos Constitucionales, Gobierno y otros	681
5. Cultura	887
6. Comercio, Turismo y P.Y.M.E.S.	896
7. Administración Financiera y Tributaria	1,390
8. Política Exterior	1,581
9. Justicia	1,780
10. Subvenciones al transporte	2,139
11. Educación	2,541
12. Servicios Sociales y Promoción Social	2,631
13. Sanidad	4,251
14. Infraestructuras	5,437
15. Fomento del empleo	5,716
16. Industria y Energía	5,771
17. Investigación civil	6,379
18. Agricultura, Pesca y Alimentación	7,707
19. Defensa	8,401
20. Seguridad ciudadana e Instituciones penitenciarias	8,418
21. Otras Prestaciones Económicas	14,385
22. Gestión y Administración de la Seguridad Social	17,297
23. Desempleo	17,702
24. Servicios de carácter general	24,643
25. Deuda Pública	31,547
26. Transferencias a otras Administraciones Públicas	49,510
27. Pensiones	144,834

Fuente: Estadísticas de los Presupuestos Generales del Estado (www.sepg.pap.hacienda.gob.es).

Sin embargo, vemos que la política de «Pensiones» es muy superior al resto y que las de «Acceso a la vivienda», «Investigación militar», «Otras actuaciones de carácter económico» y «Órganos constitucionales, gobierno y otros», son muy inferiores. Estos valores extremos, a veces denominados *valores atípicos*, tienen una influencia desproporcionada sobre la media y, por lo tanto, pueden afectar a la forma en la que la media representa los datos. En estos casos, la media se desplazará hacia los extremos (se desplazará hacia la cola)

y no proporcionará un valor «central». Esto implica que la media no siempre funciona como una medida de tendencia central y es necesario contar con procedimientos alternativos disponibles.

La segunda medida de tendencia central es la mediana. La **mediana** se define como el punto medio de la lista cuando los valores en una distribución se enumeran en orden de menor a mayor. Esta divide las puntuaciones, de modo que el 50 % de los valores en la distribución tengan valores iguales o menores a la mediana. El cálculo de la mediana requiere valores que puedan ordenarse (de menor a mayor) y se midan en una escala ordinal, de intervalo o de relación. La mediana del presupuesto de 2018 sería entonces de 5,437 euros. El valor de la política que se encuentra en el medio de la distribución, la número 14, es «Infraestructuras» ?que puede calcularse como (27+1)/2). En este caso se tiene un número impar de políticas y el cálculo es sencillo. Con un número par de valores, deben ordenarse los valores y la mediana está a medio camino entre las dos puntuaciones medias (es decir, el promedio).

Una ventaja de la mediana es que no se ve afectada por valores extremos. Por lo tanto, la mediana tiende a permanecer en el «centro» de la distribución, incluso cuando hay algunos valores extremos o cuando la distribución es muy sesgada. En estas situaciones, la mediana sirve como una buena alternativa a la media.

La tercera medida de tendencia central es la moda. La **moda** se define como la categoría o valor más frecuente en la distribución. En un gráfico de distribución de frecuencias, la moda es la categoría o puntuación correspondiente al punto más alto o máximo de la distribución.

Si quisiéramos describir la distribución del número de habitaciones por hogar que vimos en el apartado anterior, sería posible calcular la moda para los datos de la tabla 4. Podemos ver fácilmente que el número de habitaciones con mayor frecuencia es cinco. En este caso, también podríamos calcular la media. Encontramos que el número medio de habitaciones por hogar en nuestra muestra es de 5,4. La moda se usa habitualmente como una medida complementaria de tendencia central que se informa junto con la media o la mediana.

Por otra parte, si queremos describir los datos agregados en categorías (tabla 5), vemos que ya nos es posible calcular la media porque no tenemos valores numéricos, sino una categoría de «10 o más». Sin embargo, podemos ver que en estos datos la categoría que se repite con mayor frecuencia son los hogares con cinco habitaciones. La mayor ventaja de la moda es que es la única medida de tendencia central que se puede utilizar para datos medidos en una escala categórica.

Sin embargo, también debemos considerar que es posible que una distribución tenga más de una moda. En el caso de que hubiese muchos hogares pequeños (con tres habitaciones), pero también muchos hogares con seis, tendríamos

una distribución con dos valores que se repiten con la mayor frecuencia. Tal distribución se llama *bimodal*. Es importante tener en cuenta que una distribución puede tener una sola media y una sola mediana, pero dos o más modas. El término *moda* se usa a menudo para describir un pico en una distribución que no es realmente el punto más alto. Por lo tanto, una distribución puede tener una moda mayor en el pico más alto y una moda menor en un pico secundario en una ubicación diferente.

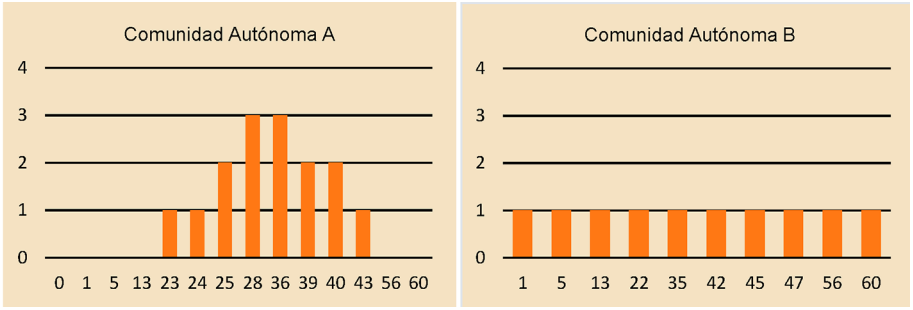
2.4. Medidas de dispersión

Las **medidas de dispersión** (también conocidas como **variación**) nos informan sobre el grado en el que los datos se agrupan sobre la media. También puede entenderse como un indicador de cómo se distribuyen los valores en una distribución.

La dispersión sirve como una medida descriptiva y como un componente importante de la mayoría de las estadísticas inferenciales. Como estadística descriptiva, mide el grado en el que los valores se distribuyen o agrupan en una distribución. Una medida de dispersión generalmente acompaña a una medida de tendencia central como estadística descriptiva básica para un conjunto de valores. La tendencia central describe el punto central de la distribución y la dispersión describe cómo los valores se distribuyen alrededor de ese punto central.

En el contexto de las estadísticas inferenciales, la dispersión proporciona una medida de la precisión con la que un valor u observación individual representa a toda la población. Las dos muestras del gráfico 2 tienen una media de 32.6. Cuando la dispersión de la población es pequeña, todas las puntuaciones se agrupan juntas, y una medida de tendencia central proporcionará una buena representación de todo el conjunto (gráfico 2A). Por el contrario, cuando la dispersión es grande y los valores están ampliamente distribuidos, es fácil que uno o dos valores extremos den una imagen distorsionada de la población general (gráfico 2B).

Gráfico 2. Ejemplos de diferentes niveles de dispersión.



La dispersión puede medirse en diferentes formas: el rango, la desviación estándar (o varianza), el coeficiente de variación, el rango intercuartil o los valores Z. En todos estos casos, la dispersión se determina midiendo la distancia. A continuación, se presentan las tres primeras formas.

El **rango** es la distancia total cubierta por la distribución, desde el puntaje más alto hasta el puntaje más bajo (utilizando los límites reales superior e inferior del rango). El rango es altamente sensible a los valores atípicos, pero no se ve afectado por la forma de la distribución. En el ejemplo del gráfico 2, el rango para la comunidad autónoma A es de 20 (valor máximo-valor mínimo), mientras que en la comunidad autónoma B tiene un valor de 59.

La **desviación estándar** mide la distancia estándar entre un valor y la media. El cálculo de la desviación estándar se puede resumir como un proceso de seis pasos:

- Calcular la media.
- Calcular la desviación (distancia desde la media) para cada valor.
- Elevar al cuadrado cada desviación.
- Sumar las diferencias al cuadrado.
- Dividir la suma por el número de observaciones.
- Calcular la raíz cuadrada de la varianza.

Estos pasos pueden resumirse en la siguiente fórmula:

$$Sy = \sqrt{\frac{\sum_{i=1}^{N} (Xi - \mu)^2}{N}}$$

Para el ejemplo del gráfico 2, la desviación estándar es de 6.83 para la comunidad autónoma A y de 16.35 para la comunidad autónoma B. Los pasos de cálculo se presentan en la tabla 7.

Tabla 7. Cálculo de la desviación estándar para dos muestras.

Comunidad Autónoma A					**Comunidad Autónoma B**			
Xi	μ	Xi - μ	$(Xi - \mu)^2$		Xi	μ	Xi - μ	$(Xi - \mu)^2$
22	32.6	-10.6	112.4		1	32.6	-31.6	998.6
24	32.6	-8.6	74.0		5	32.6	-27.6	761.8
25	32.6	-7.6	57.8		13	32.6	-19.6	384.2
25	32.6	-7.6	57.8		22	32.6	-10.6	112.4
28	32.6	-4.6	21.2		35	32.6	2.4	5.8
28	32.6	-4.6	21.2		42	32.6	9.4	88.4
28	32.6	-4.6	21.2		45	32.6	12.4	153.8

Comunidad Autónoma A					Comunidad Autónoma B			
36	32.6	3.4	11.6		47	32.6	14.4	207.4
36	32.6	3.4	11.6		56	32.6	23.4	547.6
36	32.6	3.4	11.6		60	32.6	27.4	750.8
39	32.6	6.4	41.0					
39	32.6	6.4	41.0					
40	32.6	7.4	54.8					
40	32.6	7.4	54.8					
43	32.6	10.4	108.2					

	CA A	CA B
Suma valor cuadrado de las diferencias	699.6	4010.4
Suma cuadrado de las diferencias / N (varianza)	46.6	267.4
Desviación estándar (raíz cuadrada de la varianza)	6.83	16.35

3. Relaciones entre variables categóricas

3.1. Relaciones entre variables y verificación de hipótesis

Hasta ahora se han presentado métodos de análisis estadístico que se aplican a una sola variable. Sin embargo, para la mayoría de los análisis es de interés no solo describir las variables por separado, sino poder responder a preguntas sobre las relaciones entre las variables. Por ejemplo, ¿existe alguna relación entre las variables o características? En caso de que exista una relación, ¿hasta qué punto es fuerte? ¿Puede utilizarse una variable para predecir otra?

El análisis de datos en dos dimensiones o bivariados se lleva a cabo mediante enfoques similares a los de los datos univariados. Se hacen tabulaciones, representaciones gráficas y se analizan las características numéricas. El propósito de este tipo de análisis es describir fenómenos de interés (qué está pasando) para poder observar una asociación y establecer regularidades o patrones, o explicar la relación entre dos variables (por qué está pasando).

Por lo general, entre las variables que se analizan se suele diferenciar entre:

- **Variable dependiente**: es la variable que se quiere predecir o explicar, por lo general representada por la letra Y. También puede ser descrita como el resultado de un valor conocido de la variable independiente.

- **Variable independiente**: la variable (o conjunto de variables) que predice o explica la variable Y. Por lo general se representa con la letra X.

La variable dependiente es aleatoria, esto es, por cada valor dado a la variable independiente, existen muchos posibles resultados para la variable dependiente.

En la medida en que se logra entender qué causa variabilidad en un fenómeno de interés es posible proponer o evaluar políticas públicas con mayor certeza sobre los resultados posibles. Para ello, es necesario tener muy claro un modelo de análisis, es decir, establecer cuál es la variable de interés que se quiere explicar o predecir y cuál o cuáles son las variables que explican estos cambios. Esta variable que se quiere explicar o predecir es la variable dependiente. En la gestión pública, la variable dependiente pueden ser los resultados de una política, los bienes sociales o el grado de éxito de un programa (por ejemplo, logro educativo, nivel de paro, esperanza de vida, etc.). La variable que la explica es la variable independiente. También se denomina *variable explicativa*, *causal* o *exógena*. Esta puede ser un atributo propio del objeto de estudio, como

la composición de los hogares o la densidad de población, o una variable que puede cambiarse o controlarse, como, por ejemplo, el presupuesto o el nivel de competencia de un actor responsable de la política.

Una vez establecidas las variables de interés (por ejemplo, la esperanza de vida y el impuesto a las bebidas azucaradas) es necesario definir claramente cómo creemos (o cómo esperamos) que están relacionadas. Para ello, definimos una hipótesis. La **hipótesis** es un enunciado declarativo que indica explícitamente la relación que se espera encontrar entre las variables (por ejemplo, esperamos que un aumento en el impuesto a las bebidas azucaradas haga crecer la esperanza de vida). A partir de una hipótesis que pueda ser contrastada es posible concluir sobre la relación que esperamos entre las variables. Es decir, podemos evaluar en qué medida la evidencia empírica disponible (los datos que hemos observado) permite rechazar o no la hipótesis .

Lectura recomendada

Una explicación más detallada sobre el diseño de investigación y las teorías e hipótesis en las ciencias sociales puede verse en:

I. Crespo; E. Anduiza; M. Méndez (2009). *Metodología de la ciencia política*. Madrid: Centro de investigaciones sociológicas («Cuadernos Metodológicos», 28).

A este tipo de análisis entre dos variables se le denomina **análisis bivariado** (o bivariante). La tabla 8 resume las técnicas de análisis apropiadas para cada tipo de variables.

Tabla 8. Técnicas de análisis por tipo de variable.

		Variable independente (VI)	
		Categórica	Continua
Variable dependiente (VD)	Categórica	Tablas de contingencia	(Regresión logística)
	Continua	Diferencia de medias	Correlación Regresión

A continuación, se describen los procedimientos de análisis de datos que son apropiados para las relaciones que involucran variables categóricas, esto es, las tablas de contingencia.

3.2. Tablas de contingencia: celdas, columnas, filas y marginales

Una **tabla de contingencia** es una tabulación cruzada que resume simultáneamente dos variables de interés. Se utiliza cuando alguna o las dos variables tienen escalas categóricas u ordinales. El propósito es analizar la variación conjunta de las dos variables. Se suelen agrupar las observaciones en una tabla de frecuencias bivariadas donde se presenta la distribución de frecuencias correspondiente a las parejas de valores o categorías de las dos variables.

Para construir la tabla de contingencia es necesario establecer cuáles son las variables dependiente e independiente. En este caso queremos explicar el tipo de prestación (VD) y cómo varía por género (VI). Para ello es necesario organizar los porcentajes para las categorías de la VI. De esta manera es posible comparar los tipos de prestación entre las categorías hombre y mujer. Con-

vencionalmente, la VI va en las columnas y la VD va en las filas (porcentajes por columnas), y se indica el número de observaciones para cada una de las categorías de la VI.

La tabla 9 presenta el porcentaje de beneficiarios de prestaciones por desempleo y tipo de prestación por género (SEPE, enero de 2019, total nacional).

Tabla 9. Beneficiarios de prestaciones por desempleo y tipo de prestación por género.

	Hombres	**Mujeres**	**Marginal**
Prestación Contributiva	46%	41%	43%
Subsidio por Desempleo	39%	40%	39%
Renta Agraria	4%	4%	4%
Subsidio Agrario	4%	7%	5%
Renta Activa de Inserción	7%	9%	8%
Programa de Activación para el Empleo	0%	0%	0%
Total	**872,946**	**1,025,423**	**1,898,369**

Fuente: http://www.sepe.es

En términos generales, puede decirse que existe una relación si los valores de la VD son diferentes según los valores de la VI, pero hay tres interpretaciones posibles de la tabla. La más simple es por categorías de la VI. Es posible ver, por ejemplo, que un 46 % de los hombres reciben una prestación contributiva. La segunda interpretación puede hacerse por comparación entre las categorías de la VD. Puede verse que hay más hombres que reciben la prestación contributiva que mujeres o que las mujeres reciben la renta activa de inserción en una proporción mayor que la de los hombres. La tercera comparación es con el valor global de la muestra (marginal). Puede observarse, por ejemplo, que entre los hombres hay un mayor porcentaje que recibe una prestación contributiva que en el global de la muestra (46 % > 43 %).

3.3. Medidas del grado de asociación entre variables

A partir de los análisis, es posible llegar a conclusiones sobre la dirección y la fuerza de la relación entre las variables. Para las variables nominales basta con describir qué valores de la VI están asociados con qué valores de la VD. En el caso de tener variables ordinales, es posible hablar de la dirección, es decir, de relaciones positivas o negativas.

Por otro lado, la fuerza de la relación indica en qué medida difieren los valores de la VD en las categorías de la VI. La relación es perfecta cuando todos los valores de una categoría de la VI van asociados a una categoría diferente de la VD. La fuerza de la relación no viene determinada por el nivel de significación estadística.

Podemos calibrar la fuerza de la relación a partir de la diferencia entre los porcentajes correspondientes a las diferentes categorías de la variable independiente. Por ejemplo, la diferencia por género en la prestación contributiva es de cinco puntos (46-41, relación fuerte), mientras que no hay diferencias entre hombres y mujeres para la Renta agraria ni para el Programa de activación para el empleo. Las diferencias por género para el Subsidio por desempleo, el Subsidio agrario o la Renta activa de inserción son más débiles por que alcanzan tan solo los tres puntos porcentuales.

Adicionalmente, debemos saber hasta qué punto la diferencia entre las frecuencias observadas y las esperadas es lo suficientemente grande como para decir que existe una relación con la población.

3.4. La prueba de significación Chi cuadrado

La hipótesis que hemos planteado es que existen diferencias entre géneros en las prestaciones por desempleo. El siguiente paso es evaluar la evidencia empírica para evaluar si las diferencias observables son lo suficientemente grandes como para decir que existe una relación con la población. Si se cumple esta condición, es posible afirmar que el resultado es estadísticamente significativo. Esto quiere decir que no es probable que se haya producido al azar.

Para ello se utilizará la prueba de significación Chi cuadrado:

$$\chi^2 = \sum \frac{(Fo - Fe)^2}{Fe}$$

donde *Fo* se refiere a la frecuencia observada y *Fe* a la frecuencia esperada. La frecuencia esperada se calcula así:

Fe = [(marginal fila/total)*(marginal columna/total)]*total
= (marginal fila)*(marginal columna)/total

La tabla 10 es una tabla de contingencia con dos categorías de la VD y dos de la VI, y los pasos para calcular los valores de Chi cuadrado.

Tabla 10. Cálculo de Chi cuadrado.

	X1	X2	Marginal columna			Fo	Fe	Fo - Fe	$(Fo - Fe)^2$	$(Fo - Fe)^2/Fe$
Y1	5	15	20		Y1 X1	15	10	5	25	2.5
Y2	15	5	20		Y1 X2	5	10	-5	25	2.5
Marginal fila	20	20	40		Y2 X1	15	10	5	25	2.5
					Y2 X2	5	10	-5	25	2.5

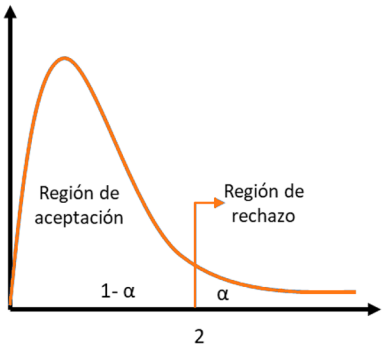
	X1	X2	Marginal columna			Fo	Fe	Fo - Fe	$(Fo - Fe)^2$	$(Fo - Fe)^2/Fe$
									Suma	10

Una vez sabemos que el estadístico Chi cuadrado (?²) tiene un valor de 10, es necesario calcular los grados de libertad a partir de los atributos de las variables (número de filas y número de columnas en la tabla cruzada).

Grados libertad = (número de filas - 1)*(número de columnas - 1)

En el ejemplo de la tabla 10, se tienen dos filas y dos columnas, así que los grados de libertad serían de 1. Con esta información, es posible calcular el nivel de significancia (a) a partir de la distribución de Chi cuadrado (o distribución de Pearson). Esta es una distribución de probabilidad continua con un parámetro que representa los grados de libertad de la variable aleatoria. Con la información disponible en la tabla de distribución de probabilidad Chi cuadrado, es posible concluir que la probabilidad de obtener un valor 10 o mayor con 1 grado de libertad es menor que p = 0.005. Por lo tanto, el valor es estadísticamente significativo, pues es menor que 0.01.

Gráfico 3. Distribución de probabilidad Chi cuadrado y tabla inversa.



Grados libertad	0.1	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75

El nivel de significancia indica el nivel de confianza que deseamos que tengan los cálculos de la prueba; es decir, si queremos tener un nivel de confianza del 95 %, el valor de significancia (a) debe ser del 0.05. Son comunes los niveles de significancia del 0.05, 0.01 y 0.1. Habitualmente, no es suficiente decir que una relación es significativa, sino que se especifica a qué nivel lo es. En las tablas de resultados, se suele indicar mediante asteriscos: * p <0.05, ** p <0.01 y *** p <0.001. ***

Es necesario tener en cuenta los siguientes puntos para interpretar correctamente el nivel de significación:

- No es reversible: p = 0.01 no quiere decir que hay un 99 % de probabilidad de que exista una relación.
- Está basado en el supuesto de que el muestreo es aleatorio.

- No dice nada sobre otras fuentes de error posibles (por ejemplo, sesgos en la muestra o errores de medida).
- No implica que el resultado sea importante en sentido práctico (significación sustantiva).
- No indica la fuerza de la relación entre las variables.
- Cuanto más grande es la muestra, más fácil es encontrar relaciones estadísticamente significativas.
- No indica que la relación sea causal.

4. Relaciones entre variables continuas

4.1. Diferencias de medias y prueba *T*

Un reto común en la Administración pública es poder llegar a conclusiones sobre las diferencias entre dos muestras para analizar los efectos de las medidas o intervenciones políticas. La comparación implica concluir si los valores medidos para una muestra son diferentes de los de la otra muestra en promedio. Por ejemplo, un técnico ambiental necesitará saber si los niveles de contaminación del aire de un municipio mejoraron después de que se regulara la circulación de vehículos antiguos. Para ello, será necesario hacer una comparación entre antes y después de la regulación, o entre dos municipios similares que tan solo difieren en la existencia de esta regulación.

En situaciones como estas, en las que se necesita saber si dos muestras son diferentes en sus medias o proporciones de muestra, la técnica apropiada es una **prueba de diferencia de medias**, que permite hacer análisis para variables de interés continuas entre grupos (variable categórica).

Cuando se busca comprobar si existen diferencias entre dos medias muestrales, el objetivo es determinar si ambas medias podrían haberse extraído de la misma población, o si las dos son tan diferentes que no podrían haberse extraído de la misma población. Para muchos problemas de gestión o evaluación de políticas, se espera que los valores para una muestra sean diferentes de los de la otra.

Al igual que con otro tipo de análisis, la pregunta de investigación debe ser muy clara. Es necesario definir la razón por la que se espera una diferencia entre los dos grupos (¿qué hace que un grupo sea diferente de otro?) y la dirección esperada de la diferencia.

Por ejemplo, si comparamos el desempeño de dos emprendedores que han recibido fondos públicos para el desarrollo empresarial, la hipótesis debe reflejar la expectativa de que el desempeño de los emprendedores que han recibido las ayudas será mayor que el de aquellos que no las han recibido. La lógica general es poder confirmar que los grupos difieren y en qué medida la inversión en programas o cambios institucionales está relacionada con una diferencia en los resultados. Se sigue la lógica de un experimento en el que el grupo que recibe la financiación se asemeja a un grupo que recibe un tratamiento y el que no la recibe se usa como un grupo de control que sirve para conocer el efecto del tratamiento. Esta lógica supone que los grupos son idénticos y que

la selección de estar en los grupos de tratamiento o control es aleatoria. El procedimiento de análisis consiste entonces en definir la variable de interés para los dos grupos.

En este caso tomaremos la utilidad neta de los emprendedores en los cinco primeros años de funcionamiento. La hipótesis que queremos rechazar es que las medias de utilidad neta son similares para los emprendedores que recibieron la financiación pública que para los que no la recibieron. Para ello seguiremos cinco pasos:

1) Calcular las medias y desviaciones estándar de la VD para los dos grupos.

Grupo	Media	Desviación estándar	Número de emprendedores
Emprendedores con financiación pública (tratamiento)	52.1	45.1	22
Emprendedores sin financiación pública (control)	27.1	26.4	22

2) Calcular el parámetro estimado (estadístico), es decir, la diferencia entre ambos grupos.

Estimación media = media control - media tratamiento = 27.1 – 52.1 = -25

Puede observarse que los emprendedores que han recibido financiación pública tienen una utilidad neta en cinco años superior a quienes no la han recibido. Sin embargo, no puede concluirse que esta diferencia sea estadísticamente significativa dado que cada grupo tiene su propia variabilidad.

3) Calcular un error estándar general o «agrupado» para ambos grupos.

Error estándar (media control - media tratamiento) =

$$\sqrt{\frac{S_{y\ Control}}{N_{Control}} + \frac{S_{y\ Tratamiento}}{N_{Tratamiento}}} = \sqrt{\frac{26.4^2}{22} + \frac{45.1^2}{22}} = 11.14$$

A partir de aquí, queremos conocer la probabilidad de que los grupos hayan sido extraídos de la misma población. Para ello es necesario recurrir a métodos paramétricos como el basado en la distribución T-Student. La distribución T-Student es similar a la distribución normal. Tiene como parámetros la media y la varianza, y depende del tamaño de la muestra a través de los grados de libertad.

4) Calcular el estadístico *t* usando la estimación media y el error estándar:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{SE} = \frac{0 \quad - \quad 25}{11.14} = \quad 2.24$$

5) Buscar la puntuación *t* de -2.24 en la tabla *T* para dos colas, para 21 grados de libertad (22 observaciones menos uno).

Web recomendada

Existen herramientas en línea que consultan los valores de las tablas de diferentes estadísticos. Un ejemplo de ello es https://www.uv.es/ceaces/scrips/tablas/tastud.htm

La probabilidad de que la diferencia entre los dos grupos sea igual a 0 es de 0.036. Esto implica que el efecto de la financiación pública de los emprendedores tiene un efecto estadísticamente significativo sobre su utilidad neta en los cinco primeros años.

4.2. Tipos de diferencias de medias

El método presentado en el apartado anterior para una prueba de diferencia de medias es tan solo una de las tres pruebas de este tipo. Puede aplicarse cuando las dos muestras son independientes y se asume que las dos varianzas de población son diferentes. Sin embargo, esta condición no se da siempre y para ello existen otras dos pruebas: una para muestras independientes con varianzas iguales y otra para muestras dependientes. Para decidir qué prueba debe aplicarse a una pregunta de investigación en particular, es necesario comprender la **diferencia entre muestras independientes y dependientes**.

Las **muestras independientes** son aquellas en las que las observaciones de las dos muestras no están «pareadas» de ninguna manera. El mejor procedimiento para obtener muestras independientes es mediante técnicas de muestreo aleatorio. Por ejemplo, si un analista encargado de los programas de ayuda a la dependencia selecciona aleatoriamente dos muestras de una base de datos de su comunidad autónoma, cada una de las cuales consta de 500 beneficiarios, la probabilidad de que exista un emparejamiento uno a uno entre los casos de la muestra es mínima. Así, el primer beneficiario de la muestra A puede ser una mujer de 70 años que vive en el municipio X y cobra la ayuda por incapacidad, mientras que la primera observación en la muestra B puede ser un hombre de 87 años, del municipio Z, que cobra la ayuda por razones de edad. Las demás observaciones de cada muestra deberían ser igualmente diversas en términos de los atributos de los individuos.

Por el contrario, en las **muestras dependientes**, cada observación de una muestra tiene un par similar en la otra. Una prueba para conocer el antes y el después de un tratamiento (por ejemplo, un cambio legislativo o una intervención social) generaría muestras dependientes si los mismos individuos se observaran antes y después.

Las técnicas para el análisis de diferencias de medias con muestras independientes con varianzas iguales y para muestras dependientes siguen lógicas similares.

Lectura recomendada

as técnicas para el análisis de diferencias de medias con muestras independientes con varianzas iguales y para muestras dependientes no se explican en esta guía, pero pueden verse en detalle en:

D. A. Lind; W. G. Marchal; S. A. Wathen (2017). *Pruebas de hipótesis de dos muestras*. En: D. A. Lind; W. G. Marchal; S. A. Wathen. *Estadística aplicada a los negocios y la economía*. México: McGraw-Hill / Interamericana Editores, SA de CV.

4.3. El análisis de correlación

El propósito del análisis de correlación es representar la asociación entre dos variables cuantitativas mediante un conjunto de técnicas numéricas que incluye el análisis gráfico y el uso de indicadores. El **análisis de correlación** sirve para indagar si los cambios en los valores de una variable están asociados con los valores de otra variable. Si esto ocurre, es posible concluir que ambas variables están correlacionadas o bien que hay correlación entre ellas.

Más concretamente, el principal objetivo del análisis de correlación consiste en determinar los siguientes elementos en la relación entre las variables analizadas:

- Si existe o no una asociación entre las variables, esto es, si la variación en los valores de una de ellas está relacionada con la variación en los valores de la otra.

- Si la relación que existe entre ellas es directa o inversa (o, dicho de otra forma, positiva o negativa), esto es, si al aumentar el valor de la variable independiente el valor de la variable dependiente también aumenta o, por el contrario, disminuye.

- Determinar la intensidad de la relación entre ambas variables, esto es, la magnitud de la variación que experimenta la variable dependiente cuando la variable independiente cambia.

Para determinar los elementos que caracterizan la asociación que existe entre dos variables, por lo general, el primer paso es trazar los datos en un **diagrama de dispersión**. Este procedimiento proporciona una representación visual de la relación entre las variables. El siguiente paso suele ser calcular el **coeficiente de correlación**, que brinda una medida cuantitativa de la fuerza de la relación entre dos variables.

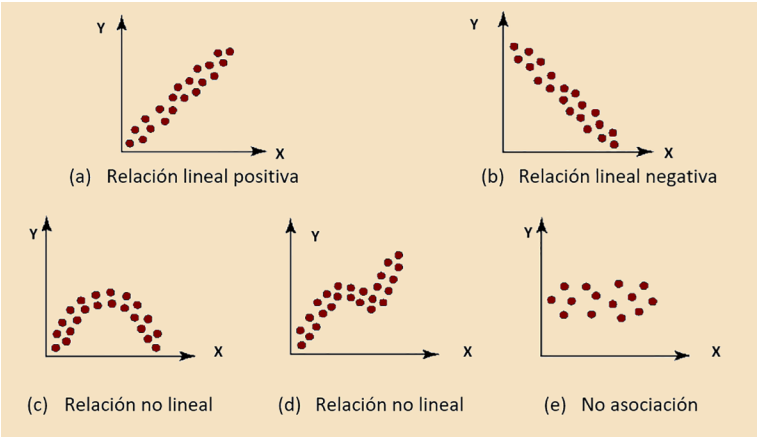
La visualización de la relación: el gráfico de dispersión

Cuando se toma una muestra de dos variables para cada observación de la población o de la muestra, se obtiene una serie de pares de datos. Estas parejas tienen la forma (X,Y) y pueden representarse como puntos en un plano bidimensional o plano cartesiano; la representación gráfica de las parejas se conoce como *diagrama de dispersión* y es una herramienta muy utilizada para conocer la tendencia de los datos antes de profundizar en el estudio de la correlación e incluso del análisis de regresión, que se explicará en la siguiente unidad.

Cada individuo en el conjunto de datos está representado por un punto en el diagrama de dispersión. Es práctica común situar la variable dependiente en el eje vertical o *Y*, y la variable independiente en el eje horizontal o *X*. Para elaborar un diagrama de dispersión simplemente basta con localizar en el eje correspondiente los valores de las variables *X* e *Y* que se observan simultáneamente para cada individuo de la población o muestra.

El diagrama de dispersión resultante sirve para evaluar visualmente si las variables pueden estar relacionadas y qué tipo de relación hay entre ellas. En el gráfico 4 pueden verse cinco ejemplos de relaciones.

Gráfico 4. Ejemplos de relaciones bivariadas para variables continuas.



Nótese que, según sea la dispersión de los datos (nube de puntos) en el plano cartesiano, puede darse el caso de que la relación sea positiva (o directa) o negativa (o inversa). La relación será positiva si cuando *X* aumenta, el valor de *Y* también aumenta, y será negativa si cuando *X* aumenta, el valor de *Y* disminuye. En los casos (a) y (b) la relación que se presenta es claramente positiva y negativa, respectivamente. Adicionalmente, en ambos casos la relación es lineal, esto es, la relación de cambio entre ambas variables es similar a una línea recta cuya pendiente o inclinación dependerá del signo de la relación. Sin embargo, en otros casos no es tan evidente si la relación es positiva o negativa, como ocurre en las relaciones no lineales. En el ejemplo (c), la relación es primero positiva y después de cierto valor de *X* se convierte en negativa. En el caso (d), la relación, que primero es positiva, se convierte en negativa y finalmente pasa a ser positiva de nuevo según aumentan los valores de *X*. En ambos casos, la relación no podría representarse con una línea recta, por lo que son no lineales ?el caso (c) representaría una relación cuadrática y el (d) una relación cúbica. También puede darse el caso que no pueda identificarse claramente un patrón en las nubes de puntos, lo que indicaría que no existe en realidad una asociación entre las variables, como se ejemplifica en el caso (e).

El diagrama, sin embargo, solo puede darnos una «sensación» visual sobre la asociación entre dos variables, pero en realidad no mide la fuerza de tal asociación. Para medir la fuerza de la relación entre *X* e *Y* podemos calcular el coeficiente de correlación.

Para ilustrar este tipo de diagramas, analizaremos, con datos del Ayuntamiento de Barcelona, la asociación que existe entre la población inmigrante de un barrio y el número de parados en ese mismo barrio.

La inmigración

La inmigración es un asunto público importante, especialmente en momentos en los que los cambios demográficos pueden representar tanto una amenaza como una oportunidad para el mercado laboral de los países y la sostenibilidad de sus sistemas de pensiones, así como en contextos en los que la llegada masiva de inmigrantes puede suponer una importante carga para las sociedades de acogida. A pesar de que la percepción de la población sobre la relación entre inmigración y desempleo suele ser negativa, múltiples estudios demuestran que los efectos negativos de la inmigración sobre el mercado laboral son nulos o, en el peor de los casos, muy pequeños, y que tienden a desaparecer con el paso del tiempo. Por otro lado, lo que se ha demostrado, sin embargo, es que los inmigrantes son parte de la población más afectada por las crisis económicas como la que ha sufrido España desde 2008, por lo que no es una sorpresa encontrar una correlación positiva entre el número de inmigrantes y el número de desempleados en una población específica.

En referencia a la escasa influencia de este fenómeno en el mercado de trabajo, recomendamos leer la entrada «De los efectos de la inmigración sobre el mercado de trabajo», publicada por Lidia Farré en el blog *Nada es gratis* (http://nadaesgratis.es/lidia-farre/de-los-efectos-de-la-inmigracion-sobre-el-mercado-de-trabajo) Y en cuanto a la afectación del paro en la población inmigrante véase, por ejemplo, «Los inmigrantes son los más afectados por la crisis y el paro, según Ranstad», publicado en *El Economista.es* en plena crisis en 2009 (https://www.eleconomista.es/economia/noticias/1590521/10/09/Los-extranjeros-son-los-mas-afectados-por-la-crisis-y-el-paro.html), o el estudio de la OCDE que reseñaba más recientemente *El Mundo* en el artículo «La crisis se ensaña más con los inmigrantes que con los españoles» (https://www.elmundo.es/espana/2014/12/01/547b907ce2704e77408b4593.html).

Para analizar esta relación en Barcelona, se examinan los datos del número de inmigrantes y el número de parados para los 73 barrios de la ciudad. La tabla 11 contiene una muestra de dichos barrios, solo los cinco barrios más grandes y los cinco más pequeños según el tamaño de la población.

Tabla 11. Inmigración y paro para una muestra de diez barrios de Barcelona.

		Población inmigrante		Población en paro	
Barrio	**Población**	Número de inmigrantes	Número de inmigrantes por cada 1.000 habitantes	Número de parados	Número de parados por cada 1.000 habitantes
La Clota	590	23	39.0	18	30.5
La Marina del Prat Vermell - ZonaFranca	1146	35	30.5	125	109.1
Vallbona	1354	65	48.0	116	85.7
Can Peguera	2216	68	30.7	121	54.6
Baró de Viver	2511	84	33.5	128	51.0
El Raval	47274	4976	105.3	3565	75.4

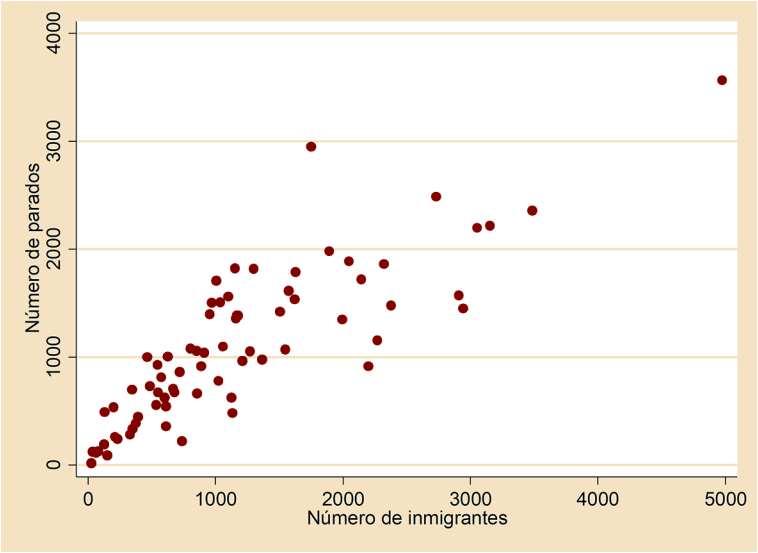
Fuente: Ayuntamiento de Barcelona (2019), Portal de estadísticas, cifras por barrios. http://www.bcn.cat/estadistica/castella/dades/barris/index.htm

		Población inmigrante		Población en paro	
La Vila de Gràcia	50670	3055	60.3	2196	43.3
La Sagrada Familia	51349	3155	61.4	2216	43.2
Sant Andreu	56695	1750	30.9	2947	52.0
La Nova Esquerra del'Eixample	57676	3486	60.4	2357	40.9

Fuente: Ayuntamiento de Barcelona (2019), Portal de estadísticas, cifras por barrios. http://www.bcn.cat/estadistica/castella/dades/barris/index.htm

En el gráfico 5 puede verse el diagrama de dispersión que muestra la relación entre ambas variables para los 73 barrios de la ciudad de Barcelona.

Gráfico 5. Diagrama de dispersión de inmigración y paro por barrios de Barcelona.



El patrón que siguen los datos parece sugerir una relación positiva entre ambas variables, aunque no todos los puntos se encuentran sobre una línea recta. Por esta razón, es necesario medir la fuerza y la dirección de esta relación entre dos variables mediante el coeficiente de correlación.

El cálculo del coeficiente de correlación *r* de Pearson

La medida estadística más utilizada para representar la relación lineal entre dos variables es el coeficiente de correlación de Pearson, que describe la fuerza de la asociación entre dos variables y se designa con la letra *r*.

Para calcular el coeficiente de correlación de Pearson entre *X* e *Y* se utiliza la relación entre la covarianza entre ambas variables y el producto de las desviaciones estándares de estas. La covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias y se calcula como el promedio del producto de las desviaciones de cada variable con respecto a su media:

Nota

Al igual que ocurre en el caso de la desviación estándar, se divide entre *n-1* en el caso de tener datos muestrales o entre *n* en el caso de datos poblacionales.

$$cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

Si bien el resultado no está acotado en ningún rango específico y el número en sí mismo no tiene ninguna interpretación, cuando la covarianza produce un valor positivo, las variables tienden a cambiar en la misma dirección, mientras que cuando produce uno negativo tienden a cambiar en la dirección opuesta.

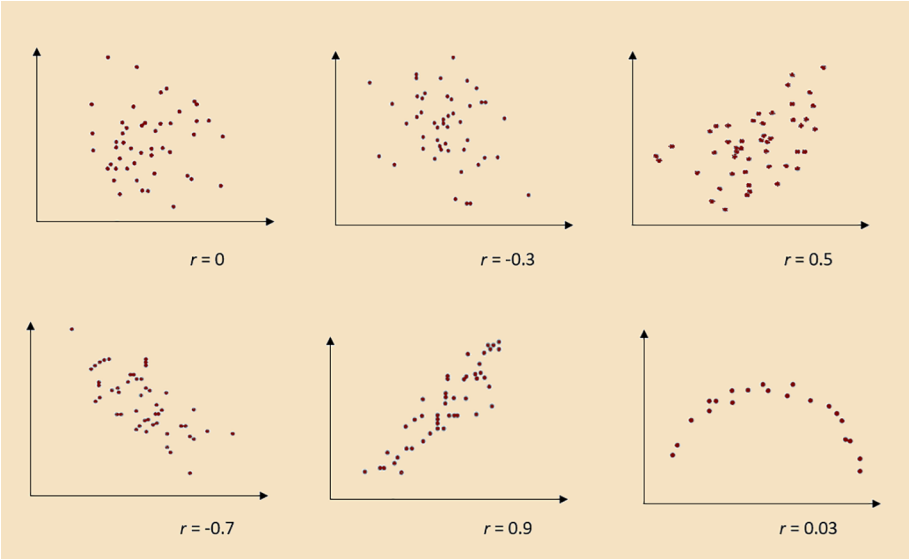
A diferencia de la covarianza, la *r* de Pearson arroja un valor acotado en un rango y permite una interpretación fácil. Este indicador compara la cantidad de variabilidad conjunta entre *X* e *Y* medida por la covarianza con la cantidad en que *X* e *Y* varían por separado. De modo que la fórmula para calcular el coeficiente de correlación es:

$$r = \frac{cov(X, Y)}{S_Y S_X}$$

El valor del coeficiente de correlación de Pearson *r* puede tomar valores desde menos uno hasta uno, esto es, -1 = r = 1. Mientras más cercano a uno sea el valor del coeficiente en cualquier dirección, más fuerte será la asociación lineal entre las dos variables. Por el contrario, mientras más cercano a 0 sea el coeficiente de correlación, más débil es la asociación entre ambas variables. Un coeficiente de correlación de 1 o bien de -1 indica una correlación perfecta, esto es, el diagrama de dispersión representaría una nube de puntos en una línea recta perfecta con pendiente positiva o negativa, respectivamente. Si es igual a 0 se concluirá que no existe relación lineal alguna entre ambas variables.

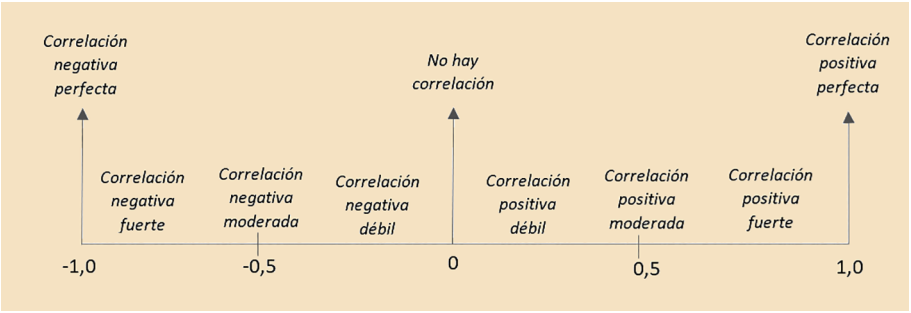
Nótese que el coeficiente de correlación de Pearson se refiere únicamente a la asociación lineal entre dos variables. Adicionalmente, la fuerza de la correlación que indica el coeficiente no depende de la dirección. Por ejemplo, un coeficiente de correlación cercano a 0 (0.07, por ejemplo) indica que la relación lineal es muy débil, y se llega a la misma conclusión si *r* fuera -0.07. Los coeficientes -0.94 y 0.94 indican una correlación muy fuerte entre las dos variables. El gráfico 6 muestra ejemplos de coeficientes de Pearson obtenidos para diferentes diagramas de dispersión. Es interesante ver que en último caso el valor de *r* es casi 0. Esto ocurre porque a pesar de que hay una asociación entre ambas variables, la asociación no es lineal y el coeficiente de correlación de Pearson solo identifica correlaciones lineales.

Gráfico 6. Ejemplos de coeficientes de Pearson para diferentes patrones de dispersión.



No existe una regla precisa para afirmar si la correlación entre las variables puede considerarse fuerte o débil, ya que la calificación depende del rigor del estudio y la experiencia del investigador para juzgar los resultados de acuerdo con las expectativas planteadas. Sin embargo, pueden seguirse ciertas convenciones para resumir la fuerza y la dirección del coeficiente de correlación, que pueden ayudar a la interpretación de un coeficiente de correlación determinado, como se ve en el gráfico 7.

Gráfico 7. Interpretación de un coeficiente de correlación de Pearson.



En resumen, las características del coeficiente de correlación son las siguientes:

- Muestra la dirección y la fuerza de la relación lineal (recta) entre dos variables cuantitativas.
- Varía de -1 a 1, ambos inclusive.
- Un valor cercano a 0 indica que hay poca asociación entre las variables.
- Un valor cercano a 1 indica una asociación directa o positiva entre las variables.
- Un valor cercano a -1 indica una asociación inversa o negativa entre las variables.
- La unidad de medida de X e Y no desempeña ningún papel en la interpretación de r.

Para concluir, debemos recordar que una asociación, por fuerte que sea, no implica necesariamente causalidad. Por ejemplo, aunque se pueda demostrar que los ingresos de profesores y el número de pacientes en instituciones psiquiátricas han aumentado proporcionalmente y que el coeficiente de correlación entre ambas variables sea positivo, no se puede concluir que una variable cause la otra. Asimismo, se ha demostrado que, aunque hay una correlación positiva entre el coste del despido y el nivel de empleo, tal asociación no implica que la reducción del coste del despido genere nuevos puestos de trabajo. Las relaciones de este tipo en las que parece que hay una relación causal que en realidad no existe se denominan **correlaciones espurias**. Lo que se puede concluir cuando se tienen dos variables con fuerte correlación es que hay una relación o asociación entre ambas, no que un cambio en una ocasiona un cambio en la otra.

Abaratamiento del despido

Un estudio del Centro de Investigación en Economía Internacional (CREI) de la Universidad Pompeu Fabra advierte que abaratar el despido no favorece por sí solo el empleo, sino que «depende crucialmente» del grado de incertidumbre del empresario ante la resolución de un posible conflicto judicial. Maïa Güell, autora del estudio, ha argumentado que estos costes pueden tener «efectos ambiguos, neutros o incluso positivos» sobre la tasa de empleo. Véase el siguiente artículo en *La Vanguardia* (31/08/2010): https://www.lavanguardia.com/economia/fiscalidad-empresa/20100831/53992656625/un-estudio-cuestiona-que-abaratar-el-despido-genere-automaticamente-empleo.html

Para continuar con el ejemplo del apartado anterior, calculemos el coeficiente de correlación de Pearson entre el número de inmigrantes y el número de parados para los 73 barrios de la ciudad de Barcelona:

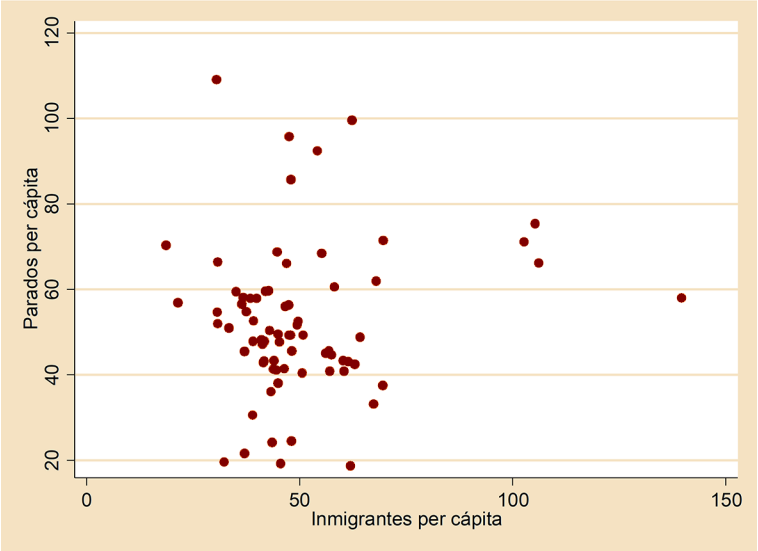
$$r = \frac{cov(X, Y)}{S_Y S_X} = \frac{570.688}{925{,}2 \times 711{,}4} = 0{,}84$$

Con este valor, puede interpretarse que existe una fuerte correlación positiva entre el número de inmigrantes y el número de parados en Barcelona.

Como se ha mencionado antes, esta interpretación no implica que haya una relación causal entre ambas variables. Además, es necesario analizar si la correlación calculada puede ser una correlación espuria. En este caso, si la relación entre el número de inmigrantes y de parados está influenciada por otros factores que aún no se han analizado, puede ser que la correlación hallada no represente realmente la intensidad de la relación entre ambas variables. Un factor que, por ejemplo, puede afectar a esta relación es el tamaño de los barrios. Si en los barrios con una mayor población es más probable que haya más inmigrantes y al mismo tiempo más parados, el tamaño de los barrios puede estar mediando la relación y distorsionando la verdadera intensidad de esta.

Para controlar por el tamaño de los barrios, es posible hacer el mismo tipo de diagramas y el cálculo de la *r* de Pearson utilizando el número de parados y el número de inmigrantes per cápita, esto es, por cada 1.000 habitantes en cada barrio. El gráfico 8 muestra el diagrama de dispersión.

Gráfico 8. Diagrama de dispersión de inmigración y paro por cada 1.000 habitantes por barrios de Barcelona.



Comparado con el gráfico anterior, en el que se mostraba una clara relación positiva entre el número de parados y el número de inmigrantes, con los datos per cápita es más difícil identificar visualmente la dirección de la relación entre ambas variables, si es que puede identificarse alguna.

$$r = \frac{cov(X, Y)}{S_Y S_X} = \frac{38{,}56}{18{,}94 \times 17{,}47} = 0{,}116$$

De hecho, si se calcula el coeficiente de correlación con los datos por cada 1.000 habitantes, se observa que la correlación es positiva pero muy débil.

Esa observación refleja la importancia de avanzar en el análisis y no sacar conclusiones simplemente basándose en la inspección visual de un diagrama de dispersión o de un coeficiente de correlación. Es fundamental analizar dos tipos de problemas.

- En primer lugar, la existencia de particularidades de los datos que pueden afectar a los resultados del análisis, como algunos problemas de medición y la existencia de datos atípicos o *outliers*. En nuestro ejemplo, por un lado, no toda la inmigración está registrada formalmente, y puede ser que los registros estadísticos no sean todo lo precisos que deberían y, por otro lado, la existencia de datos atípicos en los que la inmigración y el paro son particularmente altos, como es el caso del barrio del Raval (es el punto más alejado del origen en el primer diagrama de dispersión), puede estar

generando un coeficiente de correlación excepcionalmente alto cuando no se controla por el tamaño de la población.

- En segundo lugar, es fundamental analizar la naturaleza de la relación y los demás factores que pueden afectarla. Además del tamaño de la población, otras variables pueden afectar a la intensidad de la asociación entre el número de inmigrantes y el número de parados en los barrios de Barcelona, como las características económicas de esos barrios y la oferta de empleo, entre muchos otros.

5. El modelo de regresión lineal

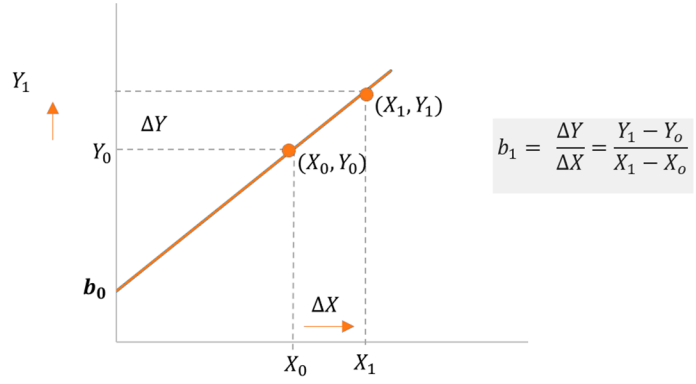
En el apartado anterior, se ha mostrado que el diagrama de dispersión y el análisis de correlación lineal ofrecen una idea bastante aproximada sobre el tipo de asociación que existe entre dos variables y la intensidad de esa asociación. Sin embargo, ambos análisis son limitados cuando se requiere una descripción más precisa de esa relación y no tienen capacidad predictiva. Por ello, una vez se ha realizado un análisis de correlación lineal, el siguiente paso es desarrollar una ecuación matemática que permita estimar el valor de una variable dependiente sobre la base del valor de otra u otras variables independientes. Este procedimiento se conoce como **análisis de regresión** y, cuando se analiza la relación entre dos variables solamente, se denomina **regresión lineal simple**.

La regresión lineal simple pretende encontrar la recta que mejor represente todos los puntos que están representados en un diagrama de dispersión. La ecuación de una línea recta se representa como:

$$Y = b_0 + b_1 X$$

donde el coeficiente b_0 representa la intersección de la línea recta con el eje vertical, esto es, el valor de la variable Y cuando la variable X toma un valor igual a 0, mientras que b_1 representa la pendiente o grado de inclinación de la línea recta, esto es, el cambio medio que se produce en Y cuando la variable X varía en una unidad.

Gráfico 9. Recta de regresión lineal.



Conociendo el valor de estos dos coeficientes, es posible describir la relación entre X e Y, así como estimar con alguna precisión el valor que tomaría Y ante diferentes valores de X.

En la situación poco probable en la que todos los puntos de la nube de dispersión se situaran sobre una misma línea recta, encontrar la ecuación que mejor representa la relación entre *X* e *Y*, es decir, encontrar los parámetros b_0 y b_1, no supondría un problema, ya que bastaría con unir los puntos para obtener la recta con mejor ajuste a la línea de puntos. Sin embargo, en una nube de puntos más realista con una mayor dispersión, es posible trazar un sinnúmero de líneas rectas diferentes. El análisis de regresión lineal simple trata de encontrar la recta que mejor represente el conjunto de datos observados, esto es, la que se ajuste a una descripción más precisa de la nube de puntos representados en el diagrama de dispersión.

Una vez se ha determinado dicha ecuación, esta puede usarse para predecir los valores de *Y* que deberían ocurrir con valores dados de *X*. Es decir, la ecuación de regresión $Y = b_0 + b_1X$ se puede usar para predecir un valor *y* individual que se espera que ocurra con un valor *x* observado:

$$\hat{y}_i = b_0 + b_1x$$

Estos valores se conocen como *valores predichos* o *esperados* de *Y* porque son lo que la línea nos lleva a esperar que estén asociados con los valores de *X*. El símbolo $\hat{y}_i$, «y sombrero», se usa para representar un valor de *Y* que se predice usando la ecuación de regresión, para que podamos distinguirlo de un valor de *Y* real.

5.1. El cálculo de los coeficientes de regresión

Como se ha explicado en el apartado anterior, la regresión lineal simple busca encontrar la recta que mejor represente todos los puntos representados en un diagrama de dispersión. Sin embargo, múltiples líneas que aproximarían razonablemente los datos observados podrían trazarse en dicho diagrama. ¿Cómo se determina cuál es la recta que mejor ajusta los datos?

Para determinar la ecuación de la línea recta que mejor se ajusta a las observaciones, el análisis de regresión simple utiliza el método de **mínimos cuadrados ordinarios**. En este método, se emplean los datos de la muestra para estimar los parámetros b_0 y b_1 que minimizan la suma de los cuadrados de las desviaciones entre los valores observados de la variable dependiente, representados por y^i, y los valores estimados de la variable dependiente a partir de los valores estimados de b_0 y b_1, representados por y_i.

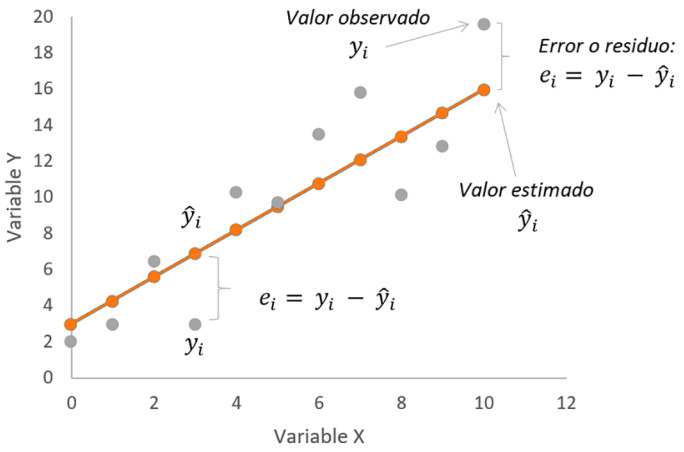
Definamos e_i como el error asociado con la observación *i*. Los errores o residuos se definen como:

$$e_i = y_i - \hat{y}_i$$

Lo que se busca es que la desviación que se obtiene entre la diferencia vertical de los valores «reales» y_i y los valores «estimados» $\hat{y}_i$ sea la menor posible. Dado que las diferencias entre ambos valores pueden ser positivas o negativas, no es correcto simplemente comparar la sumatoria de las desviaciones de diferentes rectas estimadas. Para garantizar la minimización de los errores, lo que se busca es que la suma de los cuadrados de las desviaciones, esto es, $\sum e_i^2 = (y_i - \hat{y}_i)^2$, sea mínima.

Una propiedad de los errores es que su valor esperado es igual a 0 $E(e_i) = 0$, lo que significa que el promedio de los errores es igual a 0. Esto puede entenderse más fácilmente viendo la representación de los errores en el gráfico 10.

Gráfico 10. Valores observados y estimados en la regresión lineal.



Denotaremos la media de X como $\overline{X}$, la media de Y como $\overline{Y}$, sus respectivas desviaciones estándar como S_y y S_x, y el coeficiente de correlación de Pearson como r. Se puede demostrar que los parámetros b_0 y b_1 que minimizan el cuadrado de las desviaciones se pueden calcular como:

$$b_1 = r \frac{S_y}{S_x} \qquad b_0 = \overline{Y} - b_1 \overline{X}$$

Un par de observaciones. La línea de regresión de mínimos cuadrados siempre pasa por el punto ($\overline{X}$, $\overline{Y}$), esto es, por la media de ambas variables. Además, nótese que el signo del coeficiente de correlación coincide con el signo del parámetro b_1, la pendiente de la recta. En consecuencia, si el coeficiente de correlación indica que la relación entre X e Y es positiva, la pendiente de la recta será positiva y el signo de b_1 será positivo. Lo mismo en caso de que la r de Pearson sea negativa.

Es importante subrayar que los resultados del modelo de regresión lineal con el método de mínimos cuadrados ordinarios se obtienen bajo una serie de supuestos, necesarios para poder hacer pruebas de hipótesis de la población a partir de muestras:

- La relación ente las variables *X* e *Y* es lineal en los parámetros. Este es un supuesto que no resulta muy restrictivo, ya que permite incorporar relaciones no lineales entre las variables al introducir funciones (o transformaciones) no lineales de *X* y/o *Y*.

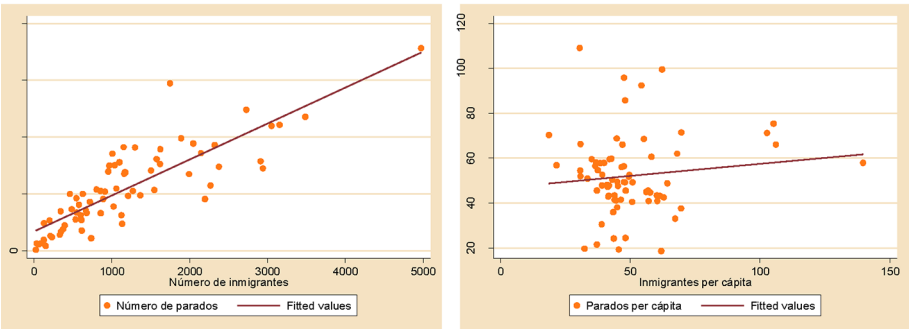
- Los errores o residuos se distribuyen normalmente alrededor de la recta de regresión poblacional.

- Las varianzas de los errores son las mismas en todos los valores de *X* (propiedad que se conoce como *homocedasticidad*).

- Los errores o residuos son independientes entre ellos: no se muestra ningún patrón definido (propiedad que se conoce como *no autocorrelación*).

Para ilustrar el cálculo de los parámetros de la línea de regresión, es posible seguir el ejemplo de la relación entre el número de inmigrantes y el número de parados en los barrios de Barcelona. El gráfico 11 muestra los dos diagramas de dispersión analizados anteriormente, incluyendo ahora la línea de regresión en ambos casos.

Gráfico 11. Línea de regresión para la relación entre inmigración y paro en Barcelona.



Como se podía anticipar a partir de los coeficientes de correlación calculados, en ambos casos se observa una relación positiva, pero con una pendiente mucho más inclinada en el primer caso, en el que no se controla por el efecto del tamaño de la población en cada barrio. Para ver la magnitud de las diferencias en cada caso, estimamos la ecuación de la regresión lineal en ambos casos asumiendo que la variable *Y* es el número de parados mientras que *X* es el número de inmigrantes.

Para los datos del número de parados y número de inmigrantes por barrio:

$$b_1 = r \ \frac{S_y}{S_x} = 0{,}84 \times \frac{950{,}22}{711{,}41} = 0.632$$

$$b_0 = \overline{Y} - b_1\overline{X} = 1.080{,}15 - 0.632 \times 1.166{,}62 = 342{,}78$$

$$Y = 342{,}78 + 0.632X$$

Para los datos del número de parados per cápita y número de inmigrantes per cápita por barrio, en ambos casos por cada 1.000 habitantes:

$$b_1 = r \ \frac{S_y}{S_x} = 0{,}116 \times \frac{17{,}47}{18{,}94} = 0{,}107$$

$$b_0 = \overline{Y} - b_1\overline{X} = 52.1 - 0.107 \times 50{,}1 = 46{,}71$$

$$Y = \ \ 46{,}71 + 0{,}107X$$

Las dos ecuaciones estimadas son muy diferentes porque los datos utilizados también lo son. ¿Cómo se interpretan estos resultados? La primera ecuación indica que, de acuerdo con la magnitud del intercepto, la media del número de parados en los barrios de Barcelona, cuando el número de inmigrantes es igual a 0, es de 342.78. En este caso, la interpretación del intercepto es simple, porque es fácil imaginarse una situación poco probable pero factible en la que algún barrio no tiene ningún inmigrante. Con otro tipo de datos la interpretación del intercepto es menos evidente y no suele comentarse demasiado. Puede darse el caso de análisis que produzcan un intercepto negativo con datos que no suelen ser negativos, o que produzcan valores que es poco probable que ocurran (como precios cero cuando se estima la demanda de cualquier producto según su precio o el de la competencia).

El parámetro más importante es la pendiente de la recta. En el primer caso, la ecuación implica que un incremento de 1 inmigrante en la ciudad está asociado a un incremento de 0.63 parados. En el segundo caso, el parámetro estimado indica que cada incremento de 1 inmigrante por cada 1,000 habitantes está asociado con un incremento de 0.107 parados por cada 1,000 habitantes. La pendiente en este caso es mucho menor, pero no pueden compararse directamente ambas estimaciones, puesto que los datos utilizados en cada caso son diferentes y, por lo tanto, la interpretación también lo es. Lo que sí puede compararse es el grado de significancia estadística de las pendientes en los dos modelos y la bondad de ajuste de ambos.

5.2. El nivel de significancia de los coeficientes de regresión

Una vez estimados los parámetros, debe contrastarse si estos parámetros son significativamente distintos de 0, esto es, si la relación estimada es estadísticamente significativa o no. Tal como se ha realizado en los análisis anteriores, se trata de un tipo de prueba de hipótesis que se conoce como **contraste de significatividad**.

Según se aplica en el análisis de regresión, la hipótesis que se contrasta es que el parámetro estimado es igual a 0 (Ho = b_1 es igual a 0). En el caso de la pendiente, por ejemplo, que el parámetro b_1 es igual a 0, esto es, que no existe asociación entre la variable *X* y la variable *Y*. El parámetro calculado es el valor de b_1 calculado por la regresión con los datos de la muestra, mientras que el valor hipotético es 0.

Como se ha explicado en el apartado «Diferencias de medias y prueba *T*», el estadístico *t* es un ejemplo de un estadístico de prueba cuya distribución bajo la hipótesis nula es conocida. La fórmula general para una prueba *t* es:

$$t = \frac{parámetro \quad calculado - valor \quad hipotético}{error \quad estándar \quad del \quad estimador}$$

Los grados de libertad para esta prueba son el número de observaciones menos dos (*n*-2).

Para entender cómo se realiza este contraste, se presenta el resultado del estadístico *t*, así como la tabla ANOVA (análisis de varianza). Estos se obtienen con programas informáticos para el análisis estadístico, ya que calcularlos manualmente es
complicado.

Se ha utilizado el programa Stata, pero cualquier programa estadístico con modelos de regresión lineal (R, SPSS, Excel) puede producir resultados similares.

La siguiente imagen muestra el resultado de la regresión para los datos del primer modelo, que relaciona el número de parados y el número de inmigrantes por barrio.

Resultados de la regresión. Datos brutos.

Source	SS	df	MS	Number of obs	=	73
				F(1, 71)	=	176.14
Model	26331446.6	1	26331446.6	Prob > F	=	0.0000
Residual	10613696.8	71	149488.687	R-squared	=	0.7127
				Adj R-squared	=	0.7087
Total	36945143.3	72	513126.991	Root MSE	=	386.64

paro	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inmig	.6320525	.0476234	13.27	0.000	.5370942	.7270108
_cons	342.7879	71.65544	4.78	0.000	199.911	485.6647

Después de comprobar que los valores estimados de los parámetros son los que se han calculado antes (Y= 342.78+0.632X), es posible ver el valor del estadístico t y su correspondiente p-valor. Como es usual en este tipo de pruebas, a un nivel de significancia del 5 %, puede rechazarse la hipótesis nula que la pendiente de esta ecuación es igual a 0, ya que el p-valor es menor que el nivel de significancia elegido.

Para los datos del número de parados per cápita y del número de inmigrantes per cápita por barrio, en ambos casos por cada 1.000 habitantes, también se comprueba que los parámetros calculados son los mismos que produce el programa (Y= 46.71+0.107X) pero, a diferencia del anterior, ahora no podemos rechazar la hipótesis nula de que la pendiente de esta ecuación es igual a 0, ya que el p-valor es mayor que el nivel de significancia elegido.

Resultados de la regresión. Datos ponderados por cada 10,000 habitantes.

Source	SS	df	MS	Number of obs	=	73
				F(1, 71)	=	0.98
Model	302.485644	1	302.485644	Prob > F	=	0.3263
Residual	21981.9546	71	309.604994	R-squared	=	0.0136
				Adj R-squared	=	-0.0003
Total	22284.4402	72	309.506114	Root MSE	=	17.596

paropc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inmigpc	.107462	.1087193	0.99	0.326	-.1093181	.3242421
_cons	46.71821	5.823272	8.02	0.000	35.10693	58.32948

Esto significa que, a pesar de que la recta que se observa en el diagrama de dispersión parece representar una relación positiva entre ambas variables, dicha relación no es estadísticamente significativa.

5.3. La bondad de ajuste

Además de estimar el valor de los parámetros de la recta de regresión y su nivel de significatividad, también debe evaluarse el grado en el que la recta se ajusta a la nube de puntos y, por tanto, la capacidad de la ecuación para hacer estimaciones correctas, lo que se conoce como la **bondad de ajuste de la regresión**, que se mide principalmente con el coeficiente de determinación.

Un aspecto útil de la regresión es que puede dividir la variación observada en *Y* (o varianza total) en dos partes: la variación en los valores predichos $\hat{Y}$ y la variación en los errores de predicción e_i.

La variación de *Y* se denomina *suma de cuadrados de la regresión* o *suma de los cuadrados totales*, que denotaremos como SSY, y se define como la suma de las desviaciones al cuadrado de la variable dependiente *Y* respecto a la media de la misma variable, $\overline{Y}$. Cuando se calcula en una muestra:

$$SSY = \sum (y_i - \overline{Y})^2$$

La suma de cuadrados totales SSY puede dividirse en dos partes: la suma de los cuadrados explicada por el modelo (SSY') y la suma de cuadrados de los errores (SSE):

- La suma de cuadrados explicados por el modelo es la suma de las desviaciones al cuadrado de los valores predichos $\hat{y}_i$ y la media de dichos valores:

$$\text{SSY}' = \sum (\hat{y}_i - \overline{\hat{y}})^2$$

- La suma del cuadrado de los errores, como lo hemos definido antes, es la suma de los errores de predicción al cuadrado:

$$\text{SSE} = \sum e_i^2 = \sum (y_i - \hat{y})^2$$

Así, la suma de cuadrados de la regresión SSY puede escribirse como:

$$\text{SSY} = \text{SSY}' + \text{SSE}$$

Esto es, la capacidad de la ecuación de regresión para predecir con precisión los valores de *Y* se mide calculando primero la proporción de la variabilidad de los valores de *Y* estimados con la ecuación de regresión y la proporción que no

puede predecirse o explicarse con el modelo. Mientras que SSY es la variación total de *Y*, SSY' es parte de esa variación explicada por el modelo y SSE es la parte de la variación no explicada.

Por lo tanto, la proporción de variación explicada puede calcularse como:

$$Proporción\ explicada = \frac{SSY'}{SSY}$$

Del mismo modo, la proporción no explicada es:

$$Proporción\ no\ explicada = \frac{SSE}{SSY}$$

La proporción explicada por el modelo es lo que se conoce como el **coeficiente de determinación** de la regresión, el cual expresa el porcentaje de variación de la variable dependiente causado o atribuido a la variación de la variable independiente.

Existe una relación importante entre la proporción de variación explicada y la correlación de Pearson: el coeficiente de determinación puede calcularse como el cuadrado del coeficiente de correlación (r^2). Por ello, la notación del coeficiente de determinación generalmente es R^2. Dado que este coeficiente se calcula como el cuadrado del coeficiente de correlación de Pearson, su valor está acotado en el intervalo entre 0 y 1. Como la *r* de Pearson, en la medida en que los puntos se acerquen a la recta, el coeficiente de determinación será más próximo a 1, y si los puntos se alejan de la recta, el coeficiente de correlación será más próximo a 0.

Para ver cómo funciona la bondad de ajuste en el ejemplo que hemos utilizado, usaremos de nuevo el resultado que produce Stata, ya que normalmente las proporciones explicadas y no explicadas no se calculan manualmente, sino que se obtienen de la tabla ANOVA que produce el comando de regresión.

ANOVA. Datos brutos.

Source	SS	df	MS		
Model	26331446.6	1	26331446.6	Number of obs	= 73
Residual	10613696.8	71	149488.687	F(1, 71)	= 176.14
				Prob > F	= 0.0000
Total	36945143.3	72	513126.991	R-squared	= 0.7127
				Adj R-squared	= 0.7087
				Root MSE	= 386.64

paro	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inmig	.6320525	.0476234	13.27	0.000	.5370942	.7270108
_cons	342.7879	71.65544	4.78	0.000	199.911	485.6647

Para los datos del primer modelo que relaciona el número de parados y el número de inmigrantes por barrio:

$$Proporción\ explicada = \frac{SSY'}{SSY} = \frac{26331446,6}{36945143,3} = 0,7127$$

$$R^2 = r^2 = 0,84^2 = 0,7127$$

De acuerdo con el coeficiente de determinación en este modelo, la variación en el número de inmigrantes en la ciudad explica el 71.3 % de la variación en el número de parados. Como vemos en la siguiente tabla, el coeficiente de determinación en el segundo modelo es solo de 0.0136. Eso implica que la variación en el número de inmigrantes por cada 1,000 habitantes en la ciudad explica solamente el 1.36 % de la variación en el número de parados per cápita.

ANOVA. Datos ponderados por cada 10,000 habitantes.

Source	SS	df	MS		
Model	302.485644	1	302.485644	Prob > F	0.3263
Residual	21981.9546	71	309.604994	R-squared	0.0136
Total	22284.4402	72	309.506114	Root MSE	17.596

paropc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inmigpc	.107462	.1087193	0.99	0.326	-.1093181	.3242421
_cons	46.71821	5.823272	8.02	0.000	35.10693	58.32948

En consecuencia, es posible concluir que la bondad de ajuste del primer modelo es mejor que la bondad de ajuste del segundo. De nuevo, no se puede concluir que el primer modelo es «mejor», porque el objetivo de ambos modelos es explicar una variable dependiente diferente. Sin embargo, mientras que el primero tiene un mayor poder predictivo, en el segundo caso diremos que el modelo estimado no es útil para predecir la variación observada en el número de parados por cada 1.000 habitantes en los barrios de Barcelona.

5.4. Introducción al análisis multivariado

Es claro que la mayoría de las relaciones de interés involucran a más de dos variables. En la práctica, vamos a incorporar tantos controles como creamos necesario. Un modelo de regresión lineal múltiple es más adecuado para análisis *ceteris paribus*, ya que permite controlar explícitamente muchos de los factores que afectan simultáneamente a la variable dependiente. Además, nos da mayor flexibilidad para modelar la relación entre Y y las X. Cuando hay más de una variable independiente x, la más utilizada es la regresión múltiple. La lógica es similar a la de la regresión univariada.

Lectura recomendada

Para saber más sobre la regresión múltiple puede consultarse:

D. A. Lind; W. G. Marchal; S. A. Wathen (2017). *Análisis de regresión múltiple*. En: D. A. Lind; W. G. Marchal; S. A. Wathen. *Estadística aplicada a los negocios y la economía*. México: McGraw-Hill / Interamericana Editores, SA de CV.

Ejercicios de autoevaluación

1. Define las variables para los siguientes conceptos:

- Ingresos
- Presupuesto
- Población objetivo de una política educativa

¿Qué tipo de variable es cada una de ellas?

2. Para los datos de la tabla 2.Tabla de datos procesados (Fragmento del estudio «3242 del Macrobarómetro de marzo de 2019» del CIS:

- ¿Qué tipo de variable es la variable «Edad»? ¿De qué otra manera podría representarse esta variable?
- ¿Qué tipo de variable es la variable «Principal_problema»?
- ¿Cuál es el valor de la variable «Estudios» para la observación 8?

3. Tienes un listado de observaciones del número de vecinos que han ido al Ayuntamiento por día para pedir solución a sus problemas en los últimos dos meses.

10	11	9	9	5	9	10	11	11	6
5	9	10	10	11	7	6	3	11	4
10	9	6	7	10	8	7	6	10	6
9	3	8	8	10	11	10	7	5	6
5	6	10	6	8	6	10	11	9	9
11	6	10	8	11	9	6	7	11	10
6	11	10	6	10	9	8	10	9	7
9	9	7	11	7	10	9	8	8	10
3	2	9	11						

Construye una tabla de distribución de frecuencias y una representación gráfica de los datos. Con esta descripción, escribe una interpretación para incluir esta estadística en el informe trimestral del Ayuntamiento.

4. Los siguientes datos representan el número de semanas que tomaron siete empresas para obtener sus licencias de actividad.

2, 11, 5, 7, 6, 7, 4

- Calcula el tiempo medio para obtener una licencia de actividad.
- Calcula la media de la muestra.
- Calcula la moda de la muestra.

5. Con el fin de hacer recomendaciones sobre un proyecto legislativo, se desea estudiar hasta qué punto existe relación entre el copago y la valoración de los servicios sanitarios. Se ha realizado una encuesta a una muestra aleatoria de 100 individuos, tal como se muestra en la siguiente tabla de frecuencias observadas. ¿Confirma esta evidencia la hipótesis planteada con un nivel de confianza del 95 %?

Valoración del servicio	**Copago sanitario**		**Total**
	Sí	No	
Malo	20	25	45

Valoración del servicio	Copago sanitario		Total
Bueno	10	45	55
Total	30	70	100

Solucionario

1.

Ingresos:

- Nivel de ingreso subjetivo (alto, medio, bajo): variable ordinal.
- Nivel de ingreso objetivo (0-800 €; 801-1.000 €; 1,001-1,200 €; 1,201-1,800 €; 1,801-3,000 €; 3,001-6,000.€; más de 6,000 €): variable ordinal.
- Ingresos mensuales (euros por mes): variable cuantitativa continua.

Presupuesto:

- Grado de ejecución presupuestal como porcentaje del presupuesto total (valor entre 0 y 100 %): variable cuantitativa continua.

Población objetivo de una política educativa:

- Elegibilidad_ Elegible /No elegible: variable categórica.
- Algún criterio de elegibilidad (por ejemplo: Nota media – valor entre 0 y 10): variable continua.

2. La variable «Edad» es de tipo cuantitativo. También podría representarse como una variable categórica en la que se agrupen valores por rangos de edad (menores de 18 años; entre 18 y 25 años; entre 26 y 40 años; entre 40 y 65 años; más de 65 años).

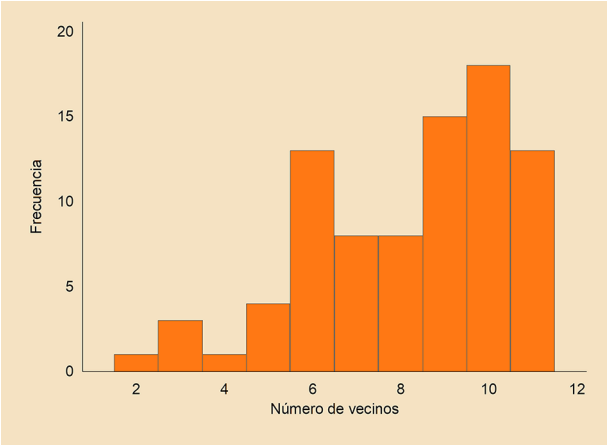
La variable «Principal_problema» es una variable categórica.

El valor de la variable «Estudios» para la observación 8 es estudios superiores, tal como puede verse en la intersección de la fila 8 y la columna «Estudios» de la tabla 2.

3. La tabla de distribución de frecuencias para el número de vecinos que han ido al Ayuntamiento por día para pedir solución a sus problemas en los últimos dos meses:

Número vecinos	Frecuencia	Porcentaje	Porcentaje acumulado
2	1	1.19	1.19
3	3	3.57	4.76
4	1	1.19	5.95
5	4	4.76	10.71
6	13	15.48	26.19
7	8	9.52	35.71
8	8	9.52	45.24
9	15	17.86	63.1
10	18	21.43	84.52
11	13	15.48	100
Total	**84**	**100**	

La representación gráfica de estos datos es un histograma:



Podemos observar que el número de vecinos que visitaron el Ayuntamiento en los últimos tres meses varía entre un mínimo de dos y un máximo de once vecinos por día. También podemos concluir que en la mayoría de los días el número de vecinos que ha visitado el Ayuntamiento en los últimos tres meses es superior a seis.

4. Para calcular la media debemos sumar los valores y dividirlos por el número de observaciones. Tenemos: (2 + 11 + 5 + 7 + 6 + 7 + 4)/7 = 42/7 = 6. Así, el tiempo medio que tomó obtener una licencia de actividad es de 6 semanas.

Para calcular la mediana, primero debemos ordenar los datos en orden creciente:

2, 4, 5, 6, 7, 7, 11

Como el tamaño de la muestra es 7, se deduce que la mediana de la muestra es el cuarto valor más pequeño. El número mediano de la muestra de semanas que tomó obtener una licencia de actividad es m = 6 semanas.

Para calcular la moda hacemos una tabla de frecuencia y vemos que el número de semanas que más se repite es 7. La moda para obtener una licencia de actividad es de 7 semanas.

Semanas	**Frecuencia**
2	1
4	1
5	1
6	1
7	2
11	1
Total	**7**

5. El primer paso para probar la hipótesis de que el copago sanitario está relacionado con la valoración de los servicios es calcular las frecuencias esperadas:

Valoración del servicio	**Copago sanitario**		**Total**		**Valoración del servicio**	**Copago sanitario**	
	Sí	No				Sí	No
Malo	20	25	45		Malo	13.5	31.5
Bueno	10	45	55		Bueno	16.5	38.5

Valoración del servicio	Copago sanitario		Total		Valoración del servicio	Copago sanitario	
Total	30	70	100				

Posteriormente. se calcula el estadístico Chi cuadrado:

$$\chi^2 = \sum \frac{(Fo - Fe)^2}{Fe} = 3.13 + 2.56 + 1.34 + 1.10 = 8.13$$

Los grados de libertad son: (n-1) x (m-1) = 1 x 1 = 1

Mirando en la tabla Chi cuadrado obtenemos que la probabilidad de obtener un valor 8,13 o mayor con 1 grado de libertad es menor que $p = 0{,}005$. Podemos llegar a la conclusión de que el valor es estadísticamente significativo, pues es menor que 0,01. Con esta información podemos concluir que la valoración del servicio varía negativamente con el copago sanitario, esto es, quienes se ven afectados por el copago dan un menor valor al servicio que quienes no se han visto afectados por este.

Grados libertad	0.1	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75

Bibliografía

Crespo, I.; Anduiza, E.; Méndez, M. (2009). *Metodología de la ciencia política*. Madrid: Centro de investigaciones sociológicas («Cuadernos Metodológicos», 28).

Manual que explica detallada y muy claramente los conceptos de la metodología de investigación en las ciencias sociales. Sus ejemplos sobre temas de ciencias políticas explican claramente las definiciones y las aplicaciones prácticas de las teorías, los conceptos, las estrategias de investigación, los datos y la contrastación de hipótesis.

Glass, G. V.; Stanley, J. C.; Gómez, E. G.; Guzmán, E. (1986). *Métodos estadísticos aplicados a las ciencias sociales*. México: Prentice-Hall Hispanoamericana.

Manual clásico de estadística con explicaciones muy completas y detalladas sobre estadística descriptiva e inferencial. En sus quinientas páginas muestra los detalles de las técnicas y proporciona ejercicios para cada tema.

Hernández, J. J. C. (2007). *Conceptos básicos de estadística para ciencias sociales*. Madrid: Delta Publicaciones.

Libro introductorio que sirve como guía para elegir los métodos estadísticos apropiados para cada tipo de problema en las ciencias sociales. Interesante por su énfasis en aprender para qué sirve y cómo es posible aplicar la estadística.

Lind, D. A.; Marchal, W. G.; Wathen, S. A. (2017). *Estadística aplicada a los negocios y la economía*. México: McGraw-Hill / Interamericana Editores, SA de CV.

Manual de estadística muy completo y con un tratamiento riguroso de la teoría. Es muy detallado en las explicaciones y contiene ejemplos y ejercicios para desarrollar en Excel. Los objetivos de aprendizaje son útiles para programar el proceso de aprendizaje.

Meier, K.; Brudney, J.; Bohte, J. (2012). *Applied statistics for public and nonprofit administration*. Boston: Wadsworth, Cengage Learning.

Un libro indispensable para el estudio de la gestión pública por su enfoque intuitivo y sus ejemplos aplicados. Desarrolla los conceptos estadísticos sin entrar en detalles matemáticos complejos.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Nueva York: Routledge.

Un manual avanzado sobre estadística aplicada. Es un complemento para quienes quieran continuar aprendiendo nuevas técnicas que no cubre este material o quieran profundizar en los conceptos y métodos. Sus ejemplos y ejercicios se explican para el programa estadístico SPSS.