El análisis exploratorio de datos

PID_00270395

Albert Padró-Solanet i Grau

Tiempo mínimo de dedicación recomendado: 4 horas

Universitat Oberta de Catalunya

Albert Padró-Solanet i Grau

Profesor de los Estudios de Derecho y Ciencia Política de la UOC. Máster en Gestión Pública UAB. Máster en Ciencia Política UAB. Licenciado en Filosofía UAB.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Albert Padró-Solanet (2020)

Primera edición: febrero 2020

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

Índice

Introducción

Este módulo complementa el que se dedica a la metodología cuantitativa de la asignatura *Metodología de las Ciencias Sociales* del grado de Criminología. Por un lado, quiere ser una profundización y una reformulación de conceptos fundamentales del módulo como la conceptualización, la operacionalización y la medición para ayudar que sean utilizados conscientemente en la práctica de la investigación empírica. Por otro lado, es un repaso de los conceptos básicos del análisis exploratorio de datos que ayude al estudiante a refrescar y a descubrir los instrumentos que la estadística ha diseñado para ayudar a entender y dar sentido a los datos recogidos en la realidad empírica.

Objetivos

Con la lectura de este módulo el estudiante logrará los objetivos siguientes:

1. Entender la relación de la operacionalización y medición de las variables con la conceptualización teórica.

2. Distinguir el nivel de medición de las variables y entender sus condicionantes.

3. Entender los conceptos de fiabilidad y validez aplicados a las medidas de los conceptos criminológicos.

4. Reconocer que los instrumentos estadísticos descriptivos están asociados al nivel de medición de las variables.

5. Conocer los principales instrumentos de la estadística descriptiva univariante según el nivel de medición de las variables: los gráficos univariantes, las medidas de centralidad y medidas de dispersión de las variables.

6. Conocer los instrumentos que muestran la asociación entre diferentes variables en los diferentes niveles de medición.

1. La conceptualización y la operacionalización de las variables

1.1. La conceptualización

La primera tarea de toda investigación científica consiste en aclarar nuestras ideas de cómo funcionan los fenómenos que queremos estudiar. Es normal que los **conceptos** utilizados en el habla ordinaria (las palabras o los símbolos que utilizamos para referirnos a ideas o representaciones mentales) tengan una multiplicidad de significados diferentes que, a veces, incluso pueden ser contradictorios. El filósofo idealista Hegel veía en esta característica del lenguaje (sobre todo, el alemán) una virtud, puesto que recogía la dinámica dialéctica de la realidad. Pero, la ciencia empírica requiere que los conceptos que utiliza sean unívocos (tengan un único significado) y sean precisos para poder construir explicaciones de la realidad claras y que, posteriormente, permitan construir hipótesis que se puedan testar empíricamente. Si no sabemos con claridad qué es lo que esperamos, ¿cómo podremos saber si lo que pensábamos de la realidad es correcto o no? Se tiene que saber con claridad a que se refieren los conceptos con los que intentamos explicar la realidad que interesa a la criminología: ¿Qué es un delito? ¿Qué es reincidencia? ¿Qué es la desorganización social? ¿Qué es sentimiento de inseguridad? Etcétera.

Sobre el concepto de delito

En el lenguaje ordinario, el concepto de **delito** se puede usar de una forma vaga para referirse a una acción que va contra la convención; en un lugar un determinado comportamiento es delictivo y en otro, no.

Por ejemplo, para muchos sicilianos, las infracciones de tránsito realmente no son delitos puesto que, para ellos, no son normas sino recomendaciones.

Pero esta ambigüedad puede crear problemas para entender con precisión cuáles son los argumentos que se usan para explicar las razones de los delitos y los pronósticos. Cada investigación tiene que hacer este esfuerzo de conceptualización de los términos que utiliza. Sobre todo si se trata de una investigación que innova en relación con las conceptualizaciones que han hecho las investigaciones previas. Naturalmente, si la investigación no tiene como punto central la innovación conceptual tenderá a utilizar los conceptos desarrollados por las investigaciones previas, de forma que sea más eficiente y comparable con los trabajos anteriores.

En una investigación sobre los factores que afectan las sentencias por **delitos sexuales a menores** dictadas por las audiencias provinciales españolas entre 2011 y 2014, la conceptualización de delito no es problemática porque se refiere directamente a los artículos del Código Penal español que los tratan. De hecho, si adoptamos el punto de vista de la investigación, el Código Penal puede ser visto como un enorme esfuerzo para conceptualizar los comportamientos que atentan contra el bien público en una determinada comunidad política. Pero en otras investigaciones que quieran tratar de, por ejemplo, **delitos violentos** tenemos que hacer un ejercicio para aclarar de qué tipo de violencia se trata, puesto que, como dejan claro los manuales de retórica, la violencia, como cualquiera otro concepto en manos de la retórica, puede querer

decir cualquier cosa. En el límite, no saludar a la vecina podría ser considerada una conducta violenta.

Sobre el concepto de reincidencia

El concepto de **reincidencia** se refiere a la repetición de un tipo de conducta delictiva. La tasa de reincidencia es la proporción de un determinado tipo de delincuente (por ejemplo, los delincuentes que se encuentran en libertad provisional) que vuelve a cometer delitos en un tiempo determinado. Para calcular esta medida se utilizan los registros de detenciones o reingresos a la prisión, pero esta forma de medir el fenómeno es problemática porque hay una parte del comportamiento delictivo que no queda registrado por parte de la policía. Por lo tanto, se tienen que pensar formas específicas de hacerlo de acuerdo con los objetivos del tipo de estudio (Capdevila Capdevila & Ferrer Puig, 2009; Capdevila Capdevila, Ferrer Puig & Luque Reina, 2005). Algunos estudios utilizan la reincidencia autoinformada: la reincidencia que informan los mismos infractores para paliar el problema de la existencia de registros de buena parte de los delitos (naturalmente, fiarse de las confesiones de los delincuentes, aunque sea en entrevistas anónimas, también es problemático). Existen diferentes conceptos de reincidencia dependiendo del punto del sistema de justicia penal que quiera ser medido y evaluado. La reincidencia policial se refiere a una nueva detención; la reincidencia judicial se refiere a un nuevo procesamiento; la reincidencia penal a una nueva pena o medida cautelar; la reincidencia jurídica se refiere a un nuevo hecho delictivo del mismo título del Código Penal. El equipo de Eulàlia Luque utilizó el concepto de reincidencia penitenciaria:

«el reingreso en un centro penitenciario de personas que previamente han sido sometidas (al menos una vez) a una pena de prisión.»

La ciencia es una tarea colectiva. La comunidad científica se especializa en campos específicos y comprueba si las pruebas hechas por otros científicos son reproducibles y si son o no correctas. Intenta proponer explicaciones y pruebas mejores y que resuelvan problemas que piensen que no se han resuelto satisfactoriamente previamente. Este trabajo colaborativo también necesita claridad, univocidad y precisión en los conceptos.

1.2. La operacionalización: de los conceptos a las variables

Esta tarea de aclaración de los conceptos se denomina **conceptualización** y es previa a la tarea de definir cómo estos conceptos pueden ser **medidos empíricamente**. La tarea de definir la forma como se tienen que medir los conceptos se denomina **operacionalización**.

La operacionalización recoge las instrucciones que indican cómo se tiene que etiquetar, medir o identificar un concepto.

La operacionalización convierte los conceptos teóricos en variables empíricas de las que podemos tener diferentes indicadores.

Tabla 1. Conceptos, variables e indicadores

Concepto	Variable	Indicador
Estabilidad residencial	Cantidad de cambio en la población de un vecindario	Tasa de cambio de población en un año = (nuevos residentes + residentes migrados)/total población vecindario
Heterogeneidad étnica	Diversidad de la composición étnica de un vecindario	% de población no ciudadana española en un vecindario
Control local	Capacidad de un vecindario de supervisar los propios miembros o forasteros	Tasa de pertenencia de los vecinos a asociaciones formales e informales (a partir de la pregunta de encuesta: ¿A qué asociaciones del vecindario pertenece?)

Sobre el concepto de desorganización social

Kornhaurser (1978: 63) define el concepto de **desorganización social** como la situación que «existe en primera instancia cuando la estructura y la cultura de una comunidad es incapaz de implementar y de expresar los valores de sus propios residentes.» El concepto de desorganización social quiere capturar la idea de que existen comunidades humanas, vecindarios, que estructuralmente no son capaces de combatir la delincuencia y lograr la aspiración a un mejor entorno. A una comunidad desorganizada le falta todo lo que caracteriza a una comunidad organizada:

1) solidaridad o un consenso sobre normas y valores esenciales (los residentes valoran las mismas cosas, como la ausencia de delincuencia);

2) cohesión, o un vínculo fuerte entre los vecinos (los vecinos se conocen y se valoran entre ellos);

3) integración, la interacción social regular.

La intuición detrás de este concepto puede ser clara, pero la operacionalización para comprobar la teoría es compleja. Por ejemplo, Sampson y Groves (1989) crearon índices en los barrios que tuvieran en cuenta el estatus socioeconómico, la heterogeneidad étnica, la movilidad residencial, la disrupción familiar y la urbanización, de la misma forma que medidas de desorganización social, etc., para comprobar la relación con las tasas de criminalidad. En estos casos tan complejos, la definición de cómo se operacionaliza el concepto de desorganización social, es una forma de definirlo (Kubrin & Wo, 2015).

El caso de la **justicia penal** es especial, porque la operacionalización del delito es lo que hace el sistema de justicia penal. Los jueces y magistrados son el equipo entrenado para atribuir a cada denuncia unos valores específicos del Código Penal a cada sentencia que pronuncian. Después de hacer el juicio y escuchar a las partes, fiscales y abogados de la defensa, se dice si se ha cometido un delito y, si es así, qué tipo de delito se ha cometido y se valoran las circunstancias en las que se ha cometido, de aquí se derivan las penas que se imponen a los acusados y las compensaciones a las víctimas. La existencia de todo este enorme aparato de operacionalización y de medición hace que parezca, desde el punto de vista de la investigación en criminología y en la justicia penal, que la operación de medición en este caso es trivial, pero esta trivialidad es aparente, porque los costes de medir correctamente han sido asumidos previamente por un conjunto de organizaciones entrenadas y especialistas en medir este delito de forma válida y fiable: cuerpos de seguridad,

juzgados y personal legal. Precisamente la pregunta que se hace respecto del funcionamiento del sistema de justicia penal es sobre la validez y fiabilidad de las medidas que resultan.

Por ejemplo, uno de los primeros estudios sobre las sentencias pidió a diferentes jueces que resolvieran un mismo caso y observaron las amplias diferencias que existían en las sentencias que pronunciaron.

A veces, la mejor estrategia para resolver los problemas de conceptualización, de aclararlos, hacerlos unívocos y menos abstractos consiste precisamente en definir el concepto a través de su operacionalización; por eso se habla de la **definición operativa de los conceptos**. La operacionalización es como **una receta de cocina** que puede seguir cualquier otro investigador para comprobar que se pueden obtener los mismos hallazgos de una investigación. La operacionalización está conectada con el enfoque teórico, es una parte esencial de la metodología y permite contrastar si se mejora el conocimiento de los fenómenos que queremos estudiar; por lo tanto, forma parte de una estrategia para permitir la acumulación del conocimiento científico en un área.

Cuando el investigador **operacionaliza** un concepto en **variables** nos dice de qué forma se tiene que medir el concepto teórico en las **unidades de análisis, observaciones, individuos, elementos o casos**. A veces estas unidades de análisis son personas que son entrevistadas en una encuesta. La operacionalización de la mala conducta en los presos consiste en la formulación de la pregunta que se realizará para medir esta mala conducta.

Por ejemplo: '¿alguna vez ha atacado algún otro preso que no le había agredido antes?'.

Otras veces son agregados de personas y la información que obtenemos se refiere a estos agregados.

Un ejemplo de agregado puede ser una prisión. El concepto de conflictividad en una prisión puede ser operacionalizado a través del porcentaje mensual de internos que son tratados en la enfermería por lesiones traumáticas en relación con el total de internos.

Inevitablemente, la operacionalización forma parte de esta decisión de diseño de investigación sobre cuáles son las unidades de análisis. Los investigadores se juegan buena parte de la relevancia de su investigación cuando eligen unas unidades u otras de análisis para testar sus teorías o explicaciones de la realidad.

Problema del muestreo

El problema de decidir cuáles son los casos adecuados para responder a las preguntas que formula una investigación se conoce como el *problema del muestreo*.

2. El nivel de medición: ¿qué precisión tenemos en la medida de las variables?

Cuando medimos los conceptos teóricos en las unidades de análisis (o casos, individuos...) lo podemos hacer con diferentes grados de precisión. Este grado de precisión se denomina **nivel de medición**.

El nivel de medición puede ser **cualitativo** o **cuantitativo**. En la medición cualitativa, las mediciones que atribuimos a las unidades de análisis (los valores de las variables en cada unidad de análisis) no pueden ser interpretadas como una expresión cuantitativa de la presencia del atributo en la unidad de análisis.

Por ejemplo, cuando una persona se pone en una balanza y vemos que pesa 60 kg, esta característica es numérica. Sabemos que esta persona pesa la mitad de otra persona que pesa 120 kg y el doble de otra que pesa 30 kg.

Este es un ejemplo de **medición cuantitativa**. Pero no siempre tenemos este mismo nivel de precisión en las mediciones.

Por ejemplo, podemos clasificar a los mismos individuos que antes hemos pesado en hombres y mujeres. Esta variable es bastante diferente de la anterior. Está claro que el número con el que identificamos cada caso en nuestra matriz de datos (1, mujer; 2, hombre) no tiene ningún significado auténticamente numérico: un hombre no es lo mismo que dos mujeres (¡por muy machista que sea!).

Cuando los identificamos con estos valores solo estamos diciendo que cada individuo **pertenece a una clase**.

Las mediciones dependen de las operacionalizaciones de los conceptos y estas del papel de los conceptos en las explicaciones que se quieren testar. Hasta hace unos años, la clasificación diádica (en dos valores: hombre y mujer) parecía «natural» e «inevitable». Los caracteres sexuales primarios servían para clasificar a la población en dos categorías, de acuerdo con su capacidad de tener hijos y alimentarlos. Hoy en día, en muchos entornos culturales y políticos se ve como una reducción artificial de la realidad del género y la clasificación es como mínimo triádica, que permite que haya parte de la población que no se quiera clasificar en ninguna de las categorías anteriores. En una investigación que utilizara el concepto psicosociológico de «femineidad», no se conformaría en una clasificación simple, ni siquiera triádica, y podría utilizar una batería de preguntas para establecer el grado o la intensidad de femineidad de los individuos de la muestra, de forma que la medida se convirtiera en cuantitativa.

En general, el nivel de medición no es una cuestión «natural» e «inevitable» del concepto medido. Depende de cuál es el grado de precisión requerida para responder a las preguntas de nuestra investigación, de nuestra capacidad para capturar la información y de los recursos disponibles.

Los niveles de medición forman una **escala**. Cada nivel superior nos ofrece un grado de información más preciso respecto de las variables que queremos medir en nuestros casos.

2.1. Nivel de medición cualitativo nominal

El **nivel de medición nominal** o **categórico** solo nos permite clasificar a los casos en diferentes grupos. Este nivel de medición es el del género en la discusión anterior: mujeres y hombres.

Para obtener una clasificación se requieren reglas sistemáticas para distinguir entre un fenómeno de otro y establecer criterios para determinar las fronteras entre las diferentes clases o los diferentes grupos.

Por ejemplo, como hemos ido comentando, el Código Penal es una herramienta de clasificación de los diferentes tipos de delitos que ofrece estas reglas sistemáticas para distinguirlos y marcar fronteras entre ellos. Otro ejemplo de variable nominal es la población de residencia de los individuos de una muestra. En una investigación donde esta pregunta fuera importante (porque, por ejemplo, se quieren conocer los niveles de seguridad percibida de los residentes) es posible que se tengan que establecer reglas claras para resolver posibles casos ambiguos. ¿Cómo se tiene que clasificar a alguien que reside entre semana en una población y solo el fin de semana vuelve a la residencia donde está legalmente registrado? Todavía otro ejemplo, un caso que ha entrado en el sistema judicial puede estar: archivado, sobreseído, en trámite, absuelto o condenado.

Desde el punto de vista lógico-matemático, la clasificación utiliza el **principio de identidad**: los casos de la misma categoría son **idénticos** en relación con la propiedad que se mide y son **diferentes** de los casos que están en las otras categorías.

Cada miembro de la categoría «mujeres» es idéntico en su «femineidad» que todos los otros miembros de la categoría o clase. Del mismo modo, un residente del barrio de Sant Roque, es idéntico en su residencia a todos los otros residentes del barrio de Sant Roque y es diferente de los residentes de los otros barrios de la muestra.

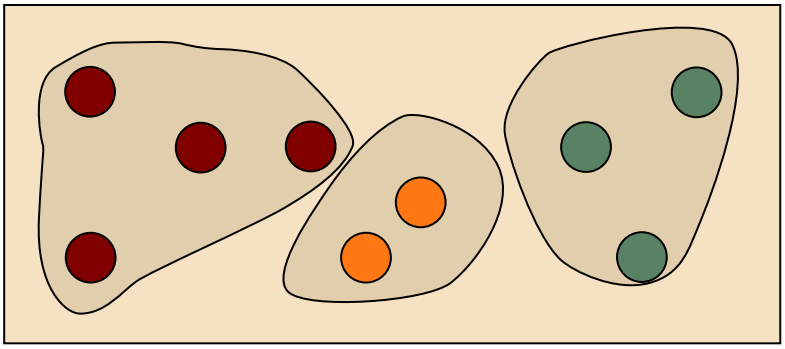
Formalmente, si hay dos categorías A y B: $A \neq B$

El conjunto de las categorías en el nivel de medida nominal tienen que cumplir dos propiedades lógicas:

1) Las categorías **tienen que ser mutuamente excluyentes**. Esto quiere decir que no pueden existir miembros que pertenezcan simultáneamente a dos categorías. Si eres hombre, no puedes ser mujer. Si eres residente de Sant Roque, no eres residente de San Juan. No puede haber solapamiento entre las categorías.

2) Las categorías **tienen que ser colectivamente exhaustivas**. Es decir, todos los miembros de la muestra tienen que estar en alguna categoría, no puede quedar ninguno que no esté clasificado en una categoría. Entre todas las categorías se engloban todos los casos de la muestra.

Figura 1



La **dicotomía**, que nos dice si un individuo forma parte de una categoría o no, es el tipo originario de las clasificaciones más complejas. La variable binaria de sexo (hombre o mujer) la podemos reducir a dos dicotomías (hombres; no hombres) o (mujeres; no mujeres). Lógicamente, cualquier clasificación puede ser reducida a un conjunto de dicotomías.

El nivel de medición nominal nos ofrece un conocimiento limitado de muchos fenómenos que queremos estudiar. Esto también implica una limitación en el tipo de análisis estadísticos posibles con esta información. El nivel nominal se encuentra en el nivel inferior de la escalera de medición. Por encima, pero todavía en el nivel cualitativo, se encuentra el nivel ordinal.

2.2. Nivel de medición cualitativo ordinal

En el nivel de medición ordinal se añade un elemento al nivel nominal: existe un orden entre las categorías. Cada elemento está solo en una categoría y entre todas las categorías engloban todos los elementos (las categorías son mutuamente excluyentes y colectivamente exhaustivas). Pero ahora podemos decir que hay una categoría que es más que todas las otras en la propiedad que se mide; después, hay otra que es menos que la categoría anterior en esta propiedad o variable, pero con más propiedad que las otras... y así hasta completar todas las categorías. Existe un orden.

Si tenemos cuatro categorías, podemos ordenarlas de acuerdo con un criterio de más/menos.

- Por ejemplo, los delitos del Código Penal pueden ser clasificados por orden de seriedad. En las investigaciones criminológicas existen muchas medidas que se recogen en las variables del nivel ordinal.
- Por ejemplo, el nivel de educación se puede recoger como: 1. «Ninguna educación formal», 2. «Educación primaria», 3. «Educación secundaria», 4. «Educación superior». En cada categoría hay un contacto más grande con el sistema educativo, pero lo medimos con unas categorías amplias, donde los valores numéricos atribuidos a cada categoría no indican una cantidad, sino que simplemente expresan un orden.

- Típicamente, muchas de las respuestas a las preguntas de las encuestas de opinión están medidas en el nivel ordinal (las denominadas escalas de Likert). Por ejemplo, las opciones de respuesta a la pregunta «¿En qué medida está de acuerdo con la afirmación 'mi barrio es seguro'?»: 1. Muy en desacuerdo; 2. En desacuerdo; 3. No lo sé; 4. De acuerdo; 5. Muy de acuerdo. Por los valores de las categorías solo sabemos que los que responden 1 están menos de acuerdo con la afirmación que los de la categoría 2, pero, aparte de esto, no podemos ser más precisos. No podemos decirlo con una cantidad numérica.

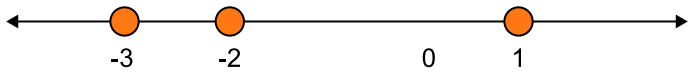
Si disponemos de más información sobre la presencia de una propiedad en las unidades de análisis podemos subir a los niveles de medición cuantitativa de la escala.

2.3. Nivel de medición cuantitativo de intervalo

El **nivel de medición cuantitativo de intervalo** se utiliza cuando no solo somos capaces de clasificar nuestros casos (gente o acontecimientos), sino que podemos establecer diferencias de grado entre estos casos respecto de la propiedad que estamos midiendo.

Una escala de intervalo requiere que los intervalos medidos tengan una misma unidad de medida y que, por lo tanto, la diferencia en la escala entre 2 y 1 sea la misma diferencia que entre 2 y 3; y la diferencia entre 1 y 3 sea dos veces la diferencia entre 1 y 2 o entre 2 y 3. No solo decimos que existe un orden. La distancia lógica entre atributos o propiedades puede ser expresada de forma significativa por intervalos estándares.

Figura 2



Las medidas de intervalo que normalmente se utilizan en las ciencias sociales consisten en medidas que provienen de índices compuestos y estandarizados, como el popular coeficiente de inteligencia.

En criminología, podríamos obtener este tipo de índices o escalas –medidas cuantitativamente a nivel de intervalo– en medidas de peligrosidad o de riesgo de reincidencia en jóvenes. Se podrían basar en la agregación de una cierta cantidad de ítems: respuestas a cuestionarios que miden la impulsividad junto con evaluaciones de los factores de riesgo o vulnerabilidad social. El resultado final sería una medida en la que sabríamos con precisión que, por ejemplo, los valores más elevados nos indican a personas con más riesgo de no lograr una rehabilitación y, por lo tanto, se podrán establecer los protocolos adecuados para tratarlos.

La diferencia entre las mediciones de nivel cuantitativo de intervalo respecto del nivel de medición siguiente, el nivel cuantitativo de razón, es que, en el nivel de intervalo no se conoce el nivel cero de presencia de la propiedad o atributo.

En el ejemplo anterior del índice de factores de riesgo, sabemos que cada vez que nos desplazamos en el índice en una unidad, estamos avanzando hacia un riesgo más grande o retrocediendo hacia un riesgo más pequeño (normalmente, en estos índices el valor cero es el valor más frecuente en la población). Lo que no sabemos es qué valor tiene el índice para el valor 0 de riesgo real.

Esta cuestión, que es importante desde el punto de vista conceptual, también tiene consecuencias en el tipo de operaciones matemáticas con sentido que se pueden realizar con estas medidas. En este tipo de medidas solo se pueden hacer sumas y restas, no multiplicaciones ni divisiones.

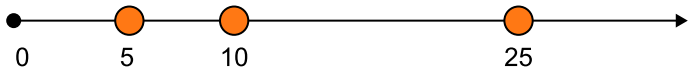
El ejemplo más típico de las escalas de intervalo son las escalas de temperatura en grados Fahrenheit o en grados Celsius. Las dos escalas son de intervalo: no tienen el 0 en el mismo lugar, ni las distancias entre grados son las mismas. Pero se puede pasar de una a otra con una fórmula matemática, una transformación lineal, clara: ºF = ºC 1,8 + 32. En las dos escalas, sabemos que si sube la temperatura 10º (entre 0º y 10º, digamos) el incremento de temperatura habrá sido de los mismos grados que si sube entre 10º y 20º. La suma y la resta en estas escalas no es un problema. El problema aparece cuando se quiere multiplicar o dividir: ¿20 ºC de temperatura es el doble de temperatura que 10 ºC? No, porque 0 ºC no es la ausencia total de calor. En la escalera absoluta de temperatura Kelvin, donde el 0 sí que es la ausencia de calor, los 0 ºC son 273 K y 10 ºC son 283 K; por lo tanto, el doble de temperatura de 10 ºC serían 566 K, ¡y no 20 ºC!

2.4. Nivel de medición cuantitativo de razón

El nivel de medición escala de razón se encuentra en el escalón más alto de la escala de medición.

El nombre de escala de razón significa que las distancias entre las categorías, no solo tienen sentido en los intervalos, sino también en las **proporciones**: el valor de 10 es el doble de 5 y este es 5 veces más pequeño que 25.

Figura 3



Muchas de las medidas de la criminología están medidas con el nivel cuantitativo de razón.

Por ejemplo, el número de veces que un delincuente ha sido arrestado previamente, se mide en la escala cuantitativa de razón. Si alguien ha sido arrestado 3 veces, ha sido 3 veces menos arrestado que otro que ha sido arrestado 9 veces.

Esta comparación es posible porque en esta variable el valor cero no es arbitrario. El número de actos delictivos en las encuestas de autoinforme delictivo o el número de victimizaciones en las encuestas de victimización también son medidas en este nivel. Otras medidas como la edad medida en años o la renta personal, también son variables cuantitativas de razón.

Del mismo modo, los valores de las variables que son el agregado de acontecimientos o de características de los individuos que los forman en diferentes unidades geográficas (distritos, municipios, comarcas, provincias, regiones, estados...) también se miden con la escala cuantitativa de razón.

Por ejemplo, la tasa de criminalidad o la tasa de reincidencia tienen este nivel de medida.

Observación

Notad que variables que en el ámbito individual normalmente son medidas con el nivel cualitativo nominal (por ejemplo, la etnia o el género), en los agregados se convierten en variables de escala de razón: por ejemplo, la tasa de masculinidad o el porcentaje de población gitana. Esto confirma, desde otro punto de vista, que el nivel de medición depende de la estrategia de investigación que elegimos; que el nivel de medición no es una cosa intrínseca de la naturaleza del concepto que queremos medir. No es una cuestión sustantiva, sino analítica.

2.5. Implicaciones del nivel de medición en el tipo de tratamiento

Para llevar a cabo los análisis estadísticos, conocer el nivel de medición de las variables es crucial para saber qué tipos de análisis tienen sentido. Las técnicas específicas para describir y para relacionar las variables dependen de su nivel de medición.

Claramente, con las variables de tipo nominal, calcular medidas de resumen como la media es totalmente absurdo. Pero a veces nos encontramos en situaciones donde variables que realmente están medidas en el nivel ordinal son tratadas como si fueran variables de tipo cuantitativo de intervalo, especialmente si hay muchos niveles de la escala ordinal. Este tipo de práctica es más frecuente en las investigaciones más antiguas, que se llevaron a cabo cuando todavía no se habían desarrollado herramientas específicas para tratar variables de tipo cualitativo, ya que estas herramientas se retrasaron respecto de las técnicas orientadas a variables cuantitativas o numéricas típicas de ciencias sociales tempranas y muy influyentes, como la economía. Si se tiene en cuenta lo que se está haciendo y se avisa al lector del análisis, una práctica de este estilo puede ser aceptable. Naturalmente, siempre es mejor utilizar las técnicas que son más adecuadas para cada nivel de medición.

Por un lado, en las variables de tipo numérico o cuantitativo se puede distinguir entre variables que solo pueden tener valores discretos (típicamente en las variables de recuento) y otras variables que pueden tener valores continuos. Esta cuestión no es muy relevante, pero puede provocar expresiones curiosas del tipo 'los hechos delictivos medios en una población son 3,7 por 1.000'. Por otro lado, se tiene que tener muy claro que las variables numéricas con valores discretos no se tienen que confundir con las medidas con el nivel ordinal.

En general, siempre es conveniente en la investigación medir los conceptos en el nivel de medición más elevado posible. Si a la hora de presentar o de analizar los datos nos damos cuenta de que hemos recogido demasiada información, siempre podremos disminuir el nivel de medición, agrupando los valores y convirtiendo las variables numéricas en variables ordinales.

A veces, la conveniencia de disminuir el nivel de medición no es debido a la necesidad de simplificar sino de la misma naturaleza del concepto que queremos medir.

Por ejemplo, la madurez de una persona podemos medirla con los años. Pero si estamos interesados en la experiencia seguramente la edad no será una buena medida (una medida válida) o una medida suficiente. Quizás es mejor componer la edad con la experiencia vital para clasificar mejor a los individuos. Por ejemplo, categorizar como **jóvenes** aquellos individuos que no han superado los 30 años y que todavía no han trabajado de forma remunerada; mientras que se clasifican como **adultos** los individuos que superen los 25 años y que ya hayan trabajado de forma remunerada. Aquí estaríamos disminuyendo el nivel de medición para tener mayor precisión para capturar el concepto de madurez.

3. La calidad de las medidas: la fiabilidad y la validez

3.1. La fiabilidad

Una medida es **fiable** si produce el mismo resultado cuando el proceso de medición se repite.

Ejemplo

Por ejemplo, si nos pesamos cada mañana antes de ducharnos y cada vez tenemos un peso bastante diferente –porque resulta que tenemos una báscula no muy buena y el peso varía según el equilibrio de los dos pies en el plato– no tendremos una medida fiable de nuestro peso (también podemos decir que nuestro instrumento de medida, la báscula, no es fiable).

Otro ejemplo cotidiano

¿Por qué se recomienda medir la fiebre de los niños pequeños con un termómetro en el ano mejor que con el termómetro en la boca? Pues, precisamente porque una medida es más fiable que la otra. No es seguro que el niño mantenga quieto el termómetro bajo la lengua durante todo el tiempo de tomar la medida.

En las ciencias sociales hay muchas fuentes que entorpecen la fiabilidad en la medición de los datos empíricos. Veamos dos ejemplos con datos de registro y con datos de encuestas.

En el campo de los registros oficiales, en un registro policial de los delitos cometidos puede haber diferentes razones que lleven a incorrecciones. Puede haber errores en la codificación de los datos. Pero también puede haber problemas de saturación del trabajo en algunos momentos, de forma que a veces no queden correctamente registrados todos los delitos. Es posible que un mismo delito pueda ser clasificado de forma diferente según los criterios que utilice el encargado de entrar los datos. El resultado es que la medida de los delitos puede no ser fiable.

Para comprobar la fiabilidad de las medidas se usan diferentes procedimientos.

- En el **procedimiento de test-retest**, repite la medición una segunda vez para ver hasta qué punto las medidas tomadas en un primer momento cambian cuando se vuelven a medir a los mismos individuos.
- En la **prueba de fiabilidad de división por la mitad** (Split-Half Check) se divide la muestra y a cada parte se le aplica una medición diferente de un mismo concepto; las diferencias entre los resultados obtenidos servirán de medida de la fiabilidad de estas medidas. Este método se desarrolló para medir la fiabilidad de los test psicológicos. Los test se dividen en dos partes, cada parte se pasa a una mitad de la muestra y se ve si los resultados de los test tienen el mismo grado de correlación, indicando que están midiendo una misma característica.

3.2. La validez

Mientras la fiabilidad hace referencia a la precisión con la que una medida empírica tiende a recoger un determinado concepto, la **validez** se refiere a si la medida realmente mide el concepto que se quiere medir.

Hipotéticamente, el problema de una falta de fiabilidad lo podríamos resolver realizando muchas mediciones y calculando la media. Naturalmente, una medida que no es fiable no es válida, porque nos puede dar unos valores que se encuentran alejados de los auténticos valores de la variable. Pero al mismo tiempo, puede haber medidas que sean fiables y que no sean válidas. Es decir, puede haber medidas que tienden a darnos siempre los mismos resultados (son fiables), pero no son válidas porque no varían de acuerdo con el concepto que se cree que miden. Para explicar esta relación entre validez y fiabilidad se usa la imagen de la capacidad de acertar una diana de un tirador (la imagen ya fue utilizada en un comentario de Galtung sobre el error aleatorio de medición). Un tirador que no tiene pulso es lo mismo que una medida que no es fiable: sus disparos están dispersos porque un temblor o una tensión no controlada en el momento de disparar hacen que el disparo se desvíe de la cota.

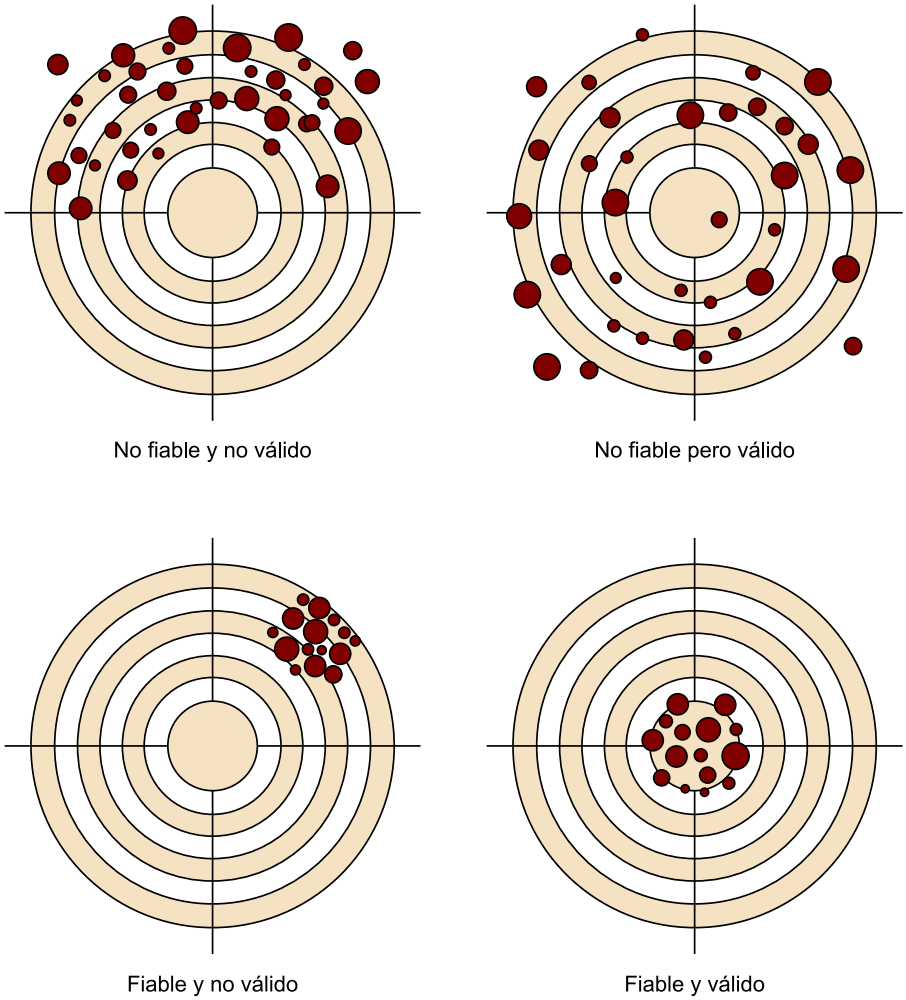
En la **figura 4** vemos en la primera fila los disparos de un tirador que no es fiable: en los dos blancos los disparos están dispersos. En el blanco de la derecha, los disparos se distribuyen alrededor de la diana. Los disparos apuntaban el objetivo –eran «válidos», en este sentido–, pero no eran fiables. En el blanco de la izquierda el tirador tampoco es fiable, pero, lo que es peor, no apunta correctamente a la diana (quizás la mira de la escopeta está desviada). En las imágenes de abajo tenemos los resultados de un buen tirador: es preciso en las dos dianas; pero apunta correctamente a la derecha y de forma desviada a la izquierda.

Una **medida válida** tiene que tener las dos propiedades: ser precisa (**fiable**) y medir lo que se supone que tiene que medir (**válida**).

El tirador de la derecha de la primera fila no es un buen tirador, aunque apunte correctamente a la diana: una medida que es válida pero no es fiable, no es una buena medida.

La distinción entre **error de medida sistemático** y **error de medida aleatorio** sirve para explicar esta relación. En los gráficos de la primera columna tenemos errores sistemáticos a la hora de medir el concepto que nos interesa: en los dos casos no estamos apuntando al concepto que queremos medir. El segundo gráfico de la primera fila muestra solo un error aleatorio. Si podemos medir un número suficientemente grande de veces, podremos descontar este error.

Figura 4. Relación entre fiabilidad y validez



La existencia de **error aleatorio** es el origen del enfoque estadístico. En las ciencias sociales, el error aleatorio no solo se relaciona con el error de medida producto del instrumento de análisis (como pasa en las ciencias físicas), sino con la complejidad inherente de los fenómenos que estudian: hay muchos factores que simultáneamente afectan a los valores de las diferentes variables. El **error sistemático** existe cuando algunos de estos factores tienen un impacto en la medida que no es neutro. Naturalmente, las cuestiones de los errores son relativas a los conceptos que quieren ser medidos (la diana del blanco).

Siguiendo el ejemplo anterior, cuando pedimos que la gente se ubique en una escala ideológica, es posible que no estemos midiendo correctamente las preferencias ideológicas, es decir, tenemos un error sistemático, producto de la crispación política. Pero, precisamente por esta razón, ¡esta medida podría convertirse en una buena medida de la crispación política!

La relación entre el concepto teórico y la medida se supone teóricamente. La medida, por ella misma, no es válida o inválida. Lo es por la interpretación que hace de ella el investigador.

Un ejemplo de problemas de validez de una medida es el de usar el número de denuncias de maltrato de género para medir el número de maltratos. A partir del momento en que el tema del maltrato fue reconocido públicamente como un achaque, se observó que el número de denuncias creció, en parte por las facilidades para hacer esta denuncia (esto es lo que buscaban este tipo de medidas que querían ayudar a desvelar la violencia que pasaba desapercibida y ayudar a tomar medidas preventivas). Desde este punto de vista,

ahora se disponía de una mejor medida de la violencia de género. El problema fue que, tal como declaró la fiscal jefe de Cataluña, una vez reconocido el delito de maltrato, este se pasó a argumentar con mucha más frecuencia en los casos de divorcio porque, de este modo, el abogado que lo esgrimía tenía una ventaja ante el cónyuge. El resultado es que, por esta interferencia, ya no estábamos seguros de tener una mejor medida de la violencia de género.

Para evitar incurrir en problemas de validez siempre se tiene que estar prevenido de la posibilidad que estos se estén produciendo, tanto en la construcción y utilización de los instrumentos de medida de las variables como para interpretar las medidas. Hay que pensar continuamente de qué manera algunos factores pueden interferir en la medida del concepto teórico: ¿cómo podría ser invalidada una medida?

Pero, además de este consejo a la precaución, ¿cómo se puede comprobar la validez de las medidas? En algunos campos se hacen estudios de validez de forma sistemática.

- Por ejemplo, en el campo de la salud se hacen estudios de validación de los diagnósticos realizados en el sistema de salud. Se realizan autopsias a una muestra de pacientes de hospitales para determinar las causas reales de las muertes y las comparan con como los habían diagnosticado. Así es posible comprobar si funcionan correctamente las nuevas herramientas de diagnosis o, también, es posible medir cuál es el porcentaje de casos que fueron tratados de forma incorrecta.
- En los estudios poselectorales hechos en Estados Unidos y en Reino Unido se han hecho comprobaciones de validez de la variable de participación. Como los registros de participación son públicos, se comprueba si la gente que ha sido encuestada ha hecho lo que ha dicho que hizo: votar o no votar. De este modo se puede tener un perfil de los grupos de ciudadanos que más mienten a la hora de decir si han votado o no (¡entre un 23 y un 28 % de los que dicen que han votado en las encuestas, en realidad, no lo han hecho!).
- De forma parecida, en las encuestas preelectorales las respuestas son reanalizadas una vez que se han hecho las elecciones, para ver qué grupos de votantes tienden a esconder su intención de voto (lo que se denomina popularmente el voto oculto). Con esta información las empresas de opinión pueden mejorar sus predicciones en las posteriores elecciones. Naturalmente, estos métodos de ajuste son válidos mientras el entorno político sea estable, en el momento en el que cambia, todas estas referencias previas y correcciones de las medidas se pueden volver en contra.

Estos son ejemplos de **validación externa** de las medidas de las variables. En la validación externa se intenta que alguna prueba del mundo real garantice que se está midiendo correctamente el concepto que queremos medir. Si comprobamos el grado en el que nuestra variable nos permite obtener medidas del mundo real, podremos denominar esta prueba de validez como validez pragmática o predictiva. No siempre se encuentran medidas externas de control de validez. Una solución es buscar otras medidas que se supone que están relacionadas con el mismo concepto que queremos medir y así se ve si las dos medidas están correlacionadas. El hecho de que las dos estén correlacionadas, es una pista para suponer que la medida que hemos hecho es una medida correcta del concepto que intentamos medir. Este tipo de prueba de validez puede ser llamada validez interna o construida y sirve de base para la construcción de medidas de conceptos con indicadores múltiples.

Por ejemplo, podemos tener tres medidas diferentes de la existencia de corrupción política en un país: la primera, la percepción de corrupción entre los trabajadores públicos según los hombres de negocios (esta es la base del índice de Transparency International); la segunda, puede ser el número de casos de corrupción destapados por los medios de comunicación; y, finalmente, la tercera medida puede ser la discrepancia entre los costes presupuestados y los costes efectivamente realizados en las obras públicas. Las tres medi-

das se supone que están relacionadas con el nivel de corrupción política en un país, pero las tres miden el concepto desde puntos de vista y con métodos muy diferentes. Por lo tanto, no esperamos que tengan exactamente los mismos valores, pero sí que esperamos que, en tanto que miden el mismo concepto de corrupción, estén mínimamente correlacionadas. Si las tres medidas fueran en conjunto en un mismo sentido (hubiera suficiente correlación entre ellas), tendríamos más confianza en que estamos midiendo un concepto difícil de forma correcta. Incluso, podríamos plantearnos construir un índice que agregara y resumiera las diferentes medidas en una única. Si, en cambio, alguna de estas medidas fuera muy discrepante, nos veríamos obligados a revisarla e incluso descartarla, porque **no es** una **medida válida** del concepto.

Finalmente, la medida más general de validez de las variables es la **validez aparente** (*face validity*), que es un término para decir que una medida es correcta porque nos lo parece como investigadores especializados en un fenómeno.

Vemos los valores de la medida y los casos a los que corresponden y nos parece que son correctos; ordena a los casos de la forma como esperamos que estén ordenados. Naturalmente, no se trata de una prueba muy fuerte, puesto que solo vale en cuanto que los lectores estén de acuerdo con ella.

Por ejemplo, en cuanto a las puntuaciones de los índices de Transparency International vemos que Italia o España quedan por debajo de los países escandinavos en percepción de corrupción, pero por encima de Nigeria o México; como nos parece que esta ordenación es verosímil, podemos decir que el índice es **aparentemente válido**.

Los problemas potenciales de validez y de fiabilidad de nuestras medidas tienen que aparecer en cualquier investigación.

4. El análisis exploratorio de los datos

4.1. La descripción univariante de los datos

Antes de medir las relaciones entre las variables, es conveniente hacer una descripción univariante de las variables que se utilizan en la investigación. Las descripciones nos permiten entender cuál es la estructura de las variables y nos servirán para entender que se produzcan algunas relaciones entre diferentes variables. Por otro lado, las descripciones univariantes son muy importantes como control de la información de nuestra base de datos, puesto que nos permiten detectar posibles errores en las codificaciones de las variables.

Para el investigador siempre es crucial visualizar las distribuciones univariantes de sus variables.

La estadística nos proporciona herramientas para **resumir** la información de las distribuciones. Sirve para ver los bosques (características generales del conjunto de datos) que hay detrás de los árboles (los casos individuales). Sirve para evitar que los casos que llamen más la atención –por la razón que sea– influyan desproporcionadamente en la valoración del grupo. Es un control para evitar equivocarnos y llegar a conclusiones erróneas en relación con una información numérica numerosa.

El enfoque exploratorio de datos pone el énfasis en conocer el máximo posible de cosas a través de los datos. Cuanto más se sepa de los datos, más útiles serán para desarrollar la teoría. Dos principios tienen que regir la exploración de los datos:

1) Primero, se tiene que ser escéptico hacia las medidas de resumen. Ninguna de ellas puede sintetizar completamente la información contenida en las variables.

2) En segundo lugar, se tiene que ser abierto de miras respecto de lo que nos pueden enseñar los datos. No tenemos que dar por supuesto que conocemos cómo son los datos. Muchas veces pueden descubrir cosas inesperadas, nuevas, respecto de las características o explicaciones de los fenómenos. Los anglosajones de esta capacidad de hacer descubrimientos por azar, inesperadamente, la llaman *serendipity* y es una calidad importante en la investigación científica tener esta disposición a estar abierto ante las nuevas ideas o los factores que pueden ayudar a explicar los fenómenos de interés.

Cuando se describen estadísticamente las variables, hay tres cuestiones que nos interesan para sintetizar las distribuciones:

- ¿Cuáles son los valores más característicos donde se encuentran mayoritariamente los casos?
- ¿En qué medida son característicos o representativos estos valores?
- ¿Cuál es la forma de la distribución de los datos?

1) En términos estadísticos, el **resumen** de una variable es la medida de **centralidad** o de localización alrededor de la que se encuentran los casos. Es el valor más representativo de la distribución de datos de una variable.

2) La **dispersión** sirve para responder a la pregunta sobre hasta qué punto los valores centrales son representativos. Si la mayoría de los casos no son como los centrales, si no se encuentran agrupados y más bien están repartidos entre todos los valores que puede adquirir la variable, los estadísticos de centralidad nos dirán poca cosa de la distribución que queremos conocer.

3) Finalmente, tenemos la **forma** de las distribuciones. ¿Cómo se distribuyen los valores alrededor de los valores centrales? ¿Se trata de una distribución simétrica o asimétrica? ¿Se trata de una distribución más bien uniforme y rectangular o hay una moda? ¿Hay más de una moda?

Los gráficos

Los **gráficos** nos ofrecen una **descripción más completa de las distribuciones** que los estadísticos de centralidad y dispersión. Para extraer el máximo de información de las variables, los análisis exploratorios ponen énfasis en las representaciones gráficas. Nos dan simultáneamente una aproximación a los tres aspectos importantes de una distribución: forma, dispersión y localización. Parece un tópico el proverbio chino que dice que una imagen vale más que diez mil palabras, pero en el análisis exploratorio de datos representar gráficamente los datos es prescriptivo.

Los estadísticos para resumir las variables están asociados al nivel de medición de estas variables. En la **tabla 2** se listan los estadísticos de centralidad y de dispersión y los gráficos adecuados para los diferentes niveles de medición.

Tabla 2. Nivel de medida de las variables y descriptivos de las distribuciones

Nivel medida variable	Descripción	Centralidad	Dispersión	Gráficos
• Cualitativa • Escala nominal	• Valores no numéricos • Ausencia de orden	• Moda (Mo)	• Razón de variación • Núm. de categorías • Índice Hirsch-Herfindahl	• Diagrama de barras • Gráfico de sectores
• Cualitativa • Escala ordinal	• Valores no numéricos • Presencia de orden	• Mediana (Med) • Moda (Mo)	• Rango (mínimo-máx.) • Cuartiles • Percentiles • Rango intercuartílico	• Diagrama de barras • Box-plot (caja)

Nivel medida variable	Descripción	Centralidad	Dispersión	Gráficos
• Cuantitativa • Escala intervalo y razón	• Cuantitativa discreta (cantidad finita o numerable de valores numéricos) • Cuantitativa continua (cualquier valor numérico en el intervalo) • Escala de intervalo (solo tiene sentido diferencia entre valores) • Escala razón (además de diferencia, razón entre valores)	• Media (μ) • Mediana (Med) • Moda (Mo)	• Rango • Cuartiles • Percentiles • Rango intercuartílico • Desviación típica • Varianza • Coeficiente de variación	• Histograma • Box-plot (caja) • Gráfico de tallo-y-hojas • (Stem-and-leaf)

4.1.1. La matriz de datos

El punto de partida en todo análisis estadístico es la **matriz de datos** en la que tenemos en las filas la lista de nuestras **unidades de análisis** (los casos, las observaciones, los elementos o los individuos) sobre las que hemos medido nuestras variables.

Tabla 3. Matriz de datos

Observaciones / Variables	Variable 1	Variable 2	...	Variable *M*
Obs. 1				
Obs. 2				
...				
Obs. *N*				

Por ejemplo, en la **tabla 4** se representa una matriz de datos en la que las unidades de análisis son los distritos de Barcelona. Esta es una tabla muy pequeña, pero nos sirve de ilustración a la hora de hacer el análisis descriptivo. En las filas tenemos los distritos. En las columnas tenemos listadas algunas variables relativas, por un lado, a la encuesta de victimización de la ciudad del año 2017 (el índice de victimización, el índice de denuncia, la percepción de la seguridad en la ciudad y la percepción de la seguridad en el barrio); por otro lado, tenemos datos del sistema de seguridad pública de Barcelona con las denuncias por infracción de la ordenanza de convivencia ciudadana, los incidentes por degradación del espacio público, los incidentes en la convivencia vecinal, los incidentes por actividades molestas en el espacio público y, finalmente, los incidentes por actividades indebidas en el espacio público. Las medidas del sistema de seguridad son del año 2018 y todas están estandarizadas en tanto por mil habitantes en el distrito (fuente: Ayuntamiento de Barcelona. Gerencia de Seguridad y Prevención).

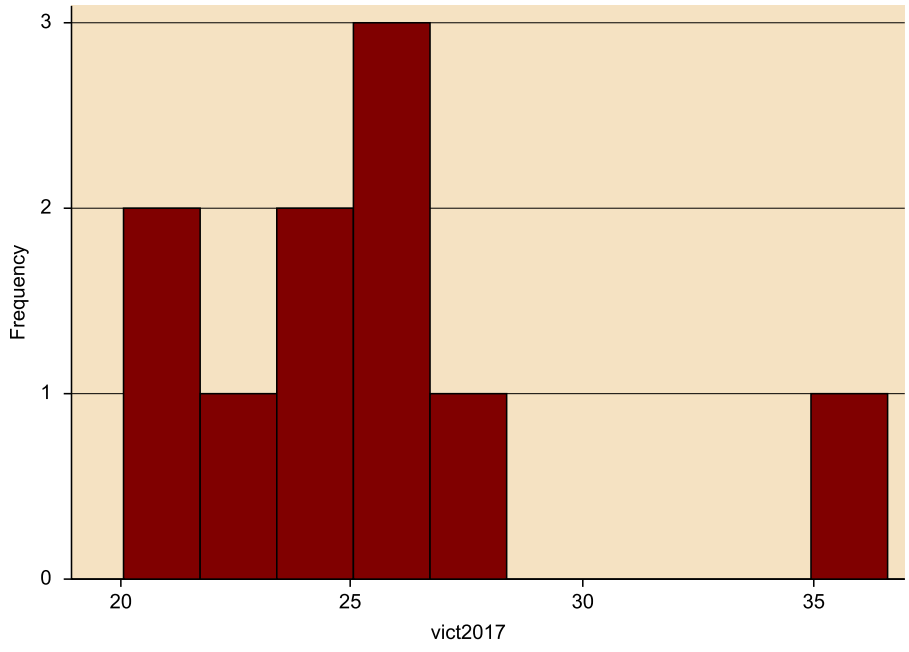
Tabla 4. Matriz de datos de seguridad del Ayuntamiento de Barcelona

	victim	de-nun	segciud	segbarrio	denun-ciasOC	incdegrEsp	IncidConv	Molesto	ActIndeb
1. Ciutat Vella	36,6	15,1	6,2	5,2	46,15	4,23	56,64	110,53	18,18
2. Eixample	28,2	17,4	6,3	6,9	2,08	1,90	26,39	24,45	9,64
3. Sants-Mont-juïc	24,7	22,6	6,1	6,2	2,31	1,95	22,80	27,05	3,90
4. Les Corts	20,1	23,4	6,4	7,3	0,59	1,41	13,58	16,85	2,36
5. Sarrià-Sant Gervasi	25,7	34,6	6,0	6,8	1,47	1,52	18,75	16,84	1,70
6. Gràcia	22,7	22,4	6,3	7,0	4,29	1,97	25,09	25,96	2,24
7. Horta-Guinar-dó	21,2	24,1	6,2	6,3	0,27	1,23	17,74	13,00	1,51
8. Nou Barris	24,1	21,9	6,2	5,9	0,29	1,68	19,70	20,77	2,22
9. Sant Andreu	26,1	29,4	6,2	6,3	0,83	1,91	15,61	17,39	1,68
10. Sant Martí	25,8	21,8	6,2	6,1	5,92	1,90	18,05	25,23	5,13

4.1.2. Las medidas de centralidad

Las medidas de centralidad más conocidas son la moda (Mo) que es el valor más frecuente de una distribución. En el caso de los distritos de Barcelona respecto del índice de victimización de la encuesta de victimización del 2017 (victim), no hay un valor que se repita más de una vez.

Figura 5. Histograma índice victimización BCN 2017



Pero con la ayuda del histograma de la **figura 5** se puede ver cómo hay un pico con 3 distritos con un índice de victimización en el rango 25-26.

La **media** aritmética (μ) es la suma de todos los valores de la distribución dividida por el número de casos o unidades de medida.

$$\mu_x = \frac{(x_1 + x_2 + \ldots + x_n)}{n} = \frac{\sum x_{y\,n}}{}$$

En el índice de victimización del año 2017 (victim), la media es de 25,52.

La **mediana** es el valor que, si ordenamos todos los casos de más pequeño a más grande, tiene el 50 % de los casos con valores superiores y el 50 % de los casos con valores inferiores. Como tenemos 10 distritos, el valor medio de la distribución no es uno de los valores de uno de los distritos, sino el punto medio del rango entre dos distritos Sarrià-Sant Gervasi y Sants-Montjuïc: 25.2.

La mediana se encuentra cerca de la media, porque se trata de una distribución relativamente simétrica. La media tiende a moverse hacia la derecha, por el efecto del distrito de Ciutat Vella, que tiene un índice de victimización extraordinariamente elevado. En general, la mediana es una medida de centralidad más «**robusta**» que la media, que se ve muy afectada por la existencia de casos con valores muy diferentes del resto de la distribución. En este caso, Ciutat Vella.

4.1.3. Las medidas de dispersión

Existen muchos estadísticos de dispersión de las variables medidas numéricamente.

El **rango** mide la distancia entre el valor máximo y el mínimo.

El índice de victimización del año 2017 (victim):

Rango = Max – Min = 36.6 – 20.1 = 16.5

El **rango intercuartílico** representa la distancia entre el primer cuartil (el valor que tiene el 25 % de los casos con valores inferiores y el 75 % de los casos con valores superiores) y el tercer cuartil (el valor que tiene el 75 % de los casos con valores inferiores y el 25 % de los casos con valores superiores):

Rango intercuartílico:

valor $Q_3 – Q_1$ = 26.1 – 26.1 = 3.4

La **varianza** es la suma ponderada por el total de casos de las diferencias al cuadrado entre los valores que adquiere la variable y su media. La base es la diferencia entre la media y cada valor, para evitar que, cuando sumemos todas estas diferencias que algunas veces son positivas y otras negativas (la media es un valor que se encuentra en medio de la distribución), resulte en un valor próximo a cero, se elevan al cuadrado las diferencias. Con esta operación todos los valores ahora son positivos y se pueden sumar para tener una estimación de la dispersión de los casos alrededor de la media:

Varianza:

$$S^2 = \frac{\sum(\mu_x - x_y)^2}{n} = 21.01$$

La **desviación típica** es el estadístico de dispersión más popular. Es la raíz cuadrada de la varianza. La varianza nos proporciona una buena medida de la dispersión, pero tiene el problema que su unidad de medida es el cuadrado del de la variable. Esto hace que no pueda ser interpretado en la escala de la distribución de los casos. La desviación típica resuelve este problema con la raíz cuadrada.

Desviación típica:

$$S = \left(\frac{\sum(\mu_x - x_y)^2}{n}\right)^{\frac{1}{2}} = 4.58$$

La desviación típica del índice de victimización en los distritos de Barcelona es 4.58, que significa una buena proporción de los casos.

La desviación típica es muy conveniente cuando queremos comparar distribuciones de una misma variable entre grupos o poblaciones diferentes. Podemos ver si, por ejemplo, la distribución de las calificaciones de dos clases son muy diferentes, a pesar de que en las dos clases la media sea la misma. Pero si queremos comparar las distribuciones de variables diferentes la desviación típica no es adecuada. El **coeficiente de variación** resuelve el problema dividiendo la desviación típica por la media. Es una medida de dispersión adimensional: sin unidades de medida, de forma que permite comparar la dispersión de cualquier tipo de distribución.

Coeficiente de variación:

$$CV = \frac{s}{\mu_x} = 0.18$$

4.1.4. Medidas de forma

Existen el coeficiente de asimetría y el coeficiente de apuntamiento que comparan las distribuciones con la distribución normal.

La medida más simple y útil de la asimetría de las variables es la diferencia entre la media y la mediana dividida por la desviación típica.

Asimetría:

$$A = \frac{\mu_x - \text{Med}}{s}$$

4.1.5. Las tablas de frecuencia y las representaciones gráficas

Una forma de representar el conjunto de los datos de una variable es a través de la **tabla de frecuencias**.

La **tabla de frecuencias** lista las frecuencias de cada valor de la variable.

Esta solución es conveniente para las variables medidas cualitativamente, que acostumbran a tener un número limitado de valores o categorías; pero para las variables de tipo cuantitativo (especialmente si la variable es continua), hay que agrupar las categorías en rangos.

La **tabla 5** muestra la tabla de frecuencias de la variable comunidad autónoma del estudio de las sentencias por delitos sexuales a menores en las audiencias provinciales españolas entre los años 2011 y 2014.

Tabla 5. Tabla de frecuencias comunidad autónoma. Sentencias por delitos sexuales a menores 2011-2014

Comunidad autónoma	Frec.	Porcentaje	Cum.
Andalucía	378	16.12	16.12
Aragón	63	2.69	18.81
Asturias	41	1.75	20.55
Baleares	117	4.99	25.54
Canarias	172	7.33	32.88
Cantabria	37	1.58	34.46
Castilla-León	78	3.33	37.78
Castilla-Mancha	106	4.52	42.3
Cataluña	330	14.07	56.38
Valencia	260	11.09	67.46

Comunidad autónoma	Frec.	Porcentaje	Cum.
Extremadura	93	3.97	71.43
Galicia	102	4.35	75.78
Madrid	287	12.24	88.02
Murcia	116	4.95	92.96
Navarra	35	1.49	94.46
País Vasco	108	4.61	99.06
La Rioja	11	0.47	99.53
Ceuta	2	0.09	99.62
Melilla	9	0.38	100

Los **diagramas de barras** son convenientes para las variables cualitativas ordinales o categóricas. La **tabla 6** es una tabla de frecuencias que reproduce la información de la variable comunidad autónoma (variable medida **desde el punto de vista nominal**) de la investigación sobre las sentencias por abusos sexuales que aparecen en la **tabla 5**. En esta tabla se ha añadido un diagrama de barras y las categorías se han ordenado en orden descendente de tamaño. De esta forma, la información contenida se puede ver fácilmente de un vistazo.

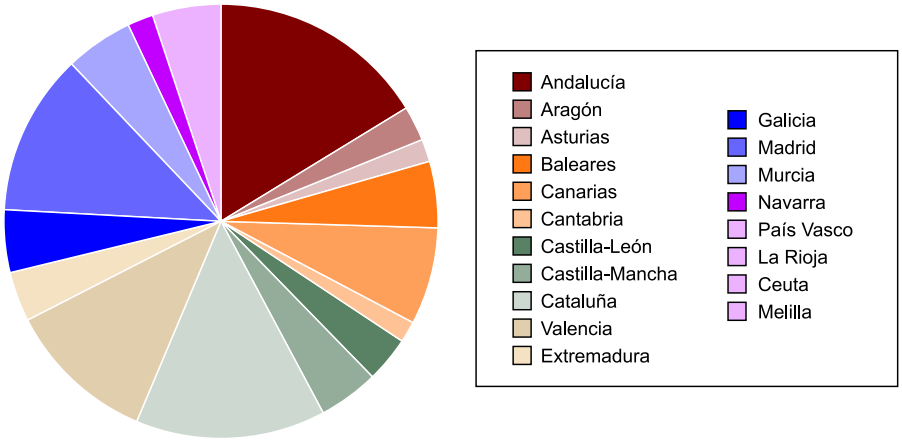
Tabla 6. Tabla de frecuencias con diagrama de barras. Sentencias por delitos sexuales a menores 2014-2016

Comunidad aut	Frec.	
Andalucía	378	**********
Cataluña	330	*********
Madrid	287	********
Valencia	260	*******
Canarias	172	*****
Baleares	117	****
Murcia	116	****
País Vasco	108	****
Castilla-Mancha	106	****
Galicia	102	****
Extremadura	93	***
Castilla-León	78	***
Aragón	63	**
Asturias	41	*
Cantabria	37	*

Comunidad aut	Frec.	
Navarra	35	*****
La Rioja	11	*
Melilla	9	*
Ceuta	2	

La **figura 6** es un **diagrama de sectores** que presenta alternativamente la información de las tablas 4 y 5. Este tipo de gráfico visualmente es atractivo y, por esta razón, es utilizado frecuentemente en los reportajes periodísticos. Pero como se puede apreciar en la figura 6, muchas veces **no es la mejor forma de presentar cuidadosamente la información**. La figura 6 nos proporciona una visión de la fragmentación de los datos en las 17 comunidades, pero **es confusa** y no se aproxima a la calidad de la representación de la tabla 6 que combina tabla de frecuencias y diagrama de barras ordenadas por frecuencia de las categorías.

Figura 6. Diagrama de sectores. Sentencias sexuales a menores



El buen gráfico

En general, en las **representaciones gráficas** intentaremos ser lo más claros y precisos posible. De este modo, mejoraremos en la eficacia de la comunicación de la información. El mejor gráfico es el que es capaz de transmitir la información relevante en el mínimo tiempo, espacio y tinta. Los gráficos eficientes son una muestra de respeto para los lectores de las investigaciones. El buen gráfico tiene que mostrar los datos y tiene que inducir a pensar en lo que es sustancial. Se tiene que evitar hacer representaciones que induzcan a una percepción distorsionada de los datos. El buen gráfico representa mucha información en poco espacio y proporciona coherencia a un gran número de datos. También puede presentar los datos a diferentes niveles de detalle.

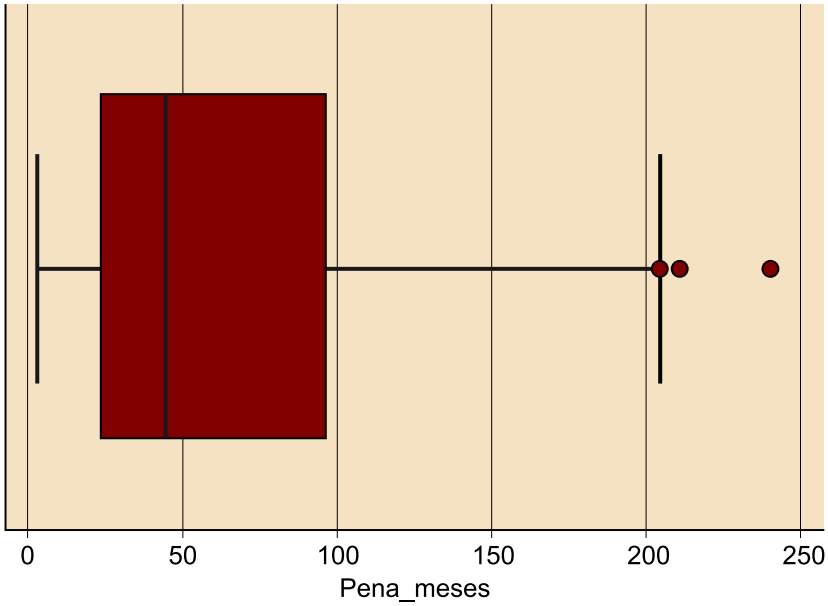
Los diagramas de sectores están **completamente contraindicados** para representar las **variables de tipo ordinal** porque hacen perder la noción de orden en las categorías. En las crónicas periodísticas de las encuestas de opinión, donde una buena parte de las preguntas son escalas de actitudes ordinales, para mejorar en la «variedad» en la presentación y así evitar «el aburrimiento» del lector, se tiende a utilizar los diagramas de sectores, que esconden alguna información más relevante de las variables: la distribución de la opinión entre los que favorecen o no una alternativa.

Para las variables ordinales, la representación gráfica preferida tendría que ser el **diagrama de barras**, a pesar de que, tal como recoge la tabla 2, también se pueden utilizar los diagramas de caja o box-plot. El **box-plot** es una representación gráfica de la localización del primero, segundo y tercer cuartiles de una variable y de su máximo y mínimo. Este tipo de representación de las variables es muy **compacto** y es muy útil cuando queremos **comparar distribuciones**

de diferentes variables. El box-plot no solo se utiliza con las variables medidas desde el punto de vista ordinal, sino que también es muy recomendable para representar las variables numéricas.

En la base de datos de las sentencias por delitos sexuales a menores de edad, tenemos la **variable numérica** de las **penas de prisión** medidas en meses, a las que han sido condenados los ofensores que se han considerado culpables.

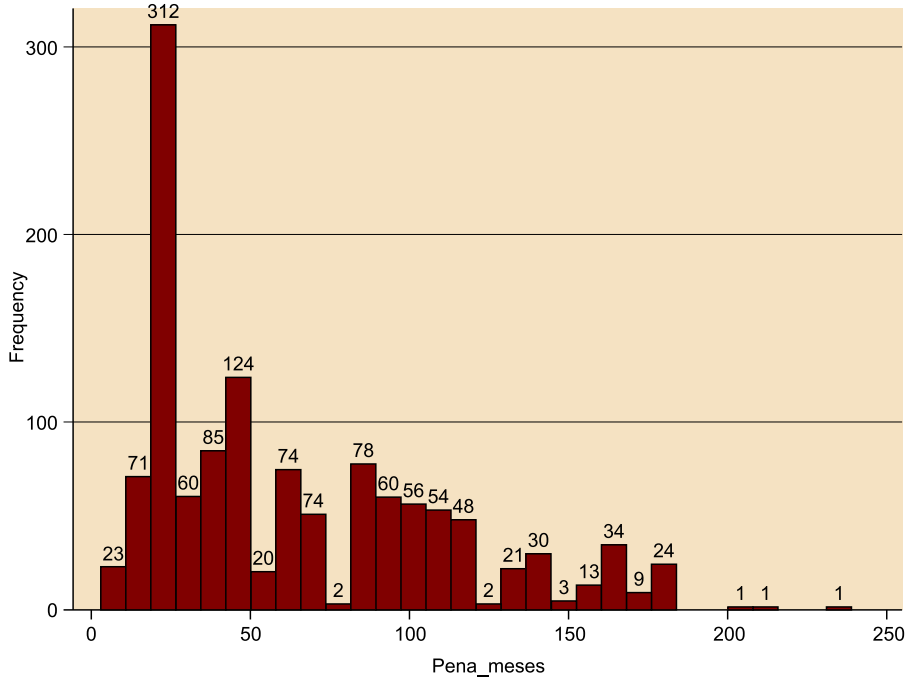
Figura 7. Box-plot pena de prisión (meses)



En la figura 7 se representa el box-plot de la variable pena de prisión. La caja central del gráfico delimitada por el primer cuartil (24 meses) y el tercer cuartil (96 meses), contiene el 50 % de los casos. La línea vertical que divide la caja central corresponde a la situación de la mediana (48 meses, que es el segundo cuartil). Por encima de la mediana, tenemos el 50 % de casos y, por debajo, tenemos el otro 50 % de casos. Las líneas delgadas verticales que están conectadas con la caja por una línea horizontal a los dos lados de la caja central (a veces denominadas bigotes) marcan el límite de los casos que tienen por debajo (3 meses) y por encima (210 meses), el 5 % de los casos. Los puntos que aparecen a la derecha son los casos extremos, los *outliers*.

Si sabemos leer el box-plot es muy fácil darse cuenta que se trata de una distribución muy asimétrica, puesto que la mayoría de los valores se encuentran agrupados en valores bajos de la distribución. Naturalmente, la representación del box-plot es esquemática (y esta es la ventaja cuando queremos comparar las distribuciones de muchas variables), pero no nos permite tener el nivel de detalle en la representación de una variable de tipo cuantitativo que nos permite el histograma (ya hemos visto uno en la figura 5).

Figura 8. Histograma pena de prisión (meses)



La figura 8 es el histograma de la misma variable (pena de prisión en meses) que el box-plot de la figura 7. Las barras del histograma recogen agrupamientos de valores de la variable y han sido etiquetadas por el número de casos que contienen. Desde este punto de vista, la información es comparable a una tabla de frecuencias. Cuando comparamos la representación del histograma con el box-plot podemos comprobar hasta qué punto nos ofrece más detalle y complejidad.

4.2. El análisis estadístico de dos variables

Para comprobar las relaciones entre las variables tenemos que tener en cuenta su nivel de medida, del mismo modo que en las medidas descriptivas univariantes.

4.2.1. Diferencia entre medias y ANOVA

La relación entre dos variables más fácil de imaginar es la que se establece entre una **variable independiente categórica** y una **variable dependiente numérica**. Esta situación puede corresponder a la situación de un experimento en el que la variable independiente (categórica dicotómica –con dos valores posibles–) indica que se ha recibido el tratamiento o no.

Por ejemplo, la variable dependiente numérica es el crecimiento de las plantas de una parcela y la variable independiente (o variable de tratamiento) es haber recibido una dosis de fertilizante o no. Si el fertilizante tiene efecto (es decir, si la variable independiente tiene relación con la variable dependiente) esperamos que las plantas que han recibido el tratamiento tengan un crecimiento medio superior. Para comprobar que el tratamiento tiene efecto, tendremos que medir el crecimiento medio en las parcelas sin tratamiento y el crecimiento medio en las parcelas con tratamiento. La **diferencia de las medias** de los dos grupos de parcelas será el efecto del tratamiento. No todas las parcelas son iguales, las hay más soleadas que otras, las hay con tierra más rica que otras pero, si hacemos muchas

parcelas y las distribuimos aleatoriamente entre el grupo de tratamiento y de control, podremos establecer sin duda el efecto medio del tratamiento sobre el crecimiento de las plantas.

Si estamos trabajando con una muestra aleatoria y queremos inferir si la diferencia entre las medias que hemos observado existe en la población, se puede hacer un test de significatividad estadística. El estadístico de la diferencia entre las medias sigue una distribución t-Student que es fácil de calcular y que está incorporada en los paquetes estadísticos más básicos.

Cuando tenemos la variable independiente categórica que tiene valores múltiples (es multinomial; por ejemplo, la confesión religiosa), se pueden analizar las diferencias entre las diferentes categorías y la variable dependiente numérica. Si estamos trabajando con una muestra aleatoria y queremos hacer el test de significatividad, tendremos que hacer un análisis de la varianza (**ANOVA**, por las siglas en inglés). El estadístico F, que relaciona la variación de los valores de los individuos dentro de las categorías con la variación total de los individuos, permite establecer los valores críticos a partir de los que se puede decir con un cierto grado de certeza (nivel de confianza) que las dos variables están relacionadas.

4.2.2. Relaciones entre variables numéricas: el gráfico de dispersión

Cuando la variable independiente y la variable dependiente son numéricas, la mejor manera de establecer si hay una relación es representarlas en un **gráfico de dispersión**. Cada observación de nuestro estudio se identifica por una coordenada cartesiana y puede ser representada en el plano. La existencia de relaciones se observa porque hay patrones definidos en la situación de los puntos.

Figura 9. Ausencia de relación y relación lineal positiva con poca dispersión

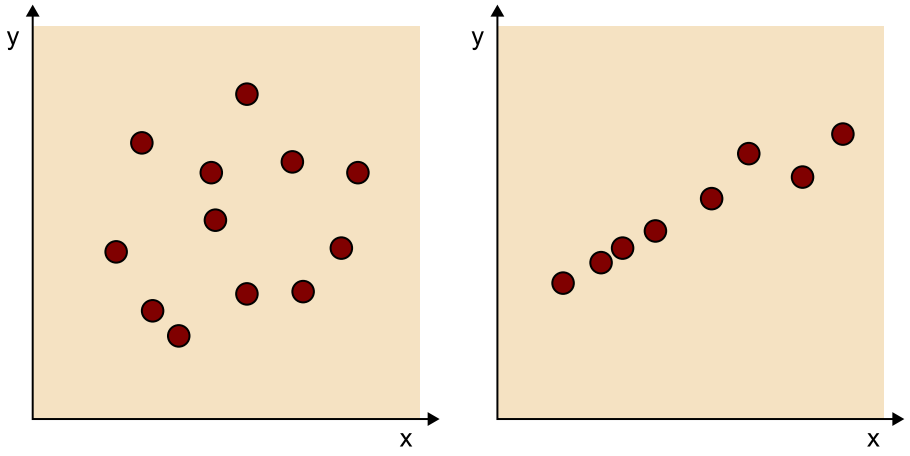
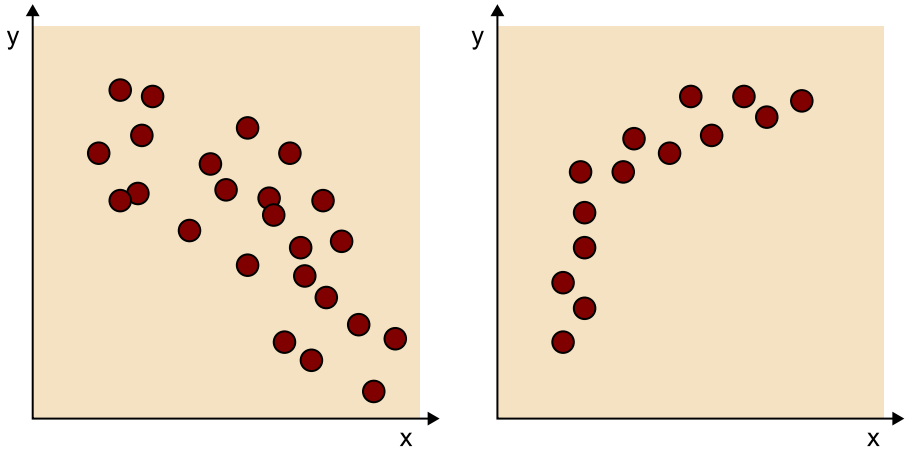


Figura 10. Relación lineal negativa con dispersión y relación curvilínea



En los seis paneles de las **figuras 9**, **10** y **11**, podemos observar diferentes tipos de relaciones entre variables numéricas. En el **primer panel** (**figura 9** izquierda) se ve que los casos se reparten sin orden en el plano. A valores elevados en el eje de la *x* podemos encontrar valores altos o bajos del eje de la *y* y, al revés, a valores bajos de la *x* podemos encontrar cualquier tipo de valor en el eje de la *y*. No conocemos mejor el valor de la *y* de un caso cuando conocemos su valor en *x*. En cambio, en el **segundo panel** (**figura 9** derecha), nos encontramos con una posición completamente diferente: si conocemos los valores de la variable independiente *x*, podemos conocer con mucha precisión cuál será su valor en la *y*.

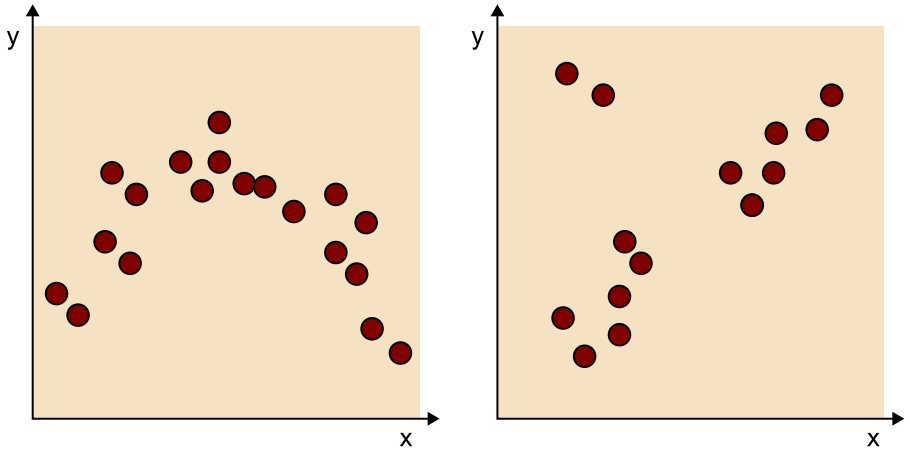
> Por ejemplo, imaginemos que la variable *x* son las horas de estudio en matemáticas y la variable dependiente *y* son las calificaciones obtenidas en las pruebas finales. Si la relación entre las dos variables es la descrita en el **panel 1**, cuando sabemos la cantidad de horas estudiando no podemos conocer cuál es la nota final que se obtendrá. La situación es completamente diferente si la relación es la que describe el **panel 2**: si conozco cuánto tiempo un estudiante ha dedicado a las matemáticas, conozco casi perfectamente la nota que obtendrá.

En el **panel 3** (**figura 10** izquierda) la relación es negativa. Cuando crece la variable independiente, decrecen los valores de la variable dependiente. Más horas de estudio no sirven para tener más nota (quizás aquí tenemos una reversión de la relación de causalidad: los estudiantes que están más bloqueados con las matemáticas le dedican más tiempo, ¡pero realmente son los que tienen peores resultados!). Además, la relación entre las variables es mucho menos precisa. La nube de puntos es más dispersa, no adopta la forma clara de línea que mostraba el **panel 2**.

Finalmente, el **panel 4** (**figura 10** derecha) muestra que los gráficos de dispersión pueden mostrar patrones de relaciones entre las variables más complejos y matizados que las relaciones lineales. En el gráfico se ve que la relación es positiva, pero el efecto del cambio en la variable independiente sobre la variable dependiente es muy fuerte en los valores más bajos, mientras que para los valores altos de *x*, la variable independiente casi no tiene efecto. En el ejemplo, esto quiere decir que las primeras horas dedicadas a estudiar matemáticas

tienen un efecto grande sobre la calificación obtenida, pero que, a partir de un cierto momento, estudiar más horas contribuye mucho menos a la calificación.

Figura 11. Relación no monótona y constatación de vacíos y de casos extremos (*outliers*)



El **panel 5** (**figura 11** izquierda) muestra una relación todavía más interesante entre las dos variables. La relación entre las variables es positiva en los valores bajos de la *x* y es negativa a partir de un cierto momento. Esto quiere decir que la relación no es monótona, no siempre va en una dirección. En nuestro ejemplo, hay un momento en el que más horas de estudio son contraproducentes, no es que el efecto de una hora de estudio más tenga un efecto menor (efecto marginal decreciente, en el lenguaje de los economistas), sino que, directamente, una hora más de estudio hace sacar peores resultados en las evaluaciones.

Finalmente, el **panel 6** (**figura 11** derecha) muestra cómo los diagramas de puntos son muy útiles para indicar anomalías en las relaciones entre las variables. Pueden ser identificados vacíos en las relaciones: hay algunos valores de algunas de las variables que no están presentes. También son buenos para visualizar la existencia de casos extremos (*outliers*).

Al hacer la descripción de un gráfico nos podemos fijar en tres aspectos:

- ¿Cuál es la tendencia que se puede detectar? Positiva, negativa, curvilínea.
- ¿Cuál es la fortaleza de esta relación?
- ¿Cuál es el grado de dispersión de los casos? ¿Hay casos extraordinarios (*outliers*)? ¿Cuáles son? ¿Qué razones parecen explicarlos? (muy a menudo estos casos son errores en la introducción de los datos).

En conjunto, los gráficos de puntos son un instrumento imprescindible para comprobar las relaciones que existen entre las variables numéricas. El nivel de medida numérico de las variables nos da mucha información de los casos individuales y esto permite que las relaciones entre las variables puedan ser muy afinadas. La visualización de las relaciones nos permite identificar anomalías (los casos extremos, los casos que se apartan de la relación normal) que nos

pueden llevar a pensar y desarrollar o refinar las explicaciones de los fenómenos que estamos estudiando. En general, el énfasis en la visualización de los datos está relacionada con la *serendipity*, un neologismo inglés que plantea la capacidad de hacer descubrimientos inesperados a lo largo de nuestra investigación.

4.2.3. Coeficiente de correlación lineal: *r* de Pearson

Para medir el grado de asociación lineal entre dos medidas numéricas se usa el coeficiente de correlación lineal de Pearson, *r*. Es una medida direccional que se encuentra acotada entre –1 y 1. El signo nos dice si la relación es negativa o positiva. Cuanto mayor es el valor absoluto de *r* , se acerca más a 1, la relación lineal entre las dos variables es más fuerte. Se tiene que ser consciente que relaciones no lineales (como la del panel 5) a pesar de que sean claras y fuertes, no aparecerán o se subestimarán en los valores de *r* . Los paquetes estadísticos nos ofrecen medidas de significatividad estadística.

4.2.4. Análisis de regresión simple (o bivariante)

Cuando, más allá de la fortaleza de la relación entre dos variables y la dirección de esta relación, queremos examinar la relación entre dos variables numéricas entre las que suponemos que existe una relación de causalidad (o una relación estructural), usamos el **análisis de regresión**.

> En el análisis de regresión se entiende que la variable dependiente *y*, y la variable independiente *x* pueden ser conectadas matemáticamente con diferentes formas funcionales.

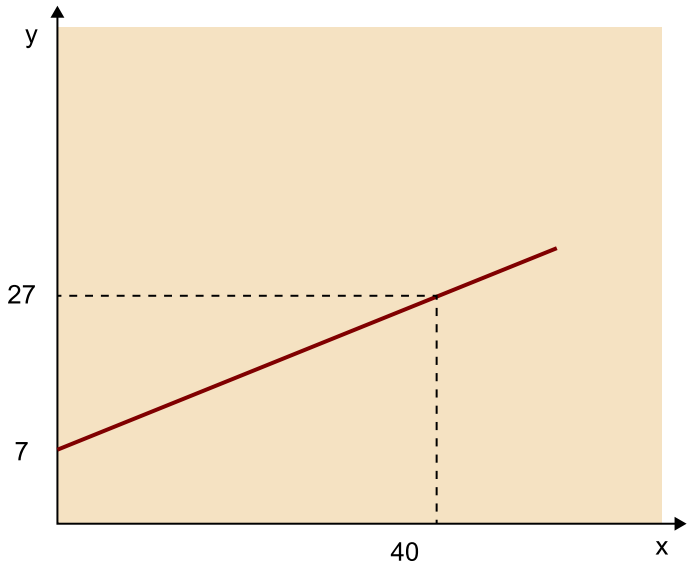
La más simple es la forma de la línea recta:

$$y = a + b\,x$$

Con esta función se pueden plantear relaciones relativamente complejas entre las variables. Para ilustrar qué hay en esta fórmula podemos utilizar un ejemplo políticamente incorrecto.

> Un despropósito del reconocido misógino Josep Pla dice que «la edad ideal de la pareja de un hombre tiene la mitad de años del hombre más siete años». Así, un hombre de 40 años tendría que hacer pareja con una mujer de 27. Esta relación se puede formular como una línea recta que establece los años de la mujer (*y*) en función de los años del hombre (*x*):
>
> Años mujer = 7 + ½ años hombre

Figura 12. Relación años mujeres respecto años hombres



En el análisis de regresión, se supone que la relación entre la variable independiente (la explicativa) y la variable dependiente (a explicar) no es completamente determinista: hay otros muchos factores que explican los valores que logra la variable dependiente (además de la variable que estamos teniendo en cuenta). Entonces, esperemos que exista una variación probabilista para cada valor del factor explicativo. El factor de error (*e*) captura el efecto de todos estos factores y nos dice que los valores de *y* que observamos son superiores o inferiores a los valores que predice la relación y están representados por la recta de regresión:

$$y = a + b\,x + e$$

En el análisis de regresión bivariante podemos averiguar la relación entre las dos variables y los diagnósticos habituales que nos ofrecen los paquetes estadísticos nos permiten ver de qué manera esta relación es fuerte y significativa.

4.3. Relaciones entre variables cualitativas

4.3.1. La tabla de contingencia

Cuando tenemos variables calitativas o categóricas (nominales u ordinales), la **tabla de contingencia** es el instrumento básico utilizado para representar la relación. Esta herramienta es comparable al diagrama de dispersión utilizado por las variables medidas numéricamente. Nos proporciona un grado parecido de riqueza y de matiz en la percepción de la relación. En los análisis de las encuestas es típico que los datos se recojan con este nivel de medida.

Una tabla de contingencia es una **tabla de doble entrada** que facilita el examen de las relaciones entre las variables.

Normalmente, las categorías de la variable independiente (*y*) se ponen en las columnas y las categorías de la variable dependiente (*x*), en las filas de la tabla.

Tabla 7. Tabla de contingencia acusado víctima y fase del juicio al que llega el caso

	Relación acusado víctima				
Fase del juicio	**Casos perdidos**	**Conocido/maestro**	**Desconocido**	**Padres/familiares**	**Total**
Sobreseimiento	2	0	11	2	15
Archivo	6	14	10	17	47
Juicio faltas	2	0	3	0	5
Juicio oral	0	3	13	14	30
Total	10	17	37	33	97

La **tabla 7** muestra la relación entre la variable tipo de relación entre la víctima y el acusado en los delitos sexuales a menores en los casos juzgados en la Audiencia Provincial de Lérida en 2011 y la fase del juicio al que llegaron. La primera casilla de la izquierda (primera columna, primera fila) nos dice que hubo 2 casos en los que no se pudo saber la relación entre víctima y acusado, en los que hubo sobreseimiento.

La última columna y la última fila son especiales porque tienen la suma de cada fila y de cada columna, respectivamente. Se denominan los marginales. En realidad, la última columna es la distribución de las categorías de la variable dependiente (el nivel de juicio al que se llega), y la última fila es la distribución de las categorías de la variable independiente (la relación entre víctima y ofensor).

Para ver el efecto de la variable independiente sobre la dependiente se tiene que eliminar el efecto de los diferentes tamaños de las categorías (hay muchos más padres o familiares y desconocidos que conocidos o maestros y casos perdidos). Por esta razón, se tiene que calcular el porcentaje de cada casilla en su columna. Es decir, el valor de la casilla por el marginal de la fila. Esto lo vemos a la **tabla 8**.

Tabla 8. Relación acusado víctima y fase del juicio al que llega el caso

	Relación acusado víctima				
Fase del juicio	**Casos perdidos**	**Conocido/maestro**	**Desconocido**	**Padres/familiares**	**Total**

	Relación acusado víctima				
Sobreseimiento	20 %	0 %	30 %	6 %	15.5 %
Archivo	60 %	82 %	27 %	51 %	48.4 %
Juicio faltas	20 %	0 %	8 %	0 %	5.1 %
Juicio oral	0 %	17 %	35 %	42 %	30.9 %
N	10	17	37	33	97

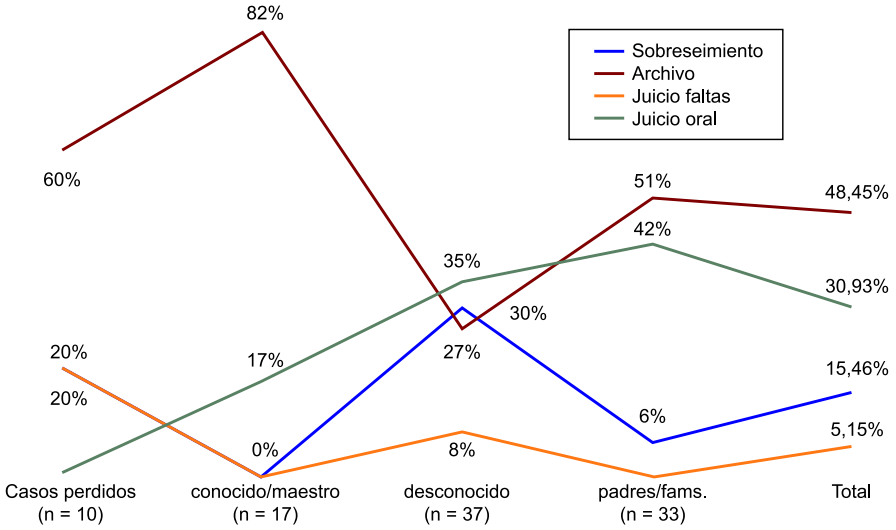
¿Cómo se lee la tabla de contingencia? El **efecto** de la variable independiente, la relación acusado víctima, sobre la variable dependiente, la fase del juicio a la que se llega, se ve **comparando los porcentajes por columna de cada fila**. Es decir, eligiendo una fila (por ejemplo, Archivo) y viendo los porcentajes de «desconocidos» que hay por cada fase del juicio. Así, los casos archivados donde el ofensor es desconocido de la víctima son un 27 %, muy por debajo de los padres o familiares que son el 51 %.

Si comparamos los porcentajes de cada categoría de la variable dependiente (filas) con su marginal, veremos si cada casilla está por encima o por debajo de la proporción en el conjunto de casos.

Cuando leemos los porcentajes por columna, lo que estamos mirando es la composición de la categoría de la variable independiente sobre la dependiente: como se distribuyen los casos en los que los ofensores son padres o familiares de la víctima respecto a la fase a la que llegan en el juicio.

La comparación de las diferencias entre los porcentajes de columna permite calcular la diferencia entre los porcentajes para cada categoría. Esta diferencia es el efecto de la variable independiente con la dependiente. Para representar esta diferencia entre las diferentes casillas de la tabla se puede dibujar un **diagrama de pendientes**, que se representa en un gráfico de líneas. La **figura 13** presenta la información de la **tabla 8** en la forma de un diagrama de pendientes. De un vistazo, se puede ver de qué forma la relación entre víctima y ofensor afecta la probabilidad de acceder a una fase superior en el juicio. Por ejemplo, destaca el hecho de que la mayor parte de los casos donde el ofensor es conocido o es el maestro de la víctima, el caso se sobresee.

Figura 13. Gráfico de pendientes: Relación víctima-ofensor y fase del juicio



Para ver si son estadísticamente significativas las relaciones, se puede hacer un test de significatividad de las diferencias entre las proporciones. Pero, en una tabla normal hay muchas casillas y esto quiere decir que se pueden hacer muchas pruebas diferentes. Si queremos un test de la relación entre dos variables categóricas, iría muy bien tener un test global.

4.3.2. El test de independencia ji cuadrado (χ^2)

El test de independencia **ji cuadrado** nos permite hacer este test global de la relación entre dos variables categóricas. El test se basa en la discrepancia entre los casos que se observan en las diferentes casillas de la tabla en relación con los casos que se esperaría observar si las dos variables fueran independientes. Si no hubiera relación entre las variables, habría poca discrepancia. La distribución de valores del estadístico de **ji cuadrado** nos dice cuál es la probabilidad de estas discrepancias si los datos provinieran de una población en la que las dos variables no estuvieran relacionadas. Si el valor de ji cuadrado supera un valor crítico, se puede afirmar, en un determinado nivel de confianza, que las dos variables no son independientes.

4.3.3. El coeficiente de contingencia y la V de Cramer

El test de ji cuadrado solo nos dice si se puede rechazar la independencia entre las variables, pero no nos dice la fortaleza de la relación cuando la comparamos con la relación que existe con otras variables. El **coeficiente de contingencia (CC)** o la **V de Cramer** son estadísticos basados en la variable ji cuadrado que están acotados entre 0 y 1. El problema del CC es que no llega generalmente al máximo de 1, mientras que la V de Cramer, sí.

4.3.4. La lambda (λ)

Alternativamente, la **lambd**a (λ) es una medida de asociación basada en la lógica de la reducción proporcional del error cuando alguna de las variables tiene un **nivel de medida nominal**. La reducción proporcional del error nos dice en que medida dejamos de equivocarnos al atribuir un valor en una variable a un individuo cuando conocemos previamente su valor en otra variable. Se trata, por esta razón, de una medida direccional: asume que hay una variable independiente y otra dependiente. El ANOVA o la regresión lineal funcionan con esta lógica.

4.3.5. La gamma (γ) y las tau (τ) de Kendall

La **gamma** (γ) y las diferentes **tau** (τ) **de Kendall** son medidas de asociación de **variables ordinales** basadas en la reducción proporcional del error. El cálculo se basa en la proporción de parejas concordantes (que indican una relación positiva entre las variables) y parejas discordantes (que indican una relación negativa) en una tabla de contingencia. Las medidas son direccionales y están acotadas entre -1 y 1. Las variedades de las tau de Kendall dependen del tipo de tabla en la que se hace el cálculo, cuadrada o rectangular (es decir, donde el número de categorías de las dos variables son iguales o diferentes).

Glosario

coeficiente de variación $m\ CV = \frac{s}{\mu_x}$

desviación típica $f\ s = \left(\frac{\sum \left(\mu_x - x_i \right)^2}{n} \right)^{1/2}$

índice HH *m* El índice Herfindahl-Hirschman mide la concentración de las variables categóricas (se desarrolló para medir la concentración industrial en diferentes sectores) $HH = \sum p_i^2$, donde p_i es la proporción de cada categoría en la variable.

mediana $f\ \mu_x = \frac{\left(x_1 + x_2 + \ldots + x_n \right)}{n} = \frac{\sum \left(x_i \right)}{n}$. Medida de centralidad del nivel de medición cuantitativo. Puede haber versiones «robustas» de la medida que evitan los efectos de los valores extremos. Por ejemplo, la media truncada (que elimina el 5 % de los valores superiores y el 5 % de los valores inferiores). Es un sistema utilizado en las competiciones deportivas (p. ej., gimnásticas o de saltos de trampolín) para evitar el impacto de votos de jueces que quieran afectar el resultado final con una valoración desproporcionada.

moda *f* Valor más frecuente. Es la única medida de centralidad de las variables categóricas nominales, pero puede ser utilizada en otras escalas de medición. En las distribuciones de probabilidad se pueden distinguir diferentes modas: las generales y las locales.

cuartil o decil *m* Valor que tiene una proporción de individuos de la distribución por debajo o por encima. Por ejemplo, el primer cuartil C_1 es el valor que tiene el 25 % de los casos de la distribución por debajo y el 75 % de los casos por encima. El sexto decil D_6 es el valor que tiene el 60 % de los casos por debajo y el 40 % de los casos por encima.

rango *m* Medida de la distancia entre el valor máximo y el mínimo.

rango intercuartílico *m* Diferencia entre los valores del tercer y primer cuartil de una distribución (rango intercuartílico = C_3-C_1).

razón de variación $f\ RV = \frac{1 - n_{Mo}}{n}$. Una medida de dispersión de las variables categóricas, donde n_{Mo} es la frecuencia de la categoría modal.

variancia $f\ s^2 = \frac{\sum \left(\mu_x - x_i \right)^2}{n}$

Bibliografía

Referencias y lecturas recomendadas

Bayens, G. J.; Roberson, C. (2011). *Criminal Justice Research Methods: Theory and Practice* (2.ª ed.). CRC Press.

Capdevila Capdevila, M.; Ferrer Puig, M. (2009). *Tasa de reincidencia penitenciaria 2008* (p. 237) [Àmbit Social i Criminològic]. Barcelona: Centre d'Estudis Jurídics i Formació Especialitzada.

Capdevila Capdevila, M.; Ferrer Puig, M.; Luque Reina, E. (2005). *La reincidencia en el delito en la justicia de menores* (p. 276). Barcelona: Centre d'Estudis Jurídics i Formació Especialitzada.

Kubrin, C. E.; Wo, J. C. (2015). «Social Disorganization Theory's Greatest Challenge: Linking Structural Characteristics to Crime in Socially Disorganized Communities». En: A. R. Piquero (ed.). *The Handbook of Criminological Theory*. Chichester: West Sussex; Afanan, MA: John Wiley & Sons.

Piquero, A. R.; Weisburd, D. (2009). *Handbook of Quantitative Criminology*. Nueva York / Dordrecht / Heidelberg / Londres: Springer.

Weisburd, D.; Britt, C. (2007). *Statistics in Criminal Justice*. Springer.