UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DEPARTAMENTO DE PROYECTOS DE INGENIERÍA
GEDCE Grupo de Estudios en Desarrollo, Cooperación y Ética

CUADERNOS DOCENTES EN
PROCESOS DE DESARROLLO
N.º 1

Metodología y Técnicas Cuantitativas de Investigación

Andrés Hueso y Mª Josep Cascant

Andrés Hueso González
Mª Josep Cascant i Sempere

METODOLOGÍA Y TÉCNICAS CUANTITATIVAS DE INVESTIGACIÓN

CUADERNOS DOCENTES EN PROCESOS DE DESARROLLO

NÚMERO 1

EDITORIAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Grupo de Estudios en Desarrollo, Cooperación y Ética
Departamento de Proyectos de Ingeniería
Universitat Politècnica de València

Primera edición, 2012

ISBN: 978-84-8363-893-4 (versión impresa)

Prefacio

¿Por qué surge este cuaderno?

La idea de este cuaderno germinó durante la preparación de una asignatura de metodología cuantitativa en el Máster sobre Políticas y Procesos de Desarrollo de la Universitat Politècnica de València. A la luz de nuestra experiencia práctica en investigaciones en desarrollo, los manuales y publicaciones sobre cuantitativa nos parecían poco adecuados a estudios en desarrollo: textos centrados exclusivamente en la estadística, otros que pretenden abarcar toda la realidad mediante números o que son solo aplicables en entornos 'sencillos' y controlables. Esto nos llevó a elaborar unos materiales específicos que con el tiempo han ido creciendo hasta convertirse en el cuaderno que ahora estás leyendo.

Este cuaderno pretende ayudar a la lectora a construir los conocimientos básicos para diseñar y realizar una investigación con técnicas cuantitativas de manera rigurosa y adecuada a los objetivos y el contexto de una investigación en desarrollo. Por el camino, trata de cuestionar y desafiar ciertos mitos que suelen acompañar a la metodología cuantitativa:

- la propia división entre lo cualitativo y cuantitativo, que eclipsa la pluralidad de estrategias de investigación y los matices y combinaciones posibles
- la identificación de la metodología cuantitativa con análisis estadístico, cuando el análisis es tan solo una de las etapas de la metodología cuantitativa
- la identificación de la metodología cuantitativa con la encuesta, como si fuese la única técnica de recogida de información
- la asociación de la investigación cuantitativa con el positivismo epistemológico, que eleva lo cuantitativo a verdad universal y –como reacción– genera rechazo a lo cuantitativo desde otras perspectivas epistemológicas
- la disociación de lo cuantitativo de la participación, el aprendizaje o el cambio social
- la rigidez metodológica, que pone los métodos por encima de los objetivos de la investigación

Así, la explicación de la metodología cuantitativa está adaptada a situaciones habituales en proyectos y procesos de desarrollo, y surge en gran medida de nuestra experiencia práctica como facilitadoras, investigadores y docentes. La metodología se trata desde una perspectiva epistemológica realista, en la que la combinación de técnicas cualitativas y cualitativas no se considera problemática, sino que resulta esencial para acercarnos a la realidad al unir extensión y profundidad.

El cuaderno se ha escrito en un tono distendido y ameno, utilizando numerosos ejemplos y pensando en un lector con poca experiencia previa en la materia. Esperamos que te sea útil, lo disfrutes y nos hagas llegar cualquier comentario a ahuesog@upvnet.upv.es.

El autor y la autora

Índice

Capítulo 1. La investigación cuantitativa

Objetivos del capítulo:	Tiempo estimado de lectura: 45 min
• Comprender críticamente los fundamentos y paradigmas en la investigación cuantitativa • Seleccionar técnicas de investigación - incluidas las cuantitativas- adecuadas al contexto de la misma	Apartados del capítulo: 1.1 Introducción 1.2 Los números en el desarrollo 1.3 Metodología de investigación cuantitativa
Índice	Capítulo siguiente. Diseño de investigación

1.1 Introducción

1.1.1 Conceptos básicos

La **metodología de investigación cuantitativa** se basa en el uso de técnicas estadísticas para conocer ciertos aspectos de interés sobre la población que se está estudiando.

Se utiliza en diferentes ámbitos, desde estudios de opinión hasta diagnósticos para establecer políticas de desarrollo. Descansa en el principio de que las partes representan al todo; estudiando a cierto número de sujetos de la población (una muestra) nos podemos hacer una idea de cómo es la población en su conjunto. Concretamente, se pretende conocer la distribución de ciertas variables de interés en una población. Dichas variables pueden ser tanto cosas objetivas (por ejemplo número de hijos, altura o nivel de renta) como subjetivas (opiniones o valoraciones respecto a algo). Para 'observar' dichas variables, o recolectar la información, se suelen utilizar distintas técnicas, como las encuestas o la medición. Como se ha dicho, no hace falta observar todos los sujetos de la población, sino solamente una muestra de la misma. Siempre que la muestra se escoja de manera aleatoria, será posible establecer hasta qué punto los resultados obtenidos para la muestra son generalizables a toda la población.

Población: es el conjunto de sujetos en el que queremos estudiar un fenómeno determinado. Puede ser una comunidad, una región, las beneficiarias de un proyecto, etc.

Sujeto: es la unidad de la población de la que buscamos información. Pueden ser familias, personas, ciudades, etc.

Muestra (aleatoria): subconjunto de sujetos seleccionados de entre la población, a fin de que lo que se averigüe sobre la muestra se pueda generalizar a la población en su conjunto

Veamos estas ideas en un ejemplo: Estamos estudiando el resultado de un proyecto de cooperación de microemprendimientos productivos para mujeres. Para ello hacemos una encuesta a unas cuantas beneficiarias del proyecto seleccionadas al azar (la muestra). Les preguntamos cuánto ha aumentado su renta y si están satisfechas o no con el proyecto. Sale como resultado un aumento de renta promedio de 25$ y un porcentaje de beneficiarias satisfechas del 85%. Ese resultado es exacto para la muestra (las beneficiarias encuestadas). Dado que 'las partes representan al todo' y que la muestra es aleatoria, podemos generalizar el resultado a toda la población (el conjunto de los beneficiarios del proyecto), en este caso con un margen de error del 2% y un nivel de confianza del 95%.

Este ejemplo sirve también para ver los principales elementos de la investigación cuantitativa. En primer lugar, la operacionalización, o traducir lo que se quiere investigar en variables (de resultado del proyecto

hemos pasado a renta y satisfacción). En segundo lugar, el muestreo, o la selección de algunos de los sujetos de entre la población (las beneficiarias escogidas para la encuesta). En tercer lugar, la recolección de la información (realización de la encuesta). En cuarto lugar, el análisis de los datos mediante la estadística descriptiva (cálculo del aumento de renta promedio y del porcentaje de beneficiarias satisfechas). En quinto lugar, la generalización a toda la población mediante la inferencia estadística (calculando el margen de error y el nivel de confianza). Estos elementos clave se irán desarrollando a lo largo de los distintos capítulos del cuaderno, haciendo énfasis en su aplicación en estudios sobre desarrollo.

1.1.2 Perspectiva epistemológica

Los antecedentes de la investigación social empírica suelen ubicarse en los siglos XVII y XVIII, con el surgimiento del movimiento de la *estadística social*. Este movimiento, donde destacan los aritméticos políticos ingleses y la escuela estadística alemana, aplicó la ciencia estadística por primera vez al estudio de los fenómenos sociales, económicos y demográficos. Hasta entonces, dichos procedimientos de medición sólo se utilizaban en las ciencias naturales. De forma general, el principio básico del que se parte es que la sociedad funciona de manera similar a la naturaleza y, por lo tanto, el método científico de las ciencias naturales (basado en la experimentación/observación y las matemáticas) es aplicable también a las ciencias sociales. La realidad social es única, observable y responde a regularidades (leyes universales). Lo cuantitativo es pues clave para conocer la realidad.

Esta perspectiva, conocida como **positivismo**, fue dominante en las ciencias sociales hasta finales del siglo XIX, cuando empezaron a tomar fuerza posturas que disentían en la equiparación del mundo social y natural. Se fue formando así la perspectiva **interpretativista**, que parte de que no existe una única realidad social, sino múltiples realidades que son experimentadas por los distintos agentes. No hay pues unas leyes universales, más bien manifestaciones específicas y singulares, por lo que lo relevante son los aspectos cualitativos, no los cuantitativos.

Aunque estas dos corrientes admiten numerosos matices y son una simplificación de las perspectivas epistemológicas, ha habido durante mucho tiempo y sigue habiendo hoy en día manifiestas diferencias entre positivistas e interpretativistas. Estas divergencias epistemológicas explican en gran parte la falta de entendimiento entre disciplinas académicas, que suelen adscribirse a una u otra corriente. Por ejemplo, los economistas suelen investigar mediante técnicas cuantitativas (encuestas) mientras los antropólogos utilizan técnicas cualitativas (observación, entrevistas, etc.).

En las últimas décadas se han tendido puentes entre ambas corrientes epistemológicas, en lo que ha venido a llamarse el **realismo**. Existe una realidad social independiente al observador, pero ésta no puede ser conocida objetivamente. Así, se puede describir la realidad, pero no aspirar a establecer la verdad sobre ella. Desde esta perspectiva, tanto lo cuantitativo como lo cualitativo tiene relevancia.

Este cuaderno se ubica en una perspectiva realista. Se conciben la metodología y las técnicas cuantitativas de investigación como herramientas significativas para describir la realidad, aunque no puedan abarcarla ni explicarla completamente. Constituyen pues una herramienta irrenunciable para la investigación de procesos de desarrollo, que se enriquece con la combinación de técnicas cualitativas y cuantitativas.

1.1.3 Delimitando 'lo cuantitativo'

El término cuantitativo parece referirse a todo lo que tenga que ver con números, mientas lo cualitativo se relaciona con palabras. Así, además de metodología de investigación cuantitativa, existen variables

cuantitativas, técnicas cuantitativas de recolección, técnicas cuantitativas de análisis… ¿Se refieren a cosas parecidas? ¿Van siempre todas de la mano? La respuesta es ¡no!

Vayamos por partes. La metodología cuantitativa, como se ha explicado anteriormente, es un conjunto de técnicas que se utiliza para estudiar las variables de interés de una determinada población. Se suelen utilizar técnicas de recolección cuantitativas (como las encuestas) y técnicas de análisis cuantitativo (estadística descriptiva e inferencial). Sin embargo, las variables pueden ser tanto cuantitativas (por ejemplo la altura) como cualitativas (por ejemplo el sexo). Por otro lado, las técnicas de análisis cuantitativo también son ampliamente utilizadas para analizar información obtenida mediante técnicas cualitativas como las entrevistas abiertas.

De hecho, autores como Sumner y Tribe (2008) rechazan la dicotomía entre metodología cualitativa y cuantitativa y distinguen cuatro dimensiones relevantes, que sirven para caracterizar las investigaciones de forma menos simplista. Así cada investigación utilizaría (1) técnicas de muestreo aleatorias o intencionales, (2) técnicas de recolección de datos estructuradas o interactivas, (3) información cuantitativa o de percepción y (4) técnicas de análisis estadísticas o sociológicas.

Más allá de utilizar unas categorías u otras, lo importante es ser consciente de los matices que pueden ocultarse tras las categorizaciones más genéricas.

1.2 Los números en el desarrollo

En el ámbito del desarrollo, los números y la estadística juegan un papel vital (ver los primeros 3 minutos de este vídeo): sirven para identificar, priorizar áreas de actuación, analizar evoluciones, fijar objetivos, evaluar indicadores, conocer el impacto, etc. Por ejemplo, en los Objetivos de Desarrollo del Milenio (ODM), cada objetivo va acompañado de una serie de indicadores estadísticos para medir el cumplimiento de las metas planteadas. Otro ámbito donde las estadísticas han sido y son vitales es el del género, donde han contribuido a mostrar las relaciones desiguales entre hombres y mujeres (hojear por ejemplo este artículo de Amartya Sen).

> Meta 2.A: Asegurar que, en 2015, los niños y niñas de todo el mundo puedan terminar un ciclo completo de enseñanza primaria
> Indicadores:
> 2.1 Tasa neta de matriculación en la enseñanza primaria
> 2.2 Proporción de alumnos que comienzan el primer grado y llegan al último grado de la enseñanza primaria
> 2.3 Tasa de alfabetización de las personas de entre 15 y 24 años, mujeres y hombres

La metodología cuantitativa tiene la virtud de plantear una serie de pasos que permiten estudiar un fenómeno de forma estandarizada, acotando en gran medida la interferencia de los sesgos –conscientes o no– del investigador. Además la comunicación de los resultados en forma de estadísticas y gráficos resulta fácil y rápida de entender para el público en general y los tomadores de decisiones. Ese potencial de neutralidad les confiere un halo de objetividad y verdad casi sacrosanta.

Esto puede atraer a personas interesadas en manipular, tergiversando los datos y extrayendo conclusiones interesadas. Por otro lado, puede provocar una 'tiranía de los números', en la que se incurre cuando se les concede un protagonismo excesivo. Una consecuencia de ello es poner en el punto de mira el resultado en vez del proceso y acabar persiguiendo el nivel de indicador marcado en vez del objetivo real, poniendo así en riesgo la sostenibilidad de los cambios. Los ODM son también un ejemplo de esto, pues se han dado casos de países donde se han realizado campañas de matriculación masiva de niños y niñas para alcanzar la meta establecida, pero no se ha prestado atención a si asisten efectivamente a clase.

Esta tiranía se traslada en muchas ocasiones a la investigación en el ámbito del desarrollo. La sed de datos cuantitativos 'objetivos' promueve un uso excesivo de encuestas con amplias muestras aleatorias

para poder obtener una alta precisión en los estimadores. Es posible que un estudio así sea lo mejor para elaborar un informe orientado a la incidencia política. Sin embargo, supone un elevado coste y no suele ser útil para comprender en profundidad ciertas problemáticas, por lo que sería desaconsejable para un estudio orientado por ejemplo al aprendizaje organizacional. La elevada complejidad de los procesos de desarrollo y el contexto de investigación habitual, pueden también dificultar la viabilidad de un estudio de este tipo. A modo de ejemplo, la falta de suficiente información fiable sobre la población a estudiar puede impedirnos realizar un muestreo aleatorio.

Lo importante es, por tanto, conocer adecuadamente las distintas metodologías y técnicas, y aplicar las más adecuadas según el tipo de estudio, los objetivos, los destinatarios, el contexto, los recursos, etc. Desde esta perspectiva, el presente cuaderno busca facilitar una comprensión global sobre la metodología y las técnicas de investigación cuantitativa.

1.3 Metodología de investigación cuantitativa

1.3.1 El proceso de investigación

El siguiente gráfico representa tentativamente 6 pasos generales en los que se podría estructurar una investigación: el problema, diseño, recolección, análisis, interpretación y diseminación. Para cada paso, se detallan algunas de las fases incluidas. El gráfico está particularizado para una investigación en desarrollo que combine técnicas cuantitativas y cualitativas desde una perspectiva epistemológica realista. Además, en el centro del ciclo está, por un lado, la perspectiva epistemológica específica, que determinará en gran medida la forma de realizar la investigación. Por otro lado, la reflexión que debe acompañar el proceso de investigación en desarrollo, en relación con qué visiones se incluyen y quién marca la agenda en la investigación.

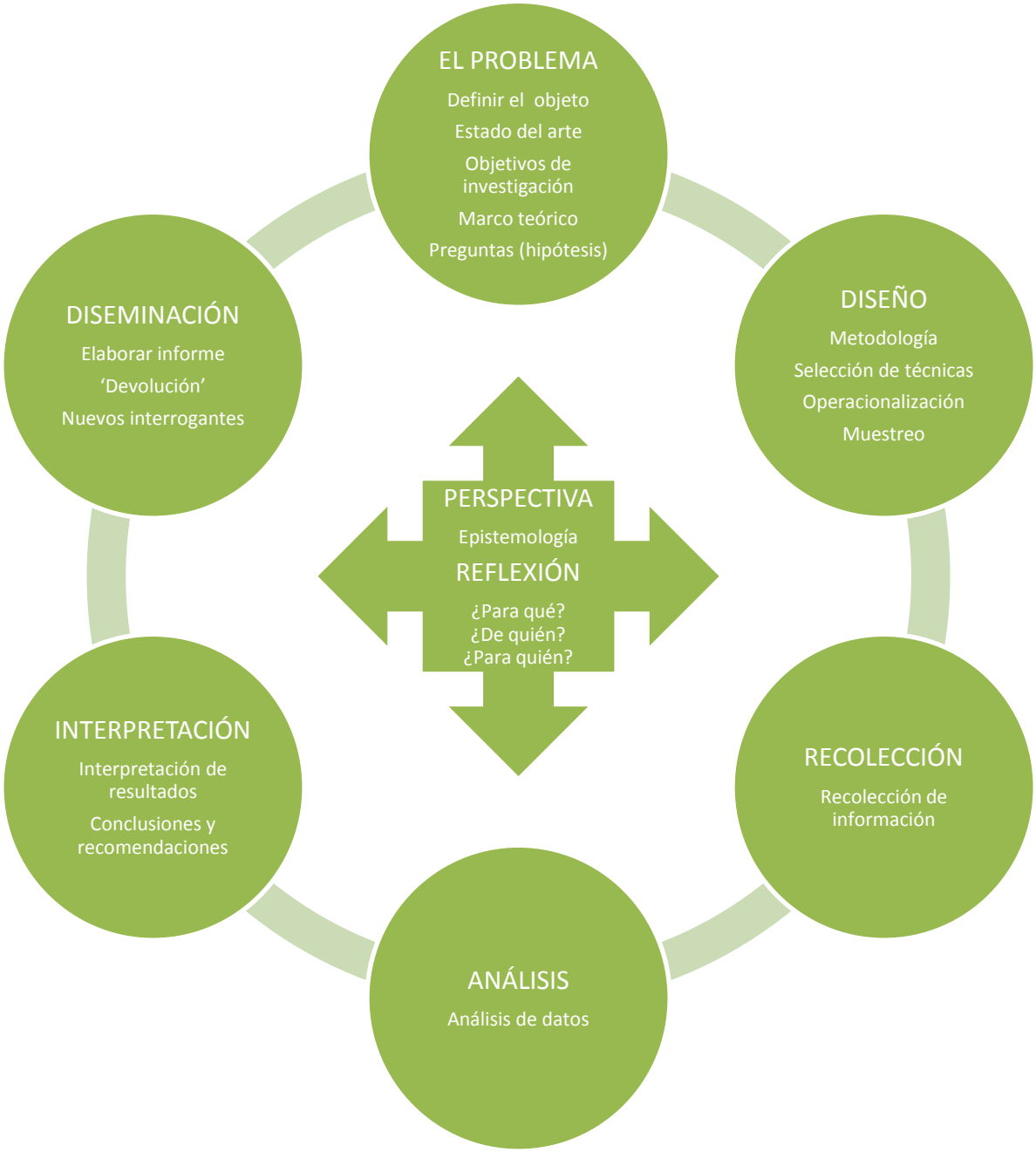


Figura 1: El proceso de investigación
Fuente: elaboración propia

Este cuaderno se centrará solamente en los pasos de diseño, recolección y análisis, por ser los más relevantes a la hora de comprender la metodología de investigación cuantitativa.

Se parte pues del punto en el que EL PROBLEMA de investigación ya está definido, el marco teórico elaborado y las preguntas de investigación planteadas.

El siguiente paso es el DISEÑO. Se debe establecer la metodología que se empleará, seleccionando las técnicas de recolección y análisis de la información. Para la parte cuantitativa de la investigación, será necesario también operacionalizar las preguntas de investigación, convirtiéndolas en indicadores o variables medibles y diseñar el muestreo o selección de unidades que facilitarán la información buscada. Todo esto se verá en el capítulo 2.

A continuación se realiza la RECOLECCIÓN, mediante técnicas de recolección de información. La más habitual en metodología cuantitativa es la encuesta. Este paso se tratará en el capítulo 3.

El paso siguiente es el ANÁLISIS. La información de encuestas o similares se analiza y sintetiza mediante la estadística descriptiva, para luego generalizar esos datos de la muestra a la población realizando estimaciones mediante la estadística inferencial, como se verá en el capítulo 4 y el capítulo 5 respectivamente.

El ciclo se completaría con los pasos de INTERPRETACIÓN, en el que se elaborarían los resultados, y DISEMINACIÓN, en el que se realizaría una devolución a los sujetos involucrados en el estudio y se prepararían materiales para la comunicación de los resultados.

Queda fuera del alcance de este cuaderno el análisis cuantitativo de información obtenida mediante técnicas cualitativas. Por otro lado, la inferencia estadística se limitará a las estimaciones, tratando los contrastes de hipótesis solo de manera superficial, ya que por el alcance y perspectiva de este cuaderno, se confiere el peso principal a la vertiente descriptiva de la metodología cuantitativa, frente a la explicativa. Por la misma razón, no se incluyen análisis de evoluciones a lo largo del tiempo.

1.3.2 Rigor

No hay un consenso claro sobre en qué consiste el rigor o la calidad de una investigación. Un prerrequisito en el que sí hay consenso es que el diseño de la investigación responda a los objetivos planteados. En cuanto al diseño en sí, según la perspectiva epistemológica y la disciplina, se suelen enfatizar unos criterios u otros. Tradicionalmente la metodología cuantitativa (desde una perspectiva positivista) debe cumplir los siguientes cuatro criterios. Nótese que no todos los autores utilizan los mismos nombres para describirlos:

- Validez: la adecuada operacionalización de las preguntas de investigación, de forma que las variables que se estudian sean relevantes y abarquen todas las dimensiones que incorporan las preguntas de la investigación.
- Generalizabilidad: también llamada validez externa, consiste en que la muestra sea representativa de la población. Para ello debe evitar sesgos a través de marcos muestrales adecuados y muestreos aleatorios.
- Fiabilidad: la medición ha de tener la precisión suficiente. Se relaciona con la minimización del error aleatorio y requiere de un tamaño de muestra suficiente.
- Replicabilidad: la posibilidad de que se pueda repetir la investigación y que los resultados no se contradigan.

Desde los estudios de desarrollo, que además suelen utilizar metodologías cualitativas y mixtas, se han propuesto otro tipo de criterios para valorar el rigor o calidad, como son la credibilidad o la contribución a la ciencia o al cambio social.

Más allá de la elección de los criterios, resulta interesante la visión del rigor propuesta por Sumner y Tribe (2008), que lo identifican con la sistematicidad. El rigor pasaría por una buena definición del problema de investigación, así como preguntas de investigación no demasiado amplias, claramente articuladas y alineadas con el problema. Además, la recolección de datos estaría alineada con la pregunta de investigación y el análisis sería consistente, utilizando técnicas estandarizadas y aceptadas. Finalmente, todo el proceso requeriría transparencia, es decir, que se expliciten los pasos de la metodología, de manera que se pueda reconstruir el proceso investigador, y que se reconozcan las limitaciones existentes. Esta visión del rigor es coherente con la posición epistemológica realista.

En la práctica, tener una visión clara del rigor será útil a la hora de diseñar la investigación. En numerosas ocasiones, no será posible cumplir con todos los criterios de calidad deseables en una investigación, por lo que será necesario priorizar. Dicha priorización dependerá de nuestra visión del rigor en la investigación, así como del paradigma y perspectiva epistemológica en el que nos situemos como investigadores. Sin embargo, es también estratégicamente aconsejable tener en cuenta la visión de rigor y la perspectiva epistemológica de los potenciales destinatarios de la investigación (sin necesariamente asumirla) y los fines del mismo. Teniendo en cuenta todos estos ingredientes, podemos diseñar una investigación que responda a la visión resultante.

Veámoslo con una serie de ejemplos:

El primer caso es una investigación que se realiza para poner de manifiesto un problema que no se quiere enfrentar desde la Administración, para así incidir políticamente. La perspectiva de aquellos en los que queremos incidir (Administración, público en general) probablemente será positivista. Puede por tanto ser interesante realizar una investigación cuantitativa, cumpliendo con los criterios de generalizabilidad y fiabilidad (muestra aleatoria y suficientemente grande), dado que la ortodoxia –y el aval de la teoría estadística– puede ser un valor estratégico para una investigación de este tipo. Esto no excluye la utilización de otras técnicas. Y tanto o más importante será la presentación de los resultados a través de materiales específicos para la incidencia política (tipo *policy brief*).

El segundo caso es una investigación que se inserta en un proceso de aprendizaje local. El criterio de rigor relacionado con la contribución (al aprendizaje en este caso) resulta probablemente más relevante que la replicabilidad. Quizá esto nos lleve a priorizar técnicas participativas, tanto para la cuantificación como para aspectos más cualitativos.

Un tercer caso sería una evaluación de un programa de desarrollo en una determinada área. Si quién encarga la evaluación la entiende como medir el impacto con una serie de indicadores, valorará los criterios tradicionales vistos anteriormente y una encuesta será suficiente. Si lo entiende como oportunidad de aprendizaje, puede que vea bien reducir el tamaño de la muestra y con ella la fiabilidad, para dedicar esos recursos a otras técnicas más interactivas y de profundización. El investigador no tiene porqué plegarse a la visión de quién encarga la evaluación, pero desde luego le será útil ser consciente de ella.

La casuística es interminable. Muchas veces ocurrirá también que no será posible cumplir con el rigor ‘tradicional’ al aplicar técnicas cuantitativas en contextos de desarrollo, donde se carece habitualmente de información sobre la población a estudiar. En estos casos, lo principal es reconocer y ser transparente en cuanto a las limitaciones metodológicas. Otra opción interesante en esos casos es la triangulación, es decir, la complementación del estudio con información obtenida mediante otras técnicas (generalmente cualitativas), para comprar los resultados de ambas técnicas.

Capítulo 2. Diseño de investigación

Objetivos del capítulo:	Tiempo estimado de lectura: 60 min
• Seleccionar técnicas de investigación - incluidas las cuantitativas- adecuadas al contexto de la misma • Elaborar cuestionarios que ayuden a responder a los aspectos investigados • Seleccionar muestras, utilizando métodos y tamaños adecuados a los objetivos de inferencia planteados	Apartados del capítulo: 2.1 Selección de metodologías y técnicas 2.2 Operacionalización 2.3 Muestreo
Capítulo anterior. La investigación cuantitativa Índice Capítulo siguiente. Recolección de información: la encuesta	

2.1 Selección de metodologías y técnicas

El diseño de la investigación incluye en primer lugar la selección de la metodología de investigación y las técnicas de recolección y análisis de la información. En segundo lugar, la operacionalización de las preguntas de investigación, convirtiéndolas en variables. En tercer lugar, el muestreo.

La metodología es la estrategia de investigación que elegimos para responder a las preguntas de investigación. Dependerá tanto de éstas como del marco teórico de la investigación. Se trata pues de optar por una estrategia de investigación general, ya sea de índole cuantitativa, cualitativa o mixta. También el nivel de complejidad y detalle deseado (desde extensa al caso de estudio) o el nivel de participación que se pretende (desde lo extractivo hasta la investigación acción participativa). En segunda instancia, se escogerán las técnicas de recolección (por ejemplo la encuesta) y las técnicas de análisis, esto es, las herramientas más específicas de investigación. Éstas también dependen de las preguntas y del marco teórico y deben ser coherentes con la metodología.

Resulta difícil aportar un criterio claro sobre qué metodología y técnicas escoger. En relación a cuándo optar por una metodología cualitativa o por una cuantitativa, un aspecto a tener en cuenta es la profundidad con la que deseamos comprender el fenómeno estudiado. Las técnicas cuantitativas son especialmente útiles para obtener una imagen general en base a ciertas magnitudes de interés. Se puede visualizar como una foto que permite apreciar todo el bosque. En cambio, las cualitativas permiten profundizar en determinados aspectos que quizá ni se tenían en cuenta inicialmente. Son técnicas que hacen *zoom* en un árbol determinado, indagan qué hay escondido detrás del follaje y amplían la perspectiva, permitiéndonos entender cómo otros ven ese árbol o bosque. Para preguntas de investigación exploratorias o descriptivas se suele tender a la metodología cuantitativa, mientras que para preguntas de mayor detalle y profundización, se suele recurrir a cualitativa.

En la práctica, en muchas investigaciones habrá preguntas de ambos tipos, y para obtener un buen estudio es muy probable que sea necesario el uso de una metodología que combine técnicas cuantitativas y cualitativas. Es muy común realizar un estudio general mediante encuesta y después utilizar técnicas cualitativas (entrevistas o grupos focales) para analizar aspectos que emergen de las encuestas. En otros casos, el proceso es inverso y se utiliza la observación participante o entrevistas para comprender el fenómeno estudiado y a partir de ahí se realiza una encuesta sobre los aspectos más relevantes.

2.2 Operacionalización

Centrándonos ya en la metodología cuantitativa, el primer paso consiste en traducir las preguntas o hipótesis de investigación en indicadores o variables que luego se 'medirán' mediante la técnica de recolección de información que se elija. En función de la abstracción de la pregunta de investigación, habrá más o menos niveles de concreción entre la pregunta y la variable.

Variable: Característica que se pretende estudiar, es decir, lo que queremos conocer del sujeto investigado. Ejemplos: nivel de renta o religión.

En muchos casos, será conveniente concretar las preguntas de investigación en sub-preguntas o preguntas de nivel inferior. A partir de cada pregunta de último nivel, se establece una serie de dimensiones o conceptos relevantes. Estos se concretan en una serie de variables.

Dichas variables se 'recolectarán' en la fase de trabajo de campo mediante las técnicas que se consideren oportunas.

Pongamos como ejemplo un estudio mediante encuesta en una comunidad cuya pregunta de investigación sea: ¿Cómo son los hábitos y prácticas diarias de la comunidad vinculados a la salud? Esta pregunta podría concretarse en varias sub-preguntas relacionadas con el tratamiento de enfermedades, las prácticas de manipulación de alimentos, el manejo de los animales, el uso y transporte de agua o las prácticas higiénicas. La última sub-pregunta sería: ¿En qué medida tienen las familias prácticas de higiene saludables? Dicha pregunta se concretaría en distintas dimensiones y éstas a su vez en una serie de variables, tal y como se ve a continuación

Sub-pregunta: ¿En qué medida tienen las familias prácticas de higiene saludables?

Dimensiones	Variables
D1. Lavado de manos	V1.1 # veces que se lavan las manos al día
	V1.2 Momentos del día en los que se lavan las manos
	V1.3 Medios que utilizan para lavarse las manos
D2. Uso de letrinas	V2.1 Tipo de letrina a la que se tiene acceso
	V2.2 # personas con las que se comparte la letrina
	V2.3 Estado físico de la letrina
	V2.4 Limpieza de la letrina
	V2.5 Proporción de veces que se usa la letrina respecto a las veces que se practica la defecación al aire libre
D3. Gestión de residuos	V3.1 Lugar de descarga del agua utilizada/sucia
	V3.2 Destino de residuos orgánicos
	V3.3 Destino de residuos no orgánicos

Cada una de las sub-preguntas del estudio se descompondría en dimensiones y variables de forma análoga. A cada variable le correspondería luego una pregunta en la encuesta.

Para realizar una buena operacionalización, es importante conocer bien la temática tratada (revisión bibliográfica y experiencia) y disponer de un marco teórico robusto. Las dimensiones deben ser relevantes para la pregunta, y el conjunto de dimensiones de una pregunta debe abarcar todos los aspectos clave de la misma. Ocurre análogamente con las variables respecto a su dimensión.

El paso de preguntas a dimensiones se utiliza también a menudo para técnicas cualitativas, por ejemplo para elaborar las preguntas en una entrevista semiestructurada. No obstante, el paso a variables es más específico de la metodología cuantitativa, pues en cualitativa no interesa concretar tanto, sino que suele interesar una mayor amplitud que permita profundizar en la comprensión del fenómeno, así como obtener respuestas no esperadas que abran nuevos caminos de investigación.

2.3 Muestreo

2.3.1 Conceptos básicos

Una vez seleccionadas las técnicas y operacionalizadas las preguntas de investigación, la última fase del diseño metodológico es la selección de los sujetos a estudiar: el muestreo.

El muestreo consiste en seleccionar una serie de sujetos para obtener información de ellos. En investigación cuantitativa, el muestreo se suele realizar con la intención de que el análisis de la muestra sirva para tener una idea más o menos aproximada de la población de la que proviene la muestra.

Repasemos los conceptos: la **población** es el conjunto de todos los sujetos, sobre los que queremos conocer cierta información relacionada con el fenómeno que se estudia. Se pone como ejemplo, una investigación sobre el nivel de ingresos familiar de la región Logone Occidental del Chad. Las familias serían los sujetos y la población sería el conjunto de familias de dicha región.

La **muestra** es el subconjunto de la población que se selecciona para el estudio, esperando que lo que se averigüe en la muestra nos dé una idea sobre la población en su conjunto. Se seleccionan muestras porque normalmente no es posible o económico estudiar todos y cada uno de los sujetos de una población (lo que sería un censo). Siguiendo con el ejemplo anterior: como sería muy caro averiguar el nivel de ingresos de todas las familias de Logone Occidental (casi 700.000 habitantes), lo normal es seleccionar unas cuantas familias (la muestra), y realizar una encuesta sobre el nivel de ingresos. A partir de los datos obtenidos se obtendría el ingreso medio muestral.

La muestra, en el caso de estudios estadísticos, descansa en el principio de que las partes representan al todo. Así, una muestra reflejará las características que definen la población de la que fue extraída. Por lo tanto, se podrían generalizar las características de la muestra a toda la población utilizando la estadística inferencial. En el caso de Logone, la inferencia nos daría información sobre la precisión con la que el ingreso medio muestral representa el ingreso medio de toda la población. Esta información de precisión se concreta en este caso en un intervalo de confianza o margen de error y un nivel de confianza o probabilidad de acertar, como veremos en el apartado 5.2.

Inferencia estadística: es el proceso de aplicar métodos estadísticos para sacar conclusiones sobre una población a partir de datos de una muestra.

Sin embargo para poder aplicar la inferencia, es decir, para poder generalizar, la muestra debe reflejar las características de la población. Para ello, debe cumplir dos condiciones.

En primer lugar, debe ser suficientemente grande (en el capítulo 5 se explican los cálculos del tamaño de muestra).

En segundo lugar, debe ser seleccionada de manera aleatoria. El muestreo se considera aleatorio (o probabilístico) cuando todos los sujetos tienen la misma posibilidad de ser escogidos para la muestra. Sería como poner todos los nombres de los sujetos en un bombo y e ir extrayéndolos al azar. En la práctica, hay diferentes tipos de muestreos aleatorios: simple, sistemático, estratificado y por etapas. En los dos últimos, no todos los sujetos tienen la misma probabilidad de formar parte de la muestra, pero como sabemos qué probabilidad tiene cada sujeto, podemos corregir la desviación mediante ponderaciones, así que se considera igualmente aleatorio.

Muestreo aleatorio: técnica de muestreo en la que cada uno de los sujetos de la población tiene la misma probabilidad de ser incluido en la muestra.

En contraposición, están los muestreos no aleatorios, más propios de técnicas cualitativas. Éstos, ni son aleatorios, ni pretenden obtener una muestra representativa de la población. Más bien, buscan seleccionar sujetos que constituyan casos paradigmáticos (primando la diversidad) o que tengan especial conocimiento sobre una cuestión (informantes clave). Se prima la calidad frente a la cantidad.

Existe un tercer grupo, que podríamos denominar pseudoaleatorio. Son muestreos que no se pueden considerar aleatorios, pero que sí pretenden obtener una muestra tan representativa de la población como sea posible, por ejemplo el muestreo por cuotas.

Otros dos conceptos importantes son:

El **marco muestral** es el conjunto de sujetos de la población realmente disponibles para la elección de la muestra. Debería coincidir con la población, pero no siempre es así, sobre todo, en los contextos de estudios de desarrollo. En Logone Occidental, el marco muestral sería completo si se dispusiera de una lista actualizada de todas las familias de la región. A partir de ahí se seleccionaría la muestra. En cambio, si se parte del listín telefónico (poco aconsejable en este caso), el marco muestral no son todas las familias de la región, sino solo las familias de la región que tienen teléfono. La disponibilidad o no de un marco muestral adecuado es importante, ya que determina las técnicas de muestreo a aplicar. En ocasiones, cabe la posibilidad de reconstruirlo (elaborar la lista de la población), como paso previo al muestreo.

La **unidad muestral** es el elemento individual que constituye el marco muestral, y sobre el que se obtendrá información. Normalmente es lo mismo que el sujeto (las familias en el ejemplo de Logone Occidental), si bien se pueden dar excepciones. Sería el caso, volviendo al mismo ejemplo, que se hiciese una encuesta por hogares, con lo que la unidad muestral sería el hogar y no la familia.

2.3.2 Muestreos aleatorios

El **muestreo aleatorio simple** consiste en escoger los sujetos de la población al azar, uno por uno. Requiere disponer de un marco muestral adecuado. En caso de tener un archivo electrónico con la lista de la población, los programas estadísticos pueden hacer dicha selección. En caso contrario, y para una población pequeña, se puede asignar un número a cada sujeto de la población y extraer números aleatorios mediante ordenadores, calculadoras, tablas o incluso un bombo hasta completar el tamaño de muestra deseado. Esta animación representa este tipo de muestreo. No es una técnica muy empleada en los estudios en el ámbito del desarrollo, dada la necesidad de una lista de la población a estudiar y su poca practicidad para grandes tamaños de muestra.

Para poblaciones grandes, se suele utilizar el **muestreo aleatorio sistemático**. Requiere disponer de un marco muestral adecuado. Se asigna un número a cada sujeto de la población, igual que en la anterior, y se extrae un solo número al azar. El sujeto al que corresponde ese número es el primero de nuestra muestra. A partir de él, se van tomando los siguientes sujetos, dejando un intervalo determinado entre

ellos. El tamaño de ese intervalo (k) se calcula dividiendo el tamaño de la población (N) entre el tamaño de muestra deseado (n). $k = N / n$. Esta animación reproduce este tipo de muestreo. Hay que prestar atención a que la lista numerada de sujetos no tenga ninguna periodicidad. Esta técnica se emplea en investigaciones en el ámbito de desarrollo cuando se dispone de una lista de la población a estudiar.

El **muestreo aleatorio estratificado** consiste en dividir la población de estudio en grupos o clases (estratos), que se suponen homogéneos con respecto a las características a estudiar. Esta homogeneidad debe existir dentro del estrato, pero no entre estratos. Para cada estrato se asigna una cuota que representa el tamaño de muestra de ese estrato, y se realiza un muestreo aleatorio sistemático. Este tipo de muestreo pretende dotar de mayor representatividad a la muestra, asegurándose de que los distintos estratos están representados adecuadamente en la muestra. Se puede estratificar, por ejemplo, según el sexo o la profesión. La lógica de los estratos tiene que ser coherente con lo que se busca. Si estudiamos el nivel educativo, se puede estratificar según el origen étnico, pero no tiene sentido estratificar por si se es zurdo o diestro. Es probable que el nivel educativo sea similar entre miembros de la misma etnia (homogeneidad en el estrato) y difiera con respecto a otras etnias (heterogeneidad entre estratos). El muestreo estratificado requiere un marco muestral muy detallado pues necesitamos, además de la lista de nombres, información de las características respecto a las que queremos estratificar.

Dentro del muestreo estratificado, existen variantes. La más común es la afijación proporcional, donde el tamaño de la muestra de cada estrato es proporcional al tamaño del estrato dentro de la población. Por otro lado está la afijación no proporcional, donde ciertos estratos están sobrerrepresentados en la muestra. Esta animación representa ese proceso. Con la afijación no proporcional se busca, por ejemplo, aumentar la representación de un estrato clave que por su pequeño tamaño podría estar muy poco representado en un muestreo no estratificado. Por ejemplo, se podría estratificar por etnias o religión aumentando el tamaño del estrato de la etnia o religión minoritaria. Cuando se opta por afijación no proporcional, para combinar los datos entre estratos será necesario ponderarlos, asignando un peso según la proporción de ese estrato en la población (ver cálculo de la media ponderada en el apartado 4.3.3). El muestreo estratificado se suele usar en estudios en el ámbito del desarrollo –cuando se dispone de información previa sobre la población– por su potencialidad a la hora de prestar atención a minorías.

El **muestreo aleatorio por etapas o conglomerados** consiste en seleccionar primero subdivisiones de la población –los conglomerados– y luego muestrear sujetos de los conglomerados elegidos. Un conglomerado es una subdivisión pre-existente o natural de la población, como la provincia o el distrito electoral. Un conglomerado debe ser heterogéneo en sí mismo; idealmente contiene toda la variabilidad de la población.

El más sencillo consta de una primera etapa en la que se muestrean los conglomerados y una segunda etapa en la que se estudian todos los sujetos de los conglomerados seleccionados (no se muestrea). Por ejemplo, si la población a estudiar es el profesorado de primaria de la ciudad, la primera etapa sería escoger unas cuantas escuelas (conglomerados) aleatoriamente y encuestar a todos los profesores y profesoras de las escuelas escogidas.

En muchas ocasiones, hay más etapas (muestreo polietápico), y se muestrea en varios niveles sucesivamente. Los conglomerados de cada etapa pueden ser, por ejemplo, regiones administrativas, áreas geográficas, edificios… En cada etapa, el muestreo puede ser simple, sistemático o estratificado.

Por ejemplo, si ahora la población a estudiar es el profesorado de primaria de todo un país, se pueden crear dos niveles de conglomerado: provincias y escuelas. En una primera etapa, se extrae una muestra aleatoria simple de provincias del país. En una segunda etapa, se extrae una muestra aleatoria de escuelas para cada provincia seleccionada, a partir del listado de escuelas disponible en las administraciones

provinciales. En la tercera etapa, a partir del listado de cada escuela seleccionada (facilitado por el director), se extrae una muestra aleatoria de los profesores a encuestar.

El muestreo por etapas se considera aleatorio si los conglomerados son heterogéneos en sí mismos y homogéneos respecto a otros conglomerados. En el nivel de conglomerado de escuelas, se concretaría en que no existan grandes diferencias entre una escuela y otra (por ejemplo, que tengan currículos similares), y dentro de cada escuela haya diversidad (por ejemplo, que haya profesores de distintas edades, con formaciones diversas, que impartan desde el primer al último curso, etc.).

Nótese que los estratos y las etapas parten, en cierto sentido, de ideas opuestas. La estratificación funciona correctamente cuando dentro del estrato hay homogeneidad, y a su vez los estratos son muy diferentes entre sí. Por el contrario, en el muestreo por etapas los conglomerados deben ser parecidos entre sí y presentar heterogeneidad dentro del propio conglomerado.

El muestreo por etapas se utiliza cuando no se dispone de una lista completa de la población a estudiar, ya que es más factible construir el marco muestral de los conglomerados seleccionados. Al establecer los niveles, solo hace falta una lista completa de los conglomerados de primer nivel, después de los seleccionados del segundo y así sucesivamente. En el ejemplo, necesitamos la lista de provincias, luego de entre las provincias seleccionadas necesitaremos la lista de escuelas y de las escuelas seleccionadas la correspondiente lista del profesorado. Esto es mucho más sencillo que conseguir una lista de todo el profesorado del país. Además, un muestreo a partir de la lista nacional generaría una muestra con sujetos tan distribuidos que acceder a ellos resultaría prohibitivamente caro.

El muestreo por etapas es uno de los más empleados en estudios en el ámbito del desarrollo, debido a las limitaciones que suelen estar presentes en estos estudios, como son la falta de información precisa sobre la población a estudiar o la falta de recursos para acceder a muchos lugares dispersos. Para el caso de Logone, se concretaría por ejemplo en seleccionar algunos departamentos, dentro de los departamentos algunas comunidades y ya en las comunidades elaborar un listado de familias y seleccionar algunas (o todas).

Hay que tener presente que si en alguna etapa el muestreo no es proporcional al tamaño del conglomerado, se deberá utilizar la ponderación para compensar los pesos (ver cálculo de la media ponderada en el apartado 4.3.3).

2.3.3 Muestreos pseudoaleatorios

Hay ciertos muestreos que no se pueden considerar aleatorios, pero que sí pretenden, en cierta medida, ser representativos, por lo que los denominaremos pseudoaleatorios. Se emplean muchas veces cuando no se dispone de marco muestral y es difícil de construir. Dentro de los pseudoaleatorios, se pueden conseguir distintos niveles de aleatoriedad, o mejor dicho, la selección de la muestra puede depender más o menos de la arbitrariedad del investigador o investigadora. Desde una visión del rigor amplia (ver apartado 1.3.2), estos muestreos pueden utilizarse cuando las condiciones de la investigación impidan realizar muestreos aleatorios. En cualquier caso, siempre se debe especificar la técnica utilizada y ser consciente de que teóricamente, no se les pueden aplicar los cálculos inferenciales, es decir, no es posible cuantificar la precisión de nuestros datos ni el tamaño de muestra requerido (ver apartado 5.2.4). Seguidamente, se presentan tres técnicas, de mayor a menor aleatoriedad.

Hay una técnica llamada **muestreo por áreas**, que permite introducir un cierto grado de aleatoriedad (espacial) cuando no se tiene información exacta de la población estudiada. Es como el muestreo por etapas, pero sin disponer de listas siquiera dentro de los niveles. Así, en el primer nivel de conglomerado

(o en aquel en que carezcamos de listado), se introduce una subdivisión previa en áreas geográficas aleatorias. Esto se haría a partir de un mapa en el que esté toda la zona a investigar, dibujando aleatoriamente líneas que la dividan en esas pequeñas áreas al azar y seleccionando aleatoriamente algunas de esas áreas. Si las áreas son suficientemente pequeñas, será posible listar a toda la población del área para hacer un muestreo aleatorio en dicha etapa. Si no, se podría introducir otra etapa identificando las comunidades existentes y seleccionándolas aleatoriamente, para luego muestrear (aleatoriamente a ser posible) dentro de las comunidades seleccionadas.

Figura 2: Mapa para muestreo por áreas
Fuente: elaboración propia

Aunque no es 100% aleatoria, esta técnica permite obtener cierto nivel de representatividad y puede ser útil a la hora de realizar el trabajo de campo en zonas relativamente pequeñas de cuya población no se haya podido obtener información previa.

El **muestreo por cuotas** es el muestreo pseudoaleatorio más utilizado, siendo el caso paradigmático los sondeos electorales en nuestro país. Como en el muestreo estratificado, se establecen estratos, que se suponen homogéneos y se asigna una cuota o tamaño de muestra de ese estrato, que ha de ser proporcional a su presencia en la población. Sin embargo, el muestreo no se hace a partir de un listado poblacional, sino que se deja al libre albedrío del encuestador o encuestadora, que tiene libertad para elegir a los sujetos (normalmente, los primeros que pasen por el lugar donde se ubica), siempre que cumpla con las cuotas de cada estrato. Por ejemplo, para muestrear por cuotas a 100 personas para una encuesta, se podría estratificar por sexo y grupo de edad, y pedir al encuestador que pregunte a 25 hombres y 25 mujeres menores de 41 años, y a 27 mujeres y 23 hombres mayores de 41 años. Estas características se irían comprobando sobre la marcha. Cuando se complete el cupo de 48 hombres, solo se encuestará a mujeres.

Como se comentaba, este muestreo es muy común en estudios de mercado y sondeos de opinión para medir la evolución de las preferencias de la gente. Aunque no sea aleatorio, si se mantiene el mismo sistema de muestreo en sucesivos sondeos, suele estimar con relativa precisión dicha evolución. En estudios en desarrollo se puede emplear cuando no sea posible realizar un muestreo aleatorio. Es conveniente dotarle de alguna forma de mayor aleatoriedad, por ejemplo, estableciendo rutas para que el encuestador tome la muestra en distintos puntos.

El **muestreo intencional** tiene un grado muy bajo de aleatoriedad, pues el equipo investigador determina la muestra según su propio criterio, aunque siempre con la intención de obtener una muestra más o menos representativa de la población. Este muestreo se da también en estudios en desarrollo, cuando se carece no solo de marco muestral, sino también de información sobre la población estudiada (mapas, grupos de población). Esto impide utilizar muestreos por cuotas o áreas. El equipo investigador, una vez llega a terreno y se hace una imagen general de la composición de la población, elige la muestra intentando minimizar los sesgos. Por ejemplo, se podría intentar que haya tantos hombres como mujeres y que haya personas de los distintos grupos sociales existentes; sería como un muestreo por cuotas aproximado, ya que no se conoce la proporción de las distintas cuotas en la población. Este muestreo se recomienda únicamente para estudios previos o que no requieran mucha precisión y requiere que cierto conocimiento de la población en el momento del muestreo.

2.3.4 Muestreos no aleatorios

Hay muestreos que ni son aleatorios ni pretenden ser representativos de la población. Estos muestreos se utilizan en estudios de carácter cualitativo, en los que el interés no es la generalización sino descubrir un significado o reflejar realidades múltiples. También pueden ser útiles como estudio previo para conocer la población y después aplicar otras técnicas de muestreo aleatorio. Por ello, y por lo importante que resulta combinar metodologías, se presentan dos de estas técnicas.

El **muestreo de bola de nieve** está indicado para estudios de poblaciones minoritarias, excluídas o invisibilizadas, como personas inmigrantes sin papeles, niñas y niños de la calle, personas con ciertas enfermedades, etc. Consiste en ir identificando los sujetos de la muestra a medida que se realizan las entrevistas. Así, se parte de unos pocos individuos de la población a los que se pueda acceder, y a través de ellos se logra contactar con otros sujetos con características similares, y así sucesivamente. Véase un ejemplo en esta animación.

En el **muestreo subjetivo o de juicio**, los sujetos de la muestra se eligen de forma razonada, en función del objetivo perseguido, y sin importar la representatividad respecto a la población. Hay muchos tipos de muestreo subjetivo, como el muestreo de casos típicos, de diversidad (sujetos que no se asemejen a la media), de informantes clave o de casos de éxito. Estos muestreos son muy comunes en estudios de sistematización o en los estudios de caso.

2.3.5 Error aleatorio y sesgos

Como se ha recalcado en los apartados anteriores, en cuantitativa el muestreo tiene como fin obtener una muestra que sea representativa de la población, es decir, que refleje las características de la misma. Esto nunca se conseguirá al 100%, pues existen dos enemigos acérrimos de las muestras que se lo impiden. Son el error aleatorio y el sesgo muestral.

El **error aleatorio** es natural e inevitable. Que una muestra aleatoria refleje más o menos las características de la población, es en parte cuestión de azar. Habrá siempre una imprecisión en las estimaciones que realicemos: el error aleatorio. Pero gracias a la inferencia estadística, podemos cuantificar esa imprecisión, y dimensionar las muestras (¡aleatorias!) para minimizar el error.

> **Error aleatorio**: imprecisión en la estimación de una variable al calcularse a partir de una muestra en lugar de a partir del conjunto de la población.

En un ejemplo extremo, si se pregunta a dos personas de una comunidad de 500 personas su satisfacción respecto a un proyecto, por muy aleatoria que sea la selección, es poco probable que su opinión represente la de toda la comunidad. El error aleatorio sería muy grande.

Este error es inevitable, siempre está ahí y afecta a cualquier tipo de muestreo. Pero si el muestreo es aleatorio, la estadística inferencial permite cuantificarlo, y minimizarlo por medio del aumento del tamaño de la muestra. Así que el error aleatorio es un enemigo relativamente fácil de manejar.

El **sesgo muestral** es un enemigo más peligroso. Ocurre cuando hay sujetos que son excluidos a priori de la muestra, es decir, que son parte de la población, pero no aparecen en el marco muestral. Es generalmente evitable, y se debe evitar, pues no tenemos herramientas para cuantificarlo y controlarlo, como en el caso del error aleatorio.

> **Sesgo muestral**: distorsión que se introduce debido a la forma en que se selecciona la muestra.

El sesgo muestral es frecuente en todo tipo de investigaciones, más aún si cabe en estudios en desarrollo, debido a diversos problemas relacionados con el marco muestral.

Muchas veces no existe o no se puede acceder al marco muestral ideal, que es una lista con todos los sujetos de la población. En esos casos, el muestreo se hace a partir de otros medios. Un ejemplo sería la lista de residentes en un determinado municipio, marco muestral que excluiría a residentes en asentamientos informales. Otro ejemplo de sesgo (esta vez causado por la practicidad) es el de las encuestas telefónicas, que usan el listín telefónico como marco muestral, dejando fuera a los que no tienen teléfono.

El sesgo es también relevante para muestras pseudoaleatorias, pues reduce la ya de por sí cuestionada representatividad de estas técnicas. Por ejemplo, en un muestreo por cuotas en una comunidad, si la encuestadora va acompañada de un miembro de una comunidad que indica a qué personas realizar la encuesta, se está introduciendo un sesgo considerable. También cuando se hacen las encuestas en horarios concretos, dejando fuera a los que están trabajando en el campo, por ejemplo.

En el momento de la recolección de información también se pueden introducir sesgos –aunque técnicamente no serían sesgos muestrales– que reducen la representatividad. Por ejemplo, si se utilizan encuestas escritas, las personas analfabetas no pueden participar adecuadamente. O cuando los entrevistadores no hablan las lenguas locales, excluyendo a indígenas monolingües.

También si hay personas que no quieren responder la encuesta o alguna pregunta, se genera una cierta distorsión. Las razones pueden ser diversas. Una podría ser que haya personas que tienen miedo a responder a preguntas que consideren sensibles (¿Cuántas hectáreas de tierra posee?). En cualquier caso, las 'no respuestas', más allá de distorsionar el marco muestral, son muchas veces una pista que indica que se han tocado temas sensibles o sobre los que hay conflictos.

Debido a preguntas mal formuladas o a que la gente no recuerde bien el asunto que se investiga, se puede recoger información errónea. Además, cuando las personas encuestadas intuyen –sea cierto o no– que la encuesta sirve para priorizar o identificar intervenciones de cooperación, es fácil que presenten una visión distorsionada (para mejor o para peor) de la realidad.

Finalmente, hay un último sesgo, relacionado tanto con los ejemplos de exclusión del marco muestral como con este último referente al momento de la recolección de la información: ocurre cuando el sujeto o unidad muestral no es la persona individual, sino el hogar o la familia, o incluso la comunidad. En muchas ocasiones, esa familia o comunidad se convierten en una caja negra, y no importa quién sea el que ha respondido; lo que haya dicho se da por válido para la familia o comunidad. Sin embargo, parece obvio que cada miembro de la familia responderá de manera diferente, sobre todo a ciertas temáticas. Este aspecto está muy vinculado al enfoque de género, puesto que mujeres y hombres (que son los que suelen responder como cabezas de familia) suelen tener visiones distintas sobre la situación familiar y las necesidades en su hogar. Esto puede llevar a que preguntas aparentemente neutrales, como la distribución de los gastos familiares, arrojen resultados dispares según si son respondidas por unas u otros. Para evitar este sesgo, se pueden registrar las características de la persona entrevistada, intentar que haya un equilibrio muestral en cuanto al sexo (o edad, o rol en el hogar) y después analizar posibles diferencias.

Como se ha dicho al principio, los sesgos se deben evitar en la medida de lo posible. Cuando no hay alternativa, es importante reconocer ese sesgo de manera transparente a la hora de presentar los resultados de la investigación. Así queda claro sobre qué población se pueden considerar representativos los resultados (personas residentes en asentamientos formales, hogares que tienen teléfono o incluso limitarlo a 'personas encuestadas'). Hay que estar atentos a los sesgos, ya que los más peligrosos son aquéllos de los que no somos conscientes.

Resumiendo, se ha visto que la representatividad de la muestra y, por tanto, la posibilidad de generalizar lo observado a toda la población, se ven amenazadas por el error aleatorio y los sesgos que se dan en el

muestreo y la recolección de información. Esto supone una llamada de atención sobre lo importante que es realizar un buen proceso de muestreo y recogida de información. La clave para reducir el sesgo muestral es intentar que el marco muestral incluya a toda la población. Finalmente, se empleará la inferencia para calcular el error aleatorio y tomar un tamaño de muestra que lo minimice.

Capítulo 3. Recolección de información

Objetivos del capítulo:	Tiempo estimado de lectura: 130 min
• Elaborar cuestionarios que ayuden a responder a los aspectos investigados • Recordar técnicas de investigación cuantitativa relacionadas con perspectivas de participación comunitaria	Apartados del capítulo: 3.1 Fuentes documentales y estadísticas 3.2 Medición y observación sistemática 3.3 Técnicas participativas 3.4 La encuesta

Capítulo anterior. Diseño de investigación Índice Capítulo siguiente. Estadística descriptiva

Una vez completado el diseño de la investigación, llega el momento de recolectar la información que se ha identificado como relevante, es decir, las variables identificadas.

Los principales métodos para la obtención de información se suelen clasificar en cualitativos o cuantitativos, aunque quizá sería mejor caracterizarlos según se recoja la información de manera más estructurada y cerrada (cuantitativos) o más abierta (cualitativos) y situarlos en un continuo en lugar de en dos grupos estancos.

Entre las técnicas de recolección de información consideradas cuantitativas destaca la encuesta. Por ello, se trata de forma más extensa. Sin embargo, hay otras técnicas de recolección cuantitativa relevantes como el uso de fuentes secundarias, la medición, la observación sistemática o los métodos participativos/visuales, a los que también es interesante que prestemos atención.

3.1 Fuentes documentales y estadísticas

La recolección de información se realiza a través de internet, bibliotecas, organismos, etc. Consiste en obtener información ya recolectada previamente, es decir, de fuentes secundarias, para luego analizarla estadísticamente. Dicha información suele presentarse en bases de datos estadísticos, que son unas tablas en las que se organizan en filas y columnas los sujetos (personas de un municipio, empresas, países…) y algunas de sus características (edad, facturación, PIB) para distintos puntos temporales.

Base de datos: también denominada banco de datos, es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

Las principales bases de datos sobre desarrollo a nivel internacional son UNdata de Naciones Unidas y el Banco Mundial. A nivel nacional, las agencias estadísticas suelen ser la fuente más completa. En el caso de España es el Instituto Nacional de Estadística.

Algunas bases de datos son interactivas y permiten al usuario crear sus propios índices o tablas con indicadores que les interesen. Un ejemplo es esta herramienta del PNUD. Otros ejemplos, que incluyen además potentes herramientas de visualización son GapMinder, que anima en el tiempo la evolución de hasta 4 variables, y WorldMapper, que crea mapas proporcionales al indicador de interés.

3.2 Medición y observación sistemática

La medición consiste en utilizar aparatos de medición para determinar la magnitud de un indicador o variable de interés. Es muy común en estudios de nutrición, en los que se pesa bebés o se mide el diámetro del brazo a niños y niñas.

La observación sistemática es un procedimiento por el cual se recoge información observable sobre un determinado aspecto de interés y de acuerdo a un procedimiento establecido. Un ejemplo sería una observación sobre los hábitos higiénicos en una escuela. El observador podría observar una clase y anotar cuántos alumnos y alumnas se lavan las manos después de ir al baño, si se les indica que se laven las manos antes de comer, etc.

El registro, para una metodología cuantitativa, debe ser inequívoco y estructurado, de manera que los datos generados sean uniformes y comparables de una observación a otra para su posterior análisis estadístico. Si la forma de observar los hábitos higiénicos difiere de una escuela a otra no podremos comparar las observaciones registradas.

Aunque suele relacionarse con conductas, también se puede aplicar a aspectos materiales. Tendría lugar por ejemplo en una evaluación de un proyecto de construcción de letrinas. El investigador, iría a los hogares muestreados y comprobaría si existe letrina, de qué materiales está hecha, etc.

La observación sistemática y la medición se utilizan en muchas ocasiones junto a la encuesta, combinando preguntas y observación con cada sujeto.

3.3 Técnicas participativas

Las técnicas participativas, como los grupos focales, se han utilizado tradicionalmente para obtener información cualitativa. Sin embargo, desde los años 90 y principalmente en el ámbito del desarrollo, se vienen desarrollando y empleando técnicas participativas que permiten también la obtención de información cuantitativa. Así, en muchos lugares se han sustituido las encuestas por diagnósticos participativos con herramientas visuales. Estos tienen el valor añadido de permitir que las personas 'investigadas' participen en mayor medida y que se recoja y analice colectivamente la información de forma simultánea. Esto puede ayudar también a corregir sesgos pues las personas se dan cuenta in situ de posibles incongruencias, puntos de vista no incluidos, etc.

Algunas de las técnicas más comunes que se utilizan a nivel de comunidad son el listado de hogares, la jerarquización de grupos de bienestar, la estimación de producción agrícola o los mapeos. Se aplican en una reunión o taller con miembros de la comunidad, en la que se realizan las dinámicas específicas de las distintas técnicas para recoger la información deseada.

Como no suelen estar presentes todos los miembros de la comunidad, en ocasiones se pueden estar excluyendo las voces de ciertos grupos de la comunidad. Es importante por tanto cuidar la composición del grupo de personas que participa en el taller, velando por que sea inclusivo y la información represente realmente a la comunidad. Aunque siempre habrá cierto sesgo, al igual que ocurre en las encuestas por hogares, donde se suele dar por válido para todo el hogar lo que dice la cabeza de la familia. La razón para darlo por válido (y no preguntar además a la pareja, hijas o ancianos), es que la cabeza de familia tiene conocimiento experto sobre el sujeto estudiado (su hogar). Con la misma lógica, en las técnicas participativas lo que dice el grupo que participa en el taller se puede dar por válido para toda la comunidad, pues dicho grupo tiene conocimiento experto sobre el sujeto estudiado (su comunidad). Para ambas

técnicas, lo importante es ser conscientes de esta limitación, cuestionarse lo grave que es –variará según la información que estemos recogiendo– y ser transparentes al respecto.

Normalmente se busca información extensiva de una región, por lo que se realizan talleres participativos en una serie de comunidades de la región, es decir, en una muestra de comunidades. A nivel de comunidad, se pueden dar dos casos, según si el sujeto a estudiar es la comunidad o la familia.

En primer lugar, cuando la comunidad es el sujeto a estudiar, se está buscando información del conjunto. Sería el caso de un mapeo para estudiar la distancia a recorrer para llegar desde la comunidad a distintos servicios (sanitario, educativo, etc.). En cada comunidad incluida en la muestra, se dibujaría el centro de la comunidad, la escuela, el centro sanitario, etc. y en pequeños grupos se estimaría la distancia a dichos lugares.

En segundo lugar, cuando el sujeto a estudiar es la familia o el hogar, la información buscada está referida a un nivel más micro. Sería el caso de un mapeo para conocer la cobertura de saneamiento, que depende de cuántos hogares tienen letrina. Se dibujaría un mapa en el que aparezcan todos los hogares de la comunidad, con distintos colores en función de si tienen o no letrina.

Figura 3: Mapeo participativo sobre saneamiento
Fuente: elaboración propia

Es importante tener presente que cuando sujeto a estudiar es la familia, los participantes en el taller o reunión no son la muestra del estudio en esa comunidad, sino que son los informantes que facilitarán la información de todas las familias de la comunidad. Siguiendo con el ejemplo anterior, no se pedirá a los participantes en el mapeo que dibujen su hogar y digan si tienen letrina o no, si no que se deben incluir en el mapa todos los hogares (estén o no presentes en el taller). Por ello, en estos casos la información que se recoge debe ser pública y conocida, pues los participantes deben aportar dicha información no solo sobre su hogar, sino también sobre los hogares de sus vecinos. Así, mientras que la posesión de letrina suele ser algo público, sería más difícil utilizar esta técnica para saber el gasto familiar en medicamentos, ya que es información más privada.

Tanto si el sujeto es la familia como si es la comunidad, las técnicas suelen realizarse en numerosas comunidades (la muestra) para obtener información extensiva en una región. Para que la información recogida participativamente en una comunidad pueda integrarse con la de otra y la podamos analizar estadísticamente, es necesario cumplir algunas condiciones: que las escalas no sean relativas (en el caso de clasificaciones de pobreza), que las dinámicas se faciliten de manera análoga en cada comunidad y que se fortalezca la fiabilidad del resultado (por ejemplo dividiendo en grupos naturales –mujeres, hombres, niños, niñas– para minimizar relaciones de poder, o en grupos mixtos para luego contrastar resultados).

Veamos un ejemplo ilustrativo real –aunque simplificado:

En 1999, debido a contradicciones sobre de población rural de Malawi (8.500.000 personas según el censo y 12.500.00 según una estimación) se decidió encargar un estudio para cuantificar la población rural y dirimir la discrepancia. Se diseñó una investigación en la que los sujetos eran las comunidades rurales y se tomó una muestra aleatoria de 54 de ellas con el fin de determinar su población. Debido a la falta de límites administrativos claros, se utilizó la técnica participativa del mapeo en cada comunidad. Una vez reunido un número considerable de miembros de la comunidad se formaron 3 ó 4 pequeños grupos y se les

pidió que hiciesen un mapa completo de la comunidad, situando todas las casas existentes y el número de personas que vivían en cada casa. El hecho de que los mapas se hiciesen en grupo permitió luego una puesta en común para contrastar y corregir errores, aumentando la fiabilidad del mapa.

Después, los investigadores calculaban el número total de habitantes de la comunidad. Una vez obtenidos los datos de las 54 comunidades, se calculó la desviación del censo respecto a los datos generados. Aplicando la inferencia estadística se pudo generalizar esta desviación y se concluyó que la población rural malawiana rondaba las 11.500.000 personas.

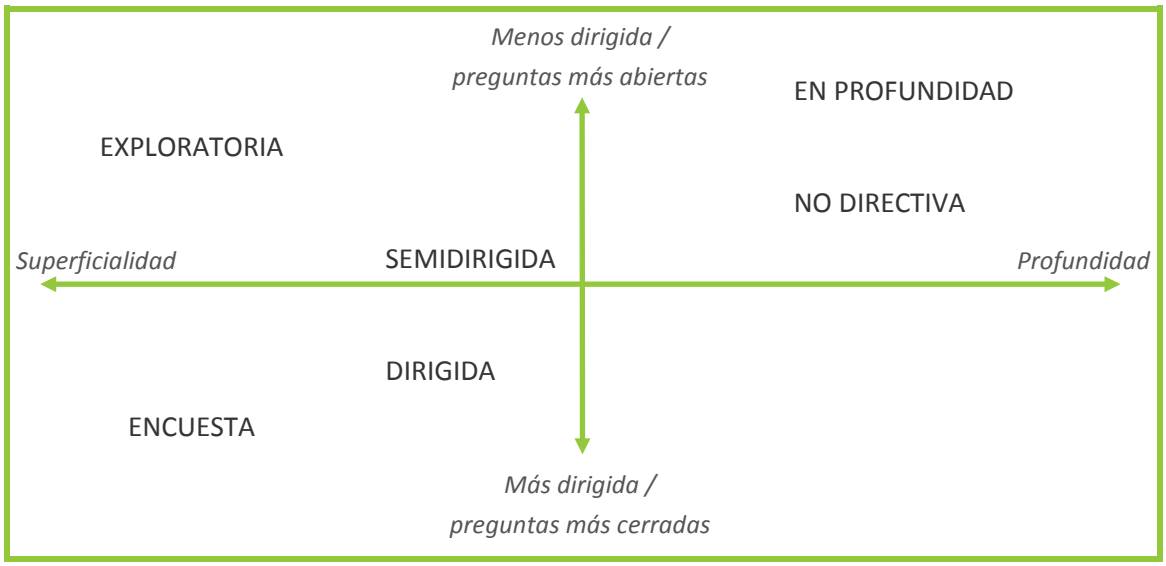
Para profundizar sobre el uso de técnicas participativas en cuantitativa, se puede consultar Barahona y Levi (2002) o Chambers (2007).

3.4 La encuesta

3.4.1 Conceptos básicos

Como se ha señalado, la técnica cuantitativa más habitual en la recolección de datos es la **encuesta**. Esta técnica, mediante la utilización de un cuestionario estructurado o conjunto de preguntas, permite obtener información sobre una población a partir de una muestra.

Las preguntas del cuestionario suelen ser cerradas en su mayoría, esto es, no se da opción a que quién responde se exprese con sus propias palabras (como en una entrevista) sino que se marcan unas opciones de respuesta limitadas entre las que elegir. Así, mediante codificación, se facilita una comparativa y análisis de datos más rápido que en las entrevistas, en detrimento eso sí, de la profundidad y matización en las respuestas. Se podría pues decir que la encuesta es una entrevista de tipo estandarizada y cerrada, cubriendo el límite opuesto a la entrevista en profundidad.



Fuente: Domínguez y Coco (2000)

Los datos que se pueden obtener con un cuestionario incluyen **datos objetivos** (hechos, cogniciones) y **subjetivos** (opiniones, actitudes):

- **Hechos** personales como la edad, nivel educativo; de contexto como tipo de vivienda, tipo de familia, y de comportamiento (reconocido o aparente) y **cogniciones**, es decir, índices de nivel de conocimiento de los temas estudiados en el cuestionario (ej. grado de conocimiento sobre la transmisión del SIDA).

- **Opiniones, actitudes, motivaciones y sentimientos**, es decir, todo lo que empuja a una determinada acción, o datos subjetivos (ej. satisfacción en la vida profesional).

Para la medición de actitudes, existen varias **escalas** para que las encuestadas/os indiquen su grado de conformidad. Entre las escalas más importantes encontramos: la escala Thurstone y la escala Guttman (afirmaciones de "de acuerdo / en desacuerdo"), la escala Likert (generalmente con cinco categorías: "muy de acuerdo", "de acuerdo", "indeciso", "en desacuerdo" y "muy en desacuerdo") y el **diferencial semántico de Osgood** (generalmente siete posiciones que median entre dos adjetivos polares, ej. progresista / conservador). Las dos últimas escalas, las de Likert y Osgood, son las más populares. Para más información, ver Cea d'Ancona (2001).

Es importante y útil **distinguir entre encuesta y cuestionario**. Si la encuesta es una técnica cuantitativa, el cuestionario es sólo *una parte* de la encuesta y hace referencia al *formulario o documento* que recoge las preguntas, que a su vez, representan unos indicadores implicados en el objetivo teórico de la encuesta.

Figura 4: Iceberg
Fuente: http://express.howstuffworks.com/gif/wq-iceberg-underwater.jpg [12-6-2012]

Cuestionario

Documento que recoge el conjunto de preguntas para una encuesta

Encuesta

Es mucho más que el cuestionario. Es la base sobre la que se sustenta el cuestionario. Abarca el diseño y aplicación del cuestionario (trabajo de campo) y el procesado de los datos obtenidos. Entendida como metodología con entidad propia, puede incluir también la operacionalización y el diseño muestral.

Como **cuestiones previas** (apartado 3.4.2) al diseño de la encuesta (o como primera fase de la metodología, si considerásemos la encuesta como una metodología), la investigación debe estar bien definida y operacionalizada. Ello supone la concreción de las preguntas o hipótesis en dimensiones e indicadores o variables concretas. Asimismo, el muestreo debe estar diseñado, y el momento y procedimiento de aplicación del cuestionario definidos.

A continuación, se procede al **diseño del cuestionario** (apartado 3.4.3) prestando atención a definir las preguntas correctamente, esto es, que sean exhaustivas, excluyentes, claras y que respondan en todo momento a las dimensiones teóricas establecidas. Normalmente se preparará una pregunta por cada variable especificada en la operacionalización. Después, se codifica el cuestionario con el fin de facilitar la medida y el análisis posterior de las respuestas (aunque algunos programas estadísticos codifican por sí mismos las respuestas).

Antes de proceder a la **aplicación del cuestionario** (apartado 3.4.4) al total de la muestra (trabajo de campo), es importante consultar a expertas/os y hacer una prueba piloto con algunos sujetos, para probar y validar el cuestionario. Después de estas pruebas y de las correcciones oportunas, ya se puede iniciar el trabajo de campo.

La última fase consiste en el **procesado de la información recogida** (apartado 3.4.5).

3.4.2 Cuestiones previas

Antes de proceder con el cuestionario, debemos tener claro qué queremos averiguar con él, es decir, qué dimensiones teóricas nos interesan. Es indispensable que cada pregunta se haya elaborado con una razón, que sirva para desvelar parte de esas dimensiones. Esto se debería de haber completado en la fase de operacionalización (paso de diseño), a partir de las preguntas de investigación y de la preparación del marco teórico (revisión bibliográfica, consulta de expertas/os y recogida de información ya existente sobre el tema). En caso contrario, o en el supuesto de que al plantear el cuestionario los indicadores o variables no nos parezcan satisfactorios, será preciso revisar y mejorar la operacionalización.

Si se comienza la evaluación por la construcción del cuestionario sin haber precisado claramente los objetivos de la evaluación, podemos incluir muchos elementos que supongan un esfuerzo baldío e incluso resulten perjudiciales, porque pueden restar claridad a las variables investigadas (García Muñoz, 2003).

Sugerencia

Hazte un guion con las dimensiones y variables obtenidas de la operacionalización. Pon por escrito la información que tu encuesta pretende sacar. Luego, basándote en el guion, redacta las preguntas. Liga cada variable a por lo menos una pregunta. Pregúntate cada vez que pienses en una posible pregunta: ¿por qué estoy preguntando esto? ¿Tiene la persona entrevistada la información solicitada?

Un aspecto importante es la forma de administración del cuestionario, pues determina en gran medida la elaboración del cuestionario. Según la presencialidad y el lenguaje, la encuesta puede ser personal, telefónica, escrita o por correo

	Presencial	A distancia
Oral	**Personal**, donde encuestador y encuestado interactúan frente a frente. Las preguntas deben redactarse en forma de conversación. El encuestador no debe sesgar o influir en las respuestas. Proporciona mayor abundancia en los datos, pues permite anotar observaciones y repreguntar. Es intensivo en recursos humanos.	**Telefónica**, donde la interacción es a distancia. El diseño es similar a la presencial, pero requiere de preguntas más breves y sencillas. Reducen el coste y el tiempo, pero pueden presentar mayor tasa de no respuesta, así como generar sesgos en contextos donde el teléfono no es universal.
Escrita	**Escrita**, donde el encuestado completa el cuestionario por sí solo. Requiere de una buena introducción e instrucciones y que las preguntas sean cuidadosamente formuladas para que no haya lugar a interpretaciones distintas. Pueden aplicarse en grupo.	**Correo (postal o electrónico)**: Similares a las escritas presenciales, pero con menor tasa de respuesta (50% aproximadamente). Por ello, las preguntas deben ser más sencillas y llamativas, y se deben incluir instrucciones motivantes. Las electrónicas permiten incluir patrones de salto complejos y aleatorización de las preguntas para eliminar las tendencias por el orden.

En cuanto al **número de preguntas** de un cuestionario, deberá tener todas las necesarias, pero "*ni una más. […Es] recomendable hacer solamente las preguntas necesarias para obtener la información desea-*

da" (García Muñoz, 2003). Un cuestionario largo produce fatiga y rechazo en el sujeto que lo completa, con el riesgo añadido que se quede incompleto o se conteste sin la debida reflexión. Así, se debe de evitar salvo que sea absolutamente necesario.

Sugerencia

Triangula con otras técnicas. El uso de un cuestionario es únicamente para variables que no se pueden obtener de otra manera. Comprueba si hay preguntas en tu cuestionario que pueden ser cubiertas con otras técnicas (ej. observación) y si hay dimensiones que se podrían investigar mediante otros métodos (ej. cualitativos) para así acortar al máximo tu cuestionario.

En cuanto al **tiempo empleado en contestar al cuestionario**, la literatura científica suele recomendar la regla de "que pueda ser contestado entre media y una hora" *(García Muñoz, 2003).* Con todo, si queremos una elevada tasa de respuesta, es mejor que no sobrepase los 10 minutos.

Aun así, tanto el número de preguntas como el tiempo empleado dependen del **grado de información y de interés de la encuestada/o**:

"Cuando se trata de hechos que le son familiares, que está deseando dar a conocer y que cree que significan para él/ella la oportunidad para hacerse oír, el sujeto responde sin fatiga en un tiempo muy superior al de la hora, pero cuando se trata de cuestiones que obligan a reflexionar, acerca de las cuales no hay una actitud definida o que no tenemos ningún motivo para expresarlas y más aún preferíamos no formularlas, las reservas e incertidumbres van haciendo dilatar las respuestas y al final el cuestionario se convierte en una tarea ingrata que procuramos terminar pronto y de cualquier modo, con lo que su validez es dudosa". (Marín Ibáñez en García Muñoz, 2003)

Otro aspecto muchas veces olvidado es la presencia de distintas lenguas. Es importante que el cuestionario esté redactado en las distintas lenguas que se utilizan en la región estudiada, y para las encuestas orales, que los encuestadores las dominen adecuadamente.

En resumen, para la elaboración del cuestionario cabe tener en cuenta: la finalidad detrás de cada pregunta (base teórica), las características de la población estudiada, su conocimiento e interés sobre el tema, el tamaño de la muestra, el presupuesto, los recursos y la forma de administración del cuestionario (correo, personal...). Todo esto definirá el número de preguntas, la duración del cuestionario y su formato.

3.4.3 Diseño del cuestionario

La elaboración formal del cuestionario abarca dos aspectos básicos: la **redacción de las preguntas** y la determinación de los **aspectos formales** del cuestionario.

Hay tres tipos de preguntas en cuanto a su redacción: cerradas, abiertas y semi-abiertas.

- Las preguntas cerradas incluyen una selección de respuestas, que pueden ser dicotómicas, es decir, de dos respuestas (ej. "sí / no"); o múltiples, o sea, un abanico de más de 2 posibilidades. Las múltiples pueden llevar un orden de menor a mayor o incluso intervalos de una característica continua, como en el ejemplo del cuadro dado abajo.
- Las preguntas abiertas no incluyen respuesta.
- Las semi-abiertas incluyen respuestas, pero dejan un espacio para otras opciones.

Tipos de preguntas	
Pregunta cerrada (dicotómica)	¿Tiene en su domicilio acceso a Internet? ( ) Sí ( ) No El SIDA se transmite por la saliva: ( ) Verdadero ( ) Falso
Pregunta cerrada (múltiple)	¿Cuánto dinero cobras al mes? ( ) Menos de 1000 ( ) De 1001 a 1500 ( ) De 1501 a 3000 ( ) Más de 3000
Preguntas semi-abiertas	¿Tiene pensado cambiar de vivienda en el futuro? ( ) Sí → ¿Por qué? ____ ( ) No
	¿Cuál es la principal exigencia en su trabajo? ( ) Conocimientos ( ) Obediencia ( ) Resistencia física ( ) Otra. Especifique: ____
Pregunta abierta cualitativa	¿Cuál es la principal exigencia en su trabajo? ____
Pregunta abierta cuantitativa	¿Cuántas horas trabaja a la semana? ____

Las preguntas abiertas cualitativas son más fáciles de formular que las cerradas, puesto que no hay que prever ningún tipo de respuesta ni investigar acerca de la exhaustividad y exclusión de categorías (ver abajo). Sin embargo, la dificultad aparece a la hora de resumir y codificar la información. También requieren más tiempo de respuesta. Normalmente, será necesaria la **inclusión de los dos tipos** de pregunta:

> "Las preguntas cerradas son más eficaces donde las posibles respuestas alternativas son conocidas, limitadas en número y claramente definidas (...). Las preguntas abiertas son adecuadas cuando el tema es complejo, cuando las dimensiones relevantes no son conocidas o cuando el interés de la investigación reside en la exploración" (García Muñoz, 2003)

Por tanto, es recomendable cerrar las preguntas lo máximo posible. Para **cerrar preguntas abiertas**, se puede aplicar el cuestionario a algunas personas (¡que no formen parte de la muestra!) a modo de prueba piloto. Se hacen las preguntas abiertas y luego se aprovechan las respuestas dadas con más frecuencia

para cerrarlas en el diseño final del cuestionario. También se puede recurrir a la ayuda de personas expertas en la materia, que puedan intuir a priori respuestas que se podrían dar.

En lo que respecta a las preguntas cerradas, Fox (en García Muñoz, 2003) advierte que son muy pocas las preguntas de opiniones o actitudes que tengan una estructura tan simple y estandarizada de "sí / no", "conforme / disconforme", "satisfecho / insatisfecho", siendo **más prudente el ofrecer un abanico de opciones**.

Incluso es recomendable adaptar y **concretar** al máximo ese **abanico de respuestas**. Es decir, en vez de operacionalizar las respuestas con una escala Likert estándar ("muy bien / bien / mal / muy mal" o "1 / 2 / 3 / 4" o "poco / a veces / mucho"), es mejor definir un abanico de respuestas personalizado a la pregunta en cuestión, explicitando con las respuestas aquello que estamos preguntando efectivamente con la pregunta. Así, se reduce más el grado de interpretación de quien responde:

Concretar las escalas Likert	
PEOR	Puntúe si está usted satisfecho con el grado de participación del alumnado en la evaluación: (1) Nada satisfecho (2) Poco satisfecho (3) Satisfecho (4) Muy satisfecho
MEJOR	Puntúe el grado de participación del alumnado en la evaluación: (1) No se les pidió la opinión (2) Dieron su opinión (3) Participaron en los comités de seguimiento (4) Formaron parte de la junta evaluadora

Concretar las escalas Likert	
PEOR	Puntúe de peor (1) a mejor (4) la variedad de los métodos usados en la calificación de los estudiantes por el profesorado: 1 2 3 4
MEJOR	Puntúe la variedad de los métodos usados en la calificación de los estudiantes por el profesorado: (1) No hubo evaluación (2) Hubo una única metodología de calificación (3) Hubo dos metodologías de calificación (4) Hubo tres o más metodologías de calificación

La **concreción** también es recomendable **con palabras** como “mayor”, “joven”, “progresista”, “mucho”, “barato”, “normalmente”, “bueno”, “malo”... Mientras que algunas personas pueden considerar que 35 años es ser “joven”, otras pueden afirmar que con esa edad ya se es “viejo”. De la misma manera, es mejor usar respuestas concretas como “diariamente”, “2-3 veces a la semana”, “dos veces al mes”, que respuestas más vagas como “con frecuencia” o “regularmente”.

Concretar los significados de las palabras	
PEOR	¿Fuma mucho tabaco? ( ) No, no fumo ( ) No, fumo poco ( ) Sí, fumo mucho
MEJOR	¿Cuánto tabaco fuma? ( ) No fumo ( ) Fumo entre 1 y 5 cigarros al día ( ) Fumo más de 5 cigarros al día

Por último, en encuestas presenciales se recomienda que cuando una pregunta en abanico incluya muchas alternativas de respuesta, se recurra al **uso de tarjetas**. Esto es, en vez de leer las distintas opciones de respuesta (y forzar a la persona encuestada a su memorización), se le entregan tarjetas para que visualice las distintas opciones mientras que se leen en voz alta. La lectura en voz alta es recomendable en todo caso para prevenir posibles problemas de visión que pueda tener la persona encuestada.

La definición de cada pregunta debe ser exhaustiva, esto es, abarcar todos los casos de respuesta que pueden darse. En ese caso, ninguna persona encuestada puede dejar de responder por no encontrar su categoría:

La definición de cada pregunta debe ser exhaustiva	
INCORRECTO	¿Cuántos embarazos ha tenido?
CORRECTO	Señale el número de veces que se ha quedado embarazada, incluyendo tanto casos en que haya tenido una hija/o como aquellos que hayan finalizado en aborto.

Una manera de asegurarse la exhaustividad, son las opciones de **“otros”** con espacio para la respuesta abierta. También existen las opciones de **“no sabe”** (desconoce), **“no contesta”** (prefiere no contestar) o **“no pertinente / aplicable”** (es una pregunta que no le corresponde – ej. embarazo a un hombre). Si no existen éstas, quien responde puede seleccionar cualquier respuesta simplemente para no dejarla en blanco.

La definición de cada pregunta debe ser excluyente, es decir, que ningún sujeto al contestar al cuestionario pueda elegir válidamente dos respuestas distintas de la misma pregunta:

La definición de cada pregunta debe ser mutuamente excluyente	
INCORRECTO	¿Cada cuánto va al centro de alfabetización? ( ) Nunca ( ) Ocasionalmente ( ) Alguna vez ( ) Todos los días de la semana
CORRECTO	¿Cada cuánto va al centro de alfabetización? ( ) Nunca ( ) Menos de una vez a la semana ( ) Entre 1 y 6 días a la semana ( ) Todos los días de la semana

Evitar hacer dos preguntas en una. Esto es muy común y confunde mucho al lector/a. Por ejemplo: ¿Puedes estudiar cuando hay un radio o una televisión prendida en tu casa? Puede que con radio sí pueda estudiar, pero no con televisión. Otro ejemplo: la información ¿es interesante e importante? Si interesante e importante son sinónimos, entonces con un solo adjetivo es suficiente. Si no, habrá que formular dos preguntas. Otro ejemplo:

Evitar hacer dos preguntas en una sola	
INCORRECTO	La atención del profesorado y de la secretaría del máster es: ( ) Suficiente ( ) Regular ( ) Insuficiente
CORRECTO	La atención del profesorado del máster es: ( ) Suficiente ( ) Regular ( ) Insuficiente La atención de la secretaría del máster es: ( ) Suficiente ( ) Regular ( ) Insuficiente

Las preguntas deben ser comprensibles para las personas encuestadas. Es necesario adaptar el lenguaje al registro de quien responde. El conocimiento y aplicación de términos locales puede ayudar en la enumeración de las preguntas así como el redactarlas de forma directa y personalizada (en 2ª persona).

Preguntas sin tecnicismos	
INCORRECTO	¿Tienes dismenorrea? ( ) Sí ( ) No
CORRECTO	¿Tienes dolores durante la menstruación? ( ) Sí ( ) No

Evitar preguntas en forma negativa que puedan confundir	
INCORRECTO	¿Piensa que no debe permitirse la publicidad de leche artificial? ( ) Sí ( ) No
CORRECTO	Marque según crea: "La publicidad a favor de la leche artificial... ( ) ...debe permitirse" ( ) ... no se debe permitir" ( ) ... no tengo opinión"

Evitar preguntas que obliguen a hacer cálculos o esfuerzos de memoria	
INCORRECTO	A lo largo del año pasado ¿cuántas veces ha ido al centro de salud?
CORRECTO	- Limitar el periodo de tiempo especificado - Buscar otras técnicas de recogida de información → ej. fichas en el centro de salud

La **prueba piloto** es esencial para adaptar las preguntas y vocabulario del cuestionario y para analizar si las personas que responden están entendiendo con las preguntas aquello que quienes las formularon pretendían que entendieran. En la aplicación piloto, es conveniente recoger todas las reacciones que manifiesten las encuestadas/os, tales como facilidad, entusiasmo, aburrimiento, incertidumbre, duda, incomprensión o fatiga. Es recomendable usar la **técnica de 'pensar en alto'** (se le pide a quien responde que nos diga en voz alta lo que está pensando cuando lee cada pregunta).

Resulta interesante el ejemplo práctico de un cuestionario que se utilizó para evaluar el impacto de varios proyectos de micro-centrales hidroeléctricas en Bolivia. El objetivo de las micro-centrales es el de proveer electricidad a comunidades rurales aisladas de la red eléctrica general y que requieren sistemas de auto-abastecimiento. El cuestionario fue inicialmente elaborado junto al PNUD Bolivia y posteriormente se realizó una prueba piloto en una comunidad no incluida en la muestra. Se puede acceder on-line a las versiones del cuestionario antes y después de una prueba piloto.

Colocación de las opciones. Es recomendable colocar las preguntas verticalmente pues en ocasiones es confuso si hay que marcar antes o después de la opción. Por otra parte, este espacio da aire al cuestionario escrito.

Evitar colocaciones de respuestas que puedan llevar a la duda	
PEOR	_____ excelente _____regular _____bueno _____malo _____pésimo
MEJOR	___excelente ___regular ___bueno ___malo ___pésimo

Intentar fusionar las preguntas filtro (aquellas que descartan a quienes no les afectan determinadas preguntas, es decir, marcan la realización o no de preguntas posteriores) para mayor celeridad en la respuesta y menor fatiga de quien responde.

Fusionar las preguntas filtro	
PEOR	¿Estás casado? ( ) Sí ( ) No En caso afirmativo ¿trabaja tu cónyuge? ( ) Sí ( ) No
MEJOR	¿Trabaja tu cónyuge? ( ) No estoy casada/o ( ) Sí ( ) No

Uso de preguntas de control. Las preguntas de control son las que pretenden comprobar la consistencia de las respuestas de la encuestada/o. Consisten en la formulación de preguntas similares, formuladas de modo distinto y en momentos distintos para estudiar la coherencia entre ambas respuestas. Se recomienda no abusar de las preguntas de control por razones de espacio y usarlas sólo con las dimensiones teóricas más importantes o más subjetivas (de opinión, no las de hechos o cognición).

Usar preguntas de control
¿Piensas cambiar de ocupación en el futuro? ( ) Sí ( ) No ( ) No sé
[Poco después en el cuestionario]: ¿Piensas seguir dedicándote a tu profesión actual a largo plazo? _____

Las **preguntas muelle, "colchón" o "amortiguadores",** son preguntas que abordan temas difíciles, formuladas de forma que reduzcan su rudeza. Veamos un ejemplo en el que se le proponen al sujeto encuestado varias actividades habituales, los días laborables, entre ellas la que nos interesa, con objeto de no dejar al descubierto su falta de interés / falta de tiempo para la formación:

Usar preguntas muelle	
PEOR	¿Repasas los temas de la formación diariamente? ( ) Sí ( ) No
MEJOR	De las siguientes actividades, ¿nos podrías indicar cuáles realizas habitualmente en la tarde-noche? (puede ser más de una) ( ) Estar con mis amistades ( ) Estar con mi familia ( ) Hacer deporte ( ) Dar un repaso a los temas de la formación ( ) Preparar la cena ( ) Otros: _____

Una buena forma de validar las preguntas es hacer pruebas piloto. Además de la descrita anteriormente, puede resultar muy útil hacer una prueba piloto de análisis, simulando la fase posterior a la recogida (análisis de datos). Se puede así visualizar de antemano las tablas, gráficos o cálculos que se obtendrán de los datos. Es una buena manera de comprobar qué es importante, qué preguntas son superfluas o qué nos hemos dejado en el tintero.

Antes de pasar a los aspectos formales del cuestionario, resumamos rápidamente las orientaciones para la redacción de preguntas:

- Utiliza principalmente preguntas cerradas
- Concreta al máximo el abanico de respuestas
- La definición de cada pregunta debe ser exhaustiva
- La definición de cada pregunta debe ser excluyente

- Evita hacer dos preguntas en una
- Las preguntas deben ser comprensibles para las personas encuestadas
- Vigila la colocación de las opciones
- Intenta fusionar las preguntas filtro
- Modera el uso de preguntas de control
- Usa preguntas muelle para abordar temas difíciles
- Realiza una prueba piloto de aplicación del cuestionario y de análisis

Los **aspectos formales** son básicos en la elaboración de cuestionarios. En efecto, la calidad de las respuestas puede verse afectada no sólo por la redacción de las preguntas, sino también por su orden y ubicación en el cuestionario (entre qué preguntas se halla y si está al principio, en medio o final del cuestionario). Algunas ideas para los aspectos formales del cuestionario son las siguientes:

Identificar el cuestionario en la 1ª página con:

- número / código del cuestionario

- fecha y lugar de la encuesta

- nombre del encuestador/a

Presentarse a una/o mismo y a la institución que representa.

Presentar brevemente la finalidad y beneficios de la encuesta.

Garantizar el anonimato (por regla general, no se piden nombres en el cuestionario)

Dar unas breves instrucciones antes de comenzar el cuestionario.

Citar un tiempo estimado de compleción.

Cuando el cuestionario se aplique **por correo**, incluir una **carta de presentación** para solicitar la cooperación de la encuestada/o, presentarse, explicar la finalidad del estudio, las instrucciones y agradecer la colaboración. Se recomienda incluir fecha, teléfono de contacto y no gastar más de una página.

Introducir los **datos socio-demográficos** de la persona encuestada que sean de relevancia para el estudio. Ejemplos: edad, sexo, nivel educativo, estado civil, lugar de nacimiento, lugar de residencia, profesión, ingresos, lengua, religión, filiación política, número de hijas/os, nacionalidad, etnia... Estas preguntas de identificación son fundamentales pues suelen constituir las variables independientes principales del análisis estadístico posterior.

Estos datos se pueden poner **al final de la encuesta**, cuando ya haya más confianza por parte de la encuestada/o para compartir esa información.

Si la encuesta es sobre la **familia** y no sobre la persona en concreto, sería recomendable saber la posición del encuestado/a en la familia. Lo ideal sería que se respondiese conjuntamente.

Numerar las preguntas y respuestas (ver codificación más abajo).

Orden de las preguntas. Las preguntas más generales y fáciles suelen colocarse primero, dejando las difíciles y embarazosas detrás. Las preguntas de hechos se suelen colocar antes que las de opinión, pues suelen contestarse más fácilmente.

La estructura, diseño y disposición de las preguntas debe ser ágil y agradable. Es importante agrupar las preguntas en secciones lógicas.

Claridad en la redacción, evitando términos técnicos especializados, abreviaciones y frases largas o difíciles (dobles negaciones, alternativas no mutuamente excluyentes, vaguedad en las afirmaciones, etc.)

Es muy importante hacer buenas **transiciones** entre temas y bloques con frases como "ahora os haremos una serie de cuestiones..." o "cambiando de tema..."

En los cuestionarios presenciales, es fundamental dejar un espacio al final para:

- duración de la entrevista

- impresiones del encuestador/a (ej. sensación de sinceridad, de que miente, etc.)

- incidencias (ej. sustitución de la encuestada/o, varias personas respondiendo a una misma encuesta, presencia de personas curiosas / maridos durante la entrevista...)

- número de intentos de localización de la entrevistada/o

- un "gracias" por escrito para recordar al encuestador/a que debe dar las gracias.

3.4.4 Aplicación del cuestionario

¿Cómo accedemos a la gente que tenemos que entrevistar? ¿Cómo conseguimos un grado suficiente de compleción de respuestas? Existe una serie de limitaciones de acceso y compleción que veremos a continuación. En los estudios de desarrollo, éstas se ven a menudo acentuadas.

Limitaciones de acceso y administración de cuestionarios. Particularidades en los estudios de desarrollo.

ACCESO y COMPLECIÓN

La primera limitación puede ser la **falta de un marco muestral** (censo o padrón) del que extraer la muestra. Existen estrategias para solventar o minimizar este problema, como se ha visto en el capítulo 2. Una opción para poblaciones pequeñas son los mapeos participativos para construir un marco muestral.

La **variedad lingüística** y la posibilidad de **necesitar traductores** para las encuestas son mayores en contextos de desarrollo. Cabe contar con esto a la hora de definir el tiempo de respuesta.

La **forma del cuestionario** (presencial auto-cumplimentado, presencial con encuestador/a, por teléfono, por Internet o por correo) suele decidirse según tema y tipo de cuestionario, sopesando ventajas e inconvenientes en cada tipo de encuesta. No obstante, en contextos de desarrollo, los cuestionarios por correo, teléfono o Internet son muchas veces impracticables, especialmente en zonas rurales.

También suele ocurrir que los cuestionarios **auto-completados** son difíciles de aplicar con personas analfabetas o de poco manejo escrito (a no ser que los cuestionarios estén bien adaptados, por ejemplo mediante el uso de visuales y tarjetas). También hay diferencias de tradición escrita versus tradición oral. En muchas zonas, hay menos costumbre de usar lápices, papel, etc. En la misma línea, hay diferencias entre la abstracción conceptual versus la metáfora / cuento a la hora de narrar. Cabe por tanto cuidar el registro y la manera de presentar la información.

Para encuestas por Internet, ver encuestafacil.com o los formularios de GoogleDocs. Para encuestas por correo o Internet, usar avisos o recordatorios. Para encuestas por teléfono o presenciales, insistir al menos una segunda vez en un horario diferente.

Privacidad difícil: ¿qué hacer si el cónyuge, familia o vecinas/os están presentes a la hora de completar el cuestionario? ¿O si responde el marido cuando se le está haciendo el cuestionario a la esposa? La solución no es fácil y dependerá de la creatividad de cada encuestador/a (por ejemplo hacer la encuesta en zonas donde sólo vayan mujeres, dejar las preguntas difíciles para un "paseo" posterior por la casa o barrio con la encuestada/o, etc.). La formación y preparación de los encuestadores es vital.

Desde el **paradigma participativo**, se ha criticado que los cuestionarios (y otros métodos cuali-cuanti) son tecnocráticos y extractivos (se analizan en oficina, no en terreno y los resultados no suelen compartirse). Si se llega a este punto, las preguntas a hacerse son: ¿Quién domina el proceso? Y ¿quién aprende y acaba conociendo los resultados? Para evitar esta extracción, es esencial la explicación de las técnicas mismas (qué es un cuestionario y para qué sirve) y la devolución de resultados.

Se puede **ofrecer una devolución** oral (reunión / taller), visual (fotos) o por escrito (copia del resumen del trabajo). Si no es posible enviarlo a todas/os los encuestados/as, se puede intentar enviar a entidades comunales de la zona (centros religiosos, asociaciones, gobierno local, escuelas, centro clínico...).

Hay que respetar el **tiempo de las personas** encuestadas con cuestionarios claros y cortos. De igual manera, conviene adaptarse al tiempo y lugar en que a las personas les vaya mejor contestar.

Se debe preparar para las encuestadas/os una breve explicación sobre la importancia de su participación y lo que se hará con los resultados.

Finalmente, hay que asegurar el anonimato de su participación.

Prevención de sesgos en la cumplimentación. A través del diseño de las preguntas se pueden controlar los posibles sesgos de cumplimentación:

A. Un sesgo habitual es el **"error de tendencia central"**, o la tendencia a elegir la respuesta de en medio. Solución: elegir un número par de opciones de respuesta, cuatro o seis, con objeto de evitar que la encuestada/o pueda responder a la opción central, sin esforzarse en reflexionar.

Evitar la categoría de respuesta intermedia	
MENOS RECOMENDABLE	En las clases teóricas del curso de formación que está haciendo, ¿toma apuntes o notas de lo que dice el formador? ( ) Nunca ( ) A veces ( ) Siempre
MÁS RECOMENDABLE	En las clases teóricas del curso de formación que está haciendo, ¿toma apuntes o notas de lo que dice el formador? ( ) Nunca ( ) Pocas veces ( ) Con frecuencia ( ) Siempre

B. Un segundo sesgo es el de **"proximidad o aprendizaje",** que induce a contestar de forma similar a las respuestas anteriores. Solución: evitar en la medida de lo posible repetir el formato en preguntas consecutivas. Por ejemplo, diseñar una pregunta con una escala Likert de positivo a negativo y la siguiente, de negativo a positivo. Este sesgo es especialmente relevante en las **preguntas batería** (conjunto de preguntas sobre la misma cuestión, que se completan unas a otras. Se suelen agrupar en un *"embudo de preguntas"*, empezando por los aspectos más generales y sencillos hasta los más concretos y complejos).

C. Otro sesgo frecuente es el de **"deseabilidad social",** o responder según lo que se considera socialmente aceptable (no lo que se siente o piensa, sino lo que haga "quedar bien"). Solución: cuidar quién realiza el cuestionario. Si la temática es sobre racismo, machismo, clasismo, etc., es recomendable que el encuestador/a tenga un **parecido socio-demográfico** con la encuestada/o.

Igualmente, las preguntas consideradas personales (ej. creencias religiosas, militancia política, ideas sobre sexo, etc.) o que se crea que puedan ser motivo de premio o sanción, deben formularse de **forma indirecta** o en **3ª persona** ("conoce a mucha gente que piense que..."; "cree que sus amigos...").

Usar preguntas indirectas	
PEOR	¿Alguna vez has robado en un gran almacén? ( ) Sí ( ) No
MEJOR	¿Conoces a alguien que haya robado en un gran almacén? ( ) Sí ( ) No

Se desaconseja usar el **tiempo condicional** ("si estuviera en esta situación...") porque se puede caer en lo normativo. Es mejor recurrir a formular las preguntas sobre lo que hacen o hicieron en una situación parecida, más que lo que harían. Se considera que la conducta pasada (qué hicieron en una situación parecida) es un buen indicador de la conducta futura, a menos que se hayan producido cambios notables en la faceta que pretendemos analizar.

También puede optarse por **preguntas muelle o colchón** o por solicitar **respuestas aproximadas**. Por ejemplo, ante la posible reticencia a indicar la cantidad exacta de ingresos, se podría formular: *"¿podría indicar, aproximadamente, cuál es la cuantía de sus ingresos mensuales?".*

D. Un último sesgo es el de la "**deformación conservadora**", donde las personas tienen más tendencia a contestar "sí" que a contestar "no". Una pregunta recibe mayor porcentaje de adhesiones cuando está formulada para contestar "sí" que cuando está formulada para contestar "no". Solución: usar preguntas equilibradas o neutrales en vez de referirse en la pregunta a sólo una de las alternativas:

Usar preguntas equilibradas	
PEOR	¿Está a favor de que la formación se haga fuera del horario de trabajo? ( ) Sí ( ) No

MEJOR	¿Está a favor o en contra de que la formación se haga en horas de trabajo? ( ) A favor ( ) En contra
	La formación, ¿debería hacerse... ( ) ...durante las horas de trabajo? ( ) ...fuera de las horas de trabajo?

E. También hay que **evitar hacer referencia a personalidades públicas**. Las preguntas no pueden apoyarse en instituciones ("la iglesia opina que..."), ideas respaldadas socialmente ("la mayoría de personas opina que...") o en evidencia comprobada científicamente, puesto que es también una forma de inducir la respuesta.

3.4.5 Procesado de la información recogida

Codificación. Ya en la fase de diseño del cuestionario, se inicia el proceso de codificación del cuestionario para posteriormente introducir los datos en las bases de datos informáticas para la fase de análisis. Codificar es dar un **número y nombre a cada pregunta** y un **número-valor a cada una de las alternativas de respuesta**:

Codificar
P22. ¿Ha realizado otro programa de formación en la empresa donde trabaja actualmente? (1) Sí (2) No (9) Ns/Nc

Codificar
P23 (PRTR): Cuando se enfrenta a un problema en su trabajo, para resolverlo recurre a: 1 Su superior inmediato 2 Su propia experiencia 3 Sus compañeros 4 Los manuales de políticas y procedimientos 5 Otra fuente (especificar)_____

Esto significa que en la pregunta 22, la variable puede adquirir los valores entre 1 y 2, y en la pregunta 23, puede tomar valores entre 1 y 5. Los "no sabe / no contesta / no pertinente o aplicable" suelen codificarse con el "0", "8" ó "9" (o con "00", "88" ó "99" si hay más de 8 ó 9 valores en las respuestas).

La codificación permite transformar las diferentes dimensiones teóricas en descriptores numéricos que son más fácilmente volcados a una aplicación informática, aunque muchos de los programas estadísticos hoy en día facilitan la introducción de datos no codificados.

¿Y qué ocurre con la codificación en el caso de las preguntas abiertas o semi-abiertas?

Para el **análisis y cierre de preguntas abiertas y semi-abiertas**, se anotará en una hoja la respuesta a la primera pregunta abierta del primer cuestionario. Si la respuesta a la primera pregunta del segundo cuestionario es similar, se anotará en la misma hoja. Si es diferente se anotará en otra hoja y así sucesivamente hasta terminar con la primera pregunta de todos los cuestionarios. Una vez terminado el análisis de la primera pregunta de todos los cuestionarios, se hará un resumen de las respuestas en cada hoja (buscando términos comunes, agrupándolas en categorías y codificando nuevas categorías) así como del número de respuestas en cada hoja. Posteriormente, se hará lo mismo con cada una de las preguntas abiertas hechas en el cuestionario.

Análisis de los altos grados de no-respuesta. Es importante cuidar el análisis de las “no respuestas”, sobre todo cuando son altas. Se puede intentar tipificar por categorías las razones por las que no hubo respuesta: no familiaridad con la cuestión, “evasiva intencional”, etc. Ello permite concretar recomendaciones y planes de acción para futuras encuestas. También se pueden usar métodos cualitativos complementarios (ej. entrevistas) para analizar qué hay detrás de esas no-respuestas.

Capítulo 4. Estadística descriptiva

Objetivos del capítulo:	Tiempo estimado de lectura: 300 min
• Analizar (calcular estadísticos, realizar estimaciones y presentar gráficamente) los datos disponibles, utilizando herramientas informáticas de manera consciente	Apartados del capítulo: 4.1 Introducción a la estadística aplicada 4.2 Conceptos básicos 4.3 Análisis unidimensional 4.4 Análisis bidimensional
Capítulo anterior. Recolección de información: la encuesta	Índice — Capítulo siguiente. Inferencia estadística

4.1 Introducción a la estadística aplicada

4.1.1 Estadística descriptiva e inferencia estadística

La estadística es una ciencia con base matemática referente a la recolección y análisis de datos. Es transversal a una amplia variedad de disciplinas, desde la física hasta las ciencias sociales y desde las ciencias de la salud hasta el control de calidad.

Se puede distinguir entre *estadística matemática*, que se refiere a las bases teóricas de la materia, y la *estadística aplicada*.

La estadística aplicada tiene entre sus funciones principales describir, explicar y predecir. Los estudios estadísticos se emplean para describir una realidad a través de la síntesis, comparación y presentación de datos económicos, políticos, sociales, etc., apoyando así el aprendizaje y los procesos de toma de decisiones.

Estadística descriptiva: rama de la estadística aplicada que se utiliza para analizar y resumir datos (de una muestra)

La estadística aplicada se divide a su vez en dos ramas:

La **estadística descriptiva**, que se dedica a los métodos de organización, descripción, visualización y resumen de datos originados a partir de la recogida de información. Los datos pueden ser resumidos numéricamente mediante estadísticos (por ejemplo la media) o gráficamente (por ejemplo mediante una pirámide poblacional).

La **estadística inferencial**, que se dedica a sacar conclusiones sobre la población a partir de los datos de una muestra.

Dentro del proceso de investigación cuantitativa, una vez recolectados los datos, llega el paso de análisis, que incluye el análisis descriptivo –usar la estadística descriptiva para resumir los datos de una muestra– y el análisis inferencial –calcular con qué precisión ese resumen es representativo de toda la población.

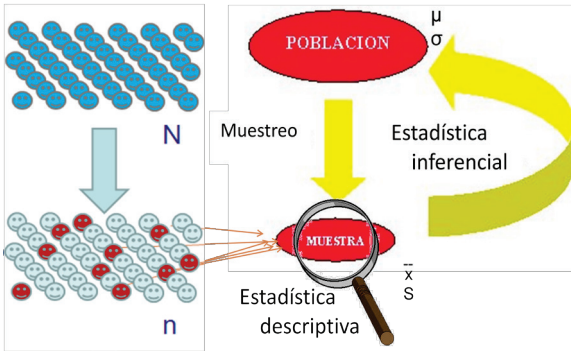


Figura 5: Estadística descriptiva e inferencial
Fuente: elaboración propia

Este capítulo se dedica a la estadística descriptiva y el siguiente, a la inferencia estadística.

Queda fuera del alcance del capítulo aprender a realizar todos los tipos de cálculos estadísticos. Más bien, se busca obtener una idea holística de la estadística aplicada al desarrollo, conocer los distintos estadísticos que existen y entender su utilidad y cuándo se deben aplicar.

4.1.2 ¿Es la estadística objetiva?

«Hay tres tipos de mentiras: mentiras pequeñas, mentiras grandes y estadísticas»

Hay una percepción general de que la estadística se usa frecuentemente de manera intencionada, encontrando maneras de interpretar los datos que sean favorables a ciertos intereses. A veces se sesga o manipula la muestra, o se extraen conclusiones poco fiables. Los medios de comunicación suelen hacerse eco de estos estudios, o simplifican otros estudios más serios. Esto lleva a muchos ciudadanos al engaño y a otros muchos a la desconfianza respecto a la estadística.

Las estadísticas muestran que casi todos los accidentes de circulación se producen entre vehículos que ruedan a velocidad moderada. Muy pocos ocurren a más de 150 Km/h. ¿Significa esto que resulta más seguro conducir a gran velocidad?

Quizá la causa del uso (mal)intencionado de las estadísticas sea su halo de neutralidad, objetividad, rigurosidad científica y verdad sacrosanta. Su comunicación al público es muy sencilla, pues son fáciles de entender. Igual de fácil es presentarlas de manera sesgada. Lo que resulta más complicado es comprender lo que se oculta, qué parte no se cuenta o qué truco se ha usado para falsearlas.

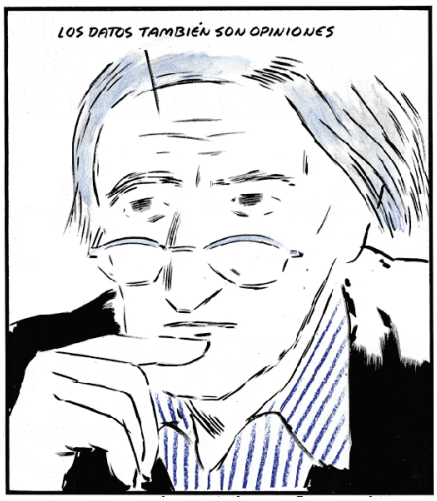


Figura 6: Viñeta de El Roto, 29-11-2010
Fuente: www.elpais.com/recorte/20101129elpepivin_3/XLCO/Ges/20101129elpepivin_3.jpg [12-6-2012]

Es importante, por tanto, analizar críticamente la información estadística que se nos presenta e indagar en la metodología del estudio en sí. Algunos trucos de los que hay que estar prevenidos son, entre otros: los ejes que no empiezan en cero (ejemplo), las mezcla de escalas (ejemplo), las muestras poco representativas (ejemplo), las preguntas tendenciosas (ejemplo), etc.

Desde el otro lado de la barrera, es importante no incurrir (consciente o inconscientemente) en este tipo de manipulaciones, y ser transparente a la hora de presentar los resultados de la investigación, detallando la metodología suficientemente. La investigación en desarrollo tiene un alto nivel de complejidad e incertidumbre, además de la habitual falta de tiempo, dinero o apoyo logístico suficientes. Aunque hay que intentar solucionar estas limitaciones para una mayor calidad del estudio, es aún más importante saber asumirlas y reconocerlas explícitamente en los informes y al presentar los resultados.

4.1.3 Programas informáticos

El rápido y sostenido incremento en el poder de cálculo de la computación desde la segunda mitad del siglo XX ha tenido un sustancial impacto en la práctica de la ciencia estadística. Un gran número de paquetes estadísticos está ahora disponible para los investigadores. Estos paquetes facilitan en gran medida la realización de cálculos de estadísticos, pruebas de hipótesis, ajustes de modelos, manejo de grandes bases de datos, representaciones gráficas, etc.

Sin embargo, a pesar de su menor potencia, en muchas ocasiones se emplean hojas de cálculo, puesto que su uso parece más sencillo y la mayoría de las personas están familiarizadas con ellas y tienen instalado en su ordenador Microsoft Excel. Hay otras hojas de cálculo similares de software libre, como Calc de Open Office. También hay hojas de cálculo que se pueden trabajar en línea de manera cooperativa, como las disponibles en GoogleDocs. Para cálculos básicos y volúmenes de datos reducidos, las hojas de cálculo pueden ser la solución más rápida y sencilla.

Para grandes volúmenes de datos sí se suelen utilizar paquetes estadísticos, ya que suelen tener una capacidad mayor. Hay muchos disponibles, que se diferencian según su potencia, su 'amigabilidad' hacia el usuario, si es software privado o libre, etc. Se destacan a continuación algunos de ellos.

SPSS se desarrolló inicialmente para las ciencias sociales y ofrece un uso sencillo de las opciones, acceso rápido a datos y características de las variables, procedimientos de análisis y generación de gráficos. Es un programa con una interfaz gráfica de usuario amigable. Es el más popular en investigaciones sociológicas.

PSPP es una alternativa al SPSS y es de software libre. Funciona prácticamente igual, aunque con menores prestaciones; solo permite hacer análisis simples.

InfoStat es un programa estadístico que también guarda cierta semejanza con el SPSS. Tiene una interfaz avanzada para el manejo de datos. Pensado para trabajar con Windows, su versión estudiantil se puede descargar gratuitamente.

Statgraphics es un programa para gestionar y analizar valores estadísticos. Destaca especialmente por sus capacidades para la representación gráfica de todo tipo de estadísticas y el desarrollo de experimentos, previsiones y simulaciones en función del comportamiento de los valores.

SAS ha sido por largos años el software más utilizado entre la comunidad estadística por su gran potencia de cálculo. Es un programa que requiere el ingreso de comandos para ejecutar gran parte de sus rutinas y opciones.

R es un programa estadístico y un lenguaje de programación de uso libre. De distribución gratuita y de código abierto, ha sido desarrollado como un gran proyecto colaborativo de estadísticos de diversos países y disciplinas. También se basa en el uso de comandos.

Existen también programas que permiten el análisis estadístico de información obtenida mediante métodos cualitativos. Los datos registrados en forma de notas tomadas durante una observación, las respuestas libres a preguntas abiertas, las transcripciones de entrevistas individuales o discusiones de grupo, los libros y los artículos periodísticos, entre otros, pueden ser procesados mediante el tratamiento cuantitativo. El procedimiento interpretativo estándar comprende: reducción de los datos, selección de palabras claves, agrupamientos de frases en dimensiones, edición de categorías exhaustivas y codificación de categorías. El análisis se transforma en una cuantificación de códigos numéricos, el recuento de códigos y la obtención de distribuciones de frecuencias. Algunos de estos programas son *Atlas.ti*, *NVivo, Sonal* o *Hyper-research*.

4.2 Conceptos básicos

Una vez recolectada la información mediante encuestas, mediciones, observación o talleres participativos, llega el paso de analizar, resumir y presentar los datos de la muestra, con la inestimable ayuda de la estadística descriptiva.

En capítulos anteriores se han definido ya los principales conceptos que se utilizan en éste. Repasamos aquí los más relevantes e introducimos algunos nuevos.

El **sujeto** es la unidad de la población de la que buscamos información. Pueden ser familias, personas, o incluso comunidades.

La **variable** es la característica que se pretende estudiar, es decir, lo que queremos conocer y vamos a observar (medir, preguntar) a cada sujeto (altura, opinión sobre algo). Hay distintos tipos, que repasamos en el apartado siguiente con varios ejemplos. Se llaman variables porque *varían*, toman valores distintos de un sujeto a otro.

Asoman en este último párrafo dos conceptos nuevos que son importantes.

El **valor** es, como su propio nombre indica, el valor obtenido para una variable determinada al recolectarla de un determinado sujeto. Por ejemplo: 173 cm (en el caso de la variable altura). Se podría traducir como resultado o respuesta. Para cada sujeto, la variable tomará un valor determinado. A las variables se les suele asignar una letra o un código como 'x' o 'alt'.

Valor: resultado de una variable al ser recolectada de un determinado sujeto.

Observación: acto de obtener el valor de la variable de un sujeto.

Estadístico muestral: medida cuantitativa calculada a partir de un conjunto de datos de una muestra.

La **observación** es al acto de preguntar o medir la variable en un sujeto. En realidad es una forma general y numerada de referirnos al sujeto encuestado. Así respecto a la variable altura, no diremos "sujeto 1: 173 cm", sino "observación 1: 173 cm".

Así, tenemos una tripla 'variable-valor-observación', que se suele representar sintéticamente con el código de la variable, el número de observación entre paréntesis o subíndice (genéricamente es 'i') y el valor correspondiente: 'var (i) = ___'. Sintetizaríamos pues el ejemplo anterior como: 'alt_1 = 173 cm', o 'x (1) = 173 cm'.

Finalmente, llamaremos **estadístico** al número que obtenemos después de resumir el conjunto de valores de una variable observados en una muestra. Aunque a veces no se explicita, los estadísticos son siempre estadísticos muestrales. Un ejemplo, será la altura media de la muestra. El estadístico sirve luego para estimar un determinado parámetro de la población de la que procede la muestra. Así, por ejemplo, del estadístico (muestral) altura media podremos estimar el **parámetro poblacional** altura media aplicando la inferencia estadística. Conviene tener en cuenta que en ese punto, ya no lo denominamos estadístico, sino parámetro.

Retomemos el ejemplo de la investigación sobre el nivel de ingresos familiar de la región Logone Occidental en Chad, para ver todos estos conceptos en la práctica:

La población será el conjunto de familias de dicha región; cada familia sería un sujeto. Si se realiza una encuesta a 1000 familias, esas 1000 familias constituyen la muestra. La variable más importante a estudiar serían los ingresos familiares, que podemos codificar como 'ingfam'. No obstante, habría otras variables interesantes como los gastos en alimentación (gastalim), la etnia, el número de miembros de la familia, el departamento, el sexo o la edad de la/el cabeza de familia.

A medida que se realiza la encuesta a las distintas familias (o sujetos) –tienen lugar las observaciones– se irían obteniendo los valores correspondientes de ingreso: Familia 1: 320.000 francos CFA; Familia 2: 325.000 francos CFA; etc.

Lo representaríamos como observaciones:

ingfam (1) = 320000 CFA

ingfam (2) = 325000 CFA

Y luego se organizarían en forma de tabla junto con el resto de variables, como base de datos para su

posterior análisis.

Observación	ingfam	etnia	numfam	…
1	322000 CFA	Baggara	3	…
2	412000 CFA	Hausas	8	…
3	354000 CFA	Masalit	5	
4	386000 CFA	Hausas	4	
…	…	…		

De dicho análisis y para cada variable, se obtendrían diversos estadísticos muestrales, por ejemplo el ingreso familiar medio (para estas 4 observaciones): 368500 CFA.

Actividad de refuerzo 1:

Explica con tus propias palabras las diferencias entre los conceptos clave (en negrita) vistos en este apartado.

Realiza este breve test.

4.2.1 Tipos de variables

Ya sabemos que las **variables** son las características que queremos estudiar, es decir, lo que queremos conocer de la muestra y la población. Pero hay muchas características distintas, así que las variables se clasifican, según su tipología, entre cualitativas y cuantitativas.

Las **variables cuantitativas** se expresan mediante números y representan cantidades (ingresos, edad, número de miembros de la familia). Pueden ser continuas o discretas.

- Una **variable cuantitativa continua** puede tomar cualquier valor real dentro de su intervalo de validez. Por ejemplo, el peso de la cosecha de trigo puede ser de 35.743,97 kilos.
- Una **variable cuantitativa discreta** sólo puede tomar ciertos valores enteros, presentando separaciones o interrupciones en la escala de valores que puede tomar. Por ejemplo, el número de miembros de la familia puede ser 1; 4; 9..., pero no puede ser 0,4.

Las **variables cualitativas** expresan características que no se pueden medir con números, como pueden ser el sexo, la etnia, o el grado de satisfacción con el nivel de ingresos. Son variables cualitativas que se analizan cuantitativamente. Se pueden codificar numéricamente sus diferentes alternativas para poder aplicar algunas operaciones con paquetes estadísticos básicos, como por ejemplo el cálculo de la moda, estadístico que veremos más adelante. Así, para la variable cualitativa 'sexo', se puede asignar el valor 1 cuando sea mujer y el valor 2 cuando sea hombre. Dentro de las variables cualitativas, distinguimos entre las ordinales y nominales.

- Una **variable cualitativa ordinal** puede tomar distintos valores ordenados siguiendo una escala establecida, aunque no es necesaria una proporcionalidad, ni que el intervalo entre mediciones sea regular. Ejemplos: el grado de satisfacción profesional puede ser: muy bajo, bajo, medio, alto o muy alto.
- Una **variable cualitativa nominal** no puede ser sometida a un criterio de orden jerárquico o proporcional. Ejemplos: la etnia o el sexo.

En estudios en desarrollo, las variables cualitativas son tan comunes como las cuantitativas. Distinguir entre ambos tipos es importante, pues las medidas, representaciones y cálculos asociados a cada una son

diferentes. Por ejemplo, la media se utiliza para variables cuantitativas, mientras que la frecuencia o el porcentaje se emplean con variables cualitativas, como veremos en los siguientes apartados.

Otra clasificación de las variables se refiere a su influencia mutua. Así, se distingue entre **variables dependientes** y **variables independientes**. El valor de la variable dependiente 'depende' en mayor o menor medida del valor de la variable independiente. Por ejemplo, si piensa que el tamaño de la familia influye en el nivel de ingresos, para estudiar esa influencia se puede tomar el número de miembros de la familia como variable independiente, y los ingresos como variable dependiente.

Actividad de refuerzo 2:

Haz un esquema con los distintos tipos de variables. Para cada tipo de variable, explica sus características más importantes y un ejemplo.

Para cada una de las variables que se presenta a continuación, di si es cualitativa nominal, cualitativa ordinal, cuantitativa discreta o cuantitativa continua: ingresos anuales, sexo, número de gallinas, hectáreas en propiedad, edad, número de miembros de la familia, nivel de satisfacción con el servicio eléctrico (alto, medio o bajo), gasto en medicinas al año, lugar de nacimiento, peso, nivel educativo alcanzado (ninguno, primaria, secundaria, superior).

4.3 Análisis unidimensional. Principales estadísticos, tablas y gráficos

Como ya hemos dicho, la estadística descriptiva pretende ayudar a analizar los datos originados a partir de la recolección de información, realizada por ejemplo mediante una encuesta. Tras una encuesta a 500 personas, ¿resulta factible o interesante revisar qué ha respondido cada uno de los sujetos a cada variable (o pregunta)? Sería muy poco práctico, y por eso utilizamos la estadística descriptiva, que nos ofrece diferentes estadísticos, tablas y gráficos para resumir y visualizar de manera sintética los resultados. A continuación, iremos conociendo algunos de ellos.

El análisis unidimensional, objeto de este apartado, es cuando se estudian las variables una por una. Cuando se estudian dos variables a la vez (por ejemplo su relación), hablamos de análisis bidimensional, que es el objeto del apartado 4.4.

En el análisis unidimensional, si la variable es cualitativa, nos interesa sobre todo conocer las frecuencias, bien en forma de porcentaje, en una tabla o en gráficos de barras o sectores. Si la variable es cuantitativa se suelen utilizar más las medidas de posición (como la media) y dispersión (como la desviación típica), representándolas mediante histogramas. Tanto o más importante que conocer cómo se calculan los distintos estadísticos tablas y representaciones, es ser capaz de seleccionarlos adecuadamente, en función del tipo de variable que se esté analizando (cuantitativa o cualitativa).

4.3.1 Las frecuencias

La **frecuencia** es un estadístico que se refiere a la cantidad de veces que una variable toma un valor determinado. Se puede expresar como un número (sale tantas veces) o como una proporción o porcentaje (sale en un tanto por ciento), es decir, como frecuencia absoluta o como frecuencia relativa.

Frecuencia: cantidad de veces que una variable toma un valor determinado

La **frecuencia absoluta** (n_i) de un valor (X_i) expresa el número de observaciones en que la variable (X) toma ese determinado valor. En forma de pregunta: ¿Cuántas veces aparece ese valor?

La **frecuencia relativa** (f_i) de un valor (X_i) es la proporción de observaciones en que la variable (X) toma ese determinado valor. Se obtiene dividiendo la cantidad de veces que aparece el valor (frecuencia abso-

luta) entre el total de observaciones, es decir, el tamaño de la muestra 'n': $f_i=n_i/n$. En forma de pregunta: ¿En qué proporción aparece ese valor?

Multiplicando la frecuencia relativa por 100, se obtiene el **porcentaje** o tanto por ciento (p_i). El porcentaje es el estadístico por excelencia de las variables cualitativas.

Si no te gustan mucho las matemáticas, no te preocupes, con un ejemplo quedará mucho más claro:

Ejemplo: En un examen de estadística los 18 alumnos y alumnas obtienen las siguientes puntuaciones (sobre 20):

18, 13, 12, 14, 11, 8, 12, 15, 5, 20, 18, 14, 15, 11, 10, 10, 11 y 13

La variable es la puntuación y tenemos 18 observaciones.

El valor 11 aparece 3 veces, así que su frecuencia absoluta es $n_i(11) = 3$.

La proporción de veces que aparece la puntuación 11, es decir, la frecuencia relativa de 11, se obtiene dividiendo por el total de observaciones: $f_i(11) = 3/18 = 0,17$. Expresado en porcentaje sería $p_i(11) = 17\%$.

Hay otra variante de las frecuencias que son las frecuencias acumuladas:

La **frecuencia absoluta acumulada** (N_i) es el número de veces que la variable toma un valor determinado o un valor menor que ese valor determinado. En forma de pregunta: ¿Cuántas veces aparece ese valor o valores menores a ése?

La **frecuencia relativa acumulada** (F_i) es la proporción de las veces que aparece ese valor o uno inferior. Al igual que antes, se obtiene dividiendo la frecuencia absoluta acumulada entre el total de observaciones (el tamaño de muestra 'n'): $F_i=N_i/n$. En forma de pregunta: ¿En qué proporción aparece ese valor o valores inferiores? Multiplicando la frecuencia relativa acumulada por 100, se obtiene el porcentaje acumulado (P_i).

Siguiendo con el ejemplo: Para calcular la frecuencia absoluta acumulada $N_i(11)$, se mira cuántas observaciones hay por debajo del 11: hay un 8, un 5 y dos 10. Por lo tanto, además de las tres veces que aparece 11, hay otras cuatro observaciones con valores inferiores a 11. La frecuencia absoluta acumulada es $N_i(11)=7$.

Como en total hay 18 observaciones, la proporción de veces que aparece la puntuación 11 o inferior, es decir, la frecuencia relativa acumulada, es $F_i(11) =7/18=0,389$. Expresado en porcentaje sería $P_i(11)=39\%$.

Las frecuencias son conceptos sencillos, pero es importante tenerlos muy claros para entender otros conceptos más avanzados.

Las frecuencias de toda una muestra se representan en una **tabla de frecuencias simple**:

Variable X	Frecuencias absolutas		Frecuencias relativas	
(Valor x_i)	Simple (n_i)	Acumulada (N_i)	Simple (f_i)	Acumulada (F_i)
X_1	n_1	$N_1 = n_1$	$f_1 = n_1 / n$	$F_1 = f_1$
X_2	n_2	$N_2 = n_1 + n_2$	$f_2 = n_2 / n$	$F_2 = f_1 + f_2$
...	...	...	...	...
X_n	n_n	$N_n = \Sigma (n_i)$	$f_n = n_n / n$	$F_n = \Sigma (f_i)$

Estas tablas recogen las frecuencias de todos los valores de una variable, y pueden estar ordenadas de distintas maneras, pudiendo incluir o no los porcentajes. Para el ejemplo que hemos empleado antes, una posible tabla (organizada de manera diferente a la anterior) sería:

	Frecuencias			Frecuencias acumuladas		
Valor	absoluta	relativa	porcentual	absoluta	relativa	porcentual
x_i	n_i	f_i	p_i	N_i	F_i	P_i
5	1	0,06	5,56%	1	0,06	5,56%
8	1	0,06	5,56%	2	0,11	11,11%
10	2	0,11	11,11%	4	0,22	22,22%
11	3	0,17	16,67%	7	0,39	38,89%
12	2	0,11	11,11%	9	0,50	50,00%
13	2	0,11	11,11%	11	0,61	61,11%
14	2	0,11	11,11%	13	0,72	72,22%
15	2	0,11	11,11%	15	0,83	83,33%
18	2	0,11	11,11%	17	0,94	94,44%
20	1	0,06	5,56%	18	1,00	100,00%

La tabla de frecuencias sirve para resumir la distribución de los resultados y se puede utilizar con variables cualitativas y con variables cuantitativas discretas (aunque no siempre es interesante). Su aplicación a variables cuantitativas continuas no resulta muy útil, puesto que suele haber muchos valores distintos y muy pocas repeticiones. Imagínese por ejemplo una tabla de frecuencias con los ingresos familiares.

En esos casos, es más interesante utilizar una tabla de frecuencias agrupada. Esta tabla se puede utilizar para cualquier tipo de variable. Para obtenerla, en vez de calcular las frecuencias para cada valor, se crean intervalos de valores para agruparlos y se calculan las frecuencias para esos intervalos, es decir, el número de observaciones con valores que se encuentran dentro de cada intervalo. El número de tramos en los que se agrupa la información es una decisión del investigador, según lo resumida que quiera tener la información. Se debe buscar un equilibrio, ya que demasiados tramos pueden complicar la lectura de los datos y demasiados pocos tramos nos hacen perder información. Es aconsejable que los intervalos tengan el mismo tamaño, aunque a veces puede ser conveniente dejar intervalos abiertos en los extremos (ver primer intervalo de la tabla siguiente).

Para el ejemplo anterior, se podría elaborar la siguiente tabla de frecuencias agrupada:

Intervalo	Frecuencias			Frecuencias acumuladas		
valores	absoluta	relativa	porcentual	absoluta	relativa	porcentual
$[x_i$ a $x_j]$	n_i	f_i	p_i	N_i	F_i	P_i
< 5	1	0,06	5,56%	1	0,06	5,56%
6 a 10	3	0,17	16,67%	4	0,22	22,22%
11 a 15	11	0,61	61,11%	15	0,83	83,33%
16 a 20	3	0,17	16,67%	18	1,00	100,00%

Otras medidas relacionadas con la frecuencia son la razón y la tasa:

La razón es una comparación por cociente entre dos cifras de diferente o similar naturaleza. Si en la clase del ejemplo hay 10 alumnas y 8 alumnos, la razón de feminidad de la clase es 10/8=1,25.

Tasa es un tipo de proporción que se calcula para una población en un periodo determinado. Por ejemplo, si en el 2009, en una región con 20.000 alumnos y alumnas matriculados, 760 no asistieron regularmente a clase, la tasa de absentismo para ese año es de 760/20000 = 0,038, aunque se suele expresar como un cociente (38/1000) para su mejor compresión: 38 de cada 1000 escolares no asistieron a clase con regularidad.

Actividad de refuerzo 3:

En un estudio preliminar para la investigación anteriormente citada sobre la chadiana región de Logone Occidental, se han realizado encuestas a 33 familias, obteniendo los resultados que se presentan bajo este cuadro y que están disponibles también en línea. A partir de esos datos:

1) elabora una tabla de frecuencias simple para la variable número de miembros de la familia,

2) elabora una tabla de frecuencias para la variable ingresos familiares. Decide razonadamente si elaboras una tabla de frecuencias simple o agrupada y explica por qué

3) y calcula los porcentajes de la variable departamento.

Familia (observación)	Número de miembros	Ingresos familiares (en miles de francos CFA)	Departamento
1	9	322	Dodjé
2	6	412	Lac Wey
3	8	354	Guéni
4	3	386	Dodjé
5	4	295	Dodjé
6	6	366	Ngourkosso
7	5	301	Guéni
8	7	345	Lac Wey
9	5	231	Dodjé
10	6	383	Lac Wey
11	4	365	Guéni
12	6	259	Lac Wey
13	7	312	Ngourkosso
14	3	346	Lac Wey
15	5	328	Dodjé
16	2	180	Lac Wey
17	7	457	Ngourkosso
18	9	320	Guéni
19	13	978	Ngourkosso
20	5	267	Guéni
21	6	401	Dodjé
22	8	326	Lac Wey
23	5	502	Lac Wey
24	6	284	Guéni
25	10	350	Ngourkosso
26	6	327	Ngourkosso
27	8	385	Lac Wey
28	12	299	Dodjé
29	26	430	Ngourkosso
30	5	333	Dodjé
31	10	310	Dodjé
32	6	361	Ngourkosso
33	4	291	Guéni

4.3.2 Representaciones gráficas de las frecuencias. La distribución

A la hora de analizar las frecuencias, puede ser interesante representar las tablas de una manera más visual, para lo que se dispone de diferentes tipos de representaciones gráficas.

El **diagrama de barras** se suele utilizar para presentar las frecuencias de variables cualitativas. Para cada valor que puede tomar la variable, se construye una barra o columna de altura proporcional a la frecuencia con la que ha aparecido. Se puede hacer a partir de una tabla de frecuencias, tanto relativas o porcentuales como absolutas. Aunque no es muy común, se pueden usar también frecuencias acumuladas.

Análogamente, el **diagrama de sectores** (más popularmente conocido como tarta) representa la frecuencia observada mediante el área de los sectores de un círculo.

Se presentan dos tablas de frecuencia con variables cualitativas y sus respectivos diagramas, a partir del ejemplo anterior y de una encuesta sobre consumo de productos de Comercio Justo.

Sexo del alumnado de la clase

Tabla de frecuencias

Sexo	Frecuencia absoluta (n_i)
Alumnos	8
Alumnas	10

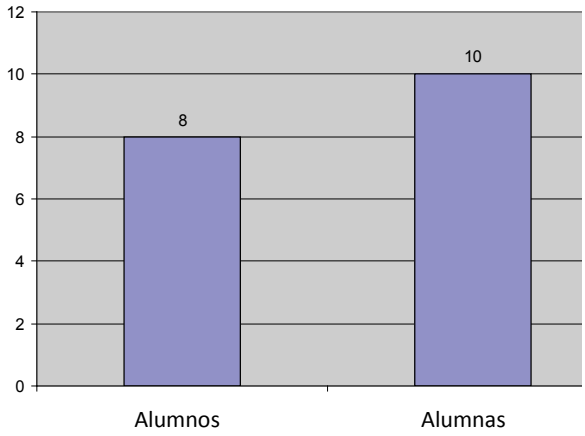


Figura 7: Diagrama de barras sobre sexo del alumnado
Fuente: elaboración propia

¿Ha consumido algún producto de Comercio Justo en el año 2007?

Tabla de frecuencias

consumCJ	absoluta (n_i)	relativa (f_i)	porcentual (p_i)
Sí	837	0,26	26,01%
No	2253	0,70	70,01%
No contesta	128	0,04	3,98%

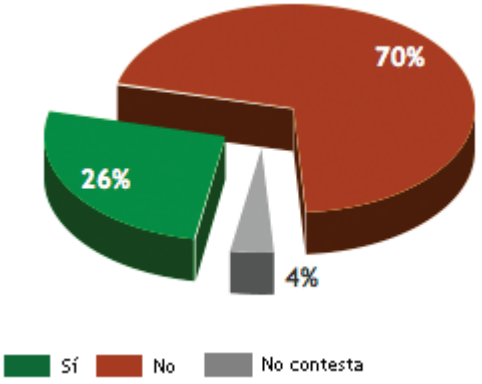


Figura 8: Diagrama de tarta sobre consumo de Comercio Justo
Fuente: barometro.fundacioneroski.es/2007/consumo-de-productos-de-comercio-justo [12-6-2012]

Para variables cuantitativas, resulta más interesante utilizar un **histograma**. A partir de una tabla de frecuencias simple (para cuantitativas discretas) o agrupada (para cuantitativas discretas y continuas), se elabora una representación gráfica en forma de columnas, cuyas alturas son proporcionales a la frecuencia (relativa o absoluta) de los valores representados. Es muy parecido a un diagrama de barras, con la diferencia de que el eje horizontal también tiene escala; es como si fuese una regla, con intervalos proporcionales numerados. Como se ve en el ejemplo más abajo, para frecuencias agrupadas las barras se sitúan en la mitad del intervalo. Se le puede añadir una línea (en azul en el ejemplo) para formar lo que se conoce como el polígono de frecuencias.

Análogamente se pueden elaborar histogramas y polígonos de frecuencias acumuladas.

A partir de la tabla anterior de frecuencias agrupadas de las puntaciones del alumnado, se obtendrían los histogramas (y polígonos) que aparecen a continuación.

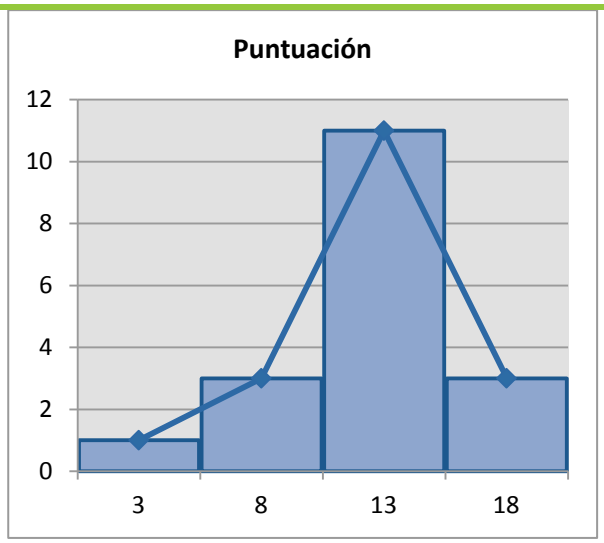


Figura 9: Histograma de frecuencias absolutas
Fuente: elaboración propia

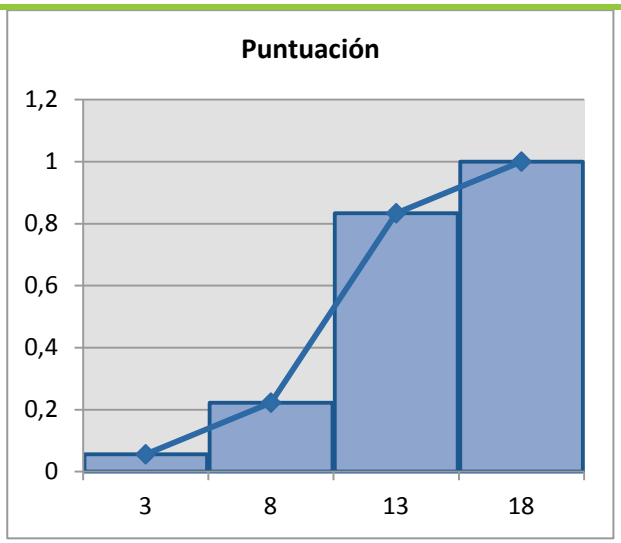


Figura 10: Histograma de frecuencias relativas acumuladas
Fuente: elaboración propia

Cuando la variable es continua y la muestra es lo bastante grande, se podría hacer un histograma con las frecuencias sin agrupar. En realidad estaríamos hablando una **distribución de frecuencias continuas**. En el eje horizontal aparecen los valores que puede tomar la variable y en el eje vertical la frecuencia (relativa) con la que aparece.

Suponiendo que se hace nuevamente el examen de estadística a un grupo muy grande de alumnos y alumnas, la distribución de frecuencias continuas quedaría tal como se aprecia en la figura siguiente. Su polígono de frecuencias sería prácticamente una curva.

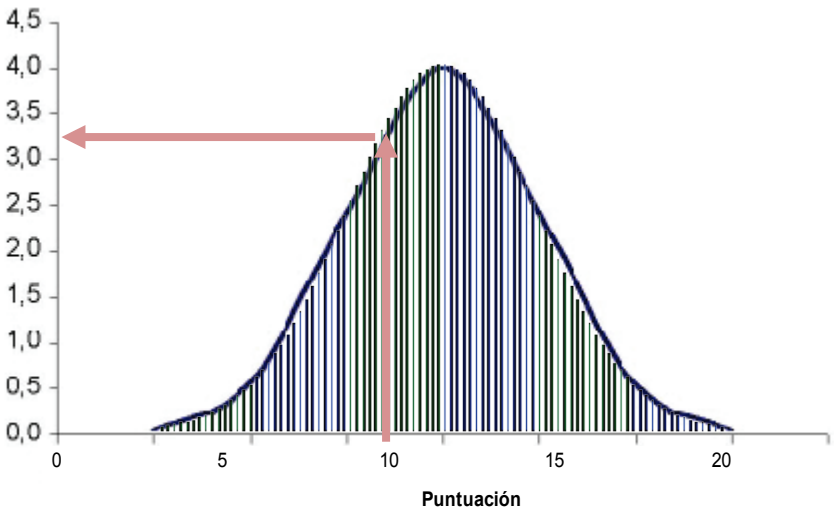


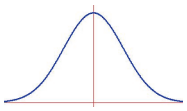
Figura 11: Diagrama de barras sobre sexo del alumnado
Fuente: elaboración propia

Esa curva se llama distribución de densidad de frecuencias o **distribución de probabilidad**, y representa en vertical la proporción con que aparece cada valor (del eje horizontal). En la curva del ejemplo, las flechas sirven para ejemplificarlo. La proporción (frecuencia relativa) de alumnos que obtienen un 10 es de 3,2% aproximadamente.

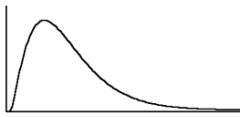
Distribución: curva que indica la probabilidad de observación de toda la gama de valores que puede presentar una variable.

Las **distribuciones** son muy útiles para visualizar rápidamente cómo se reparten los distintos valores en la muestra, aunque no suelen ser prácticas para encuestas de tamaño medio o pequeño.

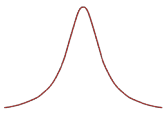
Por otro lado, es muy importante comprender el concepto de la distribución en sí, pues hay distintos modelos teóricos de distribuciones que son útiles para entender diversos fenómenos o para calcular, por ejemplo, el tamaño de muestra necesario. Así encontraremos tipos de variables que se distribuyen de forma simétrica, otras de forma asimétrica, etc.



Distribución normal



Distribución asimétrica



Distribución simétrica apuntada

Figura 12: Simetría en las distribuciones
Fuente: elaboración propia

La distribución del ejemplo anterior es una distribución simétrica, más concretamente una distribución normal. Esta distribución es muy común en la vida real y, al realizar histogramas a partir de una muestra, en muchas ocasiones tendrán una forma parecida a la distribución normal. Las notas de un examen, el peso de personas de una misma edad son ejemplos de tipos de variable que suelen presentar una distribución normal.

Otras variables pueden dar otro tipo de distribuciones.

Por ejemplo, las variables económicas como los ingresos familiares suelen presentar distribuciones asimétricas positivas, donde gran parte de la muestra tiene unos ingresos bajos (cima de la curva en el lado izquierdo) y una pequeña parte tiene ingresos muy altos (cola alargada hacia la derecha). Recordando el sistema de las flechas, vemos que con bajos ingresos (parte izquierda del eje horizontal) hay un alto porcentaje de personas (cima de la curva). Con ingresos altos, hay un porcentaje bajo (cola de la curva muy baja).

Cuando la muestra es muy homogénea, las distribuciones son más bien apuntadas, mostrando mucha concentración de los datos alrededor de la media. Es decir, si en una ciudad no hay ni ricos ni pobres, todos los habitantes tendrán una renta parecida, que coincidirá con el valor que se sitúa bajo la cima de la curva, al representar ésta el valor más observado.

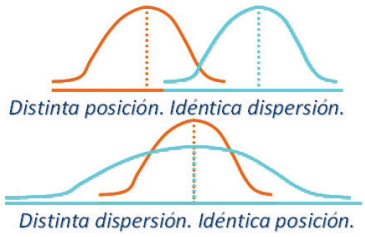


Figura 13: Posición y dispersión
Fuente: www.erasmusweb.com/admin/pdf/53/224/Tema%201%20Estadistica%20descriptiva%20unidimensional.pdf [12-6-2012]

Hay también distribuciones asimétricas negativas, distribuciones simétricas que no son normales, etc. Más adelante se profundizará en esto, pero es vital entender bien el concepto de distribución y lo que representa para avanzar en este tema, por lo que se recomienda releer este apartado hasta que quede claro.

A modo de resumen, recordar que las distribuciones son curvas que representan la proporción (frecuencia relativa) con que se observan los distintos valores de una determinada variable obtenidos de una muestra. Nos facilitan de un vistazo información sobre la posición (¿dónde se sitúa la mayoría?) y dispersión (¿están concentrados o hay muchas diferencias?).

Pero como con los vistazos no nos vale, veremos a continuación formas de medir tanto la posición como la dispersión.

Actividad de refuerzo 4:

Revisa los ejemplos de diagrama de tarta e histograma y responde para cada uno: ¿Cuál es la variable? ¿De qué tipo de variable se trata? ¿Qué valores puede tomar?

A partir de las tablas de frecuencias realizadas en las actividades de refuerzo anteriores, elabora sen-

dos histogramas de frecuencias simples.

A partir de los datos del estudio en Logone Occidental, haz un diagrama de barras o de sectores para representar cuántas familias han sido entrevistadas en cada departamento.

Explica, con tus propias palabras, lo que es una distribución.

4.3.3 Medidas de posición

Las tablas de frecuencias y las distribuciones contienen una parte considerable de la información de la muestra. Para resumirla aún más y así facilitar el manejo y comparación de las variables, se suelen emplear estadísticos que caractericen su posición y su dispersión. La dispersión se tratará en el siguiente apartado.

Posición: ubicación representativa de un conjunto de valores obtenidos de una muestra respecto a una variable.

La **posición** trata de resumir los valores que toma una variable calculando un valor promedio. Esto se entenderá más claramente a continuación, con la explicación y ejemplificación de los diferentes tipos de 'promedios' en estadística, es decir, las distintas medidas de posición central. Las más comunes son la media, la mediana y la moda.

La **media** es la medida de posición más popular. Se usa, por ejemplo, para calcular la renta per cápita de un país. La media muestral de una variable X es la suma de los valores de todas las observaciones de esa variable (el sumatorio Σ) dividida entre el tamaño de la muestra 'n'. En fórmula matemática sería:

$$\bar{x} = \Sigma(x_i) / n$$

Si no queda claro, el ejemplo de abajo será de ayuda. Es importante tener presente para qué variables tiene sentido calcular la media: ¿Se puede hacer la media de la variable sexo? ¿Y de la variable nivel educativo? En efecto, sólo se puede calcular la media de variables cuantitativas.

La **mediana** es el valor 'de en medio', es decir, el valor que tiene tantas observaciones con valores mayores que él, como menores que él. Para obtenerla, se deben ordenar de menor a mayor todas las observaciones. La mediana será el valor que deje el mismo número de observaciones a cada lado. En caso de que haya un número par de observaciones, no existirá una observación 'central', sino dos. En tal caso, la mediana es la media de esas dos observaciones. ¡De nuevo solo para variables cuantitativas!

La **moda** es otra medida de posición, que simplemente da el valor más frecuente (el que está 'de moda'). La moda se puede calcular para cualquier tipo de variable, siendo de especial utilidad para describir variables cualitativas. Acepta todo tipo de variables.

Retomando el ejemplo del examen de estadística, la media de la puntuación obtenida por el alumnado se obtendría sumando las puntuaciones y dividiéndolas entre el total de alumnos y alumnas. La suma se puede hacer indistintamente a partir de los datos (segunda línea) o de la tabla de frecuencias simples absolutas (tercera línea):

$\bar{x} = \Sigma(x_i) / n = (18+13+12+14+11+8+12+15+5+20+18+14+15+11+10+10+11+13) / 18 = 230 / 18 = 12{,}778$

Para calcular la mediana, ordenamos las puntuaciones: 5 8 10 10 11 11 11 12 **12 13** 13 14 14 15 15 18 18 20. Al haber un número par de observaciones, quedan dos observaciones 'centrales': 13 y 12, con lo que la mediana sería = (13+12) / 2 = 12,5

La moda sería la puntuación que más veces se repite, en este caso: 11

Las siguientes distribuciones nos permiten ver de manera gráfica la media, la mediana y la moda. Es un buen momento para asentar el concepto de distribución. Por ejemplo con la moda, que al ser el valor más frecuente, coincide siempre con el pico de la distribución (frecuencia más alta).

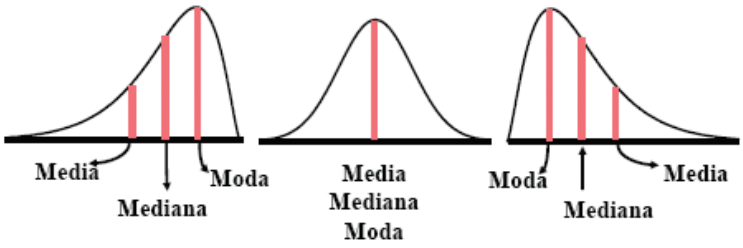


Figura 14: Media, mediana y moda en distintas distribuciones
Fuente: www.tuveras.com/estadistica/asimetria.gif [12-6-2012]

En estas distribuciones y en el cuadro siguiente, se puede ver cómo la mediana puede ser una medida interesante cuando existen valores 'extremos' que distorsionan la media.

Salarios Chicago Bulls (1997)

Jud Buechler $500.000	Luc Longley $3.184.900	S. Burrell $1.430.000	Robert Parish $1.150.000	Randy Brown $1,260,000
Jason Caffey $850.920	Scottie Pippen $2.775.000	Ron Harper $4.560.000	Dennis Rodman $4.500.000	Rusty LaRue $242,000
Michael Jordan $33.140.000	Dickey Simpkins $1.235.000	Steve Kerr $750.000	David Vaughn $693.840	Toni Kukoc $4.560.000
Joe Kleine $272.250	Bill Wennington $1.800.000	Keith Booth $597.600		

media: $3.527.862 mediana: $1.247.500 moda: $4.560.000

En ocasiones es necesario calcular la media a partir de otras medias. Para calcular, por ejemplo, la esperanza de vida en la región de Logone Occidental, se dispone de los datos de la esperanza de vida media en los 4 departamentos que la integran:

Departamento	Dodjé	Ngourkosso	Guéni	Lac Wey	Total Logone Occidental
Población (2009)	105.126	157.142	94.529	326.496	683.293
Esperanza de vida * ***(años)***	45	46	41	52	
Mortalidad infantil * ***(de cada 1000 nacidos vivos)***	99	103	105	91	

* Datos aproximados

Si se calcula la media de los 4 valores directamente, se obtendría 46 años, pero no sería correcto, puesto que la esperanza de vida en Guéni debería contar menos que la esperanza de vida en Lac Wey, dada la disparidad en número de habitantes. En estos casos, es necesario calcular la media ponderada.

La **media ponderada** se utiliza para calcular la media a partir de valores con pesos diferentes. Para ello, se debe multiplicar cada valor por su peso (en porcentaje) y después sumarlos.

Un ejemplo cercano es el cálculo de la nota de muchas asignaturas, en las que distintos ejercicios y pruebas tienen un peso determinado, y hay que multiplicar la nota de cada ejercicio por ese peso para obtener la nota de la asignatura.

Si el peso porcentual se representa con una w, la fórmula sería:

$$\bar{x}_w = \Sigma(w_j \cdot x_j) = x_1 \cdot w_1 + x_2 \cdot w_2 + \ldots + x_n \cdot w_n$$

El peso suele venir del porcentaje de personas o elementos que son representados por cada valor.

En el ejemplo que se ha puesto, los pesos se obtienen dividiendo la población del departamento por el número total de habitantes, para Dodjé:

$w_1 = 105.126 / 683.293 = 15{,}39\%$.

Departamento	Dodjé	Ngourkosso	Guéni	Lac Wey	Total Logone Occidental
Población (2009)	105.126	157.142	94.529	326.496	683.293
Peso (w_i)	15,39%	23,00%	13,83%	47,78%	

Así, la media ponderada es:

$$\bar{x}_w = 45 \cdot 15{,}39\% + 46 \cdot 23\% + 41 \cdot 13{,}83\% + 52 \cdot 47{,}78\% = 48$$

La media ponderada se utiliza mucho cuando la técnica de muestreo es estratificada o por etapas y un estrato o conglomerado esta sobrerrepresentado en la muestra (ver apartado 2.3.2). En ese caso, para calcular la media muestral, los valores se deberán ponderar con un peso inverso a su sobrerrepresentación en la muestra.

Por ejemplo, pensemos en un colegio intercultural donde se quiere estudiar el número de asignaturas pendientes y ver su relación con el origen o cultura. En ese colegio hay 1000 alumnos y alumnas, siendo un 3% gitanos, un 20% inmigrantes magrebíes y un 77% payos, y habiendo paridad entre alumnos (500) y alumnas (500). Se decide tomar una muestra estratificada por sexo y origen étnico, y un tamaño aproximado del 10% de la población, es decir, de unos 100 sujetos. Si la muestra fuese proporcional estaría compuesta por 50 alumnos y 50 alumnas, y en cuanto al origen, 3 gitanos, 20 magrebíes y 77 payos.

Por su bajo porcentaje en la escuela, hay muy pocos gitanos en la muestra (ni siquiera da para 2 gitanos y 2 gitanas), y la información que se obtendría sobre ellos sería muy pobre. Teniendo en cuenta que es un colectivo de especial interés para el estudio, se decide aumentar su representación en la muestra hasta 12 sujetos, número que los investigadores han considerado suficiente para analizar su situación. En consecuencia, mientras 1 de cada 10 payos y 1 de cada 10 inmigrantes es muestreado, lo son 4 de cada 10 gitanos.

Una vez obtenido el número de asignaturas pendientes de cada sujeto, para calcular la media, los valores de los gitanos y gitanas deberán ponderarse, multiplicándolos por w=1/4.

Ésta es una forma simplificada (pero válida) de calcular el peso. En términos generales, se debería calcular el peso de cada sujeto, como el porcentaje de representación de su grupo (estrato o conglomerado) en la población, dividido entre el porcentaje de representación de su grupo en la muestra.

$$w_g = \%_g(\text{población}) \,/\, \%_g(\text{muestra})$$

Siguiendo con el ejemplo anterior, calculemos dicho cociente para los 3 estratos:

Origen	Población (cuántos hay en el colegio)	Porcentaje de representación en la población	Muestra (cuántos elegimos para el estudio)	Porcentaje de representación en la muestra	**Peso w** (% en población / % en muestra)
Gitano	30	3,0%	12	11,0%	**0,27**
Africano	200	20,0%	20	18,3%	**1,09**
Payo	770	77,0%	77	69,7%	**1,09**
Total	1000		109		**1,00**

La relación entre los pesos es de 1 a 4, con lo que es coherente con lo visto en el ejemplo anterior.

A la hora de calcular las asignaturas pendientes medias, ponderaremos las respuestas de cada sujeto por el peso 'w' correspondientes según su origen.

Por último los índices compuestos que elaboran diversos organismos internacionales y organizaciones sociales suelen ser medias ponderadas (de manera arbitraria) de distintos indicadores que se consideran relevantes. En casos excepcionales, como el nuevo Índice de Desarrollo Humano, se utiliza la media geométrica en lugar de la media normal (aritmética). En vez de sumar los indicadores y dividirlos por el total, se multiplican y se saca su n-sima raíz (n es el número de indicadores implicados). $\bar{x}_g = \sqrt[n]{(x_1 \cdot x_2 \cdot \ldots \cdot x_n)}$. A nivel práctico, la diferencia frente a la media aritmética es que la media geométrica se reduce mucho con las diferencias entre los distintos indicadores del índice, penalizando así el desequilibrio entre dimensiones.

Existen también medidas de posición que van más allá de la posición central. Los **cuartiles** son la más relevante. La idea es análoga a la mediana, pero divide las observaciones ya no en dos, sino en cuatro partes iguales. El segundo cuartil sería igual que la mediana. El primer cuartil tiene por debajo una cuarta parte de las observaciones (se sitúa en n/4). El tercer cuartil tiene por encima una cuarta parte de las observaciones (se sitúa en 3·n/4).

En las puntuaciones ordenadas (5 8 10 10 **11** 11 11 12 **12** **13** 13 14 14 **15** 15 18 18 20), como ya se conoce la mediana, se busca el primer cuartil. Éste dejaría por debajo el 25% de las 18 observaciones, es decir, 4.5, por tanto tomamos la 5ª observación: 11. El tercer cuartil deja por debajo el 75% de las observaciones, es decir, 13.5, por tanto tomamos la 14ª observación: 15.

La información que dan los cuartiles se puede representar gráficamente a través de diagramas de caja, en los que se representan los tres cuartiles sobre un rectángulo, y los valores mínimo y máximo de la variable a la largo del eje.

Si en vez de dividir las observaciones en cuatro partes, se dividen en 5 (ó 10, ó 100), se obtendrían los quintiles (o deciles, o percentiles). Estas medidas son útiles para caracterizar la asimetría y la concentración de las distribuciones.

Figura 15: Diagrama de caja
Fuente: www.arrakis.es/~mcj/estadist/bigote3.gif [12-6-2012]

Actividad de refuerzo 5:

Para el estudio en Logone Occidental, calcula la media, la mediana, la moda y los cuartiles de las tres variables: número de miembros de la familia, nivel de ingresos y departamento.

¿Sirve de algo calcular las tres medidas de posición central o todas dan la misma información? ¿Por qué? ¿Alguna observación más?

Haz este breve test.

Calcula la mortalidad infantil en Logone Occidental a partir de los datos de la tabla anterior.

4.3.4 Medidas de dispersión

Las medidas de posición no dan excesiva información sobre cómo se distribuye la variable. Sirva de ejemplo la figura 13, donde dos distribuciones muy distintas comparten media y mediana. Haría falta, además, alguna medida de la dispersión o variabilidad de los valores observados. Esta información la proporcionan las medidas de dispersión. Las más comunes son el rango, la desviación estándar y la varianza.

Dispersión: variabilidad de las observaciones obtenidas de una muestra respecto a una variable

El **rango** es una medida sencilla e intuitiva, pues es la diferencia entre el mayor valor que toma la variable y el menor. Es una forma fácil de saber cuán dispersos están los datos, aunque no proporciona demasiada información. En una distribución de frecuencias, coincide con la base del gráfico. Volviendo al ejemplo del

examen de estadística (las puntuaciones fueron 18, 13, 12, 14, 11, 8, 12, 15, 5, 20, 18, 14, 15, 11, 10, 10, 11 y 13), el rango sería: Rango = 20 - 5 = 15.

Dado que la media es la medida de posición por excelencia, tiene sentido que haya otros parámetros que midan cuánto se desvían las observaciones respecto a la media: la varianza y la desviación estándar.

La **varianza** mide la distancia existente entre cada observación y la media. Para cada observación, se resta la media al valor observado $(x_i - \bar{x})$ y esa diferencia se eleva al cuadrado: $(x_i - \bar{x})^2$. Después de hacer esa operación para cada observación, se suma todo lo que se ha obtenido: $\Sigma(x_i - \bar{x})^2$. Para terminar se divide por el tamaño de la muestra 'n'. Así, la varianza de la variable X es:

$$S^2 = \Sigma(x_i - \bar{x})^2 / n$$

Mientras mayor es la varianza, mayor es la dispersión.

La **desviación estándar** es simplemente la raíz cuadrada de la varianza, y es la medida de dispersión de uso más generalizado en estadística, sobre todo porque es más conveniente para realizar ciertos cálculos y representaciones. Representa simplemente la distancia media entre los valores de las observaciones y la media de la variable. Cuanto mayor es la desviación estándar, más lejos están las observaciones de la media (imagen superior), y viceversa (imagen inferior). Además, se mide en las mismas unidades que la variable, por lo que es una de las medidas de variabilidad más utilizadas.

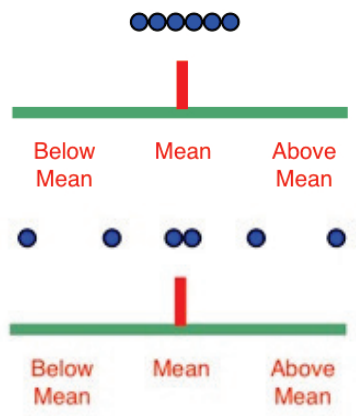


Figura 16: Desviación estándar
Fuente: www.statsmakemecry.com/smmctheblog/ 2010/8/1/top-ten-confusing-stats-terms-explained-in-plain-english-10.html [12-6-2012]

Su fórmula es parecida a la de la varianza:

$$S = \sqrt{S^2} = \sqrt{[\ \Sigma(x_i - \bar{x})^2 / n\ ]}$$

En el caso del examen de estadística (la media era 12,78):

x_i	5	8	10	10	11	11	11	12	12	13	13	14	14	15	15	18	18	20
(x_i-x)	-7,78	-4,78	-2,78	-2,78	-1,78	-1,78	-1,78	-0,78	-0,78	0,22	0,22	1,22	1,22	2,22	2,22	5,22	5,22	7,22
$(x_i-x)^2$	60,5	22,8	7,7	7,7	3,2	3,2	3,2	0,6	0,6	0,0	0,0	1,5	1,5	4,9	4,9	27,2	27,2	52,1

Sumando: $\Sigma(x_i - \bar{x})^2 = 229,1$ y dividiendo entre el tamaño de muestra se obtiene:

Varianza: $S^2 = \Sigma(x_i - \bar{x})^2 / n = 229,1 / 18 = 12,72$

Desviación estándar sería: $S = \sqrt{12,72} = 3,57$ puntos

La interpretación sería que el alumnado está mayoritariamente a 3,57 puntos de 12,78 (la nota media), es decir con una puntuación entre 12,78 y 16,35.

El cálculo de la desviación estándar es más sencillo de lo que parece por la fórmula. En cualquier caso, lo más importante es comprender claramente el concepto, pues es una medida muy utilizada, sobre todo a la hora de describir la distribución de una variable simétrica, como se verá en el siguiente apartado.

En muchas ocasiones, por ejemplo para comparar variables que no están en las mismas unidades o magnitudes, resulta interesante medir la dispersión en forma de porcentaje. Para ello, existe un estadístico derivado de desviación estándar, llamado **coeficiente de variación de Pearson**: $C_v = S / \bar{x}$. Representa la desviación típica en tanto por ciento respecto a la media.

Para el ejemplo anterior, sería: $C_v = S / \bar{x} = 3,57 / 12,78 = 27,93\%$

Actividad de refuerzo 6:

En el estudio en Logone Occidental, toma los datos de Lac Wey y Dodjé y calcula para cada departamento el rango, la desviación estándar y el coeficiente de variación de Pearson de la variable número de miembros de la familia. ¿Qué conclusiones sacas?

¿Por qué son importantes las medidas de dispersión?

Haz este breve test.

4.3.5 Forma de las distribuciones

La desviación estándar junto con la media son muy útiles para definir o describir distribuciones de datos, sobre todo cuando estas distribuciones son **simétricas**, es decir, cuando la curva es igual a un lado y a otro de la media, que está en el centro. Es el caso de la **distribución normal** explicada anteriormente.

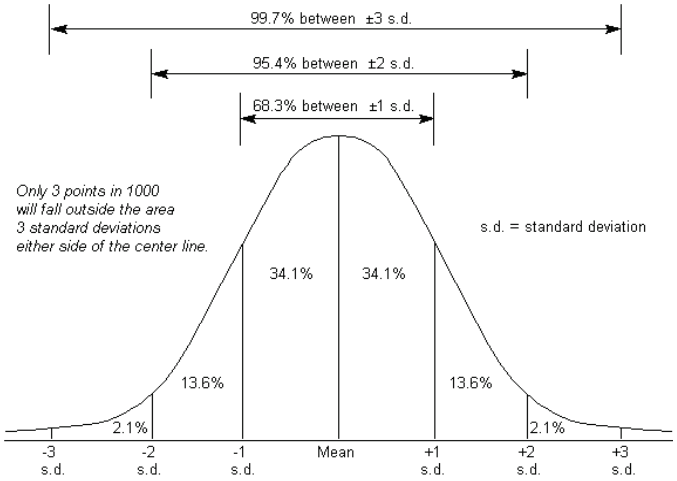


Figura 17: Distribución normal

Fuente: syque.com/improvement/Standard%20Deviation.htm [12-6-2012]

Como se ha visto en el apartado de las medidas de posición (4.3.3), en una distribución normal coinciden la media, la mediana y la moda. Además, la distribución normal se caracteriza porque el 68% de las observaciones están alrededor de la media, situadas en una horquilla de 2 desviaciones estándar (±1), como se puede leer en el gráfico (ver que la desviación estándar se denota con s.d.). El 95,4% de las observaciones tiene valores a ± 2 desviaciones estándar.

La distribución normal se llama así, porque estas características de dispersión ocurren 'normalmente' en muchas variables 'reales'. Por ejemplo, si se mide la estatura de 1000 personas de una edad determinada elegidas al azar, y se representan las observaciones en un gráfico de distribución de frecuencias, la curva que se obtendrá será muy parecida a la campana de la distribución normal. Dicho de otro manera, para una media de 175 cm y una desviación estándar de 5 cm, 680 de las 1000 personas medidas estarían entre 170 y 180 cm.

Esto es algo que se da en la naturaleza en multitud de variables. Suelen ser fenómenos determinados por una combinación compleja de múltiples factores.

También, como se ha dicho anteriormente, hay variables que no siguen la distribución normal. En la figura aparecen algunas de las más comunes. Un ejemplo de distribución asimétrica negativa serían las notas de un examen muy fácil. La distribución **bimodal** se suele dar para opiniones (numerizadas) sobre temas polarizados ideológicamente.

Las **distribuciones asimétricas** son muy importantes en desarrollo. Por ello existen parámetros para caracterizar la asimetría, es decir, para saber si la variable se distribuye de forma simétrica, asimétrica positiva o asimétrica negativa.

Se dice que la asimetría es positiva cuando la mayoría de los datos se encuentran por debajo del valor de la media; que la curva es simétrica cuando se distribuyen aproximadamente la misma cantidad de valores en ambos lados de la media; y que la asimetría es negativa cuando la mayor cantidad de datos se aglomeran en los valores mayores que la media.

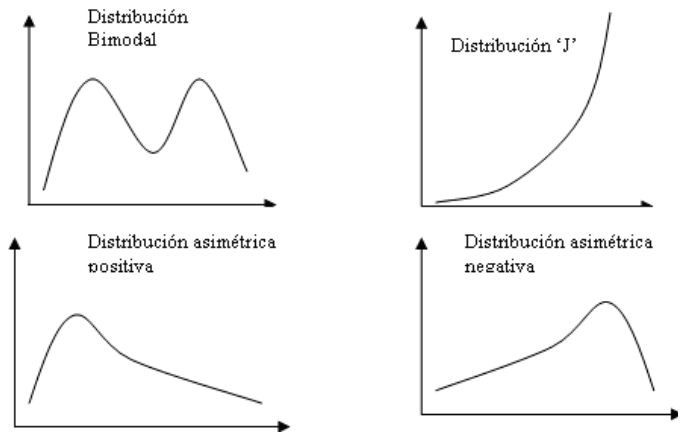


Figura 18: Distintas distribuciones
Fuente: elaboración propia

Un ejemplo de asimetría positiva sería un país con mucha desigualdad, como Nigeria: La renta per cápita media está 'inflada' porque, aunque hay muchas personas pobres, una parte de la población tiene *muuucho* dinero. Por tanto, la mayoría tiene renta per cápita por debajo de la media (asimetría positiva).

El parámetro más utilizado en la medida de la asimetría es el coeficiente de asimetría de Fisher. Aunque no hay que saberse la fórmula, no está de más conocerla:

$$g_1 = \frac{\frac{1}{n}\sum (X_i - \bar{X})^3 * n_i}{\left(\frac{1}{n}\sum (X_i - \bar{X})^2 * n_i\right)^{\frac{3}{2}}}$$

siendo X_i cada uno de los valores, $\bar{x}$ la media de la muestra y n_i la frecuencia de cada valor.

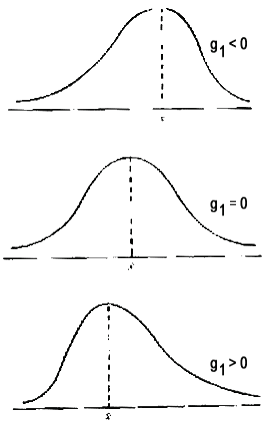


Figura 19: Asimetría de Fisher
Fuente: www.eumed.net/cursecon/dic/oc/asifisher.htm [12-6-2012]

Cuando $g_1 = 0$, la distribución es simétrica. Cuando $g_1 > 0$, la curva es asimétricamente positiva. Cuando $g_1 < 0$, la curva es asimétricamente negativa.

Como es difícil que salga exactamente 0, se considera que la curva es simétrica si g_1 está entre -0,5 y 0,5. Cuanto mayor sea el valor, más asimétrica es la curva. Aunque no es necesario saber calcularlo –las herramientas estadísticas nos pueden ayudar– sí es importante entender el concepto y lo que representa. Su importancia radica en que 'avisa' de si una distribución es muy asimétrica, lo que conllevaría que se calculen unos estadísticos específicos y se elaboren ciertas gráficas. Algunos, como la mediana o los cuartiles, ya se han tratado, pero en el apartado siguiente veremos otros especialmente interesantes, que permiten medir la concentración.

Actividad de refuerzo 7:

Explica, con tus propias palabras, lo que es una distribución normal.

Da ejemplos nuevos de variables que no sigan distribuciones simétricas.

4.3.6 Medidas de concentración

En el ámbito del desarrollo, los temas que se estudian están muy relacionados con desigualdad, desequilibrios y relaciones de poder asimétricas. Esto suele reflejarse en la distribución de las variables que se analizan, que en muchas ocasiones son igualmente asimétricas. Sirvan como ejemplos los ingresos, la propiedad de la tierra, el tiempo de permanencia en la educación formal, el gasto en servicios médicos, etc.

Por ello, se deben tener especialmente en cuenta los parámetros que ayudan a caracterizar la concentración. Ya hemos visto que la media y la desviación estándar describen bien las distribuciones simétricas, pero no son suficientes para las asimétricas. Para estas últimas, la mediana o los cuartiles son medidas útiles, pero se verá a continuación una gráfica y un índice que permiten caracterizar la concentración de la distribución de manera muy precisa. La concentración de una distribución permite saber si los valores de la variable están más o menos uniformemente repartidos a lo largo de la muestra o, en otras palabras, saber cuán equitativamente está repartida una variable. Hay varias herramientas:

La **curtosis**, mide si los valores de la distribución están más o menos concentrados alrededor de la media. Así, una distribución normal sería mesocúrtica; una distribución con los valores muy concentrados alrededor de la media sería leptocúrtica o apuntada; y una con los valores poco concentrados (mayor desigualdad), platicúrtica o chata.

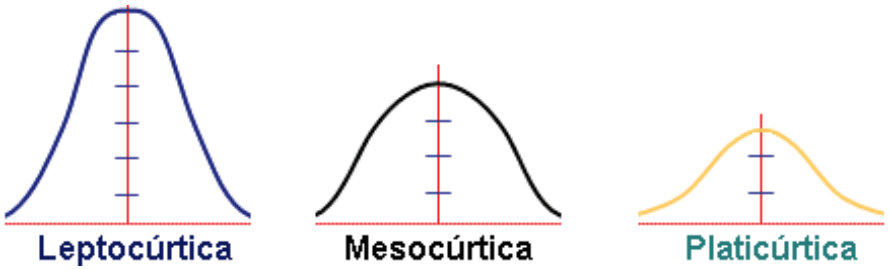


Figura 20: Curtosis

Fuente: 1.bp.blogspot.com/_2JobiRSc5DA/ShbHdqC2RhI/AAAAAAAAABo/gASGFOllcNA/s320/curtosis.gif [12-6-2012]

El coeficiente de curtosis viene definido por la siguiente fórmula:

$$g_2 = \frac{\frac{1}{n}\sum (X_i - \overline{X})^4 * n_i}{\left(\frac{1}{n}\sum (X_i - \overline{X})^2 * n_i\right)^2} - 3$$

Si $g_2 > 0$, distribución apuntada. Si $g_2 < 0$, distribución chata.

Gráficamente, la **curva de Lorenz** es la herramienta más utilizada para representar la concentración, plasmando la distribución relativa de una variable (por ejemplo los ingresos) en una región. Se elabora a partir de los percentiles de la variable (los cuartiles o, más habitualmente, los quintiles). Esta curva ayuda a responder rápidamente a preguntas como: ¿De qué porcentaje del gasto educativo del país se beneficia el 20% más pobre de la población? ¿Y el 20% más rico?

En el eje horizontal de la curva está el porcentaje acumulado de personas u hogares de la región estudiada. Se podría decir que en el eje horizontal se ordena la población de menos a más, en cuanto al valor de la variable de interés. Para, por ejemplo, los ingresos, ordenaríamos las observaciones del más pobre al más rico. El eje vertical refleja el porcentaje acumulado de la variable, es decir, cuánta riqueza total tiene la población hasta ese punto.

Véase cómo se debe leer la curva con un ejemplo: La curva roja es la curva de Lorenz, y representa la distribución de ingresos en un país. El punto marcado en ella está aproximadamente en el 0,7 del eje horizontal y en el 0,4 del eje vertical. Eso significa que el 70% más pobre del país detenta el 40% de los ingresos. Tomando otros puntos se puede ver que el 40% detenta menos del 20% de los ingresos, o que el 80% detenta poco más del 50% de los ingresos. O lo que es lo mismo, que el 20% más rico del país detenta casi la mitad de los ingresos.

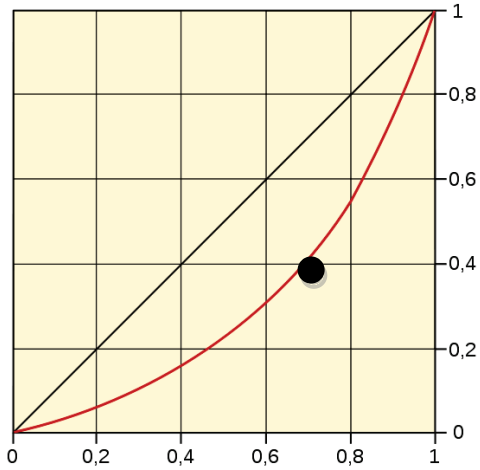


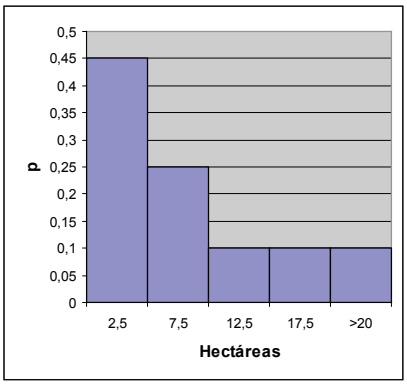
Figura 21: Curtosis
Fuente: 1.bp.blogspot.com/_2JobiRSc5DA/ShbHdqC2RhI/AAAAAAAAABo/gASGFOllcNA/s320/curtosis.gif [12-6-2012]

Si el ingreso estuviera distribuido de manera perfectamente equitativa, la curva coincidiría con la diagonal de la igualdad (línea negra que aparece en el gráfico como referencia). Cuánto más cerca esté la curva de Lorenz de la línea diagonal, menor es la desigualdad, y viceversa. Esto es muy útil para analizar varias curvas a la vez, viendo así la evolución a lo largo del tiempo de la distribución de los ingresos en un país, o comparando la distribución de ingresos entre varios países (para lo que se suele usar una variante llamada Curva de Lorenz Generalizada).

Para construir la curva a partir de una muestra, se ordenan todos los datos de menor a mayor. Se calculan los quintiles, con el fin de agrupar los datos en grupos que representen al 20% de la población. Para cada quintil, se suman todos los valores inferiores a él. Los quintiles marcan 5 posiciones en el eje horizontal (del 20% de la población cada una). Las sumas obtenidas se representan en vertical (expresadas en porcentaje), cada una sobre su quintil. Así se obtienen 5 puntos, que se unen mediante una línea, que representa la curva de Lorenz. Para mayor definición, se pueden usar deciles.

Póngase como ejemplo el estudio sobre la propiedad de la tierra en una comunidad con 20 familias. Se conocen las hectáreas que posee cada familia: 3 7 17 4 25 1 2 4 6 8 18 2 7 9 10 13 1 26 2 4

Se ha elaborado un histograma (a la derecha), comprobándose la asimetría positiva de la distribución.



Para visualizar la concentración de la propiedad de la tierra, se decide construir la curva de Lorenz.

Se ordenan los datos y se agrupan por quintiles en 5 grupos:

[1 1 2 2] [2 3 4 4] [4 6 7 7] [8 9 10 13] [17 18 25 26]

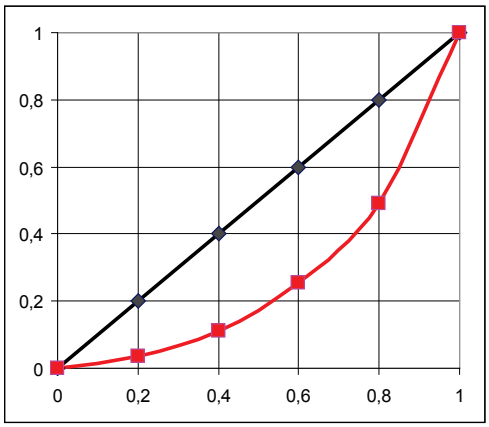
Se suman (acumulativamente) los valores para cada quintil:

[6] [19] [43] [83] [169]

Dividimos entre 169 para obtener los porcentajes:

0 0,04 0,11 0,25 0,49 1

Se representan los puntos (con los porcentajes en el eje vertical) y se traza la curva de Lorenz (en rojo):



El **índice de Gini** es un parámetro muy vinculado con la Curva de Lorenz. Mide cuánto se desvía la distribución real de recursos entre una población (la curva de Lorenz) de la igualdad total (la diagonal de la igualdad). Este índice es de uso generalizado, especialmente para medir y comparar la desigualdad distributiva de recursos. El Banco Mundial, por ejemplo, lo utiliza anualmente en sus Informes de Desarrollo para medir la desigualdad distributiva del ingreso en los países del mundo.

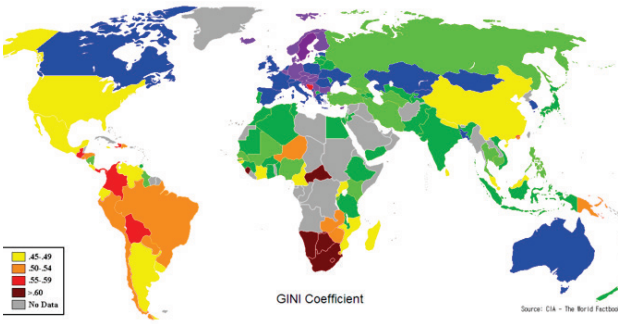


Figura 22: El índice de Gini por países en 2009
Fuente: en.wikipedia.org/wiki/File:GINIretouchedcolors.png [12-6-2012]

Geométricamente, el índice de Gini (IG) representa el área amarilla 'a' entre la curva de Lorenz y la diagonal de igualdad, en porcentaje respecto al área total del triángulo 'b' bajo la diagonal de igualdad. Así, a mayor área entre la Curva de Lorenz y la diagonal de igualdad, mayor desigualdad de distribución y mayor índice de Gini. Aunque es más importante entender el concepto que saber calcularlo, se presenta aquí la fórmula:

$$IG = 1 - (\Sigma q_i / \Sigma p_i)$$

donde p_i mide el porcentaje de observaciones de la muestra que presentan un valor igual o inferior a X_i:

$$p_i = (n_1 + n_2 + n_3 + ... + n_i)\cdot 100 / n$$

Mientras que q_i se calcula así:

$$q_i = 100\cdot[(X_1*n_1) + (X_2*n_2) + ... + (X_i*n_i)] / [(X_1*n_1) + (X_2*n_2) + ... + (X_n*n_n)]$$

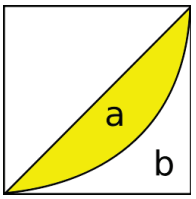


Figura 23: Curva de Lorenz
Fuente: Elaboración propia

El índice de Gini puede tomar valores entre 0 y 1. A mayor desigualdad de distribución, más cerca estará de 1. A menor desigualdad de distribución, más cerca de 0.

Actividad de refuerzo 8:

Lee este caso y propón qué herramienta estadística usarías para resumir bien las diferencias entre Brasil y Eslovaquia.

Dibuja la curva de Lorenz de los ingresos familiares en Logone Occidental. ¿Qué observas?

Lee este ejemplo práctico (a mitad de la página) de cálculo del índice de Gini y entiende cómo lo hace a partir de una tabla de frecuencias.

4.4 Análisis bidimensional

4.4.1 Frecuencias cruzadas

Hasta aquí se han visto distintos aspectos del análisis estadístico unidimensional. Sin embargo, en numerosas ocasiones interesa estudiar simultáneamente dos variables de una población. En tal caso, se habla de análisis bidimensional. Este análisis es muy útil para saber si existe o no una relación entre dos variables. Nos puede permitir comprobar, por ejemplo, si el nivel educativo está relacionado con la esperanza de vida. La clasificación de variables como dependientes o independientes es de utilidad para el análisis bidimensional. Si queremos saber cómo explicar los valores observados en una variable (la variables dependiente), podemos ir viendo si está relacionada con otras variables (variables independientes) y ver cuál la 'explica'.

Tablas de contingencia: tabla donde en cada casilla figura frecuencia de observación simultánea de determinados valores de dos variables

Según el tipo de variable se suelen utilizar tablas de contingencia (cualitativas) o diagramas de dispersión, coeficientes de correlación, etc. (cuantitativas). Vemos a continuación algunas de estas herramientas de análisis bidimensional.

La **tabla de contingencia** o tabla de frecuencias cruzadas recoge la frecuencia con la que se observa la combinación de valores posibles de las dos variables (X e Y).

X \ Y	y_1	y_2	...	y_m
x_1	$n_{1,1}$	$n_{1,2}$		$n_{1,m}$
x_2	$n_{2,1}$	$n_{2,2}$		$n_{2,m}$
...				
x_n	$n_{n,1}$	$n_{n,2}$		$n_{n,m}$

Al igual que en las tablas de frecuencia, las x_i representan los valores que va tomando la variable X; las y_j, las de la variable Y. En cada celda se pone el número de sujetos que tienen a la vez el valor x_i de su fila y el y_j de su columna, es decir, la frecuencia de la combinación de dichos valores.

Esto se verá mucho más claro con un ejemplo. La variable X se refiere al sexo y puede tomar el valor Hombre o Mujer. La variable Y se refiere al nivel educativo y puede tomar los valores: Ninguno, Primaria, Secundaria, Superior. Su tabla de contingencia (para un estudio imaginario en el que se ha tomado una muestra de la población adulta de una ciudad) sería:

Nivel educ. \ Sexo	Hombre	Mujer
Ninguno	20	20
Primaria	33	47
Secundaria	52	13
Superior	15	3

Hay 3 mujeres con estudios superiores, 33 hombres con estudios primarios, etc. Para un mejor análisis, es conveniente representar al final una columna y una fila con los totales (se denominan frecuencias marginales).

Nivel educ. \ Sexo	Hombre	Mujer	Total
Ninguno	20	20	40
Primaria	33	47	80
Secundaria	52	13	65
Superior	15	3	18
Total	120	83	203

Como se han obtenido menos datos de mujeres que de hombres, no se puede apreciar bien el nivel de estudios alcanzado según el sexo. Para ello, puede ser interesante añadir en las celdas las frecuencias relativas condicionales de nivel educativo respecto a sexo, esto es, los porcentajes respecto al total de la columna.

Nivel educ. \ Sexo	Hombre		Mujer		Total	
Ninguno	20	*17%*	20	*24%*	40	*20%*
Primaria	33	*28%*	47	*57%*	80	*39%*
Secundaria	52	*43%*	13	*16%*	65	*32%*
Superior	15	*13%*	3	*4%*	18	*9%*
Total	120	*100%*	83	*100%*	203	*100%*

Ahora ya se puede apreciar que la mayoría de mujeres solo alcanza la primaria, mientras los hombres suelen alcanzar la secundaria. Aunque en este caso no resulta de especial interés, también se pueden incluir las frecuencias relativas condicionales de sexo respecto a nivel educativo, esto es, los porcentajes respecto al total de la fila.

Cuando en una tabla comparamos dos variables binarias, existe un estadístico que permite cuantificar la relación entre ambas: el **coeficiente phi** (φ),

$$\varphi = (n_{1,1} \cdot n_{2,2} - n_{1,2} \cdot n_{2,1}) / \sqrt{(n_{1,\cdot}\, n_{2,\cdot} \cdot n_{\cdot,1} \cdot n_{\cdot,2})}$$

Nivel educ. \ Sexo	Hombre	Mujer	Total
Primaria	33	47	80
Secundaria o más	67	16	83
Total	100	63	163

En este caso:

$$\varphi = (33 \cdot 16 - 47 \cdot 67) / \sqrt{(80 \cdot 83 \cdot 100 \cdot 63)} = -2621 / 6467 = -0{,}405$$

Phi puede estar entre -1 y 1. Cuánto más cerca de 1 ó -1, más fuerte es la relación entre las variables. Si es casi 0, se considera que no hay correlación.

En las tablas de contingencia se pueden comparar variables cualitativas entre sí, cualitativas con cuantitativas y cuantitativas entre sí. Hay que recordar que las variables cuantitativas (especialmente las continuas) se deben agrupar antes. Sin embargo, hay otras formas de análisis bidimensional más adecuadas a variables continuas.

Una de ellas es el **diagrama de dispersión**. Es un diagrama que simplemente representa en el eje horizontal una variable y en el vertical, otra. Las distintas observaciones se van marcando con puntos en la intersección correspondiente de ambos valores, como se puede ver en el gráfico izquierdo, realizado con los datos de las variables ingresos y número de miembros de la familia del estudio de Logone Occidental. Se suelen utilizar solo variables cuantitativas, aunque también es posible utilizar cualitativas. Además, se podrían usar distintos tipos de puntos para comparar una tercera variable, por ejemplo, cambiando de color según el departamento. El GapMinder mencionado en el apartado 3.1.1 es una versión avanzada de gráfico de dispersión, con dos variables extras y animación a lo largo del tiempo.

Diagrama de dispersión: representación gráfica de todos los valores de dos variables en forma nube de puntos.

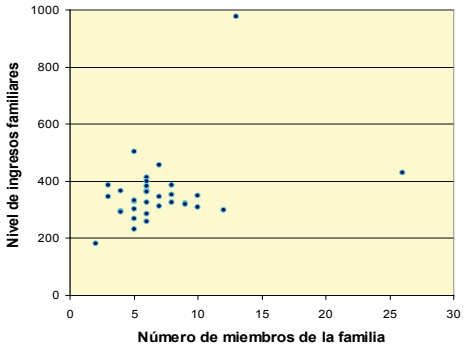


Figura 24: Diagrama de dispersión del ingreso y tamaño familiar en Logone
Fuente: elaboración propia

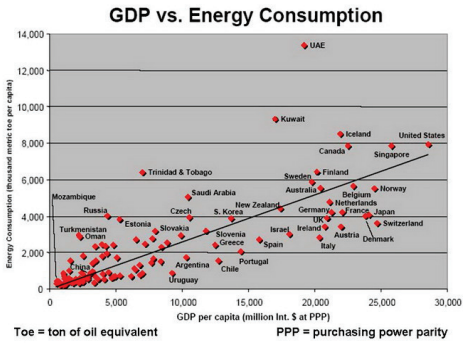


Figura 25: Diagrama de dispersión del consumo de energía y el PIB per cápita
Fuente: www.greenparty.ca/sites/greenparty.ca/files/Energy_Consumption_vs_GDP_655.jpg [12-6-2012]

4.4.2 Correlaciones

Cuando la nube de puntos formada por los datos en el diagrama de dispersión se agrupa alrededor de una línea (que no es el caso para el estudio de Logone), quiere decir que hay una relación entre una variable y otra. Es decir, que si una crece, la otra crece (o decrece). Esta relación suele ser lineal, aunque en ocasiones puede ser parabólica, hiperbólica o exponencial.

Un ejemplo de relación lineal se da entre las variables PIB per cápita y consumo energético nacionales. Como se ve en el gráfico anterior derecho, a medida que crece el PIB (eje horizontal), el consumo de energía crece proporcionalmente (eje vertical).

Modelizar dicha relación puede ser útil para realizar comparaciones y predicciones, ya que se establece una regla de relación entre ambas variables. Por ejemplo, si se cree que el PIB chino se va a duplicar en 10 años, cabe esperar que la demanda energética aumente en un 60%. Esto es muy útil para la planificación, por ejemplo ver cómo satisfacer dicha demanda… o cómo evitar que aumente tanto.

Un estadístico muy utilizado es el **coeficiente de correlación lineal de Pearson**, que sirve para cuantificar el grado de relación lineal entre las dos variables. Es análogo al coeficiente phi, pero para variables continuas. Se debe tener en cuenta que este coeficiente solo es válido si la relación entre las variables es lineal. Por ello, es interesante representar primero los datos en un diagrama de dispersión para comprobar si se agrupan más o menos respecto a una recta.

Coeficiente de correlación lineal: mide el grado de intensidad de la relación entre dos variables.

Para calcularlo, nos hace falta refrescar el cálculo de la desviación estándar y aprender el de la covarianza (S_{xy}), que se parece bastante al de la varianza (ver apartado 4.3.4 sobre medidas de dispersión). Consiste en multiplicar para cada observación 'i', la diferencia entre el valor y su media de una variable ($x_i - \overline{x}$) por la de la otra ($y_i - \overline{y}$). Se suman los productos de las diferentes observaciones y se dividen entre el tamaño de la muestra 'n':

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Finalmente, el coeficiente de correlación lineal se calcula dividiendo la covarianza entre las desviaciones estándar de las dos variables:

$$r = S_{xy} / (S_x \cdot S_y)$$

El coeficiente de correlación puede estar entre -1 y 1. Si es mayor que 0, se trata de una correlación lineal positiva (si aumenta una variable, también la otra) y, si es menor que 0, de una correlación lineal negativa. Cuánto más cerca de 1 ó -1, más fuerte es la correlación. Si es 0 ó casi 0, se considera que no hay correlación.

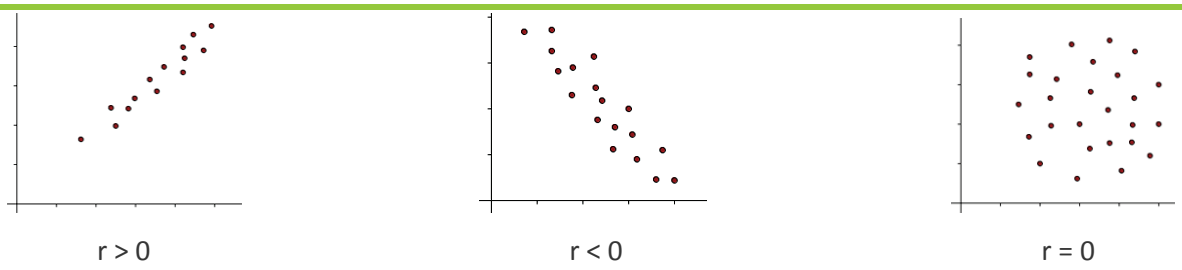
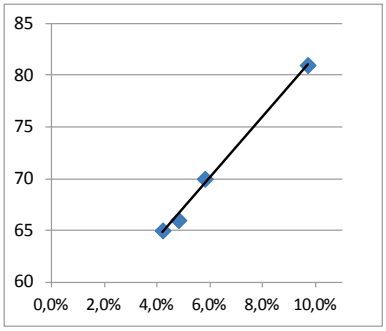


Figura 26: Coeficiente de correlación y diagramas de dispersión
Fuente: elaboración propia

Mejor veamos estos cálculos en un ejemplo en el que se estudia la relación entre porcentaje del PIB dedicado a sanidad y la esperanza de vida de cuatro países. No tiene sentido realizar análisis bidimensional con tan pocos sujetos, pero aquí se pretende solo ilustrar el cálculo.

Los valores para el año 2009 fueron los siguientes.

	% PIB salud (x)	Esperanza de vida (y)
Bolivia	4,8%	66
España	9,7%	81
India	4,2%	65
Belarus	5,8%	70



Se realiza un diagrama de dispersión primeramente, observándose cierta relación, a más % PIB dedicado a salud, mayor esperanza de vida.

Procedemos pues a calcular el coeficiente de correlación lineal para cuantificar dicha relación.

Para la variable x (% PIB dedicado a salud), calculamos, la media $\bar{x}$ es 6,1% y la desviación estándar:

$$S_x = \sqrt{(\Sigma(x_i - \bar{x})^2 / n)} = \sqrt{([(4,8-6,1)^2 + (9,7-6,1)^2 + (4,2-6,1)^2 + (5,8-6,1)^2] / 4)} =$$

$$\sqrt{([1,69+12,96+3,61+0,09] / 4)} = 2,142$$

De forma análoga para y (esperanza de vida), la media $\bar{y}$ es 70,5 y la desviación estándar

$$S_y = \sqrt{(\Sigma(y_i - \bar{y})^2 / n)} = \sqrt{([(66-70,5)^2 + (81-70,5)^2 + (65-70,5)^2 + (70-70,5)^2] / 4)} = 6,344$$

Y la covarianza:

$$S_{xy} = \Sigma [(x_i - \bar{x})\cdot(y_i - \bar{y})] / n =$$

$$= [(4,8-6,1)\cdot(66-70,5) + (9,7-6,1)\cdot(81-70,5) + (4,2-6,1)\cdot(65-70,5) + (5,8-6,1)\cdot(70-70,5)] / 4 =$$

$$= (5,85 + 37,8 + 10,45 + 0,15) / 4 = 13,56$$

Finalmente podemos calcular el coeficiente de correlación lineal:

$$r = S_{xy} / (S_x \cdot S_y) = 13,56 / (6,34 \cdot 2,14) = 0,998$$

Se trata pues de una correlación muy fuerte. Es decir, que mirando solo esos 4 países, parece hay una correlación muy fuerte entre las variables de porcentaje de gasto del PIB en salud y esperanza de vida.

A veces también es interesante calcular la **recta de regresión**, que es la recta alrededor de la cual se agrupan los puntos. Sirve para predecir el comportamiento de la variable dependiente a partir del comportamiento de la variable independiente.

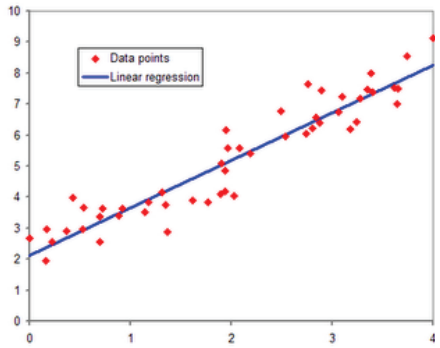


Figura 27: Recta de regresión
Fuente: metodologiaeninvestigacion.blogspot.com.es/2010/08/estadistica-inferencial-correlacion-y.html [12-6-2012]

Para quien quiera probar a realizar el cálculo, la recta de regresión se formula como $y = m \cdot x + b$

Es decir: Variable dependiente y = pendiente m · Variable independiente x + altura b

Las n observaciones nos dan pares de valores: (x_1,y_1) (x_2,y_2) … (x_n,y_n)

La pendiente se calcula como

$$m = [\, n \cdot \Sigma (x_i \cdot y_i) - \Sigma x_i \cdot \Sigma y_i \,] / [\, n \cdot \Sigma x_i^2 - (\Sigma x_i)^2 \,]$$

La altura se calcula como

$$b = (\Sigma y_i - m \cdot \Sigma x_i) / n$$

Existen otros coeficientes distintos para medir la correlación lineal. También se dan en muchas ocasiones relaciones no lineales, como las hiperbólicas, que dan lugar a curvas de regresión no lineales como la de la imagen. Aunque quedan fuera del alcance de este capítulo, es bueno ser conscientes de la diversidad de relaciones que se pueden dar.

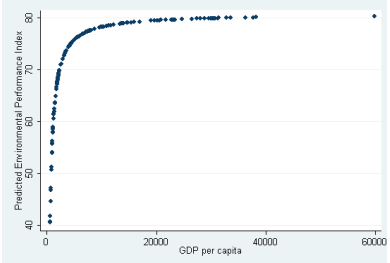


Figura 28: Relación hiperbólica
Fuente: sites.google.com/site/depken/gdp-vs-epi.png [12-6-2012]

Un ejemplo de relación hiperbólica se da entre la calidad medioambiental y PIB per cápita (a nivel país). Los países con bajos ingresos suelen tener una calidad ambiental baja. A medida que aumenta el PIB, mejora la calidad ambiental. Pero a partir de cierto momento, el aumento del PIB ya no conlleva un aumento sustancial de la calidad ambiental.

Por otro lado, existen también formas de análisis ya no bidimensional, sino multidimensional. Como es natural, hay variables que se 'explican' a partir de dos o más variables. El análisis multidimensional permite tener esto en cuenta y ver cuánto influye cada variable independiente en la variable de interés. Una de las técnicas más empleadas es el análisis de varianzas (ANOVA). Estas herramientas permiten calcular coeficientes y curvas de regresión múltiple, que quedan fuera del alcance de este capítulo.

Pero hablando ahora de la correlación en general, ¿cuál es el **significado** de la correlación en el mundo real? Recordemos que los coeficientes de correlación nos permiten saber la intensidad y la 'dirección' de la relación entre dos variables. Sin embargo, deben quedar muy claras dos cuestiones:

En primer lugar, que el coeficiente de correlación calculado a partir de una muestra es un estadístico muestral y como tal sólo es válido para la muestra. Con la inferencia estadística (contraste hipótesis), podríamos ver si dicha correlación es significativa a nivel de población. En el ejemplo anterior, aun suponiendo que la muestra fuese aleatoria, al ser tan pequeña, veríamos que no se puede generalizar esa relación entre %PIB y esperanza de vida a todos los países del mundo.

En segundo lugar, que cuando un coeficiente de correlación es alto podemos estar seguros de que las variables están relacionadas… pero no podemos saber si la relación es causal o no. De hecho, en el caso bidimensional, podrían estar pasando 3 cosas, a saber:

(a) que exista una relación causal unidireccional, es decir, una variable causa la otra, pero no viceversa. Por ejemplo, ingresos altos causan mayores gastos en actividades de ocio.

(b) que exista una relación causal bidireccional, es decir, las variables sean causa una de la otra. Por ejemplo, una alta inversión en I+D causa productividad alta, y también viceversa.

(c) que no exista una relación causal alguna. Ocurre cuando ambas variables correlacionadas son causadas por una tercera variable, que es la causa real de las dos. Por ejemplo, se podría detectar una correlación entre gastos en actividades de ocio y gastos en muebles para el hogar. ¿Habrá ahí relación causa efecto? A no ser que alguien se compre una mesa nueva para poner el parchís recién estrenado, sería absurdo pensar que un mayor gasto en actividades de ocio sea causa de un mayor gasto en muebles, o viceversa. Lo más probable es que los ingresos sean esa tercera variable no tenida en cuenta, y que mayores ingresos causen simultáneamente mayores gastos en ocio y en muebles.

Por tanto, detectar una correlación sólo permite sospechar que hay causalidad. Así, el sentido común y unos buenos marcos teóricos serán las únicas herramientas para valorar la causalidad.

Actividad de refuerzo 9:

Para el caso de Logone Occidental, haz una tabla de contingencia para las variables Departamento e Ingresos familiares. Agrupa los ingresos según consideres conveniente. Calcula las frecuencias relativas condicionales por columnas o filas, según creas que vaya a permitir un análisis más rico. ¿Qué conclusiones sacas?

Calcula el coeficiente de correlación lineal entre nivel de ingresos y número de miembros de la familia. ¿Qué te dice el resultado?

Capítulo 5. Inferencia estadística

Objetivos del capítulo:	Tiempo estimado de lectura: 150 min
• Seleccionar muestras, utilizando métodos y tamaños adecuados a los objetivos de inferencia planteados • Analizar (calcular estadísticos, realizar estimaciones y presentar gráficamente) los datos disponibles, utilizando herramientas informáticas de manera consciente	Apartados del capítulo: 5.1 La inferencia estadística 5.2 Estimación por intervalos de confianza 5.3 Contraste de hipótesis
Capítulo anterior. Estadística descriptiva	Índice

5.1 La inferencia estadística

5.1.1 El papel de la inferencia

Como se ha visto, la estadística descriptiva sirve para analizar los datos que obtenemos de una muestra. Los estadísticos muestrales obtenidos describen la muestra, pero nuestro objetivo es conocer a la población, no a la muestra. Queremos generalizar el análisis a toda la población, es decir, pasar de los estadísticos muestrales a parámetros poblacionales estimados. Pero, ¿cómo sabemos que la muestra es realmente representativa de toda la población?

Desde una perspectiva positivista, la respuesta sería que el muestreo debe ser aleatorio y no sesgado para cumplir con la generalizabilidad, y el tamaño de muestra suficientemente grande para garantizar la fiabilidad (ver apartado 1.3.2). La inferencia estadística es la herramienta estadística que valida dicha fiabilidad.

Pero repasemos primero los conceptos de sesgo y error aleatorio, vistos en el apartado 2.3.5.

El sesgo es una distorsión generada en el momento del muestreo por la forma de realizarlo o por el marco muestral. Así para que una muestra pueda ser representativa, tiene que ser no sesgada. Es decir, el marco muestral deberá ser completo (y actualizado) y la muestra deberá seleccionarse aleatoriamente. Muy importante: si el muestreo no es aleatorio, no se cumplen las condiciones para aplicar los cálculos inferenciales correctamente.

El error aleatorio es la imprecisión natural e inevitable de los estimadores muestrales, ya que siempre habrá alguna diferencia entre la muestra y la población. La gracia de la inferencia estadística es que nos permite cuantificar esa imprecisión. El error aleatorio depende –entre otras cosas– del tamaño de muestra. Si tomamos una muestra pequeña, nuestra precisión será baja, y viceversa.

La inferencia se utiliza, por tanto, para dimensionar el muestreo de manera que se minimice ese error aleatorio, aumentando la precisión de nuestro análisis. De hecho, la inferencia se utiliza ya desde la fase de diseño de la investigación, cuando realizamos el muestreo (ver apartado 2.3). Según la precisión requerida, se tomará un tamaño de muestra suficiente.

La razón de que se explique en el último capítulo es doble. Primero por una cuestión práctica; los conceptos vistos hasta aquí son necesarios para la comprensión de la inferencia. Segundo porque una vez reco-

lectada y analizada la información, vamos a recalcular la precisión con la que nuestros análisis son válidos para toda la población.

A modo de síntesis, es importante recordar que la inferencia relaciona la precisión con el tamaño de muestra, con lo que se usará tanto para calcular el tamaño de muestra como para determinar la precisión de las generalizaciones que hagamos.

5.1.2 Conceptos básicos

La estadística inferencial calcula la precisión con la que la muestra refleja ciertas características de la población. Esto se emplea principalmente para realizar dos cosas: estimaciones y contrastes de hipótesis. Vayamos por partes.

Las estimaciones sirven para responder a preguntas como: ¿hasta qué punto puedo considerar que la media muestral que he obtenido del ingreso familiar medio (por ejemplo 368500 CFA) es válida para toda la población? La respuesta que daría la inferencia es: La media poblacional estará entre 368000 y 369000, con una confianza del 95%. Es lo que se llama la **estimación por intervalos de confianza**, que es la que trataremos aquí. Consiste en dar un intervalo de confianza u horquilla alrededor del estadístico muestral (proporción, media, varianza, etc.) en el que se puede afirmar que estará el parámetro muestral, con una probabilidad de acertar determinada (el nivel de confianza). Existe también la estimación puntual, pero no se usa tanto.

Intervalos de confianza: par de números entre los cuales se estima que –con una determinada probabilidad de acierto– estará un parámetro poblacional.

Contraste de hipótesis: procedimiento para comprobar si una propiedad, que suponemos cumple una población, es coherente con lo observado en la muestra.

Los **contrastes de hipótesis** son un poco más complejos y variados. Sirven para contestar –entre muchas otras– a preguntas del tipo: ¿Puedo afirmar que la renta media de las familias de Dodjé es menor que la de las familias de Ngourkosso, tal como reflejan las medias muestrales? ¿La correlación que he observado en mi muestra entre % PIB dedicado a sanidad y esperanza de vida es significativa a nivel de la población? Debido a su complejidad, no se va a entrar en mucho detalle en los contrastes de hipótesis en este capítulo.

¡Un momento! ¿Esto de la inferencia no era para calcular el tamaño de la muestra? ¿Por qué tanto rollo? ¿Dónde está la fórmula?

Sí, la inferencia sirve para calcular el tamaño de muestra… pero para responder a la pregunta ¿qué tamaño de muestra necesito?, hace falta responder a dos preguntas previas y conocer algunos detalles.

La primera pregunta previa es: **¿para qué?**

¿Qué estás investigando? ¿Qué es lo principal que quieres saber (1) o demostrar?

Según busques averiguar una media, una proporción o una varianza, comprobar una correlación o comparar dos medias, tendrás que utilizar una fórmula diferente.

La segunda pregunta previa es: **¿con qué precisión?**

¿Cuánta precisión necesitas? Dicha precisión es inversamente proporcional al error aleatorio (2). Por tanto depende –entre otros factores– del tamaño de la muestra. En consecuencia, es requisito previo determinar un nivel aproximado de precisión para poder dimensionar la muestra. En el caso de la media de ingresos, esto se concretaría en decidir si el intervalo de confianza queremos que sea de ±500 CFA o ±3000 CFA.

Vayamos con los detalles. Necesitamos saber:

La **variabilidad** (3) de la población respecto a la variable de interés. Para una población es muy diversa, necesitaremos un tamaño de muestra grande si queremos hacernos una idea general de la misma. Si es homogénea, con menos muestra será suficiente. Una analogía: si quiero describir una pared blanca, le puedo hacer una foto con poca calidad, con unos pocos píxeles me sobra para tener una idea adecuada de la pared. Si se trata del Guernica, con pocos píxeles no consigo apreciar la riqueza del cuadro.

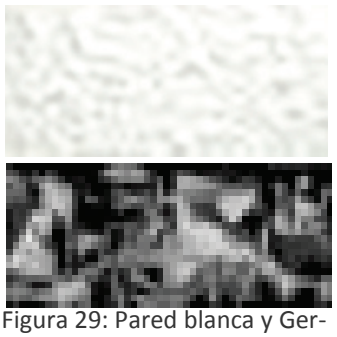


Figura 29: Pared blanca y Gernika en 1 KB
Fuente: elaboración propia

El **tipo de muestreo** (4) también influye. En muestreos estratificados y por etapas, las muestras presentan variabilidad distinta a la de aleatorios simples y sistemáticos, y esto debe ser tenido en cuenta.

El **tamaño población** (5). A medida que la población aumenta, también debe hacerlo el de la muestra. Sin embargo, para poblaciones muy grandes, el tamaño de muestra ya no se ve afectado por el tamaño de la población

¡Un momento! Acepto –a regañadientes– que para calcular el tamaño de muestra necesite (2) el error aleatorio deseado, saber (3) la variación en la población y (5) el tamaño de la población. Pero, ¿en serio me estás diciendo que hay tropecientas formas de calcular el tamaño de muestra según (1) qué quiero estimar o comprobar, (4) qué tipo de muestreo aleatorio he usado y (5) el tamaño de la población?

Pues sí, así es. Pero que no cunda el pánico. Vamos a dedicarnos aquí principalmente a los cálculos del tamaño de muestra para (1) estimar una media y estimar una proporción (según la variable sea cuantitativa o cualitativa). Asumiremos (4) muestreo aleatorio simple y (5) tamaño de población grande. De esta manera, el grueso se va a centrar en solo dos cálculos del tamaño de muestra, aunque daremos pequeñas pinceladas sobre qué hacer si no se cumplen las asunciones (4) y (5), y si queremos (1) estimar o comprobar otras cosas.

Es muy importante entender la lógica detrás del cálculo del tamaño de muestra. Es una cuestión de ideas más que de matemáticas, para las que tendremos la ayuda de programas informáticos… que solo nos será útil si comprendemos lo que están haciendo.

5.1.3 Inferencia y rigor

Hemos visto que el muestreo aleatorio e insesgado proporciona generalizabilidad y la estadística inferencial demuestra la fiabilidad (si el tamaño de muestra es suficiente). Éstos son dos de los criterios que hacen rigurosa una investigación cuantitativa, desde una perspectiva positivista.

¿Qué ocurre entonces cuando no podemos acceder a marcos muestrales de calidad o los recursos son insuficientes para grandes tamaños de muestra? ¿Al no poder inferir, ya no sería válida la información cuantitativa recolectada? Muchos investigadores opinan que sí. En estudios en desarrollo, donde los problemas de recursos y de marcos muestrales suelen ser más habituales, esto tiene como consecuencia que se renuncian a usar técnicas cuantitativas y análisis estadístico, empleándose únicamente técnicas cualitativas. Se pierde así una interesante oportunidad de combinar cualitativa y cuantitativa, y se deja de lado todo el potencial que la estadística (aún sin inferencia) ofrece a nivel de síntesis y visualización de información.

Sin embargo, desde una posición epistemológica realista (ver apartado 1.1.2), los criterios de rigor serían diferentes y las técnicas cuantitativas tienen mucho que aportar, incluso si no se cumplen las condiciones –siempre deseables– para aplicar la inferencia estadística.

Así, estudios con muestreos no estrictamente aleatorios o con muestras relativamente pequeñas, pueden aportar igualmente información relevante sobre la situación estudiada, si el proceso de investigación es sistemático, honesto y transparente. Esto se concreta en reducir la arbitrariedad en la selección de la muestra todo lo que se pueda, utilizando métodos perfectamente descritos. Los sesgos que no se hayan podido evitar, se deben reconocer explícitamente y, en las conclusiones, habría que dejar claro a quién se refieren dichas conclusiones.

Por ejemplo, si se hace una encuesta en una comunidad coincidiendo con la época de cosecha y muchas familias han emigrado (emigración estacional), éstas quedan fuera de la muestra, lo que conlleva un sesgo considerable. Los resultados serán, por tanto, válidos para miembros de la comunidad con una situación económica más desahogada, y esto debe ser reconocido claramente en los informes.

A modo de síntesis, se podría decir que la inferencia otorga una justificada garantía estadística de calidad a las investigaciones cuantitativas. Pero esto no implica que la cuantitativa sin inferencia carezca de valor.

5.2 Estimación por intervalos de confianza

A partir de una muestra se obtienen estadísticos muestrales, siendo los más destacados las medias y las proporciones. Debido al error aleatorio del que se ha hablado anteriormente, ese estadístico muestral no se corresponderá exactamente con el parámetro muestral. Por ello, se suelen utilizar los **estimadores por intervalos de confianza**, que se tratan a continuación, que son la base del cálculo de tamaño de muestra que veremos seguidamente.

La estadística inferencial permite cuantificar el error aleatorio al generalizar información de una muestra aleatoria a toda la población. El error aleatorio, se hace operativo en dos conceptos: el **intervalo de confianza** y el **nivel de confianza** (que se denota como 1-α). Estos conceptos están relacionados con la **variabilidad** de la variable y con el **tamaño de la muestra**. Así, cuando se estima un **parámetro**, éste se acompaña tanto del intervalo de confianza -que es como una horquilla o margen de error alrededor del estadístico muestral- como del nivel de confianza -que señala la probabilidad de que el parámetro que se estima esté realmente dentro del intervalo de confianza (revisar apartado 5.1.2).

Nivel de confianza: probabilidad de que el intervalo de confianza contenga al verdadero valor del parámetro poblacional.

Un ejemplo práctico podría ser un estudio sobre nutrición infantil en una región determinada. La población serían todos los niños y niñas de 9 y 10 años de dicha región. Entre otras variables, se quiere determinar el peso medio poblacional. Para ello, se realiza un muestreo aleatorio, obteniendo una media muestral de 28 kilos. Mediante los cálculos que se tratarán a continuación, se podría realizar una estimación por intervalos de confianza, estimando un peso medio poblacional de 28 kilos con un intervalo de confianza de ±1kg, y un nivel de confianza del 95%. Esto significa, que a partir de los datos obtenidos mediante la muestra, podemos estar seguros en un 95% de que el peso medio de la población estudiada está entre 27 y 29 kilos.

Tomando una muestra de mayor tamaño, se aumenta la precisión en la generalización, lo que se concreta en una reducción del intervalo de confianza (por ejemplo a 0,5 kilos) y/o un aumento del nivel de confianza (por ejemplo al 99%). El nivel de confianza y el intervalo de confianza dependen, además, de la

variabilidad de la variable y del parámetro que se pretenda estimar. Así, no se calcula de la misma manera el tamaño de muestra para estimar una proporción que para estimar una media o realizar un contraste de hipótesis.

Normalmente se observan varias variables de una muestra. Como el tamaño de muestra no puede cambiar para cada variable que queremos medir, en la práctica se fija un nivel de confianza y un intervalo de confianza para el parámetro a estimar que se considere más importante para el estudio y se calcula el tamaño de la muestra en consecuencia. Después, a partir de ese tamaño de muestra, se pueden calcular los intervalos de confianza y niveles de confianza resultantes para las demás estimaciones. Otras veces, para mantener un nivel e intervalos de confianza aceptables, en lugar de usar el parámetro más importante, se fija el tamaño de muestra según el parámetro que necesita un mayor tamaño de muestra para la confianza establecida. Así, el intervalo de confianza y nivel de confianza asociados se pueden dar por válidos para el resto de estimaciones (que serán iguales o más precisas, dada su menor variabilidad).

La fórmula del tamaño de muestra, para población infinita, muestra de al menos 30 sujetos y muestreo aleatorio simple, sería:

Tamaño de muestra n = nivel de confianza Z_α^2 · variabilidad Θ / intervalo de confianza d^2

$$n = \frac{Z_\alpha^2 \cdot \theta}{d^2}$$

En la fórmula se observa que para reducir el **intervalo de confianza** 'd', hay que aumentar el tamaño de muestra 'n'. Refuerza lo ya dicho, que a mayor tamaño de muestra, más precisa la estimación (y menor intervalo de confianza).

También aumenta el tamaño de muestra con la variabilidad poblacional 'Θ'. La variabilidad se concreta en la varianza (S^2) en el caso de variables cuantitativas y en la proporción multiplicada por su complemento (p·(1-p))en el caso de variables cualitativas. Una limitación frecuente a la hora de calcular el tamaño de muestra, es el no **conocer la variabilidad de la variable a estudiar** (por ejemplo la varianza de los gastos familiares en medicamentos, sobre los que hay pocas estadísticas). Es decir, no se sabe la varianza o proporción de la variable en la población. Como es información necesaria para este cálculo, debería hacerse un estudio previo para obtenerlo, y así poder diseñar el muestreo. Los costes de esto hacen que, para estudios con presupuesto limitado, se aproxime esta información a partir de otros estudios similares o del censo, o se estime de forma conservadora.

Por último, está el **nivel de confianza**, que debemos 'traducir' antes de introducirlo en la fórmula. La traducción nos la da dan las tablas de la distribución normal tipificada, que asocian los Z_α y los niveles de confianza. Los Z_α más relevantes se tabulan a continuación.

Nivel de confianza	99%	98%	97%	96%	95%	94%	93%	92%	91%	90%
Z_α	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695	1,645

Por convenio, se suelen tomar niveles de confianza del 95% o 99%.

Para quien quiera saber de dónde sale esto, cabe indicar que la tipificación de la distribución normal parte de un teorema llamado teorema central del límite. Se asume que aunque las variables presenten distribuciones asimétricas o binomiales, se puede utilizar la tipificación de la distribución normal para calcular el tamaño de muestra, siempre que ésta sea mayor que 30. En todo caso, la teoría estadística subyacente supera el alcance del presente capítulo.

Como hemos dicho anteriormente, la misma fórmula se utilizaría en la fase final del estudio para calcular el intervalo de confianza de las estimaciones de las distintas variables relevantes. Despejada queda como:

$$d = Z_{\alpha}\sqrt{\frac{\theta}{n}}$$

Se presentan en los siguientes apartados, los cálculos para estimar una proporción y una media, dejándose fuera el de la varianza.

5.2.1 Tamaño de muestra para estimar una proporción

Para calcular el tamaño de muestra cuando se quiere estimar una proporción de una variable cualitativa, se necesita:

El valor poblacional de la proporción (p) de la variable que se quiere medir, expresado en tanto por 1.

El intervalo de confianza o margen de error de la variable estudiada (d), expresado en porcentaje.

El nivel de confianza (1-α). Suele tomarse el 90%, el 95% o el 99%. A cada nivel de confianza, le corresponde un Z_{α} (que se consulta en la tabla).

Así, se calcula el tamaño de muestra (n) mediante la siguiente fórmula:

$$n = \frac{Z_{\alpha}^2 \cdot p \cdot (1-p)}{d^2}$$

El valor de la proporción, al no conocerse, se puede estimar a partir de otros estudios similares o en un estudio previo. Si esto tampoco es posible, se tomará el escenario más conservador, es decir, el que más variabilidad presenta. Será cuando p·(1-p) sea máximo, cosa que ocurre para p=0,5:

$p \cdot (1-p)_{max} = 0{,}5 \cdot (1-0{,}5) = 0{,}25$ (prueba con otras 'p' si no te fías).

Se toma como ejemplo un estudio sobre la incidencia del SIDA entre las mujeres de la etnia kalanga en Botswana. Se desconoce el tamaño de la población, aunque se asume suficientemente grande. La incidencia del SIDA en el país ronda el 23,9% (p). El nivel de confianza deseado es del 90% (1-α) y el margen de error, del 3% (d).

Así:

Z_{α} = 1,645 (corresponde a 90% según la tabla anterior)

p = 0,239

d = 0,03

Por tanto, $n = 1{,}645^2 \cdot 0{,}24 \cdot (0{,}76) / 0{,}03^2 = 546{,}85$

Habría que tomar una muestra aleatoria de 547 mujeres de dicha etnia para poder estimar la proporción con la precisión requerida (90%, 3%).

En caso de que se trate de una variable no dicotómica, es decir donde pueda tomar varios valores, para determinar 'p' se suele optar por el valor (o grupo de valores) más relevante para el estudio o por p=0,5. Pongamos por ejemplo una variable sobre satisfacción con un curso, que puede tomar los valores nada, poco, bastante o mucho. Quizás lo más relevante sería cuántos están bastante o muy satisfechos con el curso, por lo que 'p' sería la proporción total de población bastante o muy satisfecha. En el próximo ejemplo se utilizará un caso similar para ilustrarlo.

Dándole la vuelta a la fórmula del tamaño de la muestra, se puede calcular el intervalo de confianza:

$$d = \sqrt{\frac{Z_\alpha^2 \cdot p \cdot (1-p)}{n}}$$

Puedes ver cálculos similares en este estudio, que detalla la metodología empleada (páginas 5 y 6), describiendo las técnicas de muestreo y dando el nivel de precisión según el tamaño de muestra elegido.

Cuando la población es finita (<100000), las fórmulas incorporan el tamaño de la población N.

$$n = \frac{N \cdot Z_\alpha^2 \cdot p \cdot (1-p)}{d^2 \cdot (N-1) + Z_\alpha^2 \cdot p \cdot (1-p)} \qquad d = \sqrt{\frac{Z_\alpha^2 \cdot p \cdot (1-p)}{n} \cdot \frac{N-n}{N-1}}.$$

A raíz del estudio sobre el SIDA del ejemplo anterior, se decide replicar el estudio para las mujeres de la etnia basarwa. Por cuestiones de presupuesto, se fija una muestra de 600 mujeres. Para facilitar la comparación, se mantiene el mismo nivel de confianza (90%). Además de determinar si tienen o no la enfermedad, se les preguntará su nivel de conocimiento respecto a las vías de transmisión (muy bajo, bajo, intermedio, alto o muy alto). Se estima que el número de mujeres basarwa en el país es de menos de 20000 personas.

Los márgenes de error de la estimación de incidencia del SIDA serán (asumiendo población infinita):

$Z_\alpha = 1{,}645$

p = 0,239 (igual que en el ejemplo anterior)

n = 600

$d = \sqrt{(1{,}645^2 \cdot 0{,}24 \cdot (0{,}76) / 600)} = 0{,}0287 = 2{,}9\%$

Si se calculasen con la fórmula para poblaciones finitas, se podría ajustar mejor el margen de error:

N = 20000

$d = \sqrt{[(1{,}645^2 \cdot 0{,}24 \cdot (0{,}76) / 600) \cdot (20000-600)/(20000-1)]} = 0{,}0282 = 2{,}8\%$

Por tanto, el estudio permitiría estimar la incidencia del SIDA con un margen de error del 2,8% (nivel de confianza 90%).

En cuanto al nivel de conocimiento respecto a las vías de transmisión (variable no dicotómica), se determinó el grupo de valores 'alto' y 'muy alto' como el más relevante para el estudio. De otro estudio anterior, se sabe que solo un 8% de la población tiene un conocimiento muy alto, mientras que un 30% tiene un conocimiento alto.

Así, para la variable conocimiento de las vías de transmisión, con el mismo nivel de confianza, los cálculos serían.

$Z_\alpha = 1{,}645$

p = 0,38 (suma de 30% y 8%)

n = 600

N = 20000

Por tanto, $d = \sqrt{(1{,}645^2 \cdot 0{,}38 \cdot (0{,}62) / 600)} = 0{,}0326 = 3{,}3\%$ (población infinita)

$d = \sqrt{[(1{,}645^2 \cdot 0{,}38 \cdot (0{,}62) / 600) \cdot (20000-600)/(20000-1)]} = 0{,}0321 = 3{,}2\%$ (población finita)

5.2.2 Tamaño de muestra para estimar una media

Cuando se quiere calcular la media de una variable cuantitativa continua, es necesario conocer:

La varianza (S^2) o la desviación estándar (S) en la población. Como se ha señalado anteriormente, se puede hacer un estudio previo para determinarla, o estimarla a partir de otros estudios similares.

El intervalo de confianza deseado (2·d), lo que vendría a ser el margen de error de la media (±d).

El nivel de confianza (1-α) y su Z_α correspondiente.

Así, se calcula el tamaño de muestra (n) mediante la siguiente fórmula:

$$n = \frac{Z^2_\sigma \cdot S^2}{d^2}$$

Véase un ejemplo continuando con el estudio sobre nutrición infantil:

En la región del estudio hay unos 12.000 niños y niñas de 9 y 10 años, con lo que la población se puede considerar suficientemente grande. A partir de un estudio previo, se estima que la varianza es de 8 kg^2.

El estudio pretende determinar el peso medio, con un intervalo de confianza de 1 kg (margen de error de ±0,5kg) y un nivel de confianza del 99%.

Así:

$Z_\alpha = 2{,}576$

$S^2 = 8$

$d = 0{,}5$

$N = 12000$ (infinita)

Por tanto, $n = 2{,}576^2 \cdot 8/0{,}5^2 = 213$

Habría que tomar una muestra aleatoria de 213 niños y niñas para poder estimar una media con la precisión requerida (99%, ±0,5kg).

Cuando la población es finita (<100000), la fórmula incorpora el tamaño de la población N.

$$n = \frac{N \cdot Z^2_\sigma \cdot S^2}{d^2 \cdot (N-1) + Z^2_\sigma \cdot S^2}$$

Se puede comprobar en el ejemplo que con la fórmula no simplificada, se obtiene un resultado similar (210), con lo que la simplificación no ha supuesto un error considerable.

Dándole la vuelta a la fórmula, se pueden calcular los intervalos de confianza.

$$d = \sqrt{\frac{Z^2_\alpha \cdot S^2}{n}}$$ para población infinita

$$d = \sqrt{\frac{Z^2_\alpha \cdot S^2}{n} \cdot \frac{N-n}{N-1}}$$ para población finita

Esto sirve, por ejemplo, si se quiere estudiar varias variables de una muestra.

Si en el estudio se quiere estimar también la estatura media, se parte del tamaño de muestra n =213. A partir de otro estudio sobre estaturas, se estima que la desviación estándar de la estatura es S = 3 cm. Se desea un nivel de confianza del 99%

Así: d = 2,58·3 / √213 = 0,53

Esto representa el margen de error de la media muestral de estatura. Es decir, que si saliese una media muestral de 126,4 cm, la media poblacional que se estimaría sería 126,4 ± 0,53. Es decir, tenemos un 99% de acertar si decimos que la estatura media está en el intervalo [125,87 ; 126,96].

Aunque no se detalla aquí por su poca frecuencia, se hace notar que para muestras pequeñas (n<30), se debe utilizar la distribución t de Student en vez de la normal tipificada. El procedimiento es igual pero se sustituye Z_α por $t_{\alpha/2,n-1}$. Igual que los Z_α vienen de una tabla, los valores de la t de Student, también tienen sus propias tablas.

5.2.3 Inferencia en muestreos estratificados y por etapas

El muestreo aleatorio sistemático no presenta diferencias con respecto al muestreo aleatorio simple, así que se le aplican los mismos cálculos.

En el caso del **muestreo aleatorio estratificado**, el error aleatorio total se calcula de manera diferente. Al igual que ocurre con la media, se calcula ponderando según la representación de cada estrato. Para estimar una media, como la fórmula del tamaño de muestra depende de la varianza, la varianza (S^2_h) de cada estrato (h), el número de sujetos del estrato en la muestra (n_j) y en la población (N_j). La fórmula queda así:

$$n = \frac{\sum_{h=1}^{H} \frac{N_h^2 \cdot S_h^2}{n_h/n}}{N^2 \cdot \left(\frac{d^2}{Z_\alpha^2}\right) + \sum_{h=1}^{H} N_h \cdot S_h^2}$$

Se despejaría d para obtener el intervalo de confianza.

Para una proporción sería análogo, sustituyendo S_h^2 por $p_h \cdot (1-p_h)$.

En cualquier caso, el muestreo aleatorio estratificado genera menos error aleatorio que el muestreo aleatorio simple (si los estratos tienen lógica), por lo que si aplicamos las fórmulas del muestreo aleatorio simple, estaremos pecando de conservadores, cosa que –en este caso– no es grave.

Remarcar que estamos hablando aquí de la muestra en general. En la práctica cuando estratificamos es porque nos interesa analizar algún estrato separadamente, y requeriremos de precisiones específicas para dicho estrato (o para todos). En tal caso, aplicaríamos independientemente para cada estrato los cálculos del apartado anterior para muestreo aleatorio simple, obteniendo tamaños de muestra de estrato a estrato. En realidad, esto pasa también en una muestra no estratificada si a posteriori queremos particularizar el análisis para subgrupos en la muestra.

Para **muestreos por etapas**, la práctica más habitual es calcular la muestra como si fuese un muestreo aleatorio simple y luego multiplicarlo por un factor corrector llamado efecto de diseño. El cálculo de este factor es complejo y queda fuera del alcance del capítulo. Sí es importante saber que para un tamaño de muestra total dado, el error del muestreo por etapas es mayor que el de un muestreo aleatorio simple. Así el efecto de diseño es siempre mayor que 1. Para lo demás, requeriremos de asesoramiento experto.

El único caso asequible es el del muestreo más básico, con una primera etapa de muestreo aleatorio entre conglomerados y una segunda etapa donde todos los sujetos de los conglomerados seleccionados entran en la muestra. Al observar a todos los sujetos de los conglomerados muestreados, no hay error aleatorio en la segunda etapa. Para la primera etapa se utilizarían las fórmulas vistas en el apartado anterior, pensando que cada conglomerado es como un sujeto y por tanto el tamaño de muestra 'n' se refiere al número de conglomerados. Igualmente, las varianzas o proporciones no se obtendrían de los sujetos, sino a partir de los valores promedio de los distintos conglomerados.

5.2.4 Muestreos pseudoaleatorios y tamaño de muestra

Como sabemos, en contextos de desarrollo, hay muchos factores que pueden provocar sesgo, sobre todo por la dificultad de acceder al marco muestral.

Cuando sí se dispone de un marco muestral, aunque sesgado, podemos realizar un muestreo aleatorio, y aplicar la inferencia estadística, pero referida no ya a la población entera, sino a los sujetos que efectivamente había en el marco muestral. Por ejemplo, si en Logone muestreamos aleatoriamente a partir de un listado de familias que no se actualiza desde hace 10 años, nuestros resultados serán generalizables solo a las familias que llevan al menos diez años en la región.

En otras ocasiones no disponemos de marco muestral alguno. En esos casos una estrategia es muestrear por etapas y reconstruir marcos muestrales. Por ejemplo, seleccionaríamos comunidades de Logone a partir de una lista (sí es posible que exista una lista actualizada de comunidades de la región) y en las comunidades seleccionadas podríamos hacer un mapeo para generar una lista de hogares. Si esto tampoco es posible, tendremos que optar por muestreos pseudoaleatorios (ver apartado 2.3.3), como el muestreo por áreas o por cuotas. O renunciar a realizar el estudio cuantitativo, si la inferencia es condición sine qua non.

Entonces, si finalmente hacemos un muestreo pseudoaleatorio, ¿cómo calculamos la muestra?

Como la inferencia no se puede aplicar, las fórmulas vistas anteriormente no sirven. Tampoco sirve el criterio cualitativo de saturación aplicado en muestreos no aleatorios (subjetivo, bola de nieve, etc.). No hay reglas concretas para calcular el tamaño en muestreos pseudoaleatorios. Normalmente se suele determinar un tamaño por experiencia. Si se carece de ésta, es importante contar con el asesoramiento necesario. Otra opción es calcular el tamaño de muestra como si fuese un muestreo aleatorio, y luego sobredimensionarlo bastante.

En realidad, más importante que tener una muestra grande, es contrarrestar los posibles sesgos del muestreo. Así, deberíamos preocuparnos más en introducir una mayor aleatoriedad que en calcular el tamaño ideal de la muestra. Combinar rutas y cuotas, es una buena estrategia. También identificar posibles características clave en el problema investigado, y estratificar la muestra para poder controlarlas. Por ejemplo, en un estudio sobre oportunidades de educación, puede ser interesante establecer cuotas según el sexo, y encuestar a tantas mujeres como hombres. En uno sobre relaciones personales, resultaría interesante que entre los encuestados haya personas casadas / solteras / etc. en cantidad similar a su presencia en la población.

Un ejemplo bastante común sería un sondeo electoral con una muestra de 100 sujetos, que se podría estructurar en tres rangos de edad: 15-39 / 40-64 / >65+, y encuestar a 20/15/12 hombres y 20/17/ 16 mujeres en cada rango, ya que las mujeres viven más. Cabe recalcar la importancia de que el tamaño de las cuotas sea proporcional a su presencia en la población. Es decir, si mis cuotas son por edades, que el

porcentaje de mayores de 65 años en mi muestra sea igual al porcentaje de mayores de 65 años en la población (dato que obtendremos del censo o de registros demográficos).

Y por supuesto, es vital que el encuestador sea consciente de la necesidad de maximizar la aleatoriedad; por ejemplo que si encuentra a un grupo de personas no aproveche para encuestar a todos.

En algunos casos, podemos también introducir variables de control para a posteriori tener un argumento con el que defender que el método de selección y el tamaño de la muestra han sido adecuados. Para ello, necesitamos conocer alguna variable de la población, y ver que la variable en nuestra muestra presenta valores semejantes. Por ejemplo, podría existir información en Logone sobre el tamaño medio familiar de la región. Pongamos que es de 5,4 hijos. Si realizamos un muestreo por cuotas y el tamaño familiar medio de nuestra muestra es 5, podemos estar más contentos que si nos sale 3. Esa gran diferencia (de 3 a 5,4) sería un indicador de que nuestra muestra está sesgada o es muy pequeña.

Otra técnica para ver si el tamaño de muestra es suficiente, se inspira en la idea de la saturación. Consistiría en eliminar aleatoriamente algunos sujetos de la muestra y ver cuánto varían los estadísticos muestrales. Si la variación no es significativa, es indicio de que el tamaño de muestra es suficiente. No obstante, no nos da pistas acerca de los sesgos, que pasarían desapercibidos.

5.3 Contraste de hipótesis

Hemos visto la forma de calcular el tamaño de muestra para estimar medias y proporciones poblacionales mediante intervalos de confianza. Pero en el capítulo anterior hemos visto muchos más estadísticos que también nos gustaría generalizar. Para ello, la inferencia estadística nos ofrece los **contrastes de hipótesis**. Entre sus muchas aplicaciones, los contrastes de hipótesis permiten estudiar las diferencias entre medias o proporciones, comprobar la significación de una correlación, analizar la varianza, etc.

De forma general, el contraste de hipótesis consiste en hacer una afirmación sobre una propiedad de la población (establecer una hipótesis), y aplicar una prueba estadística para contrastar si esa afirmación es creíble; si es compatible con lo observado en la muestra.

Los contrastes de hipótesis que nos pueden resultar más interesantes son:

- Ver si una media o porcentaje supera un determinado umbral
- Comparar dos muestras
- Comprobar si una relación entre variables es significativa

Como siempre en inferencia, para que estos contrastes de hipótesis tengan potencia, necesitamos un tamaño de muestra determinado.

El mecanismo más detallado del contraste de hipótesis es el siguiente:

(1) establecer una hipótesis nula H_0 (lo que queremos comprobar, generalmente expresado en negativo)

(2) establecer la hipótesis alternativa H_1 (lo contrario a la hipótesis nula)

(3) ver qué resultados hubiese obtenido con el muestreo en caso de cumplirse la hipótesis nula

(4) utilizar un estadístico de contraste para comparar los resultados realmente obtenidos en la muestra con los de la hipótesis nula, obteniendo un estimador de la compatibilidad entre ambos

(5) En función del resultado de la prueba estadística (p-valor), acepto o no mi hipótesis

El resultado de la prueba estadística es el p-valor, o nivel de significación. Es semejante al nivel de confianza en los intervalos de confianza. Técnicamente, mide la probabilidad de haber obtenido el resultado que hemos obtenido de la muestra, si suponemos que la hipótesis nula es cierta. En otras palabras, el p-valor mide el riesgo de errar si rechazamos la hipótesis nula. Si p es bajo, puedo rechazar la hipótesis nula, y por tanto aceptar la alternativa. El valor se considera 'bajo' si es menor de 0,05… pero entendamos que un p-valor de 0,05 significa que rechazo la hipótesis nula con un riesgo del 5% de estar equivocado. Algunas investigaciones que requieren más precisión, establecen el rasero (en realidad se llama potencia de contraste) en 0,01… es decir solo quieren un 1% de riesgo de fallo cuando rechazan la hipótesis nula.

Por otro lado, si p es alto (por ejemplo 0,25), no significa que la hipótesis nula sea cierta. Significa que no tenemos suficiente evidencia para rechazarla; que si la rechazamos tenemos un 25% de probabilidad de estar fallando. Una opción es conseguir más evidencias, es decir, repetir el estudio con una muestra mayor, reduciendo así la incertidumbre.

Visto así en abstracto puede resultar lioso, así que a continuación lo veremos aplicado en algunos ejemplos prácticos de pruebas estadísticas.

Nos centraremos en establecer la hipótesis nula (1) y la hipótesis alternativa (2), y en interpretar correctamente los resultados (5). Los pasos intermedios (3) y (4) se los dejaremos a los programas estadísticos por esta vez.

Los ejemplos que veremos a continuación presentan de manera muy práctica los tres contrastes de hipótesis tipo, planteados principalmente con la intención de ilustrar el concepto de contraste de hipótesis. No se explica cómo se realizan los cálculos ya que para ello haría falta ampliar conocimientos de estadística y/o utilizar herramientas informáticas –que recordemos que son verdaderamente útiles cuando entendemos qué hacen.

Finalmente, cabe señalar que la aplicación de contrastes de hipótesis requiere de muestreos aleatorios. Se suponen en general muestras superiores a 30 y se asumen distribuciones normales o binomiales.

5.3.1 Ver si una media o porcentaje alcanza un determinado umbral

Normalmente las medias y porcentajes se calculan mediante intervalos de confianza. Pero en ocasiones, existe un interés específico por saber si un parámetro poblacional alcanza un determinado umbral o no.

Por ejemplo, supongamos que hay indicios de una epidemia de SIDA entre las mujeres de la etnia kalanga. Se considera epidemia cuando la incidencia supera el 30%. Para decidir si declarar el estado de epidemia, las autoridades sanitarias harían una encuesta a una muestra aleatoria y realizarían un contraste de hipótesis.

Lo que quieren averiguar es si las mujeres de la etnia kalanga presentan una incidencia de SIDA (IS) mayor del 30%. Así, la hipótesis nula (expresada en negativo) será:

La incidencia de SIDA entre las mujeres kalanga es igual o menor a 30%. H_0: IS≤30% (1)

La hipótesis alternativa es lo contrario: H_1: IS>30% (2)

Se seleccionaría una muestra aleatoria de tamaño 'n' y se obtendría una determinada incidencia de SIDA (is).

El paso de aplicación del contraste (4) se refleja completo para este ejemplo por ser el primero, pues ayuda a visualizar el proceso. Pero volvemos a insistir en que lo importante es comprender los conceptos más que saber realizar el cálculo.

El estadístico de contraste sería:

$$Z_\alpha < \frac{p - P}{\sqrt{\frac{P \cdot (1-P)}{n}}} = \frac{is - IS}{\sqrt{\frac{IS \cdot (1-IS)}{n}}}$$

Relaciona el porcentaje muestral (is), el porcentaje poblacional según la hipótesis nula (IS) y el tamaño de muestra (n). Una vez obtenido, en la tabla de la distribución normal tipificada obtendríamos el p-valor equivalente. (4)

Si dicho p-valor es menor de 0,05, se rechaza la hipótesis nula (y se declara la epidemia). (5)

Completemos el ejemplo:

Se toma una muestra aleatoria de 537 mujeres kalanga (dimensionada en función de los recursos disponibles), y se obtiene una incidencia de sida is=31,5%.

¿Podemos concluir que la incidencia poblacional IS > 30%?

El estadístico de contraste se calcularía así:

$$Z_\alpha < \frac{is - IS}{\sqrt{\frac{IS \cdot (1-IS)}{n}}} = \frac{0{,}315 - 0{,}3}{\sqrt{\frac{0{,}3 \cdot (1-0{,}3)}{537}}} = 0{,}758$$

En las tablas a Z =0,758 se obtiene el valor complementario del p-valor = 0,224

Es un valor muy alto, si rechazo la hipótesis nula y declaro la epidemia, mi riesgo de fallar sería del 22,4%.

Dada la situación, lo más recomendable sería repetir el muestreo con una muestra mayor, que podría dimensionar despejando n de la fórmula anterior.

El contraste para una media sería análogo, partiendo de la hipótesis nula de que la media poblacional supera cierto valor.

5.3.2 Comparar dos muestras independientes

Un organismo internacional quiere diseñar un estudio para evaluar el impacto de dos programas de microcréditos que está aplicando en Bangladesh y saber cuál es más eficaz. La eficacia del programa la miden en función del aumento de renta experimentado por la persona beneficiaria del microcrédito. Los dos programas se han promocionado conjuntamente y cada microemprendedor/a elegía el programa que le parecía más interesante. Se han concedido así miles de microcréditos de ambos programas a lo largo y ancho del país. La idea es realizar un contraste de hipótesis para comparar las medias de aumentos de ingresos entre los programas

Para ello, realizan dos muestreos entre beneficiarios del programa A y beneficiaros del programa B, con tamaños de muestra n_A=61 y n_B=61. El resultado de la encuesta arroja unos aumentos de renta $\bar{x}_A$=110$ y $\bar{x}_B$=100$, con varianzas S_A^2=35 y S_B^2=26

La hipótesis nula asume que no hay diferencia entre las medias poblacionales. H_0: $\mu_A=\mu_B$ (1)

La hipótesis alternativa es que sí hay diferencia. H_1: $\mu_B \neq \mu_A$ (2)

El paso (3) y (4) no se describen ya. Con las medias y varianzas muestrales, los tamaños de muestra y las tablas de la t de Student, se obtendría un p-valor < 0,001.

Ello nos permite rechazar la hipótesis nula y aceptar la alternativa; un programa es más eficaz que el otro (5). El riesgo de fallar al afirmar esto es ínfimo (<0,1%).

El proceso es análogo para proporciones.

5.3.3 Comprobar si una relación entre variables es significativa

En el capítulo anterior, vimos distintas formas de visualizar y cuantificar relaciones entre variables, como las tablas de contingencia y el coeficiente phi (variables cualitativas) y el diagrama de dispersión y el coeficiente de correlación (variables cuantitativas). Para determinar si dichas relaciones son estadísticamente significativas, se aplican los contrastes de hipótesis basados en la distribución Chi-cuadrado para tablas de contingencia y en la distribución t de Student para coeficientes de correlación.

Veamos un ejemplo para el coeficiente de correlación (recordemos que $r = S_{xy} / (S_x \cdot S_y)$, ver apartado 4.4.2).

Supongamos que estamos en una comunidad estudiando los distintos factores que determinan el nivel del sueldo de las personas de esa comunidad. Se exploran diversas variables mediante una encuesta a 20 personas elegidas aleatoriamente entre todas las personas con trabajo asalariado entre 30 y 50 años de edad. En el análisis descriptivo posterior se detecta una correlación bastante fuerte entre el sueldo (x) y los años de escolarización (y), con un $r_{xy}=0,885$.

Llegados a este punto, se decide realizar un contraste de hipótesis para comprobar que dicha relación es estadísticamente significativa. El nivel de significación requerido es p-valor=0,01.

La hipótesis nula asume que no hay correlación entre las variables. H_0: $r_{xy}=0$ (1)

La hipótesis alternativa es que sí hay diferencia. H_1: $r_{xy}\neq 0$ (2)

El paso (3) y (4) no se describen ya. Con el coeficiente de correlación, el tamaño de muestra y las tablas de la distribución t de Student, se obtendría un p-valor < 0,001.

Ello nos permite rechazar la hipótesis nula y aceptar la alternativa. Por tanto, sí existe una relación entre el número de años de escolarización y el sueldo. (5)

Bibliografía

Barahona, C. y S. Levi, (2002). *How to generate statistics and influence policy using participatory methods in research*. SSC, Working Paper.

Cea d'Ancona, M.A., (2001). *Metodología cuantitativa: estrategias y técnicas de investigación social.* Madrid, Síntesis.

Chambers, R., (2007). *Who Counts? The Quiet Revolution of Participation and Numbers*. IDS, Working Paper 296.

Domínguez, M. y A. Coco, (2000). *Tècniques d'investigació social I.* Barcelona, Edicions de la Universitat de Barcelona.

García Muñoz, T., (2003). *El cuestionario como instrumento de investigación / evaluación* [visitado el 08.04.2012]

Kanbur, R., (2005). *Q-Squared - Qualitative and Quantitative Poverty Appraisal: Complementarities, Tensions and the Way Forward*. Cornell University, Q-Squared Working Paper No. 1.

Mayoux, L., (2006). *"Quantitative, Qualitative or Participatory? Which Method, for What and When?"* en Desai, V. y R. B. Potter (eds.), Doing development research. Thousands Oaks, Sage Publications.

Molteberg, E. y C. Bergstrøm, (2000). *Our Common Discourse: Diversity and Paradigms in Development Studies.* Noragric.

Pulido, A., (1992). *Estadística y Técnicas de Investigación Social*. Madrid, Pirámide.

Romero,R. y L. R. Zúnica, (2005). *Métodos Estadísticos en Ingeniería*. Valencia, Editorial UPV.

Russel Bernard, H., (2002). *Research Methods in Anthropology. Qualitative and Quantitative Approaches.* AltaMira Press.

Sayer, A., (2000). Realism and social science. London, Sage Publications.

Statistical Services Centre, (2001). *Some Basic Ideas of Sampling*, Statistical Good Practice Guidelines. Reading, University of Reading.

Sumner, A. y M. Tribe, (2008). *International development studies: theories and methods in research and practice*. London, SAGE Publications Ltd.

Universitat Oberta de Catalunya, (2002). *Proyecto E-MATH: "Uso de las TIC en asignaturas cuantitativas aplicadas"*. FUOC. [visitado el 15.04.2012]

EL GRUPO DE ESTUDIOS EN DESARROLLO, COOPERACIÓN Y ÉTICA

El Grupo de Estudios en Desarrollo, Cooperación y Ética (GEDCE) de la Universitat Politècnica de València es un grupo de investigación multidisciplinar formado por profesores titulares del Departamento de Proyectos de Ingeniería, investigadores y técnicos de la UPV que, desde el año 1995, orientan su docencia, investigación y extensión social al ámbito del desarrollo, la cooperación internacional y la ética aplicada.

El Grupo de Estudios en Desarrollo, Cooperación y Ética (GEDCE) de la Universitat Politècnica de València es un grupo de investigación multidisciplinar formado por profesores titulares del Departamento de Proyectos de Ingeniería, investigadores y técnicos de la UPV que, desde el año 1995, orientan su docencia, investigación y extensión social al ámbito del desarrollo, la cooperación internacional y la ética aplicada.

El GEDCE imparte docencia de grado y posgrado relacionada con sus ámbitos de interés: desarrollo, cooperación internacional y ética aplicada. Coordina e imparte el Máster en Cooperación al Desarrollo (www.mastercooperacion.upv.es) con su especialización en Gestión de Proyectos y Procesos de Desarrollo así como el Especialista Universitario en Responsabilidad Social Corporativa. Participa en postgrados en América Latina en colaboración con universidades y organizaciones latinoamericanas.

A nivel de investigación, cuenta con diversas líneas relacionadas con la cooperación al desarrollo, las metodologías de planificación y gestión de procesos e intervenciones, la ética aplicada y la responsabilidad social corporativa, la gobernanza democrática, la educación para el desarrollo o la tecnología en el desarrollo.

Como extensión social, el grupo presta servicios de asesoría a entidades del Norte y del Sur, ONGD y administraciones públicas, de ámbito local e internacional. Participa en el diseño, la ejecución y evaluación de proyectos y presta asesoría y capacitación en gestión y organización de ONGD, metodologías de proyectos y tecnologías apropiadas a contrapartes del Sur.

Toda la información sobre el GEDCE se puede encontrar en gedce.webs.upv.es

LOS CUADERNOS DOCENTES EN PROCESOS DE DESARROLLO

Los *Cuadernos Docentes en Procesos de Desarrollo* son un espacio para el desarrollo de materiales de apoyo a la docencia que se imparte en el Máster en Cooperación al Desarrollo de la Universitat Politècnica de València. En ellos se desarrollan los contenidos centrales de las diversas asignaturas y constituyen, por tanto, el eje que articula y relaciona los diversos conceptos abordados en ellas.

Los Cuadernos publicados pueden encontrarse en cuadernos.dpi.upv.es