Introducción a la estadística industrial y la ciencia de datos utilizando Microsoft Excel y R

Manual para la Formación Profesional

Juan Riera

12/12/22

Tabla de contenidos

Prefacio

Este libro trata sobre la enseñanza de algunos métodos básicos de la estadística y la moderna ciencia de datos y su aplicación al entorno industrial. Está concebido de forma práctica con multitud de ejemplos, no sólo industriales, con el objetivo de mostrar los métodos cuantitativos de análisis, y también el razonamiento necesario para dar sentido a los resultados presentados por las herramientas de análisis de datos.

El objetivo del libro es acercar a los estudiantes de la Formación Profesional al uso de las herramientas de análisis de datos industriales. El entorno de la industria 4.0 produce un enorme y constante flujo de datos como resultado tanto de la implantación de sistemas de captura automáticos como del aumento de la tecnificación de los puestos de trabajo; se requiere por parte de los profesionales industriales que sean capaces de analizar esta enorme cantidad de datos para transformarlos en información para la decisión. En la empresa industrial actual, son los ingenieros y técnicos de planta, y no estadísticos o ingenieros informáticos, quienes participan diariamente en la presentacion y discusion de los datos y en la toma de decisiones operativas, tanto en los equipos de trabajo como ante la Dirección. Por esta razón, considero necesario proporcionar a los estudiantes de la Formación Profesional un conocimiento suficiente de los conceptos, herramientas y métodos del análisis de datos, así como de las técnicas básicas de presentación y comunicación de la información.

El control estadístico de la calidad empezó con W.E. Deming a mediados del pasado siglo XX, y fueron las empresas japonesas, sobre todo las automovilísticas, las que inicialmente recogieron el testigo de Deming y aplicaron estos principios a la mejora de la producción industrial, difundiendo su conocimiento a todos los niveles jerárquicos de las organizaciones, desde los operarios de línea hasta los más altos directivos. Desde ese momento

hasta la difusión actual de los métodos Six Sigma gracias a General Electric y Motorola, la mejora industrial de los procesos ha estado siempre apoyada en la correcta utilización de estas metodologías.

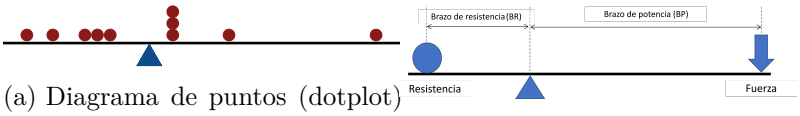
La enseñanza de los conceptos estadísticos está, casi siempre, a cargo de profesores con una gran formación en matemáticas. Estos profesores suelen identificar la comprensión de los conceptos estadísticos con su comprensión matemática. Sin embargo, cuando enseñamos estadística industrial, debemos hacer énfasis tanto en las ideas y la comprensión de los conceptos como en su utilización práctica, y reconocer que el razonamiento matemático no es el único camino para la comprensión conceptual.

La práctica de la estadística requiere buen juicio y sentido común. Dado que el buen juicio se desarrolla con la experiencia, un curso de iniciación debe presentar unas guías claras de aplicación de los métodos, y no dar por supuestas unas exigencias excesivamente altas sobre la capacidad de juicio analítico de los estudiante; no sería un planteamiento razonable. Con el fin de desarrollar esta capacidad de juicio analítico he introducido explicaciones detalladas en la mayor parte de los ejemplos. En todos los casos, los ejercicios requerirán del estudiante no sólo una resolución numérica, sino el uso del juicio analítico y la explicación verbal (o escrita) de las decisiones tomadas y conclusiones realizadas. Creo que este planteamiento será mucho más beneficioso a largo plazo que limitarse a una simple resolución numérica.

Mi experiencia industrial me ha mostrado que en las situaciones reales, sobre el terreno, la comprensión práctica de los conceptos es más importante que su rigurosa formulación matemática. Por esta razón, en el desarrollo del contenido del libro he insistido más en la forma de aplicar las herramientas y entender los análisis que en el conocimiento formal de las fórmulas estadísticas y su deducción matemática. He hecho especial hincapié en la utilización de herramientas sencillas, sobre todo gráficas, que casi siempre son una ayuda para comprender la información contenida en un conjunto de datos. El objetivo es proporcionar al estudiante las bases de la metodología del análisis de datos y del análisis estadístico, y cómo puede aplicarse a la resolución

de problemas técnicos concretos, más que el conocimiento de la teoría matemática de la estadística.

He evitado las explicaciones formales sobre temas estadísticos cuando no son indispensables para su aplicación práctica. Así, por ejemplo, al explicar la media de un conjunto de datos considero más importante entender el concepto físico de "centro de gravedad" que los conceptos estadísticos de esperanza matemática, que no se tocan en este texto. En este sentido, he intentado que el alumno aprenda a diferenciar "en qué consiste" un estadístico, de "cómo se calcula". Comprender la diferencia entre el concepto y su fórmula de cálculo es fundamental para entender en qué situación debe usarse uno u otro estadístico.



(a) Diagrama de puntos (dotplot) mostrando la posición de la media

(b) Palanca de primer grado con el fulcro como punto de equilibrio

Figura 1: Comparación gráfica del significado de la media en un diagrama de puntos (*dotplot*) con una palanca de primer grado.

Un curso de introducción a la estadística y análisis de datos industriales debe ser ante todo práctico y orientado a su aplicación en el entorno industrial real. Los principales temas de trabajo estadístico en la industria tienen que ver con la captura de datos, su almacenamiento y su depuración, su descripción utilizando gráficos, la inferencia (intervalos de confianza y tests), la construcción de modelos explicativos, el diseño de los experimentos industriales, el control estadístico de la calidad y la exposición y presentación de resultados. Dado el alcance limitado de este libro, algunos de estos temas se tratarán de forma muy ligera, y necesitarán de un estudio posterior si el alumno tiene interés en profundizar en ellos. A pesar de que los temas más especializados puedan ser importantes en algunas aplicaciones específicas, no preparan al estudiante para lo que se va a encontrar en el terreno en la mayor parte de las ocasiones. En cambio, la resolución de problemas en equipo en un entorno de aprendizaje dinámico enfrentándose a problemas exigentes, y el desarrollo de las habilidades de análisis, de síntesis y de

comunicación, tendrán un impacto mucho más positivo.

He intentado mostrar la necesidad de que los estudiantes comprendan y apliquen el método científico en el entorno industrial, y no sólo apliquen un recetario de procedimientos de manera automática. Es mucho mas importante la comprensión y adecuada utilización del método científico y de las herramientas y gráficos básicos, antes que la aplicación rutinaria y mecánica de determinadas fórmulas matemáticas o métodos sofisticados y complejos que el alumno puede no comprender en toda su profundidad.

Las industrias líderes destacan por la aplicación intensiva de métodos sofisticados, tales como Six Sigma, Lean Manufacturing, diseño robusto de productos, y otros que hacen un uso intensivo de los datos, tanto de los obtenidos en producción como de los obtenidos en la realización de experimentos bien diseñados. La mejora de la competitividad en estas empresas no se debe tanto a la aplicación de unos u otros métodos, como al desarrollo del juicio analítico de sus equipos y a la aplicación de lo aprendido a la mejora continua de los procesos industriales. Veremos que la experiencia y el conocimiento tecnológico de estos procesos son fundamentales para el desarrollo del buen juicio analítico, y, en consecuencia, para la buena interpretación de los resultados que se obtienen con las herramientas estadísticas y de análisis.

Tratándose de un libro para el uso en la Formación Profesional, considero prioritario que su estudio se oriente al desarrollo de habilidades que sean de aplicación práctica directa en el puesto de trabajo y además faciliten la empleabilidad del estudiante, y no a la obtención de conocimiento abstracto. El Informe sobre el futuro del empleo, publicado por el Foro Económico MUndial en mayo de 2023, considera que las principales habilidades clave en el trabajo del futuro serán en primer lugar el desarrollo del pensamiento analítico y, en segundo, del pensamiento crítico; unidas al desarrollo de la curiosidad y la voluntad de aprendizaje a lo largo de la vida, la eficacia, la confianza en el propio trabajo y la atención al detalle. La voluntad de este libro es proporcionar conocimientos que ayuden al estudiante a desarrollarse en esta dirección.

A quién va dirigido este libro

El libro está orientado a completar la formación técnica de los estudiantes de Formación Profesional, en las especialidades relacionadas con la actividad productiva industrial. También creo que será de utilidad para los técnicos industriales en activo que no han tenido una adecuada formación en estas metodologías, y que han encontrado dificultad para lanzarse a su aprendizaje mientras desarrollan so actividad profesional. Espero, también, que los profesores de la Formación Profesional en estos ámbitos de competencia encuentren en este documento los elementos de apoyo que les permitan integrar estas enseñanzas en sus respectivos ciclos formativos.

En todos los casos, el aprendizaje requerirá de un esfuerzo que quizás será mayor en los estudiantes que no tengan una base mínima en álgebra y cálculo. En estos casos, el trabajo en equipo y la discusión abierta entre compañeros y con los profesores ayudará a la comprensión de los conceptos.

Organización del libro

El **capítulo 1** proporciona una introducción general al **pensamiento estadístico** y su aplicación industrial. Introduce también algunos conceptos sobre la ética en análisis de datos y una reflexión sobre la honestidad del investigador o analista. Se introduce también el concepto actual de **repetibilidad** en la elaboración de los informes estadísticos.

El **capítulo 2** presenta las principales herramientas que usaremos en este libro para el análisis de los datos industriales, concretamente Microsoft Excel y R.

El **capítulo 3** trata fundamentalmente de la forma de recoger los datos y su almacenamiento. Introduce el concepto de **datos ordenados** o **arreglados** (*tidy data*), que resulta fundamental para las fases posteriores de análisis.

En el **capítulo 4** se introducen los métodos básicos para resumir tablas de datos y la **presentación mediante el uso de gráficos**, en lo que se conoce como *exploración de datos*,

que suele ser un paso previo al análisis más detallado y a la formulación de hipótesis.

El **capítulo 5** introduce el concepto de **probabilidad**, así como las distribuciones de probabilidad, necesarias para la construcción de los tests de hipótesis y, en general, de la estadística inferencial. Este es un contenido que se presenta de forma breve y sobre todo práctica.

En el **capítulo 6** se presentan los métodos para detectar la **relación entre dos variables**, haciendo énfasis en los métodos gráficos, y se discute las diferencias entre correlación y causalidad.

El **capítulo 7** introduce de manera sencilla el **análisis de la varianza**, necesario para métodos importantes en la industria como el control de la precisión analítica, que se trata en el capítulo siguiente.

El **capítulo 8** trata del análisis del sistema de medición, la calidad de las medidas y la medida de la precisión analítica. Resulta sorprendente la cantidad de laboratorios que dan soporte analítico a procesos productivos de gran impacto económico en la vida de la empresa, sin realizar nunca un autocontrol sobre el nivel de precisión de sus análisis. En este capítulo se hace una presentación básica del tema con el objetivo de que resulte útil y práctica.

El **capítulo 9** presenta una de las aplicaciones más importantes de la estadística en el entorno industrial, el **control estadístico de procesos**. Dada la importancia de este capítulo, se refuerza su contenido con numerosos ejemplos y casos prácticos, y se incluye un caso extenso para su análisis.

En el **capítulo 10** se hace una introducción al **diseño de experimentos**. La utilidad de esta técnica es primordial para el industrial, sobre todo para el área de I+D y el diseño de productos. Dado que esta técnica puede ser muy compleja en su aplicación real, se facilitan enlaces a otros recursos, como cursos, que serán útiles a los que quieran profundizar más.

Finalmente, en el **capítulo 11** se presenta el uso de los conceptos y técnicas estudiadas en los procesos de mejora de la calidad

industrial, y se introducen algunas aplicaciones prácticas de la estadística en el entorno industrial, como **Six Sigma**.

Cómo usar el libro

He intentado que cada capítulo sea lo más autocontenido posible de forma que se facilite la organización pedagógica por temas. No obstante, hay algunos contenidos que pueden necesitar contenidos de los capítulos anteriores, por lo que se sugiere estudiarlo en el orden presentado.

El libro es eminentemente práctico, con numerosos ejercicios; su resolución puede ser individual o en equipo.

Algunos recuadros utilizan códigos de color para indicar el objetivo de la información que contienen. Básicamente, los colores utilizados son:

Problema o cuestión a resolver

El recuadro azul se utilizará para proponer problemas sencillos cuya respuesta se encuentra más adelante en el texto. El objetivo de estos problemas es estimular la reflexión, aunque puede ser necesario recurrir a cálculos sencillos ayudados por las herramientas disponibles.

Respuesta al problema o cuestión a resolver

El recuadro verde se utilizará para proponer una respuesta al problema planteado; respuesta que no tiene por qué ser la única posible. Normalmente se presentará de forma oculta.

Además se incluyen diferentes tipos de avisos cada vez que se introduce algún concepto que es necesario resaltar.

! Importante

En este formato se indican cuestiones importantes

> ⚠ ¡Atención!
>
> En este formato se indican cuestiones a las que hay que prestar especial atención o que pueden inducir a error

Uso del ordenador y el *software* estadístico

En la práctica diaria, los técnicos industriales usan los ordenadores para almacenar y visualizar los datos de producción, para solucionar problemas mediante análisis estadísticos, y para presentar sus resultados de forma gráfica. De la misma forma que en el entorno industrial, en este libro se utilizarán también los ordenadores de forma habitual, y por esta razón es imprescindible que los estudiantes tengan acceso individual a un ordenador en el que esté instalado el software recomendado, y que se acostumbren a utilizarlo para resolver los problemas y casos planteados como ejercicios prácticos, individualmente y en grupo.

El estudiante que se incorpora a una empresa, sea en un laboratorio o en una planta de producción, se va a encontrar muy pronto delante de una hoja de cálculo, y debe saber cómo utilizarla correctamente. Actualmente, lo más probable es que esa hoja de cálculo sea Microsoft Excel, aunque hay otras alternativas posibles, como Google Sheets, Apple Numbers, OpenOffice Calc y algunas más. La gran dominancia en el mercado de Microsoft Excel ha hecho que todas estas herramientas sean totalmente compatibles o tengan modos de compatibilidad con Excel. Por esta razón, este libro se basa en la utilización de Excel como hoja de cálculo y herramienta principal para el almacenamiento de datos.

A lo largo del libro se presentarán informes y gráficos obtenidos con Microsoft Excel, y también con el software estadístico R. Prácticamente todos ellos pueden ser exportados a otras herramientas, como Google Sheets, OpenOffice, Minitab o Matlab, o analizarse con otros lenguajes de programación, como Python o Julia. En realidad, el método de análisis y cómo obtener un resultado correcto son aspectos más importantes que la herramienta que se utilice para ello, por lo que queda en manos del

instructor la decisión final sobre qué usar y cómo. Para facilitar este trabajo de conversión, en su caso, todo el material del libro y los datos de ejemplo estarán disponibles en un repositorio de GitHub.

Algunos ejercicios tienen que ver con la interpretación y presentación de los resultados. Es importante que estos trabajos se realicen en grupo y se haga énfasis en la comprensión del problema y en su correcta exposición; en los equipos industriales de hoy, la discusión de problemas y la exposición de resultados, en reuniones de trabajo o en paneles informativos a pie de planta, forma parte del trabajo diario. Estas habilidades de comunicación deben ser desarrolladas en los estudiantes de forma prioritaria.

La ventaja de R sobre Excel es que el código R, si está bien documentado, muestra cada paso realizado, y esto permite que otras personas puedan verificar el resultado y reproducirlo a partir de los datos originales, e incluso reutilizar los procedimientos. Utilizar código en vez de clicks de ratón es esencial para asegurar la **reproducibilidad de los análisis de datos**[1]. Por esta razón, recomiendo el uso del lenguaje R como complemento o alternativa a la hoja de cálculo, tanto para analizar como para visualizar datos. Sin embargo, como la realidad del mundo de la empresa es que los lenguajes como R están todavía poco introducidos, es inevitable mantener el uso de la hoja de cálculo; en el libro se explicarán algunas *mejores prácticas*, que permitirán el uso simultáneo de ambas herramientas de forma óptima.

Respecto a la **programación informática**, en el libro no se hace énfasis en la programación R más que como sucesión de órdenes individuales en scripts sencillos. No se busca la eficiencia computacional ni la rapidez en el cálculo, sino la comprensión de la metodología de resolución de problemas y cómo ésta se apoya en las herramientas presentadas. De la misma manera, tampoco se hace ningún uso de la programación en Excel, ya sea con **macros** o con **Visual Basic**; estos temas quedan fuera del perímetro de este libro.

[1] El concepto de reproducibilidad, cada vez más importante, se desarrolla en el capítulo 2

Un paso en la dirección de la implantación de **flujos de trabajo reproducibles** es la elaboración de informes automatizados. Estos informes incluyen el código R, los comentarios del autor en forma de texto formateado en *markdown*, y los resultados del código. Herramientas como Quarto, o Google Colaboratory, que usa la interface Jupyter, son nuevas formas de elaborar y presentar los informes y resultados estadísticos. En el entorno docente, estas herramientas abren posibilidades muy interesantes en la presentación de un ejercicio o un exámen escrito, ya que el alumno puede detallar perfectamente todos los pasos hasta llegar al resultado final, y facilita la revisión por sus compañeros o por el profesor a cargo de la asignatura.

Recursos adicionales y cómo usarlos

En este libro no se hace una introducción a R ni a Excel; se presupone que el alumno tiene un conocimiento básico de ambas herramientas. Si no tiene ninguna formación sobre el lenguaje R y el entorno RStudio recomiendo hacer alguna formación previa sencilla que introduzca los conceptos básicos. Datacamp tiene cursos gratuitos de introducción a *R*; también hay cursos de formación tanto de R como de Excel en otras plataformas web como edX, Udemy y Coursera, muchos de ellos gratuitos. El Gobierno de España, dentro de una de sus iniciativas de transformación digital, la iniciativa de datos abiertos, incluye también una amplia referencia a cursos de formación sobre R.

Todos los datos presentados en los ejemplos se incluyen en hojas de cálculo que están disponibles en GitHub. También se incluyen fuentes de datos adicionales que pueden permitir plantear nuevos ejercicios.

Al final del libro se incluye una bibliografía completa.

Sobre el libro

El libro ha sido editado en Quarto. Está disponible en PDF.

Agradecimientos

1 Introducción

Figura 1.1: Un grupo de trabajo en planta ante un panel de variables indicadoras

La estadística es la ciencia de aprender a partir de los datos. Implica la recolección, análisis y presentación de los datos, y su utilización para tomar decisiones y resolver problemas.

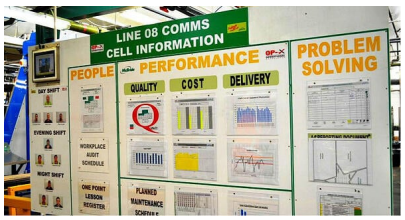
Hay muchos aspectos del trabajo industrial que implican recoger datos, trabajar con ellos y utilizarlos para resolver un problema; el uso de la estadística es sólo una herramienta más, tan importante como cualquier otra disciplina en el bagaje de conocimientos de un científico, ingeniero o técnico industrial.

Los métodos estadísticos nos ayudan a describir y comprender la **variabilidad**. Cuando hablamos de variabilidad queremos

decir que sucesivas observaciones de un mismo proceso o sistema no dan exactamente los mismos resultados. Por ejemplo, el consumo de gasolina de un coche no es siempre igual, sino que varía de manera considerable. Esta variación depende de muchos factores, como la forma de conducir, el tipo de carretera, la situación del propio vehículo (presión de neumáticos, compresión del motor, ...), la marca de la gasolina, el octanaje, o incluso las condiciones meteorológicas. Todos estos factores son causas de **variabilidad** en el consumo de gasolina. La estadística nos permite analizar estos **factores** y determinar cuáles son los más importantes o tienen mayor impacto en el consumo; una vez conocidos, podemos actuar sobre ellos.

> **!** Importante
>
> El objetivo más importante de la mejora industrial es la **reducción de la variabilidad**.

En este libro aprenderemos a utilizar herramientas diversas, tanto estadísticas como de la ciencia de datos, para realizar nuestro análisis. Para aprender de los datos necesitamos más que los simples números; para interpretarlos necesitaremos siempre el conocimiento del proceso industrial que estamos analizando.En un análisis de la producción de un producto lácteo, por ejemplo, los números significan poco sin un conocimiento del proceso; los valores de pH, temperatura o concentración de lactosa influyen en el resultado del proceso de forma diferente. Los datos son números dentro de un contexto, y necesitamos conocer este contexto para dar sentido a los números.



1.1 Pensamiento estadístico y pensamiento crítico (*Critical Thinking*)

Los ingenieros y técnicos resuelven problemas de interés para la empresa y la sociedad mediante la aplicación de los principios del método científico, siguiendo estos pasos:

1. Preparar una descripción clara y concisa del problema

2. Identificar, al menos de forma tentativa, los principales factores que afectan al problema, o que podrían tener un papel en su resolución
3. Proponer un modelo para el problema, usando conocimiento científico o tecnológico del proceso en estudio, dejando constancia de las limitaciones del modelo propuesto.
4. Realizar experimentos apropiados y recolectar datos para probar o validar el modelo tentativo o las conclusiones previas obtenidas en los pasos 2 y 3
5. Refinar el modelo sobre la base de los datos observados
6. Manipular el modelo para desarrollar una solución al problema
7. Realizar un experimento adecuado para confirmar que la solución propuesta es efectiva y eficiente.
8. Sacar las conclusiones oportunas o hacer recomendaciones basándose en la solución encontrada.

The definition of "critical thinking" used for construction of the Halpern Critical Thinking Assessment characterizes critical thinking as those cognitive skills or strategies that increase the probability of a desirable outcome. They are purposeful, reasoned, and goal directed. Critical thinking is the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions. Critical thinkers use these skills appropriately, without prompting, and usually with conscious intent, in a variety of settings. That is, they are predisposed to think critically. When we think critically, we are evaluating the outcomes of our thought processes--how good a decision is or how well a problem is solved. Critical thinking also involves evaluating the thinking process--the reasoning that went into the conclusion we've arrived at or the kinds of factors considered in making a decision. Therefore, critical thinking has to be regarded as a hierarchical multidimen-sional construct comprising the facets verbal reasoning, argument analysis skills, skills in thinking as hypothesis testing, using likelihood and uncertainty, and decision making/problem solving skills (Halpern, 1994; 1998; 2003). (APA PsycTests Database Record (c) 2019 APA, all rights reserved)

HCTA Halpern Critical Thinking Assessment (apa.org)

1.1.1 What is critical thinking?

> Critical thinking is the art of making clear, reasoned judgements based on interpreting, understanding, applying and synthesising evidence gathered from observation, reading and experimentation.
>
> Burns, T., & Sinfield, S. (2016) Essential Study Skills: The Complete Guide to Success at University (4th ed.) London: SAGE, p94.

Being critical does not just mean finding fault. It means assessing evidence from a variety of sources and making reasoned conclusions. As a result of your analysis you may decide that a particular piece of evidence is not robust, or that you disagree with the conclusion, but you should be able to state why you have come to this view and incorporate this into a bigger picture of the literature.

Being critical goes beyond describing what you have heard in lectures or what you have read. It involves synthesising, analysing and evaluating what you have learned to develop your own argument or position.

Critical thinking is important in all subjects and disciplines – in science and engineering, as well as the arts and humanities. The types of evidence used to develop arguments may be very different but the processes and techniques are similar. Critical thinking is required for both undergraduate and postgraduate levels of study.

1.1.2 What, where, when, who, why, how?

Purposeful reading can help with critical thinking because it encourages you to read actively rather than passively. When you read, ask yourself questions about what you are reading and make notes to record your views. Ask questions like:

- What is the main point of this paper/ article/ paragraph/ report/ blog?
- Who wrote it?
- Why was it written?
- When was it written?
- Has the context changed since it was written?
- Is the evidence presented robust?
- How did the authors come to their conclusions?
- Do you agree with the conclusions?
- What does this add to our knowledge?
- Why is it useful?

1.1.3 Developing an argument

Being a university student is about learning how to think, not what to think. Critical thinking shapes your own values and attitudes through a process of deliberating, debating and persuasion. Through developing your critical thinking you can move on from simply disagreeing to constructively assessing alternatives by building on doubts.

There are several key stages involved in developing your ideas and constructing an argument. You might like to use a form to help you think about the features of critical thinking and to break down the stages of developing your argument.

https://elpais.com/economia/estar-donde-estes/2021-03-24/como-aplicar-el-pensamiento-critico-en-tu-trabajo.html

Statistical and numerical data

1.2 Algunas definiciones importantes

1.2.1 Población y muestra

Una **población** es un conjunto de de personas, cosas o, en general, objetos en estudio. A veces, una población es demasiado grande para que podamos abarcarla completa; para poder estudiarla, obtenemos una **muestra**, que consiste en un subconjunto de la población que hemos seleccionado para su estudio. El proceso de obtener una muestra se llama **muestreo**, y se realiza de acuerdo con normas y procedimientos específicos.

En muchas ocasiones, cuando se recogen los datos como resultado de una experimentación, definimos la población como *todos los resultados que podríamos haber obtenido*. Llamamos a este conjunto de posibles resultados una **población conceptual**. Por ejemplo, cuando medimos el pH de varias muestras de leche, la población es el conjunto de todos los resultados posibles que podríamos haber tenido. Muchos problemas de ingeniería y tecnología se refieren a poblaciones conceptuales.

> **!** Recuerda
>
> En la mayoría de las ocasiones, nuestros datos provienen de una **muestra** obtenida de una **población**,

Cuando tomamos una muestra, debemos estar seguros de que contiene las propiedades que queremos estudiar en la población. En ese caso, decimos que la muestra es **representativa**: los individuos de la muestra son representativos de la población. Para que la muestra sea representativa, debe ser obtenida mediante un **muestreo aleatorio**. Una **muestra aleatoria simple** de tamaño n consiste en n individuos de una población, elegidos de forma que cada conjunto posible de n individuos tiene la misma **probabilidad** de ser elegido

> **!** ¿Qué es la probabilidad?
>
> Introduciremos el concepto de probabilidad con detalle en el capítulo 4

1.2.2 Parámetro y estadístico

Llamamos **estadístico** a un número que representa una propiedad o característica de la muestra. Un **parámetro** es una característica de la población, que podemos estimar a partir de la muestra mediante la obtención de un **estadístico muestral**.

1.2.3 Variables y casos

A los objetos descritos en un conjunto de datos los llamamos **casos**, de forma genérica. A veces, estos casos pueden corresponder a personas; en ese caso podemos llamarlos **individuos**. Cuando los objetos que estudiamos no son personas, como es lo habitual en el entorno industrial, utilizamos la nomenclatura genérica.

Un **atributo** es una característica que define una propiedad de un objeto, persona o cosa. Por ejemplo, edad, peso, altura, sexo, color de ojos, son atributos de una persona. Llamamos **variable** a una característica cualquiera de un individuo que puede ser medida. Una variable puede tomar diferentes valores en diferentes individuos o casos.

Según estas definiciones que acabamos de ver, una muestra está formada por un conjunto de casos, y cada caso contiene un determinado número de variables, que contienen los valores que hemos analizado o medido.

> Ejemplo 1: Muestreando una cámara de maduración de queso
>
> Imagínate que tienes que analizar el extracto seco de una producción de queso que está en fase de maduración en una cámara. Como la cámara está muy llena, es difícil acceder al interior, y decides coger tu muestra de los quesos que están más a tu alcance, justo al lado de la puerta y a la altura de la vista.¿Crees que es una buena idea? ¿Podrías definir la población en este caso?.

Respuesta al ejemplo 1: Muestreando una cámara de maduración de queso

No es una buena idea porque no tenemos garantía de que las condiciones de humedad,temperatura y circulación de aire sean las mismas en toda la cámara. Para asegurar que nuestra muestra es representativa, debemos tomar una **muestra aleatoria** de la población, que en este caso es el total de quesos en la cámara.

1.2.4 Tipos de variables

Algunas variables, como el color, sirven para clasificar los individuos en categorías. Otras, como la altura o el peso de un individuo, pueden tomar valores numéricos con los que podemos hacer cálculos. Por ejemplo, podemos sumar la altura de varias personas, pero no tiene sentido *sumar* los colores del arcoiris (aunque sí podemos *contarlos*, y hacer cálculos con estos recuentos). También podemos *categorizar* variables continuas: podemos clasificar nuestro grupo de personas en *altas* o *bajas*, y podemos contar cuántas personas entran en cada categoría.

Variables cualitativas o categóricas		Variables cuantitativas o métricas	
Nominales	**Ordinales**	**Discretas**	**Continuas**
Valores en categorías arbitrarias	Valores en categorías ordenadas	Valores enteros en escala numérica	Valores continuos en escala numérica
(sin unidades)	(sin unidades)	Unidades contadas	Unidades medidas

Una **variable categórica** coloca a un individuo en uno o más grupos o categorías

Una **variable métrica** toma valores numéricos con los que tiene sentido realizar cálculos aritméticos como sumar, restar, etc.

Variables cualitativas o categóricas		Variables cuantitativas o métricas	
Nominales	**Ordinales**	**Discretas**	**Continuas**
Valores en categorías arbitrarias	Valores en categorías ordenadas	Valores enteros en una escala numérica	Valores continuos en una escala numérica
(sin unidades)	(sin unidades)	Unidades contadas	Unidades medidas

Las variables categóricas se conocen también como **variables cualitativas** porque indican *cualidades*.

Las variables métricas se conocen también como **variables cuantitativas** porque indican *cantidades*.

⚠ Comentario: ¿Cualitativo quiere decir "que tiene calidad"?

A veces se utiliza la palabra **cualitativo** de forma incorrecta para indicar **calidad**, por ejemplo cuando alguien dice: "Este envase es muy **cualitativo**". Deberíamos decir "Este envase tiene gran calidad". **Cualitativo** no se deriva de **calidad**, sino de **cualidad**.

1.2.5 Ejemplos de variables

Para resolver

Ejemplo 1. Tiramos un dado al aire. Describe a qué corresponde la variable y el caso.
Ejemplo 2. Durante un proceso de envasado de un producto que dura una hora, controlamos el peso de cada envase cada minuto. Describe la variable y el caso. ¿Puede haber más de una variable?

Respuestas: Para resolver

Ejemplo 1: La **variable** es el resultado que obtenemos cada vez; podríamos denominarla, por ejemplo, *resultado_obtenido*. Colocaríamos este nombre en el encabezado de una columna en una hoja de cálculo. Cada tirada que hacemos es un **caso**; iríamos colocando el resultado que obtenemos cada vez en una nueva fila de nuestra hoja de cálculo.
Ejemplo 2. En este caso, la **variable** es el *peso_obtenido*, y cada pesada constituye un **caso**. Si registrásemos, además, la hora y el minuto en el que que hemos hecho cada control de peso, podríamos definir una nueva variable, que podríamos llamar *hora*, y que colocaríamos en una columna al lado del *peso_obtenido*. Incluso podríamos definir otra variable adicional, el *numero_de_pesada*, que sería un número secuencial empezando en 1 y que se incrementaría en cada pesada, de forma que al final esta variable nos daría el número de pesadas realizadas, y nos indicaría además el orden en el que las hemos realizado. Puesto que hemos realizado una pesada cada minuto, tendríamos tres variables y 61 líneas (un encabezado y 60 líneas correspondientes una a cada minuto)

2 Herramientas para el análisis de los datos industriales

2.1 Las hojas de cálculo.

La hoja de cálculo es una herramienta omnipresente hoy día en todos los ámbitos de trabajo y educativos. Desde la aparición de Visicalc, en 1978, la hoja de cálculo ha contribuido a la gestión de miles de empresas, se ha utilizado de manera general en análisis de datos y sus gráficos se han utilizado y se utilizan en publicaciones e informes de todas clases. En la década de los años 80 del pasado siglo, la hoja de cálculo Lotus 1-2-3 fue la aplicación más utilizada en los ordenadores IBM-PC y compatibles, y consiguió facturaciones millonarias para la empresa matriz. Lotus 1-2-3 dominó el mercado hasta la aparición de Microsoft Windows a finales de los años 80; el nuevo sistema operativo favoreció la implantación de Excel, que desde entonces se convirtió en la hoja de cálculo dominante.

2.2 El software estadístico R

2.3 Almacenar datos en una hoja de cálculo

Las **hojas de cálculo** son muy útiles para recoger la información de un conjunto de observaciones. De la misma manera que la gramática permite ordenar y estructurar un escrito de acuerdo a reglas comunes, veremos que hay reglas para que el almacenamiento de los datos sea lo más homogéneo posible y se reduzcan los errores al mínimo.

En este libro trataremos exclusivamente de lo que llamaremos **datos rectangulares**: grupos de valores que están asociados a

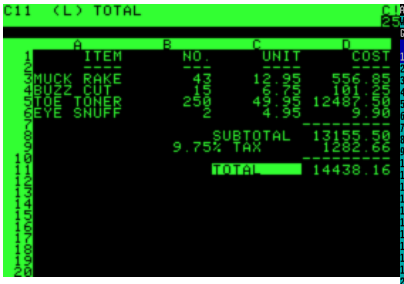


Figura 2.1: Visicalc, primera hoja de cálculo para el ordenador *Apple II* (1979)

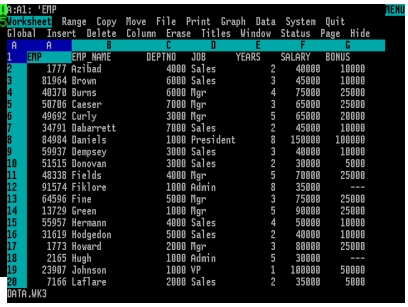


Figura 2.2: Hoja de cálculo Lotus 1-2-3 para MS-DOS (1983)

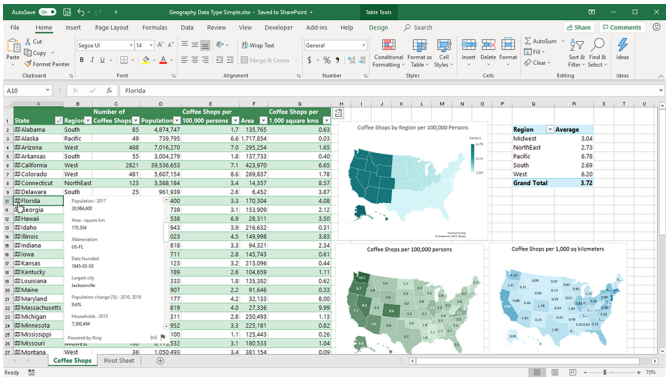


Figura 2.3: Microsoft Excel (2023)

una o más variables, y a varias observaciones. Hay muchos más datos que no se ajustan a este paradigma: imágenes, sonidos, archivos documentales de texto. Pero la forma más común de almacenar datos industriales es la de las tablas rectangulares; vamos a aprender cómo organizarlas correctamente.

2.3.1 Preparación de los datos

Los datos se pueden recoger y guardar de múltiples formas. Cuando nos incorporamos a un equipo de trabajo existente, lo más seguro es que el equipo disponga ya de un sistema de archivo de los datos, de acuerdo con sus prácticas habituales.

Cuando la recogida de datos se hace de forma manual en papel, es necesario registrar en el ordenador los datos recogidos. Lo más frecuente es que este registro se haga en hojas de cálculo, como Microsoft Excel o Google Sheets. En algunos casos, el almacenamiento se hace sobre bases de datos, genéricas o desarrolladas a medida.

Actualmente, la tendencia es recoger los datos o bien de forma automática, o bien de forma manual sobre sistemas informatizados (pantallas), lo que permite eliminar el papel y disponer directamente de los datos en un formato digitalizado.

Los equipos y líneas de producción diseñados actualmente (IoT) se interconectan con los sistemas de información y almacenan en tiempo real todos los datos necesarios, lo que libera al operario de la pesada tarea de reintroducirlos manualmente, a la vez que reduce los errores debidos a la imputación incorrecta.

En todos los casos, es imprescindible asegurar que los sistemas de información pueden exportar a ficheros de texto tipo *fichero plano* o tipo CSV, de forma que podamos importarlos tanto a Excel como a R, como veremos más adelante. Estos sistemas de exportación de datos deben diseñarse de forma flexible y abierta, para que tanto la captura como la exportación puedan modificarse y adaptar la recogida de la información a las necesidades de cada momento.

2.3.2 Diseño de la captura de información

A veces el diseño de la captura de datos sigue aproximadamente el modelo manual en papel. Se introducen los datos en la hoja de cálculo y una vez completados, se imprime el documento para su archivo.

El error más común que cometemos es tratar la hoja de cálculo como un bloc de notas, es decir, hacer anotaciones de forma libre, colocar los datos y el resultado de los análisis al lado y en cualquier parte de la hoja, y apoyarnos en el contexto para interpretar lo que hemos guardado. Pero para que el ordenador sea capaz de analizar nuestros datos de manera eficiente, debemos estructurarlos de tal forma que el programa use la información tal como nosotros queremos.

Es común utilizar una hoja para guardar múltiples tablas de datos, tal como vemos en la Figura **??**. Esta estructura, sin embargo, resulta enormemente confusa para su análisis, o lo imposibilita completamente.

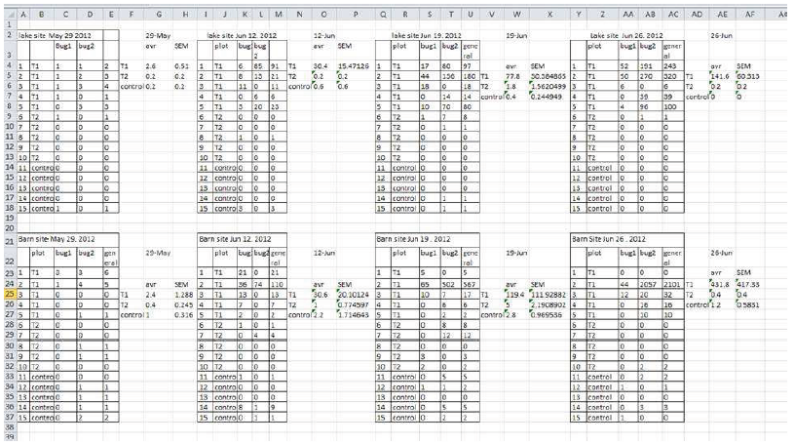
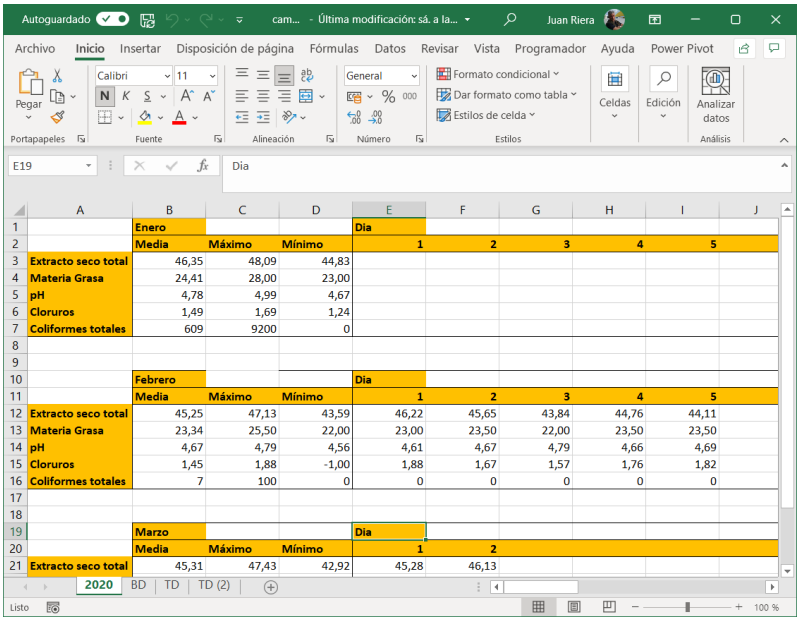


Figura 2.4: Hoja Excel desordenada: ¡No hagas esto!

En otros casos, los datos se guardan en hojas de cálculo que se componen de diferentes pestañas para cada semana, cada mes o cada año, como vemos en la Figura **??**. Sin embargo, esta forma de almacenar los datos tampoco es la óptima para su análisis.



	A	B	C	D	E	F	G	H	I	J
1		Enero			Dia					
2		Media	Máximo	Mínimo	1	2	3	4	5	
3	Extracto seco total	46,35	48,09	44,83						
4	Materia Grasa	24,41	28,00	23,00						
5	pH	4,78	4,99	4,67						
6	Cloruros	1,49	1,69	1,24						
7	Coliformes totales	609	9200	0						
8										
9										
10		Febrero			Dia					
11		Media	Máximo	Mínimo	1	2	3	4	5	
12	Extracto seco total	45,25	47,13	43,59	46,22	45,65	43,84	44,76	44,11	
13	Materia Grasa	23,34	25,50	22,00	23,00	23,50	22,00	23,50	23,50	
14	pH	4,67	4,79	4,56	4,61	4,67	4,79	4,66	4,69	
15	Cloruros	1,45	1,88	-1,00	1,88	1,67	1,57	1,76	1,82	
16	Coliformes totales	7	100	0	0	0	0	0	0	
17										
18										
19		Marzo			Dia					
20		Media	Máximo	Mínimo	1	2				
21	Extracto seco total	45,31	47,43	42,92	45,28	46,13				

Figura 2.5: Hoja Excel con una estructura no ordenada

¿Y utilizar diferentes pestañas para cada tabla? En este caso, la respuesta es sí y no. Si las diferentes tablas presentan situaciones diferentes, o datos que no son coincidentes, podemos utilizar diferentes pestañas. Pero si los datos están vinculados, por ejemplo, se corresponden con medidas hechas en fechas diferentes (meses, años), la respuesta adecuada es que las pestañas no son la forma correcta de almacenarlos datos; la forma recomendad es añadir una variable que nos permita diferenciar los datos por fecha; nuestro programa de análisis nos permitirá *filtrar* los datos según la fecha que deseemos, y todos estarán en una única tabla, facilitando la coherencia del conjunto.

Hay muchas formas de almacenar la información en una hoja de cálculo, pero hay una forma que facilita la utilización de los datos tanto por la hoja de cálculo como por otros programas de análisis, A esta forma de almacenar las tablas de datos la llamamos **datos ordenados** (*tidy data*)(Wickham 2014)

2.4 Los datos ordenados (*tidy data*)

Las reglas principales al almacenar nuestros datos en una hoja de cálculo es que columnas=variables, filas=observaciones, celdas=valores. Estas tres reglas básicas son las que hacen que nuestro conjunto de datos sea ordenado (Hadley Wickham 2017):

1. Cada variable debe tener su propia columna.
2. Cada observación debe tener su propia fila.
3. Cada valor debe tener su propia celda o casilla .

La Figura **??** muestra estas reglas de forma visual.

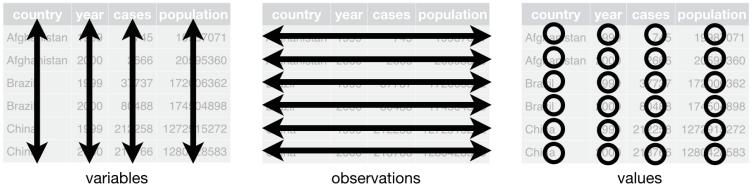
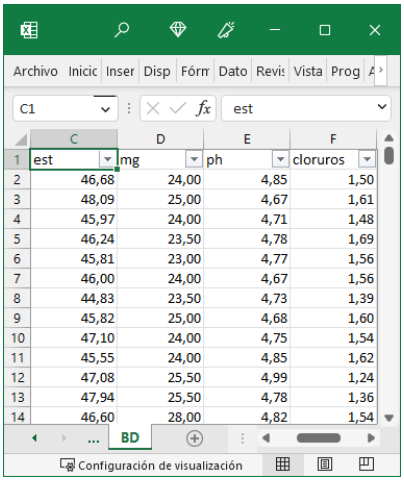


Figura 2.7: Sigue estas tres reglas para tener un conjunto de datos ordenado: las variables están en columnas, las observaciones están en filas, y los valores están en celdas. Fuente de la imagen: [@wickham2017]

Estas tres reglas están interrelacionadas porque es imposible satisfacer sólo dos de tres.



	C	D	E	F
1	est	mg	ph	cloruros
2	46,68	24,00	4,85	1,50
3	48,09	25,00	4,67	1,61
4	45,97	24,00	4,71	1,48
5	46,24	23,50	4,78	1,69
6	45,81	23,00	4,77	1,56
7	46,00	24,00	4,67	1,56
8	44,83	23,50	4,73	1,39
9	45,82	25,00	4,68	1,60
10	47,10	24,00	4,75	1,54
11	45,55	24,00	4,85	1,62
12	47,08	25,50	4,99	1,24
13	47,94	25,50	4,78	1,36
14	46,60	28,00	4,82	1,54

Figura 2.6: Hoja Excel con estructura rectangular de datos ordenados

2.5 Los nombres de las variables

Según hemos visto, existen diferentes tipos de variables, **cualitativas** (*categóricas*) y **cuantitativas** (*métricas*). Normalmente, los valores de las variables categóricas se describen mediante textos del tipo “color blanco”, “hombre”, “mujer”, “alto”, “bajo”, etc. Suelen corresponder con características descriptivas, y por lo tanto, no puede hacerse cálculos directamente con ellos, a menos que se hayan resumido, por ejemplo, mediante un conteo. Las variables métricas consisten en valores numéricos, que

pueden ser **enteros** (1;24;350) o **continuos** (1, 456;0, 35) y que sí pueden utilizarse directamente para hacer cálculos tales como sumas, etc.

Una **variable** está descrita siempre por **un nombre**, que designa la variable, y **un valor o conjunto de valores**, que corresponden a los casos. Este conjunto de valores, como acabamos de ver, pueden ser textos o números.

Ejemplos de valores de texto: "Carlos", "fruta", "Lluvia fuerte", "muy ácido", "sabor a fresa"

Ejemplos de valores numéricos: 1; 7; 10, 65

Existen también otros tipos de variables que veremos más tarde, como variables lógicas o fechas, según el tipo de dato que almacenemos en esa variable.

Siempre que sea posible, utilizaremos el nombre del atributo o característica que estamos midiendo o analizando, o su abreviatura, para designar una variable; por ejemplo, si estamos recogiendo la altura de una serie de personas, llamaremos `altura` a la variable; si estamos recogiendo el peso, usaremos el nombre `peso`, etc.

En una hoja de cálculo, colocaremos el nombre de la variable en la primera fila, e iremos añadiendo los valores debajo, un valor por línea.

A veces, asignar un nombre a una variable no es todo lo fácil que podría parecer a simple vista. Por ejemplo, ¿qué nombre daríamos a una variable que va a recoger los valores de pH de la leche en una cuba de queso en el momento de añadir el cuajo? Está claro que pH no es suficiente, porque en el proceso hay varias medidas de pH y sería bueno que pudiésemos diferenciarlas con facilidad. En un caso como éste, es probable que necesitemos utilizar varias palabras o abreviaturas que describan mejor el nombre de la variable.

	A	B
1		altura_cm
2	Luis	153
3	Ana	135
4	Iván	140
5	Lucía	140
6	Jessica	175
7	Antonio	138
8	Mikel	145
9	Marta	154
10	Carmen	152
11	Javier	159
12	María	154
13		

Para la construcción correcta de estos nombres, se han establecido un conjunto de normas, con el objetivo de evitar errores y facilitar el intercambio de los datos entre diferentes programas de análisis.

2.5.1 Reglas para los nombres de las variables

Las hojas de cálculo admiten que introduzcamos cualquier texto en una celda; no hay prácticamente ninguna limitación a los

nombres que podemos usar para nuestras variables. Excel utilizará los nombres con cualquier carácter sin inconvenientes.

Sin embargo, otros programas informáticos, entre ellos R, son mucho más restrictivos. Por esta razón, estableceremos una serie de reglas para construir los nombres de variables, que aplicaremos a nuestras tablas de Excel, y que nos permitirán intercambiarlas con otros programas, como R, con toda seguridad.

1. Un nombre válido consiste en una combinación de letras, números y signo de subrayado (_)
2. Un nombre de variable no puede empezar por un número, un punto o un signo de subrayado (_); debe empezar siempre por una letra.
3. Los nombres de variables irán siempre en minúsculas. Según esta regla, $Peso$ no es un nombre válido, pero $peso$ si lo es.

En R, las mayúsculas son significativas es decir, en R, $peso$ y $Peso$ son nombres diferentes. Por esta razón, aunque el uso de mayúsculas está permitido, nosotros adoptaremos las minúsculas de forma general)

4. No utilizaremos espacios en blanco, acentos ni caracteres especiales como $\tilde{n}$, %, guiones o paréntesis.
5. Hay veces en que nos interesa unir varias palabras para construir un nombre de variable. Se utilizan diferentes formas de unir palabras, por ejemplo:

   - un punto, como en $peso.en.cm$,
   - lo que se ha llamado *escritura de camello* (*camelCase*), que se llama así por el uso de mayúsculas y minúsculas mezcladas ($PesoEnCm$)
   - el signo de subrayado _, como en $peso\_en\_cm$

   Algunas de estas opciones son utilizadas en distintas comunidades de usuarios, por ejemplo la opción 1 es utilizada en la guía de estilo de Google, y la opción 2 es muy utilizada por los programadores del entorno de los lenguajes de Microsoft. Nosotros utilizaremos el signo de subrayado (_), que es la forma más usada en el entorno de programación de R.

6. Siempre se separarán las palabras mediante el signo de subrayado (_) para facilitar la lectura. Así, aunque *temperatura*1 es un nombre válido, preferiremos *temp_*1; es más corto y de lectura más clara. Igualmente, preferiremos *peso_empaquetado* a *pesoempaquetado*
7. Mantendremos los nombres razonablemente cortos para facilitar la lectura. Aunque podemos hacer los nombres todo lo largos que queramos, es más cómodo utilizar nombres cortos. Por ejemplo, podríamos utilizar *temperatura_de_la_leche_al_cuajar*, pero preferiremos abreviarlo como *temp_cuajo*.

Nombres no válidos:

- *peso en gramos* (contiene espacios)
- *pH_de_la_leche_en_Recepción* (demasiado largo, tiene un acento, tiene mayúsculas)
- *extracto_seco_total_a_la_salida_de_la_salmuera* (demasiado largo)

Alternativas válidas:

- *peso_g*
- *pH_leche_rec* (en este caso, de manera excepcional, podemos mantener el uso de la mayúscula por corrección formal)
- *est_salida_sal*

Un caso particular es el uso de la $\tilde{n}$, ya que no hay una alternativa fácil para el uso en las fechas (*año*). R admite el uso de la $\tilde{n}$ en los nombres de variables, por lo que podremos usarlo con cuidado, poniendo atención a los posibles errores que se pudiesen producir en algunas librerías.

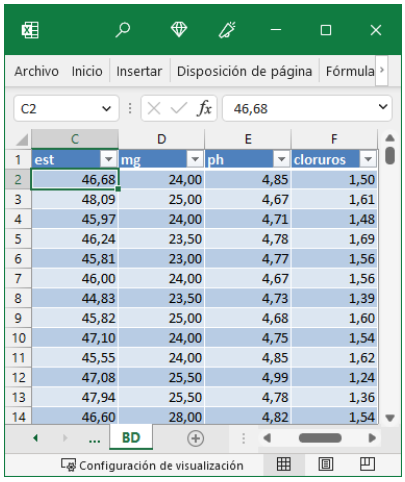
2.6 Para resolver

Poner aquí distintos ejemplos de nombres de variables para verii son válidos o no Describir medidas y preguntar cómo llamaríamos a esa variable (por ejemplo, temperatura de laleche que acabamos de descargar de una cisterna)

2.7 Datos rectangulares en Excel

La estructura de **datos ordenados** nos lleva a almacenar nuestros datos en tablas con estructura rectangular. La mejor forma de manejar los datos en Excel es convertir esta estructura en una **tabla**, para ello utilizaremos la opción `Menú>Insertar>Tabla`

Aunque en Excel no es fácil modificar esta estructura, R proporciona herramientas muy útiles que permiten intercambiar filas o columnas, lo que en ocasiones es muy útil en el análisis. Hadley Wickham (2017) proporciona métodos detallados para manejar tablas de datos ordenados.



	C	D	E	F
1	est	mg	ph	cloruros
2	46,68	24,00	4,85	1,50
3	48,09	25,00	4,67	1,61
4	45,97	24,00	4,71	1,48
5	46,24	23,50	4,78	1,69
6	45,81	23,00	4,77	1,56
7	46,00	24,00	4,67	1,56
8	44,83	23,50	4,73	1,39
9	45,82	25,00	4,68	1,60
10	47,10	24,00	4,75	1,54
11	45,55	24,00	4,85	1,62
12	47,08	25,50	4,99	1,24
13	47,94	25,50	4,78	1,36
14	46,60	28,00	4,82	1,54

Figura 2.8: Tabla Excel

2.8 De Excel a R

Una vez que tenemos nuestros datos en Excel, hay dos formas en las que podemos poner los datos a disposición de R para su análisis: exportarlos a un archivo de texto con ***formato CSV***, o leer directamente los datos de Excel desde R utilizando las funciones de la librería `tidyverse`. En ambos casos, el resultado en R es un ***dataframe*** o *cuadro de datos*, que es una estructura equivalente a la de nuestra tabla de datos en Excel.

2.8.1 Qué es un fichero plano y un fichero CSV

Se suele llamar **fichero plano** a un fichero de datos de texto sin ningún tipo de formato, donde los datos están separados por espacios o tabulaciones. Muchos equipos automáticos, como balanzas de laboratorio o básculas de proceso, producen ficheros planos de texto, que se pueden importar a Excel o R. Un fichero CSV es un fichero plano en el que los valores están separados por un carácter especial, llamado **separador**. Este separador puede ser una coma **,** (cuando los decimales se separan mediante un punto, como en EEUU) o un punto y coma **;** (cuando los decimales se separan mediante una coma, como en España)

En un fichero plano o en un fichero CSV, la primera fila puede contener los nombres de las columnas. En algunos casos, los

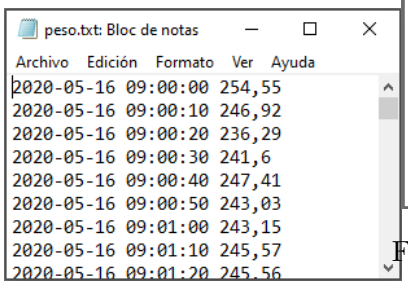


Figura 2.9: Fichero plano

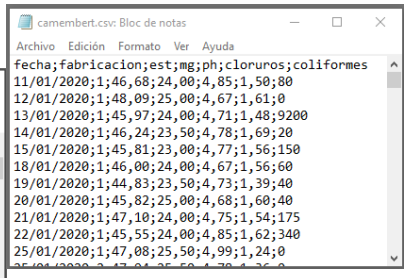


Figura 2.10: Fichero CSV separado por puntos y comas

Figura 2.11: Fichero CSV separado por comas

Figura 2.12: Tres tipos de ficheros planos de texto.

elementos de texto pueden estar entre comillas. En estos casos, los programas de importación se ocupan de la conversión de formatos.

La importación de un fichero CSV en Excel en español es directa si se ha generado con puntos y comas como separador y comas para los decimales; si no es así, nos aparecerá como un fichero plano de texto sin formato, y tendremos que realizar una conversión.

2.8.2 Cómo exportar los datos a un fichero CSV desde Excel.

2.9 La reproducibilidad de los análisis de datos

Literate programming - Wikipedia

Reproducible Research (hbiostat.org)

rr (hbiostat.org)

En el mundo científico y técnico cada vez cobra más importancia el concepto de **reproducibilidad de los análisis**, sobre todo cuando se trata de comunicar o publicar el resultado de un trabajo o de una investigación. Medios, como la prestigiosa revista *Science*, se han hecho eco de ello (Buck 2015). Por otra parte, la utilización de un flujo de trabajo basado en hojas de cálculo hace difícil garantizar esta reproducibilidad, y a veces puede llevar a cometer errores de consecuencias graves (Ferrero 2018; Ryssdal 2013).

Jesse Sadler (Sadler 2017) lo explica así:

> El peligro de la hoja de cálculo deriva de su propia estructura. La mezcla de entrada de datos, análisis y visualización hace que sea fácil confundir las celdas que contienen datos sin procesar con las que son el resultado del análisis. La forma de definir la lógica programática, tal como la selección de qué celdas se van a sumar, mediante clics del mouse, significa que una acción errónea de clic o arrastre puede

provocar errores o la sobreescritura de datos. Solo hace falta pensar en el pavor del momento en el que vas a cerrar una hoja de cálculo y el programa te pregunta si te gustaría guardar los cambios. Te hace preguntarte. ¿Quiero guardar? ¿Qué cambios hice? Debido a que la lógica en una hoja de cálculo se realiza a través de clics del mouse, no hay forma de rastrear de manera efectiva qué cambios se han realizado en una sesión o en la producción de un gráfico. Los errores cometidos con Excel pueden tener consecuencias graves, como se puso de manifiesto tras la controversia alrededor del artículo de Carmen Reinhart y Kenneth Rogoff sobre la deuda nacional de los EEUU.

Ciertamente hay razones legítimas por las que las personas usan por defecto hojas de cálculo para el análisis de datos en lugar de usar un lenguaje de programación como R. Las hojas de cálculo son mucho más atractivas y confortables de lo que cualquier lenguaje de programación podría ser para un recién llegado. Aprender a programar es intimidante y no es algo que se pueda hacer rápida o fácilmente. Las aplicaciones de interfaz gráfica de usuario (GUI) son mucho menos desalentadoras que una interfaz de línea de comandos. En segundo lugar, las hojas de cálculo son una buena herramienta para la entrada de datos, y es tentador pasar directamente al análisis de datos, manteniendo todo en el mismo documento. Finalmente, la naturaleza interactiva de las hojas de cálculo y la capacidad de crear gráficos que cambian en función de las entradas es muy atractiva, incluso si desbloquear completamente este potencial implica un conocimiento bastante complejo sobre cómo funciona el programa. La primera ventaja de las hojas de cálculo sobre la programación no se supera fácilmente, pero las dos últimas se basan en lo que creo que es un flujo de trabajo problemático. En lugar de usar un par de aplicaciones monolíticas, a menudo un conjunto de aplicaciones de oficina, para hacer todo, creo que es mejor divi-

dir el flujo de trabajo entre varias aplicaciones que hacen una cosa bien.

Crear una división clara entre la entrada y el análisis de datos es una de las principales razones por las que el análisis de datos en un lenguaje de programación es preferible al software de hoja de cálculo. Todavía uso hojas de cálculo, pero su limito su uso estrictamente a la entrada de datos. En un programa de hoja de cálculo, el análisis manipula directamente la única copia de los datos sin procesar. Por el contrario, con R se importan los datos, creando un objeto que es una copia de los datos sin procesar. Todas las manipulaciones de los datos se realizan en esta copia, y los datos originales nunca se alteran de ninguna manera. Esto significa que no hay forma de estropear los datos sin procesar. La manipulación de una copia de los datos le permite experimentar más libremente. Los errores son intrascendentes, incluso aunque a veces puedan llegar a ser frustrantes. Una línea de código que devuelve un error se puede ajustar y volver a ejecutar, repitiendo el proceso las veces necesarias hasta que se devuelva el resultado esperado.

Trabajar en una copia de los datos sin procesar puede incluso simplificar el proceso de entrada de datos. El análisis de datos tabulares en R da como resultado la creación de múltiples objetos, que se conocen como *data frames* y pueden considerarse equivalentes a tablas en una hoja de cálculo. La capacidad de dividir, muestrear y transformar el conjunto de datos original en muchos *data frames* diferentes tiene la ventaja de reducir drásticamente la complejidad de la entrada de datos. En lugar de necesitar hojas de cálculo a medida con múltiples hojas y tablas interrelacionadas, cada pieza de datos solo debe ingresarse una vez y todas las manipulaciones se pueden realizar en el código. Los diferentes *data frames* que se crean en el proceso de análisis ni siquiera tienen que ser guardados, porque son muy fácilmente

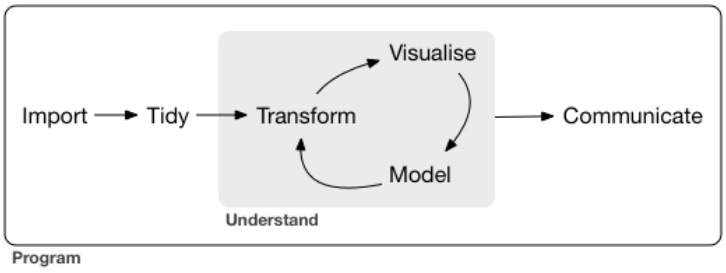
reproducidos por el script de código.

> La separación de la entrada y el análisis de los datos reduce en gran manera el potencial de errores, pero tal vez aún más significativamente, el uso de código para el análisis de datos permite la creación de investigaciones reproducibles que no son posibles en hojas de cálculo. [...] Con un lenguaje de programación, los pasos del análisis se pueden establecer claramente en el código [...] Guardar el análisis en código tiene el beneficio inmediato de que se puede volver a ejecutar fácilmente en cualquier momento que se agreguen nuevos datos. El código también se puede aplicar a un conjunto de datos completamente nuevo de una manera mucho más transparente que con las hojas de cálculo. El beneficio a largo plazo es que con el código todo el análisis se documenta en lugar de ocultarse detrás de los clics del mouse. Esto hace que sea más fácil revisar los propios análisis mucho después de haber terminado con ellos, así como que otros entiendan lo que se ha hecho y comprueben si hay errores.

2.10 El flujo de trabajo

- En qué consiste un flujo de trabajo en el análisis de datos

  Gráfico de R4DS



- Cómo estructurar un flujo de trabajo para el análisis de datos industriales

2.11 El uso de la hoja de cálculo:

- Introducción y almacenar datos
- Tablas dinámicas
- Gráficos y edición gráfica

2.12 El uso de R

- Exploración de datos (gráficos básicos)
- Manipulación de datos y exportación para su uso en Excel
- Análisis estadísticos
- Gráficos de control

Informes automatizados

3 Los datos industriales de producción

En el entorno industrial, los datos son recogidos casi siempre por uno de estos tres caminos:

- Estudio retrospectivo, basado en datos históricos
- Estudio observacional
- Experimento diseñado

Un buen sistema de recogida de datos facilitará el estudio posterior. Si ponemos poco cuidado en la toma de datos y en la forma de guardarlos, nos encontraremos después con problemas complicados de resolver en la fase de análisis o en la de interpretación , y, en algunos casos, estos problemas serán imposibles de resolver.

3.1 Estudios retrospectivos o históricos

Un **estudio retrospectivo o histórico** es el que utiliza una muestra o todos los datos históricos de un proceso, recogidos en el pasado durante un período determinado de tiempo. El objetivo de un estudio de este tipo puede ser la investigación sobre la relación entre algunas variables, o explorar la calidad de la información disponible, o construir un modelo que permita explicar el proceso tal como es actualmente, o saber si se ha desviado. Estos modelos del proceso se denominan **modelos empíricos**, porque están basados en los propios datos del proceso y no en una formulación teórica sobre el mismo.

Un estudio retrospectivo tiene la ventaja de tener a su disposición un gran número de datos que ya han sido recogidos, minimizando el esfuerzo de obtenerlos. Sin embargo, tiene varios problemas potenciales:

1. Si no disponemos de detalles suficientes, es posible que no podamos determinar si las condiciones de variación de los valores obtenidos responden a las mismas causas que en la situación actual.
2. Es posible que nos falte algún valor clave que no haya sido recogido o que lo haya sido de manera defectuosa
3. Algunas veces, la fiabilidad y validez de los datos de proceso históricos son dudosas, o al menos, cuestionables.
4. Los datos históricos no siempre se han recogido con la perspectiva actual del proceso, y es posible que no nos proporciones explicaciones adecuadas del proceso en su situación actual.
5. A veces queremos utilizar los datos históricos de proceso para fines que no estaban previstos cuando se recogieron
6. Las notas sobre los valores del proceso, incluyendo los valores anormales, pueden ser insuficientes o inexistentes, y no tenemos ninguna explicación sobre los posibles valores anómalos que detectamos en el análisis.

Usar datos históricos siempre tiene el riesgo de que, por la razón que sea, no se hayan recogido datos importantes, o que estos datos se hayan perdido, o se hayan transcrito de forma inadecuada o incorrecta. Es decir, los datos históricos pueden tener problemas de calidad de datos.

El hecho de que algunos datos se hayan recogido históricamente no siempre quiere decir que estos datos sean relevantes o útiles. Cuando el grado de conocimiento del proceso no es suficiente, o no se basa en un análisis metódico y riguroso de los datos, es posible que no se hayan recogido algunos datos que pueden ser importantes para el proceso, a veces simplemente porque son complejos o difíciles de analizar. Los datos históricos no pueden proporcionar la información que buscamos si la información de las variables clave nunca se ha recogido o se ha hecho sin una buena base experimental.

El propósito del análisis de los datos industriales es aislar las causas que están detrás de los sucesos que afectan e influyen en los procesos. En los datos históricos, estos sucesos pueden haber ocurrido semanas, meses o incluso años antes, sin que haya registros ni notas que hayan intentado explicar estas causas, y los recuerdos de las personas que han participado en ellos

se pierden con el tiempo, o se alteran involuntariamente, proporcionando explicaciones supuestamente válidas pero que en realidad son incorrectas. Por eso, con frecuencia, el análisis de los datos históricos puede poner de manifiesto hechos interesantes, pero sus causas quedan sin explicar.

Los estudios históricos pueden requerir una fase previa de preparación y depuración de datos que puede llegar a ser muy larga y tediosa. Se estima que en muchos estudios de ciencia de datos, el tiempo de preparación de los datos puede llegar al 60% del tiempo total empleado en el estudio. Las herramientas de análisis de datos son de gran ayuda en esta fase del proceso, aunque en muchas ocasiones será necesario un trabajo manual de recolección de datos en papel, hojas de cálculo diversas y otras fuentes. Esta fase es muy útil no sólo para la preparación de datos para el estudio, sino para mejorar el conocimiento de los datos, cómo se originan y cómo se almacenan. Este conocimiento siempre es de gran utilidad para mejorar los procedimientos actuales de captura de datos, facilitando la fiabilidad de los análisis futuros.

3.2 Estudios observacionales

Como su nombre indica, un **estudio observacional** simplemente observa un proceso durante un tiempo de operación en rutina. Normalmente, el ingeniero o técnico interfiere lo mínimo posible en el proceso, sólo lo suficiente para recoger la información que necesita, que en muchas ocasiones no forma parte de los controles de rutina, si piensa que esa información puede ser relevante. Si se planifican adecuadamente, los estudios observacionales proporcionan datos fiables, precisos y completos para documentar un proceso. Por otra parte, estos estudios proporcionan una información limitada sobre las relaciones entre las variables del proceso, porque es posible que durante el tiempo limitado de observación, el rango de variación de las variables no recoja todas las situaciones posibles, incluyendo situaciones extraordinarias.

3.3 Experimentos diseñados

Factores experimentales

La tercera forma de recoger información de un proceso son los **experimentos diseñados**. En un experimento de este tipo, el ingeniero o técnico hace un cambio deliberado en las variables que controla (llamadas **factores**), observa el resultado, y toma una decisión respecto a qué variable o variables son responsables de los cambios que observa en el proceso.

Una diferencia importante respecto a los estudios históricos y los observacionales es que las diferentes combinaciones de factores se aplican al azar sobre un conjunto de unidades experimentales. Esto permite establecer con precisión las relaciones causa-efecto, cosa que no suele ser posible en los estudios históricos ni en los observacionales.

4 La exploración de los datos

4.1 Describiendo un conjunto de datos

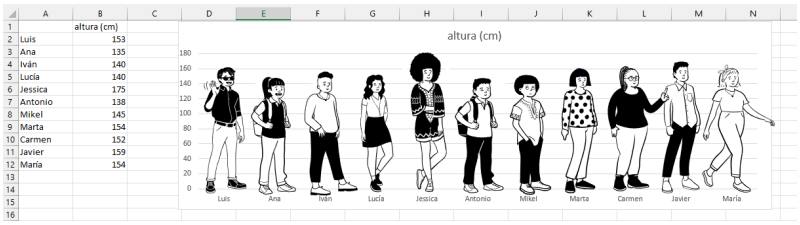
Supongamos que queremos medir la altura de los alumnos de nuestra clase. nuestro analista de la OMS ha realizado la medida de la altura de una niña siguiendo rigurosamente el método establecido, y, por lo tanto, que está razonablemente seguro de su resultado.

Al cabo de varias jornadas de trabajo, habrá realizado varias medidas, que representarán al conjunto de niños de la población en la que ha estado trabajando. Otros investigadores pueden haber estado trabajando a la vez en otras poblaciones, y al final de sus jornadas de trabajo, quieren comparar sus resultados: ¿Hay alguna de estas poblaciones en las que los niños sean significativamente más altos (o más bajos) que en las otras? ¿Cómo describir la altura de un conjunto de individuos de manera que se puedan hacer comparaciones con otros conjuntos?

Para responder a estas preguntas, vamos a cambiar el entorno de trabajo a un grupo de niños imaginario, que vamos a llamar *aula1*: son nuestros compañeros y compañeras, a los cuales realizaremos una medida de altura siguiendo el *procedimiento* especificado en nuestro *método*.

Éste es nuestro grupo de estudiantes:

Supongamos que hemos realizado las medidas. Lo primero que hacemos es registrar la altura de cada persona en una hoja de cálculo:



4.2 El diagrama de puntos o *dotplot*

El diagrama de tallo y hojas

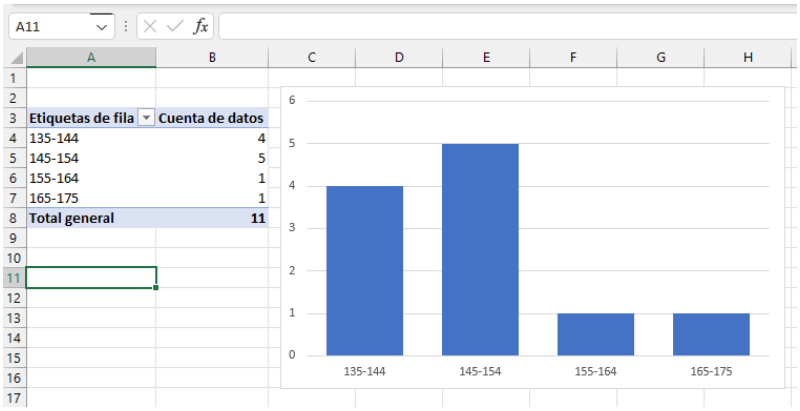
Construcción en Excel

4.3 La distribución de frecuencias

Si agrupamos nuestros valores por intervalos, y contamos el número de observaciones que aparecen en cada intervalo, obtenemos una *distribución de frecuencias*, que puede ser *absoluta* o *relativa* según que sus valores sean un recuento simple de los valores o el porcentaje que corresponde al número de observaciones en cada clase respecto al total de observaciones

Las formas más habituales de representar una distribución de frecuencias son la tabla de frecuencias o bien un gráfico de barras o histograma.

El gráfico a continuación muestra una distribución de frecuencias absoluta, calculada mediante una tabla dinámica de Excel, junto con su tabla de frecuencias absolutas.



Statistical Thinking - R Workflow (fharrell.com)

Tables can summarize frequency distributions of categorical variables, and measures of central tendency, selected quantiles, and measures of spread for continuous variables. When there is a truly discrete baseline variable one can stratify on it and compute the above types of summary measures. Tables fail completely when one attempts to stratify on a continuous baseline variable after grouping it into intervals. This is because categorization of continuous variables is a bad idea. Among other problems,

- categorization misses relationships occurring in an interval (this happens most often when an interval is wide such as an outer quartile)
- categorization loses information and statistical power because within-interval outcome heterogeneity is ignored and between-interval ordering is not utilized
- intervals are arbitrary, and changing interval boundaries can significantly change the apparent relationship
- with cutpoints one can easily manipulate results; one can find a set of cutpoints that results in a positive relationship and a different set that results in a negative relationship

4.4 Diagramas de barra

4.5 Histogramas

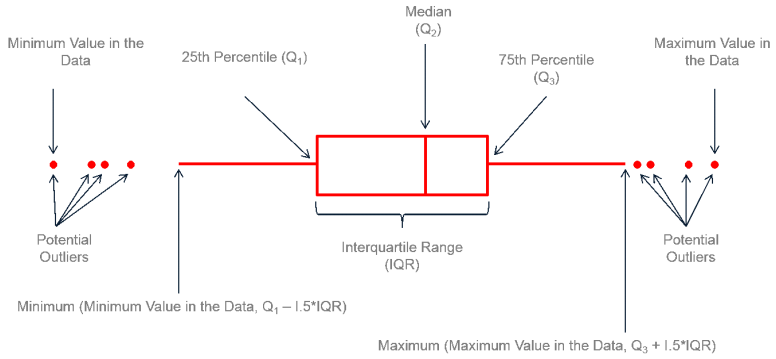
4.6 Gráficos de densidad

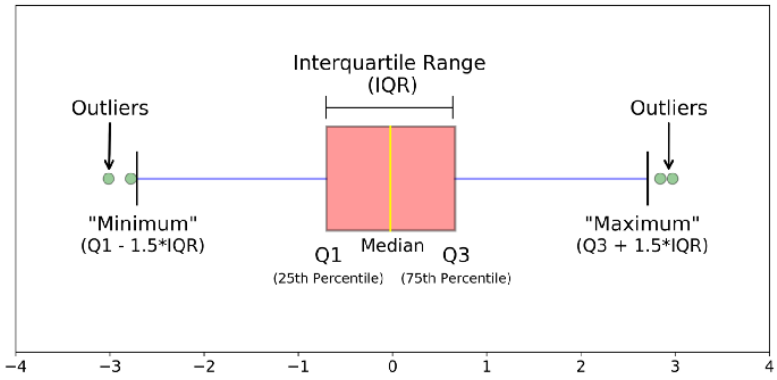
4.7 Los valores centrales: media, mediana, moda

4.8 Los valores de dispersión: varianza y desviación típica, rango intercuartil

4.9 Diagramas de caja (*box plot*)

Box Plot with Minitab - Lean Sigma Corporation





Boxplot | the R Graph Gallery (r-graph-gallery.com)

4.9.1 Gráficos de series

4.9.2 Otros gráficos

dotplot

Dot plot — geom_dotplot • ggplot2 (tidyverse.org)

4.9.3 La media o promedio: una medida central

Una primera posibilidad es suponer que en nuestro grupo de once personas no hubiese variación: que todos ellos tuviesen la misma altura. Si fuese así, podemos encontrar un valor x de altura que, repetido once veces, sea equivalente a la suma de las alturas de todos ellos. Si representamos cada alumno con una letra, la suma de sus alturas sería:

$$a + b + c + d + e + f + g + h + i + j + k$$

y el valor que buscamos sería un valor tal que, sumado once veces, el valor obtenido fuese igual a la suma de la alturas medidas:

$$a+b+c+d+e+f+g+h+i+j+k = x+x+x+x+x+x+x+x+x+x+x$$

Pero sabemos que la suma de valores repetidos es igual al valor multiplicado por el número de repeticiones:

$$a + b + c + d + e + f + g + h + i + j + k = 11x$$

Sólo tenemos que despejar la x para hallar este valor:

$$x = \frac{a + b + c + d + e + f + g + h + i + j + k}{11}$$

Este valor que hemos obtenido es lo que se conoce como *media*, *valor medio* o *promedio*, y, como hemos visto, **es aquel valor tal que repetido tantas veces como individuos tenemos, es equivalente a la suma de los valores que hemos obtenido**. La media de una muestra se representa habitualmente mediante el símbolo

$$\bar{x}$$

, y, de una manera más formal, su valor se obtiene mediante la fórmula siguiente:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

El signo $\sum$ se conoce como *sumatorio*, e indica que ese término consiste en la suma de los x valores desde el primero hasta el valor n. Expresado mediante una formulación matemática,

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

lo quiere quiere decir: *"la suma de todos los valores observados dividido entre el número de estos valores"*.

La **media** es lo que conocemos como un *valor central*, ya que representa el centro de nuestro conjunto de números. Como es el centro de nuestro conjunto de datos, *equidista* de todos los valores, o lo que es lo mismo, la suma de las distancias de todos los valores a este valor central es *cero*. Más adelante veremos la importancia de este hecho, al hablar de la dispersión y las formas de cálculo de la misma. La *media*, junto con otras medidas como la *mediana* y la *moda* se conocen en estadística como **medidas de tendencia central**. Como hemos dicho, la media de una **muestra** se representa como $\bar{x}$, mientras que la

media de una **población** se representa con la letra griega *mu*: μ. En ambos casos, el cálculo se realiza de forma idéntica.

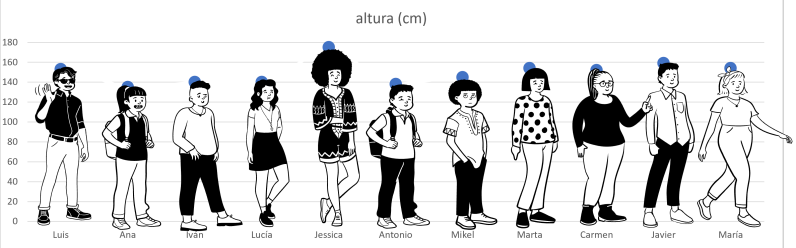
Volvamos a nuestro ejemplo para realizar los cálculos según el modelo que hemos descrito. En nuestro caso, la *altura media* de nuestros alumnos (la *media* de nuestro conjunto de números) se calcula como:

$$\bar{x} = \frac{153 + 135 + 140 + 140 + 175 + 138 + 145 + 154 + 152 + 159 + 154}{11} = 149,54$$

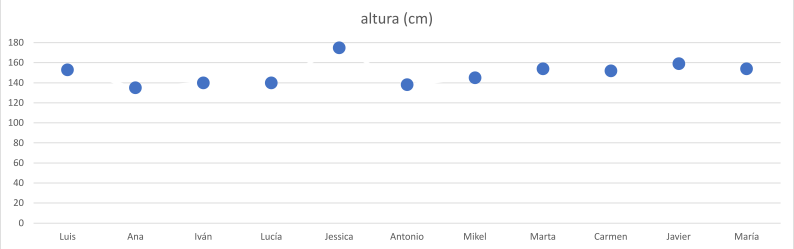
Utilicemos una hoja de cálculo para guardar nuestros valores.

La fórmula para obtener la media en la hoja de cálculo, por ejemplo en la versión en español de *Microsoft Excel*, es `=PROMEDIO(...)`, donde los puntos suspensivos deben sustituirse por el rango a calcular. En nuestro ejemplo, introduciríamos la fórmula en la celda `B13`como `=PROMEDIO(B2..B12)` (Para más detalles, verificar la hoja Excel adjunta).

Para representar más cómodamente nuestros valores, dibujamos un punto a la altura de cada alumno,

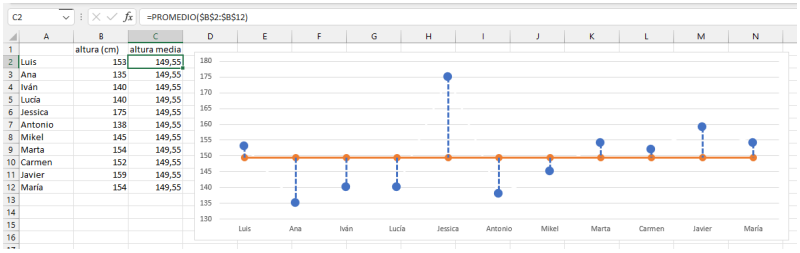


y eliminamos del gráfico los dibujos de nuestros alumnos; así hemos convertido nuestro dibujo en un *diagrama de puntos*:



Para representar la media, aunque la media es un valor único, necesitamos añadir una columna a la derecha de nuestros datos,

que rotulamos en la fila 1, celda C como `altura media`, e introducimos en cada una de las celdas desde `C2`hasta `C12`la fórmula del promedio, con le valor de nuestro rango de datos (Verificar hoja de cálculo). A continuación, designamos nuestro rango de datos para hacer un gráfico de puntos, y hacemos un *zoom* en los valores de manera que el eje Y se escale mejor entre los valores mínimo y máximo. Por último, hacemos unos ajustes en el formato para dibujar las líneas verticales que nos representan la distancia de cada valor a la media.



Si verificamos el eje Y , veremos que en este gráfico hemos ajustado la escala respecto al gráfico anterior, situando el mínimo en 130. Esto permite visualizar las diferencias con mucha más claridad. Hemos representado la media $\bar{x}$ como una línea, y hemos dibujado unas líneas que unen cada valor individual con la media, que se sitúa en el valor $149,55$, tal como calculamos más arriba.

Hemos representado la media como una serie de puntos unidos por una línea amarilla. Tal como hemos visto cuando hacíamos la descripción de este parámetro, representamos un conjunto de valores idénticos, ya que según hemos visto, la media **es aquel valor tal que repetido tantas veces como individuos tenemos, es equivalente a la suma de los valores reales que hemos obtenido**

Representamos en azul nuestros valores, uniendo cada valor con la línea media mediante una línea de puntos vertical. A partir de ahora, por conveniencia, eliminaremos los puntos en la linea media, dejando sólo la línea.

Esta línea azul de puntos representa la *distancia* de cada valor a la media. Usaremos esta distancia para calcular una *distancia media*, que será una medida de la dispersión de nuestros valores.

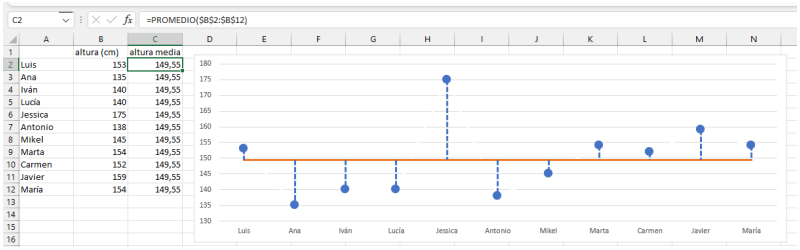


Figura 4.1: Hoja de cálculo con los valores y el gráfico de puntos

Recordemos que estamos intentando encontrar la forma de describir nuestro conjunto de números con un valor, con el fin de que nuestros analistas de la OMS puedan comparar la información de diferentes grupos de niños y ayudara determinar su situación nutricional.

Hemos visto que para describir un conjunto de números, en nuestro ejemplo, las medidas de la altura de un grupo de estudiantes, existe un valor, la *media* de este conjunto, que nos describe el centro de los valores. En nuestro ejemplo, si nuestro grupo tuviese un solo niño, éste tendría $149,55$ *cm* de altura.

¿Es suficiente con este valor para describir el conjunto de valores? Vamos a ver que no: diferentes conjuntos de valores pueden proporcionar el mismo *valor medio*, y sin embargo los grupos pueden ser muy diferentes.

Veamos un caso extremo. Comparemos dos grupos, uno formado por individuos iguales y otro formado por diez individuos iguales y uno distinto. Para ello usaremos nuestra hoja de cálculo:

¿Podemos describir adecuadamente los valores de la altura de cada uno de los grupos utilizando el valor medio? Parece evidente que no, ya que a partir de diferentes valores de altura estamos obteniendo el mismo valor medio. Sin embargo, uno de los grupos *es más alto* que el otro, si no fuera por un sólo individuo que aparentemente distorsiona el cálculo. Podríamos incluir nuestro grupo original, y veremos que los tres grupos son diferentes, aunque su valor medio es idéntico.

Si nos ayudamos de un gráfico equivalente al que hemos utilizado antes, vemos estas diferencias con claridad:

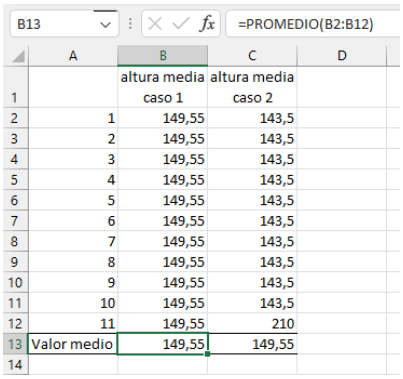


Figura 4.2: Dos grupos de valores con la misma media

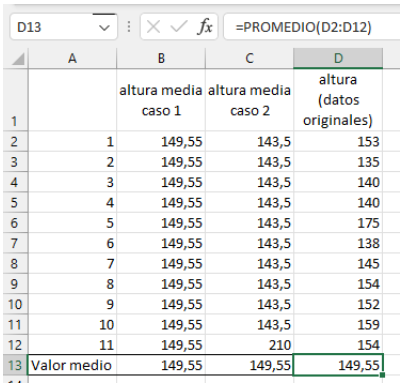


Figura 4.3: Tres grupos de valores con la misma media

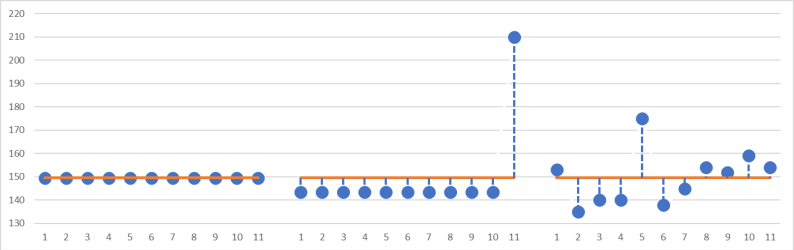


Figura 4.4: Gráfico de tres grupos de valores

Aunque el valor medio de estos tres grupos de datos es idéntico, parece claro que los tres grupos son muy distintos en su composición, y por lo tanto la *media* no es suficiente para describir con suficiente precisión cada uno de los grupos. Necesitamos un valor adicional, que nos indique de qué forma los valores se alejan del valor medio. Para ello, vamos a introducir un concepto nuevo: la *medida de la dispersión*, que nos indica precisamente la distancia de los valores al valor medio, e introduciremos también la *distribución de frecuencias*, que nos permite representar *la forma* en la que se distribuyen nuestros valores.

La media como centro de gravedad: Physics Simulation: Center of Mass (physicsclassroom.com)

4.9.4 Las medidas de dispersión: la desviación típica

Como hemos visto en el apartado anterior, diferentes conjuntos de datos pueden tener el mismo valor medio y sin embargo ser muy diferentes. En la última gráfica que hemos visto, el primer grupo se caracteriza por tener todos sus valores idénticos; el segundo tiene todos sus valores idénticos menos uno, que está muy apartado del resto, y el tercero tiene todos sus valores diferentes.

Ahora que conocemos cómo calcular un valor resumen de un conjunto de datos, podríamos utilizar una medida semejante para describir de qué forma en cada caso los valores se separan de la media. Podríamos utilizar una *distancia media*: calculamos las diferencias entre cada valor y la media, y hacemos su promedio: esto debería darnos una indicación de la magnitud de la separación de los valores en cada uno de los tres grupos.

Usemos la hoja de cálculo para ello:

D14 =PROMEDIO(D2:D12)

	A	B	C	D	E	F	G	H	I	J	K	L
1		altura media caso 1	media del grupo	diferencia		altura media caso 2	media del grupo	diferencia		altura (datos originales)	media del grupo	diferencia
2	1	149,5454545	149,55	0,00		143,5	149,55	-6,05		153	149,55	3,45
3	2	149,5454545	149,55	0,00		143,5	149,55	-6,05		135	149,55	-14,55
4	3	149,5454545	149,55	0,00		143,5	149,55	-6,05		140	149,55	-9,55
5	4	149,5454545	149,55	0,00		143,5	149,55	-6,05		140	149,55	-9,55
6	5	149,5454545	149,55	0,00		143,5	149,55	-6,05		175	149,55	25,45
7	6	149,5454545	149,55	0,00		143,5	149,55	-6,05		138	149,55	-11,55
8	7	149,5454545	149,55	0,00		143,5	149,55	-6,05		145	149,55	-4,55
9	8	149,5454545	149,55	0,00		143,5	149,55	-6,05		154	149,55	4,45
10	9	149,5454545	149,55	0,00		143,5	149,55	-6,05		152	149,55	2,45
11	10	149,5454545	149,55	0,00		143,5	149,55	-6,05		159	149,55	9,45
12	11	149,5454545	149,55	0,00		210	149,55	60,45		154	149,55	4,45
13												
14			Promedio	0,00			Promedio	0,00			Promedio	0,00

Figura 4.5: Tres grupos de valores en la hoja de cálculo

Algo parece que no está funcionando aquí: el promedio de las diferencias es cero en los tres casos; no podemos usar este cálculo para calcular la dispersión. Pero esto es esperable: ya que la media es un valor central, como hemos visto antes, la suma de las diferencias de todos los valores respecto de su media debe ser forzosamente cero, y esto es lo que estamos obteniendo.

Para encontrar una solución, vamos a recurrir al viejo teorema de Pitágoras, que si recuerdas, nos dice que, en un triángulo rectángulo, el cuadrado de la hipotenusa es igual a la suma de los cuadrados de los catetos (una explicación gráfica muy divertida en el anexo …):

$$h^2 = a^2 + b^2$$

Esta fórmula es la base del cálculo de la distancia entre dos puntos:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

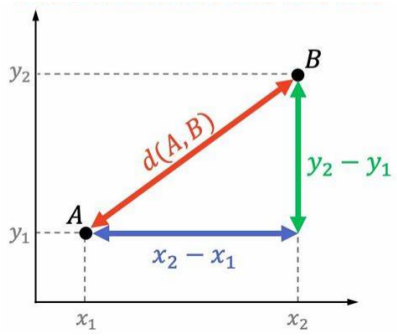


Figura 4.6: Distancia entre dos puntos

¿Podemos adaptar esta fórmula para el cálculo de nuestra distancia media? La respuesta es **sí**. En nuestro caso, sólo necesitamos la coordenada X, ya que sólo estamos calculando la distancia en una dimensión. Si tenemos en cuenta un solo punto, esta distancia d sería:

$$(d\ del\ valor\ 1\ a\ la\ media)^2 = (x_1 - \bar{x})^2$$

¡El hecho de elevar al cuadrado las diferencias nos da la solución! Las diferencias negativas ya no son un problema porque

sabemos que al elevar un numero negativo al cuadrado, el resultado es positivo; de esta manera conseguimos que las diferencias no se anulen. Ahora sí podemos calcular una distancia media $\bar{d}$ entre el conjunto de puntos y su media, calculando el promedio de las diferencias elevadas al cuadrado:

$$(\bar{d} \textit{ de los } n \textit{ valores a la media})^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

y utilizando la notación que hemos aprendido antes,

$$(\bar{d} \textit{ de los } n \textit{ valores a la media})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Al igual que en el cálculo de la distancia entre dos puntos, sólo tenemos que extraer la raíz cuadrada de este valor para obtener la distancia media, que es el parámetro que estábamos buscando.

La

$$(\bar{d} \textit{ de los } n \textit{ valores a la media})^2$$

se conoce en estadística como **varianza**, y su raíz cuadrada es lo que se conoce como **desviación típica**. La varianza de una población se representa en estadística con el signo de la letra griega *sigma* minúscula elevada al cuadrado, σ^2, y la desviación típica, mediante la letra σ. En el caso de una muestra, la varianza se representa como s_x^2, y la desviación típica, como s_x. En nuestro caso, utilizaremos la primera notación; más adelante veremos los conceptos de **población** y **muestra** y explicaremos el concepto de **grados de libertad**. Veremos también que la fórmula para el cálculo de la desviación típica muestral es ligeramente diferente de la de su equivalente poblacional, y explicaremos por qué.

Es importante resaltar que la desviación típica *es una medida de la distancia media de los valores de una población a su media*, y por lo tanto tiene dimensión, la misma que las medidas originales. La varianza, al estar elevada al cuadrado, no tiene una dimensión, o, mejor dicho, tiene la de la medida al cuadrado.

Con estos nuevos hallazgos, recalculamos nuestra hoja de cálculo:

M15 =RAIZ(M13)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		altura media caso 1	media del grupo	diferencia	diferencia ^2	altura media caso 2	media del grupo	diferencia	diferencia^2	altura (datos originales)	media del grupo	diferencia	diferencia^2
2	1	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	153,00	149,55	3,45	11,93
3	2	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	135,00	149,55	-14,55	211,57
4	3	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	140,00	149,55	-9,55	91,12
5	4	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	140,00	149,55	-9,55	91,12
6	5	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	175,00	149,55	25,45	647,93
7	6	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	138,00	149,55	-11,55	133,30
8	7	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	145,00	149,55	-4,55	20,66
9	8	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	154,00	149,55	4,45	19,84
10	9	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	152,00	149,55	2,45	6,02
11	10	149,55	149,55	0,00	0,00	143,50	149,55	-6,05	36,55	159,00	149,55	9,45	89,39
12	11	149,55	149,55	0,00	0,00	210,00	149,55	60,45	3654,75	154,00	149,55	4,45	19,84
13	Promedio	149,55		0,00	0,00	149,55		0,00	365,48	149,55		0,00	122,07
14	Varianza	0,00				365,48				122,07			
15	Desviación típica	0,00				19,12			19,12	11,05			11,05

Figura 4.7: Tres grupos de valores en la hoja de cálculo, con la misma media y distinta desviación típica

Vamos a analizar con detalle esta tabla.

En la columna `J` tenemos nuestra población original de 11 alumnos, con las alturas que hemos medido. En la columna `B` hemos supuesto que todos los alumnos fuesen iguales, con la misma altura del valor medio de los datos originales. En la columna `F` hemos simulado otro grupo, con todos los valores iguales excepto uno, y con la misma media que los otros dos grupos.

A la derecha de cada columna de medias, tenemos la columna de diferencias (columnas `D`, `H` y `L`), y en la fila `13`, nuestro primer intento de calcular una dispersión media; intento fallido, puesto que obteníamos el valor 0 para los tres grupos.

En la siguiente columna a la derecha, para los tres grupos (columnas `E`, `I`y `M`), hemos elevado al cuadrado la distancia de cada valor a la media, siguiendo los hallazgos que nos ha proporcionado el teorema de Pitágoras y la fórmula de la distancia entre dos puntos. En la fila `13` de estas columnas, calculamos el promedio de la distancia a la media al cuadrado: esta vez el resultado ya no es cero, sino que obtenemos el valor de la **varianza**, de acuerdo con la fórmula que hemos deducido más arriba. En la fila `14` (columnas B, `F` y `J`)utilizamos la fórmula de la hoja de cálculo para la **varianza poblacional** (más detalles posteriormente), y vemos que coincide exactamente con el promedio de las diferencias al cuadrado, tal como debe ser, ya que en eso consiste la fórmula que hemos deducido.

Por último, en la fila `15`calculamos la desviación típica de ambas formas, con la fórmula de la hoja de cálculo para la **desviación típica poblacional** (columnas `B`, `F`y `J`), que Excel llama

desviación estándar, y como la raíz cuadrada del promedio calculado antes (columnas `E`, `I`y `M`). De nuevo, ambos valores coinciden exactamente, como esperamos.

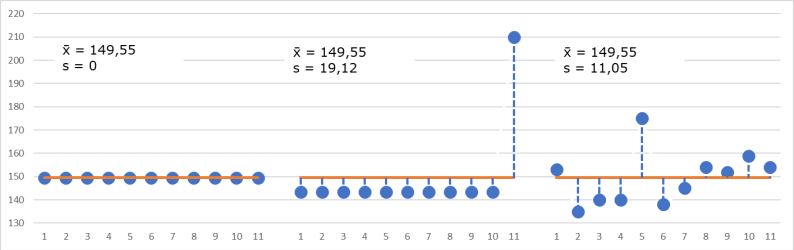


Figura 4.8: Gráfico con tres conjuntos de datos con la misma media y diferente desviación típica

Ahora sí tenemos una forma más completa de describir nuestro conjunto de valores. Aunque el valor medio es el mismo en los tres casos, la *dispersión* de los valores es muy distinta.

¿Son suficientes estos dos parámetros que hemos calculado para describir un conjunto de datos? La respuesta a esta pregunta es sí y no. La explicación es que, más allá de los valores numéricos que hemos obtenido, la visualización gráfica de los valores nos debe hacer reflexionar.

En el primer grupo, todos los valores son iguales a la media. La variación es cero. Son valores que hemos simulado en nuestra hoja de cálculo, pero difícilmente en el mundo real encontraremos una población en la que todos sus valores, en este caso, la altura de un grupo de alumnos, sean idénticos.

En el segundo grupo, todos los valores son idénticos, salvo uno, que se distancia mucho. ¿Debemos aceptar esto como bueno? En realidad, ¿es cierto que el valor medio de este grupo sea el mismo que el del primero? Para responder a esta pregunta debemos recurrir a nuestra experiencia, la estadística no nos da *fórmulas mágicas*. Pero, con un poco de sentido común, parece que el caso extremo que aparece en este grupo no es coherente con el resto de valores. Es lo que se llama un *valor anormal* o *extraño* (en inglés, *outlier*), y debe hacernos reflexionar sobre si el valor es correcto y realmente pertenece a esta población, o es un error de medida. O, simplemente, un valor que corresponde a otro grupo y que por error hemos situado en éste. La decisión