

Estadística Aplicada a las Ciencias y la Ingeniería

Emilio L. Cano

2023-04-16

Índice general

Bienvenida

Bienvenido/a a “Estadística Aplicada a las Ciencias y la Ingeniería” por Emilio L. Cano.

Este libro incluye los contenidos habitualmente presentes en el currículo de asignaturas de **Estadística** de los grados Ciencias e Ingenierías de universidades españolas. Aunque no aparezca en el título, el manual incluye también los contenidos de **Probabilidad** necesarios. Si bien existe abundante material bibliográfico que cubre los contenidos de estas asignaturas, quería elaborar un material propio que no fuera solamente para mis clases sino algo más *global*. En los últimos años ya lo hice para asignaturas de grado y Máster en ADE (?). Por otra parte, me motiva cubrir el hueco de los materiales de acceso gratuito con la opción de comprar una edición impresa¹ y con el enfoque que se menciona en el siguiente apartado. Por otra parte, los libros publicados originalmente en inglés y traducidos al español a menudo me resultan lejanos a nuestro idioma (por muy buenas que sean las traducciones, los ejemplos en *acres* no son muy intuitivos para un lector español). Espero que también sirva para lectores de otros países de habla hispana.

Estándares y software

Los contenidos de este libro se basan en dos paradigmas que están presentes en los intereses de investigación y docencia del autor: los **estándares** y el **software libre**. En lo que se refiere a estándares, la notación utilizada, definiciones y fórmulas se ajustarán el máximo posible a la utilizada en normas nacionales e internacionales sobre metodología estadística. Estas normas se citarán pertinentemente a lo largo del texto. En cuanto al software libre, se proporcionarán instrucciones para resolver los ejemplos que ilustran la teoría utilizando software libre. No obstante, el uso del software es auxiliar al texto y se puede seguir sin necesidad de utilizar los programas. Según lo que proceda en cada caso, se utilizará software de hoja de cálculo, el software estadístico y lenguaje de programación **R** (?), y el software de álgebra computacional **Máxima**². Respecto al software

¹A la espera de encontrar editorial.

²<http://maxima.sourceforge.net/es/>

de hoja de cálculo, las fórmulas utilizadas se han probado en el software libre **LibreOffice**³, en **Hojas de Cálculo de Google**⁴ y también en **Microsoft EXCEL**⁵ que, aunque no es software libre, su uso está más que generalizado y normalmente los estudiantes disponen de licencia de uso a través de su universidad. En caso de que el nombre de la función sea distinta en EXCEL, se indicará en el propio ejemplo.

Las normas son clave para el desarrollo económico de un país. Estudios en diversos países, incluido España, han demostrado que la aportación de la normalización a su economía es del 1 % del PIB⁶. La Asociación Española de Normalización (UNE) es el organismo legalmente responsable del desarrollo y difusión de las normas técnicas en España. Además, representa a España en los organismos internacionales de normalización como ISO⁷ y CEN⁸.

Las normas sobre estadística que surgen de ISO las elabora el *Technical Committee* ISO TC 69⁹ *Statistical Methods*. Por su parte, el subcomité técnico de normalización CTN 66/SC 3¹⁰, Métodos Estadísticos, participa como miembro nacional en ese comité ISO. Las normas que son de interés en España, se ratifican en inglés o se traducen al español como normas UNE. Para una descripción más completa de la elaboración de normas, véase ?.

Estructura del libro

Este libro se ha elaborado utilizando el lenguaje *Markdown* con el propio software **R** y el paquete **bookdown** (?). Se incluyen una gran cantidad de ejemplos resueltos tanto de forma analítica como mediante software. En algunos casos se proporciona el uso de funciones en hojas de cálculo (y el resultado obtenido con un recuadro). En otros, código de R, que aparecen en el texto sombreados y con la sintaxis coloreada, como el fragmento a continuación donde se puede comprobar la sesión de R en la que ha sido generado este material. Obsérvese que los resultados se muestran precedidos de los símbolos #>.

```
sessionInfo()
#> R version 4.2.3 (2023-03-15)
#> Platform: x86_64-apple-darwin17.0 (64-bit)
#> Running under: macOS Big Sur ... 10.16
#>
#> Matrix products: default
```

³<https://es.libreoffice.org>

⁴<https://www.google.es/intl/es/sheets/about/>

⁵<https://products.office.com/es-es/excel>

⁶<http://www.aenor.es/DescargasWeb/normas/como-beneficia-es.pdf>

⁷<https://www.iso.org/>

⁸<https://www.cen.eu/>

⁹<https://www.iso.org/committee/49742/x/catalogue/>


¹⁰<https://www.une.org/encuentra-tu-norma/comites-tecnicos-de-normalizacion/comite/?c=CTN%2066/SC%203>

```

#> BLAS: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
#> LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
#>
#> locale:
#> [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
#>
#> attached base packages:
#> [1] stats      graphics  grDevices  utils      datasets
#> [6] methods   base
#>
#> other attached packages:
#> [1] flextable_0.8.4  fontawesome_0.5.0
#>
#> loaded via a namespace (and not attached):
#> [1] zip_2.2.2      Rcpp_1.0.10     compiler_4.2.3
#> [4] later_1.3.0    base64enc_0.1-3  gfonts_0.2.0
#> [7] tools_4.2.3    digest_0.6.31    uuid_1.1-0
#> [10] evaluate_0.20  jsonlite_1.8.4   memoise_2.0.1
#> [13] lifecycle_1.0.3  rlang_1.0.6      shiny_1.7.4
#> [16] cli_3.6.0      rstudioapi_0.14  yaml_2.3.7
#> [19] crul_1.3        curl_5.0.0        xfun_0.36
#> [22] fastmap_1.1.0   officer_0.5.2     knitr_1.42
#> [25] xml2_1.3.3      gdtools_0.3.0     systemfonts_1.0.4
#> [28] askpass_1.1     grid_4.2.3        glue_1.6.2
#> [31] httpcode_0.3.0  data.table_1.14.6 R6_2.5.1
#> [34] rmarkdown_2.20  bookdown_0.32     magrittr_2.0.3
#> [37] promises_1.2.0.1 ellipsis_0.3.2    htmltools_0.5.5
#> [40] mime_0.12       xtable_1.8-4      httpuv_1.6.8
#> [43] openssl_2.0.5   cachem_1.0.6      crayon_1.5.2

```

Normalmente, la descripción o enunciado de los ejemplos se incluyen en bloques con el siguiente aspecto:



Esto es un ejemplo. A continuación puede mostrarse código o no. Los ejemplos pueden ir precedidos por un icono para identificar su campo de aplicación, por ejemplo 🌿 Biología, 🏠 Ciencia y tecnología de Alimentos, o 🌳 Ciencia e Ingeniería Ambiental.

Cuando el ejemplo incluya explicaciones sobre cómo resolverlo con software, estas explicaciones aparecerán en bloques con el siguiente aspecto:

HOJA DE CÁLCULO

La función **FACT** obtiene el factorial de un número x ($x!$):



=FACT(5) 120

También se incluirán con el formato anterior indicaciones para usar la calculadora científica, cuando esto sea posible.

El texto incluye otros bloques con información de distinto tipo, como los siguientes:



Este contenido se considera avanzado. El lector principiante puede saltarse estos apartados y volver sobre ellos en una segunda lectura.



Estos bloques están pensados para incluir información curiosa o complementaria para poner en contexto las explicaciones.

Este volumen cubre los contenidos de asignaturas básicas de Estadística en un amplio rango de grados. Puede servir también como repaso para alumnos de posgrado o incluso egresados que necesiten refrescar conocimientos o aprender a aplicarlos con software moderno. Un segundo volumen cubrirá en el futuro métodos y modelos avanzados para entornos más exigentes.

El libro está dividido en 4 partes. La primera parte está dedicada a la Estadística Descriptiva, y consta de un capítulo introductorio seguido de sendos capítulos para el análisis exploratorio univariante y bivalente. La segunda parte trata la Probabilidad en 4 capítulos, uno introductorio, dos dedicados a las variables aleatorias univariantes y bivariantes respectivamente, y finalmente un capítulo que trata los modelos de distribución de probabilidad. En la tercera parte se aborda la inferencia estadística, con una introducción al muestreo y la estimación puntual, seguida de capítulos dedicados a los contrastes de comparación de grupos, análisis de regresión y diseño de experimentos. La última parte está dedicada al control estadístico de la calidad, en la que, tras un capítulo introductorio, se tratan las dos herramientas más importantes en este campo: el control estadístico de procesos (SPC, *Statistical Process Control*, por sus siglas en inglés) y los muestreos de aceptación o, dicho de otra forma, la inspección por muestreo. Finalmente, una serie de apéndices con diverso material complementan el libro en su conjunto.

Sobre el autor

Emilio López Cano, Estadístico y entusiasta de R. Actualmente soy Profesor Contratado Doctor en la Escuela Técnica Superior de Ingeniería Informática e

investigador en el Data Science Laboratory de la Universidad Rey Juan Carlos. Mis intereses de investigación incluyen Estadística Aplicada, Aprendizaje Estadístico y Metodologías para la Calidad. Previamente he sido profesor e investigador en la Universidad de Castilla-La Mancha, donde sigo colaborando en docencia e investigación, y Estadístico en empresas del sector privado de diversos sectores.

Presidente del subcomité técnico de normalización UNE (miembro de ISO) CTN 66/SC 3 (Métodos Estadísticos). Profesor en la Asociación Española para la Calidad (AEC). Presidente de la asociación Comunidad R Hispano.

Más sobre mí, información actualizada y publicaciones: <http://emilio.lcano.com>. Contacto: emilio@lcano.com

El material se proporciona bajo licencia CC-BY-NC-ND. Todos los logotipos y marcas comerciales que puedan aparecer en este texto son propiedad de sus respectivos dueños y se incluyen en este texto únicamente con fines formativos. Se ha puesto especial cuidado en la adecuada atribución del material no elaborado por el autor, véase el Apéndice ???. Aún así, si detecta algún uso indebido de material protegido póngase en contacto con el autor y será retirado. Igualmente, contacte con el autor **si desea utilizar este material con fines comerciales**.



Este obra está bajo una licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional.

Agradecimientos

Este libro es el resultado de años de trabajo en la docencia, investigación y transferencia de conocimiento en el campo de la Estadística. Está construido a partir de las contribuciones a lo largo de los años de compañeros y amigos como Javier M. Moguerza, Andrés Redchuk, David Ríos, Felipe Ortega, Mariano Prieto, Miguel Ángel Tarancón, Víctor M. Casero, Virgilio Gómez-Rubio, Matías Gámez, y muchos otros (perdón a l@s omitid@s por no ser más exhaustivo).

Especial agradecimiento a toda la comunidad del software libre y lenguaje de programación R, y en particular al *R Core Team*, al equipo de Posit (antes RStudio) y a los amigos de R Hispano.

Parte I

Estadística descriptiva

Capítulo 1

Introducción

1.1. Estadística y análisis de datos

1.1.1. ¿Qué es la Estadística?

Antes de introducirnos en el estudio de la Estadística y sus métodos, vamos a intentar tener una visión de todo lo que abarca. Así pues, ¿qué es la Estadística? La primera fuente que podemos consultar es la definición de la Real Academia Española, y encontramos estas acepciones:

estadístico, ca

La forma f., del al. Statistik, y este der. del it. statista ‘hombre de Estado’.

1. adj. Perteneciente o relativo a la estadística.
2. m. y f. Especialista en estadística.
3. f. **Estudio de los datos** cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas.
4. f. Conjunto de **datos** estadísticos.
5. f. Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener **inferencias** basadas en el **cálculo de probabilidades**.

RAE

Las acepciones que nos interesan son sobre todo la tercera y la cuarta, en las que aparecen conceptos que veremos en este capítulo introductorio y en los que profundizaremos en el resto del libro. La tercera acepción, “Conjunto de **datos** estadísticos”, es lo que muchas personas entienden cuando oyen la palabra

Estadística: La estadística del paro, la estadística de los precios, etc. Pero la Estadística es mucho más amplia. En primer lugar, esos “datos estadísticos” han tenido que ser recopilados y tratados de alguna forma antes de llegar a su publicación. Además, los datos estadísticos así entendidos son el resultado de un estudio pormenorizado (acepción 3) y normalmente de la aplicación de técnicas de **inferencia** (acepción 5). Algunas de estas técnicas forma parte de lo que vulgarmente se conoce como “la cocina” de las estadísticas.

Podemos hablar entonces de la Estadística, de forma muy resumida, como la ciencia de analizar datos. Encontramos a menudo¹ una definición de la Estadística como “la ciencia que establece los métodos necesarios para la recolección, organización, presentación y análisis de datos relativos a un conjunto de elementos o individuos”. Pero esta definición se centra solo en los métodos. Una definición más completa sería la siguiente:

[...] la estadística es la parte de la matemática que estudia la **variabilidad** y el proceso aleatorio que la genera siguiendo leyes de **probabilidad**.

Esta variabilidad puede ser debida al azar, o bien estar producida por causas ajenas a él, correspondiendo al **razonamiento estadístico** diferenciar entre la variabilidad casual y la variabilidad causal.

?

Aquí vemos uno de los conceptos clave que guiará todo el estudio y aplicación de la Estadística: la variabilidad es la clave de todo. Entender el concepto de variabilidad ayudará enormemente a entender los métodos por complejos que sean.

Variation is the reason for being of statistics

?

La Estadística ha sido siempre importante en los estudios de Ciencias e Ingeniería. No obstante, en los últimos tiempos la alta disponibilidad tanto de datos como de tecnología para tratarlos, hace imprescindible un dominio de las técnicas estadísticas y su aplicación en el dominio específico.

1.1.2. Los dos grandes bloques de la Estadística

La Estadística se divide en dos grandes bloques de estudio, que son la **Estadística Descriptiva** y la **Inferencia Estadística**. A la Estadística Descriptiva también se la conoce como *Análisis Exploratorio de Datos* (EDA, *Exploratory Data Analysis*, por sus siglas en inglés). Esta disciplina tuvo un gran desarrollo gracias al trabajo de Tukey (?), que todavía hoy es una referencia. Pero en los últimos años ha cobrado si cabe más importancia por la alta disponibilidad de datos y la necesidad de analizarlos.


¹Por ejemplo en el Curso de Estadística Práctica Aplicada a la Calidad de la AEC.

La **Estadística Descriptiva** se aplica sobre un conjunto de datos concretos, del que obtenemos resúmenes numéricos y visualización de datos a través de los gráficos apropiados. Con la Estadística Descriptiva se identifican **relaciones** y **patrones**, guiando el trabajo posterior de la Inferencia Estadística.

La **Estadística Inferencial** utiliza los datos y su análisis anterior para, a través de las Leyes de la **Probabilidad**, obtener conclusiones de diverso tipo, como explicación de fenómenos, confirmación de relaciones de causa-efecto, realizar predicciones o comparar grupos. En definitiva, tomar decisiones por medio de modelos estadísticos y basadas en los datos.

1.1.3. La esencia de la Estadística

La figura ?? representa la esencia de la Estadística y sus métodos. Estudiamos alguna **característica** observable en una serie de **elementos** (sujetos, individuos, ...) identificables y únicos. Los datos que analizamos, provienen de una determinada **población** que es objeto de estudio. Pero estos datos, no son más que una **muestra**, es decir, un subconjunto representativo de la población. Incluso cuando “creemos” que tenemos todos los datos, debemos tener presente que trabajamos con muestras, ya que generalmente tomaremos decisiones o llegaremos a conclusiones sobre el futuro, y esos datos seguro que no los tenemos. Por eso es importante considerar siempre este paradigma población-muestra, donde la población es desconocida y sus propiedades teóricas. La **Estadística Descriptiva** se ocupa del análisis exploratorio de datos en sentido amplio, que aplicaremos sobre los datos concretos de la muestra en esta unidad y la siguiente. La **Inferencia Estadística** hace referencia a los métodos mediante los cuales, a través de los datos de la muestra, tomaremos decisiones, explicaremos relaciones, o haremos predicciones sobre la población. Para ello, haremos uso de la **Probabilidad**, que veremos más adelante, aplicando el método más adecuado. En estos métodos será muy importante considerar el método de obtención de la muestra que, en términos generales, debe ser representativa de la población para que las conclusiones sean válidas.

 En un ensayo clínico, se eligen una serie de participantes en el estudio a los que se le suministran distintos tratamientos según el diseño del ensayo. Los participantes en el estudio son sujetos que constituyen la **muestra**. A través de los resultados de esta muestra, obtendremos conclusiones para toda la **población**, que estará definida en el propio ensayo clínico. Por ejemplo, en el estudio del efecto de un determinado tratamiento para la diabetes, la población serían todos los enfermos de diabetes.



Otro concepto clave inherente a la Estadística, es que casi siempre estaremos investigando sobre esta fórmula:

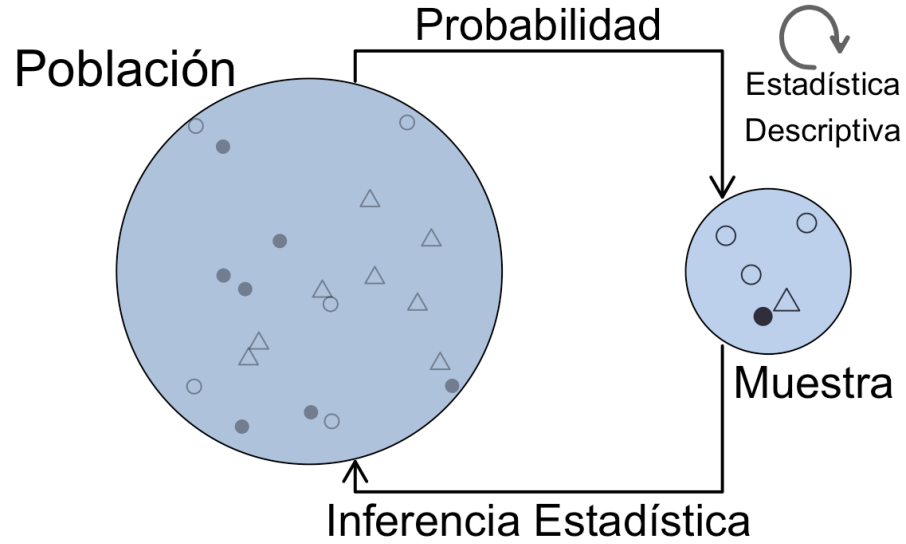


Figura 1.1: La esencia de los métodos estadísticos

$$Y = f(X)$$

Es decir, buscamos encontrar la relación entre una variable respuesta Y y una o varias variables explicativas X . Casi toda la Ciencia de Datos consiste en encontrar esa f . Es fundamental interiorizar este concepto para después aplicar el método adecuado, ya que según sean la/s Y , la/s X y el objetivo de nuestro estudio, los caminos pueden ser muy diferentes.

El origen del término *Data Science* se suele atribuir a Bill Cleveland tras la publicación de su artículo “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics” en 2001 (?)², aunque lo anticipó Tukey 40 años antes en “The Future of Data Analysis” (?). No obstante, es a partir del año 2010, con la irrupción del *Big Data* y la necesidad de analizar grandes cantidades de datos, cuando se empieza a popularizar el término intentando dar una definición gráfica de la profesión (*Data Scientist*). Así, es muy común presentar la ciencia de datos como la intersección de los conocimientos informáticos, los conocimientos estadístico-matemáticos, y el conocimiento de la materia en estudio (negocio, campo científico, etc.). Así, la persona de ciencias o ingeniería, con evidentes conocimientos en su campo, que adquiera conocimientos de Estadística y sea capaz de utilizar software avanzado como R, es uno de los perfiles más demandados.

²En el seno de los laboratorios Bell, como muchos otros avances de la Ciencia Estadística (por ejemplo SPC, *Statistical Process Control*, o S, el precursor del software estadístico y lenguaje de programación R.)


Paralelamente a la Ciencia de Datos, aparecen términos más recientes como *Big Data*, *Internet of Things* o Industria 4.0. Detrás de todos ellos, está el análisis estadístico. Y la mayoría de las veces es suficiente aplicar los métodos más básicos para solucionar los problemas o demostrar las hipótesis.

1.2. Los datos y su organización

1.2.1. Características y variables

Las **características** que observamos en los **elementos** de la muestra (o que estudiamos en una población) pueden ser distintos tipos. Nos referiremos genéricamente a estas características como **variables**, aunque en algunos ámbitos como el Control Estadístico de Procesos (SPC, *Statistical Process Control* por sus siglas en inglés) este término se refiere solo a las variables continuas que ahora definiremos.

Denotaremos las variables con letras mayúsculas del alfabeto latino (X , Y , A , ...). Cuando observamos la característica, la variable toma un **valor**. Estos valores pueden ser agrupados en **clases**, de forma que cada posible valor pertenezca a una y solo una clase. En ocasiones los datos con los que trabajamos están ya clasificados en clases. Las variables pueden tomar cualquier valor en su **dominio**, es decir, el conjunto de **posibles** valores que puede tomar la variable. Veremos más adelante cómo cuantificar esas posibilidades a través de la Probabilidad.



Cuando se recogen datos utilizando cuestionarios, a menudo en las preguntas para recoger características cuantitativas se ofrece elegir un intervalo en vez de preguntar el **valor** exacto. Por ejemplo, al preguntar la edad de una persona, se pueden dar las opciones: 1) menos de 20 años; 2) entre 20 y 40 años; 3) entre 40 y 60 años; 4) Más de 60 años. Así, si una persona tiene 30 años, el **valor** de la variable es 30 (en el caso de la encuesta no lo conoceremos exactamente) que pertenece a la **clase** “entre 20 y 40 años”.

1.2.2. Parámetros y estadísticos

Distinguiremos la caracterización de las variables que estudiamos en la población de las observadas en la muestra denotándolas por **parámetros** y **estadísticos** respectivamente. Los parámetros son valores teóricos, casi siempre desconocidos, sobre los que haremos inferencia. Los denotaremos por letras griegas minúsculas, como por ejemplo μ para la media poblacional. Un estadístico es una función definida sobre los datos de una **muestra**. Pueden ser valores de más de una variable, y los resumiremos en un único valor, resultado de aplicar esa función. Los estadísticos tomarán valores distintos dependiendo de la muestra concreta. Esto hace que sean a su vez variables, y que tengan una distribución en el

muestreo que nos permitirá hacer inferencia sobre la población. Los denotaremos con letras latinas, como por ejemplo \bar{x} para la media muestra.

La figura ?? representa la esencia de la estadística relacionando parámetros y estadísticos. Además de la equivalencia entre parámetros y estadísticos, la distribución de frecuencias de los datos de la muestra representada en el histograma se corresponde con la distribución de probabilidad teórica de la población.

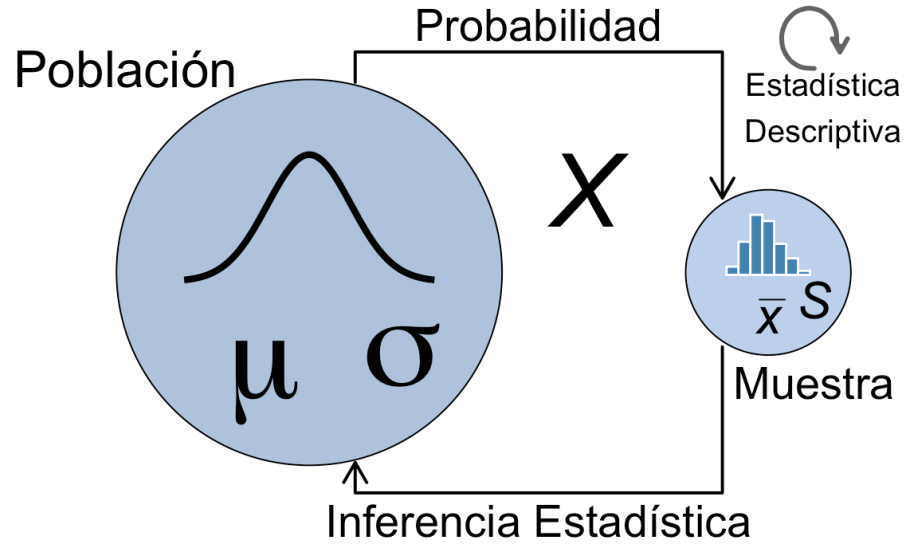


Figura 1.2: La esencia de los métodos estadísticos

1.2.3. La inferencia y sus métodos

Existen dos grandes grupos de métodos para hacer la inferencia sobre la población. La **estadística paramétrica** asume que la característica sigue una determinada distribución de probabilidad. Esta distribución de probabilidad depende de unos **parámetros** (por ejemplo, la media y la desviación típica). La inferencia se hace en base a esos parámetros, y se asumen ciertas hipótesis de partida que se deben comprobar. La **estadística no paramétrica** no asume ninguna distribución de probabilidad para la característica. Los métodos se basan en estadísticos de orden (cuantiles) y no hace falta cumplir ninguna hipótesis.

Por otra parte, se pueden seguir dos enfoques bien diferenciados a la hora de hacer inferencia. Por una parte, el **enfoque frecuentista** asume que los parámetros son valores fijos desconocidos, de los que estimamos su valor. Esta estimación está ligada a una incertidumbre (error) derivada del muestreo. Por otra parte, en el **enfoque bayesiano** los parámetros no son valores fijos desconocidos, sino variables aleatorias de las que se estima su distribución de proba-

Tabla 1.1: Tabla rectangular bien organizada

| maquina | merma1 | merma2 | manchas | defecto | defecto2 | temp |
|----------|--------|--------|---------|---------|----------|------|
| maquina1 | 5.377 | 4.007 | 11 | No | 0 | 15.7 |
| maquina1 | 6.007 | 4.598 | 7 | Sí | 1 | 18.8 |
| maquina1 | 4.822 | 5.742 | 9 | No | 0 | 13.9 |
| maquina1 | 6.014 | 3.960 | 6 | Sí | 1 | 18.5 |
| maquina1 | 3.892 | 5.268 | 6 | No | 0 | 12.0 |
| maquina1 | 5.379 | 5.913 | 9 | No | 0 | 17.3 |

bilidad. Y a partir de esa distribución de probabilidad, se hace la inferencia. En este libro no se tratarán los métodos bayesianos.

1.2.4. Organización de los datos

Hemos hablado de características de forma aislada. Pero normalmente no estudiamos una sola característica de la población, sino que observamos varias características, teniendo así en la muestra un **conjunto de variables** relativas a una serie de elementos. Cuando analizamos una única variable, aislada del resto, estaremos haciendo análisis **univariante**. Cuando analizamos más de una variable, estaremos haciendo **análisis multivariante**. Casi siempre un estudio estadístico incluye análisis univariante y multivariante.

Para poder analizar los datos de forma eficiente, debemos organizarlos siguiendo los principios *Tidy data*. Así, dispondremos los datos en forma de tablas (datos rectangulares), donde tengamos una columna para cada variable (mismo tipo de datos) y una fila para cada observación (elemento, individuo). El analista y software deben entender lo mismo, lo que podríamos decir que es preparar los datos para las máquinas y no para los humanos. Esta sería la “capa de datos”, después puede haber una “capa de presentación”, independiente de la anterior. Aquí puede jugar un papel importante los metadatos: diccionarios de datos para consultar sobre las variables (unidades, descripciones, etc.)

La tabla ?? muestra las primeras filas de una tabla de datos bien organizada. Cada fila representa un solo elemento, cada columna una sola variable, sin mezclar datos. Los nombres de las variables son cortos pero informativos.



1.2.5. Tipos de datos y escalas

Las características que observamos pueden ser de distintos tipos. La correcta identificación del tipo de variable es crucial para hacer un correcto análisis, ya que los métodos pueden ser muy distintos.

La primera diferenciación que haremos será entre variables **cuantitativas** y

cualitativas. Las variables cuantitativas o numéricas se pueden expresar con un número que además tiene una escala métrica (se pueden medir diferencias entre individuos). A su vez, pueden ser **continuas** o **discretas**. Las variables continuas pueden tomar cualquier valor en un intervalo (teóricamente infinitos valores). Las variables discretas pueden tomar un número de valores finito o infinito numerable, pero no toma valores entre un valor y otro.

Las variables **cualitativas** o categóricas son etiquetas sin sentido numérico en las que podemos clasificar a los elementos. Si el número de posibles etiquetas son dos, estaremos ante variables dicotómicas, que en algunos casos podremos codificar como ceros y unos si presenta o no presenta la característica principal. Las variables multinivel presentan más de dos posibles etiquetas. En ambos casos se trata de una escala nominal. Las variables ordinales son aquellas en las que las etiquetas se pueden ordenar, de forma que tenemos una escala ordinal.

Además de las variables propiamente dichas, nuestro conjunto de datos puede tener otras características como marcas de tiempo e identificadores, que serán útiles para aplicar los métodos, pero no serán objeto de análisis.

En ocasiones es útil transformar las variables de un tipo a otro. Por ejemplo:

- Fechas a categóricas (etiqueta de mes, día de la semana, ...)
- Cuantitativas a cualitativas (clases, intervalos)
- Ordinales como numéricas: con precaución, sobre todo si hay pocos datos (<100). Se pueden combinar en índices.
- Variables calculadas con otras (por ejemplo, IMC)

En los siguientes capítulos abordaremos el análisis de todos estos datos.

1.3. La Estadística y el método científico

La estadística es un pilar fundamental del método científico. El método científico se aplica también en el desarrollo tecnológico. Por tanto, la correcta aplicación de los métodos estadísticos es imprescindible para el avance de la ciencia y la técnica.

1.3.1. El método científico

El método científico se puede resumir en los siguientes pasos:

1. Hacerse una pregunta
2. Realizar investigación de base
3. Plantear una hipótesis
4. Comprobar la hipótesis con experimentos
5. Analizar resultados y extraer conclusiones
6. Comunicar resultados

La pregunta que nos hacemos (1) depende del campo de aplicación, y aquí todavía no aparece la Estadística (a menos que sea una investigación sobre los propios métodos estadísticos). Durante la investigación de base (2), realizamos **análisis exploratorio de datos** e identificamos **relaciones**. Posiblemente, esta primera investigación nos hace cambiar la pregunta del primer paso. Plantear una hipótesis (3) significa formalizarla en términos de Hipótesis nula, H_0 , e hipótesis alternativa, H_1 , que se comprobarán con los **datos** empíricamente. El planteamiento de la hipótesis determina el **método estadístico** a utilizar, y el diseño del experimento (en sentido amplio). Para comprobar la hipótesis con experimentos (4) es fundamental un diseño adecuado para que los resultados sean válidos, así como la correcta **organización de los datos** recogidos según los protocolos establecidos. Estos protocolos incluyen conceptos estadísticos como **aleatorización** y bloqueo, entre otros. Analizar resultados (5a) no se puede hacer sino con técnicas estadísticas, y estos resultados deben contarle al experto la historia con suficiente evidencia para extraer conclusiones (5b). Intervienen aquí el análisis exploratorio, los contrastes de hipótesis y la validación de los modelos. Por último, podemos aprovechar las herramientas estadísticas modernas para comunicar resultados (6), por ejemplo mediante **Informes reproducibles** RMarkdown, Gráficos efectivos y resultados clave. Los resultados negativos (cuando no conseguimos demostrar lo que buscábamos en la hipótesis) es un aspecto a considerar también, para utilizar como lecciones aprendidas y conocimiento general.

1.3.2. Investigación reproducible

Los informes reproducibles mencionados en el párrafo anterior hacen referencia al enfoque de **Investigación reproducible** en el cual se puedan reproducir los resultados, bien los mismos investigadores en otro momento, o terceras partes interesadas para verificar la validez de los resultados. Para esto es necesario utilizar software estadístico basado en *scripts* en los que se pueda consultar toda la lógica del análisis (frente a software de “ventanas” donde se pierde la trazabilidad). Este código se puede mezclar con la propia narrativa del informe (antecedentes, interpretación, conclusiones, etc.) de forma que, dados los mismos datos, se obtenga el mismo informe. Incluso, dados otros datos, se podría replicar el estudio de forma instantánea. El enfoque “copy-paste” alternativo, en el que vamos añadiendo a un informe los resultados en un momento dado, son fuente de inconsistencias, errores, desactualización y falta de reproducibilidad, y en los que cualquier cambio requiere mucho esfuerzo.

1.4. Estadística, Calidad y Sostenibilidad

La es una herramienta fundamental en muchos procedimientos relacionados con la Calidad, y es por eso que se habla de Control Estadístico de la Calidad.

1.4.1. Calidad y variabilidad

Todos tenemos nuestra percepción de la calidad. Pero veamos primero la definición estandarizada de calidad que tenemos en la norma ISO 9001.

Calidad: Grado en el que un conjunto de **características** inherentes de un objeto cumple con los **requisitos**

ISO 9001:2015 3.6.2

Los requisitos son **especificaciones** de la característica, que pueden ser bilaterales o unilaterales.

En la figura ?? vemos dos distribuciones de datos del tipo que vamos a ver en el libro³. Los dos conjuntos de datos correspondientes a la medición de la variable peso tienen **la misma media**: 10 g. Sin embargo, la de la izquierda tiene una **desviación típica** (medida de la variabilidad) igual a 0.6 g, menor que la de la derecha que es 1 g. Si las líneas rojas son nuestros **límites de especificación**, podemos ver cómo en el proceso de la derecha algunos de los elementos de nuestro proceso no satisfacen los requisitos. En este ejemplo se ve claramente cómo reducir la variabilidad mejora la calidad ¡sin hacer nada más! (ni nada menos).

En general, las CTQs (*Critical to Quality* características críticas para la calidad) tendrán un valor objetivo (*target*, T), o valor nominal, que es el ideal. Ante la imposibilidad de tener procesos exactos, se fijan unos límites de especificación o límites de tolerancia dentro de los cuales el producto o servicio es conforme, mientras que es no conforme cuando el valor de la CTQ está fuera de dichos límites. Se utilizan los símbolos L y U para designar los límites de control inferior y superior respectivamente.

La Calidad se mide como la pérdida total que un producto causa a la sociedad

Genichi Taguchi

Debemos considerar que la falta de calidad no produce pérdidas sólo cuando el producto no cumple con las especificaciones, sino que, a medida que nos alejamos del valor objetivo, esa pérdida aumenta, y además no lo hace de manera lineal, es decir, proporcional, sino que es mayor cuanto más nos alejamos del objetivo. Es lo que se conoce como la **función de pérdida de Taguchi** (*Taguchi's Loss Function*). Taguchi consideraba la calidad como la consecución de un objetivo de calidad, no como una tolerancia, y la falta de calidad como una pérdida para la sociedad. El producto *perfecto* no produce pérdidas (*loss*), mientras que cualquier desviación del objetivo produce una pérdida para la sociedad, que aumenta a medida que esa desviación es mayor (?). La figura ?? representa este coste para la sociedad (línea azul discontinua), que se produce siempre que no se consigue el objetivo, frente al coste *contable* (línea punteada gris), que solo se

³Los gráficos son **histogramas**, que también describiremos después.

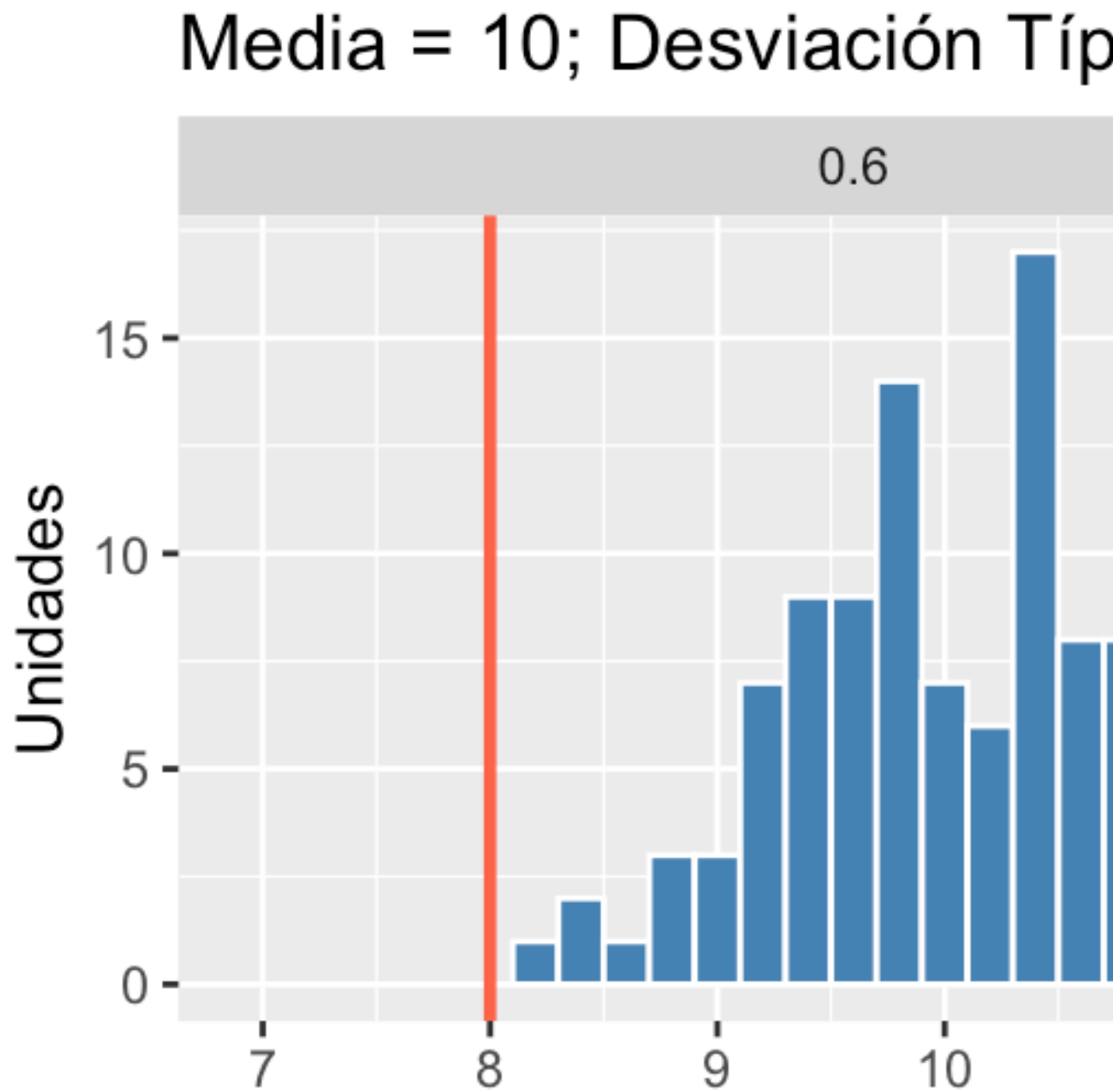


Figura 1.3: Procesos con la misma media y distinta variabilidad

produce con las no conformidades. El análisis de la función de pérdida es una herramienta muy útil en proyectos de mejora, véase ?.

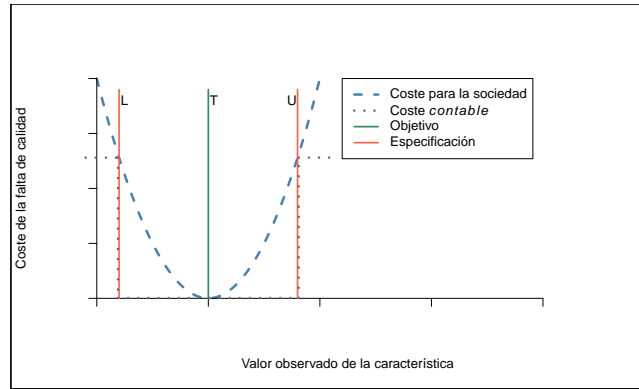
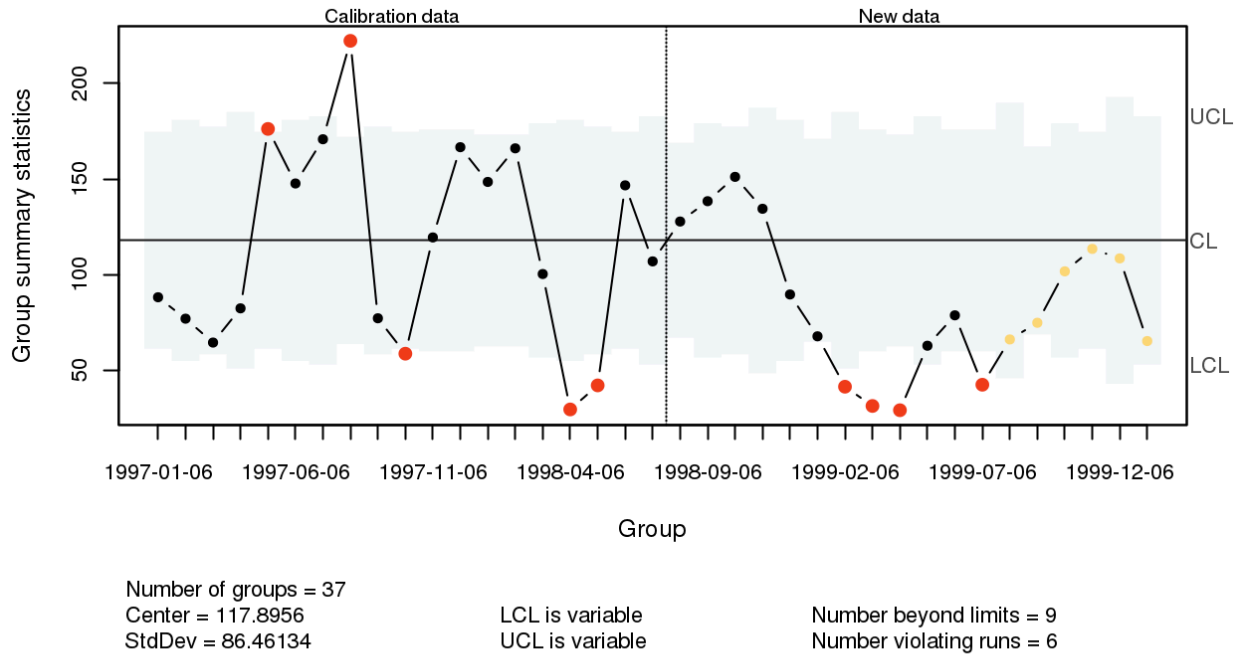
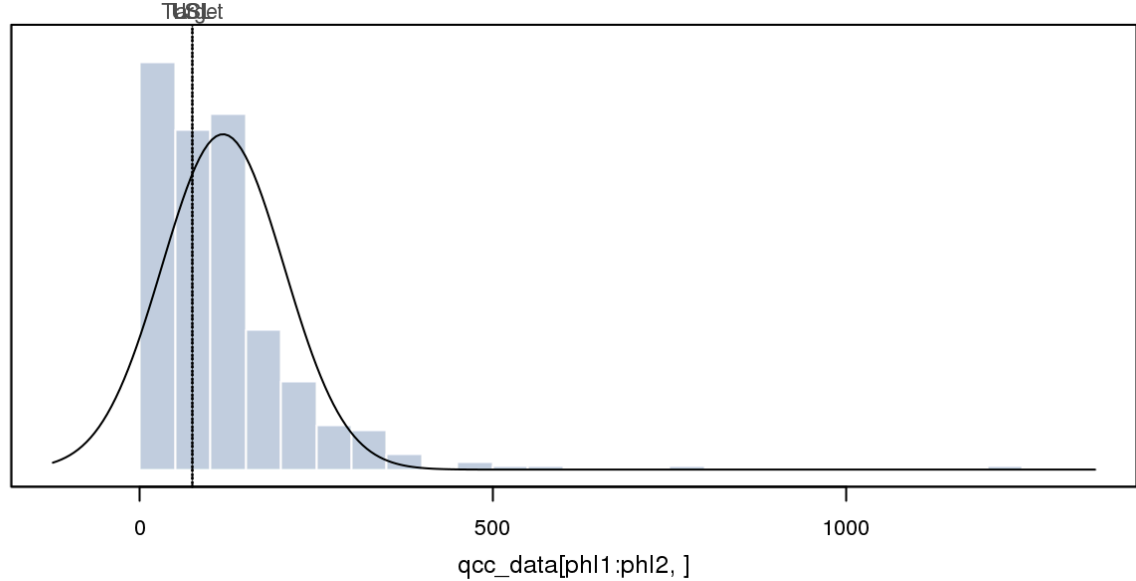


Figura 1.4: Función de pérdida de Taguchi

1.4.2. Métodos estadísticos para la calidad

Existen métodos estadísticos específicos para el control y mejora de la calidad. Las dos principales herramientas del Control Estadístico de Procesos (SPC, *Statistical Process Control*) son los **gráficos de control** y el **análisis de la capacidad del proceso**. La figura ?? muestra un ejemplo de ambas. El gráfico de control de la parte superior sirve para monitorizar las muestras (subgrupos de los que se calcula un estadístico) con el objetivo de detectar el cambio con respecto a su situación de control estadístico. Así, los límites son “la voz del proceso”. La parte inferior representa “la voz de cliente”, comparando las especificaciones con la variabilidad del proceso, y calculando los índices de capacidad que son la medida real de calidad a largo plazo (frente a la mera contabilización de las unidades defectuosas y su cuantificación monetaria). Estas técnicas se combinan con otras tanto exploratorias como de inferencia para controlar y mejorar la calidad.

Otra técnica de calidad en la que la Estadística juega un papel fundamental es la **inspección por muestreo**, también conocida como muestreos de aceptación. La aceptación de unidades o lotes de producto, se puede hacer con inspección completa, comprobando si los productos están dentro de los límites de especificación. Esto a veces es muy caro o directamente imposible, por lo que se recurre al muestreo. El análisis se puede hacer por atributos (variables cualitativas y por variables (variables cuantitativas). La base de estos métodos reside en la probabilidad de aceptar/rechazar un lote defectuoso/correcto, desde el punto de vista del consumidor/productor. Existen una gran variedad de planes de muestreo específicos, como planes simples, planes dobles y múltiples o planes secuenciales. Muchos están descritos en las normas clásicas MIL-STD, que evolucionaron a las series de normas ISO 2859 e ISO 3951.

**Process capability analysis**

Number of obs = 364
 Center = 117.8956
 StdDev = 86.46134

Target = 74
 LSL = 73.95
 USL = 74.05

Cp = 0.000193
 Cp_l = 0.169
 Cp_u = -0.169
 Cp_k = -0.169
 Cpm = 0.000172

Exp<LSL 0.31%
 Exp>USL 0.69%
 Obs<LSL 0.39%
 Obs>USL 0.6%

Figura 1.5: Gráficos de control y capacidad del proceso

En los llamados ensayos inter-laboratorios también se aplican técnicas estadísticas como el análisis del sistema de medición (MSA, *Measurement Systems Analysis*), estudios de precisión y exactitud, estudios R&R (*Reproducibility & Repeatability*), o validación de laboratorios. En la mayoría de los casos lo que se utiliza es Diseño y Análisis de Experimentos.

1.4.3. Metodologías y estándares

Las normas sobre métodos estadísticos que elabora ISO emanan del comité ISO TC69, del que hay un subcomité “espejo” en UNE (entidad acreditada de normalización en España), el subcomité UNE CT66/SC3. La propia ISO 9000 hace mención a los métodos estadísticos, y existe un informe técnico, UNE-ISO TR 1017 sobre “Orientación sobre las técnicas estadísticas para la Norma ISO 9001:2020”. Algunas universidades disponen del catálogo de normas UNE en sus bases de datos para el acceso de docentes y estudiantes.

La metodología Seis Sigma y el ciclo DMAIC aplican el método científico a la mejora de la calidad, utilizando el lenguaje de las empresas. Lean Six Sigma es una evolución en la que se añade a Seis Sigma los principios de *Lean Manufacturing*.

1.5. Objetivos de Desarrollo Sostenible (ODS)

El 25 de septiembre de 2015, los líderes mundiales adoptaron un conjunto de **objetivos globales** para erradicar la pobreza, proteger el planeta y asegurar la prosperidad para todos como parte de una nueva agenda de desarrollo sostenible. Cada objetivo tiene **metas específicas** que deben alcanzarse en los próximos 15 años.

Naciones Unidas

1.5.1. Los 17 ODS

Esta iniciativa de la ONU (*Sustainable Development Goals*, SDG) plantea 17 objetivos generales, que se detallan en 169 metas concretas. Estos objetivos van más allá del medio ambiente, que probablemente es lo primero que nos viene a la cabeza⁴. Los 17 objetivos son los siguientes, y se esquematizan en la figura ??.

1. **Fin de la pobreza** - Poner fin a la pobreza en todas sus formas en todo el mundo
2. **Hambre cero**- Poner fin al hambre, lograr la seguridad alimentaria y la mejora de la nutrición y promover la agricultura sostenible
3. **Salud y bienestar**- Garantizar una vida sana y promover el bienestar para todos en todas las edades

⁴(<https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/>)

4. **Educación de calidad**- Garantizar una educación inclusiva, equitativa y de calidad y promover oportunidades de aprendizaje durante toda la vida para todos
5. **Igualdad de género**- Lograr la igualdad entre los géneros y empoderar a todas las mujeres y las niñas
6. ***Agua limpia y saneamiento****- Garantizar la disponibilidad de agua y su gestión sostenible y el saneamiento para todos
7. **Energía asequible y no contaminante**- Garantizar el acceso a una energía asequible, segura, sostenible y moderna para todos
8. **Trabajo decente y crecimiento económico**- Promover el crecimiento económico sostenido, inclusivo y sostenible, el empleo pleno y productivo y el trabajo decente para todos
9. **Industria, innovación e infraestructura**- Construir infraestructuras resilientes, promover la industrialización inclusiva y sostenible y fomentar la innovación
10. **Reducción de las desigualdades**- Reducir la desigualdad en y entre los países
11. **Ciudades y comunidades sostenibles**- Lograr que las ciudades y los asentamientos humanos sean inclusivos, seguros, resilientes y sostenibles
12. **Producción y consumo responsables**- Garantizar modalidades de consumo y producción sostenibles
13. **Acción por el clima**- Adoptar medidas urgentes para combatir el cambio climático y sus efectos
14. **Vida submarina**- Conservar y utilizar en forma sostenible los océanos, los mares y los recursos marinos para el desarrollo sostenible
15. **Vida de ecosistemas terrestres**- Proteger, restablecer y promover el uso sostenible de los ecosistemas terrestres, gestionar sosteniblemente los bosques, luchar contra la desertificación, detener e invertir la degradación de las tierras y detener la pérdida de biodiversidad
16. **Paz, justicia e instituciones sólidas**- Promover sociedades, justas, pacíficas e inclusivas para el desarrollo sostenible, proporcionar a todas las personas acceso a la justicia y desarrollar instituciones eficaces, responsables e inclusivas en todos los niveles
17. **Alianzas para lograr objetivos**- Fortalecer los medios de ejecución y revitalizar la Alianza Mundial para el Desarrollo Sostenible

1.5.2. Estadística y sostenibilidad

La Estadística, y su aplicación en la Ciencia y la Ingeniería, puede hacerse presente en los ODS. Algunos ejemplos serían los siguientes:

- Al realizar investigación sobre algún aspecto de los ODS, irremediablemente utilizaremos la Estadística. Nos podemos proponer nuestras propias líneas de investigación y desarrollo tecnológico desde el punto de vista de uno o varios ODS
- Tener presentes los ODS para ser sostenible en los propios análisis. Por



Figura 1.6: Objetivos de Desarrollo Sostenible. Fuente: un.org

ejemplo reduciendo el uso de papel o energía, pero también utilizando lenguaje inclusivo o teniendo en cuenta a minorías.

- Relacionar con ODS e intentar contribuir sea cual sea el objetivo de la investigación
- Siempre podemos hacernos la pregunta: ¿Cómo puede contribuir este trabajo/estudio/investigación/... a conseguir los Objetivos de Desarrollo Sostenible?

Capítulo 2

Análisis exploratorio univariante

2.1. La importancia del análisis exploratorio

El análisis exploratorio de datos, y en particular su visualización, es el primer análisis que se debe hacer sobre cualquier conjunto de datos antes de abordar otras técnicas estadísticas, sean sencillas o complejas. La “historia” que nos esté contando el gráfico de los datos, nos guiará hacia las técnicas de aprendizaje estadístico más adecuadas. Incluso, en muchas ocasiones será suficiente el análisis exploratorio para tomar una decisión sobre el problema en estudio. La figura ?? representa la esencia de la Estadística y sus métodos. Los datos que analizamos, provienen de una determinada **población**. Pero estos datos, no son más que una **muestra**, es decir, un subconjunto de toda la población. Incluso cuando “creemos” que tenemos todos los datos, debemos tener presente que trabajamos con muestras, ya que generalmente tomaremos decisiones o llegaremos a conclusiones sobre el futuro, y esos datos seguro que no los tenemos. Por eso es importante considerar siempre este paradigma población-muestra. La **Estadística Descriptiva** se ocupa del análisis exploratorio de datos en sentido amplio, que aplicaremos sobre los datos concretos de la muestra en este capítulo y el siguiente. La **Inferencia Estadística** hace referencia a los métodos mediante los cuales, a través de los datos de la muestra, tomaremos decisiones, explicaremos relaciones, o haremos predicciones sobre la población. Para ello, haremos uso de la **Probabilidad**, que veremos en el capítulo ??, aplicando el método más adecuado. En estos métodos será muy importante considerar el método de obtención de la muestra que, en términos generales, debe ser representativa de la población para que las conclusiones sean válidas. En este tercer módulo del curso veremos algunos de estos métodos.

El análisis exploratorio se realiza básicamente mediante dos herramientas: los

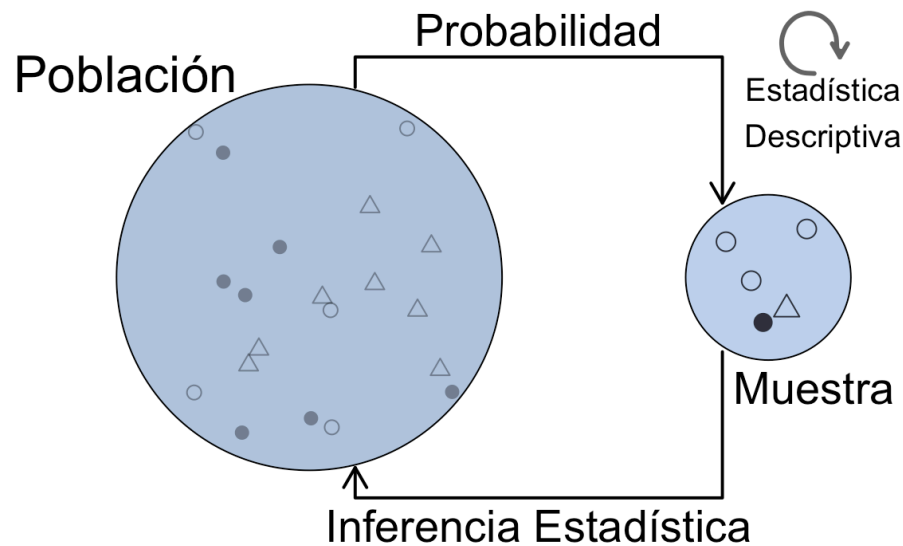


Figura 2.1: La esencia de los métodos estadísticos

resúmenes numéricos y las visualizaciones gráficas. Pero antes de aprender a hacer análisis exploratorio con R, vamos a resaltar la importancia, dentro del análisis exploratorio, de las representaciones gráficas. Para ello utilizaremos un conjunto de datos llamado “el cuarteto de Anscombe” (?), disponible con el nombre `anscombe` en el paquete `datasets` de R base. La tabla ?? muestra este conjunto de datos.

Son 11 filas de 8 variables numéricas, aunque las tres primeras son idénticas. Ya sabemos resumir los datos con la media de cada variable:

```
library(dplyr)
anscombe %>% summarise(across(.fns = mean))
#> Warning: There was 1 warning in `summarise()`.
#> i In argument: `across(.fns = mean)`.
#> Caused by warning:
#> ! Using `across()` without supplying `.cols` was deprecated
#> in dplyr 1.1.0.
#> i Please supply `.cols` instead.
#>   x1 x2 x3 x4      y1      y2 y3      y4
#> 1  9  9  9  9 7.500909 7.500909 7.5 7.500909
```

Vemos que la media de las cuatro primeras variables es idéntica, 9. Pero los datos son muy distintos en la cuarta variable. Las cuatro últimas variables también tienen una media prácticamente idéntica. Sin embargo los datos también son muy distintos. La figura ?? es un gráfico de los que aprenderemos a hacer enseguida, y representa en el eje vertical los valores de las variables, y en el eje

Tabla 2.1: Conjunto de datos 'anscombe'

| x1 | x2 | x3 | x4 | y1 | y2 | y3 | y4 |
|----|----|----|----|-------|------|-------|-------|
| 10 | 10 | 10 | 8 | 8.04 | 9.14 | 7.46 | 6.58 |
| 8 | 8 | 8 | 8 | 6.95 | 8.14 | 6.77 | 5.76 |
| 13 | 13 | 13 | 8 | 7.58 | 8.74 | 12.74 | 7.71 |
| 9 | 9 | 9 | 8 | 8.81 | 8.77 | 7.11 | 8.84 |
| 11 | 11 | 11 | 8 | 8.33 | 9.26 | 7.81 | 8.47 |
| 14 | 14 | 14 | 8 | 9.96 | 8.10 | 8.84 | 7.04 |
| 6 | 6 | 6 | 8 | 7.24 | 6.13 | 6.08 | 5.25 |
| 4 | 4 | 4 | 19 | 4.26 | 3.10 | 5.39 | 12.50 |
| 12 | 12 | 12 | 8 | 10.84 | 9.13 | 8.15 | 5.56 |
| 7 | 7 | 7 | 8 | 4.82 | 7.26 | 6.42 | 7.91 |
| 5 | 5 | 5 | 8 | 5.68 | 4.74 | 5.73 | 6.89 |

horizontal los nombres de cada variable. Vemos que, a pesar de tener medias prácticamente iguales, los datos son muy diferentes.

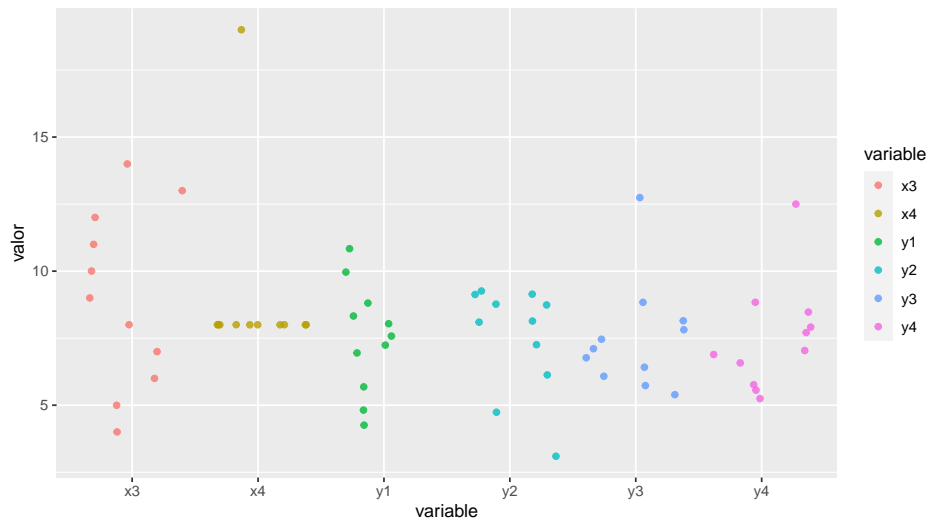


Figura 2.2: Representación de las variables del cuarteto de Anscombe

Pero si en el análisis por separado ya se ve la necesidad de hacer un gráfico, cuando analizamos las variables conjuntamente, todavía es más evidente. La figura ?? muestra los cuatro gráficos que constituyen “El cuarteto de Anscombe”, y que se puede obtener de la propia ayuda del conjunto de datos (`example(anscombe)`). La línea de regresión que se ajusta es prácticamente la misma (veremos la regresión en el capítulo ??). Además, si calculáramos los coeficientes de correlación entre las variables “x” e “y” de los cuatro gráficos, obtendríamos el mismo valor: 0,8163.

Anscombe's 4 Regression data sets

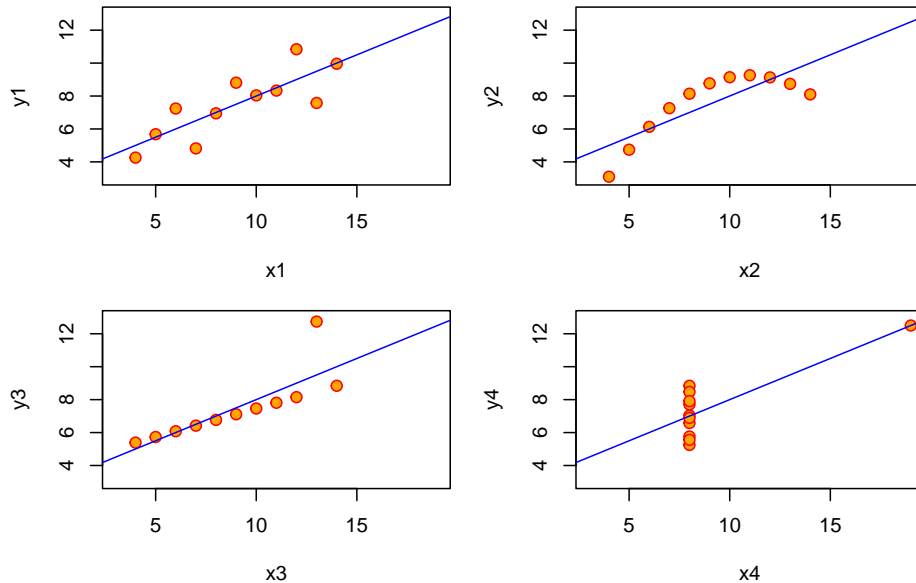


Figura 2.3: Los cuatro gráficos que constituyen ‘El cuarteto de Anscombe’

Es evidente que la relación entre las variables es muy distinta en cada uno de los casos, y si no visualizamos los datos para elegir el mejor modelo de regresión y después interpretarlo, podemos estar tomando decisiones erróneas.



El cuarteto de Anscombe es muy ilustrativo, os animo a explorar también *The Datasaurus Dozen*: (?) en <https://www.autodeskresearch.com/publications/samestats>.

2.2. Calidad de datos

Una vez hemos identificado los tipos de variables del problema de análisis de datos que queremos abordar, es necesario que tengamos los datos correctamente en el software que vamos a utilizar, es decir, es muy importante comprobar continuamente la **calidad en los datos**. La importación de datos siempre puede dar problemas (y por *Murphy*, los dará). Por eso siempre deberíamos comprobar la estructura de los datos después de importar un conjunto de datos (al menos la primera vez). Uno de los errores más comunes es que el tipo de datos importado no se corresponda con el que conceptualmente debe tener la variable. Esto no produce ningún error al importar, pero sí al analizar los datos. Otros problemas de calidad tienen que ver con valores atípicos (*outliers*) y con valores perdidos

(*missing*).

2.2.1. Datos atípicos

A medida que llevamos el análisis de datos a aplicaciones reales, es más fácil que aparezcan observaciones que *estropean* el análisis porque se salen de lo esperado en relación con el resto de datos. La parte 4 de norma UNE-ISO 16269 (?), un **valor atípico** es un “Miembro de un pequeño subconjunto de observaciones que parece ser inconsistente con el resto de una muestra dada”. La identificación de valores **candidatos** a ser considerados como atípicos es una labor muy importante para el analista, ya que pueden influir tanto en los resultados del análisis como en la técnica a utilizar. Estos valores identificados como posibles valores atípicos deben ser investigados y determinar cuál es la causa de esta posible desviación. Se suele atribuir a una de las siguientes causas:

1. *Error de medida o de registro.* Esto puede ser debido a la observación del dato o al propio registro.
2. *Contaminación.* Los datos provienen de más de una distribución. Por ejemplo, por estar mezclando datos de grupos que tienen distintas medias. Entonces, los valores de la distribución *contaminante* aparecerán como valores atípicos en la distribución de interés.
3. *Suposición incorrecta sobre la distribución.* La característica en estudio de la población se supone que sigue una determinada distribución (por ejemplo normal) pero en realidad sigue otra (por ejemplo exponencial). Entonces los valores que *parecen* atípicos para la distribución normal, son perfectamente compatibles con la distribución verdadera.
4. *Observaciones excepcionales.* Estos no son verdaderos valores atípicos, simplemente han ocurrido por azar, aunque sea muy improbable su ocurrencia.

En el primer caso, hay que encontrar el valor correcto y si esto no es posible, dar el valor por perdido (*missing*). En el segundo, hay que estratificar los datos y realizar el análisis por grupos, separando las distribuciones. Si son solo unos pocos datos los que por error han contaminado la muestra, se pueden eliminar o dar por perdidos. En el tercer caso, se modifican las asunciones sobre el modelo de distribución subyacente en la población. En el último caso los valores deberían permanecer en la muestra, aunque generalmente se etiquetan erróneamente como valores atípicos por su excepcionalidad.

El análisis de los valores atípicos es importante por varios motivos. Por una parte, puede dar lugar a descubrimientos interesantes al investigar por qué han ocurrido (por ejemplo, se ha hecho algo diferente y un proceso ha mejorado). Por otra parte, muchas medidas y métodos estadísticos son muy sensibles a observaciones atípicas, y entonces es posible que haya que usar alternativas robustas. Y en todo caso, nos ayuda a determinar la adecuada distribución de probabilidad.

La observación de los datos con métodos gráficos a menudo proporciona suficiente información para identificar valores candidatos a ser atípicos. En concreto, el gráfico de cajas diseñado por John W. Tukey (?) y recogido en la norma UNE-ISO 16269 (?) marca estos valores de forma clara (véase el apartado ?? para una completa explicación de su construcción e interpretación).

Aparte de los métodos gráficos, existen diversos contrastes de hipótesis para determinar si existen valores atípicos en una muestra de datos dada una distribución de probabilidad. La norma UNE-ISO 16269 (?) recoge métodos para la distribución normal y también para otros modelos de distribución, así como un método general para distribuciones desconocidas y el test de Cochran para varianza atípica. El paquete `outliers` de R contiene varias funciones para realizar contrastes de hipótesis sobre valores atípicos a un conjunto de datos, incluidos el test de Grubbs y el test de Cochran.

En cuanto al tratamiento de datos que contienen valores candidatos a ser atípicos pero de los que no se ha podido identificar una causa válida para eliminarlos, deberíamos recurrir al **análisis de datos robusto**, de forma que las observaciones atípicas no influyan demasiado en los resultados, pero sin eliminarlas. Otra alternativa es realizar el análisis con y sin valores candidatos a ser atípicos y comprobar cómo varía ese resultado.

Entre las medidas de centralización robustas se encuentran la mediana y la media recortada (véase ??), aunque hay otras. También para la estimación de la dispersión se encuentran estimadores robustos como la Mediana de las medianas de las desviaciones absolutas de los pares (?).

Lo dicho hasta ahora sirve para detectar atípicos para una característica. En conjuntos multivariantes, se pueden observar valores atípicos con respecto a más de una variable. En particular, en modelos de regresión puede haber observaciones influyentes (que posiblemente no son atípicas en la variable aislada) que influyen en la estimación de los parámetros de forma que el resultado no es representativo del conjunto de datos. Los gráficos de diagnóstico de R para los modelos lineales proporcionan un gráfico señalando las observaciones influyentes según la distancia de Cook. También el paquete `car` contiene una función (`outlierTest`) con la que podemos obtener la observación más extrema para la regresión.

Por último, podemos detectar observaciones atípicas con respecto a todo un conjunto multivariante de datos en escala métrica. Para ello, lo que se hace es reducir este conjunto multivariante en univariante, obteniendo unas distancias de las observaciones a la media muestral del conjunto de datos, estandarizada mediante la matriz de varianzas-covarianzas de la muestra. Entonces aquellas observaciones muy alejadas de esos valores centrales pueden estudiarse como candidatos a ser valores atípicos multivariantes. En ? se proporciona un contraste de hipótesis y un método gráfico para identificar estos valores atípicos. En el apartado ?? se proporcionan las funciones necesarias para calcular la distancia de Mahalanobis.

2.2.2. Valores perdidos (missing values)

La ausencia de valores para determinadas observaciones de nuestra muestra es otro de los problemas habituales que surgen con los datos. Al igual que con los valores atípicos, un valor perdido puede ser fruto de un error en la recogida o registro de los datos. Si ese error es recuperable, bastará con añadir el verdadero valor a nuestro conjunto de datos. Si el valor se da definitivamente por perdido, entonces podemos seguir dos caminos:

1. Realizar el análisis sin considerar las observaciones con valores perdidos.
2. Imputar un valor a las observaciones perdidas.

El primer caso merece la siguiente consideración. Cuando estamos analizando una sola característica, este camino es único. Por ejemplo, en un conjunto de 100 observaciones donde faltan 2, se calcula la media con las 98 restantes. O en un gráfico, se representan solo los valores existentes. Pero cuando estamos analizando un conjunto multivariante, podemos tener valores perdidos en todas las variables, o solo en algunas. Entonces podemos tomar diferentes decisiones a este respecto. Por ejemplo, si queremos calcular una matriz de correlaciones, podemos considerar solo las observaciones en las que hay valores para todas las variables, o eliminar solo los pares de observaciones relevantes para cada coeficiente de correlación entre dos variables¹.

El segundo camino es más complicado y requiere a su vez elegir el método de imputación del valor perdido. La imputación más sencilla es simplemente asignar la media o la mediana como valor representativo de toda la variable. Pero cuando tenemos conjuntos multivariantes, puede ser más adecuado hacer una imputación en función de la información disponible en otras variables. Por ejemplo, si tenemos una variable de tipo atributo, la media del grupo al que pertenece la observación será generalmente más adecuada que la media global.

En R tenemos varias alternativas para la imputación de valores perdidos. La función `impute` del paquete `Hmisc` realiza imputaciones sencillas (por defecto la mediana). El paquete `mice` realiza imputaciones utilizando datos multivariantes con un buen número de opciones.

La investigación de los valores perdidos y su tratamiento adecuado debe ser siempre una fase importante del proyecto de análisis de datos. Además, este análisis se puede solapar con el análisis de los valores atípicos, por ejemplo cuando un valor atípico se determina que es un dato erróneo pero no podemos asegurar cuál es el valor verdadero, entonces tenemos que considerarlo como perdido y aplicar lo aquí visto.

2.2.3. Errores comunes

Aparte de los errores en los datos que ya se han tratado, hay que evitar algunos errores demasiado comunes a la hora de abordar el análisis de datos, y especial-

¹En R, la función `cor` controla este comportamiento mediante el argumento `use`.

mente la interpretación de resultados. En este apartado se mencionan algunos de los más importantes.

1. Confundir correlación con causalidad.

Cuando realizamos una regresión de una variable respuesta Y sobre una o varias variables *explicativas* X , tendemos a pensar que X es la causa de la variación de Y . Esto no siempre es así, y deberíamos tenerlo presente incluso en aplicaciones en las que conocemos los procesos y “estamos seguros” de que es así. Para confirmar que una relación es de causa-efecto, deberíamos recurrir al Diseño de Experimentos, donde además podremos estudiar las interacciones.

2. Falta de parsimonia.

La parsimoniosidad es un principio científico (véase ?)² que, aplicado a la Estadística, significa seleccionar el modelo más reducido y simple posible que consiga explicar el fenómeno a estudiar, frente a modelos más complejos (con muchas variables) con una mínima o nula ganancia de poder predictivo. En modelos de regresión múltiple, por ejemplo, ninguna variable cuyo coeficiente no sea significativo se debería incluir en el modelo final al que llega la investigación.

2. Interpretación de porcentajes sin fijarse en el tamaño.

Este error común viene explicado por la **paradoja de Simpson** (?). Esta paradoja aparece cuando hay un atributo *oculto* que no se tiene en cuenta a la hora de interpretar porcentajes, pudiendo darse el caso de que otro atributo presenta un porcentaje mayor en una categoría, pero si se analizan por separado los porcentajes para las categorías del atributo oculto, resulta que el porcentaje de la categoría que era mayor globalmente, es menor en TODOS los grupos del atributo oculto.

3. Informar los valores medios pero no la dispersión.

La media por sí sola no debería llevar a conclusión alguna. Siempre se debe analizar conjuntamente la centralidad y dispersión de los datos, ya que un valor medio puede estar calculado con valores muy extremos y ocultar mucha información.

4. Pasar por alto las hipótesis del modelo.

Muchos modelos estadísticos requieren, para ser válidos, que se cumplan ciertas condiciones. Si utilizamos un método que requiere normalidad, debemos comprobar que los datos provienen de una distribución normal. Ante la duda, debemos comprobar que un método no paramétrico conduce a resultados similares.

5. Sobreajuste (*overfitting*).

El sobreajuste aparece cuando en un modelo predictivo conseguimos estimar perfectamente los valores de la muestra, pero el modelo utilizado no sirve para generalizar a nuevos casos. En *Machine Learning* es muy fácil conseguir un

²En igualdad de condiciones, la explicación más sencilla suele ser la más probable.

modelo perfecto para los datos utilizados, pero pésimo para nuevos casos. El paradigma de entrenamiento y validación consigue evitar el sobreajuste.

6. Utilizar muestras sesgadas como si fueran aleatorias

Los métodos probabilísticos de uno u otro modo se basan en que los datos provienen de muestras aleatorias. A pesar de que en muchas situaciones de análisis de datos esto no lo podamos ni siquiera soñar, es importante tenerlo en mente para, a la hora de interpretar resultados y llegar a conclusiones, hacer una reflexión sobre cuánto nos estamos alejando de esa aleatoriedad. Por ejemplo, si estoy haciendo un estudio de los clientes de una empresa y solo analizo las transacciones de la primera semana del mes, tengo una muestra sesgada porque no tengo representado el resto del mes (posiblemente con un comportamiento diferente).

2.3. Componentes de un gráfico

Dejando aparte las visualizaciones en tres dimensiones, animaciones 3D y realidad virtual, la visualización de datos que hacemos en la práctica totalidad de los casos es en dos dimensiones, es decir, en el plano. Vamos a pensar en este plano como si fuera un “lienzo” de pintor, independientemente de que el resultado lo vayamos a ver impreso en un papel o en una pantalla. Este lienzo se irá “poblando” de “capas” a medida que el pintor vaya añadiendo cosas. Siguiendo con el símil, empezaremos preparando un espacio para los símbolos con los que representaremos los datos, es decir, unos **ejes**: horizontal (X) y vertical (Y). A partir de aquí, representaremos los datos con algún **símbolo geométrico**, como un punto, una línea, o cualquier otro. Podremos añadir colores a los símbolos y otras características como transparencia o tamaño. También añadiremos anotaciones al gráfico, como las marcas en los ejes, títulos o incluso texto dentro del gráfico.

La figura ?? es una ilustración de Allison Horst³ que simboliza este paradigma de lienzo y capas. Si pensamos en los distintos elementos del gráfico y los relacionamos con las variables que estamos analizando, será mucho más fácil hacer el gráfico adecuado e interpretarlo.

2.4. Notación

Antes de comenzar a hacer resúmenes de los datos, vamos a definir la notación que utilizaremos. Representamos las variables con letras mayúsculas latinas del final el alfabeto como X, Y, \dots ⁴. Cada uno de los posibles valores que toma la variable X se representa por x_i . Así, i es el identificador o índice para cada observación o clase. El número total de observaciones **en la muestra** lo representamos por n , mientras que si tenemos una enumeración de toda la población

³<https://github.com/allisonhorst/stats-illustrations>

⁴Para atributos, a veces se utilizan las primeras letras del alfabeto: A, B, \dots



Figura 2.4: El dispositivo gráfico como lienzo al que añadimos capas

en estudio, denotaremos el número total de individuos por N . El número de clases o niveles de una variable categórica o numérica agrupado es k . $n_i, i = 1, \dots, k$ es el número de observaciones en la clase i . Si agrupamos los datos numéricos en intervalos (clases), $c_i, i = 1, \dots, k$ es la marca de clase, es decir, el punto central del intervalo.

Para representar los parámetros (recordemos, desconocidos) utilizamos letras griegas. Por ejemplo, μ es la media poblacional, y σ^2 la varianza poblacional. Para representar estadísticos (recordemos, calculados con los datos de la muestra) se representan con letras minúsculas. Por ejemplo, \bar{x} es la media muestral de la variable X , y s^2 : representa la varianza muestral (cuasivarianza). s es la desviación típica muestral

Para representar que un estadístico es un estimador, utilizamos la notación $\hat{[\cdot]}$, que simboliza un estimador de \cdot . Por ejemplo, $s = \hat{\sigma}$ quiere decir que la desviación típica muestral s es un estimador de la desviación típica poblacional σ .

🌲 Supongamos que tenemos que hacer un estudio de las emisiones de dióxido de carbono (CO_2) en las granjas de porcino de una determinada región. Este es un ejemplo en el que podemos enumerar la población (a partir de registros oficiales u otras fuentes). Supongamos que existen 1 000 granjas. Entonces, $N = 1\,000$. En vez de analizar el 100 % de las granjas, se decide hacer un muestreo, por ejemplo, del 10 % de las granjas^a. Entonces, $n = 100$. La región está dividida en tres zonas, y definimos el atributo $A \in \{Z1, Z2, Z3\}$. Entonces, para este atributo $k = 3$. Si en la muestra tenemos el doble de granjas en la zona 1 que en cualquiera de las otras dos, entonces $n_1 = 50, n_2 = 25$ y $n_3 = 25$.

Una vez realizadas las mediciones de emisiones en cada granja de la muestra, tendremos valores $x_i, i = 1, \dots, n$. Podremos agrupar estos valores en k' intervalos (clases) de los que podremos calcular las marcas de clase $c_i, i = 1, \dots, k'$. Como solo hemos medido las emisiones en una muestra, desconocemos el verdadero valor de la media de la población, μ , y entonces lo estimaremos con la media muestral: $\hat{\mu} = \bar{x}$.



^aEn el capítulo ?? estudiaremos cómo decidir el tamaño de la muestra.

2.5. Análisis exploratorio de variables cualitativas

Cuando nuestra variable no se expresa con números, sino con etiquetas de una determinada característica observada en cada uno de los elementos en los que se observa la característica, el resumen numérico que utilizamos es la tabla de frecuencias. Esta tabla de frecuencias se puede representar gráficamente con un gráfico de barras o con un gráfico de sectores. Este último no es recomendable ya que proporciona la misma información que el gráfico de barras y es mucho más difícil para el ojo humano distinguir ángulos que alturas. En variables cualitativas, llamamos a la categoría más frecuente **moda** de la variable.

Para construir la tabla de frecuencias, contamos el número de elementos de cada clase (n_i) que pertenecen a cada una de las clases (c_i), que son las **frecuencias absolutas**. Se pueden calcular también frecuencias relativas ($f_i = n_i/n$) y acumuladas, tanto para las absolutas (N_i) como para las relativas (F_i). No obstante, estas frecuencias acumuladas solo tienen sentido cuando la variable está en escala ordinal.

Los datos que se utilizarán en este capítulo para ilustrar los ejemplos se pueden descargar e importar con el siguiente código.

```
library(dplyr)
download.file("https://lcano.com/data/eaci/lab.xlsx",
             destfile = "lab.xlsx")
lab <- readxl::read_excel("lab.xlsx") |>
mutate(fecha = as.Date(fecha))
```



El laboratorio de una fábrica de quesos recoge datos de los análisis realizados a muestras de quesos de su producción. Se dispone de un conjunto de datos con 1171 filas y 12 columnas. La tabla ?? muestra las primeras filas de este conjunto de datos.

La columna **tipo** toma tres valores: A, B y C. La tabla ?? muestra una tabla de frecuencias completa, donde se puede ver de un vistazo, por ejemplo, que la clase con más quesos en el conjunto de datos es el tipo C. Las frecuencias relativas se pueden traducir fácilmente a porcentajes.



```
#> Warning: Since gt v0.6.0 the `fmt_missing()` function is deprecated
#> and will soon be removed.
#> * Use the `sub_missing()` function instead.
```

| fecha | codigo | est | mg | sal | ph | ebacteria | analista | tipo | bacteriax | imperfe |
|------------|--------|-------|------|------|------|-----------|------------|------|-----------|---------|
| 2013-11-01 | 1 | 33.50 | 14.0 | | 6.64 | <10 | analista_9 | C | 8606 | |
| 2013-11-01 | 2 | 31.05 | 13.0 | | 6.65 | <10 | analista_9 | C | 3055 | |
| 2013-11-01 | 3 | 31.42 | 13.0 | 1.20 | 6.66 | <10 | analista_9 | C | 17153 | |
| 2013-11-01 | 4 | 31.00 | 13.0 | | 6.60 | <10 | analista_9 | C | 46089 | |
| 2013-11-01 | 5 | 31.54 | 13.5 | | 6.60 | <10 | analista_9 | C | 6488 | |
| 2013-11-01 | 6 | 30.51 | 12.5 | | 6.63 | <10 | analista_9 | C | 9639 | |
| 2013-11-01 | 7 | 32.30 | 13.0 | | 6.64 | <10 | analista_9 | C | 1398 | |
| 2013-11-01 | 8 | 31.27 | 12.5 | | 6.63 | <10 | analista_9 | C | 14768 | |
| 2013-11-01 | 9 | 31.10 | 12.5 | 1.14 | 6.62 | <10 | analista_9 | C | 6644 | |
| 2013-11-01 | 10 | 30.76 | 12.5 | | 6.64 | <10 | analista_9 | C | 1887 | |

| tipo | n | f | N | F |
|------|-----|------|------|------|
| A | 175 | 0,15 | 175 | 0,15 |
| B | 148 | 0,13 | 323 | 0,28 |
| C | 848 | 0,72 | 1171 | 1,00 |

R

La función `table` de R crea tablas de frecuencias absolutas. Si el resultado se lo pasamos a la función `prop.table()`, las convierte en tabla de frecuencias relativas. La función `addmargins()` añade totales. Para obtener frecuencias acumuladas, podemos usar la función `cumsum`.

Las expresiones siguientes son ejemplos de uso de estas funciones. La tabla ?? se ha obtenido utilizando funciones del paquete `dplyr`:

```
lab |> count(tipo) |>
  mutate(f = n/nrow(lab), N = cumsum(n),
         F = cumsum(f))
```



```
table(lab$tipo)
#>
#>   A   B   C
#> 175 148 848
prop.table(table(lab$tipo))
#>
#>           A           B           C
#> 0.1494449 0.1263877 0.7241674
addmargins(table(lab$tipo))
#>
#>   A   B   C Sum
#> 175 148 848 1171
cumsum(table(lab$tipo))
#>   A   B   C
#> 175 323 1171
```

La representación gráfica adecuada para variables cualitativas es el **gráfico de barras**. En este gráfico, representamos las categorías en el eje horizontal (X) y las frecuencias en el eje vertical (Y), y representamos barras cuya altura representa la frecuencia. Se pueden representar frecuencias absolutas o relativas. Los gráficos de sectores también pueden representar variables cualitativas, aunque no se recomiendan porque el ojo humano no es tan bueno distinguiendo ángulos como alturas. En todo caso, si se usa se deberían incluir los valores (frecuencias o porcentajes). El gráfico de barras se puede representar también invirtiendo los ejes (a veces mejora la visualización de las etiquetas), representando líneas en vez de barras, u ordenando las barras según la frecuencia (por defecto este orden es arbitrario, muy a menudo alfabético según las etiquetas).

Un aspecto importante de los gráficos de barras es que debe haber un espacio entre las barras, puesto que son variables cualitativas en las que no tiene sentido representar la continuidad que expresarían las barras adyacentes.

R

La tabla de frecuencias ?? se puede representar con el siguiente código cuyo resultado se muestra en la figura ??.

El segundo fragmento de código produce la figura ??, que representa un gráfico de sectores con etiquetas realizado con el paquete {ggs-tatsplot}.



```
library(ggplot2)
lab |>
  ggplot(aes(x = tipo)) +
  geom_bar(fill = "#CB0017") +
  theme_bw() +
  labs(title = "Tipos de queso",
       x = "Tipo",
       y = "Frecuencia absoluta")
```

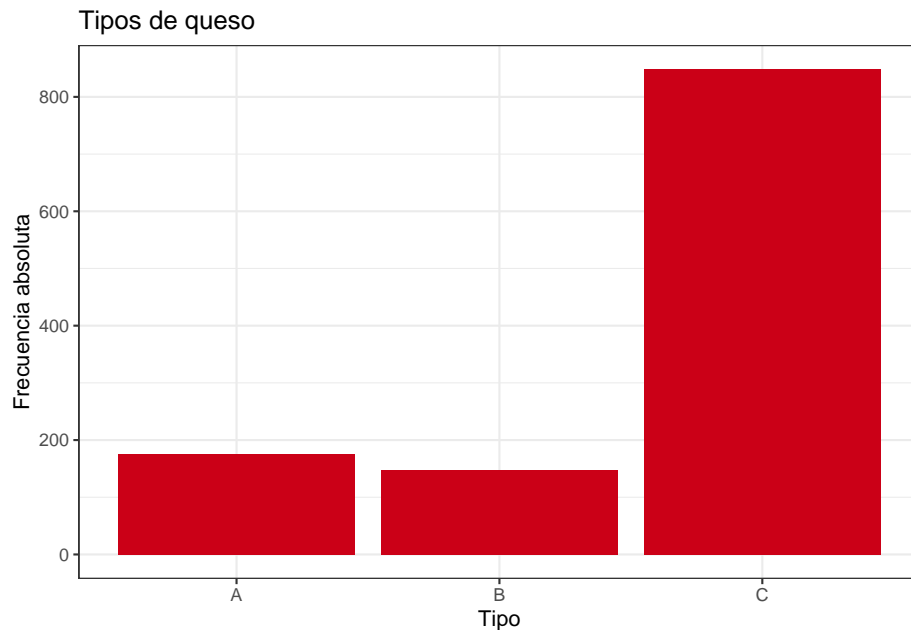


Figura 2.5: Ejemplo gráfico de barras variable cualitativa

```
library(ggstatsplot)
lab %>% ggpiestats(x = tipo, title = "Fabricación de quesos",
                  legend.title = "Tipo de queso",
                  bf.message = FALSE,
                  results.subtitle = FALSE)
```

Fabricación de quesos

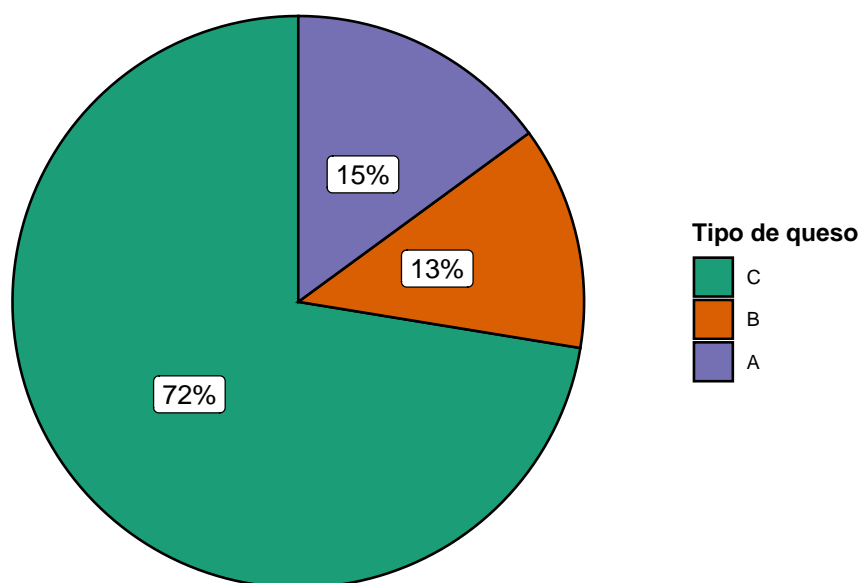


Figura 2.6: Gráfico de sectores con etiquetas

2.6. Análisis exploratorio de variables cuantitativas

2.6.1. Resúmenes de variables discretas

En el caso de variables discretas, se puede realizar el mismo análisis exploratorio que para las variables categóricas, es decir, una tabla de frecuencias y su correspondiente gráfico de barras. En este caso denotamos cada uno de los posibles valores como $x_i, i = 1, \dots, k$, siendo k el número de valores distintos que toma la variable discreta. La diferencia principal es que en este caso la tabla y el gráfico deben estar ordenados de mayor a menor según los valores numéricos que toma la variable. Aquí las frecuencias acumuladas cobran más sentido, sobre todo las relativas. Así, F_i se pueden interpretar como la proporción (o porcentaje si multiplicamos por cien) de observaciones que toman valores menores o iguales que x_i . La idea detrás de este concepto es muy importante y nos volverá a aparecer en el capítulo ?? cuando definamos la función de distribución de probabilidad.

En cuanto al gráfico, de nuevo aquí es importante decir que debe haber una separación entre las barras, porque por su naturaleza, no hay valores entre un valor y otro de la variable, y así queda bien representado que es una variable discreta.

Cuando el número de posibles valores es muy grande, aunque la variable sea discreta se puede tratar como si fuera continua, resumiendo en tablas de frecuencias por intervalos e histogramas, para facilitar su interpretación. Pero no se debe perder nunca de vista la naturaleza de la variable.

En variables discretas, también podemos resumir los datos con el valor más frecuente, es decir, la **moda**. También se podrán resumir los datos mediante los estadísticos y con el gráfico de cajas que se explicarán en el apartado siguiente de variables continuas.



La variable **imperfecciones** es un recuento de defectos en una inspección visual. Vemos en la tabla de frecuencias ?? que tenemos 10 valores posibles: desde cero imperfecciones hasta 9 imperfecciones. La moda es el 2, ya que es el valor que más se repite. Además, es única. Vemos además que el 94,9 % de los quesos tienen 4 o menos imperfecciones. O lo que es lo mismo, el 5,1 % de los quesos tiene más de 4 imperfecciones. La figura ?? es la representación gráfica de esta tabla de frecuencias, en este caso representada en horizontal.

| $\backslash (x_i \backslash$ | $\backslash (n_i \backslash$ | $\backslash (F_i \backslash$ |
|------------------------------|------------------------------|------------------------------|
| 0 | 146 | 0,125 |
| 1 | 312 | 0,391 |
| 2 | 339 | 0,681 |

| | | |
|---|-----|-------|
| 3 | 215 | 0,864 |
| 4 | 99 | 0,949 |
| 5 | 41 | 0,984 |
| 6 | 14 | 0,996 |
| 7 | 1 | 0,997 |
| 8 | 3 | 0,999 |
| 9 | 1 | 1,000 |

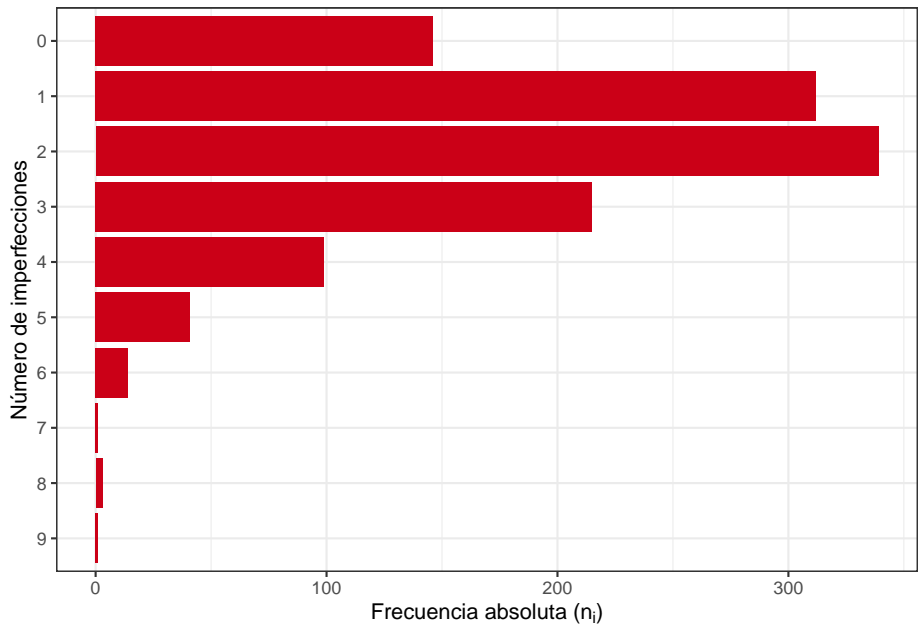


Figura 2.7: Gráfico de barras de la variable discreta imperfecciones

2.7. Resúmenes de variables continuas

Como se ha dicho anteriormente, lo que sigue también aplica a variables discretas, especialmente lo referido a las medidas de resumen.

2.7.1. Tablas de frecuencias

Si intentáramos hacer una tabla de frecuencias de una variable continua, es muy posible que no se repitiera ningún dato, y tendríamos una tabla con todos los valores que se han producido y frecuencia 1. O en todo caso, algunos valores repetidos, según el número de observaciones y la precisión en la medición. Como esto no tiene sentido, en variables continuas (o discretas con muchos posibles valores) es agrupar los datos en k **intervalos** (clases). Hay varios criterios válidos para realizar esta división. Un criterio bastante aceptado es el siguiente:

- Si $n \leq 100$, $k \approx \sqrt{n}$
- Si $n > 100$, $k \approx 1 + \log_2 n$

Como la amplitud total de los datos (también llamado rango o recorrido) es $A = x_{max} - x_{min}$, es decir, la diferencia entre el máximo y el mínimo de los datos, entonces la amplitud de cada clase es $a_i = A/k$ (en el caso más habitual en el que todos los intervalos tienen la misma amplitud. A menudo la amplitud del intervalo se redondea para una mejor lectura e interpretación de la tabla.

Los intervalos se suelen tomar abiertos por la izquierda y cerrados por la derecha, y los límites se representan por $L_i, i = 0, \dots, k$, donde L_0 puede ser el mínimo (o el valor redondeado inmediatamente inferior según se haya decidido en la amplitud). L_k entonces será el máximo, o un valor superior según el redondeo indicado.

La marca de clase es el punto central del intervalo, es decir, la media aritmética de los extremos:

$$c_i = \frac{L_{i-1} + L_i}{2}$$

La frecuencia absoluta de cada clase i , n_i , es el número de observaciones cuyo valor numérico de la variable está dentro del intervalo. La frecuencia relativa, n_i/n , y las acumuladas se calcularían sumando las frecuencias de las clases inferiores. De nuevo resaltamos la importancia del concepto de frecuencia acumulada, como proporción de observaciones que toman valores menores o iguales que el límite superior del intervalo.

En la práctica, sería muy raro que tuviéramos que calcular la tabla de frecuencias “a mano”. El software estadístico se encargará de crear las clases para obtener la tabla de frecuencias, con algún método por defecto o indicando el número o amplitud de los intervalos. Lo que sí es importante es que el analista, a la vista del resumen (tabla o histograma) decida si cambia esta división por defecto por otra que cuente mejor la historia de los datos.



No obstante, sí es importante conocer el proceso de creación de la tabla, para entender mejor esa historia.

Tabla 2.5: Tabla de frecuencias por intervalos

| $\backslash (L_{i-1}, L_i]$ | $\backslash (n_i)$ | $\backslash (F_i)$ |
|-----------------------------|--------------------|--------------------|
| (6.35,6.4] | 1 | 0.001 |
| (6.4,6.45] | 0 | 0.001 |
| (6.45,6.5] | 3 | 0.003 |
| (6.5,6.55] | 54 | 0.050 |
| (6.55,6.6] | 184 | 0.207 |
| (6.6,6.65] | 404 | 0.552 |
| (6.65,6.7] | 369 | 0.867 |
| (6.7,6.75] | 129 | 0.977 |
| (6.75,6.8] | 21 | 0.995 |
| (6.8,6.85] | 6 | 1.000 |

La tabla ?? muestra una tabla de frecuencias de la variable **ph** del conjunto de datos de ejemplo de la producción de quesos. Vemos que es aproximadamente simétrico, donde los valores centrales son los más frecuentes, y a menudo que nos alejamos de estos valores centrales disminuye la frecuencia. Parece que puede haber un valor extremo por la izquierda. Un dato importante es que aproximadamente la mitad de las observaciones están por debajo de la clase más frecuente.



2.7.2. El histograma y el gráfico de densidad

La representación gráfica de la tabla de frecuencias por intervalos de una variable numérica es el **Histograma**. Este gráfico es uno de los más importantes en Estadística Descriptiva, y prácticamente lo primero que hay que hacer al analizar una variable numérica. De forma análoga a las variables cualitativas y discretas, en el eje Y se representan las clases. En este caso, al ser intervalos continuos, se representan en espacios entre ellos. En el eje Y se representan las frecuencias (absolutas o relativas) de cada clase. La geometría serán barras, en este caso **sin espacio entre ellas** para representar la continuidad. Si hay un intervalo sin barra, será porque no hay ninguna observación que tome valores dentro de ese intervalo ($n_i = 0$).

El histograma nos proporciona un resumen muy completo de la variable, buscaremos la siguiente interpretación:

- Valores mínimo y máximo (estarán dentro del primer y último intervalo respectivamente)
- Valores más frecuentes: estarán en los intervalos con las barras más altas
- Valores centrales: Intervalos entorno a los que se distribuyen las barras

- Valores poco frecuentes: estarán en los intervalos con las barras más bajas
- Valores extremos (alejados del resto): barras muy bajas en los extremos
- Asimetría: Los datos serán simétricos si los valores a ambos lados de los valores centrales se distribuyen de forma parecida.
- Forma: Identificaremos si tiene forma de campana (normal, gaussiana), exponencial, uniforme, etc.

La figura ?? representa la tabla de frecuencias ???. Vemos más claramente la forma aproximadamente simétrica del histograma, el valor extremo a la izquierda (aunque no se aprecia la barra). El intervalo más frecuente parece repartir el resto a ambos lados de forma homogénea, disminuyendo la frecuencia a medida que nos alejamos de estos valores centrales. En resumen, la típica forma de campana de Gauss.

La figura ?? representa el histograma de la variable `bacteri`. Es una variable discreta pero con muchos valores distintos, por lo que es mejor la representación del histograma que la del gráfico de barras. Vemos una distribución típicamente exponencial, con valores bajos muy frecuentes y altos muy poco frecuentes, altamente asimétrica.

La figura ?? representa el histograma de la variable `sal`. Muestra una distribución aproximadamente uniforme hasta un valor 1, y después también pero con menos frecuencia, con algunos valores más allá de 1.2. Esto puede estar indicando una mezcla de poblaciones (por ejemplo que los distintos tipos de quesos tengan una receta distinta).



Una representación alternativa al histograma es la línea de densidad, que sustituye las barras por una línea continua, generalmente suavizada, que nos da una idea de la forma de la distribución de forma más esquemática. Esta línea de densidad se puede superponer al histograma, o sustituirlo rellenando el área que queda por debajo de la curva.

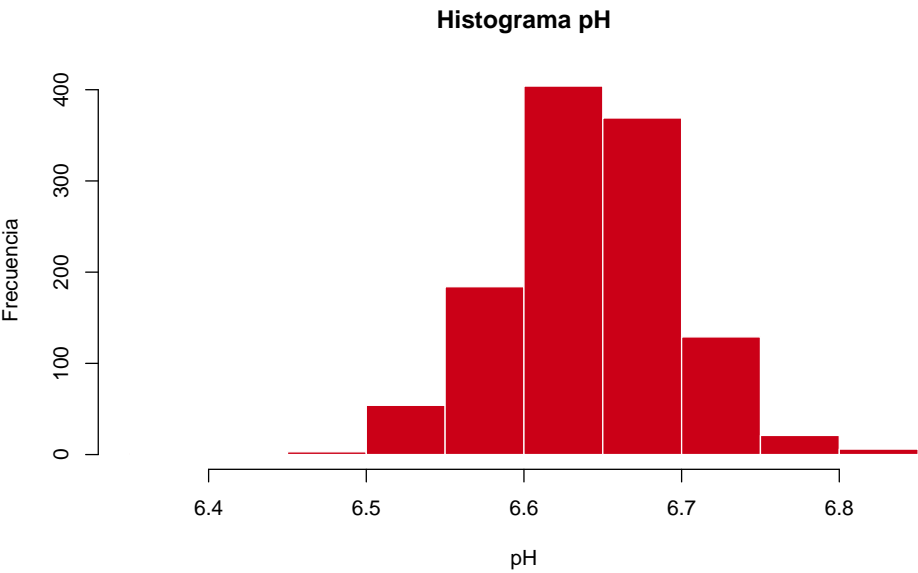


Figura 2.8: Histograma de la variable ph

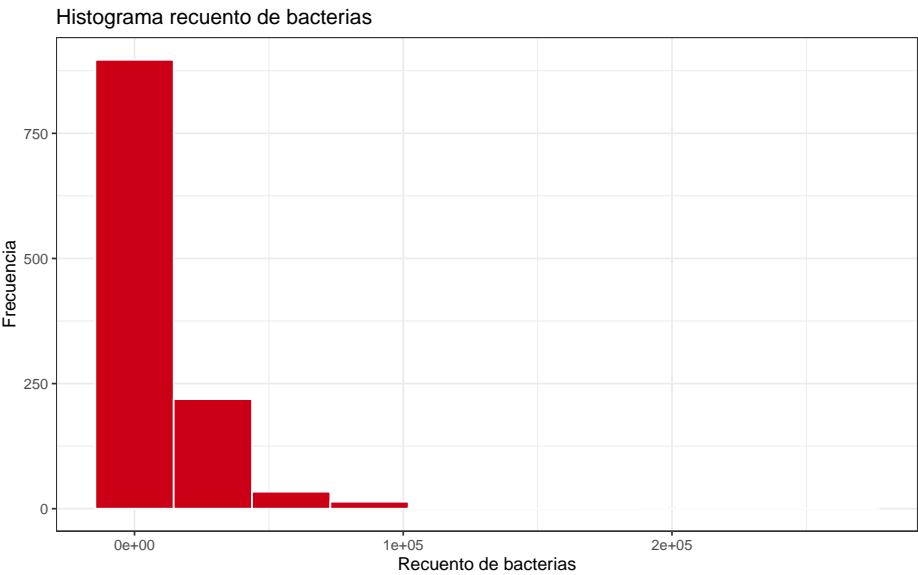


Figura 2.9: Histograma de la variable bacteri

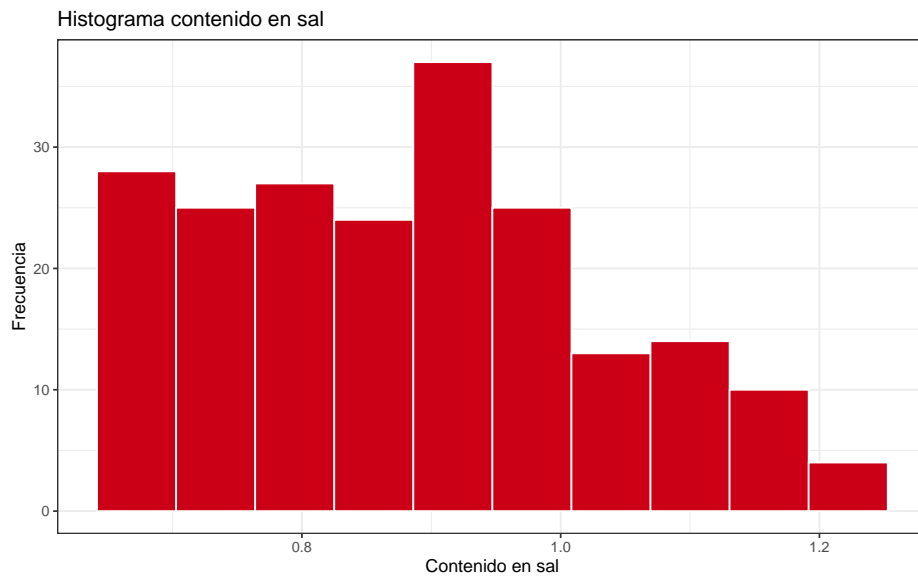


Figura 2.10: Histograma de la variable sal

El gráfico de arriba en la figura ?? muestra el histograma de la figura ?? con la línea de densidad superpuesta. Aunque decíamos que era bastante uniforme, la línea suavizada parece que sugiere dos “picos” en la parte izquierda. Si no tenemos más información, tendríamos que plantearnos que puede haber mezclados grupos que son diferentes en cuanto al comportamiento de la variable a medir. El gráfico de abajo en la figura ?? muestra un gráfico de densidades distinguiendo entre los tipos de queso (hemos “mapeado” en nuestro lienzo, el color a la variable tipo). Vemos claramente cómo el tipo de queso C tiene un nivel de sal más alto que los otros dos, que sí parecen tener una distribución más similar. Con esta separación, la variable es más simétrica y podríamos aproximar a alguna distribución de probabilidad como veremos en el capítulo ??.



2.7.3. Medidas de tendencia central

La tabla de frecuencias y el histograma es un buen resumen de una variable. Pero podemos resumir o describir la variable con una serie de medidas características que resumen algún aspecto en concreto con un solo número. El primer grupo de medidas que podemos calcular son las medidas de centralización. Es decir, los valores centrales entorno a los que varían los datos.

Ya conocemos la **moda**, que es el valor más frecuente en variables cualitativas

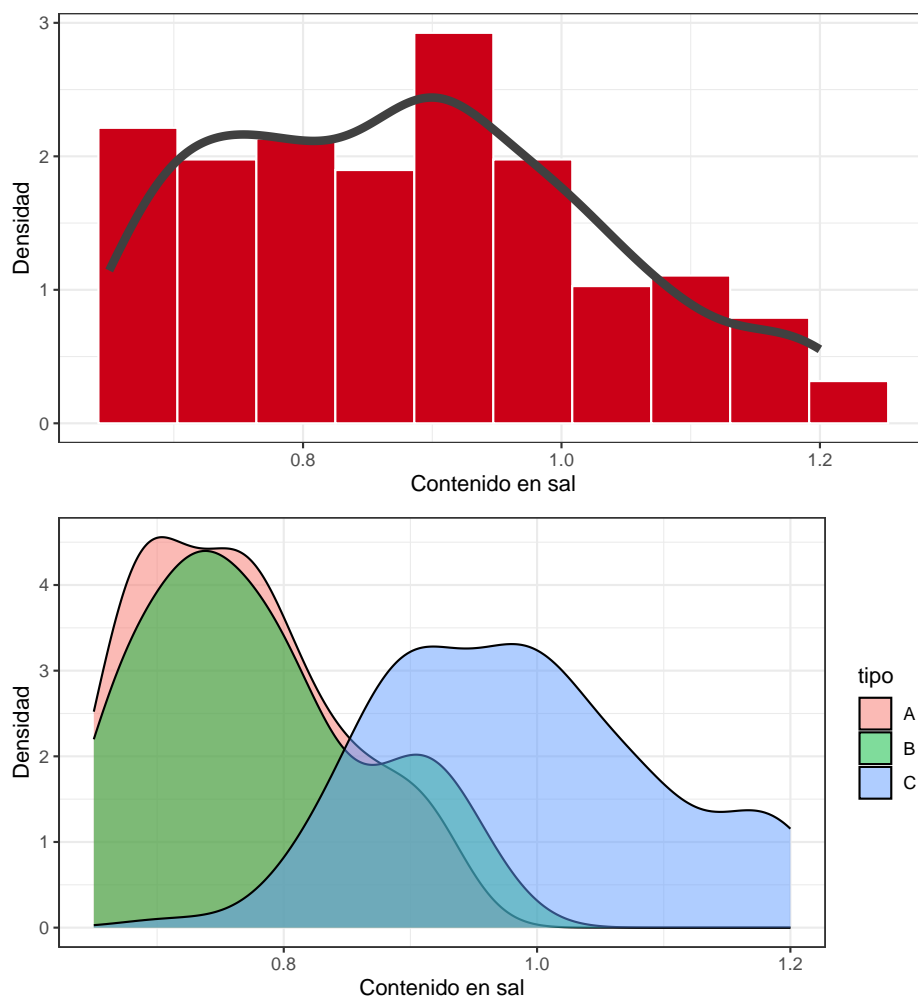


Figura 2.11: Ejemplo de gráficos de densidad

y en variables numéricas discretas. En variables numéricas continuas, hablaremos de **intervalo modal**, que será el intervalo con la frecuencia más alta. Si tuviéramos que elegir un solo número, podríamos elegir la marca de clase como valor representativo, o utilizar la siguiente fórmula, con la que se obtiene un valor más próximo al intervalo adyacente más frecuente (?):

$$Mo = L_i \frac{(n_i - n_{i-1}) + (n_i - n_{i+1})}{(n_i - n_{i-1})} \cdot a_i$$

Es importante resaltar que la moda en variables continuas va a ser diferente segúnelijamos los intervalos en los que dividimos el rango.

La **media aritmética** es sin duda la medida de centralización más conocida y más importante. Cuando disponemos de todos los valores $x_i, i = 1, \dots, n$ de la variable, la calculamos con la siguiente fórmula:

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

La media es el centro de gravedad de los datos, el valor promedio. Está muy influenciada por observaciones extremas, por lo que es poco representativa en distribuciones asimétricas o donde hay mucha dispersión. Mantiene la linealidad, es decir: si X e Y son dos variables con medias \bar{x} e \bar{y} respectivamente:

$$Y = a + bX \implies \bar{y} = a + b\bar{x}$$

Un concepto clave de la media es que los valores de la variable **varían** entorno a ella, por arriba y por abajo, y las diferencias con la media se compensan, de forma que:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Si en vez de todos los datos tenemos una tabla de frecuencias, para variables discretas la calcularemos como:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i = \sum_{i=1}^k f_i x_i,$$

Siendo $x_i, i = 1, \dots, k$ los posibles valores que toma la variable. Para datos agrupados:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i c_i = \sum_{i=1}^k f_i c_i,$$

Siendo $c_i, i = 1, \dots, k$ las marcas de clase. Nótese que, al utilizar tablas de frecuencias en vez de todos los valores, se pierde precisión en el cálculo. A veces esta fórmula la utilizamos también para calcular la **media ponderada** cuando tenemos unos pesos para cada valor. Por ejemplo, en encuestas donde cada entrevistado representa a un número determinado de individuos en estudio.

Otra variante es la **media recortada** o media robusta. Para su cálculo, se eliminan un porcentaje de observaciones (por ejemplo, el 5%) a ambos extremos, quedando así menos “expuesta” a observaciones extremas, y ganando en representatividad.

La media tiene muy buenas propiedades matemáticas y representa bien los datos en variables simétricas poco dispersas. No obstante hay que tener cuidado al interpretarla porque puede ser un valor sin sentido práctico. En un caso extremo, imaginemos una variable que toma valores -1 y 1 con la misma frecuencia. La media será 0, un valor que no puede ni siquiera tomar la variable.

Aunque en estadística será raro que nos las encontremos, existen otras dos medias que en ciertas aplicaciones de la ingeniería y las ciencias o de la economía son muy útiles:

Media Geométrica

$$m_g = (\prod_{i=1}^n x_i)^{\frac{1}{n}}$$

Media armónica

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



La **mediana** es otra medida de posición central que indica el valor que divide los datos en dos mitades: los que son menores que la mediana y los que son mayores que la mediana. La principal ventaja es que está muy poco influenciada por valores extremos. Para calcularla, se ordenan todos los datos $x_i, i = 1, \dots, n$ de menor a mayor. Si el número de datos es impar, el que ocupa la posición central, $[n/2] + 1$ es la mediana. Si el número de datos es par, la mediana es la media aritmética de los dos valores centrales, $n/2$ y $n/2 + 1$.

Si tenemos una tabla de frecuencias de una variable discreta, la mediana es el primer valor x_i que cumpla $N_i \geq n/2$ o, equivalentemente, $F_i \geq 0,5$. Si lo que tenemos es una tabla de frecuencias por intervalos de una variable continua, entonces podemos tomar como intervalo mediano el primero para el cual $F_i \geq 0,5$ y usar la marca de clase de ese intervalo c_i como valor representativo. También se puede utilizar la siguiente fórmula, aunque al igual que con la moda, el valor va a depender de la manera en que hayamos construido los intervalos:

$$Me = L_i + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i.$$

Supongamos un proceso en el que se produce una merma. Se extrae una muestra de 50 observaciones y se obtienen las mediciones de la merma de la tabla ??, en el orden en el que aparecen (por filas). La tabla ?? muestra los datos ordenados por filas, resaltando los dos valores centrales (el número 24 y el número 25 de orden). Entonces la mediana es:

$$Me = \frac{4,979 + 5,015}{2} = 4,997.$$

La figura ?? muestra gráficamente cómo la mediana divide los datos en dos mitades.

La principal ventaja de la mediana es que no está afectada por valores extremos, y siempre es más representativa de los datos. En variables simétricas, coincide con la media y también con la moda. El inconveniente es que tiene muy malas propiedades matemáticas y es más difícil de tratar en inferencia.

La media de esta variable sería 4.968. Cuando la media y la mediana están próximas, como es este caso, indica **simetría** en los datos.



| V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 5.377 | 6.007 | 4.822 | 6.014 | 3.892 | 5.379 | 4.347 | 4.599 | 4.104 | 4.979 |
| 6.075 | 4.115 | 5.432 | 4.140 | 5.067 | 4.962 | 5.429 | 5.172 | 4.709 | 5.393 |
| 4.654 | 4.408 | 5.634 | 4.844 | 5.015 | 4.259 | 4.437 | 4.118 | 4.469 | 4.329 |
| 5.377 | 4.679 | 5.716 | 4.688 | 5.114 | 5.132 | 5.215 | 4.258 | 5.090 | 6.031 |
| 5.363 | 4.756 | 4.758 | 5.923 | 5.258 | 4.443 | 4.845 | 5.046 | 5.322 | 5.187 |

| V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 3.892 | 4.104 | 4.115 | 4.118 | 4.140 | 4.258 | 4.259 | 4.329 | 4.347 | 4.408 |
| 4.437 | 4.443 | 4.469 | 4.599 | 4.654 | 4.679 | 4.688 | 4.709 | 4.756 | 4.758 |
| 4.822 | 4.844 | 4.845 | 4.962 | 4.979 | 5.015 | 5.046 | 5.067 | 5.090 | 5.114 |
| 5.132 | 5.172 | 5.187 | 5.215 | 5.258 | 5.322 | 5.363 | 5.377 | 5.377 | 5.379 |
| 5.393 | 5.429 | 5.432 | 5.634 | 5.716 | 5.923 | 6.007 | 6.014 | 6.031 | 6.075 |

2.7.4. Medidas de posición

La mediana es una medida de posición, que se encuentra en el 50 % de los datos, y también se llama **cuantil** 0,5. Otras dos medidas básicas de posición son el

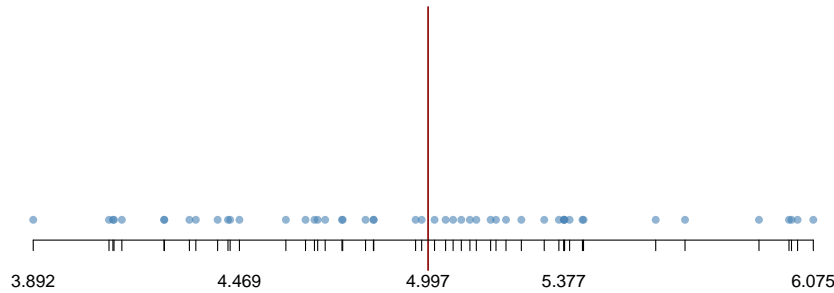


Figura 2.12: Significado gráfico de la mediana y los cuantiles

máximo, x_{max} , y el **mínimo**, x_{min} , que nos informan de los extremos de los datos. Los **cuantiles** son los cuantiles que dejan a la izquierda de su posición el 25 % (primer cuartil, Q_1) y el 75 % (tercer cuartil, Q_3) de los datos respectivamente. La mediana es también el segundo cuartil, Q_2 . Se pueden definir análogamente terciles y deciles, así como percentiles P_p , que es el valor que deja por debajo de sí mismo el p % de los datos, $0 < p < 100$. Los cuantiles, q_p , son equivalentes a los percentiles, expresados en tanto por uno, $q_p, 0 < p < 1$.

Las medidas de posición se representan con el gráfico de cajas y bigotes, pero antes de definirlo necesitamos conocer las medidas de dispersión.

2.7.5. Medidas de dispersión

Las medidas de centralización y posición no son suficientes para resumir una variable, se debe acompañar de medidas de dispersión. Vamos a ver a continuación las más importantes. Las medidas de dispersión nos dan una idea de cómo es la **variación** de los datos alrededor de los valores centrales. Pensemos que una misma medida central, como por ejemplo la media, puede provenir de datos muy próximos a ella, o muy lejanos.

La medida más básica de dispersión que podemos calcular es el **rango**, también conocido como el recorrido. Se define como la diferencia entre el máximo y el mínimo:

$$R = \max_i x_i - \min_i x_i$$

La figura ?? muestra los valores de la merma estudiada en el ejemplo anterior. El rango sería:



$$R = 6,075 - 3,892 = 2,183$$

El rango es una medida muy pobre porque solo utilizamos dos valores de todos

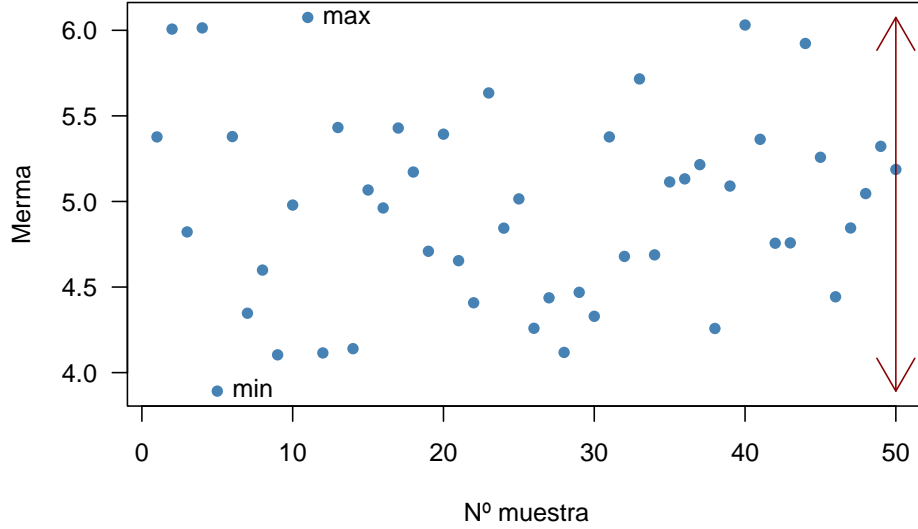


Figura 2.13: Representación del rango

los del conjunto de datos. Sería mejor una medida que mida la dispersión con todos los datos. Una posibilidad sería hacer un promedio de las diferencias con algún valor central, por ejemplo, la media. Pero ya vimos que ese promedio es igual a cero, por lo que tendríamos que usar otra medida. Por ejemplo, las medias de esas desviaciones en valor absoluto. A esta medida la llamaremos **desviación media absoluta**, DMA o MAD por sus siglas en inglés:

$$DMA = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|.$$

La **desviación absoluta mediana** es la mediana de las desviaciones con la mediana:

$$DAM = Me|x_i - Me_x|, i = 1, \dots, n.$$

Estas dos medidas son bastante robustas, pero hacen uso de la función valor absoluto, que no tiene buenas propiedades matemáticas.

La medida de la variabilidad más importante es la **Varianza**, que es el promedio de las desviaciones al cuadrado con respecto a la media. Si tuviéramos una población con N valores X_i y media conocida μ :

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

Pero en general, trabajaremos con muestras. Incluso cuando tenemos todo el universo en estudio, podemos considerar que es una muestra de los distintos estados en los que se puede encontrar. Entonces, la varianza muestral, también conocida como cuasivarianza, en muestras de tamaño n se define como:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

La varianza se expresa en las unidades de la variable al cuadrado. Una medida de la variabilidad en las mismas unidades que la variable (y que la media) es la **desviación típica**, que no es más que la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}; \quad s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

La figura ?? esquematiza las desviaciones a la media (línea roja horizontal) del ejemplo de la merma. La suma de todas las desviaciones es cero, ya que se compensan las positivas con las negativas. La desviación media absoluta es 0.464. La desviación absoluta mediana es 0.381. La varianza y la desviación típicas muestrales son:



$$s^2 = 0,321, \quad s = 0,566.$$

Nótese la relación entre la varianza y la cuasivarianza:

$$s_{poblacional}^2 = s_{muestral}^2 \cdot \frac{n-1}{n}.$$

Si desarrollamos la fórmula de la varianza, llegamos al siguiente cálculo abreviado:

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{N} = \frac{1}{N} \sum_{i=1}^n X_i^2 - \mu^2,$$

que en el caso de la cuasivarianza quedaría:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \right] = \frac{\sum_{i=1}^n x_i^2}{n-1} - \frac{n \cdot \bar{x}^2}{n-1},$$

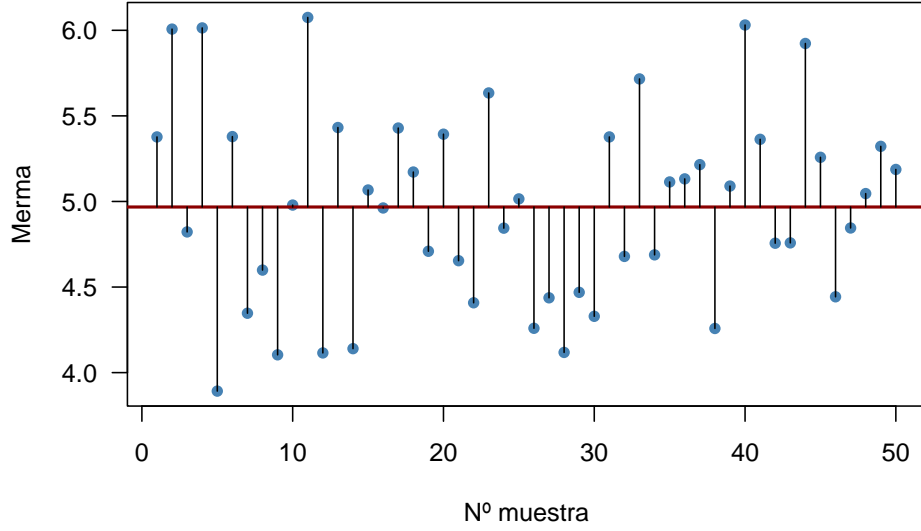


Figura 2.14: Desviaciones de la media

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right).$$

Para variables discretas con datos agrupados y frecuencias absolutas, podríamos calcularlo así:

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n n_i x_i^2 - n\bar{x}^2 \right) = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n n_i x_i^2 - \bar{x}^2 \right),$$

o con frecuencias relativas:

$$s^2 = \frac{n}{n-1} \left(\sum_{i=1}^n f_i x_i^2 - \bar{x}^2 \right).$$

Análogamente, en variables continuas agrupadas en intervalos, sustituiríamos x_i por las marcas de clase c_i . Recordemos que de esta forma perderemos precisión.

Lo normal es que en nuestro análisis de datos trabajemos con datos rectangulares como se describía en el apartado ??, y usemos software que haga los cálculos. No obstante, en ocasiones tenemos los datos solo en forma de tablas de frecuencias, por ejemplo de publicaciones oficiales o en revistas científicas, y podemos analizarlos aunque se pierda algo de precisión. También aparecen este tipo de datos cuando provienen de encuestas en las que los encuestados responden directamente un intervalo (por ejemplo de ingresos, gastos, cantidades consumidas, etc.)



En todo caso, y aunque no tengamos que aplicar nunca la fórmula “a mano” es muy importante saber cómo se hace el cálculo para entender los métodos en los que después se utilizan estos estadísticos.

Al igual que la media, la varianza y la desviación típica son muy sensibles a datos extremos. Una propiedad importante de la desviación típica y la media es que entre la media y 2 desviaciones típicas $(\bar{x} - 2s; \bar{x} + 2s)$, tenemos al menos el 75 % de los datos.

La varianza no mantiene la transformación lineal. Si $Y = a + bX$ y la varianza de X es s_X , entonces la varianza de Y es: $b^2 s_X^2$

De las propiedades de la media y la varianza se deduce también que si:

$$Z = \frac{X - \bar{x}}{s},$$

entonces $\bar{z} = 0$; $s^2 = 1$. Esta transformación es lo que llamamos “tipificar” (o escalar) la variable. Es útil para comparar distintas escalas, y mantiene la estructura correlación al aplicarlo a más de una variable. Veremos también su utilidad en variables aleatorias en el capítulo ??.

La desviación típica está en las unidades de la variable, y podemos relativizarla a la propia variable. Pero si queremos comparar la variabilidad de dos variables, necesitamos una medida comparable. El **coeficiente de variación** es una medida adimensional que nos sirve para este cometido:

$$CV = \frac{s}{|\bar{x}|}$$

Es también útil para comparar la misma variable en grupos distintos, cuando la media no es igual en todos ellos. Por otra parte, al comparar dos procesos (o tratamientos) en los que se consigue un cambio en la media, el objetivo de mantener la variabilidad se suele fijar en términos del coeficiente de variación, ya que al cambiar la tendencia central del proceso también puede cambiar la varianza.

La última medida de variabilidad que veremos es el **rango intercuartílico**. Se define como la diferencia entre el primer y tercer cuartil, y representa el rango del 50 % de los datos centrales (alrededor de la mediana):

$$IQR = Q_3 - Q_1$$

2.7.6. El gráfico de cajas y bigotes

Ahora que hemos visto las medidas de dispersión, podemos definir otro gráfico muy esclarecedor para variables numéricas (tanto discretas como continuas). Se trata del **gráfico de cajas y bigotes**, aunque abreviaremos en general como gráfico de cajas. Este gráfico representa, generalmente en el eje vertical, los siguientes estadísticos:

- El mínimo (extremo bigote inferior)
- El primer cuartil (borde inferior de la caja)
- La mediana (línea cruzando caja)
- El tercer cuartil (borde superior de la caja)
- El máximo (extremo bigote superior)
- Si existen candidatos a ser considerados atípicos (*outliers*):
 - El bigote llega hasta el último valor en “barreras”
 - Los valores fuera de las barreras se representan mediante puntos

El gráfico de cajas por sí solo puede estar ocultando información, por lo que se pueden utilizar variantes como los gráficos de violín, o complementar con un gráfico de densidad.

Las barreras se calculan como una distancia de 1.5 veces el rango intercuartílico desde el primer y tercer cuartil, es decir:

$$Q_1 - 1,5 \cdot IQR; Q_3 + 1,5 \cdot IQR.$$

La figura ?? muestra el gráfico de cajas del ejemplo de la merma añadiendo un punto más extremo que los que teníamos. El rango intercuartílico para estos datos es:

$$IQR = Q_3 - Q_1 = 0,811$$

Con él se calculan las barreras, aunque no se representarán en el gráfico de cajas. Como todos los puntos por abajo están dentro de estas barreras, el bigote inferior llega hasta el mínimo. Sin embargo por arriba, como hay un valor más allá de la barrera, el bigote llega hasta el último valor dentro de la barrera, y el valor que hay fuera se representa con un punto.



Gráfico de caja para datos de merma

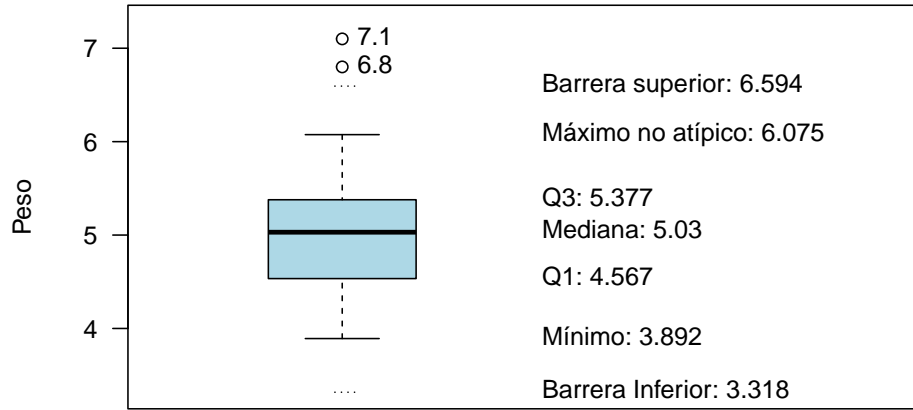


Figura 2.15: Explicación de los estadísticos representados en el gráfico de cajas

2.7.7. Medidas de forma

Aunque el histograma nos da una idea de la forma de la distribución de los datos, podemos cuantificar esta forma principalmente con dos medidas. El **coeficiente de asimetría**, γ_1 nos indicará en qué medida los datos son simétricos. Esto sucede cuando la mediana es igual a la media. Si no son simétricos, el coeficiente nos indicará el tipo de asimetría.

$$\gamma_1 = \frac{m_3}{s^3},$$

donde $m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$

La interpretación es la siguiente:

- $\gamma_1 = 0 \Rightarrow$ Simétrica.
- $\gamma_1 < 0 \Rightarrow$ Asimétrica negativa (valores bajos menos frecuentes que valores altos).
- $\gamma_1 > 0 \Rightarrow$ Asimétrica positiva (valores bajos más frecuentes que valores altos).

La figura ?? muestra histogramas de variables con distinta asimetría, y la tabla ?? los valores de sus coeficientes de asimetría.

| Tipo | γ_1 |
|---------------------|------------|
| Asimetrica_negativa | -1,917 |
| Asimetrica_positiva | 1,943 |
| Simetrica | 0,089 |

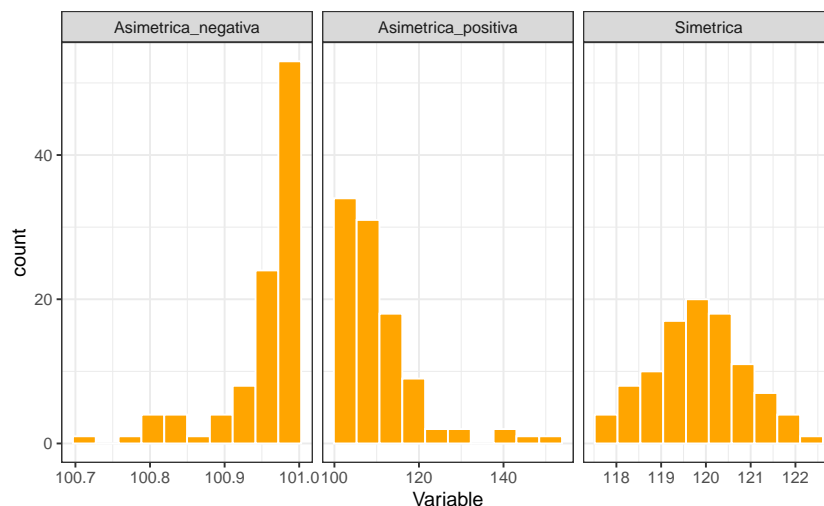


Figura 2.16: Histogramas de variables con distintos tipos de asimetría

El **coeficiente de apuntamiento o curtosis** nos indica cómo de “apuntados” son los datos (muy concentrados alrededor de la media):

$$\gamma_2 = \frac{m_4}{s^4} - 3,$$

donde $m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4$

La interpretación es la siguiente:

- $\gamma_2 = 0 \Rightarrow$ Igual de apuntada que la distribución normal.
- $\gamma_2 < 0 \Rightarrow$ Más aplastada que la normal.
- $\gamma_2 > 0 \Rightarrow$ Más apuntada que la normal.

La figura ?? muestra histogramas de variables con distinta curtosis, y la tabla ?? sus valores del coeficiente de curtosis.

| Tipo | γ_2 |
|--------------|------------|
| Leptocúrtica | 5,587 |
| Normal | 0,033 |
| Platicúrtica | -1,167 |

2.7.8. Resumen numérico

Las medidas y gráficos vistos hasta el momento, no se suelen (ni deben) utilizar de forma aislada. En el análisis exploratorio de datos, se suelen obtener todos los resúmenes numéricos de la variable de interés, posiblemente por grupos si

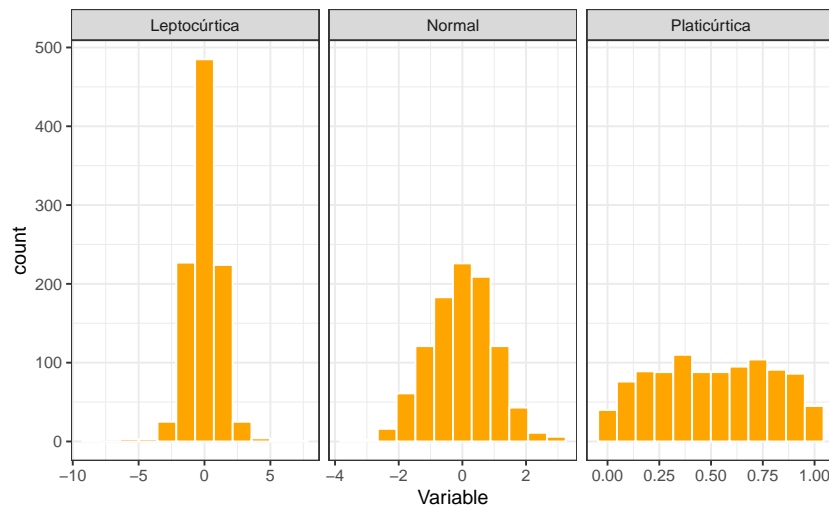


Figura 2.17: Histogramas de variables con distintos tipos de curtosis

hay alguna variable categórica que pueda presentar distintas distribuciones de la variables numéricas según la categoría. Este resumen y los gráficos adecuados (cajas, histogramas, densidades, barras) describen la variable por completo. Este análisis se puede hacer de más de una variable para estudiarlas conjuntamente.



La tabla ?? muestra un resumen de las variables **est** y **ph** del conjunto de datos de laboratorio de la fábrica de quesos.

| | ph | sal |
|-------------|--------|--------|
| Mean | 6,648 | 0,883 |
| Std.Dev | 0,055 | 0,145 |
| Min | 6,360 | 0,650 |
| Q1 | 6,610 | 0,760 |
| Median | 6,650 | 0,880 |
| Q3 | 6,680 | 0,990 |
| Max | 6,840 | 1,200 |
| MAD | 0,044 | 0,163 |
| IQR | 0,070 | 0,230 |
| CV | 0,008 | 0,165 |
| Skewness | -0,012 | 0,338 |
| SE.Skewness | 0,071 | 0,169 |
| Kurtosis | 0,846 | -0,723 |
| N.Valid | 1,171 | 207 |
| Pct.Valid | 100,0 | 17,7 |

| | Tipo A | Tipo B | Tipo C |
|-------------|--------|--------|--------|
| Mean | 0,764 | 0,778 | 0,987 |
| Std.Dev | 0,077 | 0,088 | 0,109 |
| Min | 0,650 | 0,650 | 0,710 |
| Q1 | 0,700 | 0,720 | 0,900 |
| Median | 0,760 | 0,760 | 0,990 |
| Q3 | 0,810 | 0,830 | 1,070 |
| Max | 0,930 | 0,970 | 1,200 |
| MAD | 0,089 | 0,089 | 0,133 |
| IQR | 0,110 | 0,110 | 0,170 |
| CV | 0,101 | 0,113 | 0,110 |
| Skewness | 0,475 | 0,473 | 0,258 |
| SE.Skewness | 0,333 | 0,340 | 0,234 |
| Kurtosis | -0,822 | -0,867 | -0,588 |
| N.Valid | 51 | 49 | 107 |
| Pct.Valid | 29,1 | 33,1 | 12,6 |

2.7.9. Gráficos dependientes del tiempo

Otra visualización básica para una variable numérica es la visualización secuencial de las observaciones, bien a través de puntos o a través de líneas. En el eje X se representa el orden de las observaciones, y en el eje Y la escala de la variable. El orden de las observaciones nos pueden indicar cuándo se ha producido un cambio u otros patrones. La figura ?? representa un gráfico secuencial de puntos que podría estar indicando un patrón durante los días de la semana. Cuando las observaciones tienen una marca de tiempo (fecha o fecha y hora) entonces estamos ante una serie temporal, como la de la figura ??.

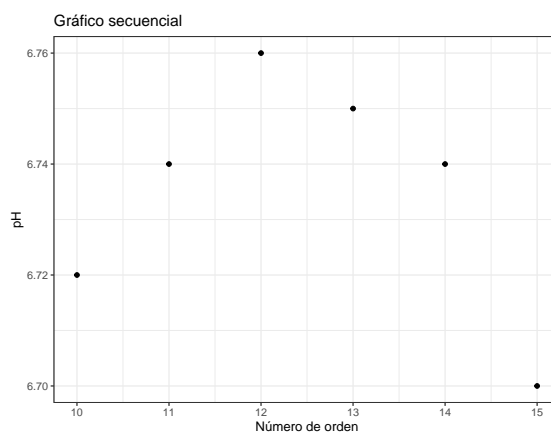


Figura 2.18: Gráfico de puntos secuencial

#> `summarise()` has grouped output by 'tipo'. You can
 #> override using the `.groups` argument.

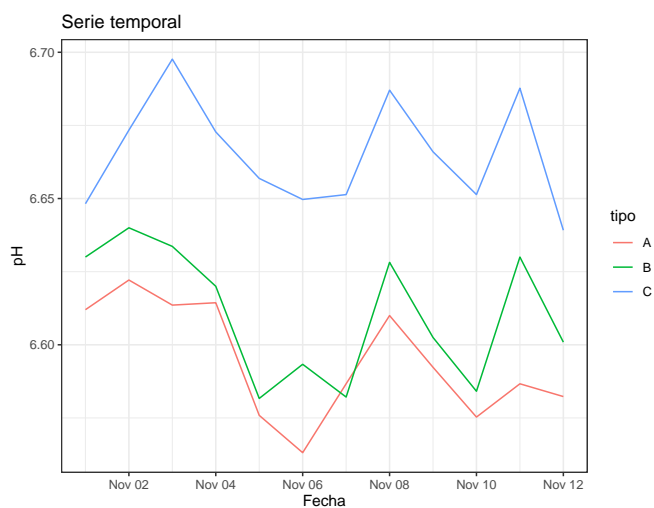


Figura 2.19: Gráfico de serie temporal

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE, comment = "")
library(tidyverse)
#> -- Attaching packages ----- tidyverse 1.3.2 --
#> v ggplot2 3.4.1      v purrr 1.0.1
#> v tibble 3.1.8       v dplyr 1.1.0
#> v tidyr 1.3.0        v stringr 1.5.0
#> v readr 2.1.3       v forcats 0.5.2
#> -- Conflicts ----- tidyverse_conflicts() --
#> x purrr::compose() masks flextable::compose()
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag() masks stats::lag()
library(knitr)
library(flextable)
library(gt)
library(gridExtra)
#>
#> Attaching package: 'gridExtra'
#>
#> The following object is masked from 'package:dplyr':
#>
#> combine
library(equationmatic)
```


Capítulo 3

Análisis exploratorio bivariante

3.1. Datos bivariantes y multivariantes

El análisis univariante es muy útil para describir una única característica de la población en estudio. Pero rara vez estudiamos una característica aislada, y lo habitual es tener conjuntos de datos con varias variables (cuantitativas y cualitativas) que podemos estudiar por separado (análisis univariante) o conjuntamente (análisis multivariante).

El caso especial del **análisis bivariante** es cuando estudiamos dos características a la vez: X, Y . Nos interesa la **relación** entre ellas, para lo que realizaremos resúmenes numéricos y gráficos. Los datos bivariantes se encontrarán como pares de valores (x_i, y_i) para cada observación $i = 1, \dots, n$.

Cuando estudiamos más de dos variables, tenemos datos multivariantes. En este caso, estudiamos las relaciones “dos a dos” entre las variables (como en el caso bivariante) y la estructura conjunta. Hay algunas técnicas multivariantes específicas para este último caso. En este capítulo nos vamos a centrar solo en el primer caso.

3.2. Frecuencias conjuntas, marginales y condicionadas

El primer resumen que podemos hacer de datos bivariante es la tabla de frecuencias conjunta. Igual que en el caso univariante n es número total de datos, es decir, la frecuencia total. La frecuencia absoluta conjunta, n_{ij} , es el número de observaciones en la clase i de X y en la clase j de Y . La frecuencia relativa

Tabla 3.1: Tabla de contingencia (frecuencias absolutas) de los analistas y el tipo de queso.

| | A | B | C |
|-------------|----|----|-----|
| analista_10 | 52 | 47 | 219 |
| analista_13 | 42 | 36 | 198 |
| analista_6 | 44 | 32 | 235 |
| analista_9 | 37 | 33 | 196 |

Tabla 3.2: Tabla de frecuencias relativas de los analistas y el tipo de queso.

| | A | B | C |
|-------------|------|------|------|
| analista_10 | 0.04 | 0.04 | 0.19 |
| analista_13 | 0.04 | 0.03 | 0.17 |
| analista_6 | 0.04 | 0.03 | 0.20 |
| analista_9 | 0.03 | 0.03 | 0.17 |

conjunta es $f_{ij} = \frac{n_{ij}}{n}$.

3.2.1. Distribución conjunta de frecuencias

Las frecuencias conjuntas se representan en una tabla de doble entrada, con los valores de una variable en filas y los de la otra en columnas. En el interior, se ponen las frecuencias conjuntas (absolutas, relativas o ambas). Si las dos variables son cualitativas, la tabla se denomina **Tabla de contingencia**.



La Tabla ?? muestra la tabla de contingencia de los analistas y el tipo de queso en el ejemplo del análisis de la producción de quesos. Asignamos la variable X al Analista (en filas) y la variable Y al Tipo (en columnas). La Tabla ?? muestra la tabla de frecuencias relativas de los mismos datos. El número total de datos es $n = 1171$.

En el caso de variables continuas, debemos tener los datos agrupados en intervalos (clases).



La tabla ?? contiene las frecuencias absolutas conjuntas de las variables:

- $X = \text{ph}$ (filas);
- $Y = \text{est}$ (columnas)