

Estadística Aplicada a las Ciencias y la Ingeniería

Emilio L. Cano

2021-10-01

Índice general

| | |
|---|-----------|
| Bienvenida | 5 |
| Estándares y software | 5 |
| Estructura del libro | 6 |
| Sobre el autor | 8 |
| Agradecimientos | 8 |
| | |
| I Estadística descriptiva | 9 |
| | |
| 1. Introducción | 11 |
| 1.1. Estadística y análisis de datos | 11 |
| 1.2. Los datos y su organización | 15 |
| 1.3. La Estadística y el método científico | 18 |
| 1.4. Estadística, Calidad y Sostenibilidad | 19 |
| 1.5. Objetivos de Desarrollo Sostenible (ODS) | 24 |
| | |
| 2. Análisis exploratorio univariante | 29 |
| | |
| 3. Análisis exploratorio bivalente | 31 |
| | |
| II Probabilidad | 33 |
| | |
| 4. Introducción a la Probabilidad | 35 |
| | |
| 5. Variable aleatoria univariante | 37 |
| | |
| 6. Variable aleatoria bivalente | 39 |
| | |
| 7. Modelos de distribución de probabilidad | 41 |
| | |
| III Inferencia estadística | 43 |
| | |
| 8. Muestreo y estimación | 45 |

| | |
|---|---------------|
| 9. Comparación de grupos | 47 |
| 10. Modelos de regresión | 49 |
| 11. Diseño de experimentos | 51 |
| IV Control estadístico de la calidad | 53 |
| 12. Introducción | 55 |
| 13. Control Estadístico de Procesos | 57 |
| 14. Inspección por muestreo | 59 |
| A. Símbolos, abreviaturas y acrónimos | 61 |
| A.1. Acrónimos | 61 |
| A.2. Letras griegas | 61 |
| A.3. Símbolos | 62 |
| B. Tablas estadísticas | 63 |
| B.1. Distribución normal | 63 |
| B.2. Resumen modelos de distribución de probabilidad | 65 |
| C. Repaso | 67 |
| C.1. Logaritmos y exponenciales | 67 |
| C.2. Combinatoria | 67 |
| D. Ampliación | 71 |
| D.1. Función característica | 71 |
| D.2. Cambio de variable | 71 |
| D.3. Variables aleatorias unidimensionales mixtas | 71 |
| D.4. Variables aleatorias bidimensionales mixtas | 71 |
| D.5. Algunos modelos de distribución continuos más | 71 |
| D.6. Modelos de distribución de probabilidad multivariantes | 75 |
| D.7. Modelos de distribución de probabilidad relacionadas con la normal | 75 |
| D.8. Simulación de variables aleatorias | 75 |
| E. Demostraciones | 77 |
| E.1. Variable aleatoria discreta | 77 |
| F. Créditos | 79 |

Bienvenida

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 (López Cano, 2018, 2019). 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** (R Core Team, 2021), y el software de álgebra computacional **Máxima**². Respecto al software de hoja de cálculo, las fórmulas utilizadas se han probado en el software libre **LibreOffice**³, en **Hojas de Cálculo de Goo-**

¹A la espera de encontrar editorial.

²<http://maxima.sourceforge.net/es/>

³<https://es.libreoffice.org>

gle⁴ 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 Cano et al. (2015).

Estructura del libro

Este libro se ha elaborado utilizando el lenguaje *Markdown* con el propio software **R** y el paquete **bookdown** (Xie, 2021). 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 `#>{r}` `sessionInfo()`

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.

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>

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 bivariante. 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

Actualmente soy Profesor Ayudante 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 F. 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* y al equipo de RStudio.

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.

Ocaña-Riola (2017)

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

Cano et al. (2012)

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 (Tukey et al., 1977), 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 1.1 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:

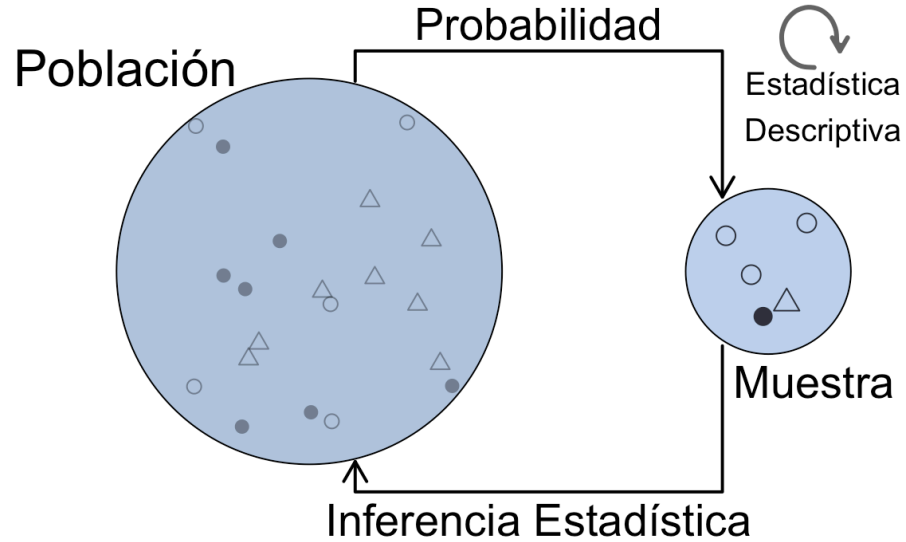


Figura 1.1: La esencia de los métodos estadísticos

$$Y = f(X)$$

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 (Cleveland, 2001)², aunque lo anticipó Tukey 40 años antes en “The Future of Data Analysis” (Tukey, 1962). 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 1.2 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.

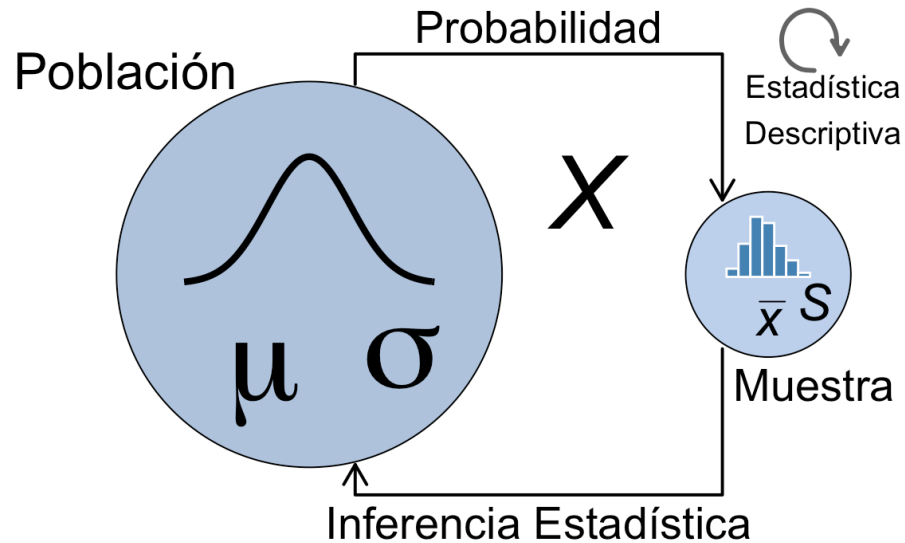


Figura 1.2: La esencia de los métodos estadísticos

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 1.1 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 .red[características] inherentes de un objeto cumple con los .red[requisitos]

ISO 9001:2015 3.6.2

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

En la figura 1.3 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 (Taguchi et al., 2007). La figura 1.4 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

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

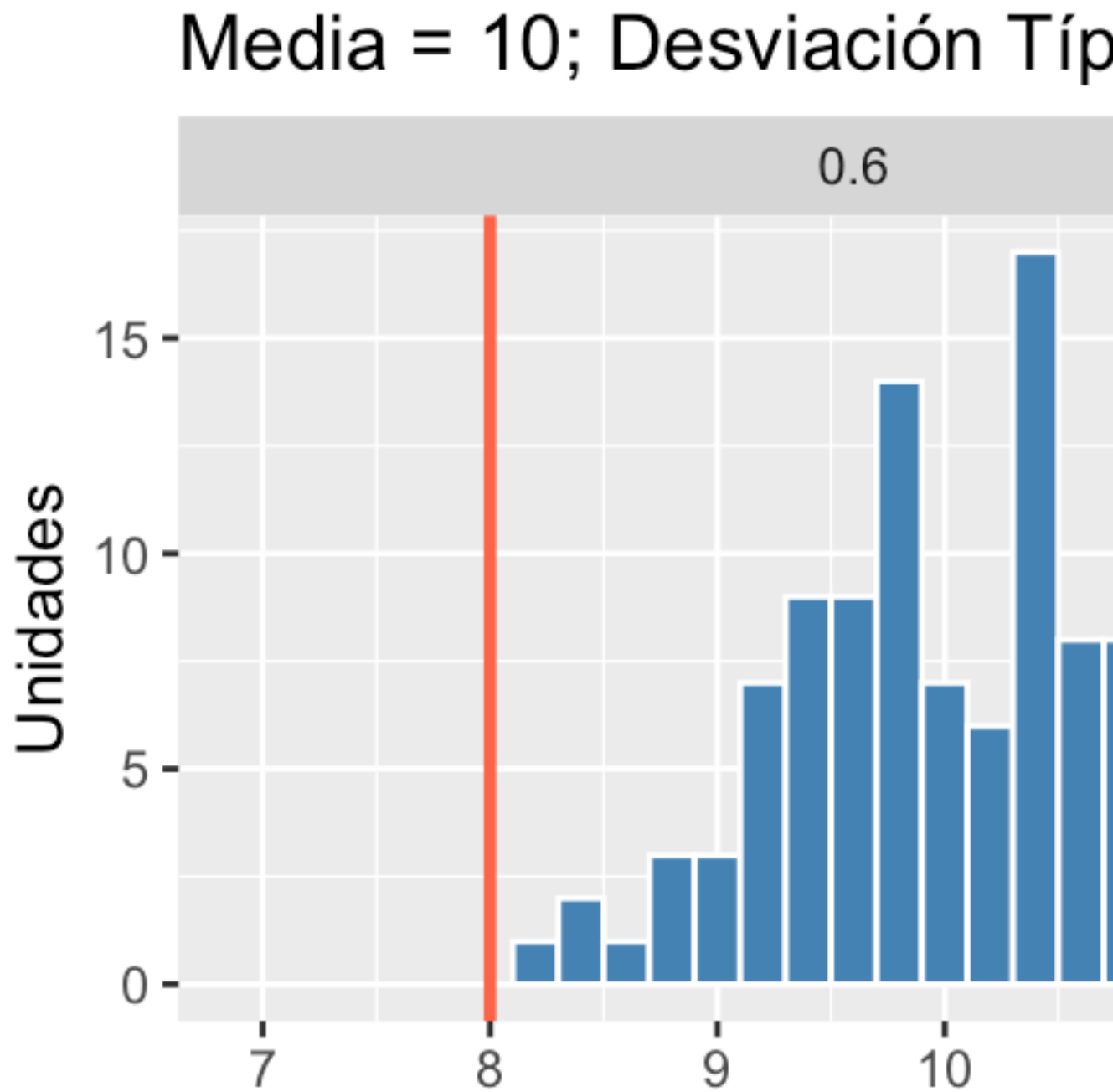


Figura 1.3: Procesos con la misma media y distinta variabilidad

gris), que solo se produce con las no conformidades. El análisis de la función de pérdida es una herramienta muy útil en proyectos de mejora, véase Cano et al. (2012).

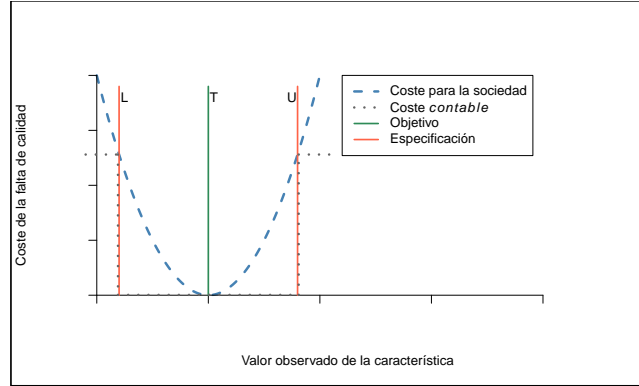
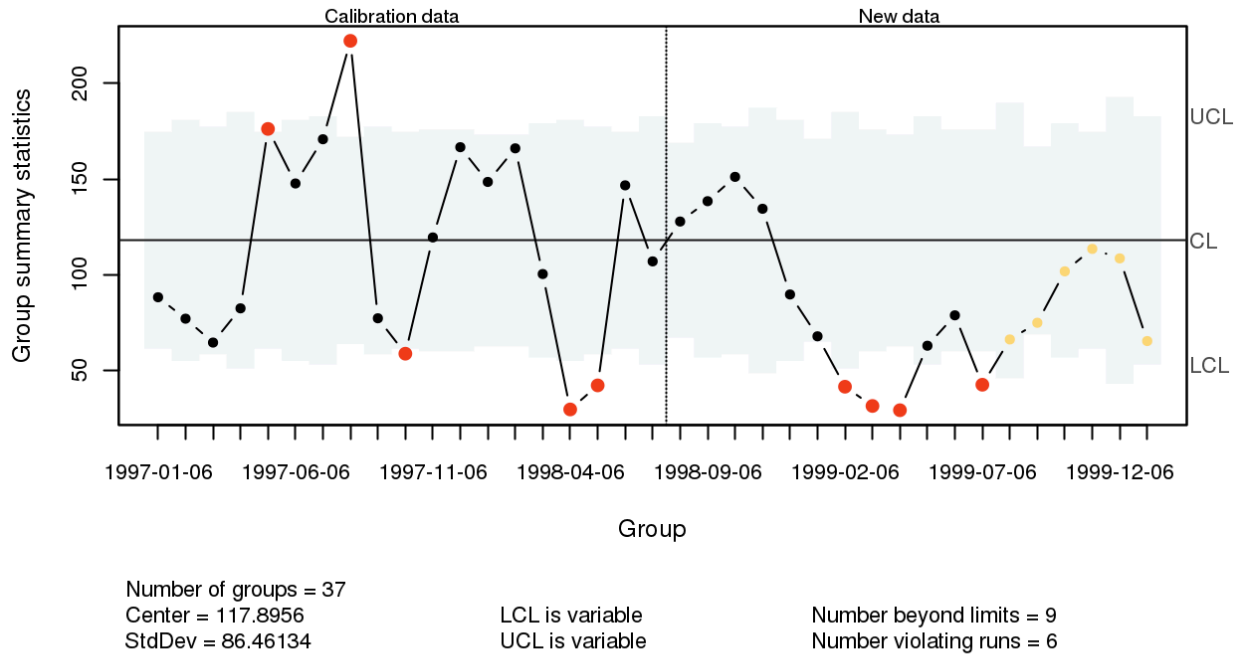
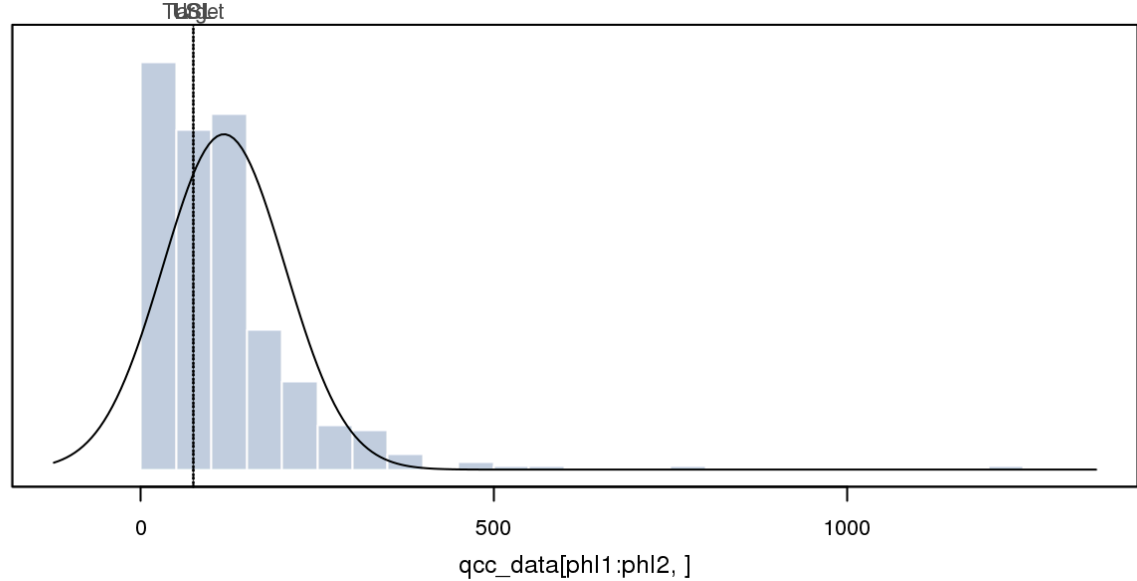


Figura 1.4: Función de pérdida de Taguchi

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 1.5 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

**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

las series de normas ISO 2859 e ISO 3951.

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 1.6.

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

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

3. **Salud y bienestar**- Garantizar una vida sana y promover el bienestar para todos en todas las edades
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



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

uno o varios ODS

- Tener presentes los ODS para ser sostenible en los propios análisis. Por 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

Resúmenes numéricos

Resúmenes gráficos

Valores atípicos

Valores perdidos

En preparación.

Capítulo 3

Análisis exploratorio bivariante

Representación gráfica

Correlación

Regresión

Intro multivariante

En preparación.

Parte II

Probabilidad

Capítulo 4

Introducción a la Probabilidad

En preparación.

Definiciones

Propiedades

Probabilidad total y Bayes

Capítulo 5

Variable aleatoria univariante

En preparación.

Definición

Función de distribución

VA discreta

VA continua

Capítulo 6

Variable aleatoria bivalente

En preparación.

Distribución conjunta

Correlación y regresión

Capítulo 7

Modelos de distribución de probabilidad

En preparación.

Introducción

Modelos discretos

Modelos continuos

Modelos multivariantes*

Parte III

Inferencia estadística

Capítulo 8

Muestreo y estimación

En preparación.

Muestreo estadístico

Estimación y contrastes

Estadísticos

Estimadores puntuales (medias, proporciones, varianzas)

Estimación por intervalos

Estimación no paramétrica

Inferencia Bayesiana*

Capítulo 9

Comparación de grupos

En preparación.

Comparación de atributos

Comparación de dos grupos

Comparación de más de dos grupos

Capítulo 10

Modelos de regresión

En preparación.

Regresión lineal simple

Regresión no lineal

Regresión lineal múltiple

Otros modelos*
(GLM, GAM, ...)

Capítulo 11

Diseño de experimentos

En preparación.

Intro

Diseños factoriales

Diseños 2^k

Diseños fraccionales

Parte IV

Control estadístico de la calidad

Capítulo 12

Introducción

En preparación.

Historia de la calidad

Estadística y calidad

Gestión de la calidad

Mejora de procesos vs control de calidad

Metodologías

Intro Six Sigma*

Capítulo 13

Control Estadístico de Procesos

En preparación.

Intro SPC

Gráficos de control

Capacidad y rendimiento

Capítulo 14

Inspección por muestreo

En preparación.

Intro

Planes para atributos

Planes para variables

Apéndice A

Símbolos, abreviaturas y acrónimos

A.1. Acrónimos

| Acrónimo | Descripción |
|----------|-----------------------------|
| SPC | Statistical Process Control |

A.2. Letras griegas

| Letra | Se lee |
|-----------|-----------------------------|
| α | alfa |
| β | beta |
| γ | gamma |
| Γ | Gamma* |
| λ | lambda |
| η | eta |
| μ | mu |
| ω | omega |
| Ω | Omega* |
| σ | sigma |
| Σ | Sigma* |
| ρ | ro |
| θ | zeta (<i>theta</i> , teta) |
| ξ | xi |
| χ | chi (o <i>ji</i>) |

| Letra | Se lee |
|---------------|---------|
| π | pi |
| ε | épsilon |

* Mayúsculas

A.3. Símbolos

| Símbolo | Se lee |
|-----------------|--|
| \emptyset | Conjunto vacío o suceso imposible |
| \aleph | Aleph |
| \wp | Probabilidad (como función) |
| : | Tal que |
| $P(\cdot)$ | Probabilidad de \cdot (sucesos) |
| $P[\cdot]$ | Probabilidad de \cdot (variables aleatorias) |
| $E[\cdot]$ | Esperanza de \cdot |
| \cdot | <i>lo que sea</i> (representa cualquier objeto matemático) |
| | Condicionado a |
| \sum | Sumatorio |
| $\sum_{i=1}^n$ | Sumatorio desde i igual a uno hasta n |
| \prod | Producto |
| $\prod_{i=1}^n$ | Producto desde i igual a uno hasta n |
| \forall | Para todo |
| \in | Pertenece/perteneciente |
| \exists | Existe |
| \Rightarrow | Implica/entonces |
| ∂ | Derivada parcial |
| \simeq | Aproximadamente igual ¹ |
| \approx | Aproximadamente ² |
| \equiv | Equivalente |
| \mathbb{R} | Conjunto de los números reales |
| \cup | Unión |
| \cap | Intersección |
| \subset | Incluido |
| \subseteq | Incluido o igual |

¹En este libro se usa sobre todo para indicar que se ha redondeado un número decimal

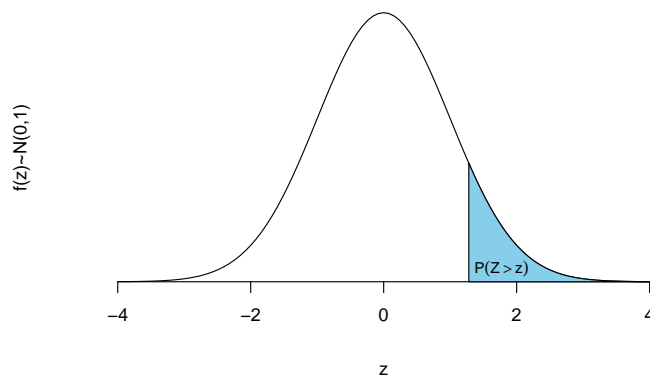
²En este libro se puede utilizar para tomar el entero superior o inferior según el contexto

Apéndice B

Tablas estadísticas

B.1. Distribución normal

La siguiente tabla contiene la probabilidad de la cola superior de la distribución normal estándar $Z \sim N(0; 1)$, es decir $1 - F(z) = P[Z > z]$.



[illegible]

B.2. Resumen modelos de distribución de probabilidad

| Distribución | Probabilidad/Densidad/Distribución |
|-------------------------------------|--|
| $\text{Bernoulli}(\mathit{Ber}(p))$ | $X = \begin{cases} 1 & \text{con probabilidad } p \\ 0 & \text{con probabilidad } 1-p \end{cases}$ |

Apéndice C

Repaso

Este apéndice cubre algunas cuestiones matemáticas básicas que el lector de este libro con seguridad habrá aprendido con anterioridad. Se incluyen como referencia para facilitar el repaso a aquellos que lo necesiten.

C.1. Logaritmos y exponenciales

C.2. Combinatoria

Una de las definiciones de probabilidad implica **contar** el número de veces que puede ocurrir un suceso determinado. Por tanto, en muchas ocasiones el cálculo de probabilidades empieza contando las posibilidades de que ocurra un suceso. La Combinatoria es la parte de la Matemática discreta que nos ayuda en esta tarea. Incluimos un breve resumen con ejemplos de las fórmulas más habituales y su cálculo con R.

C.2.1. Ejemplo ilustrativo

Habitualmente se utilizan ejemplos de juegos de azar para introducir el cálculo de probabilidades, como lanzando monedas y dados, o combinaciones de cartas en barajas de naipes. Para darle un enfoque práctico, utilizaremos a lo largo del módulo un ejemplo ilustrativo que, aunque totalmente inventado, se puede encontrar el lector en el futuro con ligeras variaciones según su ámbito de actuación. Utilizaremos en lo posible las cifras usadas en los problemas de azar para ver la utilidad de aquéllos ejemplos en casos más prácticos.

Datos básicos:

- 52 posibles usuarios de un servicio
- La mitad son mujeres

- 4 directivos, 12 mandos, resto operarios
- 13 jóvenes, 26 adultos, 13 mayores (5, 18 y 3 mujeres en cada grupo respectivamente)
- 1 de cada seis hombres contratará el servicio (el doble si es mujer)

Nótese cómo podemos *traducir* el concepto de servicio a cualquier ámbito: usuarios de salud o educación, enfermos de una determinada patología, equipos de una infraestructura, etc. Asimismo las categorías pueden ser cualesquiera aplicables a los elementos de los conjuntos.

C.2.2. Principio básico de conteo

Definición: Realizamos k experimentos sucesivamente, cada uno de ellos con n_i posibles resultados ($i = 1, \dots, k$). Entonces el número total de resultados posibles es:

$$n_1 \cdot n_2 \cdot \dots \cdot n_k$$

Ejemplo: Resultados posibles si tomamos al azar un individuo y observamos su grupo de edad y si contratará o no el servicio.

Código

```
3*2
#> [1] 6
```

C.2.3. Permutaciones

Definición: De cuántas formas posibles podemos ordenar un conjunto de n elementos sin repetirlos.

$$P_n = n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 2 \cdot 1$$

Ejemplo: De cuántas formas podemos ordenar un conjunto de tres individuos, uno de cada categoría laboral.

Código

```
factorial(3)
#> [1] 6
```

C.2.4. Variaciones (muestreo sin reemplazamiento)

Definición: De cuántas formas posibles podemos seleccionar una muestra de n elementos de un conjunto total de m , sin que se repitan. Una ordenación distinta, es una posibilidad distinta.

$$V_{m,n} = m \cdot (m-1) \cdot (m-2) \cdot \dots \cdot (m-n+1) = \frac{m!}{(m-n)!}$$

Ejemplo: De cuántas formas podemos seleccionar una muestra de 5 individuos en nuestro conjunto de 52 sin que se repitan (por ejemplo para asignar un ranking)

Código

```
factorial(52)/factorial(52-5)
#> [1] 311875200
```

C.2.5. Variaciones con repetición (muestreo con reemplazamiento)

Definición: De cuántas formas posibles podemos seleccionar una muestra de n elementos de un conjunto total de m , pudiéndose repetir. Una ordenación distinta, es una posibilidad distinta.

$$VR_{m,n} = m^n$$

Ejemplo: De cuántas formas podemos seleccionar una muestra de 5 individuos en nuestro conjunto de 52 pudiéndose repetir (por ejemplo para asignar premios consecutivamente)

Código

```
52^5
#> [1] 380204032
```

C.2.6. Combinaciones (muestras equivalentes)

Definición: De cuántas formas posibles podemos seleccionar una muestra de n elementos de un conjunto total de m , sin importar el orden.

$$C_{m,n} = \binom{m}{n} = \frac{m!}{n!(m-n)!}$$

$\binom{m}{n}$ se lee *m sobre n*, y se le conoce como *número combinatorio*. Algunas propiedades importantes de los números combinatorios:

$$\binom{m}{m} = \binom{m}{0} = 1.$$

$$\binom{m}{1} = \binom{m}{m-1} = m.$$

$$\binom{m}{n} + \binom{m}{n+1} = \binom{m+1}{n+1}$$

Por otra parte, por convenio se tiene que:

$$0! = 1,$$

$$\text{si } a < b \implies \binom{a}{b} = 0.$$

Ejemplo: De cuántas formas podemos seleccionar una muestra de 5 individuos en nuestro conjunto de 52 sin importar el orden (por ejemplo para asignar premios de una sola vez)

Código

```
choose(52, 5)
#> [1] 2598960
```

C.2.7. Combinaciones y permutaciones con repetición

Las combinaciones y permutaciones también se pueden dar con repetición, siendo las fórmulas para calcularlas las siguientes:

$$CR_{m,n} = C_{m+n-1,n} = \frac{(m+n-1)!}{n! \cdot (m-1)!}$$

$$PR = \frac{n!}{a! \cdot b! \cdot \dots \cdot z!}$$

La primera situación es aquella en la que los elementos se pueden repetir, pero no nos importa el orden en que lo hagan. La segunda aparece cuando el elemento A del conjunto total de elementos aparece a veces, y así sucesivamente.

Apéndice D

Ampliación

En este apéndice se incluyen temas avanzados que pueden ser útiles al lector más allá de un curso básico de estadística para ciencias o ingeniería, y que no se han incluido en el cuerpo de los capítulos para mantener el nivel de una asignatura de grado.

D.1. Función característica

D.2. Cambio de variable

D.3. Variables aleatorias unidimensionales mixtas

D.4. Variables aleatorias bidimensionales mixtas

D.5. Algunos modelos de distribución continuos más

D.5.1. Distribución Beta

La distribución Beta se utiliza en problemas de inferencia relativos a proporciones, especialmente en inferencia bayesiana.

$$X \sim Be(\alpha, \beta)$$

Función de densidad

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} & \text{si } 0 < x < 1 \\ 0 & \text{resto} \end{cases}$$

En matemáticas, la función Gamma (Γ) es una integral indefinida que tiene entre otras las siguientes propiedades:

- $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx, \quad \alpha > 0$
- $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$
- $n \in \mathbb{N} - \{0\} \implies \Gamma(n) = (n-1)!$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

**** Características****

- Esperanza: $E[X] = \frac{\alpha}{\alpha+\beta}$
- Varianza: $Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Caso particular: $Be(1, 1) = U(0, 1)$.

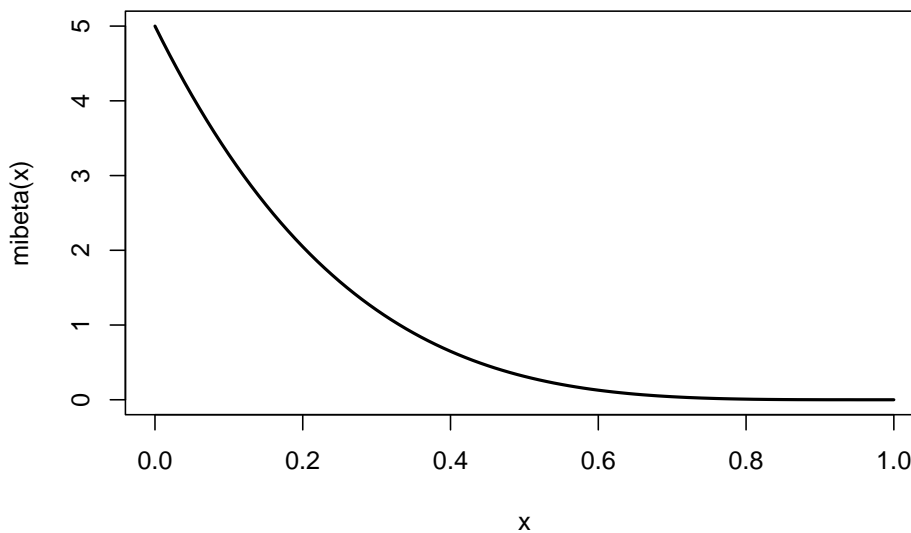
Ejemplo

X : Proporción de clientes que contratarán el servicio

$X \sim Be(1, 5)$

Código

```
mibeta <- function(x) dbeta(x, 1, 5)
curve(mibeta, lwd = 2)
```



D.5.2. Distribución Gamma

La distribución Gamma se utiliza, entre otros, para modelizar tiempos de espera hasta que suceden α eventos en un proceso de Poisson. De hecho, en inferencia bayesiana gamma es la distribución a priori de la distribución de Poisson.

$$X \sim Ga(a, b)$$

Función de densidad

$$f(x) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} & \text{si } 0 < x < \infty \\ 0 & \text{resto} \end{cases}$$

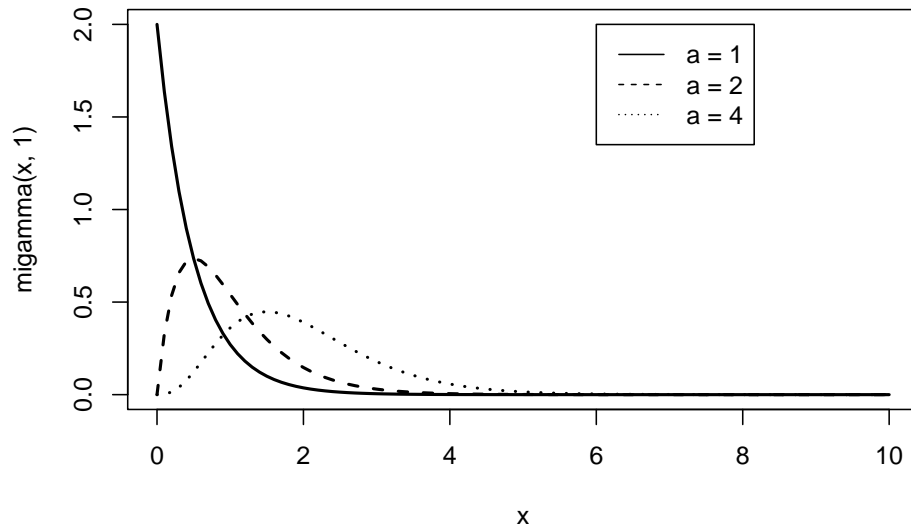
Características

- Esperanza: $E[X] = \frac{a}{b}$
- Varianza: $Var[X] = \frac{a}{b^2}$
- $\int_0^{\infty} x^{a-1} e^{-bx} dx = \frac{\Gamma(a)}{b^a}$
- La exponencial es un caso particular

Código

```
migamma <- function(x, a) dgamma(x, a, 2)
curve(migamma(x, 1), lwd = 2, xlim = c(0,10),
      main = "Distribución Gamma b = 2")
curve(migamma(x, 2), lwd = 2, add = TRUE, lty = 2)
curve(migamma(x, 4), lwd = 2, add = TRUE, lty = 3)
legend(x = 6, y = 2, c("a = 1", "a = 2", "a = 4"), lty = 1:3)
```

Distribución Gamma b = 2



D.5.3. Distribución de Weibull

La distribución Gamma presenta algunos inconvenientes al modelizar tiempos de vida, y por eso algunas veces se prefiere la distribución de Weibull, que básicamente sirve para lo mismo. Véase Ugarte et al. (2015) para los detalles.

$$X \sim We(a, b)$$

Función de densidad

$$f(x) = \begin{cases} \frac{a}{b} \left(\frac{x}{b}\right)^{a-1} e^{-(x/b)^a} & \text{si } x > 0 \\ 0 & \text{resto} \end{cases}$$

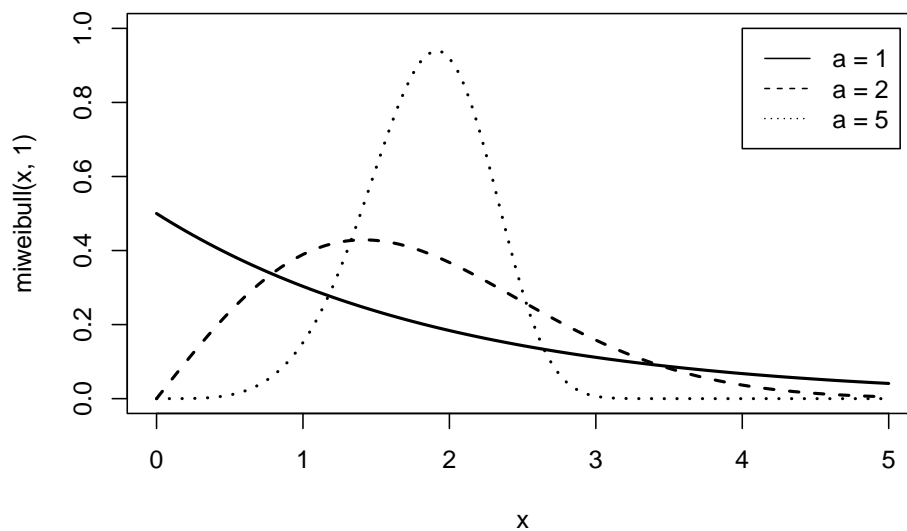
Características

- Esperanza: $E[X] = b\Gamma\left(1 + \frac{1}{a}\right)$
- Varianza: $Var[X] = b^2 \left(\Gamma\left(1 + \frac{2}{a}\right) - \left(\Gamma\left(1 + \frac{1}{a}\right)\right)^2 \right)$

Código

```
miweibull <- function(x, a) dweibull(x, a, 2)
curve(miweibull(x, 1), lwd = 2, xlim = c(0,5),
      ylim = c(0, 1),
      main = "Distribución Weibull b = 2")
curve(miweibull(x, 2), lwd = 2, add = TRUE, lty = 2)
curve(miweibull(x, 5), lwd = 2, add = TRUE, lty = 3)
legend(x = 4, y = 1, c("a = 1", "a = 2", "a = 5"), lty = 1:3)
```

Distribución Weibull $b = 2$



D.6. Modelos de distribución de probabilidad multivariantes

D.7. Modelos de distribución de probabilidad relacionadas con la normal

D.8. Simulación de variables aleatorias

$U(0; 1)$: Generador de probabilidades aleatorias. Dada cualquier función de distribución F , se pueden generar valores de esa VA obteniendo $F^{-1}(U(0; 1))$

Apéndice E

Demostraciones

Em este apéndice se incluyen aquellas demostraciones de teoremas y propiedades no incluidas en los capítulos para mantener el carácter práctico del mismo.

E.1. Variable aleatoria discreta

E.1.1. Función de probabilidad

E.1.2. Esperanza

E.1.3. Varianza

Apéndice F

Créditos

Los gráficos y diagramas generados son creación y propiedad del autor, salvo que se indique lo contrario. Su licencia de uso es la misma que la del resto de la obra, véase el Prefacio.

La imagen de la portada es de dominio público, obtenida en pixabay.com, gracias al usuario Manuchi.

Las imágenes de tipo *clipart* usadas en esta obra y las fotografías no atribuidas pertenecen al dominio público gracias a openclipart.org, unplash.com o pixabay.com.

The R logo is (c) 2016 The R Foundation.

Bibliografía

- Cano, E. L., Moguerza, J. M., and Corcoba, M. P. (2015). *Quality Control with R. An ISO Standards Approach*. Use R! Springer.
- Cano, E. L., Moguerza, J. M., and Redchuk, A. (2012). *Six Sigma with R. Statistical Engineering for Process Improvement*, volume 36 of *Use R!* Springer, New York.
- Cleveland, W. (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review*, 69(1):21.
- López Cano, E. (2018). Estadística económica y empresarial. Libro de apuntes con licencia Creative Commons.
- López Cano, E. (2019). Análisis de datos con r aplicado a la economía, la empresa y la industria. Libro de apuntes con licencia Creative Commons.
- Ocaña-Riola, R. (2017). La necesidad de convertir la estadística en profesión regulada. *Estadística Española*, 59(194):193–212.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Taguchi, G., Chowdhury, S., and Wu, Y. (2007). *Taguchi's quality engineering handbook*. John Wiley.
- Tukey, J. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.
- Tukey, J. W. et al. (1977). *Exploratory data analysis*, volume 2. Reading, Mass.
- Ugarte, M., Militino, A., and Arnholt, A. (2015). *Probability and Statistics with R, Second Edition*. CRC Press.
- Xie, Y. (2021). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.24.