

Estadística Aplicada a las Ciencias y la Ingeniería

Emilio L. Cano

2022-02-22

Índice general

Bienvenida	7
Estándares y software	7
Estructura del libro	8
Sobre el autor	10
Agradecimientos	11
I Estadística descriptiva	13
1. Introducción	15
1.1. Estadística y análisis de datos	15
1.2. Los datos y su organización	19
1.3. La Estadística y el método científico	22
1.4. Estadística, Calidad y Sostenibilidad	23
1.5. Objetivos de Desarrollo Sostenible (ODS)	28
2. Análisis exploratorio univariante	33
2.1. La importancia del análisis exploratorio	33
2.2. Calidad de datos	36
2.3. Componentes de un gráfico	41
2.4. Notación	42
2.5. Análisis exploratorio de variables cualitativas	43
2.6. Análisis exploratorio de variables cuantitativas	48
2.7. Resúmenes de variables continuas	49
3. Análisis exploratorio bivariante	71
3.1. Frecuencias conjuntas, marginales y condicionadas	71
3.2. Datos bivariantes y multivariantes	71
3.3. Tabla de frecuencias conjunta	71
3.4. Ejemplo: tabla de contingencia	72
3.5. Ejemplo: Variables continuas	72
3.6. Frecuencias marginales	73
3.7. Ejemplo frecuencias marginales	73

3.8. Ejemplo distribuciones marginales	74
3.9. Distribuciones condicionadas	74
3.10. Ejemplo distribuciones condicionadas	74
3.11. Independencia de variables	75
3.12. 2.3. Representación gráfica conjunta	76
3.13. Gráficos de barras para frecuencias conjuntas	76
3.14. El gráfico de dispersión	76
3.15. Gráfico de dispersión “enriquecido”	77
3.16. Gráficos de cajas por grupos	79
II Probabilidad	81
4. Introducción a la Probabilidad	83
4.1. Introducción	83
4.2. Sucesos aleatorios	84
4.3. Definiciones de probabilidad y sus propiedades	92
4.4. Probabilidad condicionada y sus consecuencias	102
5. Variable aleatoria univariante	113
5.1. Concepto y definición de variable aleatoria	113
5.2. Función de distribución	116
5.3. Variable aleatoria discreta	117
5.4. Variable aleatoria continua	121
5.5. Características de una variable aleatoria	136
6. Variable aleatoria bivariante	153
7. Modelos de distribución de probabilidad	155
7.1. Introducción	155
7.2. Modelos de distribución de probabilidad discretos	156
7.3. Modelos de distribución de probabilidad continuos	174
7.4. Otros modelos de distribución de probabilidad	192
7.5. Convergencia de variables aleatorias	192
7.6. Distribuciones relacionadas con la normal	192
III Inferencia estadística	193
8. Muestreo y estimación	195
9. Comparación de dos grupos	197
10. Análisis de la Varianza	199
10.1. Introducción	199
10.2. Análisis de la varianza de un factor	200
10.3. Análisis de la varianza de varios factores	213

10.4. Introducción a los modelos mixtos: efectos fijos y efectos aleatorios	217
10.5. Análisis multivariante de la varianza	219
11. Diseño de experimentos	221
11.1. Introducción	221
11.2. Bases del DoE: origen, importancia, objetivos y requerimientos .	221
11.3. Importancia del diseño	222
11.4. Planificación de la experimentación	223
11.5. Tipos de diseños de experimentos	228
12. Modelos de regresión	241
IV Control estadístico de la calidad	243
13. Introducción	245
14. Control Estadístico de Procesos	247
15. Inspección por muestreo	249
A. Símbolos, abreviaturas y acrónimos	251
A.1. Acrónimos	251
A.2. Letras griegas	251
A.3. Símbolos	252
B. Tablas estadísticas	253
B.1. Distribución normal	253
B.2. Resumen modelos de distribución de probabilidad	255
C. Repaso	257
C.1. Logaritmos y exponentiales	257
C.2. Combinatoria	257
D. Ampliación	261
D.1. Función característica	261
D.2. Cambio de variable	261
D.3. Variables aleatorias unidimensionales mixtas	261
D.4. Variables aleatorias bidimensionales mixtas	261
D.5. Algunos modelos de distribución continuos más	261
D.6. Modelos de distribución de probabilidad multivariantes	265
D.7. Modelos de distribución de probabilidad relacionadas con la normal	265
D.8. Simulación de variables aleatorias	265
E. Demostraciones	267
E.1. Variable aleatoria discreta	267

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 pertinente 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 #>.

```
sessionInfo()
#> R version 4.1.2 (2021-11-01)
#> Platform: x86_64-apple-darwin17.0 (64-bit)
#> Running under: macOS Big Sur 10.16
#>
#> Matrix products: default
#> BLAS:    /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
#> LAPACK:  /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
#>
```

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/>

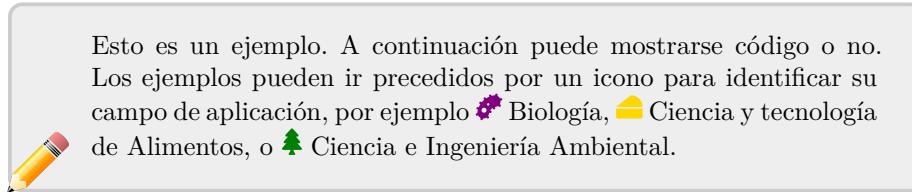
⁸<https://www.cen.eu/>

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

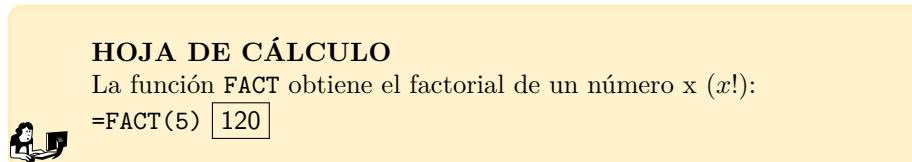
¹⁰<https://www.une.org/encuentra-tu-norma/comites-tecnicos-de-normalizacion/comite/?c=CTN%2066/SC%203>

```
#> locale:
#> [1] es_ES.UTF-8/es_ES.UTF-8/es_ES.UTF-8/C/es_ES.UTF-8/es_ES.UTF-8
#>
#> attached base packages:
#> [1] stats      graphics   grDevices utils      datasets
#> [6] methods    base
#>
#> other attached packages:
#> [1] fontawesome_0.2.2
#>
#> loaded via a namespace (and not attached):
#> [1] bookdown_0.24.3 digest_0.6.29  magrittr_2.0.2
#> [4] evaluate_0.14  rlang_1.0.1   stringi_1.7.6
#> [7] cli_3.1.1    rstudioapi_0.13 rmarkdown_2.11
#> [10] tools_4.1.2   stringr_1.4.0  xfun_0.29
#> [13] yaml_2.2.2    fastmap_1.1.0 compiler_4.1.2
#> [16] htmltools_0.5.2 knitr_1.37
```

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

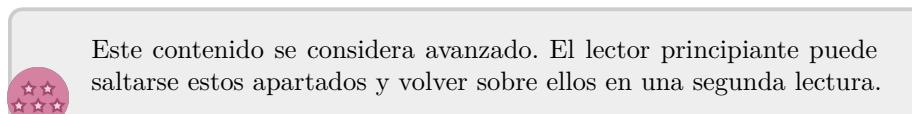


Cuando el ejemplo incluya explicaciones sobre cómo resolverlo con software, estas explicaciones aparecerán en bloques con el siguiente aspecto:



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:





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 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.

rado 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.

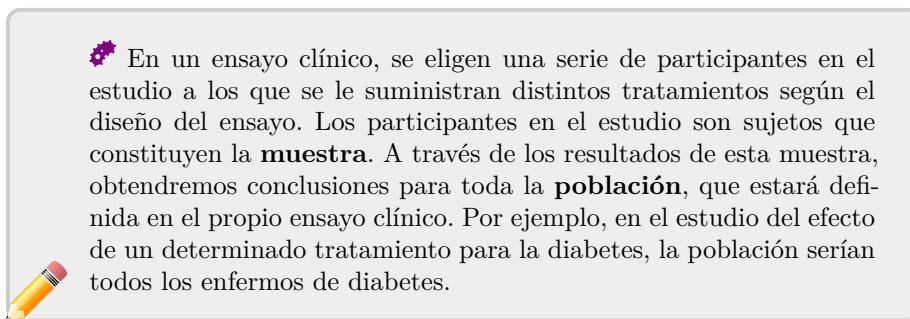
¹Por ejemplo en el Curso de Estadística Práctica Aplicada a la Calidad de la AEC.

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 2.1 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 este 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.



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

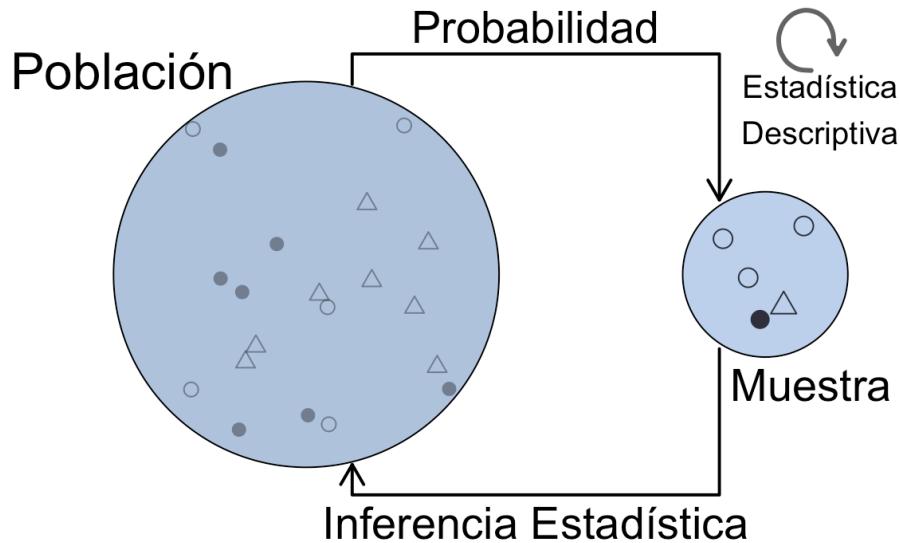


Figura 1.1: La esencia de los métodos estadísticos

$$Y = f(X)$$

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.

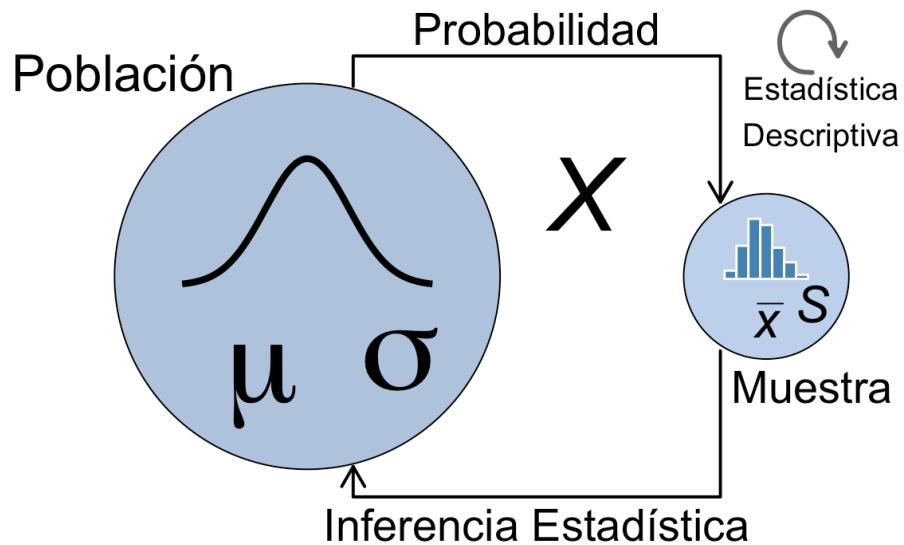


Figura 1.2: La esencia de los métodos estadísticos

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 frequentista** 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 **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 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.

Media = 10; Desviación Típica = 0.6

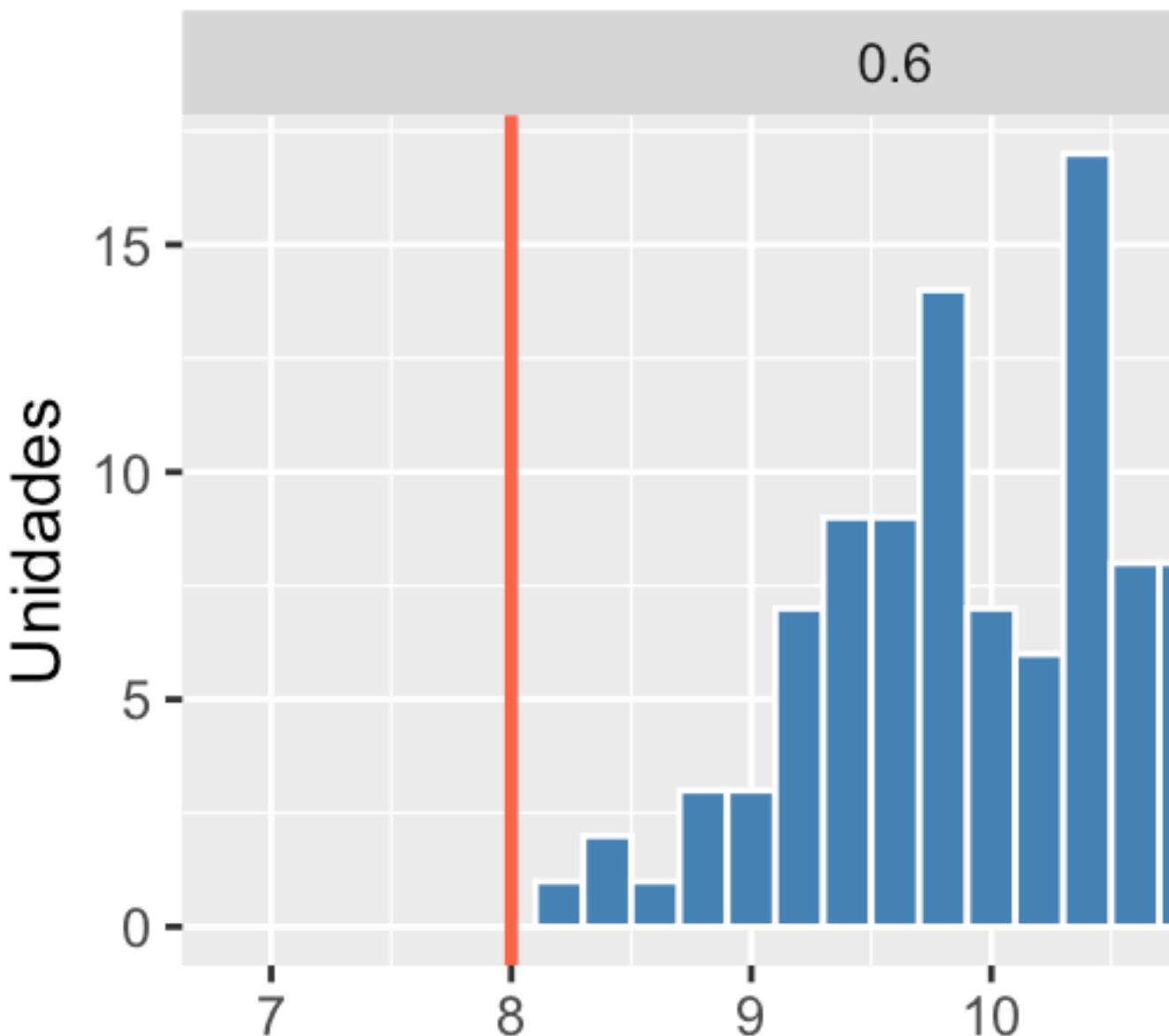


Figura 1.3: Procesos con la misma media y distinta variabilidad

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).

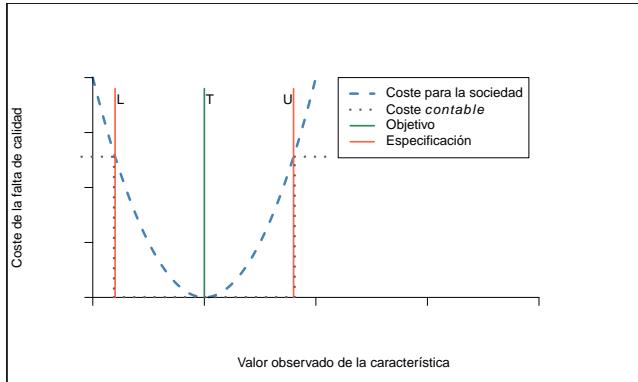


Figura 1.4: Función de pérdida de Taguchi

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

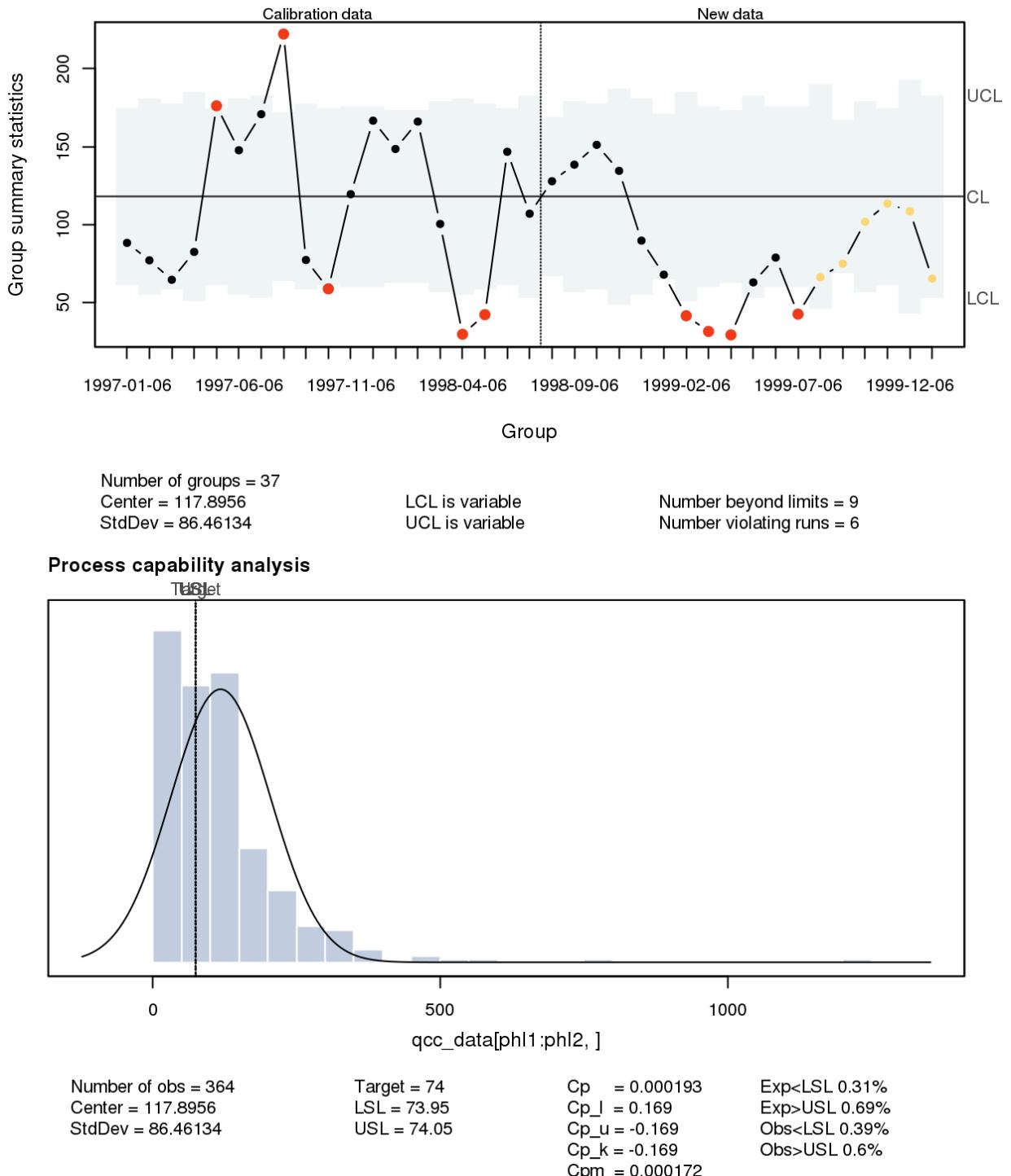


Figura 1.5: Gráficos de control y capacidad del proceso

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
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 2.1 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 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 en la unidad 4, 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

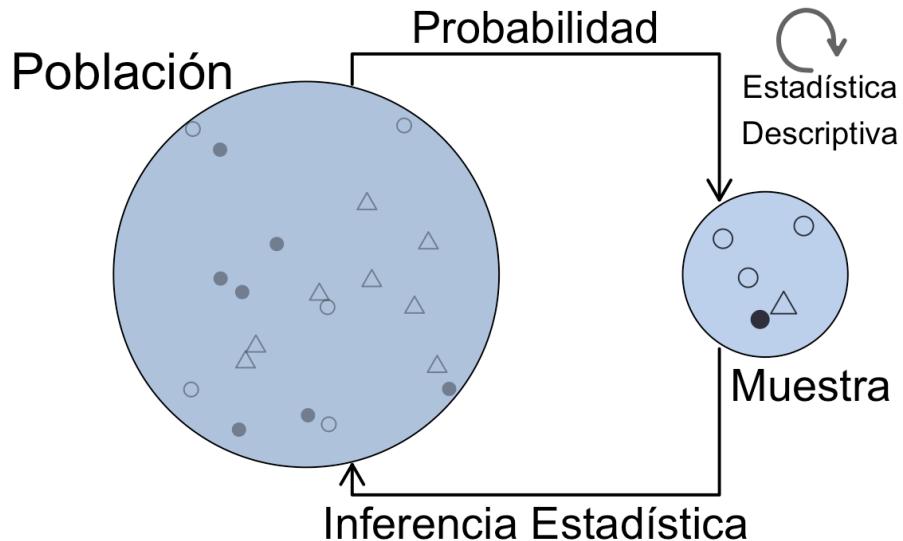


Figura 2.1: La esencia de los métodos estadísticos

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” (Anscombe, 1973), disponible con el nombre `anscombe` en el paquete `datasets` de R base. La tabla 2.1 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))
#>   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 2.2 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 horizontal los nombres de cada variable. Vemos que, a pesar de tener medias prácticamente iguales, los datos son muy diferentes.

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 2.3 muestra los cuatro gráficos que constituyen “El cuarteto de Anscombe”, y que se puede obtener de la propia ayuda del conjunto de datos

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

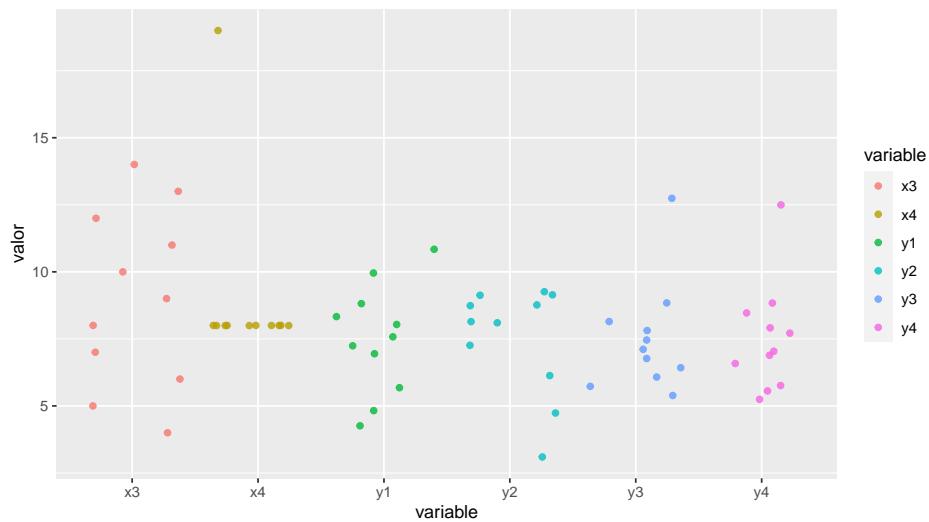


Figura 2.2: Representación de las variables del cuarteto de Anscombe

(`example(anscombe)`). La línea de regresión que se ajusta es prácticamente la misma (veremos la regresión en la unidad ??). Además, si calculáramos los coeficientes de correlación entre las variables “x” e “y” de los cuatro gráficos, obtendríamos el mismo valor: 0.8163.

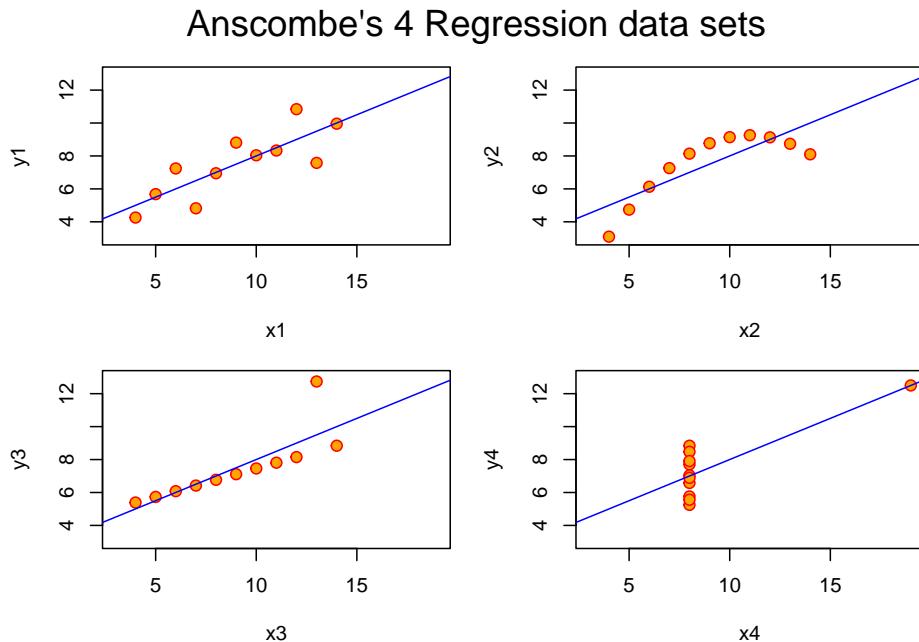


Figura 2.3: Los cuatro gráficos que constituyen ‘El cuarteto de Anscombe’

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*: (Matejka and Fitzmaurice, 2017) 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 (ISO, 2010), 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 (Tukey et al., 1977) y recogido en la norma UNE-ISO 16269 (ISO, 2010) 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 (ISO, 2010) 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 (ISO, 2010).

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 ISO (2010) 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

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

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 especialmente 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 seguro” 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 Wikipedia (2018))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** (Wikipedia, 2019). 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.

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

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 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 2.4 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.

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



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

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 del 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 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

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

[·], que simboliza un estimador de ·. 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 8 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.



fecha	codigo	est	mg	sal	ph	ebacterias	analista	tipo	bacteriasx	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 2.5.

El segundo fragmento de código produce la figura 2.6, que representa un gráfico de sectores con etiquetas realizado con el paquete {ggstatsplot}.



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

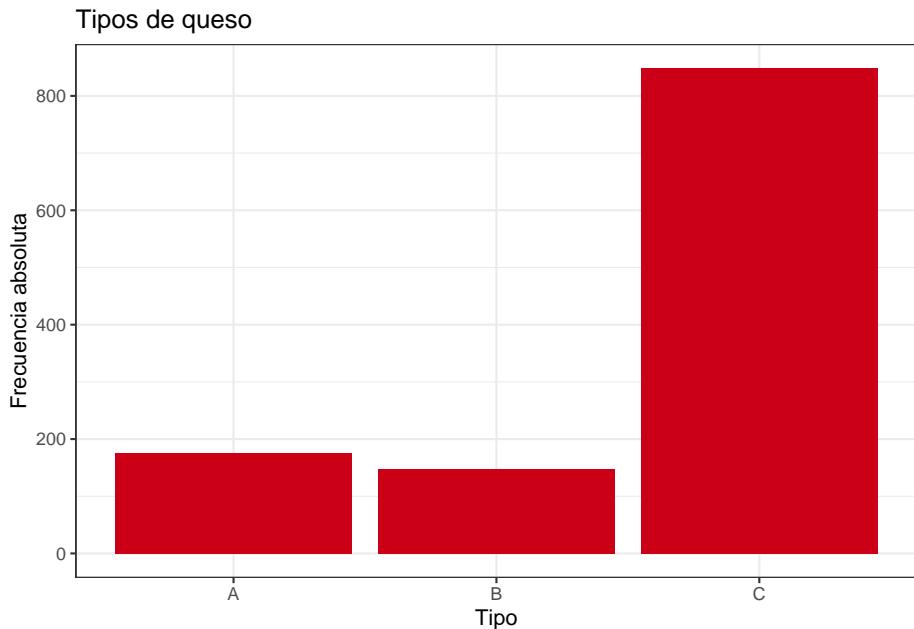


Figura 2.5: Ejemplo gráfico de barras variable cualitativa

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

Fabricación de quesos

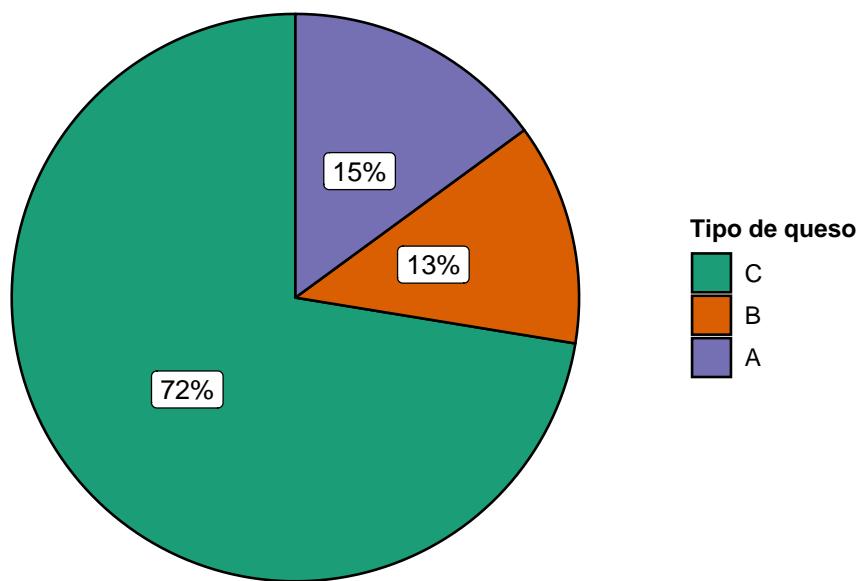


Figura 2.6: Gráfico de sectores con etiquetas

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 5 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 2.7 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

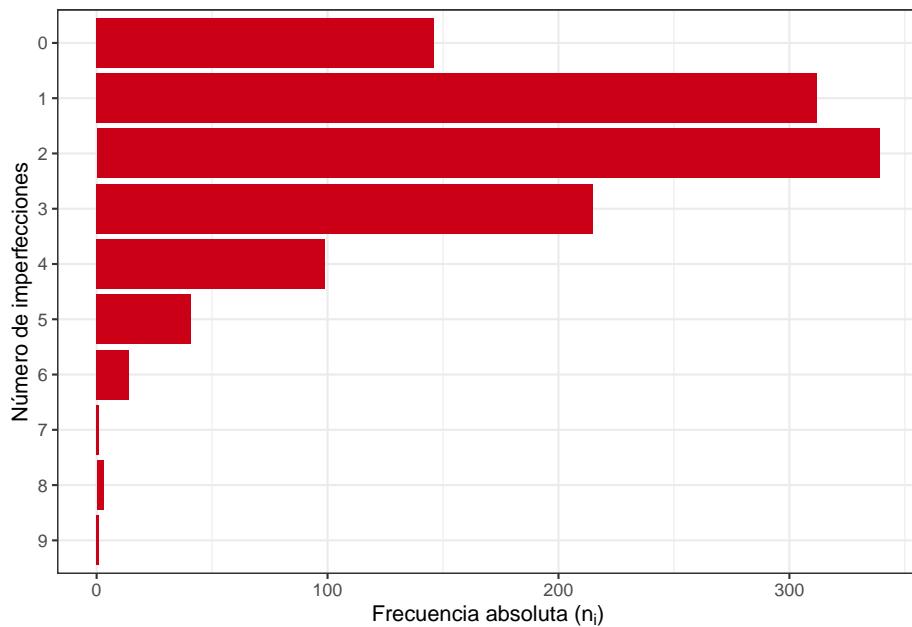


Figura 2.7: Gráfico de barras de la variable discreta imperfecciones

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 de frecuencias por intervalos

$\backslash((L_{\{i-1\}}, L_i]\backslash)$	$\backslash(n_i\backslash)$	$\backslash(F_i\backslash)$
(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 2.5 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 con 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, buscamos 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 2.8 representa la tabla de frecuencias 2.5. 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 2.9 representa el histograma de la variable `bacteriax`. 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 2.10 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.

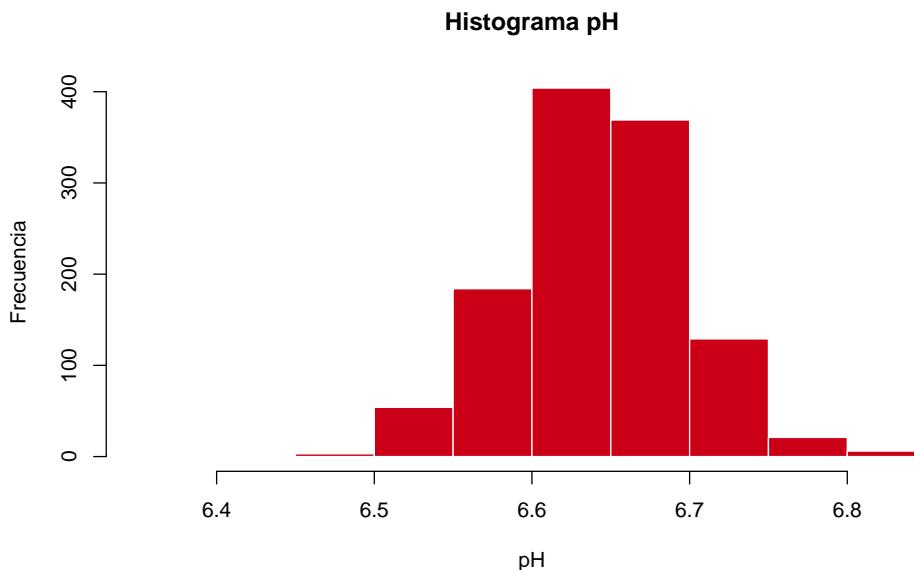


Figura 2.8: Histograma de la variable ph

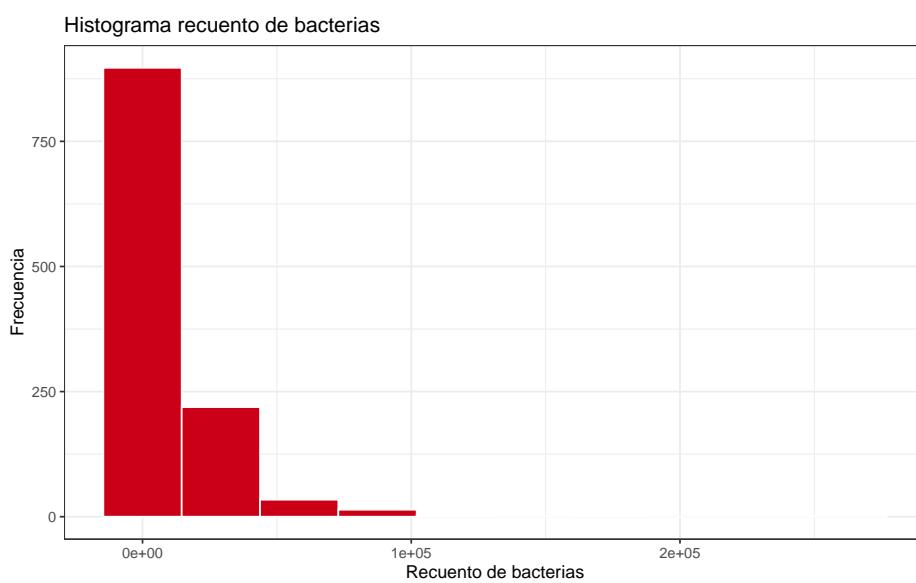


Figura 2.9: Histograma de la variable bacterias

Histograma contenido en sal

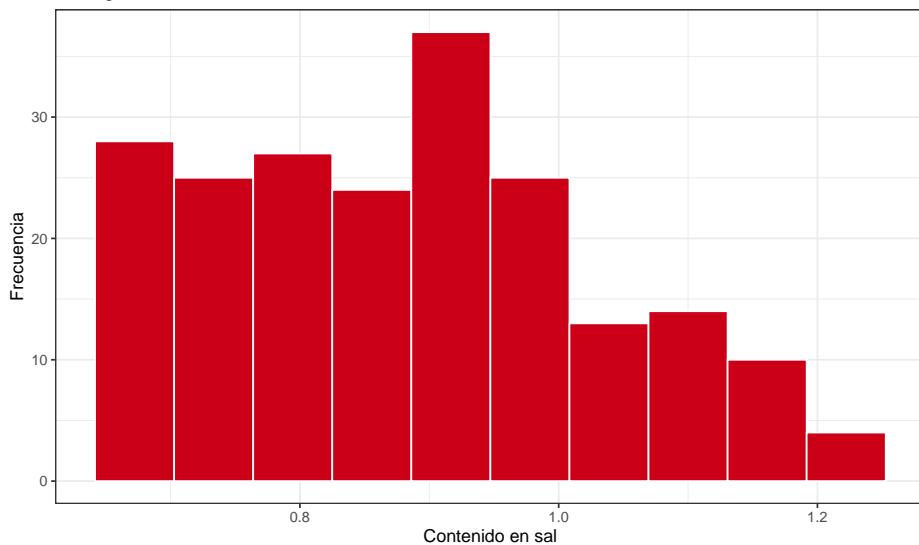


Figura 2.10: Histograma de la variable sal

El gráfico de arriba en la figura 2.11 muestra el histograma de la figura 2.10 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 2.11 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 5.



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

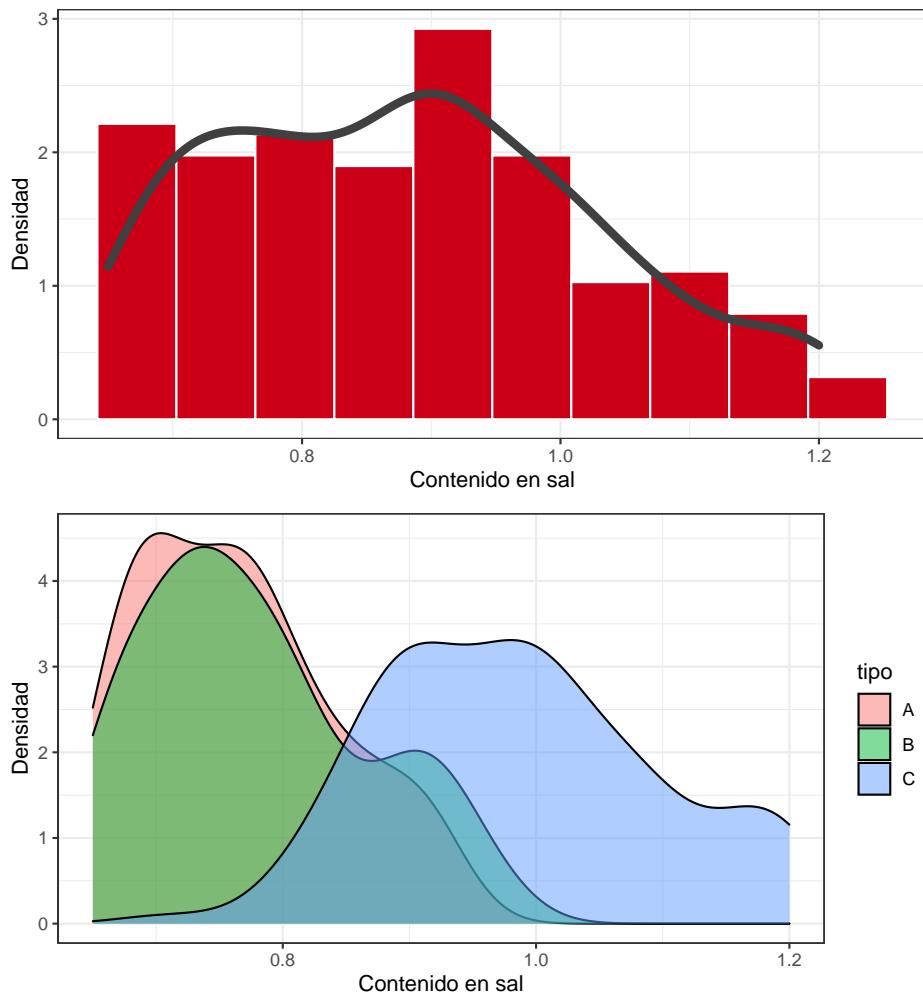


Figura 2.11: Ejemplo de gráficos de densidad

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 (Sarasola, 2018):

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

Es importante resaltar que la moda en variables continuas va a ser diferente según elijamos 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 2.12 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	5.379	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

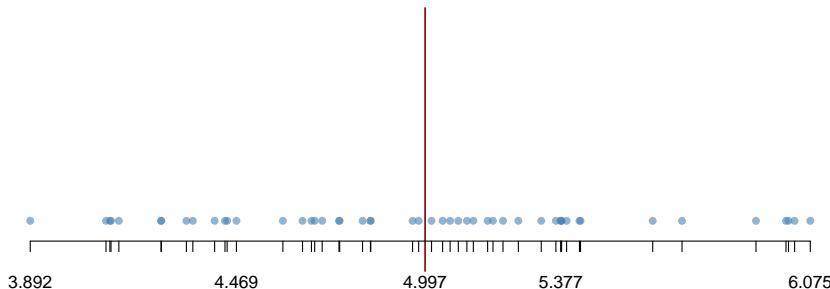


Figura 2.12: Significado gráfico de la mediana y los cuantiles

máximo, x_{max} , y el **mínimo**, x_{min} , que nos informan de los extremos de los datos. Los **cuartiles** 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 2.13 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

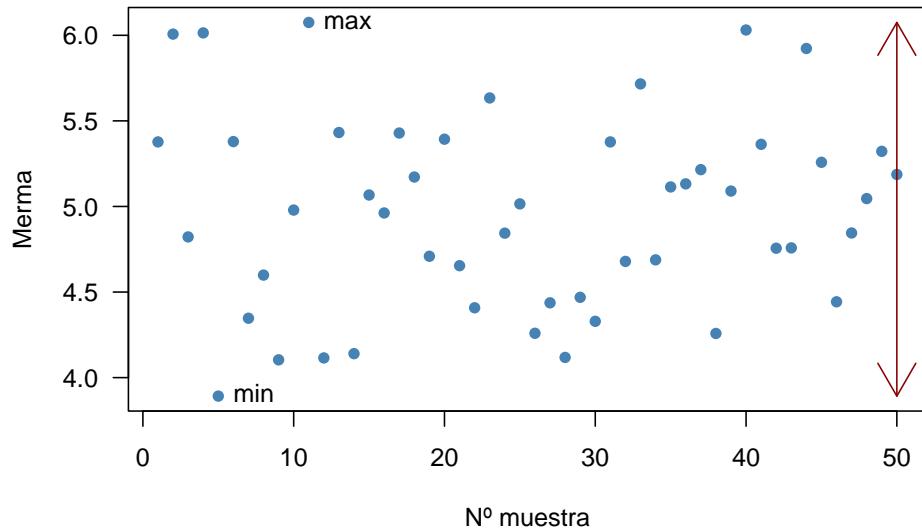


Figura 2.13: Representación del rango

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:

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

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:

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

La figura 2.14 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:



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

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:

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

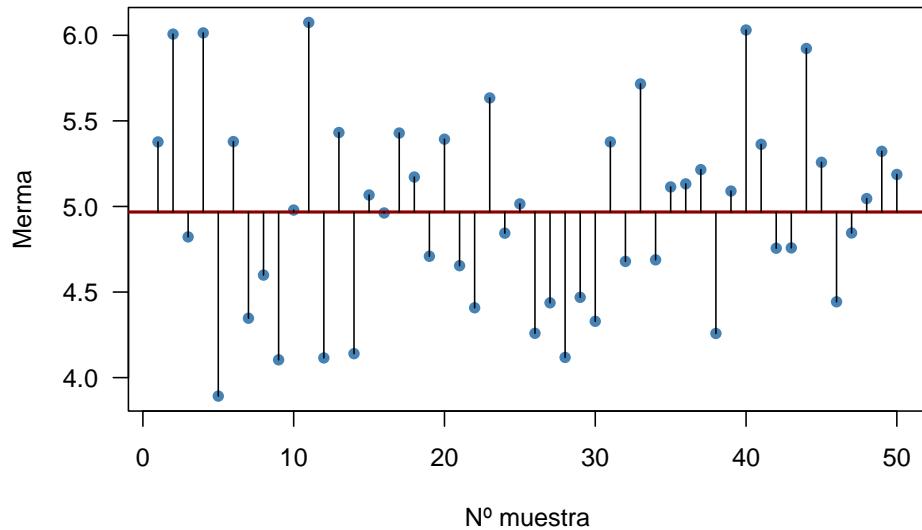


Figura 2.14: Desviaciones de la media

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

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

o con frecuencias relativas:

$$s^2 = \sum_{i=1}^n f_i x_i^2 - \bar{x}^2.$$

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 1.2.4, 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 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^2 , 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 $\bar{z} = 0; s = 1$. 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 5.

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 2.15 muestra el gráfico de cajas del ejemplo de la mermelada 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

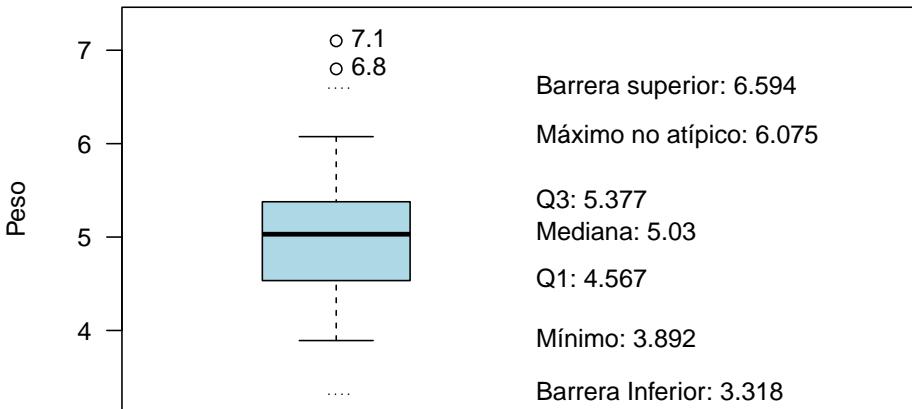


Figura 2.15: Explicación de los estadísticos representados en el gráfico de cajas

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.

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

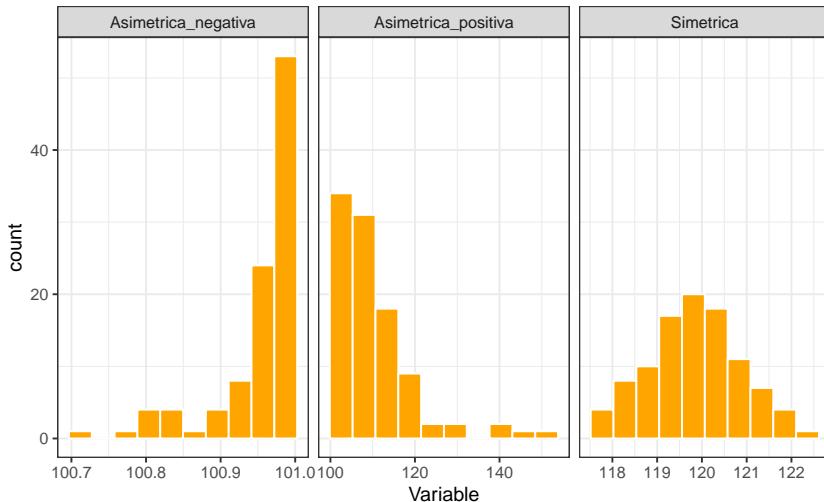
donde $m_3 = \frac{1}{n} \sum_{i=1}^n (x - \bar{x})^3$

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

Tipo	$\backslash(\backslash gamma_1\backslash)$
Asimetrica_negativa	-1,917
Asimetrica_positiva	1,943
Simetrica	0,089



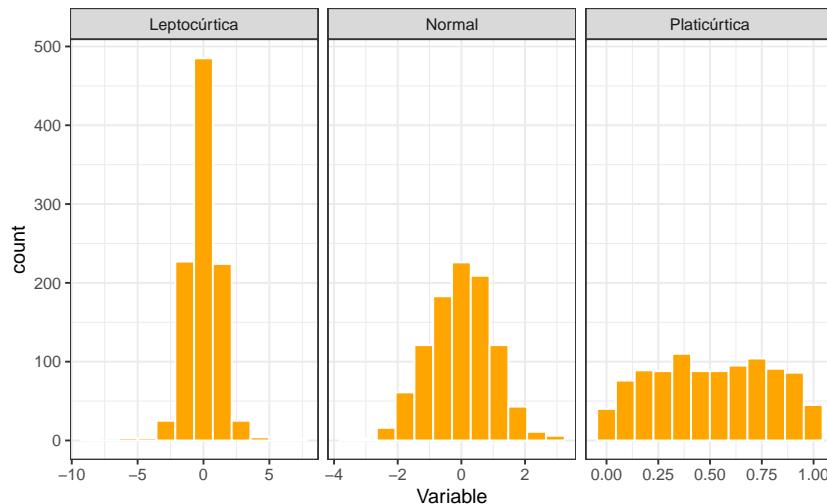


Figura 2.17: Histogramas de variables con distintos tipos de curtosis

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 2.18 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 2.19.

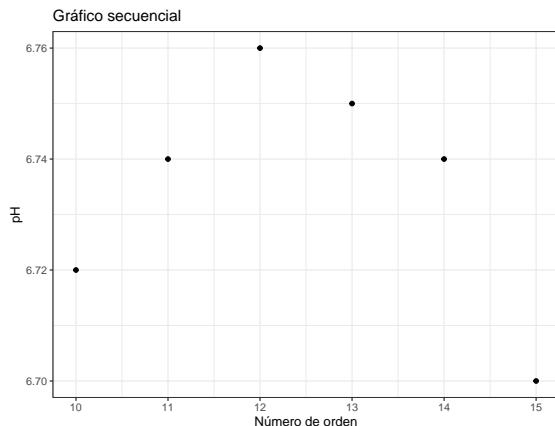


Figura 2.18: Gráfico de puntos secuencial

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

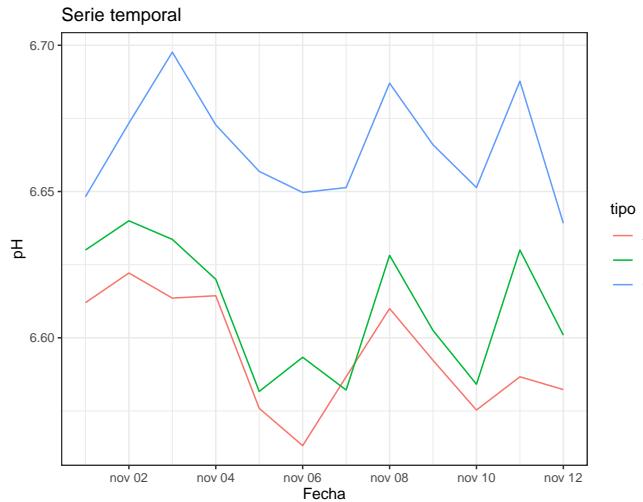


Figura 2.19: Gráfico de serie temporal

```
knitr:::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE, comment = "")
library(tidyverse)
#> -- Attaching packages ----- tidyverse 1.3.1 --
#> #> v ggplot2 3.3.5     v purrrr   0.3.4
#> #> v tibble  3.1.6     v dplyr    1.0.8
#> #> v tidyverse 1.2.0    v stringr  1.4.0
#> #> v readr   2.1.2     v forcats 0.5.1
#> #> -- Conflicts ----- tidyverse_conflicts() --
#> #> x dplyr::filter() masks stats::filter()
#> #> x dplyr::lag()   masks stats::lag()
library(knitr)
library(flextable)
#>
#> Attaching package: 'flextable'
#> The following object is masked from 'package:purrr':
#>
#>     compose
```


Capítulo 3

Análisis exploratorio bivariante

3.1. Frecuencias conjuntas, marginales y condicionadas

3.2. Datos bivariantes y multivariantes

3.2.1. Datos bivariantes

- Estudiamos dos características a la vez: X, Y
- Nos interesa la **relación** entre ellas
- Pares de valores x_i, y_i

3.2.2. Datos multivariantes

- Más de dos variables
- Estudiamos las relaciones “dos a dos” y la estructura conjunta

3.3. Tabla de frecuencias conjunta

- Frecuencia total n : número total de datos
- Frecuencia absoluta conjunta: n_{ij} número de observaciones en la clase i de X .red[y] en la clase j de Y .
- Frecuencia relativa conjunta: $f_{ij} = \frac{n_{ij}}{n}$

3.3.1. Distribución conjunta de frecuencias

- Tabla de doble entrada, valores de una variable en filas y de la otra en columnas. En el interior, las frecuencias conjuntas (absolutas, marginales o ambas)
- Si las dos son cualitativas, se llama .red[Tabla de contingencia]

3.4. Ejemplo: tabla de contingencia

- X : Analista
- Y : Tipo

3.4.1. Absolutas

```
kable(table(lab$analista, lab$tipo))
```

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

3.4.2. Relativas

```
kable(prop.table(table(lab$analista, lab$tipo)), digits = 2)
```

	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

3.5. Ejemplo: Variables continuas

- Debemos tener los datos agrupados en intervalos (clases)
- $X = \text{ph}$ (filas); $Y = \text{est}$ (columnas)

```
histo <- hist(lab$ph, plot = FALSE)
histo2 <- hist(lab$est, plot = FALSE, breaks = 4)
lab <- lab |>
  mutate(clase_ph = cut(ph, breaks = histo$breaks)) |>
  mutate(clase_est = cut(est, breaks = histo2$breaks))
levels(lab$clase_ph) <- c(levels(lab$clase_ph), "(6.4,6.45]")
```

```
kable(table(lab$clase_ph, lab$clase_est))
```

	(28,30]	(30,32]	(32,34]	(34,36]	(36,38]
(6.35,6.4]	0	0	1	0	0
(6.4,6.45]	0	0	0	0	0
(6.45,6.5]	1	2	0	0	0
(6.5,6.55]	3	50	1	0	0
(6.55,6.6]	17	129	36	2	0
(6.6,6.65]	18	195	167	20	4
(6.65,6.7]	9	160	182	18	0
(6.7,6.75]	2	53	70	4	0
(6.75,6.8]	0	9	10	2	0
(6.8,6.85]	0	4	1	0	1

3.6. Frecuencias marginales

- Si partimos de la distribución conjunta, podemos obtener la de cada una de las variables (marginal) y estudiarla como en el apartado anterior
- Basta con hacer las sumas por columnas (X) o por filas (Y):
- Frecuencias marginales absolutas de X : $n_{i \cdot} = \sum_{j=1}^{n_j} n_{ij}$
- Frecuencias marginales absolutas de Y : $n_{\cdot j} = \sum_{i=1}^{n_i} n_{ij}$
- Igual para relativas

3.7. Ejemplo frecuencias marginales

```
tabla <- table(lab$clase_ph, lab$clase_est) |> addmargins()
tabla |> kable()
```

	(28,30]	(30,32]	(32,34]	(34,36]	(36,38]	Sum
(6.35,6.4]	0	0	1	0	0	1
(6.4,6.45]	0	0	0	0	0	0
(6.45,6.5]	1	2	0	0	0	3
(6.5,6.55]	3	50	1	0	0	54
(6.55,6.6]	17	129	36	2	0	184
(6.6,6.65]	18	195	167	20	4	404
(6.65,6.7]	9	160	182	18	0	369
(6.7,6.75]	2	53	70	4	0	129
(6.75,6.8]	0	9	10	2	0	21
(6.8,6.85]	0	4	1	0	1	6
Sum	50	602	468	46	5	1171

3.8. Ejemplo distribuciones marginales

3.8.1. De X

```
tabla[, 6]
(6.35,6.4] (6.4,6.45] (6.45,6.5] (6.5,6.55] (6.55,6.6]
      1          0          3         54        184
(6.6,6.65] (6.65,6.7] (6.7,6.75] (6.75,6.8] (6.8,6.85]
     404        369        129        21         6
Sum
1171
round(tabla[, 6] / tabla[11, 6], 2)
(6.35,6.4] (6.4,6.45] (6.45,6.5] (6.5,6.55] (6.55,6.6]
      0.00       0.00       0.00       0.05       0.16
(6.6,6.65] (6.65,6.7] (6.7,6.75] (6.75,6.8] (6.8,6.85]
     0.35       0.32       0.11       0.02       0.01
Sum
1.00
```

3.8.2. De Y

```
tabla[11,]
(28,30] (30,32] (32,34] (34,36] (36,38]      Sum
      50       602      468      46       5     1171
round(tabla[11,] / tabla[11, 6], 2)
(28,30] (30,32] (32,34] (34,36] (36,38]      Sum
     0.04      0.51      0.40      0.04      0.00     1.00
```

3.9. Distribuciones condicionadas

- Distribución de una variable condicionada a algún/os valores de la otra
- Se representa por $Y|X = x_i$
- Se lee “Distribución de Y condicionada a que X es igual a x_i
- Son variables univariantes que se pueden estudiar como en el apartado anterior

3.10. Ejemplo distribuciones condicionadas

3.10.1. De $X|Y \in (30, 32]$

```
tabla[, 2]
(6.35,6.4] (6.4,6.45] (6.45,6.5] (6.5,6.55] (6.55,6.6]
```

```

      0          0          2          50         129
(6.6,6.65] (6.65,6.7] (6.7,6.75] (6.75,6.8] (6.8,6.85]
    195       160        53         9         4
Sum
602
round(tabla[, 2] / tabla[2, 6], 2)
(6.35,6.4] (6.4,6.45] (6.45,6.5] (6.5,6.55] (6.55,6.6]
    NaN       NaN        Inf        Inf        Inf
(6.6,6.65] (6.65,6.7] (6.7,6.75] (6.75,6.8] (6.8,6.85]
    Inf       Inf        Inf        Inf        Inf
Sum
Inf

```

3.10.2. De $Y|X \in (6.55, 6.6]$

```

tabla[5,]
(28,30] (30,32] (32,34] (34,36] (36,38]      Sum
    17     129     36      2      0     184
round(tabla[5,] / tabla[5, 6], 2)
(28,30] (30,32] (32,34] (34,36] (36,38]      Sum
    0.09   0.70   0.20   0.01   0.00   1.00

```

3.11. Independencia de variables

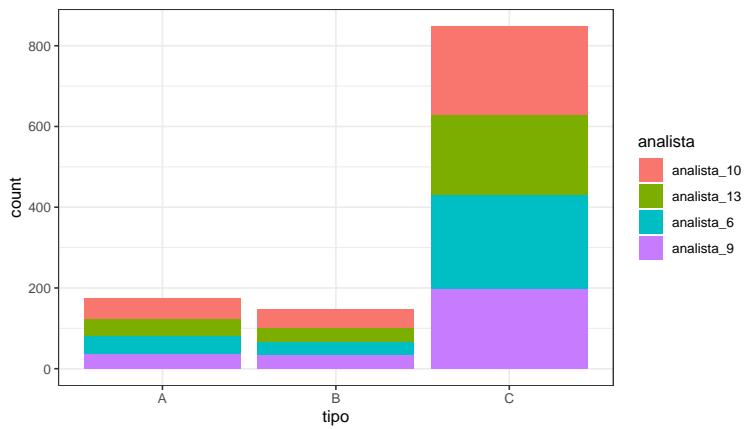
- Cuando la distribución de una variable X no varía en función de los valores de otra variable Y , entonces las variables X e Y son estadísticamente independientes.
- Equivalentemente, la distribución de $X|Y = y_j$ es igual para cualquier valor de y_j
- Si son independientes, se cumple:

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j}$$

3.12. 2.3. Representación gráfica conjunta

3.13. Gráficos de barras para frecuencias conjuntas

```
lab |>
  ggplot(aes(x = tipo, fill = analista)) +
  geom_bar() +
  theme_bw()
```

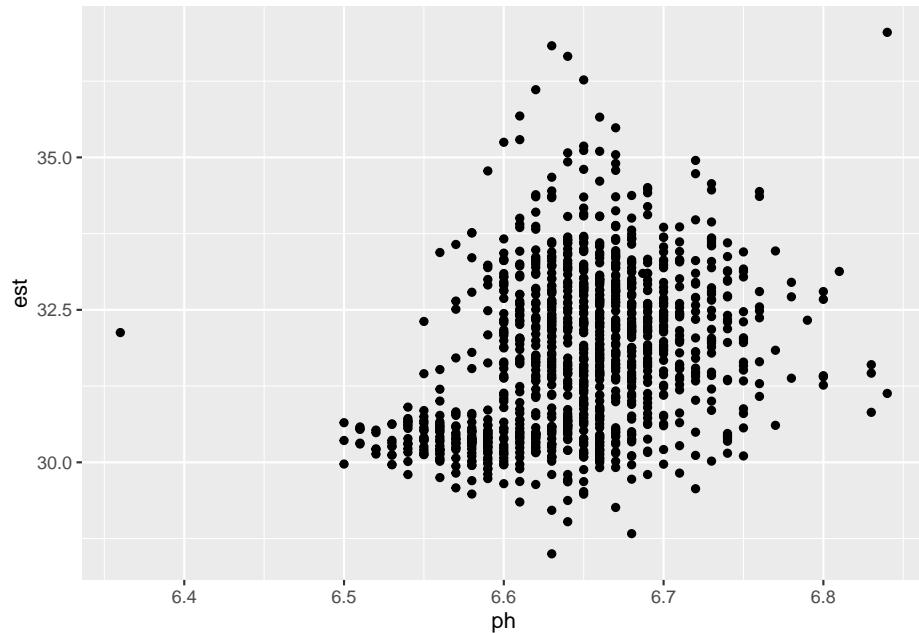


Varias variantes: apiladas, pegadas, saturadas, horizontales

3.14. El gráfico de dispersión

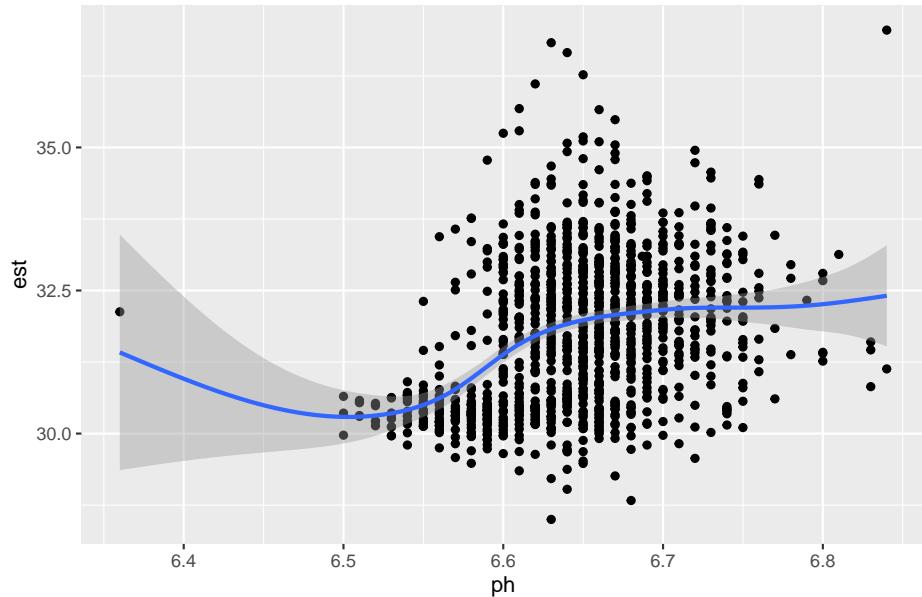
- Ejes cartesianos X , Y
- Una variable en cada eje
- Un punto en cada par (x_i, y_i)
- A simple vista se puede ver si hay relación lineal o de otro tipo

```
lab |>
  ggplot(aes(x = ph, y = est)) +
  geom_point()
```



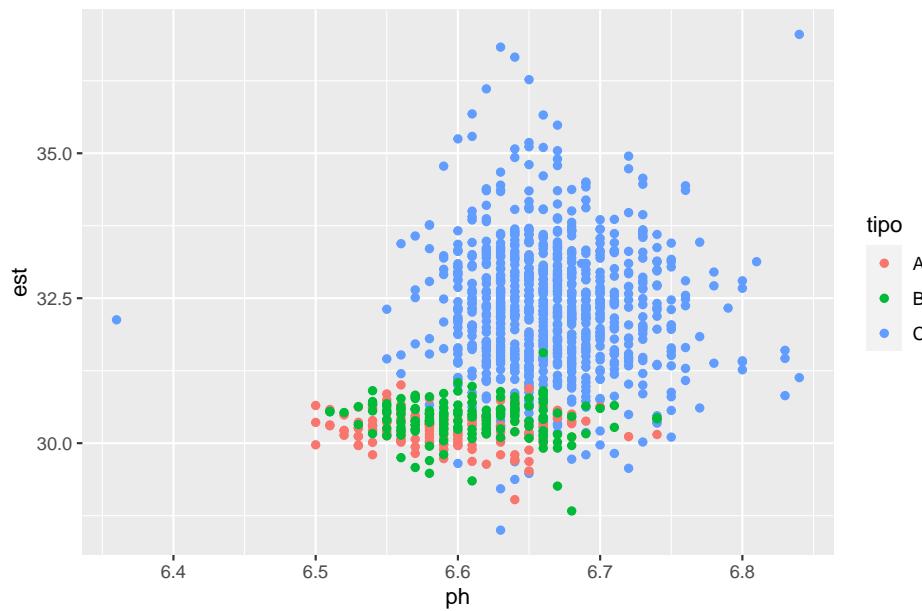
```
lab |>
  ggplot(aes(x = ph, y = est)) +
  geom_point() +
  geom_smooth() +
  labs(title = "Gráfico de dispersión con ajuste de regresión")
```

Gráfico de dispersión con ajuste de regresión



```
lab |>
  ggplot(aes(x = ph, y = est, col = tipo)) +
  geom_point() +
  labs(title = "Gráfico de dispersión identificando grupos")
```

Gráfico de dispersión identificando grupos

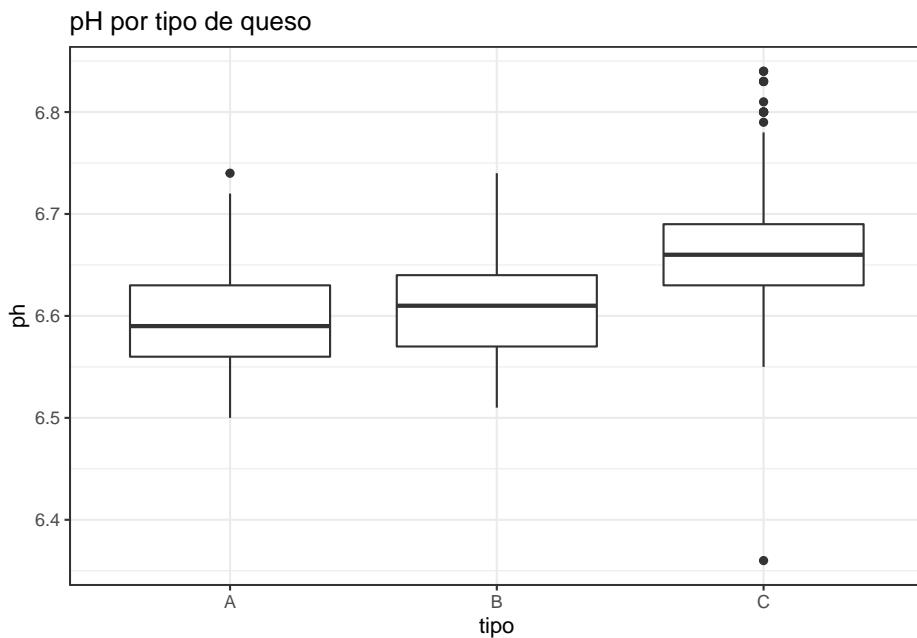


```
geom_boxplot() +  
theme_bw() +  
labs(title = "pH por tipo de queso")
```

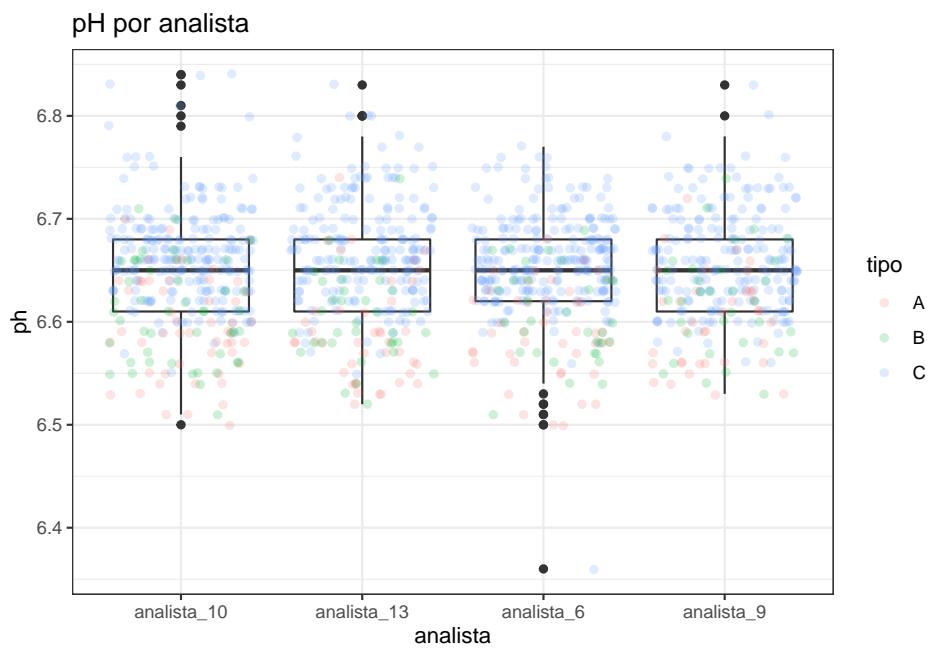
3.16. GRÁFICOS DE CAJAS POR GRUPOS

79

3.16. Gráficos de cajas por grupos



```
lab |>  
ggplot(aes(x = analista, y = ph)) +  
geom_boxplot() +  
geom_jitter(aes(col = tipo), alpha = 0.2) +  
theme_bw() +  
labs(title = "pH por analista")
```



3.16.1. Intro multivariante

Parte II

Probabilidad

Capítulo 4

Introducción a la Probabilidad

4.1. Introducción

En los capítulos anteriores, hemos visto cómo mediante la **Estadística Descriptiva** estudiamos variables estadísticas describiéndolas y representándolas. Mediante la **Estadística Inferencial** lo que tratamos es de inferir (estimar, predecir) las propiedades de una población basándonos en una muestra de datos. La Teoría de Probabilidades y el Cálculo de Probabilidades son las bases en las que se sustentan estos métodos, partiendo de la estimación del modelo de datos, es decir, la distribución de probabilidad de una determinada característica en la población. En este capítulo estudiaremos los conceptos fundamentales del **Cálculo de Probabilidades**.

Estándares de aplicación

En este capítulo se han aplicado los siguientes estándares:

- **UNE-ISO 3534-1:** Estadística. Vocabulario y símbolos. Parte 1, Términos estadísticos generales y términos empleados en el cálculo de probabilidades

Estadística y Cálculo de Probabilidades

La figura 4.1 representa la esencia de la Estadística, esto es, su relación con la probabilidad y la inferencia, a través de la población y la muestra.

Es decir, partiendo de los datos de la muestra, estimaremos el modelo de distribución de probabilidad que sigue la variable en estudio en toda la población. A partir de ahí, podremos estimar sus parámetros, calcular probabilidades y

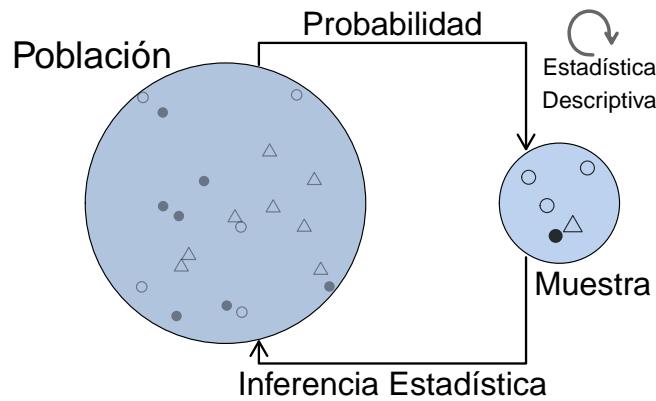


Figura 4.1: Relación entre la Estadística Descriptiva, el Cálculo de Probabilidades y la Estadística Inferencial

realizar contrastes de hipótesis usando técnicas de inferencia estadística. La Estadística Descriptiva sobre los datos de la muestra es una tarea permanente. Necesitamos en primer lugar una definición de la Probabilidad y sus propiedades.

4.2. Sucesos aleatorios

Definamos un **experimento** como cualquier actividad que deriva en un resultado observable e identificable, al que llamaremos **suceso**. Estos resultados pueden ser deterministas o aleatorios. **Sucesos deterministas** son los resultados de aquellos experimentos que, bajo las mismas condiciones, producen el mismo resultado. Por ejemplo, si observamos el número de eclipses de sol que se producen en los próximos 12 meses, el resultado es determinista. Por contra, **Sucesos aleatorios** son aquellos que están sujetos a incertidumbre. La mayoría de los experimentos no son deterministas sino **aleatorios**. Por ejemplo, el resultado al lanzar un dado, observar si un cliente compra o no al entrar a una tienda, etc.

Llamamos **sucesos elementales** a cada uno de los resultados posibles de un experimento. Al ser aleatorios, no conocemos cuál de ellos va a ser el resultado final del experimento, pero sí podemos conocer la probabilidad de que se produzca cada uno de los resultados¹. Por ejemplo: en una clase de 50 alumnos, si observamos el número de alumnos que obtiene sobresaliente en un curso, no sabemos cuántos van a ser. Pero sí podemos saber cuál es la probabilidad de cada uno de los resultados posibles, en este caso entre 0 (ninguno) y 50 (todos) en base a lo que ha sucedido en años anteriores.

Así, la **Probabilidad** es una medida del **grado de incertidumbre** sobre el

¹Muchas veces, lo que tendremos es una estimación o idea aproximada de esas probabilidades

resultado de un experimento aleatorio. Los posibles resultados de un experimento aleatorio forman un conjunto, y la teoría de probabilidades se sustenta en la teoría de conjuntos. A continuación vamos a definir formalmente los sucesos en términos de **conjuntos**.

Espacio muestral, Ω

Conjunto de todos los resultados posibles

— ISO 3534-1 2.1

Ω estará formado por los posibles resultados del experimento o sucesos elementales ω_i .

Suceso, A

Subconjunto del espacio muestral

— ISO 3534-1 2.2

Suceso complementario, A^c

Espacio muestral excluyendo el suceso dado

— ISO 3534-1 2.3

Así, un suceso cualquiera estará formado por uno o varios sucesos elementales ω_i del espacio muestral. Un suceso A ocurre si ocurre alguno de los sucesos elementales que lo componen.

4.2.1. Sucesos notables

Los siguientes sucesos tienen especial importancia en el cálculo de probabilidades:

- Suceso $A \subseteq \Omega$.
- Suceso complementario² A^c .
- Suceso seguro Ω .
- Suceso imposible \emptyset .

La figura 4.2 representa el espacio muestral, un suceso cualquiera A y su complementario A^c . El suceso imposible no aparece representado, pero en realidad sería:

$$\emptyset = \Omega^c$$

Habitualmente se utilizan ejemplos de juegos de azar para introducir el cálculo de probabilidades, como lanzamiento de monedas y dados, o combinaciones de cartas en barajas de naipes. Los ejemplos con juegos de azar tienen la ventaja de que son fáciles de comprender.

²También se suele representar por \bar{A} o A^* .

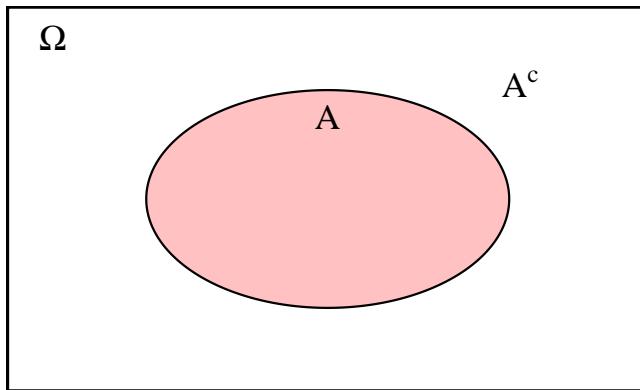


Figura 4.2: Representación del espacio muestral, un suceso cualquiera y su complementario

Lanzamiento de un dado. El experimento consiste en lanzar un dado una vez; Los sucesos elementales son los resultados del 1 al 6; El espacio muestral es el conjunto de todos los sucesos elementales, es decir, $\Omega = \{1, 2, 3, 4, 5, 6\}$; Si definimos el suceso A “que salga número par”, entonces $A = \{2, 4, 6\}$; el suceso A ocurre si sale un 2, un 4, o un 6.



La aplicación de la probabilidad en casos distintos a los juegos de azar, sigue las mismas leyes, y los ejemplos se pueden asimilar a situaciones reales de la empresa o cualquier otro ámbito. A continuación se describe 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.

En un estudio se cuenta con un conjunto de 52 sujetos, los cuales están clasificados según alguna característica. Vamos a considerar el *experimento* de observar un sujeto (por ejemplo cuando entra en la página web del estudio) y clasificarlo según un criterio determinado. Tendremos los siguientes sucesos:

- 52 posibles sujetos en estudio, (Ω)
- La mitad son mujeres (M)
- 4 investigadores (I), 12 técnicos (T), resto pacientes (P)
- 13 jóvenes (J), 26 adultos (A), 13 mayores (R); 5, 18 y 3 mujeres en cada grupo respectivamente
- 1 de cada seis hombres (H) responderá al tratamiento (S), el doble si es mujer



¿Con qué juegos de azar relacionarías cada uno de los sucesos anteriores? Piensa algunos ejemplos de sucesos en el entorno empresarial con datos similares. El siguiente puede ser un ejemplo más real.



Estudiamos una serie de proyectos de inversión y para ello queremos seleccionar dos de un total de cinco proyectos. El espacio muestral, si asumimos que no nos importa el orden en el que se seleccionan y etiquetamos los proyectos con los números del 1 al 5, es $\Omega = \{(1, 2), (1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5)\}$. Es decir, el espacio muestral tiene 10 elementos. El mero recuento se puede realizar mediante técnicas de combinatoria, véase al apéndice C.2. En este caso, $C_{5,2} = \binom{5}{2} = 10$.

CALCULADORA

5 \boxed{nCr} 2 \rightarrow 10

HOJA DE CÁLCULO

=COMBIN(5;2) $\boxed{10}$

[EXCEL] =COMBINAT(5;2) $\boxed{10}$

R

La función `choose` obtiene el número de combinaciones como se ilustra a continuación.



`choose(5, 2)`

#> [1] 10

4.2.2. Operaciones con sucesos

Como se ha comentado anteriormente, los sucesos son conjuntos. Y como tales, aplican las operaciones y propiedades de la teoría de conjuntos.

Unión de sucesos. Dados dos sucesos A y B , definimos $A \cup B$ ³ como el suceso que se cumple si:

- Ocurre A , o
- Ocurre B , o
- Ocurren A y B a la vez

El suceso unión contiene los sucesos elementales comunes y los no comunes, véase la figura 4.3.

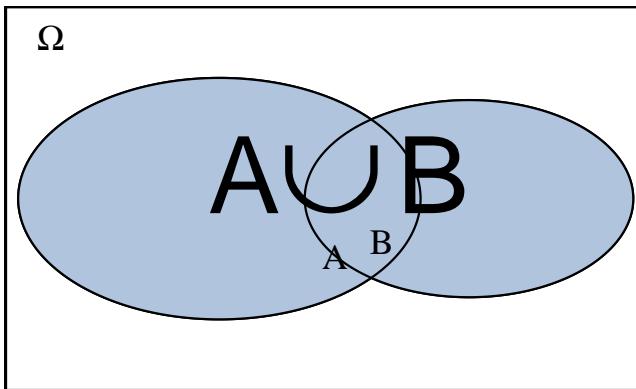


Figura 4.3: Representación de la unión de dos sucesos

El suceso “ser investigador **o** mujer” en nuestro ejemplo de los sujetos en estudio ($M \cup I$) incluirían a lo resultados elementales correspondientes con todas las mujeres (incluidas directivas) y los directivos hombres.

Intersección de sucesos. Dados dos sucesos A y B , definimos $A \cap B$ ⁴ como el suceso que se cumple si ocurren A y B simultáneamente. El suceso intersección contiene únicamente los sucesos elementales comunes a ambos sucesos, véase la figura 4.4

Las operaciones de unión e intersección entre dos sucesos se extienden inmediatamente a más de dos sucesos.

El suceso “ser hombre” **y** “ser investigador”, se corresponde con la intersección ($I \cap M^c$), e incluiría solo a los resultados del experimento en el que los potenciales usuarios hombres son directivos.

³En ocasiones se utiliza la notación $A + B$ para la unión de sucesos.

⁴En ocasiones se utiliza la notación $A \cdot B$ o simplemente AB para la intersección de sucesos.

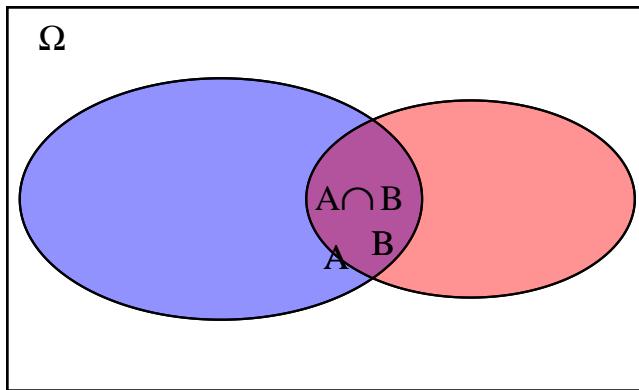


Figura 4.4: Representación de la intersección de dos sucesos

Sucesos disjuntos. Dos sucesos A y B son disjuntos o mutuamente excluyentes si:

$$A \cap B = \emptyset.$$

Un suceso A **está contenido** en otro suceso B , $A \subset B$ si siempre que se produce A , se produce también B .

Diferencia de sucesos. El suceso diferencia $A - B$ es el suceso que se produce cuando ocurre A y no ocurre B . Se verifica:

$$A - B = A \cap B^c.$$

La figura 4.5 muestra una representación de sucesos disjuntos, sucesos incluidos en otros sucesos y diferencia de sucesos.

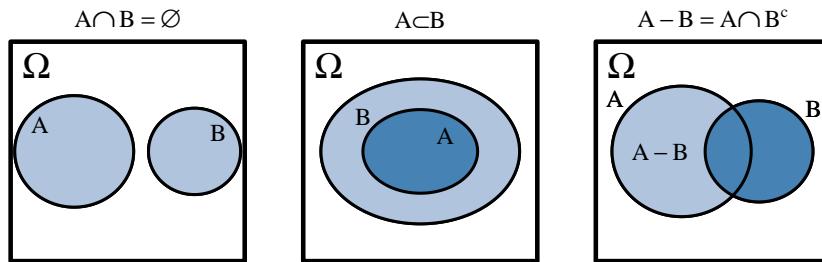


Figura 4.5: Representación de sucesos disjuntos (izquierda), suceso contenido en otro suceso (centro) y diferencia de sucesos (derecha)

El suceso “ser hombre” y el suceso “ser mujer” son sucesos disjuntos ($H \cap M = \emptyset$); El suceso “ser mujer joven” está incluido en el suceso “ser mujer”, e incluye a las mujeres jóvenes; El suceso “Ser hombre joven”, se podría representar como $J - M = J \cap M^c$.



Partición del espacio muestral. Dada una colección de sucesos A_1, A_2, \dots , decimos que es una partición del espacio muestral Ω si:

- $A_1, A_2, \dots : A_i \subset \Omega \forall i$
- $A_i \cap A_j = \emptyset \forall i \neq j$,
- $\bigcup_i A_i = \Omega$.

La figura 4.6 representa gráficamente una partición del espacio muestral Ω en cinco sucesos A_1, \dots, A_5 .

Nótese que los sucesos elementales de un experimento ω_i constituyen una partición del espacio muestral.

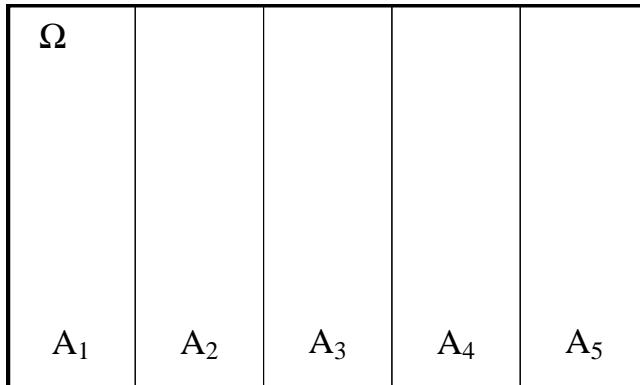


Figura 4.6: Representación de una partición del espacio muestral

De la teoría de conjuntos se deducen fácilmente las siguientes propiedades de las operaciones con sucesos:

- **Commutativa:**
 - $A \cup B = B \cup A$.
 - $A \cap B = B \cap A$.
- **Asociativa:**
 - $A \cup (B \cup C) = (A \cup B) \cup C$.
 - $A \cap (B \cap C) = (A \cap B) \cap C$.
- **Distributiva:**
 - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
 - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
- **Leyes de De Morgan:**

- $(A \cup B)^c = A^c \cap B^c$.
- $(A \cap B)^c = A^c \cup B^c$.
- $A \cup A = A \cap A = A \cup \emptyset = A \cap \Omega = A$.
- $A \cup \Omega = \Omega$.
- $A \cap \emptyset = \emptyset$.

4.2.3. Clasificación de los espacios muestrales

La primera clasificación que haremos de un espacio muestral es en función de su *tamaño*:

- **Finito:** consta de un número finito de sucesos elementales. Por ejemplo el lanzamiento de un dado: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- **Infinito numerable:** el resultado del experimento tiene (al menos teóricamente) infinitos posibles resultados, pero se pueden numerar. Por ejemplo el número de piezas correctas hasta que se produce un fallo: $\Omega = \{0, 1, 2, 3, \dots\}$.
- **Infinito no numerable:** el resultado del experimento tiene infinitos posibles resultados, que no se pueden numerar. Por ejemplo el tiempo hasta el fallo en el ejemplo anterior⁵: $\Omega = [0, \infty)$.

Otro ejemplo de espacio muestral infinito no numerable consistiría en el resultado de un experimento consistente en realizar una medición de una magnitud continua que pueda tomar cualquier valor entre, por ejemplo, 10 y 20: $\Omega = x \in \mathbb{R}, 10 \leq x \leq 20$. Una partición de este espacio muestral sería $A_1 = [10, 15]$, $A_2 = (15, 20]$.) Nótese que los números (reales, naturales, etc.) son también conjuntos, y por tanto las operaciones relacionadas con sucesos se extienden fácilmente a estos conjuntos.

Definimos una sigma álgebra de sucesos σ -álgebra o \aleph (aleph) como un conjunto de sucesos que verifican las siguientes propiedades:

- Pertenece a \aleph ,
- Si un suceso pertenece a \aleph , entonces su suceso complementario también pertenece a \aleph ,
- Si $\{A_i\}$ es un conjunto de sucesos en \aleph , entonces la unión $\bigcup_i A_i$ y la intersección $\bigcap_i A_i$ pertenecen a \aleph .

⁵Nótese que si midiéramos el tiempo, por ejemplo, en horas, sí podríamos numerar los posibles valores (0, 1, ...). Pero esto es solo debido a la precisión con la que medimos, ya que teóricamente podríamos añadir toda la precisión necesaria. Esto será importante en los siguientes capítulos cuando diferenciemos las variables aleatorias discretas y continuas.

Nótese la diferencia entre Ω y \mathfrak{N} . Mientras el espacio muestral Ω es el conjunto de todos los sucesos elementales del experimento, la σ -álgebra de sucesos \mathfrak{N} es el conjunto de todos los sucesos que podemos crear a partir del espacio muestral Ω y las operaciones de unión, intersección y complementariedad con esos sucesos. El par (Ω, \mathfrak{N}) se dice que es un **espacio probabilizable**.

Observamos al azar el tipo de participante en el estudio de uno tomando al azar. Entonces los posibles resultados del *experimento* o sucesos elementales es:

$$\Omega = \{I, T, P\}$$

Haciendo todas las operaciones posibles de unión, intersección y complementariedad, podemos llegar fácilmente a la siguiente σ -álgebra de sucesos:

$$\mathfrak{N} = \{I, T, P, (I \cup T), (I \cup P), (T \cup P), \emptyset, \Omega\}$$

4.3. Definiciones de probabilidad y sus propiedades

Ya hemos dicho anteriormente que la probabilidad es una medida del grado de incertidumbre sobre el resultado de un experimento. Ahora necesitamos formalizar la definición de probabilidad con el fin de trabajar matemáticamente con ella.

4.3.1. Definición clásica o de Laplace

La definición *clásica* de la probabilidad, también conocida como definición de Laplace⁶, requiere disponer de un espacio muestral finito referido a un experimento en el que todos los resultados posibles son igualmente probables. Bajo estas condiciones, la probabilidad de un suceso cualquiera A se obtiene como el cociente entre el número de casos *favorables* al suceso, dividido por el número total de casos *posibles* del experimento. Así:

$$P(A) = \frac{\text{casos favorables a } A}{\text{casos posibles}}.$$

Utilizaremos la definición de Laplace para asignar probabilidades a sucesos cuando tengamos una enumeración completa del espacio muestral como en los ejemplos anteriores.

⁶Pierre-Simon Laplace (1749–1827), astrónomo y matemático francés. https://es.wikipedia.org/wiki/Pierre-Simon_Laplace.

En el lanzamiento de un dado equilibrado de seis caras, la probabilidad de sacar un seis es igual al cociente entre los casos favorables a sacar un 6 (1) y los casos posibles del experimento (6):

A : Sacar un 6 en el lanzamiento de un dado

$$P(A) = \frac{\text{casos favorables a } A}{\text{casos posibles}} = \frac{1}{6} \simeq 0,1667.$$



En el ejemplo de los sujetos en estudio, la probabilidad de que un sujeto al azar sea investigador es el cociente entre los casos favorables a ser investigador (4) y los casos posibles (52):

$$P(I) = \frac{4}{52} = 0,0769$$



Casi dos siglos antes de que Laplace publicara su *Teoría Analítica de las probabilidades*, Pascal y Fermat intercambiaron correspondencia para intentar resolver los problemas que el *Caballero de Méré* le planteó al primero. Este personaje era un jugador profesional de la época que planteaba estos problemas en términos de si tenía ventaja al apostar a unos u otros resultados en el lanzamiento de dos dados. Este fue para muchos el origen de la teoría de la probabilidad. Una historia más detallada puede encontrarse en Corbalán and Sanz (2010).



4.3.2. Definición frecuentista o empírica

La definición clásica de probabilidad se encuentra con dificultades para asignar probabilidades a medida que los experimentos alcanzan cierta complejidad. Por una parte, no siempre tenemos una descripción completa del espacio muestral, o, simplemente, es infinito, con lo cual no podemos aplicar la fórmula de Laplace. Otras veces no tenemos la información disponible necesaria. Pensemos en la situación habitual descrita en la figura 2.1 al principio de este capítulo. Queremos asignar una probabilidad a un suceso referido a nuestra **población** objeto de estudio. Sin embargo, no tenemos información de los casos posibles y favorables a la ocurrencia del suceso. A lo sumo, tenemos acceso a una **muestra** de datos de la población, a la que podemos aplicar el experimento y obtener las **frecuencias** de ocurrencia de los sucesos en cuestión. Pues bien, la definición frecuentista nos dice que si observamos la frecuencia de ocurrencia del suceso A , llamémosle $n(A)$, en un número grande de experimentos n , la frecuencia relativa de ocurrencia del suceso A *tiende* a la probabilidad del suceso A . Matemáticamente:

$$P(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n}.$$

En experimentos fáciles de realizar, se puede comprobar *empíricamente*. Por ejemplo, podemos lanzar una moneda e ir anotando la frecuencia de caras con cada repetición. Este tipo de experimentos son también fáciles de realizar mediante simulación. En la siguiente aplicación se puede simular la elección de elementos de un conjunto⁷.

Relative frequency approach to Probability

Set elements (separated by commas)

Element to compute probability (one of the above)

Sample size from one to

Seed

En la práctica, utilizaremos esta definición para asignar probabilidades a sucesos en base a datos históricos, experiencia previa, etc. En muchas ocasiones, estos datos están disponibles en forma de porcentajes, y bastará con dividir por 100 para transformarlos en una frecuencia relativa, que se tomará como probabilidad.

⁷Si no estás leyendo la versión html del libro puedes ver la aplicación en el siguiente enlace:
https://elcano.shinyapps.io/probability_as_relative_frequency/

En nuestro ejemplo de los sujetos en estudio, podemos disponer de datos históricos que nos digan que 17 de 100 sujetos varones respondieron al tratamiento en un estudio similar. De ahí podemos asignar al suceso A = “el sujeto masculino responde al tratamiento” una probabilidad $P(A) = \frac{17}{100} \approx \frac{1}{6}$, equivalente a “uno de cada 6” que se decía en la descripción del ejemplo.



Históricamente, el 1 % de las piezas producidas en una fábrica tienen algún tipo de defecto. Entonces, la probabilidad de que una pieza tomada al azar tenga defecto (D) es $P(D) = \frac{1}{100} = 0,01$.



4.3.3. Definición subjetivista

En las dos definiciones anteriores de probabilidad, hemos asignado probabilidades a sucesos en base a unos determinados datos, bien de recuento de posibilidades, bien de frecuencias relativas. En ocasiones, no se dispone de absolutamente ningún dato de este tipo. Entonces las probabilidades se han de asignar de forma subjetiva, fijadas por un individuo en particular como su *grado de creencia* acerca de la ocurrencia de un suceso. El individuo fija un valor entre cero y uno en base a la evidencia de que dispone, que puede incluir juicios personales, y también interpretaciones *a priori* sobre las dos concepciones anteriores de la probabilidad, clásica y frecuentista. Por ejemplo, puede considerar la frecuencia relativa de fenómenos similares, y combinar esta información con sus conocimientos y percepciones sobre la materia de estudio.

El enfoque subjetivista tiene especial interés en fenómenos que no se prestan a repetición, así como en métodos de estadística Bayesiana, donde se fija una probabilidad *a priori* de los parámetros de la población⁸. Existen métodos específicos para asignar probabilidades subjetivas de forma racional, que quedan fuera de los objetivos de este libro, véase, por ejemplo, de Finetti (1992).

¿Cuál es la probabilidad de que me contraten en mi primera entrevista de trabajo? ¿Cuál es la probabilidad de que un proyecto de inversión determinado sea rentable? Podemos *asignar* probabilidades, pero no tenemos información previa acerca de las frecuencias relativas o casos favorables/posibles.



⁸En el enfoque *frecuentista*, que es el que sigue este libro, los parámetros de la población son fijos, aunque desconocidos.

4.3.4. Definición en ISO 3534-1

La definición estandarizada que proporciona la norma UNE-ISO 3534-1 es la siguiente para la probabilidad de un suceso A :

Probabilidad de un suceso A ; $P(A)$

Número real del intervalo cerrado $[0, 1]$ asignado a un suceso

— ISO 3534-1 2.5

Nótese que en el estándar no se entra en detalles matemáticos por el bien de la aplicabilidad en los procesos empresariales. No obstante, esta definición es en esencia compatible y congruente con el resto de definiciones de probabilidad.

4.3.5. Definición axiomática

Si bien todas las definiciones anteriores son válidas y útiles en determinados contextos, todas presentaban problemas para desarrollar una teoría de probabilidades que se pudiera aplicar a cualquier espacio probabilizable. La siguiente definición axiomática⁹ resolvió estos problemas.

Una probabilidad φ es una función:

$$\begin{aligned}\varphi : \aleph &\longrightarrow [0, 1] \\ A &\longrightarrow P(A)\end{aligned}$$

que cumple:

1. **Primer axioma:** $\forall A \in \aleph \exists P(A) \geq 0$.
2. **Segundo axioma:** $P(\Omega) = 1$.
3. **Tercer axioma:** Dada la sucesión $A_1, \dots, A_i, \dots : A_i \in \aleph \forall i, A_i \cap A_j = \emptyset \forall i \neq j$, se cumple:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

En lenguaje natural, el primer axioma indica que a cada suceso le podemos asignar un número no negativo llamado “probabilidad del suceso A ”; el segundo axioma asigna al suceso seguro una probabilidad igual a 1; el tercer axioma establece la forma de calcular probabilidades a la unión de sucesos **disjuntos** o mutuamente excluyentes, mediante la suma de sus respectivas probabilidades. Nótese que la formulación del axioma es válida para espacios muestrales infinitos (numerables y no numerables).

⁹O axiomática de *Kolmogorov*, por Andrei Nikolaevich Kolmogorov (1903–1987), matemático ruso.

A partir de estos tres axiomas, se deducen los siguientes teoremas:

1. Dados n sucesos disjuntos dos a dos $A_1, \dots, A_n : A_i \cap A_j = \emptyset \forall i \neq j$:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i).$$

2. $P(A^c) = 1 - P(A)$.
3. $P(\emptyset) = 0$.
4. Dados $A_1, A_2 : A_1 \subset A_2 \implies P(A_1) \leq P(A_2)$.
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

$$6. P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n-1} P\left(\bigcap_{i=1}^n A_i\right).$$

El primer teorema particulariza el tercer axioma a un conjunto finito de sucesos disjuntos del espacio muestral. El segundo teorema es una de las propiedades que más aplicaremos en cálculo de probabilidades, y nos indica cómo calcular la probabilidad de un suceso restándole a 1 la probabilidad de su complementario. El tercer teorema es una consecuencia del anterior y del primer axioma, por los cuales la probabilidad del suceso imposible es cero. El cuarto teorema es de vital importancia cuando trabajemos con variables aleatorias y nos viene a decir que si un suceso está contenido en otro, la probabilidad del primero no puede ser mayor que la del segundo. Los teoremas quinto y sexto nos permiten calcular probabilidades de la unión de cualesquiera conjuntos, sean o no disjuntos. Una consecuencia fundamental de las propiedades de la probabilidad es:

$$\boxed{0 \leq P(A) \leq 1}.$$

La demostración de estos teoremas se puede encontrar, entre otros, en Ugarte et al. (2015). Así mismo, se puede comprobar fácilmente cómo las definiciones clásicas y frequentistas cumplen todas estas propiedades y por lo tanto son coherentes con la definición axiomática de la probabilidad.

Lanzamiento de un dado de seis caras. Sean los siguientes sucesos:

- A_1 : “número impar”; $A_1 = \{1, 3, 5\}$.
- A_2 : “número par”; $A_2 = \{2, 4, 6\}$.
- A_3 : “número mayor que 4”; $A_3 = \{5, 6\}$.
- A_4 : “número menor o igual que 4”; $A_4 = \{1, 2, 3, 4\}$.

Podemos calcular cualquiera de estas probabilidades por la definición de Laplace, ya que los resultados elementales del experimento son equiprobables. Así:

$$P(A_1) = \frac{1}{2} = 0,5 = P(A_2); P(A_3) = \frac{2}{6} \simeq 0,3333; P(A_4) = \frac{4}{6} \simeq 0,6667.$$

Por simple enumeración de los casos posibles podemos calcular las probabilidades de los siguientes sucesos:

- $A_1 \cup A_3$: “número impar o mayor que cuatro”; $A_1 \cup A_3 = \{1, 3, 5, 6\}$; $P(A_1 \cup A_3) = \frac{4}{6} \simeq 0,6667$.
- $A_1 \cap A_3$: “número impar y mayor que cuatro”; $A_1 \cap A_3 = \{5\}$; $P(A_1 \cap A_3) = \frac{1}{6} \simeq 0,1667$.
- Y así sucesivamente para cada posible suceso A subconjunto del espacio muestral $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Ahora bien, también podemos aplicar las propiedades de la probabilidad sin necesidad de enumerar o contar todas las posibilidades.

Por ejemplo, conocidos $P(A_1)$, $P(A_3)$ Y $P(A_1 \cap A_3)$:

- $P(A_1 \cup A_3) = P(A_1) + P(A_3) - P(A_1 \cap A_3) = 0,5 + 0,3333 - 0,1667 \simeq 0,6667$,

que conduce, obviamente, al mismo resultado. A medida que aumentan la complejidad de los experimentos, con espacios muestrales más grandes, o incluso infinitos, se hace difícil o imposible trabajar con enumeraciones, y es donde hay que aplicar la definición axiomática de la probabilidad.



En nuestro ejemplo del estudio, podríamos estar interesados en el suceso “ser mujer o joven”. Este suceso se correspondería con el suceso $M \cup J$. Para calcular esta probabilidad, tendríamos en cuenta, según los datos del ejemplo, que $P(M) = \frac{1}{2} = 0,5$, $P(J) = \frac{13}{52} = 0,25$, y $P(M \cap J) = \frac{5}{52} \simeq 0,0962$. Entonces:



$$P(M \cup J) = P(M) + P(J) - P(M \cap J) = 0,5 + 0,25 - 0,0962 \simeq 0,6538.$$

En los anteriores ejemplos hemos utilizado solamente el teorema referido a la probabilidad de la unión de sucesos. El teorema de la probabilidad del suceso complementario va a ser la propiedad que más utilizaremos en cálculo de proba-

bilidades, dado que, en muchas ocasiones, es más sencillo abordar el problema desde el punto de vista del suceso complementario. Un ejemplo es la *paradoja de los cumpleaños*.

Si el día de nuestro cumpleaños asistimos a algún evento en el que haya más de 30 personas, es muy probable que nos canten el cumpleaños feliz a más de una persona. Supongamos una clase de 30 alumnos. ¿Cuál es la probabilidad de que al menos dos alumnos cumplan años el mismo, día?. Abordar el problema directamente implicaría gran cantidad de consideraciones y costosos cálculos hasta llegar a la solución, porque habría que considerar todos los casos posibles y después calcular probabilidades de uniones e intersecciones. Sin embargo, se resuelve de forma casi inmediata sin consideramos la probabilidad del suceso complementario. Es decir, si:

A : Al menos dos personas de un grupo de 30 cumplen años el mismo día,
entonces el suceso complementario es:

A^c : No hay dos personas en un grupo de 30 que cumplen años el mismo día.

Nótese cómo la probabilidad sería igual a 1 si el grupo de personas fuera de 365 personas o más^a, ya que en ese caso el suceso sería un suceso seguro. En este caso, el espacio muestral estará compuesto por el número de maneras que tendríamos de ordenar 30 fechas de nacimiento dentro de un año (día-mes), para un conjunto total de 365 días diferentes que tiene el año. Obviamente se pueden repetir las fechas, y por tanto el número total de casos posibles se corresponde con las variaciones con repetición de 365 elementos tomados de 30 en 30:

$$VR_{m,n} = m^n = 365^{30} \simeq 7,392 \cdot 10^{76}.$$

Para calcular el número de casos favorables a que nadie cumpla años el mismo día, fijamos el cumpleaños de la primera persona. Entonces la siguiente persona pueden cumplir años cualquiera de los 364 días restantes; fijados los dos primeros, la tercera persona puede cumplir años cualquiera de los 363 días restantes, y así sucesivamente. Por tanto, los casos favorables son las variaciones (sin repetición):

$$V_{m,n} = 365 \times 364 \times \dots \times (365 - 30 + 1) \simeq 2,171 \cdot 10^{76}$$

y entonces:

$$P(A) = 1 - P(A^c) = 1 - \frac{2,171 \cdot 10^{76}}{7,392 \cdot 10^{76}} \simeq 0,7063.$$

Intuitivamente nos parecería una probabilidad demasiado alta para un grupo tan pequeño de personas, por eso nos sorprendemos cuando escuchamos un *cumpleaños feliz* el día de nuestro cumpleaños en un lugar concurrido y no es para nosotros. Como vemos, no es tan difícil.



^aSi no tenemos en cuenta los años bisiestos.

Para obtener los casos favorables, si intentamos utilizar la fórmula de las variaciones utilizando los factoriales (ver apéndice C.2), la calculadora y el software pueden devolver un error, por no poder calcular el factorial de 365.

HOJA DE CÁLCULO

Disponemos en el rango A1:A30 los números del 365 (m) al 336 (m - n + 1). Entonces podemos obtener la probabilidad del ejemplo como:

=1-PRODUCTO(A1:A30)/(365^30)

MAXIMA

Maxima sí puede trabajar con números grandes, la siguiente expresión devuelve la probabilidad pedida:

1 - (factorial(365)/factorial(365-30))/365^30;

R

El siguiente código realiza los cálculos paso a paso y devuelve la probabilidad pedida. Cambiando el valor 30 por otro número de personas cualquiera, se puede ver cómo aumenta la probabilidad.



```
ncumple <- 30
cposibles <- 365^ncumple
cfavorables <- prod(365:(365 - ncumple + 1))
prob_ninguno <- cfavorables/cposibles
prob_alguno <- 1 - cfavorables/cposibles
prob_alguno
#> [1] 0.7063162
```

Una vez definida la medida de probabilidad φ con los axiomas y propiedades anteriores, llamamos **espacio de probabilidad** a la terna:

$$(\Omega, \mathcal{N}, \varphi).$$

El estándar UNE-ISO 3534-1 recoge la definición axiomática de la probabilidad de la siguiente forma:

sigma álgebra de sucesos; σ -álgebra; sigma campo; σ -campo;

\mathcal{N}

Conjunto de sucesos con las siguientes propiedades:

- a) Pertenece a \mathcal{N} ;
- b) Si un suceso pertenece a \mathcal{N} , entonces su suceso complementario también pertenece a \mathcal{N} ;
- c) Si $\{A_i\}$ es un conjunto de sucesos en \mathcal{N} , entonces la unión $\bigcup_i A_i$ y la intersección $\bigcap_i A_i$ de los sucesos pertenecen a \mathcal{N} .

— ISO 3534-1 2.69

Medida de probabilidad \wp

Función no negativa definida sobre la sigma álgebra de sucesos tal que

a) $\wp(\Omega) = 1$

donde Ω denota el espacio muestral

b) $\wp\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \wp(A_i)$

donde $\{A_i\}$ es una secuencia de pares de sucesos disjuntos

— ISO 3534-1 2.70

Espacio de probabilidad (o espacio probabilístico); $(\Omega, \mathcal{N}, \wp)$

Espacio muestral, una sigma álgebra de sucesos asociada, y una medida de probabilidad.

— ISO 3534-1 2.68

4.4. Probabilidad condicionada y sus consecuencias

4.4.1. Probabilidad condicionada

El concepto de **probabilidad condicionada** es uno de los más importantes en teoría de la probabilidad. En ocasiones, la ocurrencia o no de ciertos sucesos del espacio muestral puede estar afectada por otros sucesos del espacio muestral. Por ejemplo, desde el punto de vista de la definición de probabilidad de Laplace, en experimentos secuenciales A_1, \dots, A_n , es posible que los resultados de los sucesivos experimentos influyan en los resultados de los siguientes, y entonces hablaremos, por ejemplo, de la probabilidad del suceso A_2 condicionada a que ha ocurrido el suceso A_1 , y la calcularemos enumerando los casos favorables y los casos posibles bajo el supuesto de haber sucedido A_1 . Esta situación aparece, por ejemplo, en los problemas de urnas. Desde el punto de vista de la definición frequentista de la probabilidad, podemos considerar un experimento en el que se observen un suceso A en distintos grupos o localizaciones, siendo B el suceso que indica la pertenencia a ese determinado grupo o característica. Se pueden considerar las frecuencias relativas del suceso A sólo para aquellos experimentos en los que ha sucedido B , y llamar a estas frecuencias¹⁰ *frecuencias de A condicionadas a B* , $fr_{A|B}$. Estas frecuencias relativas las podemos calcular dividiendo el número de veces que ocurren tanto A como B (n_{AB}) entre el número total de veces que ocurre B , (n_B):

¹⁰Nótese la analogía con las frecuencias marginales utilizadas en el capítulo 3.

$$fr_{A|B} = \frac{n_{AB}}{n_B}.$$

Ahora bien, como $fr_A = \frac{n_A}{n}$, $fr_B = \frac{n_B}{n}$ y $fr_{AB} = \frac{n_{AB}}{n}$, se tiene:

$$fr_{A|B} = \frac{n \cdot fr_{AB}}{n \cdot fr_B} = \frac{fr_{AB}}{fr_B}.$$

Es decir, la frecuencia condicionada es igual a la frecuencia conjunta dividido por la frecuencia marginal del suceso condicionante. Así pues, dado que para un número grande de realizaciones del experimento, las frecuencias relativas equivalen a la probabilidad, podemos definir la probabilidad del suceso A condicionada al suceso B como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

siempre y cuando $P(B) > 0$. Se demuestra fácilmente¹¹ que esta definición de probabilidad condicionada cumple que dado un suceso $A \in \mathcal{N}$, $(\Omega, \mathcal{N}, \wp(\cdot|A))$ es un espacio de probabilidad.

La tabla 4.1 contiene las frecuencias con las que se han observado los sucesos *aprobar* y *suspender* dos elementos evaluables de una asignatura: un examen y un trabajo.

Designemos AE y SE a los sucesos “aprobar el examen” y “suspender el examen” respectivamente, y AT y ST a los sucesos “aprobar el trabajo” y “suspender” el trabajo respectivamente. La probabilidad de aprobar el examen será:

$$P(AE) = \frac{40}{100} = 0,4.$$

Si incluimos más información a modo de condición, podemos calcular por ejemplo la probabilidad de aprobar el examen condicionado a que se ha aprobado el trabajo:



$$P(AE|AT) = \frac{P(AE \cap AT)}{P(AT)} = \frac{30/100}{35/100} \simeq 0,8571.$$

¹¹Comprobando que se cumplen los tres axiomas de la definición axiomática.

Datos ejemplo probabilidad condicionada

	Trabajo aprobado	Trabajo suspenso
Examen aprobado	30	10
Examen suspenso	5	55

En nuestro ejemplo de sujetos en estudio aparece la probabilidad condicionada de la siguiente forma. Se dice que uno de cada seis hombres responde al tratamiento. Si definimos S como el suceso “responder al tratamiento”, entonces $P(S|H) = \frac{1}{6} \simeq 0,1667$. Por otra parte, si quisieramos calcular la probabilidad de que un sujeto sea mujer, condicionado a que es joven, entonces $P(M|J) = \frac{P(M \cap J)}{P(J)} = \frac{5/52}{13/52} \simeq 0,3846$.



4.4.2. Probabilidad de la intersección de sucesos

La definición de probabilidad condicionada a la que hemos llegado, nos permite calcular la probabilidad de la intersección de dos sucesos cualesquiera sin más que despejar de la fórmula. Además, tendremos dos formas de calcularla, según conozcamos $P(A|B)$ o $P(B|A)$:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A).$$

Recuerda que $A \cap B$ significa A y B , mientras que $A|B$ significa A si ocurre B .

La probabilidad condicionada aparece en los muestreos sin reemplazamiento. Se suele asociar a los problemas *de urnas*, o también a la extracción de cartas de una baraja. Por ejemplo, podemos calcular la probabilidad de sacar dos figuras seguidas de una baraja de cartas francesa, con 52 cartas en total de las cuales 12 son figuras (J, Q, K de cada uno de los cuatro palos). Entonces, si definimos A_1 como “sacar figura en la primera extracción” y A_2 como “sacar figura en la segunda extracción”, entonces lo que buscamos es la probabilidad de que ocurran los dos sucesos, $P(A_1 \cap A_2)$:



$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{12}{52} \cdot \frac{11}{51} = \frac{11}{221} \simeq 0,0498.$$

En nuestro ejemplo de los sujetos en estudio, ¿cuál es la probabilidad de que un sujeto al azar sea mujer y además responda al tratamiento?

$$P(M \cap S) = P(S|M) \cdot P(M) = \frac{2}{6} \cdot \frac{1}{2} = \frac{1}{6} \simeq 0,1667.$$

A partir de la probabilidad condicionada se llega a la **regla de la cadena** para calcular la probabilidad de la intersección de una serie de sucesos. La regla consiste en ir multiplicando cada vez la probabilidad del suceso A_i condicionada a la intersección de todos los anteriores.

$$P\left(\bigcap_{i=1}^n A_i\right) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \dots \cdot P\left(A_n|\bigcap_{i=1}^{n-1} S_i\right).$$

Por ejemplo, en una urna hay 5 bolas rojas y 3 bolas blancas. Hacemos 3 extracciones. Si en una extracción sale blanca, devolvemos la bola a la urna y metemos 2 bolas blancas adicionales. ¿Qué probabilidad hay de sacar 3 blancas seguidas?

Si definimos los sucesos A_1 , A_2 y A_3 como “sacar bola blanca en la primera, segunda y tercera extracción respectivamente”, entonces estamos buscando:

$$P(A_1 \cap A_2 \cap A_3),$$

que utilizando la regla de la cadena calcularemos como:

$$P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2).$$

En la situación inicial hay 3 de ocho bolas blancas. En el segundo experimento, si hemos sacado blanca, la devolvemos y añadimos dos más, es decir tenemos 5 de diez bolas blancas. Si la segunda vuelve a ser blanca, entonces en el tercer experimento tenemos 7 de 12 bolas blancas. Por lo tanto:

$$P(A_1 \cap A_2 \cap A_3) = \frac{3}{8} \cdot \frac{5}{10} \cdot \frac{7}{12} = \frac{7}{64} \simeq 0,1094.$$

4.4.3. Independencia de sucesos

Si bien en muchas ocasiones el conocimiento de ciertos eventos afectan a la probabilidad de ocurrencia de otros, esto no siempre tiene por qué ser así. En estos casos, diremos que dos sucesos son independientes si el conocimiento de

la ocurrencia de uno de ellos no modifica la probabilidad de aparición del otro. Por tanto, en esos casos:

$$P(A|B) = P(A) \quad \text{y} \quad P(B|A) = P(B).$$

Entonces, por la propia definición de la probabilidad condicionada, se tiene que si dos sucesos son independientes, entonces:

$$P(A \cap B) = P(A) \cdot P(B).$$

Esta fórmula, que es una definición en sí misma de independencia de sucesos, nos proporciona también un método para comprobar si dos sucesos son independientes o no conocidas las probabilidades de los mismos y la de la intersección¹².

Para más de dos sucesos, la regla de la cadena explicada más arriba se extiende inmediatamente de forma que la probabilidad de la intersección de n sucesos independientes es el producto de sus probabilidades:

$$P(A_1 \cap \dots \cap A_n) = P(A_1) \cdot \dots \cdot P(A_n).$$

Y en el caso particular de que los n sucesos sean equiprobables, tales que $P(A_i) = p \forall i$, entonces:

$$P(A_1 \cap \dots \cap A_n) = p^n.$$

El lanzamiento sucesivo de una moneda o de un dado son claros ejemplos de sucesos independientes.

En el lanzamiento de un dado dos veces seguidas (o lo que es lo mismo, en el lanzamiento de dos dados), el resultado del primero no influye en el segundo. Por tanto, la probabilidad de sacar dos seises en el lanzamiento de dos dados es:

$$P(A_1 \cap A_2) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \simeq 0,0278.$$

Nótese que podemos llegar fácilmente al mismo resultado enumerando los posibles resultados, pero con más esfuerzo. Además, en espacios muestrales más grandes se complica enormemente la enumeración.



¹²Comprobar, por ejemplo, la independencia de los sucesos “aprobar el trabajo” y “aprobar el examen” en el ejemplo anterior.

4.4.4. Probabilidad condicionada e independencia en ISO 3534-1

La norma UNE-ISO 3534-1 recoge las definiciones de probabilidad condicionada e independencia de la siguiente forma:

Probabilidad condicionada; $P(A|B)$

Probabilidad de la intersección de A y B dividida por la probabilidad de B .

— ISO 3534-1 2.6

Sucesos independientes

Par de sucesos tal que la probabilidad de la intersección de los dos sucesos es el producto de las probabilidades individuales.

— ISO 3534-1 2.4

4.4.5. Probabilidad total y fórmula de Bayes

La probabilidad condicionada nos permite calcular probabilidades de sucesos de los que tenemos información *parcial*, en el sentido de que conocemos su probabilidad *condicionada* a algún otro suceso del espacio muestral, pero queremos saber la probabilidad *total* del suceso, independientemente de aquellos sucesos. Las condiciones para que podamos calcular la probabilidad total de este suceso, llamémosle B , son:

- Disponer de una **partición** de sucesos del espacio muestral A_1, A_2, \dots, A_n tales que $A_i \cap A_j = \emptyset \forall i \neq j$ y $\bigcup_{i=1}^n A_i = \Omega$.
- Conocer las probabilidades de cada uno de esos sucesos que forman la partición, $P(A_i)$.
- Conocer las probabilidades del suceso de interés condicionadas a cada uno de los sucesos que forman la partición del espacio muestral, es decir, $P(B|A_i)$.

Entonces, según el **teorema de la probabilidad total**, se verifica que:

$$P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i).$$

En efecto, podemos ver gráficamente en la figura 4.7 que cada sumando de la fórmula de la probabilidad total se corresponde con las intersecciones del suceso de interés B con cada uno de los sucesos de la partición A_i . Como estas intersecciones son sucesos disjuntos, la probabilidad de su unión es la suma de sus probabilidades por las propiedades de la probabilidad.

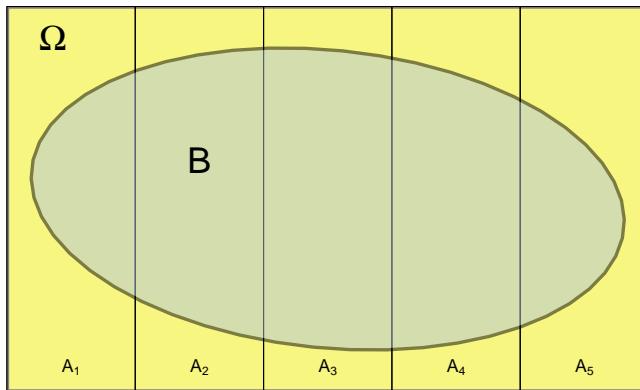


Figura 4.7: Representación del espacio muestral particionado más otro suceso

El desarrollo de la fórmula de la probabilidad condicionada a partir de la situación descrita para calcular la probabilidad total, nos permite *darle la vuelta* a la condición y encontrar probabilidades de los sucesos de la partición A_i condicionados a que se haya producido el suceso B . Partimos de la propia definición de $P(A_i|B)$:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)}.$$

Pero a su vez, la probabilidad del numerador la podemos escribir como $P(A_i \cap B) = P(B|A_i) \cdot P(A_i)$, y la probabilidad del denominador, aplicando la fórmula de la probabilidad total, es $P(B) = \sum_{i=1}^n P(B/A_i) \cdot P(A_i)$. Lo que da lugar a la fórmula de Bayes o **Teorema de Bayes**:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{i=1}^n P(B/A_i) \cdot P(A_i)},$$

siempre que $P(B > 0)$, que se puede expresar de forma simplificada como:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)}$$

En una empresa que produce componentes electrónicos tomamos 5 lotes de producto, cada uno compuesto de 50 componentes. Hay dos tipos de lotes. Los del tipo 1 (A_1) tienen 48 componentes correctos y 2 defectuosos. Los del tipo 2 (A_2) tienen 45 componentes correctos y 5 defectuosos. Tenemos 3 lotes tipo 1 y 2 lotes tipo 2. Si se toma uno de los 5 lotes al azar y se saca de éste una pieza, ¿qué probabilidad hay de que ese componente sea defectuoso?

La figura 4.8 representa la partición del espacio muestral de este ejemplo.

En este ejemplo se dan todos los elementos que habíamos descrito para calcular la probabilidad total del suceso B : “el componente es defectuoso”. Tenemos información parcial, en el sentido de que conocemos las probabilidades de ser defectuoso para cada uno de los tipos de lote, es decir $P(B|A_1) = \frac{2}{50} = 0,04$ y $P(B|A_2) = \frac{5}{50} = 0,1$. También conocemos las probabilidades de los dos sucesos que constituyen la partición, $P(A_1) = \frac{3}{5} = 0,6$ y $P(A_2) = \frac{2}{5} = 0,4$. Entonces, por el teorema de la probabilidad total:

$$P(B) = P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) = 0,04 \cdot 0,6 + 0,1 \cdot 0,4 = 0,064.$$

Supongamos ahora que se extrae del conjunto de todos los lotes un componente al azar, y resulta que es defectuoso. ¿Cuál es la probabilidad de que esa pieza provenga de un lote del tipo 1?

Nótese que ahora lo que buscamos es $P(A_1|B)$, como conocemos las $P(B|A_i)$ y $P(A_i)$, entonces podemos aplicar la fórmula de Bayes. Como ya hemos calculado antes la probabilidad total de B , podemos usar la fórmula *abreviada*:

$$P(A_1|B) = \frac{P(B|A_1) \cdot P(A_1)}{P(B)} = \frac{0,04 \cdot 0,6}{0,064} = 0,375.$$



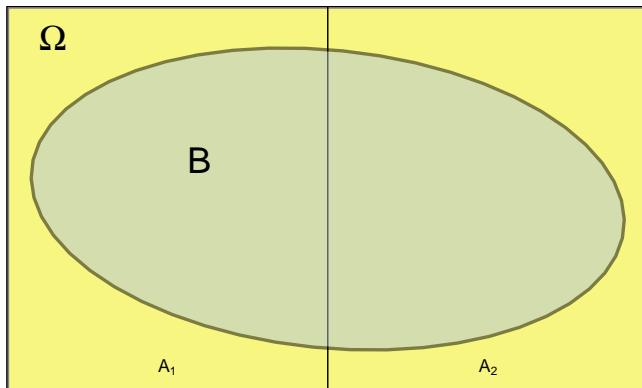


Figura 4.8: Representación del espacio muestral del ejemplo de los componentes electrónicos

En nuestro ejemplo, conocíamos las probabilidades de que un sujeto responda al tratamiento según si es hombre o mujer. También conocemos la probabilidad de que el sujeto sea hombre o mujer. Entonces podemos calcular la probabilidad de que un sujeto (independientemente de si es hombre o mujer) responda al tratamiento como:

$$P(S) = P(S|M) \cdot P(M) + P(S|H) \cdot P(H) = \frac{2}{6} \cdot \frac{1}{2} + \frac{1}{6} \cdot \frac{1}{2} = 0,25.$$

Si un sujeto responde al tratamiento, la probabilidad de que sea mujer es:

$$P(M|S) = \frac{P(S|M) \cdot P(M)}{P(S)} = \frac{\frac{2}{6} \cdot 0,5}{0,25} \simeq 0,6667.$$



El problema de Monty Hall

Monty Hall es el nombre del presentador del concurso televisivo estadounidense *Let's make a deal* que se emitió entre 1963 y 1990. En alguna de las fases del programa, el concursante tiene que elegir una entre tres puertas, dos de las cuales tienen detrás una cabra, mientras que la otra tiene un coche. Una vez elegida la puerta, el presentador muestra el contenido de una de las otras dos puertas, que contiene una cabra. Entonces el concursante tiene la opción de cambiar su puerta por la otra que queda cerrada. ¿Es más ventajoso cambiar de puerta o quedarse con la elegida inicialmente? ¿O da lo mismo?



La solución puede parecer contraintuitiva, aunque tanto desde el razonamiento a través de las frecuencias como con un desarrollo matemático se llega a la misma conclusión. Y la clave está en la **probabilidad condicionada**.



El problema de *Monty Hall* dio lugar a historias curiosas que se pueden consultar en Corbalán and Sanz (2010). Por ejemplo, el gran matemático Paul Erdős solo aceptó como buena la solución real tras comprobarla en una simulación por ordenador. Invito al lector a que concurre en la aplicación que se muestra a continuación¹³ durante un buen número de jugadas y compruebe a través de las frecuencias relativas qué estrategia ofrece mayor probabilidad de ganar el coche.

¹³accesible también en https://elcano.shinyapps.io/monty_hall

Capítulo 5

Variable aleatoria univariante

Trabajar con sucesos y todas sus combinaciones posibles puede resultar muy costoso, o incluso imposible. Con las variables aleatorias pasamos del ámbito de los sucesos a los números reales, de forma que podemos hacer cálculos numéricos. El interés de las variables aleatorias es poder modelizar la incertidumbre mediante ellas. Es importante tener en cuenta que las propiedades de las variables aleatorias son teóricas. Mediante la inferencia estadística, podremos utilizar datos empíricos de muestras para obtener conclusiones sobre la variable aleatoria que caracteriza a la población, recuérdese la figura 4.1 al principio del capítulo 4.

La figura 5.1 muestra la relación de las variables aleatorias con la población. En una muestra tenemos datos con los que calculamos estadísticos (media, varianza, etc) de esos datos. Representamos las frecuencias mediante histogramas. Por su parte, la población sigue una distribución de probabilidad teórica, con unas características teóricas (media, varianza, etc.). Ambos “mundos” se relacionan mediante la inferencia estadística, que no se trata en este texto.

5.1. Concepto y definición de variable aleatoria

Las variables aleatorias son variables numéricas cuyos valores vienen determinados por el azar. Utilizaremos letras mayúsculas X, Y, \dots para representar variables aleatorias, y letras minúsculas x, y, \dots para representar a los valores que toman. En definitiva, asignamos un número a cada posible resultado del experimento. Matemáticamente, una variable aleatoria es una función definida sobre el espacio muestral Ω perteneciente a un espacio de probabilidad $(\Omega, \mathcal{F}, \varphi)$ y que toma valores en el **conjunto** de los números reales \mathbb{R} :

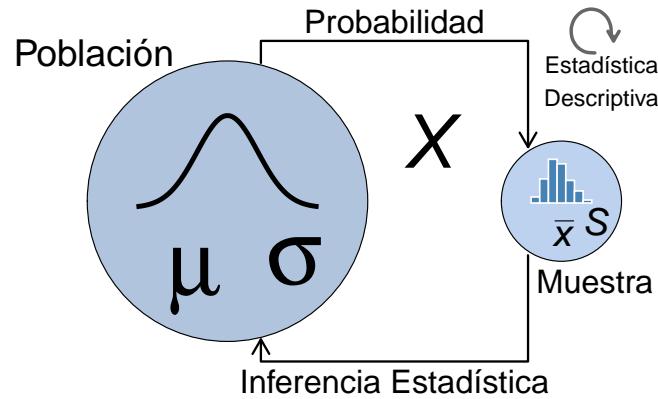


Figura 5.1: Variables aleatorias vs. datos empíricos

$$\begin{aligned} X : \Omega &\longrightarrow \mathbb{R} \\ \omega &\longrightarrow X(\omega) \end{aligned}$$

La variable aleatoria así definida cumple las siguientes características:

1. La imagen de cada elemento del espacio muestral, $X(\omega)$, es única .
2. La inversa de la variable aleatoria aplicada a cualquier intervalo de \mathbb{R} pertenece a la sigma álgebra de sucesos \mathfrak{N} .

$$M \in \mathbb{R} \implies X^{-1}(M) \in \mathfrak{N}.$$

La medida de probabilidad φ del espacio de probabilidad $(\Omega, \mathfrak{N}, \varphi)$ se aplica entonces a intervalos de los números reales en vez de a sucesos de \mathfrak{N} :

$$M \in \mathbb{R} \implies \varphi(M) = P[X \in M].$$

A esta medida de probabilidad inducida por una variable aleatoria se le suele denominar **modelo de distribución de probabilidad**.

Consideremos un experimento consistente en lanzar una moneda equilibrada al aire tres veces. El espacio muestral de este experimento aleatorio es el siguiente:

$$\Omega = \{(+, +, +), (c, +, +), (+, c, +), (+, +, c), (c, c, +), (c, +, c), (+, c, c), (c, c, c)\}$$

Definamos ahora la variable aleatoria “Número de caras” en el experimento anterior. La variable aleatoria quedaría definida como sigue:

$$\begin{array}{rcl} X : \Omega & \longrightarrow & \mathbb{R} \\ (+, +, +) & \longrightarrow & 0 \\ (c, +, +) & \longrightarrow & 1 \\ (+, c, +) & \longrightarrow & 1 \\ (+, +, c) & \longrightarrow & 1 \\ (c, c, +) & \longrightarrow & 2 \\ (c, +, c) & \longrightarrow & 2 \\ (+, c, c) & \longrightarrow & 2 \\ (c, c, c) & \longrightarrow & 3 \end{array}$$

Por tanto, el campo de variación de la variable aleatoria X o imagen de X ($Im(X)$) es:

$$Im(X) = \{0, 1, 2, 3\}$$

Ahora, basándonos en el espacio de probabilidad $(\Omega, \mathcal{N}, \varphi)$, podemos calcular probabilidades sobre cualquier subconjunto de \mathbb{R} , por ejemplo:

$$P[X = 0] = \frac{1}{8}; \quad P[X \geq 2] = \frac{1}{2}; \quad P[X > 10] = 0.$$

5.1.1. Tipos de variables aleatorias

Las variables aleatorias quedan definidas por su **campo de variación** y el **conjunto de probabilidades** que toman. El campo de variación es el recorrido de la variable aleatoria, es decir, los valores que puede tomar. El conjunto de probabilidades es el definido por la medida de probabilidad φ .

De acuerdo a la naturaleza de su campo de variación, las variables aleatorias pueden ser principalmente de dos tipos:

- Discretas: toman un conjunto de valores numerable (x_i).
- Continuas: toman un conjunto de valores no numerable (x).

También hay variables aleatorias mixtas, que no se tratan en este texto.

5.1.2. Operaciones con variables aleatorias

En general, una función de variables aleatorias es otra variable aleatoria. Sobre una o varias variables aleatorias podemos definir funciones. En particular, en los próximos apartados de este capítulo definiremos **funciones** para realizar cálculo de **probabilidades** sobre los valores que puede tomar la variable aleatoria, **transformaciones** de la variable aleatoria para calcular características de las mismas, y **combinaciones** de variables aleatorias y sus propiedades. Matemáticamente:

$$\begin{aligned} X &: \Omega \longrightarrow \mathbb{R} \\ g(X) &: \mathbb{R} \longrightarrow \mathbb{R} \\ g(X, Y) &: \mathbb{R}^2 \longrightarrow \mathbb{R} \end{aligned}$$

Los siguientes son ejemplos de funciones aplicadas a variables aleatorias:

$$X^2; \quad 1,5 \cdot X; \quad aX + b; \quad X \cdot Y; \quad \dots$$

5.1.3. Variables aleatorias y conjuntos

El paso de sucesos a variables aleatorias nos va a permitir operar con probabilidades de la misma forma que hacíamos con los sucesos. Las mismas operaciones que hacíamos con sucesos, las vamos a poder realizar con subconjuntos de los números reales, ya que, en definitiva, \mathbb{R} es un conjunto. Así, el complementario de un suceso, pasa a ser el complementario de un intervalo o conjunto de intervalos de los números reales, la unión de dos sucesos pasa a ser el conjunto de números que pertenecen a alguno de los dos subconjuntos de números reales, y la intersección de sucesos pasa a ser el conjunto de números que pertenecen a los dos subconjuntos de números reales. Algunos ejemplos:

- Complementario de un suceso: $[X \leq 1]^c = X > 1$.
- Unión de sucesos: $[10, 20] \cup (15, 30) = [10, 30)$.
- Intersección de sucesos: $[10, 20] \cap (15, 30) = (15, 20]$.

5.2. Función de distribución

Definamos una función sobre una variable aleatoria que le asigne a cada valor de X , x , la probabilidad de que la variable aleatoria X tome valores menores o iguales que dicho valor x :

$$F(x) = P[X \leq x].$$

F será por tanto una función cuyo dominio es \mathbb{R} y recorrido el intervalo $[0, 1]$:

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0, 1] \\ x &\longrightarrow F(x) \end{aligned}$$

Propiedades de la función de distribución:

- Está acotada en el intervalo $[0, 1]$: $0 \leq F(x) \leq 1$.
- Monótona no decreciente: $a < b \implies F(a) \leq F(b) \quad \forall a, b \in \mathbb{R}$.
- Continua por la derecha: $\lim_{x \rightarrow x_0^+} F(x) = F(x_0)$.
- $\lim_{x \rightarrow \infty} F(x) = 1$.
- $\lim_{x \rightarrow -\infty} F(x) = 0$.

Una consecuencia de estas propiedades es la siguiente, que nos proporciona una forma de calcular probabilidades para cualquier intervalo a partir de la función de distribución:

$$P[a < X \leq b] = F(b) - F(a).$$

Podemos comprobarlo fácilmente pensando en los números reales como conjuntos y aplicando las propiedades de la probabilidad. Efectivamente, a partir de $F(b)$, que se corresponde con la $P[X \leq b]$, y como $a < b$:

$$X \leq b = (-\infty, b] = (-\infty, a] \cup (a, b].$$

Como $(-\infty, a]$ y $(a, b]$ son conjuntos disjuntos (sucesos mutuamente excluyentes):

$$P[X \in (-\infty, b)] = P[X \in (-\infty, a)] + P[X \in (a, b)] \implies P[X \in (a, b)] = P[X \in (-\infty, b)] - P[X \in (-\infty, a)] = F(b) - F(a).$$

Además, por las propiedades de la probabilidad¹:

$$P[X > a] = 1 - F(a),$$

dado que $P[X > a] = P[(X \leq a)^c] = 1 - P[X \leq a] = 1 - F(a)$.

5.3. Variable aleatoria discreta

Son variables aleatorias discretas aquellas que pueden tomar un conjunto de valores finito o infinito numerable, x_i , $i = 1, 2, \dots, n$ o $i = 1, 2, \dots, \infty$. Formalmente, son aquellas cuya **función de distribución no es continua**. Esta discontinuidad es de salto finito, y los saltos se producen en los valores que toma la variable, x_i . A cada posible valor x_i se le asigna una probabilidad $p(x_i) = P[X = x_i]$. Los saltos son de longitud igual a $p(x_i)$.

¹O visto de otro modo, haciendo $b = \infty$, y entonces $F(b) = 1$.

5.3.1. Función de probabilidad

Dado que la variable aleatoria X no toma valores entre x_{i-1} , y x_i , podemos definir la **función de probabilidad de una variable aleatoria discreta** como:

$$\begin{aligned} p : \mathbb{R} &\longrightarrow [0, 1] \\ X &\longrightarrow p(x_i) \end{aligned}$$

$$p(x_i) = P[X = x_i] = P[x_{i-1} < X \leq x_i] = F(x_i) - F(x_{i-1}).$$

Nótese que la expresión anterior demuestra la magnitud de los saltos en las discontinuidades de la función de distribución de una variable aleatoria discreta. La función de probabilidad de una variable aleatoria discreta también se puede llamar función de cuantía o función de masa de probabilidad. Se puede encontrar también la notación abreviada p_i para referirse a $p(x_i)$.

Para que una función $p(x_i)$ sea función de probabilidad debe cumplir las siguientes condiciones:

- $p(x_i) \geq 0 \forall i.$
- $\sum_{i=1}^{\infty} p(x_i) = 1.$

A partir de la función de probabilidad podemos llegar a la función de distribución de cualquier variable aleatoria discreta como sigue:

$$F(x_i) = \sum_{j=1}^i p(x_j),$$

esto es, *acumulando* la probabilidad de los valores iguales o inferiores a cada valor x_i .

En el experimento descrito anteriormente de lanzar tres monedas, definímos la variable aleatoria:

X : Número total de caras.

La función de probabilidad de esta variable aleatoria la podemos calcular por la definición de Laplace contando los casos favorables de Ω para cada valor de la variable aleatoria X , y sería la siguiente:

$$\begin{aligned} p : \mathbb{R} &\longrightarrow [0, 1] \\ X &\longrightarrow p(x_i) = P[X = x_i] \\ 0 &\longrightarrow P[X = 0] = \frac{1}{8} \\ 1 &\longrightarrow P[X = 1] = \frac{3}{8} \\ 2 &\longrightarrow P[X = 2] = \frac{3}{8} \\ 3 &\longrightarrow P[X = 3] = \frac{1}{8} \end{aligned}$$

La figura 5.2 representa gráficamente la función de probabilidad^a.

A partir de la función de probabilidad, aplicando que $F(x_i) = \sum_{j=1}^i p(x_j)$, la función de distribución sería la siguiente:

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0, 1] \\ X &\longrightarrow F(x_i) = P[X \leq x_i] \\ 0 &\longrightarrow P[X \leq 0] = \frac{1}{8} \\ 1 &\longrightarrow P[X \leq 1] = \frac{4}{8} = \frac{1}{2} \\ 2 &\longrightarrow P[X \leq 2] = \frac{7}{8} \\ 3 &\longrightarrow P[X \leq 3] = \frac{8}{8} = 1 \end{aligned}$$

La figura 5.3 representa gráficamente la función de distribución, que se puede expresar de la siguiente forma:

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{1}{8} & \text{si } 0 \leq x < 1 \\ \frac{1}{2} & \text{si } 1 \leq x < 2 \\ \frac{7}{8} & \text{si } 2 \leq x < 3 \\ 1 & \text{si } x \geq 3 \end{cases}$$

^aNormalmente se representan líneas verticales en vez de únicamente el punto para una mejor visualización.



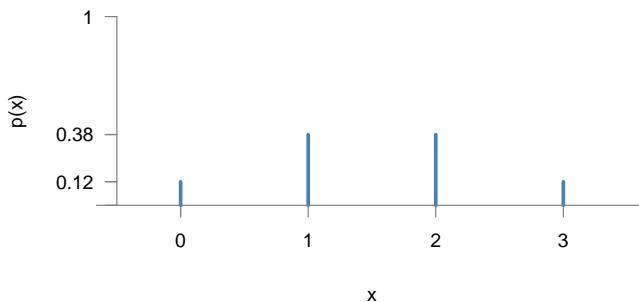


Figura 5.2: Representación de la función de probabilidad para el experimento de las monedas

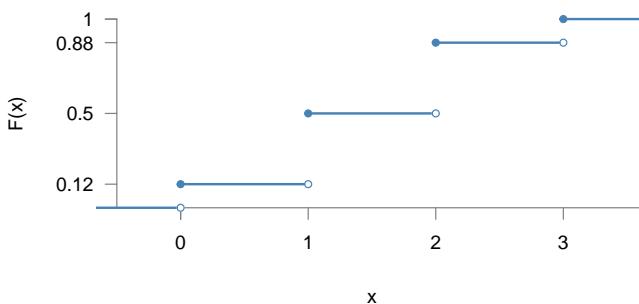


Figura 5.3: Representación de la función de distribución para el ejemplo de las monedas

Otros ejemplos de variables aleatorias discretas serían aquellos en los que realizamos recuentos u observamos características, por ejemplo:

- Número de defectos por m^2 en una superficie.
- Indicador de pieza correcta/incorrecta.
- Puntuación en una escala de valoración (por ejemplo *Likert*).
- Número de clientes que llegan a un banco cada hora.
- Número de unidades defectuosas en un lote de productos.

Funciones de probabilidad y distribución para la variable discreta del ejemplo ilustrativo

x_i	$p(x_i)$	$F(x_i)$
20	0,6923	0,6923
36	0,2308	0,9231
60	0,0769	1,0000

Vamos a ampliar el ejemplo ilustrativo de los sujetos en estudio descrito en el capítulo de introducción a la probabilidad. Supongamos que se envía a los sujetos una serie de mensajes de seguimiento por correo electrónico de forma *aleatoria*, y que el número de mensajes recibidas por un cliente puede ser 20, 36, o 60, con probabilidades $\frac{36}{52}$, $\frac{12}{52}$ y $\frac{4}{52}$ respectivamente. Podemos así definir la variable aleatoria X : *Número de mensajes remitidos por correo electrónico en un año a los sujetos en estudio*, cuyo campo de variación es $\{20, 36, 60\}$ y cuyas funciones de probabilidad y distribución se representan respectivamente en las figuras 5.4 y 5.5. La tabla 5.1 muestra ambas funciones numéricamente.

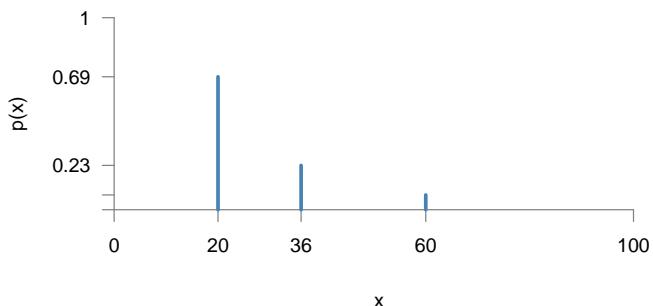


Figura 5.4: Representación de la función de probabilidad de una variable aleatoria discreta

5.4. Variable aleatoria continua

Son variables aleatorias continuas aquellas que pueden tomar un **número infinito no numerable de valores**. Formalmente, son aquellas cuya función de distribución, $F(x)$, es continua y derivable en todo su dominio. Al ser la variable continua, puede tomar cualquier valor en un intervalo de su dominio. Podemos utilizar las propiedades de la función de distribución para calcular probabilidades en cualquier intervalo de la siguiente forma:

$$P[a < X \leq b] = F(b) - F(a). \quad (5.1)$$

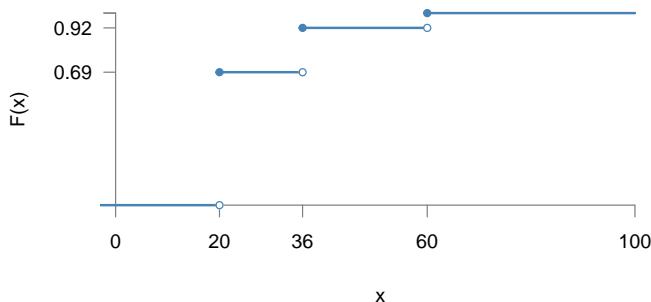


Figura 5.5: Representación de la función de distribución de la variable aleatoria discreta del ejemplo ilustrativo

Ahora bien, por ser F continua, entre a y b siempre hay masa de probabilidad, y no podemos obtener una función de probabilidad como en el caso discreto. Esto es porque no existe un valor anterior a uno dado x . Podemos *acercar* los dos extremos del intervalo tanto como queramos, por ejemplo en la ecuación (5.1) para calcular $P[X = b]$ podríamos buscar el valor anterior a b aproximando a a b , pero siempre habría un valor más allá, y finalmente la **probabilidad para un valor concreto de la variable aleatoria es igual a cero**.

$$P[X = x] = 0 \quad \forall x \in \mathbb{R}.$$

Intuitivamente, podemos entender esta característica pensando en la definición de probabilidad de Laplace. Si tuviéramos que *contar* el número de casos favorables para que la variable aleatoria tome un valor concreto, sería 1. Pero los casos posibles, por ser variable continua, son infinitos, y por tanto la probabilidad sería $\frac{1}{\infty}$.



Es importante señalar que, en la práctica, el número de valores de una variable aleatoria que podamos *medir* será finito, pero la variable aleatoria seguirá siendo continua conceptualmente, y la aplicación de sus propiedades nos permitirá resolver aquellos problemas prácticos, aunque el aparato de medida utilizado no nos permita ir más allá de cierta precisión.

Piensa en tu marca de cerveza favorita (o cualquier otra bebida), por ejemplo en el formato de 33 cl (tercio). Cuando la pides en un bar, ¿cuál crees que es la probabilidad de que la botella tenga **exactamente** 33 cl?

En realidad, si medimos con una precisión, por ejemplo, de un decimal, podemos obtener mediciones de 33,0 cl. Pero las mediciones están sujetas a un error, y en realidad lo que nos está diciendo esa medición es que el volumen está entre 32,95 y 33,05, intervalo del cual sí podemos calcular su probabilidad.



Entonces nos surge la siguiente pregunta: si no podemos calcular la probabilidad de los *sucesos* individuales, ¿cómo saber qué valores son más probables? ¿dónde se concentra la probabilidad? Precisamente la continuidad de la función de distribución nos proporciona la herramienta matemática para resolver estas cuestiones. La figura 5.6 muestra la representación gráfica de la función de densidad $F(x)$ de una determinada variable aleatoria X .

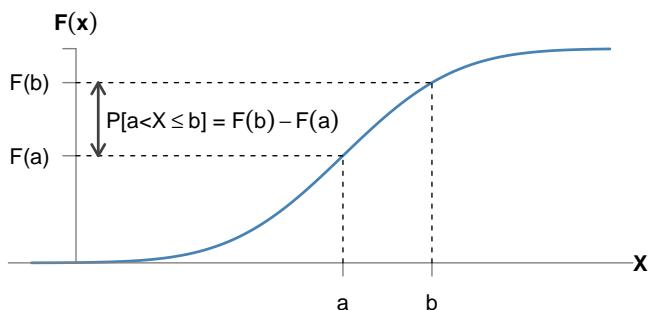


Figura 5.6: Función de distribución de una variable continua y probabilidad de un intervalo

Como podemos ver, la probabilidad en un intervalo cualquiera, es el **cambio** que se produce en la función de distribución entre los extremos del intervalo. Si acercamos los extremos del intervalo, es decir, hacemos que b tienda a a , obtenemos la tasa **instantánea** de cambio en un punto, que representa la masa de probabilidad en ese punto, y que es la derivada de la función de distribución en ese punto:

$$\lim_{b \rightarrow a} \frac{F(b) - F(a)}{b - a}.$$

La derivada de una función en un punto se corresponde con la pendiente de la recta tangente a la función en ese punto. En nuestro caso, la pendiente de la recta tangente a la función de distribución es la que nos proporciona la “densidad” de probabilidad. Se muestra a continuación una aplicación interactiva para visualizar el concepto de derivada como pendiente de la recta tangente^a, que se puede aplicar a cualquier función de distribución.



^aaccesible también en <https://elcano.shinyapps.io/derivada/>

Derivada como pendiente de la recta tangente

Valor en el eje de ordenadas

-10 10 50

Función

$x^3 - 2x^2 + 5$

Dominio (límite inferior)

Dominio (límite superior)

Recorrido (límite inferior)

Recorrido (límite superior)

5.4.1. Función de densidad

Si pensamos en la probabilidad como la tasa de cambio en la función de distribución, entonces podemos definir la densidad de probabilidad como la derivada de la función de distribución, y calcular así probabilidades con esa función, llamada **función de densidad** que se representa por $f(x)$:

$$f(x) = \frac{dF(x)}{dx}.$$

Además, por el teorema fundamental del cálculo, podemos obtener la función de distribución a partir de la función de densidad mediante la integral:

$$F(x) = \int_{-\infty}^x f(t)dt = P[X \leq x].$$

La figura 5.7 representa la relación entre la función de densidad y la función de distribución. Como se puede apreciar, el área debajo de la curva de la función de densidad $f(t)$ se corresponde con las distintas probabilidades de que la variable aleatoria tome valores en los intervalos que encierran dicha área.

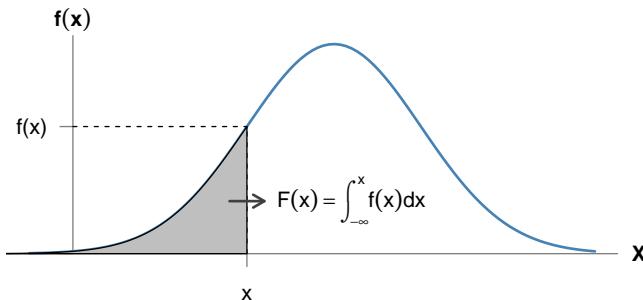


Figura 5.7: Relación entre las funciones de densidad y de probabilidad

Para que una función $f(x)$ sea función de densidad, tiene que cumplir las siguientes condiciones:

1. $f(x) \geq 0$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

La primera condición impone la condición evidente de que la masa de probabilidad sea positiva. La segunda condición la impone el segundo axioma de la probabilidad y las propiedades de la función de distribución, ya que:

$$\int_{-\infty}^{\infty} f(x)dx = F(\infty) = P[X \leq \infty] = P(\Omega) = 1.$$

Esto implica que cualquier función $g(x)$ definida positiva en un determinado intervalo, se puede convertir en una función de densidad multiplicándola por una constante k calculada como:

$$k = \frac{1}{\int_{-\infty}^{\infty} g(x)dx}.$$

Supóngase que la empresa de servicios de nuestro ejemplo quiere hacer una campaña para aplicar entre un 5 % y un 25 % de descuento a sus clientes de forma aleatoria y lineal de forma que haya más descuentos bajos que altos. Entonces la función de densidad para la variable aleatoria $X = \text{Descuento aplicado a un cliente}$ se puede modelizar mediante una recta con esta forma:

$$f(x) = \begin{cases} k(25 - x) & \text{si } 5 \leq x \leq 25 \\ 0 & \text{resto} \end{cases}$$

que, para que sea función de densidad, debe cumplir que:

$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

y por tanto:

$$k = \frac{1}{\int_5^{25} (25 - x)dx} = 0,005.$$



La figura 5.8 representa esta función de densidad.

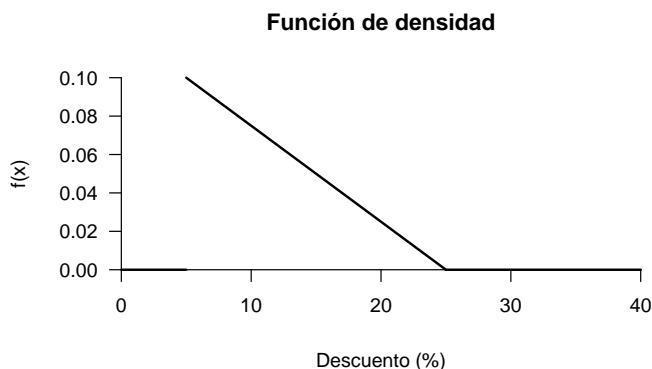


Figura 5.8: Función de densidad del ejemplo de los descuentos

Aunque las integrales que se presentan en este texto son inmediatas, en la práctica el uso del software para resolverlas es más rápido y productivo. Se aconseja al lector realizar las comprobaciones por sí mismo.

MAXIMA

Maxima resuelve integrales de forma simbólica. En el caso de una integral definida obtenemos directamente el área bajo la curva de una función. En el ejemplo:

```
1 / integrate(25-x, x, 5, 25);
```

R

R puede calcular integrales definidas mediante métodos numéricos. El código a continuación muestra la expresión para calcular la integral buscada y el valor de k .



```
integral <- integrate(f = function(x) { 25 - x },
                      lower = 5,
                      upper = 25)
integral
#> 200 with absolute error < 2.2e-12
k <- 1/integral$value
k
#> [1] 0.005
```

Sea la variable aleatoria X con función de densidad:

$$f(x) = \begin{cases} \frac{1}{8}x & \text{si } 0 < x < 4 \\ 0 & \text{resto} \end{cases}$$

Comprobar que es función de densidad, y obtener la función de distribución.

Para comprobar si es función de densidad, verificamos las dos condiciones:

1. $f(x) \geq 0$ según está definida (véase la figura 5.9)
2. Integral en todo \mathbb{R} :

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^4 \frac{1}{8}x dx = \left[\frac{x^2}{16} \right]_0^4 = 1.$$

Para calcular la función de distribución, tenemos en cuenta que:

$$F(x) = \int_{-\infty}^x f(t)dt = P[X \leq x].$$

Como lo que queremos obtener es una **función**, y no un número, el límite superior de la integral definida es variable (x). la función f la ponemos en función de t simplemente para no utilizar el mismo símbolo x y evitar confusiones.

Si tenemos una función de densidad definida por trozos, tendremos que ir *acumulando trozos*. Recorremos de menor a mayor los intervalos de \mathbb{R} realizando la integral definida completa para los intervalos anteriores al que estamos considerando. Entonces, para nuestra función:

- Si $x \leq 0$:

$$F(x) = \int_{-\infty}^x 0 dt = 0.$$

- Si $0 < x < 4$:

$$F(x) = \int_{-\infty}^0 0 dt + \int_0^x \frac{1}{8}t dt = \left[\frac{t^2}{16} \right]_0^x = \frac{x^2}{16}.$$

- Si $x > 4$:

$$F(x) = \int_{-\infty}^0 0 dt + \int_0^4 \frac{1}{8}t dt + \int_4^x 0 dt = \left[\frac{t^2}{16} \right]_0^4 = 1.$$

Expresamos por tanto la función de distribución, cuya representación aparece en la figura 5.10, de la siguiente forma para todos sus intervalos:

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \frac{x^2}{16} & \text{si } 0 < x < 4 \\ 1 & \text{si } x \geq 4 \end{cases}$$



MAXIMA

La integral definida para comprobar que vale uno sería:

```
integrate((1/8)*x, x, 0, 4);
```

Podríamos obtener la expresión de la función de distribución en el intervalo en que está definida con la siguiente expresión:

```
integrate((1/8)*t, t, 0, x);
```

R

El código a continuación realiza la comprobación de que la integral vale 1. R no puede hacer cálculo simbólico para obtener una expresión de la función de distribución. No obstante, se puede crear una función que obtenga valores de la función de distribución para utilizar posteriormente, o representarla gráficamente.



```
integrate(f = function(x) { (1/8)*x },
          lower = 0,
          upper = 4)
#> 1 with absolute error < 1.1e-14
Fx <- function(x) {
  integrate(f = function(t) { (1/8)*t },
            lower = 0,
            upper = x)
}
Fx(2)
#> 0.25 with absolute error < 2.8e-15
```

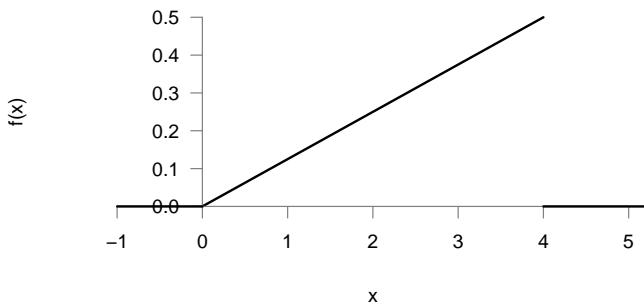


Figura 5.9: Representación de la función de densidad del ejemplo

Nótese que la función de densidad **no es una probabilidad**, y, por tanto, podría tomar valores mayores que 1. Por otra parte, la función de densidad puede ser discontinua.

Es fácil comprobar que:

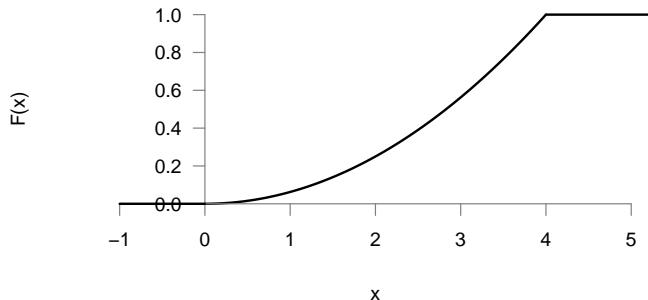


Figura 5.10: Representación de la función de distribución del ejemplo

$$P[a < X \leq b] = \int_a^b f(x)dx.$$

Lo que nos proporciona una forma de calcular probabilidades de una variable aleatoria continua mediante la función de densidad (aunque no conocemos la función de distribución). Las probabilidades son, por lo tanto, equivalentes al **área bajo la curva** de la función de densidad, que, esta vez sí, tiene que ser menor o igual que 1. Utilizando las propiedades de la probabilidad, podemos calcular probabilidades de cualquier intervalo utilizando tanto la función de densidad como la función de distribución, tal y como se resume en la figura 5.11.

Para mejorar la comprensión de la función de densidad, cuya importancia es vital en el cálculo de probabilidades, vamos a relacionarla con otros conceptos ya conocidos por el lector. En primer lugar, en el tránsito de variables discretas a continuas, hemos pasado del sencillo cálculo del sumatorio (\sum) al intimidante cálculo con integrales (\int). Sin embargo, una integral es en realidad una *suma infinita* de áreas bajo la curva cuando tomamos intervalos cada vez más pequeños. En segundo lugar, recordemos la definición de probabilidad como **frecuencia relativa** en el límite. Entonces decíamos, que si pudiéramos repetir un experimento un número grande de veces, la frecuencia relativa de ocurrencia de un suceso tenía a la probabilidad de ese suceso. En el marco de las variables aleatorias, tendríamos un número grande de realizaciones de la variable aleatoria, es decir, de números reales. Como sabemos por la estadística descriptiva, estos valores los podemos agrupar en intervalos y contar las frecuencias de los valores dentro de cada intervalo, representándolos en un **histograma**. Pues bien, si tenemos muchos números, y hacemos la amplitud de los intervalos muy pequeños, entonces el histograma de los datos se parece cada vez más a la función de densidad de la variable aleatoria que describe el experimento², recuérdese la

²Se parecerá más cuantos más datos tengamos, pero téngase en cuenta que la forma del histograma, con datos empíricos, será *aproximada* (nunca exacta) a la forma de la función, *teórica*.

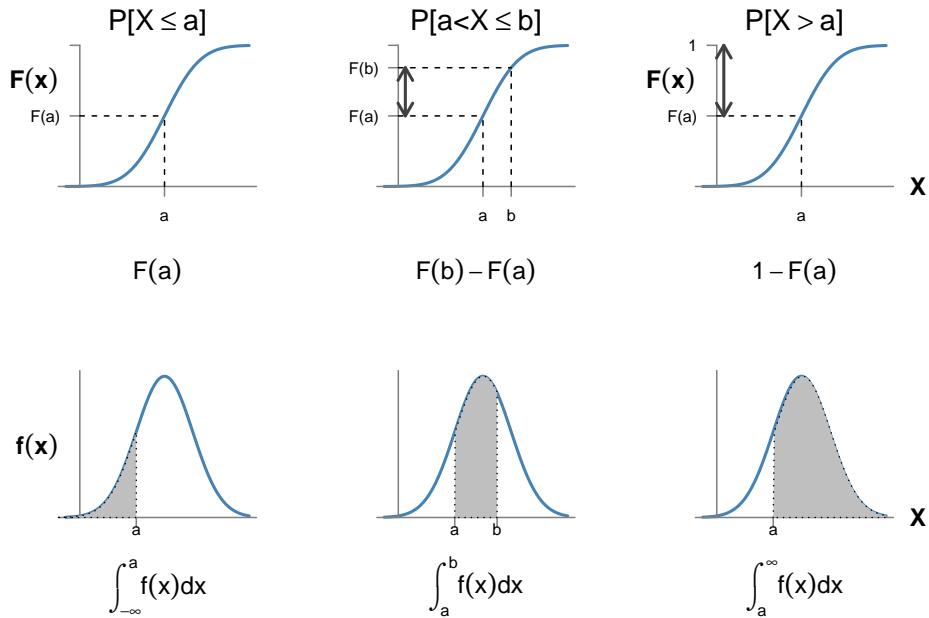


Figura 5.11: Cálculo de probabilidades de una variable continua

Figura 5.1. Además, el área de las barras del histograma representa también las probabilidades de los intervalos que podamos formar. La figura 5.12 muestra esta relación entre frecuencias y función de densidad en un determinado experimento.

Una última consideración en cuanto a las variables aleatorias continuas es la irrelevancia práctica de incluir o no el símbolo igual en las desigualdades. Si bien en las variables aleatorias discretas sí habrá una diferencia numérica que puede ser importante en aplicaciones prácticas, en las variables aleatorias continuas la utilización del símbolo \leq o el símbolo $<$, o sus contrarios \geq y $>$ es irrelevante para el cálculo. Efectivamente, como la probabilidad en un punto, $P[X = x] = 0$, entonces se cumple para variables continuas que:

$$P[X \leq x] = P[X < x]; \quad P[X \geq x] = P[X > x].$$

Pero mucho cuidado porque esto **no pasa con las variables aleatorias discretas**. Además, siempre es preferible utilizar los símbolos de forma adecuada aunque no tenga consecuencias prácticas.

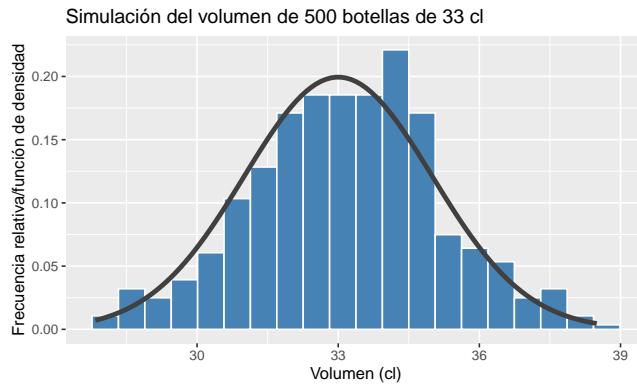


Figura 5.12: Frecuencias, histograma y función de densidad

Sea la variable aleatoria del ejemplo anterior, con las siguientes funciones de densidad y de distribución:

$$f(x) = \begin{cases} \frac{1}{8}x & \text{si } 0 < x < 4 \\ 0 & \text{resto} \end{cases}; F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \frac{x^2}{16} & \text{si } 0 < x < 4 \\ 1 & \text{si } x \geq 4 \end{cases}$$

Calcular:

$$P[1 < X < 2].$$

Lo podemos hacer a través de la función de densidad:

$$P[1 < X < 2] = \int_1^2 f(x)dx = \int_1^2 \frac{1}{8}x dx = \left[\frac{x^2}{16} \right]_1^2 = \frac{2}{8} - \frac{1}{16} = \frac{3}{16},$$

y también a través de la función de distribución:

$P[1 < X < 2] = F(2) - F(1) = \frac{4}{16} - \frac{1}{16} = \frac{3}{16}.$

MAXIMA

La probabilidad pedida se calcularía simplemente:

`integrate((1/8)*x, x, 1, 2);`

R

El código a continuación calcula la probabilidad pedida.

```
integrate(f = function(x) { (1/8)*x },
           lower = 1,
           upper = 2)
#> 0.1875 with absolute error < 2.1e-15
```

El tiempo de duración (en minutos) de la visita de un potencial usuario de un servicio tras seguir el link de una oferta es una variable aleatoria X que sigue una distribución de probabilidad según la siguiente función de densidad:

$$f(x) = \begin{cases} 2e^{-2x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

La representación gráfica de esta función aparece en la figura 5.13. Podemos comprobar que es una función de densidad verificando que cumple los dos requisitos. Es una función exponencial multiplicada por un número positivo, por tanto es siempre positiva. Comprobemos el área debajo de la curva para todo su dominio:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \iff \int_0^{\infty} 2e^{-2x}dx = [-e^{-2x}]_0^{\infty} = 1$$

La función de distribución de esta variable aleatoria será:

$$F(x) = \int_{-\infty}^x f(t)dt = \int_0^x 2e^{-2t}dt = [-e^{-2t}]_0^x = 1 - e^{-2x},$$

y su representación gráfica es la que se muestra en la figura 5.14. ¿Qué porcentaje de visitantes abandonarán probablemente la página antes de 10 segundos? (nótese que 10 segundos = 10/60 minutos). Dado que tenemos la función de distribución, es más sencillo obtenerlo a través de esta que resolviendo la integral:

$$P[X < 10/60] = F(10/60) = 1 - e^{-2 \cdot 10/60} = 0,2835.$$

Como la pregunta se hace en términos de porcentaje, la respuesta sería aproximadamente un 28.35 % de los visitantes.



MAXIMA

Las siguientes expresiones obtienen en Maxima los resultados del ejemplo.

```
integrate(2*exp(-2*x), x, 0, inf);
integrate(2*exp(-2*t),t, 0, x);
integrate(2*exp(-2*x), x, 0, 10/60);
```

R

En el siguiente código de R se realizan los cálculos explicados en el ejemplo.



```
integrate(function(x) 2*exp(-2*x), 0, Inf)
#> 1 with absolute error < 5e-07
integrate(function(x) 2*exp(-2*x), 0, 10/60)
#> 0.2834687 with absolute error < 3.1e-15
```

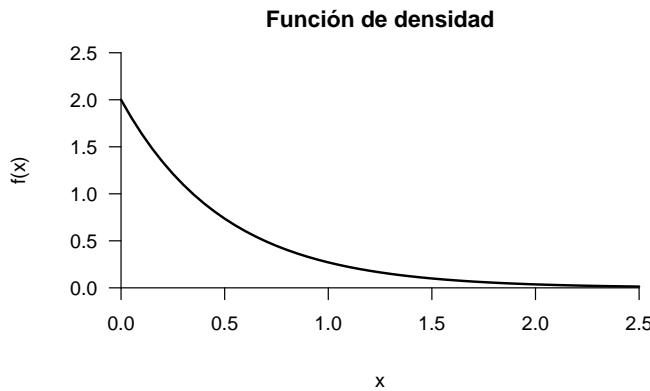


Figura 5.13: Representación de la función de densidad del ejemplo ilustrativo

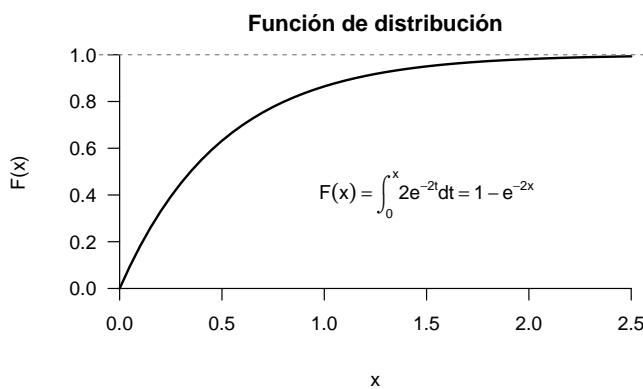


Figura 5.14: Representación de la función de distribución del ejemplo ilustrativo

5.5. Características de una variable aleatoria

Al igual que con los datos concretos de una muestra podemos calcular estadísticos que resumen la información, las variables aleatorias constan de **parámetros** de centralización, posición y forma que caracterizan la variable aleatoria a través de su distribución de probabilidad. A través de los posibles valores de una variable aleatoria y sus probabilidades podemos definir estas características. Las más importantes son la esperanza (media) y la varianza. Una vez más, téngase en cuenta que estos parámetros de la variable aleatoria son **valores teóricos de la variable aleatoria**, generalmente referidas a una población de la cual tenemos sólo información parcial a través de una muestra (recuérdese la figura 5.1).

5.5.1. Esperanza Matemática

La Esperanza matemática se define sobre una función $g(x)$ de una variable aleatoria X como:

$$E[g(X)] = \int_{\mathbb{R}} g(x)dF(x),$$

que en el caso discreto resulta en:

$$E[g(X)] = \sum_i g(x_i) \cdot p(x_i),$$

y en el caso continuo:

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)dx.$$

Así, la esperanza va a ser un número, ya sea calculado como suma de términos en el caso discreto, o como suma infinita a través de la integral definida. El uso de integrales no debe intimidar, ya que no se trata más que de áreas debajo de una curva, cuyo cálculo con el software apropiado es muy sencillo. Se puede ver como la suma de los valores de la variable aleatoria (o una función de ella) multiplicado por sus probabilidades. El resultado va a ser el *valor esperado*, que se corresponde con la media de la distribución, su valor central.

El uso de la palabra esperanza en este ámbito tiene su origen, cómo no, en los juegos de azar. Así, se hablaba de la esperanza de ganar en el juego (y la ganancia que se esperaba tener era el resultado), y también del *temor*, cuando la esperanza era negativa.



La esperanza se define como hemos visto sobre una función cualquiera de la variable aleatoria $g(x)$. Si $g(x) = x$, entonces tendremos la esperanza de la propia variable aleatoria. Se cumplen las siguientes propiedades para la esperanza matemática:

- La esperanza de una constante es esa misma constante:

$$c \text{ constante} \implies E[c] = c.$$

- Sea una variable aleatoria que es suma de n variables aleatorias. Entonces su esperanza es la suma de las esperanzas de dichas variables aleatorias:

$$E \left[\sum_{i=1}^n X_i \right] = \sum_{i=1}^n E[X_i].$$

- Sea una variable aleatoria que es producto de n variables aleatorias. Entonces su esperanza es el producto de las esperanzas de dichas variables aleatorias si y solo si dichas variables aleatorias son independientes:

$$E \left[\prod_{i=1}^n X_i \right] = \prod_{i=1}^n E[X_i] \iff X_i \text{ independientes.}$$

- La esperanza de una variable aleatoria es su valor central:

$$E[X] = \mu \implies E[X - \mu] = 0.$$

A la variable aleatoria transformada $X - \mu$ se le denomina **variable aleatoria centrada**, y su media es cero.

- Sea una variable aleatoria que es una transformación lineal de otra variable aleatoria. Entonces su esperanza es la misma transformación lineal de la esperanza de la variable original:

$$a, b \text{ constantes} \implies E[a + bX] = a + bE[X]. \quad (5.2)$$

- Si la integral no existe, la variable aleatoria no tiene esperanza. Esto puede pasar cuando no existe la integral que la define³.

5.5.2. Momentos de variables aleatorias unidimensionales

Hemos visto que la esperanza se define sobre una determinada función $g(x)$ de la variable aleatoria. Los momentos se definen sobre unas funciones muy específicas y que nos van a permitir caracterizar a las variables aleatorias. Se define el momento de orden r respecto al origen de una variable aleatoria X , α_r , como:

³En el caso de variables discretas, cuando la suma no converge.

$$\alpha_r = E[X^r].$$

El momento de orden 1 respecto del origen es la media de la variable aleatoria, μ :

$$\alpha_1 = \boxed{E[X] = \mu}.$$

El momento de orden r respecto de la media μ de una variable aleatoria, μ_r se define como:

$$\mu_r = E[(X - \mu)^r].$$

Nótese que, en realidad, $X - \mu$ es una **variable aleatoria centrada** cuya esperanza es igual a cero por las propiedades de la esperanza enumeradas anteriormente, y por tanto:

$$X - \mu \implies \mu_1 = E[X - \mu] = 0.$$

En el caso discreto, estos momentos se calcularán respectivamente como:

$$\alpha_r = \sum_i x_i^r p(x_i),$$

y

$$\mu_r = \sum_i (x_i - \mu)^r p(x_i).$$

En el caso continuo, se calcularán respectivamente como:

$$\alpha_r = \int_{-\infty}^{\infty} x^r f(x) dx,$$

y

$$\mu_r = \int_{-\infty}^{\infty} (x - \mu)^r f(x) dx.$$

Se verifica la siguiente relación entre los momentos respecto del origen y los momentos respecto de la media que nos ayudarán, como veremos posteriormente, a simplificar los cálculos:

$$\mu_r = \alpha_r - \binom{r}{1} \alpha_1 \alpha_{r-1} + \binom{r}{2} \alpha_1^2 \alpha_{r-2} + \cdots + (-1)^r \alpha_1^r = \sum_{k=0}^r (-1)^k \binom{r}{k} \mu_k \alpha_{r-k}. \quad (5.3)$$

También se verifica que:

- Si existe α_r , entonces existen también todos los α_s tales que $s < r$.
- Si existe μ_r , entonces existen también todos los μ_s tales que $s < r$.

En resumen, podemos calcular momentos respecto de la media (que requieren cálculos más costosos) a través de momentos respecto del origen (cuyos cálculos son más sencillos). Y si existe un momento, todos los de orden inferior también existen.

5.5.3. Medidas de centralización de una variable aleatoria

Ya hemos visto que la esperanza matemática de una variable aleatoria se corresponde con su valor central, al que denominaremos **media**, y es el parámetro de centralización de la variable aleatoria. Es importante no confundir este valor medio o esperado de la variable aleatoria, que es teórico, referido a una población, con la media de unos datos concretos, que es empírica, calculada para una muestra.

$$\mu = \alpha_1 = E[X],$$

cuyo cálculo para variables discretas es el siguiente:

$$\mu = \boxed{E[X] = \sum_i x_i p(x_i)},$$

y para variables continuas:

$$\mu = \boxed{E[X] = \int_{-\infty}^{\infty} x f(x) dx},$$

La **mediana** es otra medida de centralización, y es el valor de la variable aleatoria que *divide* la probabilidad del espacio muestral en dos **mitades**. Por tanto, será el primer valor de la variable aleatoria para el cual la función de distribución vale 0,5:

$$Me = \inf x : F(x) \leq 0,5.$$

En variables discretas a menudo la mediana se puede obtener simplemente de la tabla de valores de $F(x)$. Un método más general consiste en obtener la función inversa de la función de distribución, $F^{-1}(x)$, que estará en función de la probabilidad acumulada, y sustituir la probabilidad por 0,5:

$$F(x) = p \Leftrightarrow x = F^{-1}(p) \Rightarrow Me = F^{-1}(0,5).$$

Cuando no es posible despejar la x hay que recurrir a métodos numéricos para obtener la inversa de la función de distribución. La figura 5.15 muestra gráficamente la mediana como inversa de la función de distribución en $F(x) = 0,5$.

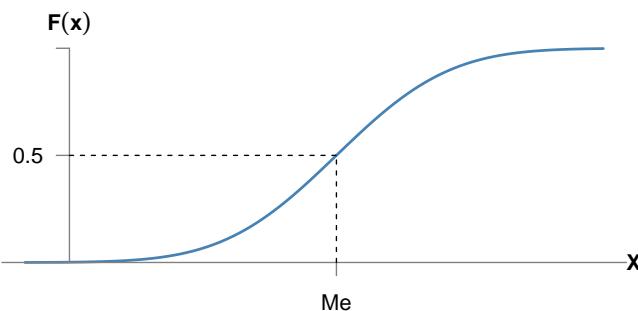


Figura 5.15: La mediana a partir de la inversa de la función de distribución

Por último, la **moda** de una variable aleatoria es el valor donde la función de probabilidad o la función de densidad tienen su máximo. La moda puede no ser única, y en particular para variables continuas se suele hablar de distribuciones *bimodales* o *multimodales* cuando tienen más de un máximo local (aunque solo uno de ellos sea el máximo absoluto). En la figura 5.16 se representan las tres medidas de una determinada variable aleatoria con una determinada función de densidad. Nótese que en una distribución de probabilidad asimétrica, como es la que se representa, la media se desplaza hacia la cola más larga.

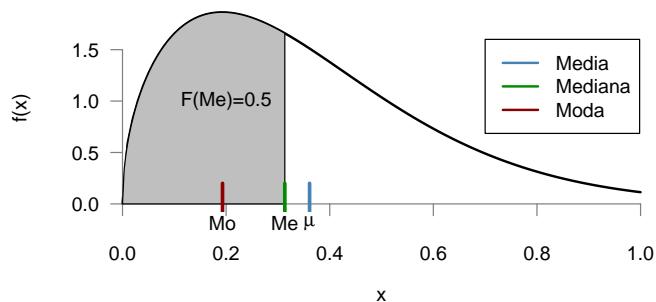


Figura 5.16: Medidas de centralización de una variable aleatoria

La media de la variable aleatoria *número de caras* del experimento descrito más arriba y consistente en lanzar una moneda tres veces, es la siguiente:

$$\mu = E[X] = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1,5.$$

Para obtener la mediana, miramos en la función de distribución el primer valor para el que $F(x) \geq 0,5$, y entonces la mediana es 1. La moda es el valor más frecuente, mirando en la función de probabilidad vemos que los valores 1 y 2 tienen la frecuencia más alta. Como vemos, la moda puede no ser única (sí lo son siempre la media y la mediana.)



HOJA DE CÁLCULO

Disponemos los posibles valores de la variable x_i en la primera columna, las probabilidades en la segunda columna. En la tercera columna calculamos $x_i \cdot p_i$, y sumamos los valores.

R

Podemos guardar los valores y sus probabilidades en sendos vectores y calcular la esperanza calculando la suma del producto de ambos vectores, como se muestra en el siguiente código.



```
x_i <- 0:3
p_i <- c(1/8, 3/8, 3/8, 1/8)
Ex <- sum(x_i*p_i)
Ex
#> [1] 1.5
```

La media de la variable aleatoria definida por la función de densidad:

$$f(x) = \begin{cases} \frac{1}{8}x & \text{si } 0 < x < 4 \\ 0 & \text{resto} \end{cases}$$

Se calcula de la siguiente forma:

$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^4 \frac{1}{8}x^2 dx = \left[\frac{x^3}{24} \right]_0^4 = \frac{64}{24} = \frac{8}{3} \simeq 2,67.$$

Para obtener la mediana, tendríamos que obtener la inversa de la función de distribución, $F^{-1}(p)$, y sustituir p por 0,5. En este caso es sencillo, basta con despejar x de la función de densidad (nos centramos solo en el tramo donde la densidad es mayor de cero):

$$F(x) = p = \frac{x^2}{16} \iff x = F^{-1}(p) = +\sqrt{16p}$$

Tomamos solo la raíz positiva puesto que sabemos que la variable está entre 0 y 4. Entonces la mediana de esta variable aleatoria es:

$$F^{-1}(0,5) = +\sqrt{16 \cdot 0,5} = 2\sqrt{2} \approx 2,8284.$$

En cuanto a la moda sería 4, ya que es el valor donde la función de densidad es máximo, al ser una recta de pendiente positiva entre 0 y 4 (véase la figura 5.9).



MAXIMA

La siguiente expresión devuelve el valor la integral definida con el resultado de la esperanza:

R

El código a continuación obtiene la esperanza de la variable aleatoria.



```
integrate(function(x) x*(1/8)*x, 0, 4)
#> 2.666667 with absolute error < 3e-14
```

Vamos a calcular las medias de las variables aleatorias de los ejemplos de sujetos en estudio. Para la variable aleatoria discreta:

X : Número de mensajes remitidos por correo electrónico en un año a los sujetos,
la media sería:

$$\mu = E[X] = \sum_{i=1}^3 x_i p_i = 20 \cdot \frac{36}{52} + 36 \cdot \frac{12}{52} + 60 \cdot \frac{4}{52} \simeq 26,7692.$$

Para la variable aleatoria continua:

X : Tiempo de duración de la visita a la web de un sujeto,
la media sería:

$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} x \cdot 2e^{-2x}dx = 0,5,$$



resolviendo la integral por partes y aplicando la regla de Barrow.

HOJA DE CÁLCULO (variable discreta)

Disponemos los posibles valores de la variable x_i en la primera columna, las probabilidades en la segunda columna. En la tercera columna calculamos $x_i \cdot p_i$, y sumamos los valores.

MAXIMA

La esperanza de la variable continua la podemos obtener con la siguiente expresión:

`integrate(x*2*exp(-2*x), x, 0, inf);`

R

Para la variable discreta, podemos guardar los valores y sus probabilidades en sendos vectores y calcular la esperanza calculando la suma del producto de ambos vectores, como se muestra en el siguiente código. Para la variable continua, utilizamos la función `integrate` para calcular la integral. Nótese que se pueden usar límites infinitos.



```
x_i <- c(20, 36, 60)
p_i <- c(36/52, 12/52, 4/52)
Ex <- sum(x_i*p_i)
Ex
#> [1] 26.76923

integrate(function(x) x*2*exp(-2*x), 0, Inf)
#> 0.5 with absolute error < 8.6e-06
```

5.5.4. Medidas de dispersión de una variable aleatoria

La **varianza** es el parámetro de dispersión de la variable aleatoria. Se define como el momento de orden 2 respecto de la media, y se representa por σ^2 .

$$V[X] = \sigma^2 = \mu_2 = E[(X - \mu)^2],$$

que para variables discretas se calcula como:

$$\sigma^2 = V[X] = \sum_i (x_i - \mu)^2 p(x_i),$$

y para variables continuas como:

$$\sigma^2 = V[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

Aplicando la relación entre los momentos respecto del origen y los momentos respecto de la media de la ecuación (5.3) resulta que:

$$\mu_2 = \alpha_2 - \alpha_1^2,$$

y podemos calcular la varianza con la siguiente expresión *abreviada*:

$$\boxed{\sigma^2 = E[X^2] - E[X]^2},$$

donde

$$\alpha_2 = E[X^2] = \sum_i x_i^2 p(x_i)$$

para variables discretas y:

$$\alpha_2 = E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx$$

para variables continuas.

La varianza de una variable aleatoria cumple además las siguientes *propiedades*:

- La varianza de una constante es nula:

$$V[c] = 0.$$

- La varianza de una variable aleatoria es siempre positiva:

$$V[X] \geq 0.$$

- Sea una variable aleatoria que es una transformación lineal de otra variable aleatoria. Entonces su varianza es:

$$a, b \text{ constantes} \implies V[a + bX] = b^2 V[X].$$

Nótese la diferencia con la esperanza de la transformación lineal vista en la ecuación (5.2).

La **desviación típica** de la variable aleatoria es la raíz cuadrada positiva de la varianza. La desviación típica viene expresada en las mismas unidades que la variable aleatoria, mientras que la varianza está expresada en las unidades de la variable aleatoria al cuadrado.

$$\sigma = +\sqrt{V[X]}.$$

Una característica adimensional de la variabilidad es el **coeficiente de variación**, que es el cociente entre la desviación típica y la media de la variable aleatoria. Si la media fuera negativa, se suele expresar en valor absoluto:

$$CV = \frac{\sigma}{\mu}.$$

Para calcular la varianza de la variable aleatoria *número de caras* del experimento consistente en lanzar una moneda tres veces, primero calculamos el momento de orden 2 respecto del origen:

$$\alpha_2 = E[X^2] = 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{3}{8} + 3^2 \cdot \frac{1}{8} = 3.$$

Como ya sabíamos que la media era $\mu = 1,5$, entonces la varianza es:

$$\sigma^2 = \alpha_2 - \mu^2 = 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4} = 0,75.$$

La desviación típica y el coeficiente de variación serán:

$$\sigma = \sqrt{3/4} \simeq 0,8660; CV = \frac{\sigma}{\mu} \simeq 0,5774$$

Para calcular la media de la variable aleatoria definida por la función de densidad:

$$f(x) = \begin{cases} \frac{1}{8}x & \text{si } 0 < x < 4 \\ 0 & \text{resto} \end{cases}$$

calculamos también en primer lugar el momento de orden 2 respecto del origen, en este caso a través de la integral:

$$\alpha_2 = E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^4 \frac{1}{8} x^3 dx = \left[\frac{x^4}{32} \right]_0^4 = \frac{256}{32} = 8.$$

Como la media era $\mu = \frac{8}{3}$, la varianza es:

$$\sigma^2 = \alpha_2 - \mu^2 = 8 - \left(\frac{8}{3}\right)^2 = \frac{8}{9} \simeq 0,8889.$$

La desviación típica y el coeficiente de variación serán:

$$\sigma = \sqrt{8/9} \simeq 0,9428; CV = \frac{\sigma}{\mu} \simeq 0,3536$$



HOJA DE CÁLCULO (variable discreta)

Si tenemos dispuestos los valores y probabilidades como se indicó más arriba, podemos añadir dos columnas con el cálculo de x_i^2 y $x_i^2 \cdot p_i$ en cada fila, sumar esta última para obtener α_2 y a continuación restarle la media calculada anteriormente elevada al cuadrado, para obtener la varianza.

MAXIMA

Para la variable continua obtenemos α_2 con la siguiente expresión, y después podemos hacer operaciones para calcular todas las características:

```
integrate(x^2*(1/8)*x, x, 0, 4);
```

R

El siguiente código realiza todos los cálculos para obtener los distintos parámetros de dispersión.



```
alpha_2 <- integrate(function(x) x^2*(1/8)*x, 0, 4)$value
alpha_1 <- integrate(function(x) x*(1/8)*x, 0, 4)$value
varianza <- alpha_2 - alpha_1^2; varianza
#> [1] 0.8888889
desv.tip <- sqrt(varianza); desv.tip
#> [1] 0.942809
cv <- desv.tip/alpha_1; cv
#> [1] 0.3535534
```

Vamos a calcular las varianzas de las variables aleatorias de los ejemplos de sujetos en estudio. Para la variable aleatoria discreta:
 X : Número de mensajes remitidos por correo electrónico en un año a los sujetos,
el momento de orden dos con respecto al origen sería:

$$\alpha_2 = E[X^2] = \sum_{i=1}^3 x_i^2 p_i = 20^2 \cdot \frac{36}{52} + 36^2 \cdot \frac{12}{52} + 60^2 \cdot \frac{4}{52} \simeq 852,9231.$$

Y entonces, la varianza es:

$$\sigma^2 = \alpha_2 - \mu^2 = 852,9231 - (26,7692)^2 = 136,333.$$

Para la variable aleatoria continua:

X : Tiempo de duración de la visita a la web de un sujeto,
el momento de orden dos sería:

$$\alpha_2 = E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^{\infty} x^2 \cdot 2e^{-2x} dx = 0,5,$$

y entonces la varianza es:

$$\sigma^2 = \alpha_2 - \mu^2 = 0,5 - (0,5)^2 = 0,25.$$



HOJA DE CÁLCULO (discreta)

Procederíamos igual que en el ejemplo anterior, calculando en columnas, sumando totales y finalmente aplicando la fórmula.

MAXIMA

La siguiente expresión calcularía α_2 , y a partir de ahí se aplican las fórmulas para obtener los distintos parámetros.

`integrate(x^2*2*exp(-2*x), x, 0, inf);`

R

El código a continuación calcula la varianza de la variable aleatoria de forma análoga al ejemplo anterior.



```
alpha_2 <- integrate(function(x) x^2*2*exp(-2*x), 0, Inf)$value
alpha_1 <- integrate(function(x) x*2*exp(-2*x), 0, Inf)$value
varianza <- alpha_2 - alpha_1^2; varianza
#> [1] 0.25
```

A partir de la variable aleatoria anterior:

X : Tiempo de duración de la visita a la web de un sujeto,
supongamos que esta visita se produce siempre después de haber visto un anuncio de 10 segundos, y queremos estudiar la variable:

Y : Tiempo total de conexión con el servidor en segundos.

Esta nueva variable aleatoria se puede expresar como:

$$Y = 10 + 60 \cdot X$$

y tendrá una determinada distribución de probabilidad cuya determinación no se trata en este texto. En cualquier caso, a través de las propiedades de la esperanza y la varianza, sí podemos calcular el valor de estas características para la nueva distribución. Así:

$$E[Y] = 10 + 60 \cdot E[X] = 10 + 60 \cdot 0,5 = 40,$$

$$V[Y] = 60^2 \cdot V[X] = 60^2 \cdot 0,5 = 1800.$$

5.5.5. Variable aleatoria estandarizada

La última de las propiedades de la varianza enumeradas anteriormente, es decir:

$$a, b \text{ constantes, } Y = a + bX \implies V[Y] = b^2 V[X],$$

nos va a permitir *escalar* cualquier variable aleatoria transformándola en otra que tenga desviación típica igual a uno. Efectivamente, si en la transformación lineal anterior hacemos $b = \frac{1}{\sigma}$:

$$V[Y] = \left(\frac{1}{\sigma}\right)^2 \cdot \sigma^2 = 1.$$

Si aplicamos esta transformación a una variable aleatoria centrada $X - \mu$, entonces tenemos una **variable aleatoria estandarizada** con media cero y desviación típica 1 y que normalmente se denota por Z :

$$Z = \frac{X - \mu}{\sigma} \implies \mu_Z = 0; \sigma_Z = 1.$$

Utilizaremos esta transformación para realizar cálculo de probabilidades del modelo de distribución normal en el capítulo 7. Además, tiene mucho interés en Estadística inferencial y en técnicas multivariantes que no se tratan en este texto.

5.5.6. Otros parámetros

Al igual que se definió la mediana, podemos definir cualquier **cuantil** X_p para una probabilidad dada:

$$X_p = \inf x : F(x) \leq p$$

Por ejemplo, los cuantiles 0,25 y 0,75 serían los valores $X_{0,25}$ y $X_{0,75}$ que dejan por debajo una probabilidad de 0,25 y 0,75 respectivamente⁴. El método más general para calcular cualquier cuantil consiste en obtener la inversa de la función de distribución $x = F^{-1}(p)$ y dar valores a p (véase el ejemplo de la mediana más arriba). La figura 5.17 muestra la representación del cuantil 0,75 en relación con una determinada función de densidad.

También se pueden calcular a partir de los momentos otros parámetros como los coeficientes de asimetría y de curtosis de una variable aleatoria, que no se tratan en este texto.

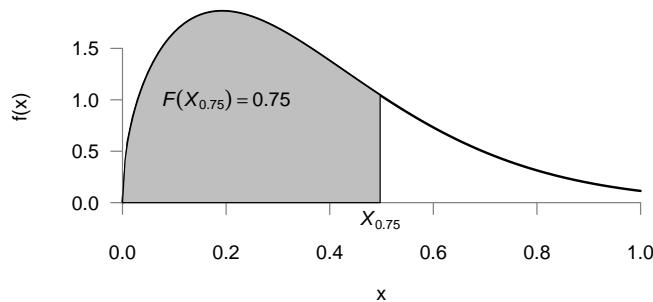


Figura 5.17: Cuantiles de una variable aleatoria

5.5.7. Desigualdad de Chebyshev

En ocasiones, es posible que conozcamos la media, μ , y la varianza, σ^2 , de una variable aleatoria, pero no conocemos nada sobre su distribución de probabilidad. En estos casos, no podemos calcular la probabilidad en un intervalo, pero podemos acotar la probabilidad entre dos valores entorno a la media conocida. La fórmula general para esta acotación es la siguiente:

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2},$$

conocida como **desigualdad de Chebyshev**⁵ y que nos permite acotar la probabilidad de una variable aleatoria de dos formas:

⁴A estos valores se les conoce como cuartiles

⁵Podemos encontrar diversas grafías del apellido de este matemático ruso (1821-1894) como Chebyshov, Tchebychev, Tchebycheff, Tschebyscheff, Chebyshev o Čebišev.

- La probabilidad de que la variable aleatoria tome valores más extremos de k desviaciones típicas desde la media es, como mucho, $\frac{1}{k^2}$, véase la figura 5.18:

$$P[\mu - k\sigma \geq X \geq \mu + k\sigma] \leq \frac{1}{k^2}.$$

- La probabilidad de que la variable aleatoria tome valores dentro de k desviaciones típicas desde la media es, como poco, $1 - \frac{1}{k^2}$, véase la figura 5.19:

$$P[\mu - k\sigma < X < \mu + k\sigma] \geq 1 - \frac{1}{k^2}.$$

Si lo que queremos es acotar la probabilidad para un valor concreto x , entonces podemos encontrar primero k despejando de $\mu + k\sigma = x$ y después aplicar las propiedades de la probabilidad para encontrar una cota.

De la desigualdad de Chebyshev se deduce que, por ejemplo, para cualquier variable aleatoria la probabilidad de que esa variable aleatoria tome valores entre su media y dos desviaciones típicas es de, al menos, 0,75:

$$P[\mu - 2\sigma < X < \mu + 2\sigma] \geq 1 - \frac{1}{2^2} = 0,75.$$

También podemos determinar mediante esta desigualdad, entre qué valores estará, al menos, una determinada probabilidad.

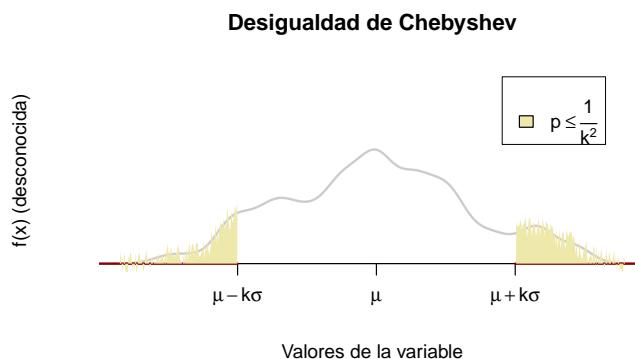


Figura 5.18: Cota superior externa Desigualdad de Chebyshev

Desigualdad de Chebyshev

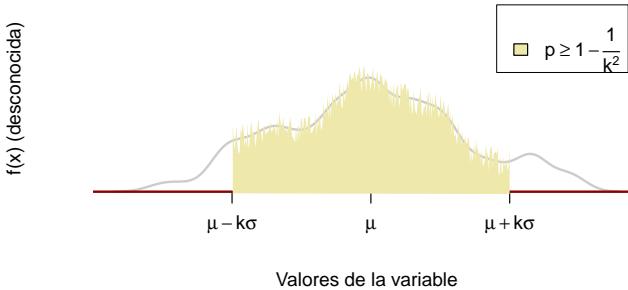


Figura 5.19: Cota inferior interna Desigualdad de Chebyshev

Se sabe que la media de una variable aleatoria es 9 y su varianza 4. ¿Entre qué dos valores tendremos, al menos, una probabilidad de 0.75?

De la propiedad que acabamos de ver, esto se cumple para $k = 2$, y por tanto esos valores serán $\mu \pm 2\sigma = 9 \pm 2 \cdot \sqrt{4} = [5, 13]$.

¿Entre qué valores tendremos una probabilidad de, al menos, 0.84?

Para contestar a esta pregunta, calculamos primero k teniendo en cuenta:

$$P[\mu - k\sigma < X < \mu + k\sigma] \geq 1 - \frac{1}{k^2} \implies 1 - \frac{1}{k^2} = 0,84 \implies k = 2,5,$$

y calculamos los valores como:

$$\mu \pm k\sigma = 9 \pm 2,5 \cdot \sqrt{4} = [4, 14].$$

Entre 4 y 14 tendremos, al menos, una probabilidad de 0,84. Y a la inversa, más allá de estos valores tendremos, como mucho, una probabilidad de 0,16.

¿Cuál sería la probabilidad de que esta variable aleatoria tome valores mayores de 15?

No podemos contestar exactamente a esta pregunta puesto que no disponemos de la distribución de probabilidad. Pero sí podemos dar una cota de dicha probabilidad. En este caso tenemos que obtener k sabiendo cuánto vale $\mu + k\sigma$:

$$\mu + k\sigma = 15 \iff k = \frac{15 - 9}{2} = 3$$

Entonces:

$$P[9 - 3 \cdot 2 > X > 9 + 3 \cdot 2] \leq \frac{1}{k^2} \iff P[3 > X > 15] \leq \frac{1}{9} \approx 0,1111$$

Nótese que esto significa que:

$$P[X \leq 3] + P[X \geq 15] \leq 0,1111,$$

y por tanto si la suma de dos números es menor que 0.1111, entonces cada uno de ellos será como mucho ese valor, y podemos asegurar que la probabilidad de que esta variable aleatoria con media 9 y varianza 4 tome valores de 15 es menor de 0.1111.



Capítulo 6

Variable aleatoria bivariante

En preparación.

Distribución conjunta

Correlación y regresión

Capítulo 7

Modelos de distribución de probabilidad

7.1. Introducción

En el capítulo 5 vimos que una variable aleatoria unidimensional se puede modelizar por cualquier función de distribución de probabilidad que cumpla los requisitos básicos de la probabilidad así, tenemos infinitas funciones de probabilidad para variables aleatorias discretas, o de densidad para variables aleatorias continuas. Sin embargo, la mayoría de los fenómenos de interés estudiados mediante la probabilidad se ajustan a un reducido conjunto de modelos de distribución de probabilidad o familias de distribuciones para los que se han determinado sus características principales, facilitando así el trabajo con variables aleatorias. En este capítulo revisaremos los más importantes para variables aleatorias discretas.

El primer paso para identificar el modelo de distribución de probabilidad más adecuado, es describir claramente la variable aleatoria X , y de ahí deducir cuál es el modelo adecuado. Para cada modelo, se conoce su función de probabilidad o de densidad que contiene un número muy reducido de **parámetros**. A partir de esta función de probabilidad o de densidad, se deducen sus características, por ejemplo la media y la varianza, que quedan expresadas en función de dichos parámetros. Una vez identificado el modelo de distribución de probabilidad, hay que establecer los parámetros concretos que caracterizan la variable aleatoria concreta de interés. En este libro se asumen como conocidos (o deducibles fácilmente de la descripción del problema), aunque en aplicaciones reales se deberán estimar a partir de muestras representativas de la población con técnicas de inferencia estadística, que no se tratan en este texto. Una vez determinados los parámetros, podemos calcular fácilmente las características de la variable aleatoria con las fórmulas dadas, así como realizar cálculo de probabilidades

utilizando todo lo aprendido hasta ahora.

Para indicar que una variable aleatoria X sigue una determinada distribución de probabilidad, utilizamos la siguiente notación:

$$X \sim \mathcal{Distr}(\theta),$$

donde \mathcal{Distr} identifica el modelo de distribución de probabilidad, y θ es el vector de parámetros con los que queda totalmente definida la distribución de probabilidad de la variable aleatoria X según ese modelo de distribución. Tanto para los modelos de distribución de probabilidad discretos de este capítulo, como en los continuos del siguiente, se proporciona la función de probabilidad o de densidad de los mismos, así como la esperanza y la varianza que se deduce de las mismas (aunque no se incluye dicha deducción). El resto de características de cada modelo se puede obtener igualmente a partir de su distribución de probabilidad. Tampoco se incluyen las demostraciones de que, obviamente, las funciones de densidad y de probabilidad de cada distribución cumplen las propiedades para ser una Ley de probabilidad.

7.2. Modelos de distribución de probabilidad discretos

7.2.1. Distribución de Bernoulli

Las distribuciones de probabilidad discretas se basan de una forma u otra en procesos de Bernoulli. Un proceso de Bernoulli consiste en realizar un experimento que tiene dos resultados posibles. A uno le llamamos éxito y al otro le llamamos fracaso, y conocemos la probabilidad del suceso *éxito*, a la que llamamos p .

Dado un proceso de Bernoulli aislado, podemos definir la variable aleatoria X que toma el valor 1 si el experimento es un éxito, y 0 si el experimento es un fracaso.

$$X = \begin{cases} 1 & \text{si éxito con probabilidad } p \\ 0 & \text{si fracaso} \end{cases}$$

Entonces las probabilidades para los dos posibles valores de la variable serán:

$$P[X = 1] = p; \quad P[X = 0] = 1 - p,$$

y diremos que X sigue una distribución de Bernoulli de parámetro p :

$$X \sim Ber(p); \quad 0 < p < 1.$$

Algunas veces se utiliza la notación $q = 1 - p$. Una expresión general para la **función de probabilidad** es la siguiente:

$$P[X = x] = p^x(1 - p)^{1-x}; \quad x = 0, 1.$$

Las características de posición y dispersión de esta variable aleatoria se deducen fácilmente:

- Media: $\mu = E[X] = p$.
- Varianza: $\sigma^2 = V[X] = p \cdot (1 - p)$.

La distribución de Bernoulli aparece en los procesos de clasificación de observaciones (individuos, empresas, etc.) en una de dos categorías.

En el ejemplo de los potenciales usuarios de nuestro servicio, dedujimos en el capítulo 4 que la probabilidad de que un cliente tomado al azar contrate el servicio era 0,25. Entonces la variable aleatoria:

$$X : \begin{cases} 0 & \text{el cliente no contrata} \\ 1 & \text{el cliente contrata} \end{cases}$$

sigue una distribución de probabilidad de Bernoulli de parámetro $p = 0,25$, su media es $\mu = 0,25$, su varianza $\sigma^2 = 0,1875$ y su función de probabilidad:

$$P[X = x] = 0,25^x \times 0,75^{1-x}$$

El interés de la distribución de Bernoulli también está en las distribuciones de probabilidad derivadas de ella cuando repetimos el proceso bajo distintas condiciones. En los siguientes apartados veremos algunas de estas distribuciones que se extienden a partir de la de Bernoulli.

7.2.2. Distribución binomial

Partiendo de un proceso de Bernoulli, consideraremos la repetición del experimento n veces, y que el resultado de cada experimento es **independiente** de los demás. Entonces, la variable aleatoria X : *Número de éxitos en n pruebas independientes de Bernoulli con probabilidad de éxito p cada una de ellas* sigue una distribución de probabilidad binomial de parámetros n y p :

$$X \sim Bin(n; p); \quad n > 0, \quad 0 < p < 1.$$

Nótese que la distribución de Bernoulli es un caso particular de la binomial cuando $n = 1$.

$$Ber(p) = Bin(1; p).$$

A su vez, la distribución binomial es la suma de n variables aleatorias independientes de Bernoulli:

$$\Rightarrow Bin(n; p) = \sum_{i=1}^n X_i : X_i \sim Ber(p) \forall i,$$

de donde llegamos a la siguiente expresión de la función de probabilidad:

$$P[X = x] = \binom{n}{x} \cdot p^x \cdot (1-p)^{(n-x)}; x = 0, 1, \dots, n,$$

donde:

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

conocido como número combinatorio o coeficiente binomial. En el apéndice C.2 se encuentran algunas propiedades de este coeficiente, que se puede calcular fácilmente en las calculadoras científicas con la tecla **nCr**.

Nótese que en la fórmula de la función de probabilidad de la distribución binomial aparecen muchos conceptos de probabilidad aprendidos hasta ahora. Como son sucesos independientes, p^x es la probabilidad de la intersección de x éxitos, y $(1-p)^{n-x}$, la probabilidad de la intersección de $n-x$ fracasos. Entonces $p^x \cdot (1-p)^{(n-x)}$ es la probabilidad de una de las ordenaciones posibles. Como el orden de éxitos y fracasos nos da igual, la probabilidad que nos interesa es la probabilidad de la unión de todas las ordenaciones posibles que, como son sucesos disjuntos, se corresponde con la suma de probabilidades. Estas probabilidades son todas iguales, y el número de ordenaciones posibles es $\binom{n}{x}$, por eso multiplicamos.

La figura 7.1 muestra gráficamente la distribución de probabilidad para varios valores de n y p .

```
#>
#> Attaching package: 'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
```

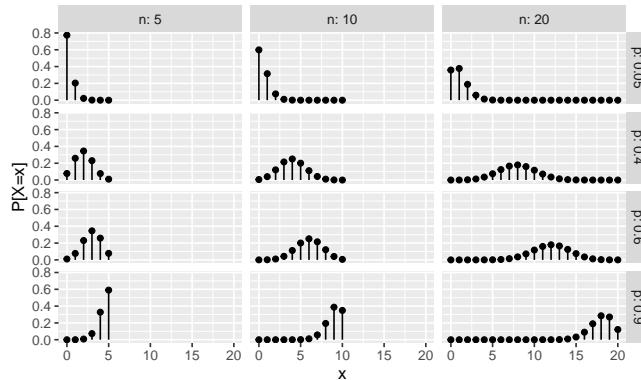


Figura 7.1: Representación de la función de probabilidad del modelo binomial

Las características principales de la distribución binomial se deducen fácilmente aplicando las fórmulas de la esperanza matemática vistas en el capítulo 5, y son:

- Media: $\mu = E[X] = n \cdot p$.
- Varianza: $\sigma^2 = V[X] = n \cdot p \cdot (1 - p)$.

La distribución binomial, además, cumple la propiedad aditiva, es decir, la suma de m variables aleatorias binomiales con idéntico parámetro p y, posiblemente, distintos parámetros n_j , $j = 1, \dots, m$, es una distribución binomial de modo que:

$$Y = \sum_{j=1}^m X_j, \quad X_j \sim Bin(n_j; p) \implies Y \sim Bin\left(\sum_{j=1}^m n_j; p\right).$$

Esta propiedad, que iremos viendo en casi todos los modelos, es muy importante porque nos permite resolver problemas de probabilidad en los que se repiten las realizaciones de las variables aleatorias, lo que nos interesa es el total. No hay que confundir la **suma** de variables aleatorias con la **mezcla** de poblaciones en los que hay que aplicar los teoremas de la probabilidad total y de Bayes.

Supongamos que la probabilidad de que un estudiante acabe un grado en Ciencias es de 0,4. Tomamos al azar un grupo de 5 estudiantes. ¿Cuál es la probabilidad de que ninguno obtenga el grado? ¿Y la probabilidad de que al menos dos lo obtengan?

Si definimos la variable aleatoria X : *Número de estudiantes que obtienen el grado de un grupo de 5*, entonces X sigue la distribución:

$$X \sim Bin(5; 0,4); x = 0, 1, 2, 3, 4, 5$$

y por tanto las probabilidades pedidas son, respectivamente:

$$P[X = 0] = \binom{n}{x} \cdot p^x \cdot (1 - p)^{(n-x)} = \binom{5}{0} \cdot 0,4^0 \cdot (0,6)^5 \simeq 0,0776.$$

$$\begin{aligned} P[X \geq 2] &= 1 - P[X < 2] = 1 - [P[X = 0] + P[X = 1]] = \\ &= 1 - \left[0,0778 + \binom{5}{1} \cdot 0,4^1 \cdot 0,6^4 \right] \simeq 0,6630. \end{aligned}$$



HOJA DE CÁLCULO

En las aplicaciones de hoja de cálculo, tenemos funciones que devuelven la densidad (probabilidad en modelos discretos) y la probabilidad acumulada (función de distribución) de los modelos de distribución de probabilidad más utilizados. Puede diferir el nombre de la función entre diferentes programas. En Hojas de Cálculo de Google y LibreOffice se obtendrían las probabilidades del ejemplo así:

```
=BINOM.DIST(0;5;0,4;0)
=1-BINOM.DIST(1;5;0,4;1)
Mientras que en EXCEL la función se llama DISTR.BINOM.N:
=DISTR.BINOM.N(0;5;0,4;)
=1-DISTR.BINOM.N(1;5;0,4;VERDADERO)
```

R

En R, para cada modelo de distribución de probabilidad tenemos una función que empieza por `d` y devuelve la “densidad” (probabilidad en el caso de discretas) y otra que empieza por `p` y devuelve la “probabilidad (acumulada)”, es decir, la función de distribución (o su complementario). Después de la `d` o la `p` vendrá el nombre (o abreviatura) del modelo de probabilidad, por ejemplo para la binomial `binom`. Entonces la función `dbinom` devuelve la probabilidad para un valor de la variable aleatoria. A las funciones hay que pasarle también los parámetros del modelo de distribución. En el caso de la binomial, el parámetro `p` y el parámetro `n`. A continuación se muestran las expresiones que calculan las probabilidades del ejemplo. Véase cómo la segunda probabilidad se puede calcular de varias formas, utilizando el complementario como en la hoja de cálculo, el argumento `lower.tail` de la función `dbinom`, o sumando las probabilidades para los valores que cumplen la condición.



```
dbinom(x = 0, size = 5, prob = 0.4)
#> [1] 0.07776
1 - pbinom(q = 1, size = 5, prob = 0.4)
#> [1] 0.66304
pbinom(q = 1, size = 5, prob = 0.4, lower.tail = FALSE)
#> [1] 0.66304
sum(dbinom(x = 2:5, size = 5, prob = 0.4))
#> [1] 0.66304
```

Selecciono 10 potenciales sujetos del estudio al azar. ¿Cuál es la probabilidad de que al menos uno responda al tratamiento?

En términos de variable aleatoria:

- X : Número de éxitos en 10 experimentos independientes de Bernoulli con probabilidad de éxito 0.25
- $X \sim \text{Bin}(10; 0,25)$
- $P[X \geq 1] = 1 - P[X < 1] = 1 - P[X = 0] \simeq 1 - 0,0563 \simeq 0,9437$



HOJA DE CÁLCULO

[LibreOffice] =1-BINOM.DIST(0;10;0,25;1)

[EXCEL] =1-DISTR.BINOM.N(0;10;0,25;VERDADERO)

R

La siguiente expresión calcula la probabilidad buscada. El lector puede probar otros caminos para llegar a la misma probabilidad, como en el ejemplo anterior.



```
binom(q = 0, size = 10, prob = 0.25, lower.tail = FALSE)
#> [1] 0.9436865
```

Hay tres consideraciones muy importantes a la hora de resolver ejercicios en variables discretas:

1. Es muy importante tener claro cuáles son los posibles valores de la variable aleatoria, y así saber qué probabilidades hay que calcular.
2. Es posible llegar al resultado de varias formas posibles, y hay que pararse a pensar cuál será la más rápida, usando las propiedades de la probabilidad (principalmente: probabilidad del suceso complementario y probabilidad de la unión de sucesos disjuntos).
3. Al cambiar de una probabilidad a la del suceso contrario, es muy importante tener en cuenta si las desigualdades incluyen el símbolo igual.

7.2.3. Distribución de Poisson

La distribución de Poisson surge inicialmente como distribución límite de la binomial cuando n tiende a infinito y p se mantiene estable. Posteriormente se vio que describe muy bien los procesos donde se cuentan el número de ocurrencias de un evento por unidad (de tiempo, espacio, ...). La probabilidad de ocurrencia en un instante concreto es muy baja, pero en un intervalo determinado es muy probable que suceda varias veces. Bajo estas circunstancias, la variable aleatoria:

X : Número de eventos por unidad

sigue una distribución de Poisson:

$$X \sim Poiss(\lambda); \lambda > 0,$$

donde el único parámetro λ es la media y la varianza de la variable aleatoria. Es decir, se producen, de media, λ eventos por unidad de tiempo, superficie, etc. La distribución de Poisson tiene la siguiente función de probabilidad:

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}; x = 0, 1, \dots, \infty.$$

 El estadístico ruso L. Bortkewicz explicó en 1898 que la distribución de Poisson describía muy bien el número de muertes producidas por picaduras de caballo en las guerras prusianas.

La figura 7.1 muestra gráficamente la distribución de probabilidad para varios valores de n y p . Se representan valores desde $x = 0$ hasta $x = \mu + 4\sigma$. Aunque teóricamente los posibles valores son hasta infinito, a partir de ese valor la probabilidad es prácticamente cero. Para valores de λ grandes, esto también sucede en los valores de x bajos.

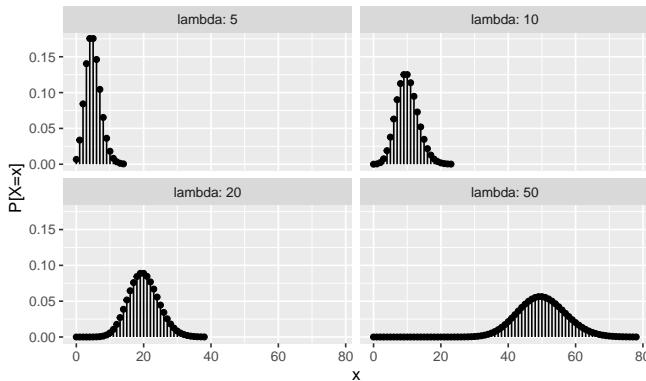


Figura 7.2: Representación de la función de probabilidad del modelo de Poisson

Las características principales de la distribución de Poisson son las siguientes:

- Media: $\mu = E[X] = \lambda$.
- Varianza: $\sigma^2 = V[X] = \lambda$.

Como la binomial, también cumple la propiedad aditiva de modo que, para m variables aleatorias independientes de Poisson:

$$Y = \sum_{j=1}^m X_j, X_j \sim Poiss(\lambda_j) \implies Y \sim Poiss\left(\sum_{j=1}^m \lambda_j\right).$$

En una parada de autobús llegan de media cuatro autobuses cada hora. Cuál es la probabilidad de llevar una hora y que no haya pasado ninguno todavía?

Si X : número de autobuses que pasan en una hora, entonces:

$$X \sim Poiss(4),$$

y entonces lo que queremos saber es:

$$P[X = 0] = \frac{e^{-\lambda} \lambda^x}{x!} = \frac{e^{-4} \cdot 4^0}{0!} \simeq 0,0183.$$



HOJA DE CÁLCULO

En este caso la función si es la misma en Excel y en las hojas de cálculo libres.

=POISSON.DIST(0;4;0)

R

La siguiente expresión calcula la probabilidad pedida. Nótese que ahora se utiliza `pois` en el nombre de la función.



```
dpois(x = 0, lambda = 4)
#> [1] 0.01831564
```

La tasa media semanal de visitas de un cliente a la página web de ofertas es igual a 8. Calcular la probabilidad de que un posible cliente acceda menos de 3 veces en una semana. En términos de variable aleatoria, tenemos que:

- X : Número de visitas por semana a la web de oferta
- $X \sim Poiss(8)$
- $P[X < 3] = P[X \leq 2] = \sum_{x=0}^2 P[X = x] = P[X = 0] + P[X = 1] + P[X = 2] \simeq 0,0003 + 0,0027 + 0,0107 = 0,0138$

Supongamos que estamos interesados en las visitas que un cliente hace a la página web durante cuatro semanas. Y queremos saber la probabilidad de que acceda 30 veces. Entonces aplicamos la propiedad aditiva de la distribución de Poisson, y definimos:

Y : Número de visitas en cuatro semanas $= X_1 + X_2 + X_3 + X_4$,
donde

X_i : Número de visitas en el día i , $i = 1, 2, 3, 4 \sim Poiss(8)$

Entonces:

$$Y \sim Poiss(32),$$

y la probabilidad buscada es:



$$P[Y = 30] = \frac{e^{-32} \cdot 32^{30}}{30!} \simeq 0,0681.$$

HOJA DE CÁLCULO

=POISSON.DIST(2;8;1)
=POISSON.DIST(30;32;0)

R

Las siguientes expresiones obtienen las probabilidades pedidas a través de la función de distribución y de probabilidad.



```
ppois(q = 2, lambda = 8)
#> [1] 0.01375397
dpois(x = 30, lambda = 32)
#> [1] 0.06814215
```

La distribución de Poisson se puede utilizar como aproximación de la distribución binomial bajo ciertas condiciones. En la práctica, para $n \geq 100$ y $p \leq 0,05$, se puede utilizar la aproximación:

$$X \sim Bin(n; p) \rightsquigarrow Poiss(\lambda = np),$$

siempre y cuando np tenga sentido como parámetro λ , es decir, no excesivamente

grande ni excesivamente pequeño. La figura 7.3 muestra la función de distribución de una variable aleatoria binomial con parámetros $n = 100$, $p = 0,05$ y su aproximación por una Poisson de parámetro $\lambda = 5$.

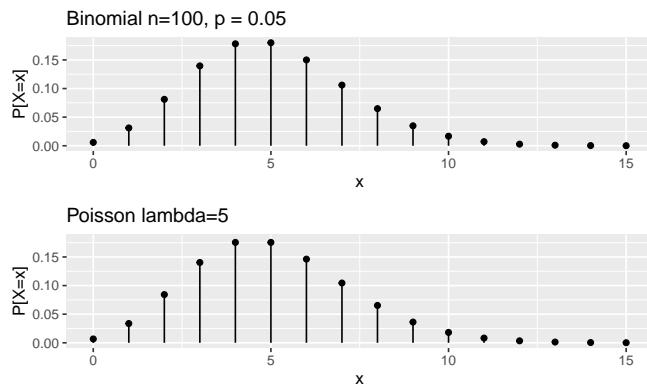


Figura 7.3: Aproximación a binomial por la Poisson

Supongamos que tenemos en la página web del estudio un formulario de contacto, y que sabemos por históricos que el 1% de los sujetos de nuestro servicio que entran al formulario, terminan enviando una reclamación.

Tomamos al azar 100 potenciales usuarios. ¿Cuál es la probabilidad de que menos de 3 hayan puesto una reclamación?

La variable aleatoria con la que podemos modelizar este problema es:

X : Número de clientes de una muestra de 100 que pone una reclamación, que sigue una distribución binomial de parámetros $n = 100$, $p = 0,01$.

Como se dan los requisitos, podemos hacer la aproximación a la distribución de Poisson, y entonces:

$$X \sim Poiss(\lambda = 1),$$

y la probabilidad pedida la podemos aproximar como:

$$P[X < 3] = \sum_{x=0}^2 \frac{e^{-1} 1^x}{x!} \simeq 0,9199.$$



R

Utilizando software, podemos hacer los cálculos exactos. Vemos que, en este caso concreto, nos estaremos equivocando en el tercer decimal.



```
pbinom(2, 100, 0.01)
#> [1] 0.9206268
ppois(2, 1)
#> [1] 0.9196986
```

7.2.4. Distribución binomial negativa

La distribución binomial negativa describe procesos en los que realizamos sucesivos experimentos independientes de Bernoulli, con probabilidad de éxito p . Pero no sabemos cuántos vamos a realizar, porque lo que nos interesa es el número de fracasos x hasta que se produzcan c éxitos. También se puede expresar como el número total de pruebas necesarias $x + c$ hasta obtener c éxitos. Así, definimos la variable aleatoria:

X : Número de fracasos hasta c éxitos

que sigue el modelo de distribución de probabilidad binomial negativa con parámetros c y p :

$$X \sim BN(c; p); c > 0; 0 < p < 1.$$

Nótese que X puede tomar, teóricamente, cualquier valor mayor o igual que 0 (no tiene límite). Su función de probabilidad es:

$$P[X = x] = \binom{x + c - 1}{x} \cdot p^c \cdot (1 - p)^x; x = 0, 1, 2, \dots, \infty,$$

donde:

$$\binom{x + c - 1}{x} = \frac{(c + x - 1)!}{x!(c - 1)!}.$$

Si nos fijamos detenidamente en la función de probabilidad, podemos hacer el mismo análisis que en la binomial, multiplicando las probabilidades de cada experimento independiente de Bernoulli para una ordenación posible, y sumando las probabilidades de cada ordenación. La diferencia está en que el último experimento es siempre un éxito (habremos llegado al éxito número c , y paramos). Si se da $X = x$, entonces habremos realizado un total de $x + c$ pruebas de Bernoulli.

El término *negativa* viene de la siguiente forma alternativa de escribir su función de probabilidad:

$$P[X = x] = \binom{-c}{x} \cdot p^c \cdot (1-p)^x.$$

La figura 7.4 muestra gráficamente la distribución de probabilidad para varios valores de c y p . Se representan valores desde $x = 0$ hasta $x = 20$. Aunque teóricamente los posibles valores son hasta infinito, a partir de cierto valor (dependiendo de los parámetros) la probabilidad es prácticamente cero. Para valores de p pequeños, esto también sucede en los valores de x bajos.

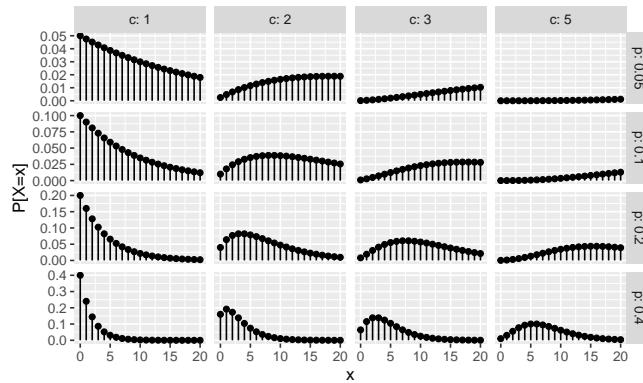


Figura 7.4: Representación de la función de probabilidad del modelo binomial negativo

La media y la varianza de una variable aleatoria que sigue un modelo binomial negativo son:

- Media: $\mu = E[X] = \frac{c \cdot (1-p)}{p}$.
- Varianza: $\sigma^2 = V[X] = \frac{c \cdot (1-p)}{p^2}$.

Se cumple la propiedad aditiva de forma similar a como lo hacía en la distribución binomial. Es decir, la suma de m variables aleatorias binomiales negativas con el mismo parámetro p y parámetros c_j que pueden ser diferentes, es una variable aleatoria que sigue también una distribución binomial negativa con el mismo parámetro p :

$$Y = \sum_{j=1}^m X_j, \quad X_j \sim BN(c_j; p) \implies Y \sim BN\left(\sum_{j=1}^m c_j; p\right).$$

Dos equipos de balonmano A y B se disputan la final de liga al mejor de 7 partidos. El factor campo no influye y el equipo A tiene una probabilidad de ganar un partido de 0,6. ¿Cuál es la probabilidad de que el equipo A gane la liga en 5 partidos?

Para plantear el problema en términos de variable aleatoria, tenemos que pensar a qué llamamos éxito y a qué llamamos fracaso, definir la variable aleatoria, y decidir cuál es el valor del que queremos calcular la probabilidad. Como la pregunta se plantea para el equipo A, que tiene una probabilidad de ganar un partido de 0,6, cada partido es un experimento independiente de Bernoulli con probabilidad de éxito $p = 0,6$, que vamos realizando uno tras otro. Si la liga se disputa al mejor de 7, quiere decir que la ganará el primero que gane 4. Por tanto, repetiremos el experimento de Bernoulli que hemos definido hasta tener 4 éxitos ($c = 4$). Como el suceso que nos interesa es que el equipo A gane la partida en $x + c = 5$ partidos, esto significará que habrá perdido $5 - 4 = 1$ partido (un fracaso). Si definimos la variable aleatoria

X : Número de partidos que pierde A antes de ganar el cuarto,
entonces

$X \sim BN(c = 4; p = 0,6)$,

y por tanto buscamos la probabilidad de que pierda solo uno es la probabilidad de que la variable aleatoria sea igual a uno:

$$P[X = 1] = \binom{4}{1} \cdot 0,6^4 \cdot (0,4)^1 \simeq 0,2074.$$



HOJA DE CÁLCULO

En hojas de cálculo de Google hay que quitar el último argumento de la fórmula.

=NEGBINOM.DIST(1;4;0,6;0)

R

La siguiente expresión obtiene la probabilidad pedida.

```
dnbinom(x = 1, size = 4, prob = 0.6)
#> [1] 0.20736
```

En nuestro ejemplo ilustrativo, se seleccionan sujetos al azar y de forma independiente. ¿Cuál es la probabilidad de que se necesiten más de 10 extracciones para que haya 4 mujeres?

El experimento de Bernoulli consiste en observar si un sujeto es mujer (éxito) u hombre (fracaso). Y se repite hasta que hayamos observado $c = 4$ mujeres. Entonces

- X : Número de fracasos en pruebas independientes de Bernoulli con probabilidad de éxito $1/2$ hasta el cuarto éxito
- $X \sim BN(4; 1/2)$

Nótese que aquí se está planteando la pregunta en términos de número total de experimentos, es decir, $x + c > 10$, y entonces buscamos $x > 10 - 4$:

$$P[X > 6] = 1 - P[X \leq 6] = 1 - \sum_{x=0}^6 P[X = x] =$$

$$= 1 - (0,0625 + 0,125 + 0,1563 + 0,1562 + 0,1367 + 0,1094 + 0,082) = 0,1719$$



HOJA DE CÁLCULO

=1-NEGBINOM.DIST(6;4;0,5;1)

En hojas de cálculo de Google no está el argumento para calcular acumulado, por lo que habría que calcular primero las probabilidades (desde cero hasta 6), sumar y restarlo de 1.

R

Con la siguiente expresión calculamos la probabilidad a través del complementario de la función de distribución.



```
pnbinom(q = 6, size = 4, prob = 0.5, lower.tail = FALSE)
#> [1] 0.171875
```

```
pnbinom(q = 6, size = 4, prob = 1/2, lower.tail = FALSE)
#> [1] 0.171875
```

```
qnbnom(p = 0.95, size = 4, prob = 1/2)
#> [1] 9
```

Un caso particular de la distribución binomial negativa cuando $c = 1$, es la **distribución geométrica**. Es decir, nos interesan el número de fracasos hasta obtener el primer éxito y entonces:

X : Número de fracasos hasta obtener el primer éxito en una serie de pruebas independientes de Bernoulli con probabilidad de éxito p :

$$X \sim Ge(p); 0 < p < 1,$$

cuya función de probabilidad se simplifica bastante, ya que solo hay una ordenación posible de los éxitos y fracasos:

$$P[X = x] = p \cdot (1 - p)^x; \quad x = 0, 1, \dots, \infty.$$

En la figura 7.4 la primera columna se corresponde con distribuciones geométricas. La media y varianza de una distribución geométrica son:

- Media: $\mu = E[X] = \frac{1-p}{p}$.
- Varianza: $\sigma^2 = V[X] = \frac{1-p}{p^2}$.

Observamos los sujetos que inician sesión en la página web del estudio, y nos interesa si es un investigador o no. ¿Cuál es la probabilidad de que se lleguen menos de 5 sujetos hasta que llega el primer investigador? ¿Cuál sería el número esperado de no investigadores hasta que llegue el primer investigador?

La probabilidad de éxito es $p = 4/52$, y el suceso que nos interesa se corresponde con $x + 1 < 5$. Entonces:

- X : Número de fracasos en pruebas independientes de Bernoulli con probabilidad de éxito $4/52$ hasta el primer éxito
- $X \sim Ge(4/52)$
- $P[X < 4] = P[X \leq 3] \simeq 0,0769 + 0,071 + 0,0655 + 0,0605 \simeq 0,2739$

A la segunda pregunta damos respuesta calculando la media:

$$\mu = \frac{1-p}{p} = \frac{1-(4/52)}{4/52} = 12,$$

Es decir, en promedio el primer directivo será el número 13 (ya que 12 es el número medio de no directivos)



HOJA DE CÁLCULO

No hay una fórmula específica para la distribución geométrica, pero podemos usar la de la binomial negativa con parámetro $c = 1$.

=NEGBINOM.DIST(3;1;4/52;1)

R

La siguiente expresión obtiene la probabilidad pedida.

```
pgeom(q = 3, prob = 4/52)
#> [1] 0.273975
```

7.2.5. Distribución hipergeométrica

La distribución hipergeométrica es el equivalente a la binomial cuando las pruebas de Bernoulli no son independientes. Se asemeja a los problemas de urnas con bolas blancas y negras, o aquellos en los que realizamos muestreos sin reposición. La variable aleatoria se define en los siguientes términos: tenemos una conjunto de N elementos (por ejemplo bolas) de los cuales M son de una determinada clase A (por ejemplo blancas). Por tanto, $N - M$ no son de la clase A (por ejemplo negras). Extraemos n elementos sin reposición de este conjunto, y lo que nos interesa es el número de elementos de la muestra que cumplen la característica. Entonces podemos definir la variable aleatoria:

X: Número de elementos de la clase A obtenidos en un muestreo sin reemplazo de tamaño n de un conjunto con N elementos totales de los que M son de dicha categoría A.

Que sigue una distribución geométrica de parámetros N , M y n .

$$X \sim HG(N; M; n); \quad N > M; \quad N \geq n.$$

La distribución hipergeométrica tiene la siguiente función de probabilidad:

$$P[X = x] = \frac{\binom{N-M}{n-x} \cdot \binom{M}{x}}{\binom{N}{n}}, \quad \max(0, n + M - N) \leq x \leq \min(M, n).$$

La figura 7.1 muestra gráficamente la distribución de probabilidad para varios valores de M y n y $N = 20$.

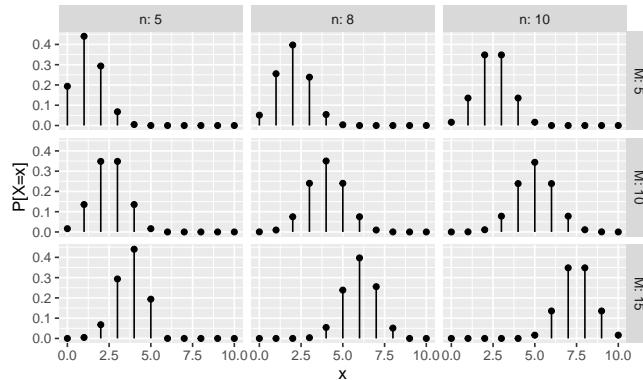


Figura 7.5: Representación de la función de probabilidad del modelo hipergeométrico

La media y la varianza de la distribución hipergeométrica son las siguientes:

- Media: $\mu = E[X] = M \cdot \frac{n}{N}$.

- Varianza: $\sigma^2 = V[X] = \frac{M \cdot (N-M) \cdot n \cdot (N-n)}{N^2 \cdot (N-1)}$.

Nótese que la distribución hipergeométrica no asume la independencia de los sucesivos experimentos. No obstante, es asintótica a la distribución binomial $\text{Bin}(n; p = \frac{M}{N})$ si p se mantiene estable. Se suele considerar apropiada la aproximación si $\frac{n}{N} < 0,1$.

En una comunidad de vecinos con 50 propietarios, 30 están de acuerdo en instalar un ascensor, y el resto no. En el descanso, cinco vecinos (al azar) se salen a fumar a la puerta. ¿Cuál es la probabilidad de que de esos cinco solo uno esté de acuerdo en instalar el ascensor?

Definimos la variable aleatoria:

X : Número de vecinos de esos cinco que están de acuerdo en instalar el ascensor. Entonces:

$$X \sim HG(N = 50; M = 30; n = 5),$$

y la probabilidad que buscamos es:



$$P[X = 1] = \frac{\binom{50-30}{5-1} \cdot \binom{30}{1}}{\binom{50}{5}} = \frac{4845 \cdot 30}{2118760} \simeq 0,0686.$$

HOJA DE CÁLCULO

[EXCEL] =DISTR.HIPERGEOM.N(1;5;30;50;0)

[LibreOffice] =HYPGEOM.DIST(1;5;30;50;0)

[Hojas de Cálculo de Google] =HYPGEOM.DIST(1;5;30;50)

R

La parametrización en R es ligeramente distinta, aunque obviamente equivalente, a la que hemos usado aquí, que se corresponde con la que aparece en la definición 2.48 de la norma ISO 3534-1. Además, utiliza los términos utilizados en problemas de urnas, de forma que los argumentos de la función son:

- x : El valor (*quantile*) para el cual hay que calcular la probabilidad.
- m : Número de bolas blancas (*white balls*), que se corresponde con nuestro parámetro M .
- n : Número de bolas negras (*black balls*), que se corresponde con $N - M$ según nuestra parametrización.
- k : Número de bolas extraídas, que se corresponde con nuestro parámetro n .



La siguiente expresión calcula la probabilidad del ejemplo.

Se asignan 10 premios a potenciales usuarios del servicio, pero no se pueden repetir ganadores. ¿Cuál es la probabilidad de que exactamente un directivo sea premiado?

Recordemos que teníamos 52 potenciales usuarios, de los cuales 4 eran directivos. Conocemos la composición exacta del conjunto, y es un muestreo sin reemplazamiento, por tanto la distribución adecuada es la hipergeométrica. Además, no podríamos usar la aproximación de la binomial, ya que $10/52 \not\leq 0,1$.

En términos de variable aleatoria, definimos:

X : Número de directivos en una muestra sin reemplazamiento de tamaño 10 realizada sobre un conjunto de 52 personas de las que 4 son directivos.

Entonces:

- $X \sim HG(N = 52; M = 4; n = 10)$
- $P[X = 1] \simeq 0,4240$


`dhyper(x = 1, m = 4, n = 52-4, k = 10)
#> [1] 0.4240465`

7.3. Modelos de distribución de probabilidad continuos

7.3.1. Introducción

En este apartado vamos a revisar algunas distribuciones de probabilidad continuas que tienen interés en ciencias e ingeniería. Al igual que en los modelos de distribución de probabilidad discretos, un conjunto de parámetros determinan completamente la distribución de probabilidad. Entonces tendremos la función de densidad, o la función de distribución, o ambas, en función de la variable x y también del conjunto de parámetros θ . Entonces, para valores concretos de los parámetros, podremos calcular probabilidades o determinar las características de la variable aleatoria en estudio. Para algunas distribuciones de probabilidad se han tabulado los valores de la función de distribución o su complementario, y tradicionalmente se han utilizado estas tablas para resolver problemas de probabilidad. Actualmente se pueden realizar los cálculos con el uso de software. Por tanto, seguiremos utilizando la notación vista en el apartado 7.2 para indicar la distribución de probabilidad continua que sigue la variable aleatoria X :

$$X \sim \mathcal{Distr}(\theta),$$

donde \mathcal{Distr} identifica el modelo de distribución de probabilidad, y θ es el vector de parámetros. Entonces las expresiones de la función de densidad y de distribución contendrán los parámetros: $f(x|\theta)$, $F(x|\theta)$.

En este capítulo veremos con detalle las distribuciones uniforme, exponencial y normal. Existen otros modelos de distribución de probabilidad continuos univariantes y multivariantes que se mencionan al final del capítulo.

7.3.2. Distribución uniforme

La distribución uniforme se caracteriza por tener una densidad constante en un intervalo $[a, b]$. Si una variable aleatoria X sigue una distribución uniforme en el intervalo entre a y b lo expresamos así:

$$X \sim U(a; b); \quad a < b; \quad a, b \in \mathbb{R}.$$

La función de densidad de una variable aleatoria continua que sigue un modelo uniforme tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{resto} \end{cases}$$

y la función de distribución se obtiene fácilmente a partir de esta:

$$F(x) = \int_a^x \frac{1}{b-a} dt = \left[\frac{t}{b-a} \right]_a^x = \frac{x}{b-a} - \frac{a}{b-a} = \frac{x-a}{b-a},$$

quedando en su forma completa como:

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x < b \\ 1 & \text{si } x \geq b \end{cases}$$

La media y la varianza de una variable aleatoria uniforme se deducen fácilmente a partir de su función de densidad:

- Media: $\mu = E[X] = \frac{a+b}{2}$.
- Varianza: $\sigma^2 = V[X] = \frac{(b-a)^2}{12}$.

El modelo de distribución uniforme es muy útil para simular probabilidades y variables aleatorias a través de la $U(0; 1)$. También se suele utilizar cuando conocemos el rango de valores pero no tenemos información sobre cuáles de esos valores son más probables. La figura 7.6 muestra la representación de las funciones de densidad y distribución de una variable aleatoria que sigue una distribución continua uniforme.

```
#>
#> Attaching package: 'gridExtra'
#> The following object is masked from 'package:dplyr':
```

```
#>  
#>     combine
```

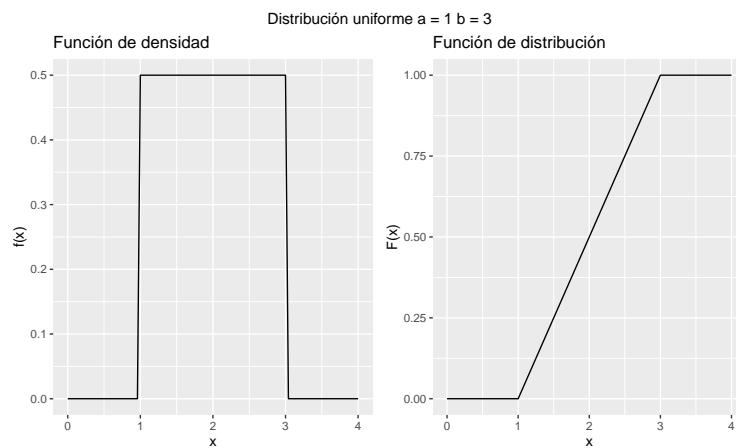


Figura 7.6: Representación gráfica de las funciones de densidad y distribución de una variable aleatoria uniforme

El volumen anual de ventas de un almacén se distribuye uniformemente entre 380 y 1200 miles de euros. ¿Cuál es la probabilidad de que las ventas sean superiores a 1000 miles de euros? ¿Cuáles son las ventas esperadas en un año?

Definimos la variable aleatoria:

X : ventas del almacén un año $X \sim U(380; 1200)$

Entonces la función de densidad es:

$$f(x) = \frac{1}{1200 - 380}, \quad 380 < x < 1200,$$

la probabilidad pedida:

$$P[X > 1000] = \int_{1000}^{1200} \frac{1}{820} dx = \frac{200}{820} \simeq 0,2439.$$

Pero también podemos calcularla más fácilmente utilizando la función de distribución, que conocemos:

$$P[X > 1000] = 1 - P[X \leq 1000] = 1 - F(1000) = 1 - \frac{1000 - 380}{1200 - 380} \simeq 1 - 0,7561 \simeq 0,2439.$$

y las ventas esperadas son la media de la variable aleatoria:

$$\mu = E[X] = \frac{380 + 1200}{2} = 790 \text{ miles de euros.}$$

La figura 7.7 representa la función de densidad y la probabilidad pedida como área bajo la curva.

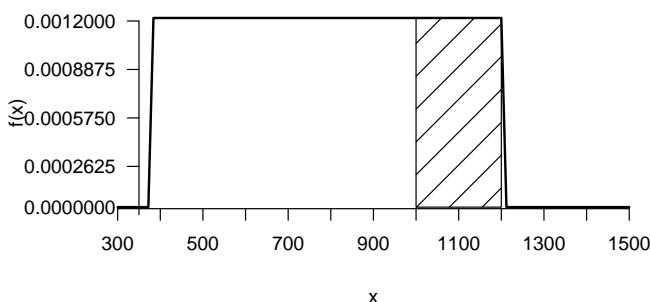


Figura 7.7: Ejemplo distribución uniforme

HOJA DE CÁLCULO

No hay funciones específicas para obtener la probabilidad de una variable aleatoria uniforme, aunque se puede insertar una fórmula con la función de distribución y a partir de ahí calcular probabilidades, por ejemplo, si en la celda A1 tenemos el valor 1000, en la celda A2 el parámetro $a = 380$ y en la celda A3 el parámetro $b = 1200$, entonces en otra celda podemos calcular la probabilidad del ejemplo como:

$$= 1 - (A1 - A2)/(A3 - A2)$$

R

 La función `punif` devuelve la función de distribución uniforme.

```
punif(q = 1000, min = 380, max = 1200, lower.tail = FALSE)
#> [1] 0.2439024
```

Si la proporción de video visualizado por un sujeto que sigue el mensaje se distribuye de forma uniforme, ¿cuál es la probabilidad de que un visitante de la web del estudio vea más del 90 % del vídeo?

En términos de variable aleatoria:

X : Proporción de video visualizado, $X \sim U(0; 1)$.

Entonces:

$$P[X > 0,9] = \int_{0,9}^1 dx = 0,1.$$

O bien:



$$P[X > 0,9] = 1 - F(0,9) = 1 - \frac{0,9 - 0}{1 - 0} = 0,1.$$

R

Análogamente al ejemplo anterior:

```
punif(q = 0.9, min = 0, max = 1, lower.tail = FALSE)
#> [1] 0.1
```

7.3.3. Distribución exponencial

Cuando en un proceso de Poisson observamos el tiempo que transcurre entre un evento y otro, aparece la distribución exponencial. También modeliza bien tiempos de vida, por ejemplo de componentes electrónicos. La distribución exponencial solo tiene un parámetro:

$$X \sim \text{Exp}(\beta), \beta > 0.$$

El parámetro β del modelo de distribución exponencial representa, al igual que en la distribución de Poisson, la tasa media de eventos por unidad de tiempo. Una variable aleatoria que sigue un modelo de distribución exponencial tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \beta e^{-\beta x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

La función de distribución se obtiene fácilmente a partir de la función de densidad:

$$F(x) = \int_{-\infty}^x f(t)dt = 1 - e^{-\beta x}, \quad x > 0.$$

La figura 7.8 muestra la representación de las funciones de densidad y distribución de una variable aleatoria que sigue una distribución continua exponencial.

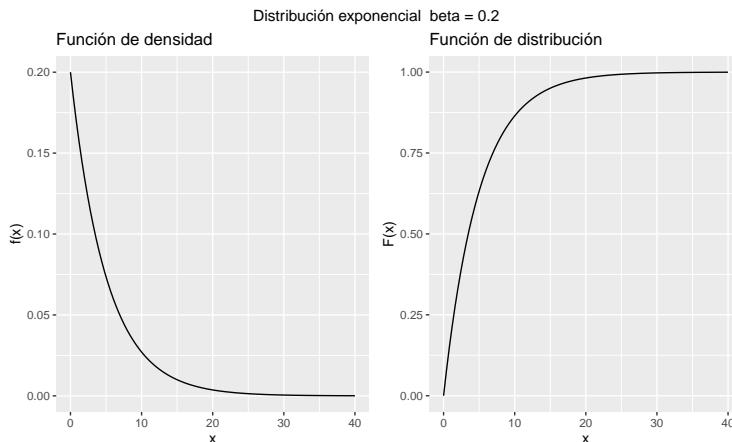


Figura 7.8: Representación gráfica de las funciones de densidad y distribución de una variable aleatoria exponencial

La media y la varianza de una variable aleatoria que sigue el modelo exponencial son:

- Media: $\mu = E[X] = \frac{1}{\beta}$.
- Varianza: $V[X] = \frac{1}{\beta^2}$

Se dice que la exponencial es una variable aleatoria *sin memoria*, en el sentido de que el tiempo que haya tardado en ocurrir un evento, es independiente de cuándo sucedió el anterior:

$$(P[X > t_2 + t_1 | X > t_1] = P[X > t_2]).$$

La distribución exponencial es un caso particular de la distribución gamma, que no vemos en este texto. La distribución gamma modeliza el tiempo transcurrido hasta ocurrir un número determinado de eventos.

El tiempo en horas que se tarda en arreglar una máquina sigue una distribución exponencial de parámetro $\beta = 4$. ¿Cuál es la probabilidad de que una avería tarde más de una hora en ser reparada?

$$X \sim \text{Exp}(4),$$

$$P[X > 1] = 1 - \int_0^1 4e^{-4x} dx = 1 - [-e^{-4x}]_0^1 = 1 - (-e^{-4} - (-e^0)) = e^{-4} \simeq 0,0183.$$

Es más sencillo si lo resolvemos con la función de distribución:

$$P[X > 1] = 1 - F(1) = 1 - (1 - e^{-4 \cdot 1}) = \simeq 0,0183.$$

La figura 7.9 muestra la representación gráfica de la función de densidad del ejemplo.



HOJA DE CÁLCULO

[EXCEL] =1-DISTR.EXP.N(1; 4; 1)

[LibreOffice] =1-EXPON.DIST(1;4;1)

R

La función `pexp` obtiene la función de distribución del modelo exponencial.

```
pexp(q = 1, rate = 4, lower.tail = FALSE)
#> [1] 0.01831564
```

En ocasiones nos interesa calcular la inversa de la función de distribución. Es decir, encontrar un valor de la variable aleatoria para el cual se cumple alguna condición de probabilidad, como en el siguiente ejemplo.

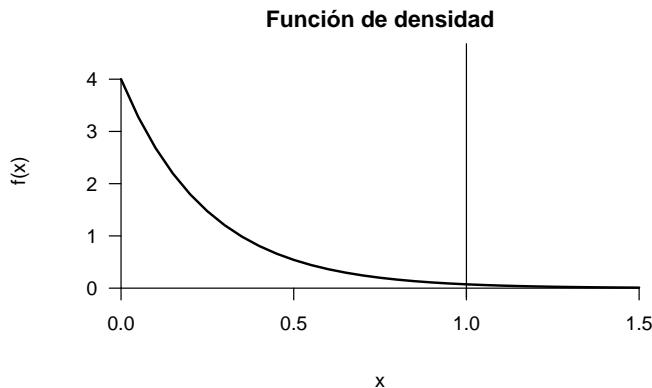


Figura 7.9: Representación de la función de densidad del modelo exponencial del ejemplo

El tiempo que permanece un visitante en la web del estudio sigue una distribución exponencial. La tasa media de abandonos es de 2 cada minuto. ¿Cuánto tiempo permanece como máximo el 95 % de los usuarios antes de abandonar?

En términos de variable aleatoria:

X : Tiempo hasta abandonar la web tras hacer clic en el mensaje,
 $X \sim \text{Exp}(2)$.

En este caso, lo que nos interesa es obtener el cuantil 0.95, es decir, el valor x de la variable aleatoria para el cual $P[X > x] = 0,05$, o lo que es lo mismo, $P[X \leq x] = 0,95$. como tenemos la expresión de la función de distribución, no hay más que despejar y tenemos:

$$F(x) = 0,95 \Leftrightarrow 1 - e^{-2x} = 0,95 \Leftrightarrow x = 1,498 \text{ minutos.}$$

También nos podemos preguntar cuánto tiempo permanece un visitante, en promedio, en la web. Entonces calculamos la esperanza:

$$\mu = \frac{1}{\beta} = 0,5$$



R

La función `qexp` obtiene la inversa de la función de distribución del modelo exponencial.

```
qexp(p = 0.95, rate = 2)
#> [1] 1.497866
```

7.3.4. Distribución normal

Sin duda, la distribución normal (o gaussiana) es el modelo de distribución de probabilidad continuo más importante de todos. Gracias al teorema central del límite que veremos en el capítulo 7.5, muchas situaciones se aproximan a la distribución normal¹. El modelo de distribución normal queda determinado por dos parámetros, que son su media μ y su desviación típica σ :

$$X \sim N(\mu; \sigma); \mu \in \mathbb{R}, \sigma > 0.$$

La función de densidad de una variable aleatoria que sigue el modelo de distribución normal tiene la siguiente función de densidad:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty.$$

La figura 7.10 muestra la función de densidad y la función de distribución para unos valores determinados de σ y μ . La función de distribución se ha obtenido por métodos numéricos, ya que no es posible obtener una expresión analítica de $F(x)$ al no existir una primitiva de $f(x)$.

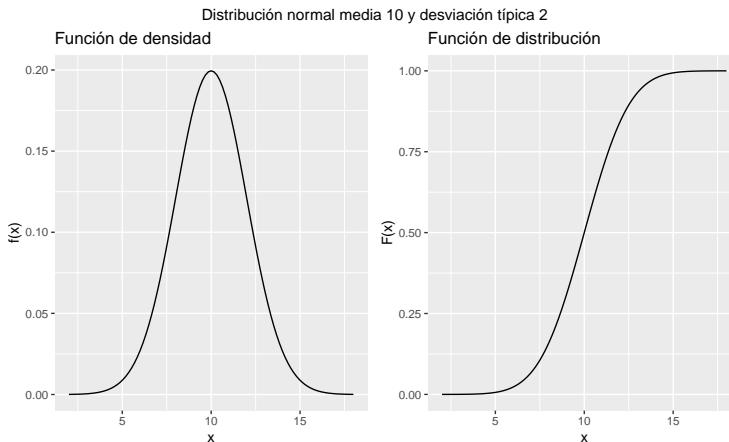


Figura 7.10: Representación gráfica de las funciones de densidad y distribución de una variable aleatoria normal

La distribución normal es simétrica respecto de la media, siendo la mediana y la moda igual a ella. Esta importante propiedad implica que $P[X \leq \mu] = 0,5$. Cuanto más cerca de la media estén los valores, más probables son, y a medida

¹Básicamente implica que la suma de muchas variables aleatorias, tengan la forma que tengan, seguirá una distribución normal. En muchas situaciones, la variable aleatoria será realmente la suma de muchas características y circunstancias, y por eso se distribuyen con el modelo de distribución normal.

que nos alejamos de la media, cada vez son más improbables, de hecho como vemos en la figura 7.11 entre la media y dos desviaciones típicas tenemos más del 95 % de la probabilidad, y la probabilidad de que la variable aleatoria tome valores más allá de tres desviaciones típicas desde la media es de solo 0.0027. La función de densidad presenta puntos de inflexión en $\mu \pm \sigma$.

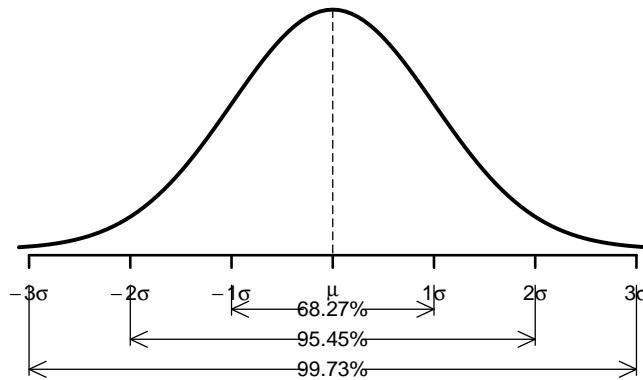


Figura 7.11: Función de densidad de la distribución normal

El modelo de distribución normal cumple la propiedad aditiva, de modo que si tenemos las variables aleatorias:

$$X_j \sim N(\mu_j; \sigma_j) \quad \forall j = 1, \dots, n,$$

entonces la variable aleatoria:

$$Y = a + \sum_{j=1}^n b_j X_j,$$

no siendo todos los b_j nulos, se distribuye también como una distribución normal, y por tanto por las propiedades de la esperanza y la varianza que vimos en el capítulo 5:

$$Y \sim N\left(a + \sum_{j=1}^n b_j \mu_j; \sqrt{\sum_{j=1}^n b_j^2 \sigma_j^2}\right).$$

Un caso particular del modelo de distribución normal, es la **distribución normal estándar**, cuyos parámetros serán $\mu = 0$ y $\sigma = 1$, y que vamos a representar en este texto como Z^2 :

$$Z \sim N(0; 1).$$

²En otros textos la encontramos representada por la letra griega ϕ .

La función de densidad en este caso quedaría:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty.$$

Nótese que, al ser la distribución normal simétrica, se cumple que $P[Z \leq 0] = 0,5$.

Trabajar con variables aleatorias estandarizadas es conveniente en muchas situaciones. En particular, se ha utilizado tradicionalmente para obtener probabilidades por medio de tablas estadísticas que contienen probabilidades de la distribución normal estandarizada, bien la función de distribución $F(z) = P[Z \leq z]$ o su complementario $P[Z > z]$. A través de estas tablas podemos hacer cálculo de probabilidades para cualquier variable aleatoria normal, con cualesquiera μ y σ , ya que se cumple, según la aditividad y las propiedades de la esperanza y la varianza:

$$X \sim N(\mu; \sigma) \implies Z = \frac{X - \mu}{\sigma} \sim N(0; 1).$$

Ya vimos en el capítulo 5 que podemos estandarizar cualquier variable aleatoria. Si estandarizamos una distribución normal con cualesquiera parámetros μ y σ , entonces tendremos variables aleatorias *estandarizadas*.

A la hora de calcular probabilidades de la distribución normal, nos encontramos que la función de densidad no es integrable, es decir, no podemos encontrar una primitiva. Entonces, en vez de utilizar integrales se utilizan métodos numéricos o tablas como se ha descrito anteriormente.

El procedimiento para calcular probabilidades de variables aleatorias que siguen el modelo de distribución normal es el siguiente:

1. Determinar los parámetros de la distribución μ y σ (para el alcance de este capítulo, vendrán dados).
2. Tipificar el/los valores de la variable X para los que se quiere calcular la probabilidad ($X \rightarrow Z$).
3. Utilizando las propiedades de la probabilidad, transformar la expresión de la probabilidad que se quiere calcular en expresiones compatibles con la tabla a utilizar, por ejemplo $P[Z \leq z]$
4. Buscar dentro de la tabla las probabilidades que se necesiten para los valores z y hacer los cálculos.

Para la operación inversa del cálculo de cuantiles a partir de una probabilidad, procedemos de la siguiente forma:

1. Tipificar la variable aleatoria, obteniendo una expresión $z = \frac{x-\mu}{\sigma}$, donde x es el valor que queremos encontrar.

2. Expresar la probabilidad en forma compatible con la tabla a utilizar, por ejemplo $P[Z \leq \frac{x-\mu}{\sigma}] = p$.
3. Buscar dentro la tabla la probabilidad deseada p .
4. Encontrar el valor de z que se corresponde con dicha probabilidad, y despejar x de la expresión $z = \frac{x-\mu}{\sigma}$.

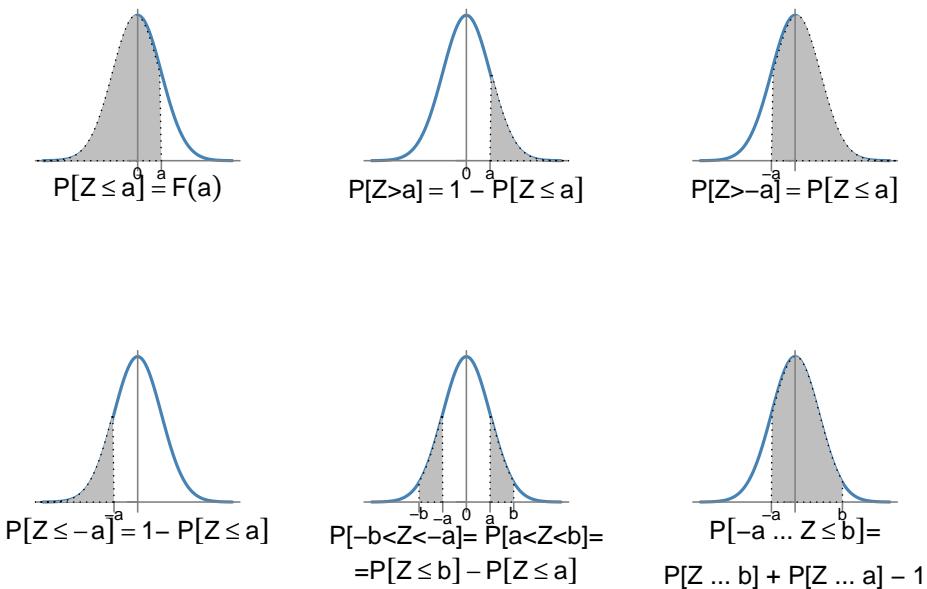
En lo que sigue, utilizaremos la tabla de la cola inferior de la distribución normal estandarizada, disponible en el apéndice B. En esta tabla tenemos, para valores de $z > 0$, $P[Z \leq z]$. Con estos valores, seremos capaces de calcular cualquier probabilidad utilizando las siguientes propiedades y gracias a la simetría de la distribución. Dados $a < b$ positivos, debemos expresar cualquier probabilidad de forma que podamos buscarla en la tabla:

- En la tabla tenemos $P[Z \leq b]$ o $P[Z \leq a]$.
- $P[Z > a] = P[Z \leq -a] = 1 - P[Z \leq a]$.
- $P[Z > -a] = P[Z \leq a]$.
- $P[-b \leq Z \leq -a] = P[a \leq Z \leq b] = P[Z \leq b] - P[Z \leq a]$.
- $P[-a \leq Z \leq b] = P[Z \leq b] + P[Z \leq a] - 1$.

La figura 7.12 resume estos cálculos. Ayudará al lector pensar en la probabilidad en términos de área bajo la curva de la función de densidad, teniendo en cuenta que el área total debe ser igual a la unidad, y que el área por encima y por debajo de cero es 0,5. La misma lógica se aplicaría en el caso de utilizar una tabla con la cola superior que podamos encontrar en alguna otra bibliografía.

```
#> Warning in mtext(expression("P" * group("[", "-a Z" <= b,
#> "]) * "="), : font metrics unknown for Unicode character
#> U+22264
#> Warning in mtext(expression("P" * group("[", "-a Z"
#> <= b, "]") * "="), : conversion failure on '-a Z' in
#> 'mbcsToSbcs': dot substituted for <e2>
#> Warning in mtext(expression("P" * group("[", "-a Z"
#> <= b, "]") * "="), : conversion failure on '-a Z' in
#> 'mbcsToSbcs': dot substituted for <89>
#> Warning in mtext(expression("P" * group("[", "-a Z"
#> <= b, "]") * "="), : conversion failure on '-a Z' in
#> 'mbcsToSbcs': dot substituted for <a4>
#> Warning in mtext(expression("P" * group("[", "-a Z" <= b,
#> "]) * "="), : font metrics unknown for Unicode character
#> U+22264
#> Warning in mtext(expression("P" * group("[", "-a Z"
#> <= b, "]") * "="), : conversion failure on '-a Z' in
#> 'mbcsToSbcs': dot substituted for <e2>
#> Warning in mtext(expression("P" * group("[", "-a Z"
#> <= b, "]") * "="), : conversion failure on '-a Z' in
```

```
#> 'mbcsToSbcs': dot substituted for <89>
#> Warning in mtext(expression("P" * group("[", "-a Z"
#> <= b, "]") * "="), : conversion failure on '-a Z' in
#> 'mbcsToSbcs': dot substituted for <a4>
#> Warning in mtext(expression("P" * group("[", "-a Z"
#> <= b, "]") * "="), : conversion failure on '-a Z' in
#> 'mbcsToSbcs': dot substituted for <e2>
#> Warning in mtext(expression("P" * group("[", "-a Z"
#> <= b, "]") * "="), : conversion failure on '-a Z' in
#> 'mbcsToSbcs': dot substituted for <89>
#> Warning in mtext(expression("P" * group("[", "-a Z"
#> <= b, "]") * "="), : conversion failure on '-a Z' in
#> 'mbcsToSbcs': dot substituted for <a4>
#> Warning in mtext("P[Z b] + P[Z a] - 1", 1, 3):
#> conversion failure on 'P[Z b] + P[Z a] - 1' in
#> 'mbcsToSbcs': dot substituted for <e2>
#> Warning in mtext("P[Z b] + P[Z a] - 1", 1, 3):
#> conversion failure on 'P[Z b] + P[Z a] - 1' in
#> 'mbcsToSbcs': dot substituted for <89>
#> Warning in mtext("P[Z b] + P[Z a] - 1", 1, 3):
#> conversion failure on 'P[Z b] + P[Z a] - 1' in
#> 'mbcsToSbcs': dot substituted for <a4>
#> Warning in mtext("P[Z b] + P[Z a] - 1", 1, 3):
#> conversion failure on 'P[Z b] + P[Z a] - 1' in
#> 'mbcsToSbcs': dot substituted for <e2>
#> Warning in mtext("P[Z b] + P[Z a] - 1", 1, 3):
#> conversion failure on 'P[Z b] + P[Z a] - 1' in
#> 'mbcsToSbcs': dot substituted for <89>
#> Warning in mtext("P[Z b] + P[Z a] - 1", 1, 3):
#> conversion failure on 'P[Z b] + P[Z a] - 1' in
#> 'mbcsToSbcs': dot substituted for <a4>
```

Figura 7.12: Cálculo de probabilidades de la distribución $N(0; 1)$

En un curso de reciclaje dirigido a teleoperadores las puntuaciones obtenidas en el test final se distribuyen siguiendo un modelo normal de media 5 y desviación típica 2. Con menos de tres puntos un teleoperador no promociona. ¿Cuál es la probabilidad de que un teleoperador no promocione? ¿Cuál es la puntuación mínima que han obtenido el 3% de los teleoperadores mejor preparados?

La variable aleatoria es:

X : Calificación obtenida por el teleoperador, $\sim N(5; 2)$,
y lo que buscamos es la probabilidad de obtener menos de tres puntos:

$$P[X < 3] = P\left[\frac{X - 5}{2} < \frac{3 - 5}{2}\right] = P[Z < -1] = 1 - P[Z \leq 1] = \boxed{0,1587}.$$

A la segunda pregunta contestamos de manera inversa. Tenemos una probabilidad $p = 0,03$, y buscamos el valor x de la variable que cumple:

$$P[X \leq x] = 1 - 0,03,$$

o lo que es lo mismo:

$$P\left[Z \leq \frac{x - \mu}{\sigma}\right] = 0,97,$$

Buscamos esta probabilidad en el interior de la tabla^a, en este caso el valor más próximo redondeando a dos decimales es 0,9699, que se corresponde con un valor $z = 1,88$. Entonces tenemos:

$$z = \frac{x - \mu}{\sigma} \Leftrightarrow 1,88 = \frac{x - 5}{2} \Leftrightarrow x = 1,88 \cdot 2 + 5 = \boxed{8,76},$$

Es decir,

$$P[X > 8,76] \simeq 0,03.$$

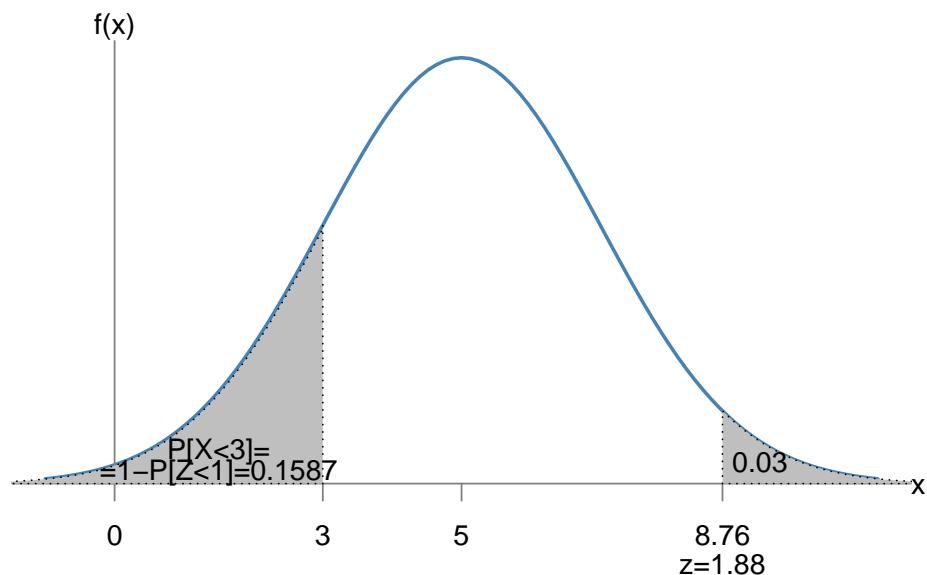


Figura 7.13: Ejemplo de cálculo de probabilidad y cuantil de la normal

Al utilizar software, no es necesario estandarizar. Le pasaremos directamente los parámetros de la distribución normal a la función correspondiente.

HOJA DE CÁLCULO

[LibreOffice] =NORM.DIST(3;5;2;1)

[EXCEL] =DISTR.NORM.N(3;5;2;1)

Para obtener el cuantil, tenemos que pasar como argumento de probabilidad 1-0.03, ya que siempre da la cola inferior.

[LibreOffice] =NORM.INV(0,97;5;2)

[EXCEL] =INV.NORM(0,97;5;2)

R

Con la función `pnorm` calculamos la probabilidad, y con la función `qnorm`, el cuantil.

```
pnorm(q = 3, mean = 5, sd = 2)
#> [1] 0.1586553
qnorm(p = 0.03, mean = 5, sd = 2, lower.tail = FALSE)
#> [1] 8.761587
```

El peso de los paquetes que contienen los pedidos que recibe un laboratorio se distribuye según una distribución normal de media 1,8 y desviación típica 0,5 kg. ¿Cuál es la probabilidad de que un paquete esté entre 1 y 2 kilos?

Definimos la variable aleatoria:

$X : \text{Peso de los paquetes}, X \sim N(1,8, 0,5)$.

Entonces:

$$\begin{aligned} P[1 \leq X \leq 2] &= P\left[\frac{1-1,8}{0,5} \leq \frac{X-\mu}{\sigma} \leq \frac{2-1,8}{0,5}\right] = \\ &= P[-1,6 \leq Z \leq 0,4] = P[Z \leq 4] - P[Z \leq -1,6] = \\ P[Z \leq 4] - (1 - P[Z \leq 1,6]) &= 0,6554 + 0,9452 - 1 = \boxed{0,6006}. \end{aligned}$$

¿Por debajo de qué peso estarán probablemente al menos el 95 % de los paquetes?

Buscamos el valor de x que cumpla:

$$P[X < x] = 0,95$$

Buscamos esta probabilidad en el interior de la tabla, y hay dos valores que nos servirían si redondeamos a dos decimales: 0,9495, correspondiente a $z = 1,64$ y 0,9505, correspondiente a $z = 1,65$. Vamos a tomar este último para asegurarnos la probabilidad de 0,95^a. Solo nos queda igualar este valor a la x estandarizada y depear:

$$z = \frac{x - \mu}{\sigma} \iff 1,65 = \frac{x - 1,8}{0,5} \iff x = 1,65 \cdot 0,5 + 1,8 = 2,625.$$

Entonces, el 95 % de los paquetes pesan más de 2,625 kg.



^asi tomáramos 1.64, los valores que están por debajo serían el 94.95 %

Resolvemos de forma análoga al ejemplo anterior. Nótese cómo ahora calculamos el cuantil exacto para la probabilidad de 0.95. Como las funciones nos dan la función de distribución, aplicamos que $P[a \leq x < b] = F(b) - F(a)$.

HOJA DE CÁLCULO

[LibreOffice] NORM.DIST(1;1,8;0,5;1) [LibreOffice] =NORM.INV(0,95;1,8;0,5) [EXCEL] =DISTR.NORM.N(2;1,8;0,5;1) [EXCEL] =INV.NORM(0,95;1,8;0,5)	=NORM.DIST(2;1,8;0,5;1) - NORM.DIST(1;1,8;0,5;1) =DISTR.NORM.N(2;1,8;0,5;1) - =DISTR.NORM.N(1;1,8;0,5;1)
--	---



R

```
pnorm(2, 1.8, 0.5) - pnorm(1, 1.8, 0.5)
#> [1] 0.6006224
qnorm(p = 0.95, 1.8, 0.5)
#> [1] 2.622427
```

7.3.5. Mezcla de poblaciones y adición de variables aleatorias

Vamos a ilustrar con un ejemplo más completo la propiedad de la **aditividad** de variables aleatorias normales. Es importante no confundir la aditividad con la **mezcla** de poblaciones. En ambos casos el problema al que nos enfrentamos puede estar referido a una característica que se observa en dos grupos, y a veces es difícil distinguir si tenemos que resolverlo mediante la probabilidad total y el teorema de Bayes, o mediante la suma de variables aleatorias. Para diferenciarlo, debemos entender bien el planteamiento del problema. Algunos indicios que nos ayudarán son:

- Mezcla de poblaciones: Hay dos o más grupos en los que se observan elementos tomados al azar. La característica tiene distinta distribución de probabilidad en cada grupo, pero la probabilidad de interés se refiere a las poblaciones mezcladas (probabilidad total) o a la probabilidad de pertenecer a uno de los grupos, condicionado a que se ha producido algún evento de interés.
- Suma de variables aleatorias: Hay dos o más variables aleatorias (que se pueden referir a grupos distintos, y de ahí la posible confusión con la mezcla de poblaciones). Pero lo que nos interesa es estudiar la variable aleatoria que resulta de hacer operaciones con esas variables aleatorias (por ejemplo, sumarlas).

En el siguiente ejemplo se plantean preguntas que abordan los dos problemas.

Una empresa de comercio minorista tiene tres tiendas (A, B y C) en una determinada ciudad. El tiempo que se tarda en atender a un cliente se distribuye según una distribución exponencial de media 2 minutos, 4 minutos y 5 minutos en las tiendas A, B y C respectivamente. La tienda C atiende a tantos clientes como A y B juntas (que atienden al mismo número de clientes). Si llamamos T_A , T_B , T_C a las variables aleatorias “tiempo en ser atendido en la tienda A, B, o C” respectivamente, entonces:

$$T_A \sim Exp(0,5),$$

$$T_B \sim Exp(0,25),$$

$$T_C \sim Exp(0,2).$$

Se considera que un cliente estará insatisfecho si se tarda más de 8 minutos en atenderle.

Por otra parte, las ventas diarias de cada tienda, V_A , V_B y V_C , son independientes, y se distribuyen según una distribución normal con los siguientes parámetros en miles de unidades monetarias (u.m.):

$$V_A \sim N(\mu = 100; \sigma = 10),$$

$$V_B \sim N(\mu = 150; \sigma = 20),$$

$$V_C \sim N(\mu = 140; \sigma = 40).$$

Cuestión 1:

- a) ¿Cuál es la probabilidad de que un cliente de la empresa no esté satisfecho con el tiempo de servicio?
- b) Recibimos una queja de un cliente insatisfecho con el tiempo de servicio. ¿Cuál es la probabilidad de que sea un cliente de la tienda A?

Cuestión 2:

Las tiendas A y B son propiedad 100 % de la empresa. Pero de la tienda C la empresa realmente solo recauda el 50 %, ya que el otro 50 % es de otro socio. Por otra parte, la empresa recibe unos ingresos fijos de 25.000 u.m. diarios de una tienda franquiciada en otra ciudad.

- a) ¿Qué distribución de probabilidad siguen las ventas totales de la empresa, teniendo en cuenta su participación en las tiendas?
- b) ¿Cuál es la probabilidad de que un día cualquiera esas ventas totales sea de menos de 300.000 u.m.?

Para resolver cada cuestión, tenemos que pensar si estamos ante una mezcla de poblaciones, o una suma de variables. Al estar los dos problemas planteados, es fácil de ver. Pero si solamente nos estuvieran preguntando por una de las dos cosas, pueden surgir dudas.

La primera cuestión es un típico problema de probabilidad total y Teorema de Bayes en el que tenemos una partición del espacio muestral en tres tiendas, y conocemos las probabilidades *a priori*. En cuanto al suceso de interés (cliente insatisfecho), conocemos las distribuciones de probabilidad de cada tienda, y tendremos que calcular las probabilidades condicionadas a cada tienda.

En la segunda cuestión lo que tenemos es una combinación lineal de variables aleatorias, porque las ventas totales serán la suma de las ventas de las tiendas. Además, una de las variables estará multiplicada por un coeficiente, y tenemos también una constante que sumar.

Pasemos entonces a resolver cada cuestión.



Las probabilidades de este ejemplo se resuelven de forma análoga a los anteriores. Se deja como ejercicio para el lector comprobar por sí mismo los resultados ofrecidos a través del programa de su elección.

7.4. Otros modelos de distribución de probabilidad

Los modelos vistos en este capítulo y el anterior cubren la mayoría de los problemas cotidianos de modelización. Existen otros modelos de distribución que se aplican a problemas específicos. Para finalizar este capítulo, se proporciona una breve descripción de las que aparecen en la norma ISO 3534-1.

- **Distribución multinomial.** Es el equivalente multivariante a la distribución binomial, donde no solamente hay dos resultados posibles sino más de dos. Entonces tenemos un vector aleatorio con tantas componentes como clases posibles (resultados del experimento). Cada componente del vector aleatorio sigue una distribución binomial.
- **Distribución lognormal.** Una variable lognormal, al transformarla mediante el logaritmo será una normal.
- **La distribución Gamma.** Ya se ha comentado que es una generalización de la distribución exponencial, y modeliza el tiempo hasta k eventos
- **La distribución Beta.** Es muy útil para modelizar proporciones y probabilidades.
- **La distribución de Weibull.** También se utiliza para modelizar tiempos de vida, y es muy flexible describir formas muy diferentes de la distribución mediante el ajuste de sus parámetros. Es también una distribución de valores extremos (tipo III). La norma incluye otras dos distribuciones de valores extremos: Tipo I (Gumbel) y Tipo II (Fréchet).
- **La distribución normal multivariante.** Se aplica a vectores aleatorios donde todas sus componentes son variables aleatorias normales.
- **La distribución multinomial.** Se aplica a características cualitativas multiclas.

En el apéndice B se puede encontrar un resumen de todas las distribuciones de probabilidad y sus principales características.

7.5. Convergencia de variables aleatorias

7.6. Distribuciones relacionadas con la normal

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 dos grupos

En preparación.

Comparación de atributos

Comparación de dos grupos

Comparación de más de dos grupos : remitir a ANOVA

Capítulo 10

Análisis de la Varianza

10.1. Introducción

El análisis de la varianza es una técnica estadística de análisis de dependencias, donde se busca explicar una o varias variables cuantitativas a partir de una o varias variables cualitativas o factores. Es decir, buscamos un modelo del tipo:

$$\mathbf{Y} = f(\mathbf{X}) + \varepsilon \quad (10.1)$$

donde \mathbf{Y} es un vector de variables respuesta numéricas que queremos explicar o predecir, y \mathbf{X} es un vector de variables predictivas cualitativas con las que pretendemos explicar las variables respuesta. Como todo modelo estadístico, está sujeto a un error ε .

En realidad el análisis de la varianza incluye un conjunto amplio de técnicas cuyo análisis varía ligeramente según la naturaleza y el número de variables en los vectores aleatorios \mathbf{Y} y \mathbf{X} . En los apartados siguientes se irán detallando los distintos modelos desde el más sencillo al más complejo. A medida que se avance en los modelos, se presentan más ejemplos y menos teoría, ya que el fundamento es muy similar y se puede consultar en la bibliografía citada.

La técnica del análisis de la varianza se puede abordar desde dos perspectivas: explicativa y predictiva. Desde una perspectiva explicativa, se puede aplicar la técnica para realizar estudios observacionales a datos ya existentes. Estos estudios observacionales confirmarán la **relación** entre las variables. Desde el punto de vista predictivo, se pueden diseñar experimentos antes de la recogida de datos. Estos estudios predictivos permiten confirmar la **relación de causa-efecto** entre las variables.

10.2. Análisis de la varianza de un factor

El caso más sencillo que podemos aplicar al modelo general (10.1), es aquel en el que tenemos una única variable continua en el vector aleatorio \mathbf{Y} y una sola variable cualitativa (factor) en el vector aleatorio \mathbf{X} con $k > 2$ niveles. Cuando la variable cualitativa tiene solamente dos posibles niveles, podemos utilizar simplemente contrastes de hipótesis para la comparación de poblaciones mediante test paramétricos como el de la t de Student, o no paramétricos como el test de Wilcoxon, o el test de Wilcoxon-Mann-Whitney.

 Una granja experimental quiere estudiar el efecto que tiene el tipo de fertilizante utilizado en el cultivo de una determinada variedad de plantas y su peso en su punto óptimo de recolección. Para ello diseña un experimento en el que selecciona doce semillas aleatoriamente de un determinado lote. Se asigna aleatoriamente cada semilla a una maceta. Y a cada maceta, se le asigna también aleatoriamente un tipo de fertilizante entre tres varieaddes: A, B y C. El peso de cada planta en gramos se recoge en la tabla 10.1, que están guardados en el data frame `danovaa`. Vamos a utilizar este ejemplo a lo largo de este capítulo.

^aSe puede descargar el conjunto de datos de <http://emilio.lcano.com/b/adr/datos/danova2.rds>

 El primer paso en toda técnica estadística es hacer un análisis exploratorio. Como son muy pocos puntos por cada grupo, vamos a obtener un resumen numérico y a representarlos todos con un gráfico de puntos (figura 10.1).

A la vista de las medias parece que el peso medio con el fertilizante C es menor. Por otra parte parece que hay menos variabilidad con también con el fertilizante C. Una vez ajustado el modelo lo comprobaremos numéricamente.



```
library(tidyverse)
danova |>
  group_by(Fertilizante) |>
  summarise(Peso_medio = mean(Peso),
            Desv.Tipica = sd(Peso))
#> # A tibble: 3 x 3
#>   Fertilizante Peso_medio Desv.Tipica
#>   <fct>        <dbl>      <dbl>
#> 1 A             143.       28.6
#> 2 B             125.       20.6
#> 3 C              110.      8.77
```

Peso de la planta a su recogida

Fertilizante	Peso
A	137.4
A	176.2
A	113.5
A	138.6
A	178.7
A	114.6
B	105.3
B	105.7
B	127.5
B	156.8
B	115.2
B	140.8
C	102.4
C	106.4
C	106.8
C	127.0
C	106.2
C	108.9

```
anova |>
  ggplot(aes(x = Fertilizante,
             y = Peso)) +
  geom_point(alpha = 0.5,
             col = "orangered")
```

10.2.1. Modelo

Tenemos una variable Y que toma valores reales y una variable cualitativa o factor X con niveles $1, 2, \dots, i, \dots, k$. La variable Y toma valores y_{ij} , $j = 1, \dots, n_i$ en el nivel i del factor X , siendo n_i el número de observaciones en el nivel i del factor X . Cuando todos los niveles tienen el mismo número de observaciones, $n_i = n_{i'}$ $\forall i, i'$, decimos que el diseño está balanceado o equilibrado. El modelo puede escribirse de dos formas:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad (10.2)$$

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad (10.3)$$

En la ecuación (10.2), μ es la media de la variable Y , mientras que α_i es el **efecto** en la media del nivel i , es decir, cuánto aumenta o disminuye la media

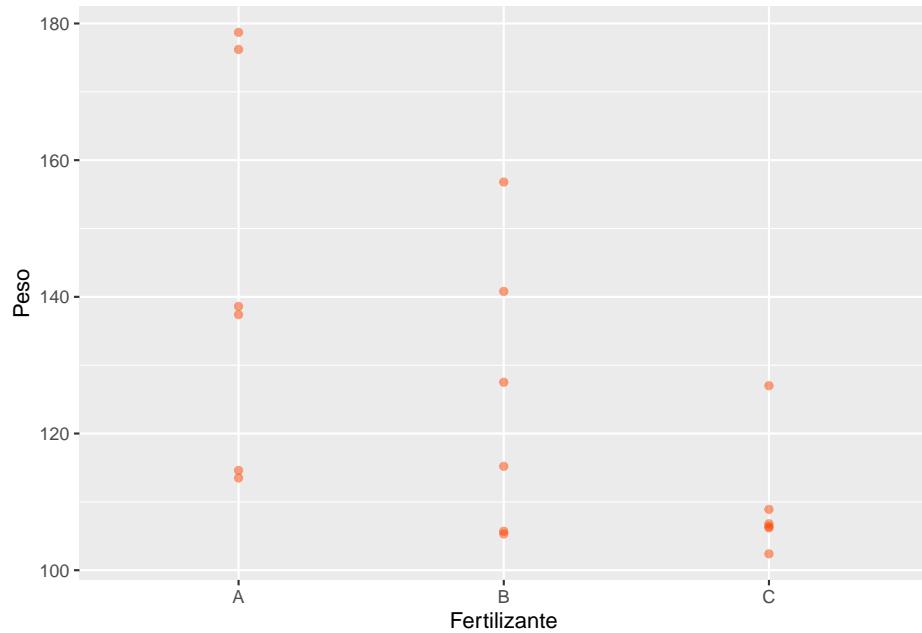


Figura 10.1: Gráfico de puntos del experimento en plantas

de Y por pertenecer a la categoría i . En la ecuación (10.3), μ_i es la media de la variable Y para el nivel i del factor X , de donde tenemos que el efecto es:

$$\alpha_i = \mu_i - \mu,$$

y el término de error, que representa toda la variabilidad que no explica el modelo, es:

$$\varepsilon_{ij} = y_{ij} - \mu_i.$$

Se cumple que:

$$\sum_i \alpha_i = 0; \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Podríamos representar matemáticamente nuestro ejemplo como:

$$Peso = \mu + \alpha_{Fertilizante} + \varepsilon,$$

o bien como:



$$Peso = \mu_{Fertilizante} + \varepsilon.$$

 El modelo ANOVA se ajusta en R con la función `aov`. El siguiente código ajusta el modelo de nuestro ejemplo, pero de momento solo lo guardamos, veremos en los siguientes apartados cómo extraer información e interpretarla.

```
modelo.aov <- aov(Peso ~ Fertilizante, danova)
```

10.2.2. Estimación de los parámetros

Si tenemos un total de n datos de los cuales hay n_i de cada nivel i , de forma que $\sum_i n_i = n$, los estimadores obtenidos tanto por el método de mínimos cuadrados como por el método de máxima verosimilitud (véase por ejemplo Lawson (2015)) son los siguientes:

$$\hat{\mu}_i = \bar{y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i},$$

$$\hat{\mu} = \bar{y}_{..} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}}{n},$$

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu},$$

es decir, las medias dentro de cada nivel del factor, y la media total. Con estos estimadores de los parámetros, la estimación de valores de Y vendrá dada por:

$$\hat{y}_{ij} = \hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i,$$

y por tanto los residuos del modelo son:

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

Nótese que, si tenemos k niveles solo tenemos que estimar $k - 1$, ya que:

$$\sum_i \alpha_i = 0.$$

Esta última restricción hace que no se puedan estimar los efectos con la representación matricial completa:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{pmatrix}$$

Debido a la singularidad de la matriz $\mathbf{X}^T \mathbf{X}$ (véase por ejemplo Lawson (2015)), lo que se hace es fijar uno de los niveles del factor como nivel “base”, y estimar la media del nivel base y los efectos de los otros niveles con respecto a la media del nivel base, es decir:

$$\hat{\beta} = \begin{pmatrix} \hat{\mu} + \hat{\alpha}_1 \\ \hat{\alpha}_2 - \hat{\alpha}_1 \\ \hat{\alpha}_3 - \hat{\alpha}_1 \end{pmatrix}$$

El nivel base que se toma en R de un factor no ordenado es el primero en orden alfabético, y se puede cambiar con la función `relevel`.

Sobre el objeto `modelo.aov` que guardamos antes, podemos aplicar funciones genéricas que devuelvan ciertos resultados. Por ejemplo, la función `coef` nos devuelve los estimadores de los coeficientes, teniendo en cuenta que toma `A` como nivel de referencia del factor `Fertilizante`. Podemos comprobar cómo se corresponden los coeficientes con los efectos estimados. La función `confint` nos devuelve un intervalo de confianza para los parámetros. Vemos que, con respecto al nivel base `A`, el Peso medio de la planta se reduce en casi 18 gramos cuando el fertilizante es el `B` y más de 33 cuando es el `C`. Se pueden visualizar los efectos con el paquete `effects`, como se muestra en la figura 10.2.



```

## El primer nivel es el de referencia
levels(danova$Fertilizante)
#> [1] "A" "B" "C"
## Estimación de los coeficientes
coef(modelo.aov)
#> (Intercept) FertilizanteB FertilizanteC
#> 143.1667     -17.9500      -33.5500
## Efecto nivel 1
a1 <- coef(modelo.aov)[1] - mean(danova$Peso); a1
#> (Intercept)
#> 17.16667
## Efecto nivel 2
a2 <- coef(modelo.aov)[2] + a1; a2
#> FertilizanteB
#> -0.7833333
## Efecto nivel 3
a3 <- coef(modelo.aov)[3] + a1; a3
#> FertilizanteC
#> -16.38333
## Comprobación
danova |>
  group_by(Fertilizante) |>
  summarise(Medias = mean(Peso)) |>
  mutate(Efectos = Medias - mean(danova$Peso))
#> # A tibble: 3 x 3
#>   Fertilizante Medias   Efectos
#>   <fct>        <dbl>    <dbl>
#> 1 A            143.    17.2
#> 2 B            125.   -0.783
#> 3 C            110.   -16.4
# aggregate(Peso ~ Fertilizante, danova, mean)$Peso - mean(danova$Peso)
## Intervalo de confianza
confint(modelo.aov, alpha = 0.99)
#> 2.5%    97.5%
#> (Intercept) 124.89704 161.436297
#> FertilizanteB -43.78716  7.887158
#> FertilizanteC -59.38716 -7.712842

plot(effects::effect("Fertilizante", modelo.aov))

```

10.2.3. Contrastes

En el análisis de la varianza de un factor, lo que nos interesa demostrar es que hay diferencias entre las medias de Y para distintos niveles del factor X (la X explica la Y). Si no hubiera diferencias entre los niveles, entonces las medias μ_i

Fertilizante effect plot

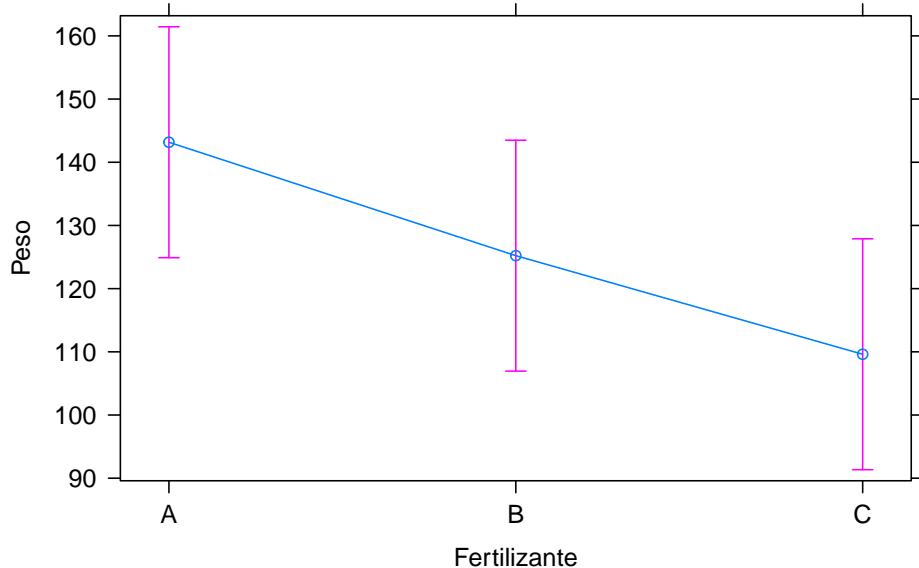


Figura 10.2: Visualización de los efectos

serían iguales, o lo que es lo mismo, los efectos α_i serían nulos. Por tanto, la hipótesis nula del modelo ANOVA de un factor es:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k,$$

o equivalentemente:

$$H_0 : \alpha_i = 0 \quad \forall i.$$

Nótese que la hipótesis alternativa es que hay diferencia entre los niveles, pero eso no significa que todos los niveles sean diferentes. Es decir, si rechazamos la hipótesis nula, al menos dos grupos serán distintos. Y por tanto tenemos evidencia para aceptar la alternativa:

$$H_1 : \alpha_i \neq 0 \text{ para algún } i,$$

Para contrastar la hipótesis nula, dividimos la variabilidad total de los datos entre la variabilidad que existe “dentro” de los grupos y la variabilidad que existe “entre” los grupos. Esta variabilidad la representamos por las sumas de cuadrados, de manera que la suma de cuadrados total (*SCT*) la podemos descomponer en la suma de cuadrados entre grupos (*SCE*) más la suma de cuadrados dentro de grupos (*SCD*):

Contenido de la tabla ANOVA

	GL	SC	CM	F	p-valor
factor	\$k-1\$	SCE	CME	F	p
residuos	\$n-k\$	SCD	CMD		

$$\sum_{ij} (y_{ij} - \bar{y}_{..})^2 = \sum_i n_i (\bar{y}_{i\cdot} - \bar{y}_{..})^2 + \sum_{ij} (y_{ij} - \bar{y}_{i\cdot})^2,$$

$$SCT = SCE + SCD.$$

Los grados de libertad de la suma de datos total es $n - 1$, que se descomponen también en $k - 1$ grados de libertad para la suma de cuadrados entre grupos y $n - k$ grados de libertad para la suma de cuadrados dentro de los grupos, de forma que podemos calcular los cuadrados medios totales, entre grupos y dentro de grupos:

$$CMT = \frac{\sum_{ij} (y_{ij} - \bar{y}_{..})^2}{n - 1},$$

$$CME = \frac{\sum_i n_i (\bar{y}_{i\cdot} - \bar{y}_{..})^2}{k - 1},$$

$$CMD = \frac{\sum_{ij} (y_{ij} - \bar{y}_{i\cdot})^2}{n - k},$$

Entonces, si se cumplen las condiciones para aplicar el modelo y la hipótesis nula es cierta, el estadístico:

$$F = \frac{CME}{CMD}$$

sigue una distribución F con $k - 1$ y $n - k$ grados de libertad, y podemos bien realizar el contraste de hipótesis para un nivel de confianza determinado, bien interpretar el p-valor, para llegar a una conclusión o decisión. En Análisis de la Varianza aquí descrito se resume normalmente en la llamada tabla ANOVA (ver tabla 10.2), que incluye las sumas de cuadrados, cuadrados medios, estadístico F y el p-valor.

La función genérica `summary` aplicada a un modelo `aov` devuelve precisamente la tabla ANOVA. En nuestro ejemplo, el p-valor es pequeño, menor de 0.05 (aunque por poco). Luego para un nivel de significación del 5% podemos rechazar la hipótesis nula y aceptamos que hay diferencias en las medias^a.



^aCon cautela, por lo ajustado del p-valor.

```
summary(modelo.aov)
#>           Df Sum Sq Mean Sq F value Pr(>F)
#> Fertilizante  2   3382   1691.2   3.836 0.0451 *
#> Residuals     15   6612    440.8
#> ---
#> Signif. codes:
#> 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como se ha indicado anteriormente, al rechazar la hipótesis nula aceptamos que no todos los niveles del factor producen una misma respuesta en la variable Y . Una vez rechazada la hipótesis nula, tendremos que comprobar qué niveles son realmente distintos para, en última instancia, sacar conclusiones o tomar decisiones. Un enfoque erróneo sería realizar comparaciones con el test de la t de Student para cada par de niveles del factor. En su lugar, utilizamos el método HSD¹ de Tukey, que utiliza el estadístico de los rangos *estudentizados*, $R/\hat{\sigma}$ para realizar todos los contrastes siguientes:

$$H_0 : \mu_i = \mu_j \quad \forall i \neq j.$$



La función `TukeyHSD` realiza los contrastes dos a dos de un modelo ANOVA, como en el siguiente ejemplo. La salida proporciona las diferencias entre cada par de niveles, un intervalo de confianza y el p-valor del contraste. Vemos que hay diferencias significativas entre los fertilizantes A y C, pero no en el resto de comparaciones. Estas diferencias se pueden visualizar utilizando la función `plot`, que produce la figura 10.3.

```
TukeyHSD(modelo.aov)
#> Tukey multiple comparisons of means
#> 95% family-wise confidence level
#>
#> Fit: aov(formula = Peso ~ Fertilizante, data = danova)
#>
#> $Fertilizante
```

¹Honestly Significance Difference

```
#>      diff     lwr      upr     p adj
#> B-A -17.95 -49.4362 13.536201 0.3274991
#> C-A -33.55 -65.0362 -2.063799 0.0361280
#> C-B -15.60 -47.0862 15.886201 0.4236517
plot(TukeyHSD(modelo.aov))
```

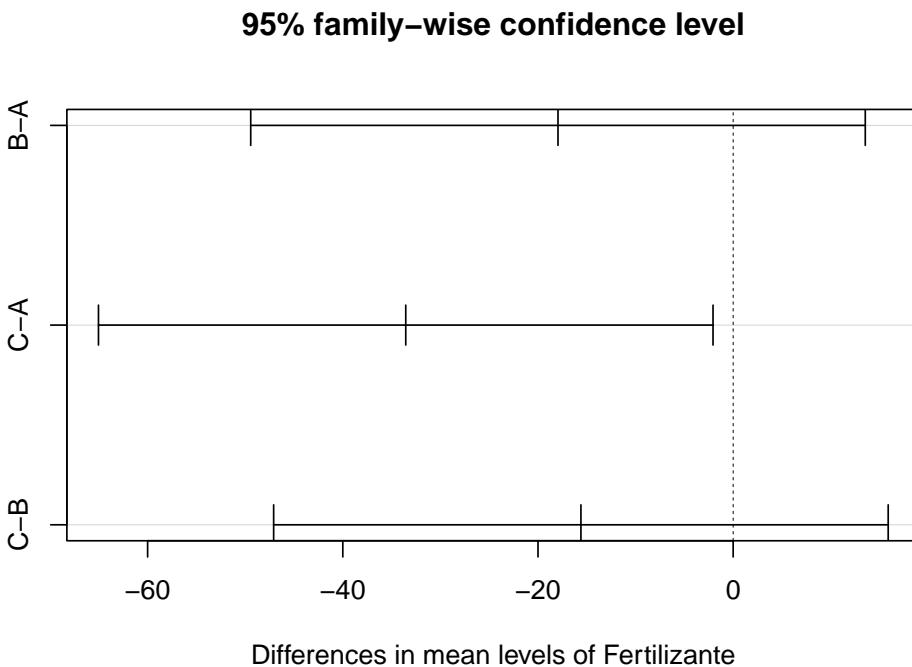


Figura 10.3: Visualización de las diferencias por pares

Recordemos que en un estudio observacional si se rechaza la hipótesis nula estamos confirmado que existe **relación** entre el factor y la variable. Para confirmar la relación causa-efecto del factor sobre la variable, el ANOVA de un factor debe realizarse a partir de un experimento diseñado correctamente, véase ??.

10.2.4. Validación del modelo y alternativas

Las hipótesis del test de la F y de los tests HSD de Tukey son la igualdad de varianzas entre grupos, y la normalidad de los residuos. La igualdad de varianzas se pueden comprobar fácilmente con el test de Bartlett². La normalidad de los residuos se puede verificar con alguno de los múltiples tests de normalidad existentes, como por ejemplo el de Kolmogorov-Smirnoff, el de Shapiro-Wilk o el de Anderson-Darling. La función genérica `plot` de R sobre un modelo ANOVA

²Hay otras alternativas, como el test de Levene en la función `levene.test` del paquete `car`.

guardado produce una serie de gráficos de diagnóstico que nos dan una idea a veces suficiente del cumplimiento de las hipótesis. La función `autoplot()` del paquete `{ggfortify}` (Horikoshi and Tang, 2022) los realiza con `{ggplot2}`.

La función genérica `residuals()` sobre el objeto `modelo.aov` devuelve los residuos del modelo. Podemos entonces hacer un contraste de normalidad. El p-valor es grande, mucho mayor de 0,05, por lo que no podemos rechazar que los residuos sean normales. La función `bartlett.test()` contrasta la hipótesis de homogeneidad de varianzas. El p-valor de este contraste es grande, mayor de 0.05 (aunque no mucho mayor). Podemos aceptar la igualdad de varianzas y por tanto las conclusiones de las pruebas de Fisher y de Tukey son válidas.

La figura 10.4 muestra los gráficos de diagnóstico generados con la función `ggfortify::autoplot()`. El gráfico superior izquierdo debería mostrar una línea recta si el modelo lineal se ajusta, y también se aprecia la homogeneidad de varianzas. A menudo cuando no se cumple esta hipótesis el gráfico muestra forma de embudo. El gráfico inferior izquierdo muestra la misma información pero a otra escala, donde tampoco se aprecian varianzas distintas. El gráfico superior derecho es un gráfico cuantil-cuantil para comprobar la normalidad de los residuos, en este caso se ajusta bastante bien. El gráfico inferior derecho sirve para detectar observaciones con gran influencia en el modelo. En este caso no se etiqueta ninguna.



```

shapiro.test(residuals(modelo.aov))
#>
#> Shapiro-Wilk normality test
#>
#> data: residuals(modelo.aov)
#> W = 0.92539, p-value = 0.161
bartlett.test(Peso ~ Fertilizante, data = danova)
#>
#> Bartlett test of homogeneity of variances
#>
#> data: Peso by Fertilizante
#> Bartlett's K-squared = 5.3317, df = 2, p-value =
#> 0.06954

library(ggfortify)
autoplot(modelo.aov)

```

En caso de no cumplimiento de las hipótesis, podemos realizar contrastes no paramétricos. Para el contraste de hipótesis de igualdad entre los niveles del factor, utilizamos el contraste de Kruskal-Wallis, que también dispone de un método para realizar comparaciones múltiples, como veremos en los ejemplos.

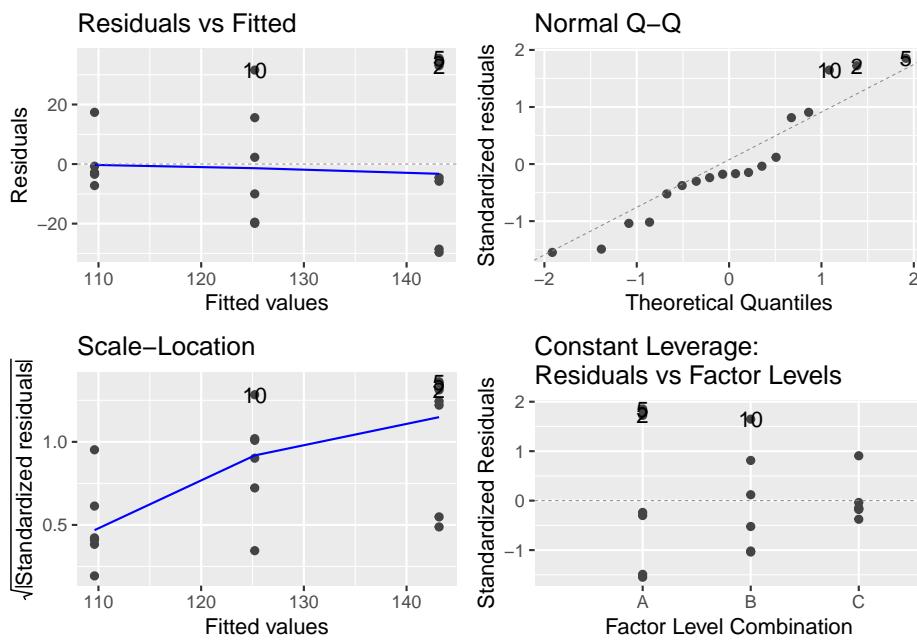


Figura 10.4: Gráficos de diagnóstico modelo ANOVA

Es importante tener en cuenta que la función `kruskal.test()` de R requiere que la variable cualitativa sea de tipo factor y no carácter.

Otra alternativa es transformar los datos originales para conseguir normalidad y/o homogeneidad de varianzas y realizar el análisis con los datos transformados.

En el conjunto de datos hay una variable que no hemos usado. Es el factor **Tierra**, que tiene tres niveles (tabla 10.3). Se deja al lector como ejercicio realizar un análisis exploratorio de los datos. Ajustamos el modelo ANOVA para estudiar el efecto de este factor en el Peso de la planta. Toma los valores Z1, Z2, Z3, según la zona de origen de la tierra utilizada en la maceta. En este caso, el contraste de hipótesis no permite rechazar la igualdad de medias. No obstante, al validar las hipótesis del modelo^a vemos que no se cumple la hipótesis de normalidad, por lo que hay que hacer el contraste de Kruskal-Wallis para confirmar que la Tierra no explica el Peso. Obtenemos un p-valor muy grande, por lo que no podemos rechazar que los datos vengan de la misma población, y por tanto llegamos a la conclusión de que no hay diferencias. Aunque en este caso no sería necesario seguir, por completitud del ejemplo vamos a realizar las comparaciones por pares utilizando la función `kruskalmc()` del paquete `{pgirmess}` (Girraudoux, 2021). Obtenemos en este caso las diferencias observadas y críticas (para un determinado nivel de significación) y un indicador (TRUE/FALSE) de diferencia significativa. Como era de esperar, no se encuentra ningún par de niveles con diferencias significativas.



^aSe deja como ejercicio al lector analizar los gráficos de diagnóstico.

```

modelo.aov2 <- aov(Peso ~ Tierra, danova)
summary(modelo.aov2)
#>           Df Sum Sq Mean Sq F value Pr(>F)
#> Tierra       2    43    21.6   0.033  0.968
#> Residuals   15  9951   663.4
shapiro.test(residuals(modelo.aov2))
#>
#> Shapiro-Wilk normality test
#>
#> data: residuals(modelo.aov2)
#> W = 0.84513, p-value = 0.007103
bartlett.test(Peso ~ Tierra, danova)
#>
#> Bartlett test of homogeneity of variances
#>
#> data: Peso by Tierra
#> Bartlett's K-squared = 3.4019, df = 2, p-value =
#> 0.1825
kruskal.test(Peso ~ Tierra, danova)
#>
#> Kruskal-Wallis rank sum test
#>
#> data: Peso by Tierra

```

Peso de las plantas y origen de la tierra

Tierra	Peso
Z2	137.4
Z2	176.2
Z3	113.5
Z1	138.6
Z2	178.7
Z1	114.6
Z2	105.3
Z2	105.7
Z3	127.5
Z1	156.8
Z2	115.2
Z1	140.8
Z2	102.4
Z1	106.4
Z2	106.8
Z3	127.0
Z1	106.2
Z2	108.9

```
#> Kruskal-Wallis chi-squared = 0.5653, df = 2, p-value
#> = 0.7538
pgirmess::kruskalmc(Peso ~ Tierra, data = danova)
#> Multiple comparison test after Kruskal-Wallis
#> p.value: 0.05
#> Comparisons
#>      obs.dif critical.dif difference
#> Z1-Z2 1.9444444   6.735838    FALSE
#> Z1-Z3 0.1666667   9.037076    FALSE
#> Z2-Z3 1.7777778   8.520237    FALSE
```

10.3. Análisis de la varianza de varios factores

La principal diferencia con el análisis de la varianza de un factor es que, además de los efectos principales de cada uno de los factores, es decir, cuánto varía la media según los niveles del factor, se puede estudiar el efecto de las interacciones entre factores. Intuitivamente, la interacción entre factores es similar a la que podemos observar en la vida diaria, por ejemplo un tranquilizante tiene un efecto positivo sobre el bienestar de una persona. Una copa de vino en determinadas circunstancias también. Pero utilizados conjuntamente, producen una interacción que afecta negativamente en el bienestar de la persona. Del mismo modo, podemos observar que una variable dependiente Y toma mejores valores para

ciertos niveles de los factores X_1 y X_2 . Pero se deben estudiar las interacciones, porque puede ser que dichos niveles combinados produzcan peor resultado.

El modelo de dos factores, que en la literatura en inglés de encuentra como *two-way anova*, es el siguiente:

$$Y = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon,$$

donde $(\alpha\beta)_{ij}$ representa el efecto de la interacción, y el resto de términos tienen la misma interpretación que en el ANOVA de un factor (o *one-way anova*). En la tabla ANOVA se añaden nuevas filas y contrastes para los efectos principales y las interacciones. Las hipótesis del modelo son las mismas, luego comprobamos normalidad de los residuos y homogeneidad de variazas. Para este último caso, utilizamos mejor el test de Levene que permite incluir el término de la interacción.

En el ajuste del modelo y estimación de efectos, se toma como base el primer nivel de todos los factores.

Para especificar modelos con más de un factor e interacción en R, ampliamos el lado derecho de la fórmula del modelo. Los nuevos efectos se añaden “sumando”. La interacción se expresa separando los factores con dos puntos. Si utilizamos el símbolo asterisco, entonces el modelo incluye todos los efectos principales y las interacciones. Por ejemplo, para dos factores **a** y **b**, el modelo completo se puede expresar como **a*b** o como **a + b + a:b**.

En el apartado anterior hemos analizado por separado el Peso frente a los factores Fertilizante y Tierra. Pero deberíamos analizarlos en un modelo multifactorial.

 Las siguientes expresiones crean el modelo multifactorial, realizan los contrastes, calcula los efectos y representa las interacciones.

```
modelo.aov3 <- aov(Peso ~ Fertilizante*Tierra, danova)
summary(modelo.aov3)
#>                               Df Sum Sq Mean Sq F value    Pr(>F)
#> Fertilizante                  2   3382   1691.2   9.674 0.00572 ***
#> Tierra                         2     43     21.6   0.124 0.88521
#> Fertilizante:Tierra           4   4996   1248.9   7.145 0.00712 ***
#> Residuals                      9   1573    174.8
#> ---
#> Signif. codes:
#> 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
shapiro.test(residuals(modelo.aov3))
#>
```

```
#> Shapiro-Wilk normality test
#>
#> data: residuals(modelo.aov3)
#> W = 0.90975, p-value = 0.08516
car::leveneTest(Peso ~ Fertilizante*Tierra, danova)
#> Levene's Test for Homogeneity of Variance (center = median)
#>      Df F value Pr(>F)
#> group   8  0.5554 0.7902
#>      9
coef(modelo.aov3)
#>             (Intercept)          FertilizanteB
#>             126.60000           22.20000
#>             FertilizanteC          TierraZ2
#>             -20.30000           37.50000
#>             TierraZ3 FertilizanteB:TierraZ2
#>             -13.10000           -77.56667
#> FertilizanteC:TierraZ2 FertilizanteB:TierraZ3
#>             -37.76667           -8.20000
#> FertilizanteC:TierraZ3
#>             33.80000
effects::allEffects(modelo.aov3)
#> model: Peso ~ Fertilizante * Tierra
#>
#> Fertilizante*Tierra effect
#>          Tierra
#> Fertilizante   Z1      Z2      Z3
#>       A 126.6 164.1000 113.5
#>       B 148.8 108.7333 127.5
#>       C 106.3 106.0333 127.0
with(danova,
  interaction.plot(x.factor = Fertilizante,
    trace.factor = Tierra,
    response = Peso,
    las = 1))
```

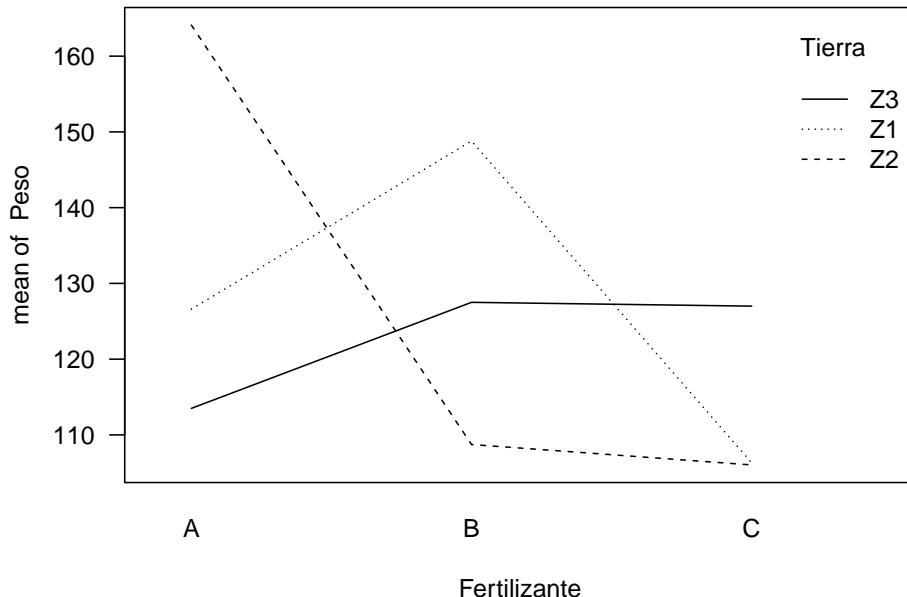


Figura 10.5: Visualización de las interacciones entre iluminación y Tierra.

La tabla ANOVA nos muestra que el término de la interacción es altamente significativo, con un p-valor muy bajo, incluso inferior a 0,01. También el tipo de fertilizante. No lo es el efecto principal de la Tierra, pero como este factor está en una interacción significativa, no deberíamos eliminarlo del modelo.

Las hipótesis del modelo y el modo de proceder son análogos al caso unifactorial. Se verifica la normalidad de los residuos y la homogeneidad de varianzas. El mayor Peso medio de la planta se consigue con la combinación fertilizante A y Tierra Z2.

El gráfico de las interacciones (figura 10.5) muestra claramente las grandes diferencias entre las combinaciones de factores. Si no hubiera interacción, las líneas serían paralelas. En el apartado anterior habíamos llegado a la conclusión de que el mejor fertilizante para conseguir el peso más alto era el fertilizante “A”, pero era indiferente utilizar tierra de una zona u otra. Si no analizáramos las interacciones, podríamos cometer el error de utilizar para las plantas fertilizante A y tierra de la zona Z3, lo que sería una pésima decisión ya que esta interacción baja drásticamente el peso de las plantas.



El caso de dos factores con interacción se extiende fácilmente a más de dos factores añadiendo términos a la fórmula. No obstante, las interacciones de más de dos factores se suele despreciar. En el modelo resultante final se deberían elimi-

nar los efectos no significativos, aunque manteniendo aquellos efectos principales que intervengan en alguna interacción.

Si no se cumplieran las hipótesis del modelo podríamos usar el contraste de Kruskal-Wallis para los efectos principales. Para la interacción también hay métodos paramétricos, aunque no están tan extendidos, véase una revisión en Feys (2016). En muchas ocasiones una transformación de Box-Cox en la variable respuesta es suficiente para ajustar un modelo válido, véase ??.

10.4. Introducción a los modelos mixtos: efectos fijos y efectos aleatorios

El modelo de análisis de la varianza y los contrastes vistos hasta ahora asumen que los individuos se han asignado aleatoriamente a los distintos niveles. Incluso en estudios observacionales, podemos asumir que esta asignación se ha hecho de forma aleatoria y controlada, y son efectos fijos.

Pero esto no siempre se puede asumir, o directamente la naturaleza del propio factor es aleatoria, y el tratamiento que le tenemos que dar a los factores es distinto. En particular, los efectos de estos factores no son una constante que queramos estimar, sino variables aleatorias de las cuales queremos estudiar su varianza en el modelo. El modelo para un factor fijo α y otro aleatorio β sería el siguiente:

$$Y = \mu + \alpha_i + \beta_j + \varepsilon,$$

donde el efecto aleatorio $\beta \sim N(0, \sigma_\beta^2)$. No tiene por tanto sentido estimar el efecto, cuya media es cero, sino la variabilidad, y ver si es importante con respecto al resto de factores.

Imaginemos por un momento que la variable Tierra de nuestro ejemplo se refiera a la tierra de una parcela determinada que no hemos podido elegir, sino que es algo aleatorio (por ejemplo en granjas experimentales distantes). Está claro que este factor no es algo que podamos controlar, y por tanto su efecto es aleatorio. En este caso tendríamos que ajustar un modelo mixto. En este ejemplo vemos que el factor Tierra afecta poco o nada a la variable respuesta, ya que la varianza es prácticamente nula^a. Podemos obtener una estimación de los efectos con la función `ranef` del paquete `lme4`.



^aAunque aparezca un cero, es debido al redondeo.

```
library(lme4)
#> Loading required package: Matrix
#>
```

```

#> Attaching package: 'Matrix'
#> The following objects are masked from 'package:tidyverse':
#>
#>     expand, pack, unpack
modelo.mixto <- lmer(Peso ~ 1 + Fertilizante + (1|Tierra), danova)
#> boundary (singular) fit: see help('isSingular')
summary(modelo.mixto)
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: Peso ~ 1 + Fertilizante + (1 / Tierra)
#>   Data: danova
#>
#> REML criterion at convergence: 139.3
#>
#> Scaled residuals:
#>   Min     1Q Median     3Q    Max
#> -1.4130 -0.4437 -0.1580  0.5838  1.6924
#>
#> Random effects:
#> Groups   Name        Variance Std.Dev.
#> Tierra   (Intercept) 0.0      0
#> Residual           440.8    21
#> Number of obs: 18, groups: Tierra, 3
#>
#> Fixed effects:
#>             Estimate Std. Error t value
#> (Intercept) 143.167    8.571 16.703
#> FertilizanteB -17.950   12.122 -1.481
#> FertilizanteC -33.550   12.122 -2.768
#>
#> Correlation of Fixed Effects:
#>          (Intr) FrtlzB
#> FertilizntB -0.707
#> FertilizntC -0.707  0.500
#> optimizer (nloptwrap) convergence code: 0 (OK)
#> boundary (singular) fit: see help('isSingular')
ranef(modelo.mixto)
#> $Tierra
#>   (Intercept)
#> Z1            0
#> Z2            0
#> Z3            0
#>
#> with conditional variances for "Tierra"

```

Los modelos mixtos (o incluso aleatorios puros) tienen muchas aplicaciones tanto

en modelos biológicos como en cualquier otro ámbito. Algunos ejemplos son:

- Efectos aleatorios puros, como el que se ha mostrado en el ejemplo.
- Modelos de panel, donde tenemos mediciones en diferentes períodos, y la unidad observable dentro del tiempo forman el factor aleatorio
- Medidas repetidas, donde el individuo es el factor aleatorio
- Modelos anidados y jerarquizados, donde unos niveles están dentro de otros

Como se ha podido comprobar, el análisis de modelos mixtos no es tan sencillo como el de efectos fijos. Para una revisión más completa se recomienda consultar el libro de Faraway (2016).

10.5. Análisis multivariante de la varianza

Hasta ahora hemos analizado el efecto que uno o varios factores tienen sobre una única variable Y . En ocasiones, tenemos en el lado izquierdo de la fórmula un vector aleatorio \mathbf{Y} con p variables respuesta Y_1, \dots, Y_p con cierta estructura de correlación y queremos determinar si esta variable multivariante se comporta de forma distinta para los distintos niveles de las variables en el vector de variables independientes \mathbf{X} . El método calcula una matriz de errores y una matriz de hipótesis³, y mediante el cálculo de un estadístico (por defecto Pillai⁴) se contrasta la hipótesis.

En el conjunto de datos de ejemplo tenemos otra variable en escala métrica que es la Pureza de un determinado compuesto en la planta. El siguiente ejemplo realiza el análisis multivariante de la varianza para estas dos variables agrupadas en una matriz. El modelo se puede ajustar con la función `aov`, o bien con la función `maov`, que es realmente un *wrapper* de la anterior. Los contrastes multivariantes se obtienen con la función `summary.manova`, si no le añadimos `.manova` tenemos los contrastes univariantes para cada componente del vector aleatorio \mathbf{Y} . El resultado que obtenemos en este caso es muy similar al obtenido en el ejemplo univariante. En ocasiones puede suceder que los contrastes univariantes no son significativos pero sí lo es el contraste multivariante.



```

Y <- as.matrix(danova[, c("Peso", "Pureza")])
modelo.manova <- aov(Y ~ Fertilizante*Tierra, data = danova)
summary.manova(modelo.manova)
#>           Df Pillai approx F num Df den Df

```

³Ver: <https://rpubs.com/aaronsc32/manova-test-statistics>

⁴ver una descripción de las alternativas en <https://rpubs.com/aaronsc32/manova-test-statistics>.

```

#> Fertilizante      2  0.97640   4.2925     4    18
#> Tierra            2  0.14749   0.3583     4    18
#> Fertilizante:Tierra 4  1.14085   2.9877     8    18
#> Residuals          9
#>                         Pr(>F)
#> Fertilizante       0.01300 *
#> Tierra              0.83494
#> Fertilizante:Tierra 0.02561 *
#> Residuals
#> ---
#> Signif. codes:
#> 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
summary(modelo.manova)
#> Response Peso :
#>                               Df Sum Sq Mean Sq F value    Pr(>F)
#> Fertilizante           2 3382.3 1691.17  9.6743 0.005724 **
#> Tierra                  2   43.2   21.61  0.1236 0.885213
#> Fertilizante:Tierra    4 4995.8 1248.95  7.1446 0.007121 **
#> Residuals               9 1573.3  174.81
#> ---
#> Signif. codes:
#> 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Response Pureza :
#>                               Df Sum Sq Mean Sq F value    Pr(>F)
#> Fertilizante           2 411.96 205.980 31.0211 9.168e-05 ***
#> Tierra                  2   3.53   1.765  0.2658 0.7724260
#> Fertilizante:Tierra    4 496.88 124.219 18.7077 0.0002184 ***
#> Residuals               9  59.76   6.640
#> ---
#> Signif. codes:
#> 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Capítulo 11

Diseño de experimentos

11.1. Introducción

El Diseño y Análisis de Experimentos (que abreviaremos como DoE), como cualquier otra técnica estadística, se basa en el estudio de la variabilidad. DoE es la herramienta más potente para la mejora, lo que ha llevado a algunos autores a llamarlo “the jewel of quality engineering” (Ver por ejemplo Allen (2010)).

En apartados anteriores del libro hemos aprendido las herramientas básicas para **analizar la variabilidad** de los datos. En este apartado vamos a revisar las técnicas de Diseño de Experimentos y su posterior análisis. Demasiado a menudo los esfuerzos se centran en intentar analizar un experimento sin diseño, lo que provoca frustración en los equipos involucrados en el análisis de datos. Vamos a mostrar la importancia de la fase de diseño, así como su planificación y correcta ejecución. No obstante la parte de análisis es igualmente importante, sobre todo en lo que concierne a la correcta interpretación de los resultados.

11.2. Bases del DoE: origen, importancia, objetivos y requerimientos

El DoE moderno surge a principios del siglo XX de la mano de Ronald A. Fisher cuando trabajaba en el “Rothamsted Experimental Station” en Inglaterra. Sus estudios se centraban en reducir la variación natural y prevenir la confusión con la variación de los restantes efectos. En última instancia, detectar las relaciones causa-efecto con el menor esfuerzo experimental.

Básicamente, necesitamos el DoE frente a estudios observacionales u otras estrategias como “un factor cada vez” para estudiar las interacciones y encontrar relaciones de causa-efecto con el menor uso de recursos posible. Así, podremos tomar decisiones respaldadas por los datos.

El objetivo del diseño de experimentos es encontrar los niveles de ciertos factores que optimizan una determinada característica medible. Esto se consigue con un método sistemático¹ que evita salidas en falso y respuestas incompletas. Mediante la reducción del error experimental se consigue evitar la confusión de los efectos y anular los efectos sin interés para el estudio.

Para empezar, lo primero que necesitamos es definir los datos del problema objeto de estudio y disponer de una forma de obtenerlos adecuadamente, en particular:

- Una variable respuesta en escala métrica
- Factores controlables
- Posiblemente, otros factores aleatorios

Esta recogida de datos se debe realizar de forma sistemática y teniendo en cuenta los tres pilares del DoE: aleatorización, bloqueo y replicación.

11.3. Importancia del diseño

Con la experimentación básicamente controlamos los niveles a los que operan ciertos factores controlables, a la vez que se asignan dichos niveles (configuraciones, tratamientos, etc.) a las unidades experimentales. Esto permite, unido a las apropiadas estrategias de aleatorización, bloqueo y replicación, realizar predicciones acerca del desempeño de un determinado proceso. Estas predicciones así establecidas serán el resultado de la identificación de una relación causa-efecto, que no se puede conseguir simplemente analizando datos recogidos sin diseño. En los estudios observacionales:

- Recogemos información
- No controlamos factores
- Análisis descriptivos
- Descubrir *relaciones*

Mientras que con experimentos diseñados:

- Se controlan los factores
- Se analizan efectos
- Incluidas las interacciones
- Se verifica la relación *causa-efecto*

Si la experimentación se lleva a cabo variando una vez cada factor, buscando el valor óptimo para la respuesta para cada factor individualmente dejando fijos el resto arbitrariamente, estaremos obviando un aspecto fundamental: el efecto de las interacciones. La interacción es el efecto que tiene un factor a distintos niveles de otros factores. Por otra parte, el número de experimentos necesarios para llegar a conclusiones válidas es mucho mayor (y por tanto el experimento más costoso). Con diseño de Experimentos obtenemos el mayor

¹En realidad, el método científico.

número de combinaciones posibles para estimar interacciones, con el mínimo número de experimentos.

El análisis de datos, por muy sofisticado que sea, no puede nunca arreglar un experimento mal diseñado (chapucero, según Lawson)

Sometimes the only thing you can do with a poorly designed experiment is to try to find out what it died of
R.A. Fisher

As we know from Murphy's Law, if anything can go wrong it will, and analysis of data can never compensate for botched experiments

Lawson (2015)

El análisis de la varianza sin diseño de experimentos tiene algunas limitaciones importantes. Sin Diseño de Experimentos, los datos pueden ser inconsistentes o incompletos, al no incluir factores de ruido o Factores latentes. Si tenemos variables correlacionadas, y alguna de ellas no se mide, su efecto puede quedar enmascarado por las otras, como en el ejemplo de la figura 11.1, donde si miramos solo la relación de la variable respuesta con el factor 1 (gráfico de la izquierda), podemos llegar a la conclusión errónea de que el factor 1 es determinante. Pero podría ser que la causa real sea el factor 2, que no ha sido medido y está muy correlacionado con el factor 1. En el gráfico de la derecha vemos que la variable respuesta crece en el mismo sentido que los factores 1 y 2, pero podría ser que el factor 2, no medido al principio, sea la causa, y no el que realmente se ha medido.

Por otra parte, el rango de valores de la variable respuesta está limitado por su rango normal de operación, que puede ocultar relaciones más amplias. En la figura 11.2, el gráfico de la derecha se corresponde con el rango de variación normal de los factores de un proceso. En el de la izquierda, ampliamos el rango de posibles valores de la variable, y vemos una relación más clara, que queda oculta en el otro caso.

11.4. Planificación de la experimentación

El conocimiento de la materia (*subject matter knowledge*) en cuestión es fundamental para desarrollar cambios que resulten en mejoras. Sin embargo, es necesario otro tipo de conocimiento (*profound knowledge*), en el que se incluye la Estadística. Combinar ambos conocimientos, lleva a una mayor capacidad de mejora. Estas ideas, originarias de Deming, se recogen en Moen et al. (2012). Algunas capacidades necesarias fruto de esta combinación son:

- Entender las interdependencias entre los sistemas donde se lleva a cabo la experimentación;

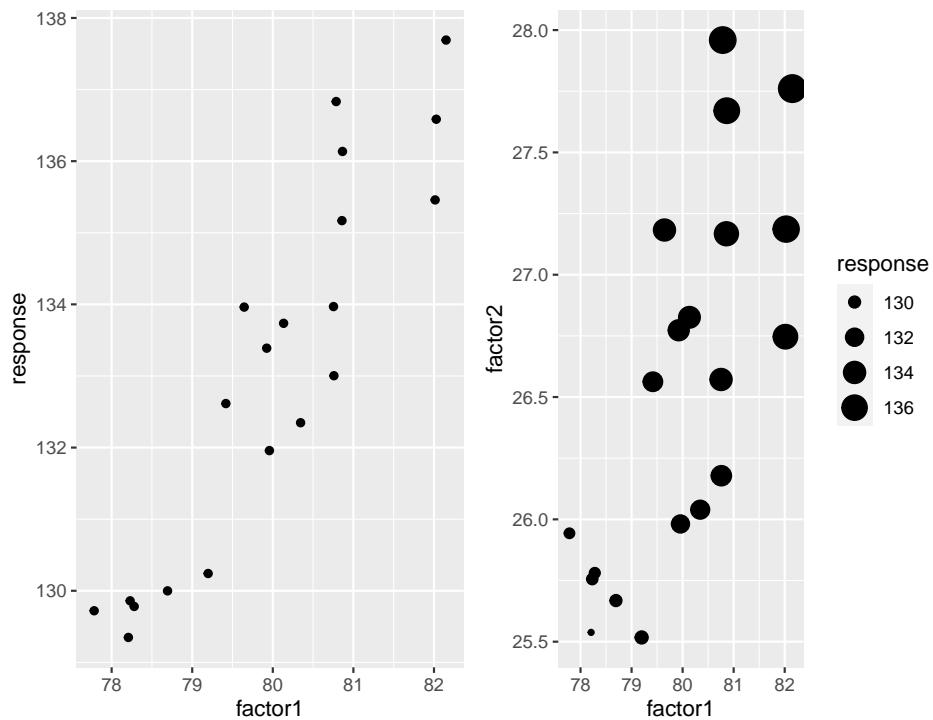


Figura 11.1: Efecto de no medir un factor

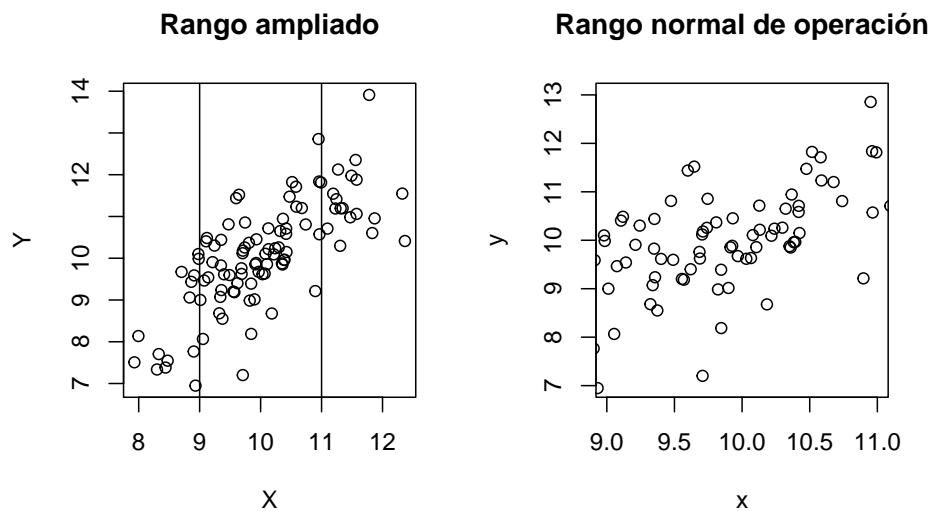


Figura 11.2: Efecto de la limitación del rango de valores

- Entender la relación entre las predicciones y el conocimiento del sistema que se quiere cambiar;
- Entender el efecto temporal de los cambios;
- Entender la importancia de la estabilidad del proceso;
- Entender la extrapolación de los resultados de las pruebas para mejorar el sistema.

En general, se pueden seguir tres estrategias de planificación para el diseño de experimentos. Sin planificación se pueden ir cambiando niveles de factores cada vez y haciendo pruebas (ensayo-error), definitivamente poco efectivo. Una planificación completa desde el inicio puede llevar a no explorar alternativas surgidas durante la experimentación, y por tanto a no cumplir los objetivos. La estrategia óptima la secuencia, es decir, llevar a cabo un número de experimentos al inicio, cuyas conclusiones supondrán la planificación de una segunda fase donde centrarnos en los factores realmente relevantes y hacer análisis más detallados y precisos. En las primeras fases se suelen realizar diseños de *screening* para descartar factores no significativos. En realidad, es la aplicación del método científico, en un proceso iterativo de aprendizaje como se muestra en la figura 11.3.

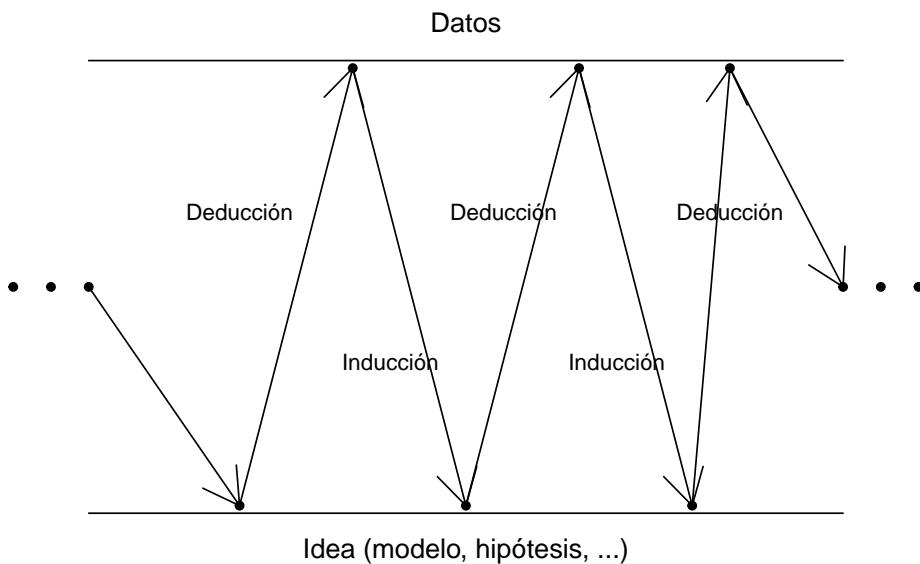


Figura 11.3: Método iterativo de aprendizaje

En Moen et al. (2012) se propone el ciclo PDSA (*Plan-Do-Study-Act*) para la mejora que se muestra en la figura 11.4. Básicamente consiste en:

1. Planifica un cambio o prueba, dirigido a la **mejora**
2. Lleva a cabo el cambio o prueba (corto alcance)
3. Estudia el resultado: ¿qué has aprendido? ¿qué ha ido mal?
4. Actúa:

- Adopta el cambio
- Abandónalo
- Empieza el ciclo de nuevo

¡Documenta todas las acciones de mejora!

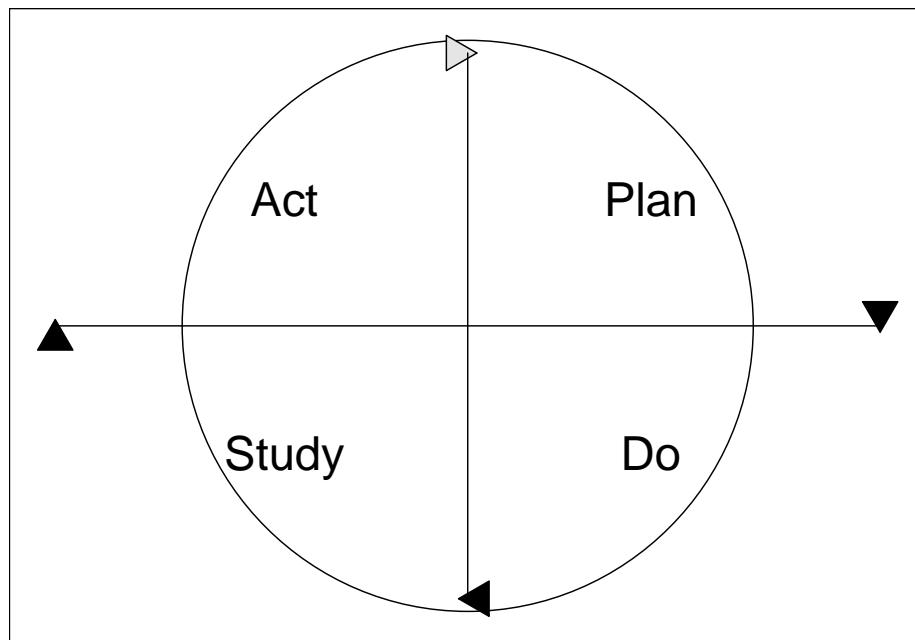


Figura 11.4: Ciclo PDSA para la mejora

Una buena forma de empezar el ciclo es a partir de un análisis de causa y efecto, por ejemplo con un diagrama de Ishikawa como el que aparece en la figura 11.5.

Lo siguiente probablemente sería determinar el presupuesto/recursos disponibles, en especial determinar el número de experimentos que se pueden realizar realísticamente.

Hasta ahora, hemos ido mencionando algunos conceptos básicos del diseño de experimentos. Ahora vamos a definirlos un poco más formalmente.

- **variable respuesta:** La variable de interés que pretendemos mejorar. Será una cuantificación de alguna característica de calidad, en sentido amplio.
- **factor:** Variable independiente que puede ser causa de la respuesta. La inferencia que haremos con DoE será confirmar o rechazar esta hipótesis.
- **variable de bloque:** Variable que no tiene interés en la investigación, pero puede influir en la respuesta. Mediante la formación de bloques confundimos su efecto con los factores que realmente nos interesan.
- **variable ruido:** Variable que puede influir en la respuesta, pero de la que no tenemos control.

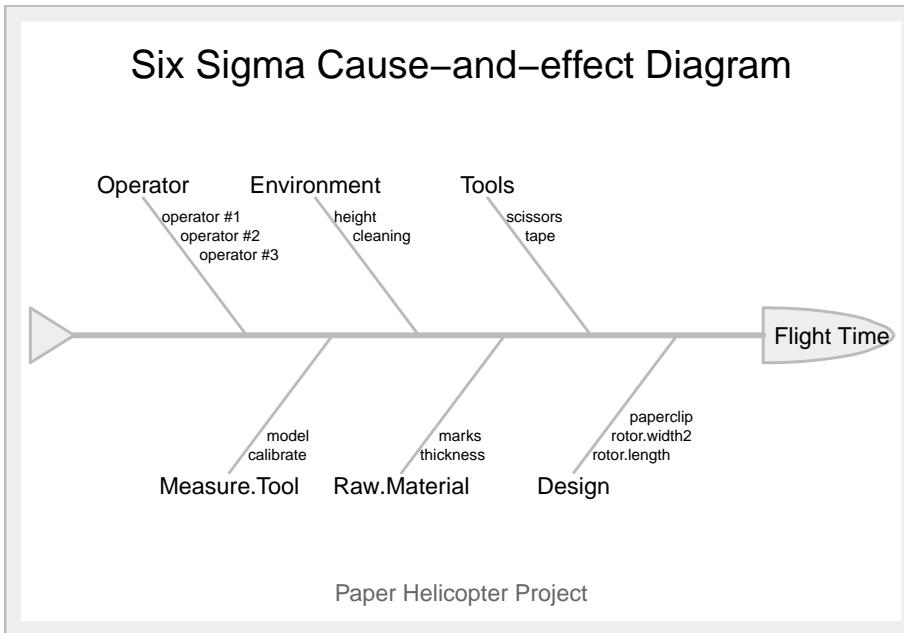


Figura 11.5: Ejemplo diagrama de causa-efecto

- **nivel:** Valor que fijamos de un factor. En variables cualitativas, una categoría. En variables cuantitativas, un valor numérico determinado fijado con antelación. A menudo se le llama también tratamiento.
- **unidad experimental:** La división más pequeña posible de unidades de un experimento tal que a dos cualesquiera se les pueden aplicar distintas combinaciones de factores y niveles.
- **unidad observable:** Cada uno de los elementos que forman la unidad experimental. A veces, un tratamiento no se puede aplicar a un solo elemento, sino a varios a la vez.
- **bloque:** Grupos de unidades experimentales que son tratados de forma similar en el experimento.
- **efecto:** El principal resultado de interés del experimento: qué pasa con la variable respuesta.
- **réplica:** Repetición de un experimento sobre una misma combinación de factores y niveles, a diferentes unidades experimentales.
- **repetición:** Repetición de la medición de la respuesta con las mismas condiciones experimentales, a la misma unidad experimental.
- **aleatorización:** Asignación de niveles y bloques a unidades experimentales de forma aleatoria

Al utilizar un modelo para simplificar una realidad, estamos cometiendo un error. El **error experimental** es aquel que se debe exclusivamente a las réplicas de las mismas condiciones experimentales. En cada diseño el error experimental se

calcula de una forma distinta, de forma que se separa de la variabilidad total para ver cuánta variaación se debe al modelo y poder así tomar decisiones. Así, en el modelo:

$$Y = f(X) + \varepsilon$$

- Y es la variable respuesta
- X es el conjunto de variables predictivas
- f función lineal, exponencial, etc.
- ε es una variable aleatoria

Se separa el error de la variabilidad total.

Los siguientes principios son cruciales a la hora de diseñar el experimento.

- **Aleatorización.** Los tratamientos deben ser asignados de forma aleatoria a las unidades experimentales. Esto incluye bloques, factores controlables, anidamientos, etc.
- **Formación de bloques.** Cuando no se puedan replicar exactamente las condiciones experimentales (por ejemplo, días diferentes), se deben organizar en bloques.
- **Réplicas.** Para poder estimar el error experimental y hacer contrastes de hipótesis, es necesario tener más de una *corrida* de cada combinación de tratamientos.

Lawson (2015) propone la siguiente *checklist* a la hora de planificar experimentos:

1. Definir objetivos
2. Identificar unidades experimentales
3. Definir variable respuesta medible y con sentido
4. Identificar los factores controlables y latentes
5. Ejecutar pruebas piloto
6. Hacer diagrama de flujo para cada experimento
7. Elegir el diseño experimental
8. Determinar el número de réplicas necesarias
9. Aleatorizar las condiciones experimentales a las unidades experimentales
10. Definir método de análisis de datos
11. Calendario y presupuesto para la ejecución

11.5. Tipos de diseños de experimentos

11.5.1. Experimentos con un factor

Podemos comparar una variable a distintos niveles de un solo factor. El contraste de la t de Student es la técnica utilizada para dos niveles. Para más niveles, utilizamos el análisis de la varianza de un factor (véase sec:anova1). Cuando hay

algún factor más que no es de interés, pero puede afectar a la variable resuesta, se debe introducir como variable de bloque. Los diseños de cuadrados latinos y cuadrados greco-latino se utilizan para introducir dos o tres factores de bloque respectivamente.

El diseño experimental para el ANOVA de un factor sigue las siguientes pautas:

1. Se quiere estudiar el efecto de un solo factor sobre una población. No hay otros factores controlables que puedan influir.
2. Se realiza el plan de recogida de datos, posiblemente con prueba piloto.
3. Se decide el número de unidades experimentales del experimento.
4. Se asignan **aleatoriamente** las unidades a los niveles del factor.
5. Se recogen los datos (experimento físico, cuestionario, etc.)
6. Se realiza un análisis descriptivo, sobre todo gráfico, de los datos recogidos.
7. Los datos se verifican y se preparan adecuadamente para el análisis.
8. Se ajusta el modelo.
9. Se comprueba la validez del modelo. Si no es válido, se busca modelo alternativo de análisis.
10. Se estiman los parámetros.
11. Se comprueba las hipótesis principal.
12. Si hay diferencias, se realizan comparaciones por pares.
13. Se comprueba la significación práctica y se obtienen conclusiones o se toman decisiones.

11.5.2. Diseños multifactoriales

Cuando analizamos más de un factor a varios niveles, aplicamos lo explicado en el apartado 10.3. Recordemos que en estos diseños es de vital importancia estudiar las interacciones.

11.5.3. Diseños factoriales a dos niveles 2^k

Un tipo especial de diseño multifactorial es aquél en el que todos los factores tienen solamente dos niveles. El número de experimentos necesarios para probar todas las combinaciones de niveles para k factores es 2^k , de ahí su nombre.

Diseño factorial 2^2

Modelo:

$$y_{ijk} = \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

Datos:

- Aleatorizar tratamientos
- Realizar k réplicas
- Número de experimentos: $k \times 2^2$

Análisis: - Efectos principales - Interacción

En el siguiente código se analiza un experimento con dos factores A y B a dos niveles, + y -. Se muestran los gráficos de los efectos.



```

library(xtable)
library(DoE.base)
library(effects)
library(reshape2)
datosf22 <- scan(text =
  "- - 28
  - - 25
  - - 27
  + - 36
  + - 32
  + - 32
  - + 18
  - + 19
  - + 23
  + + 31
  + + 30
  + + 29
",
  what = list(character(), character(), numeric()),
  sep = "\t")
datosf22 <- as.data.frame(datosf22)
colnames(datosf22) <- c("A", "B", "respuesta")
datosf22$replica <- rep(1:3, 4)
library(knitr)
kable(dcast(datosf22, A + B ~ replica, value.var = "respuesta"))

\begin{table border="1">
| A | B | 1 | 2 | 3 |
| --- | --- | --- | --- | --- |
| - | - | 28 | 25 | 27 |
| - | + | 18 | 19 | 23 |
| + | - | 36 | 32 | 32 |
| + | + | 31 | 30 | 29 |

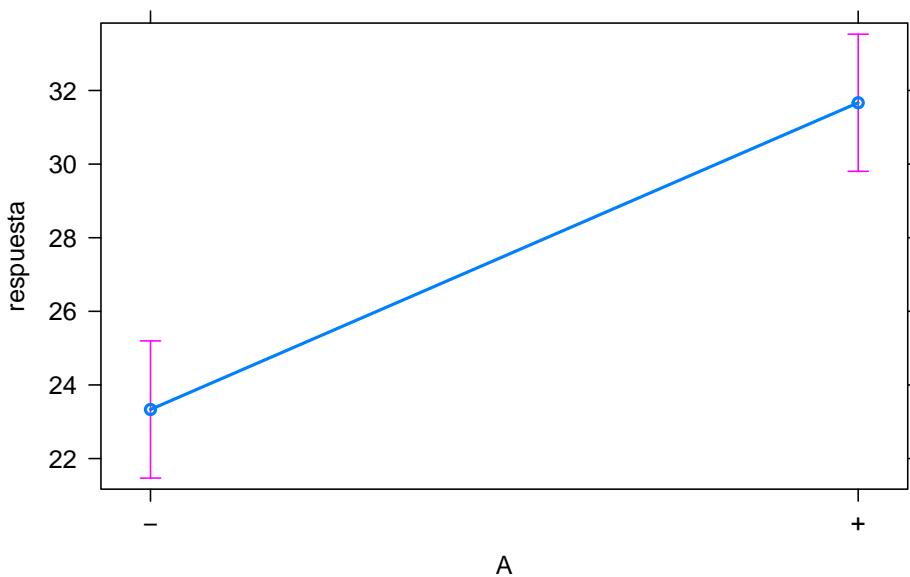


modelof22 <- lm(respuesta ~ A + B + A*B, data = datosf22)
# kable(anova(modelof22))
anova(modelof22)
#> Analysis of Variance Table

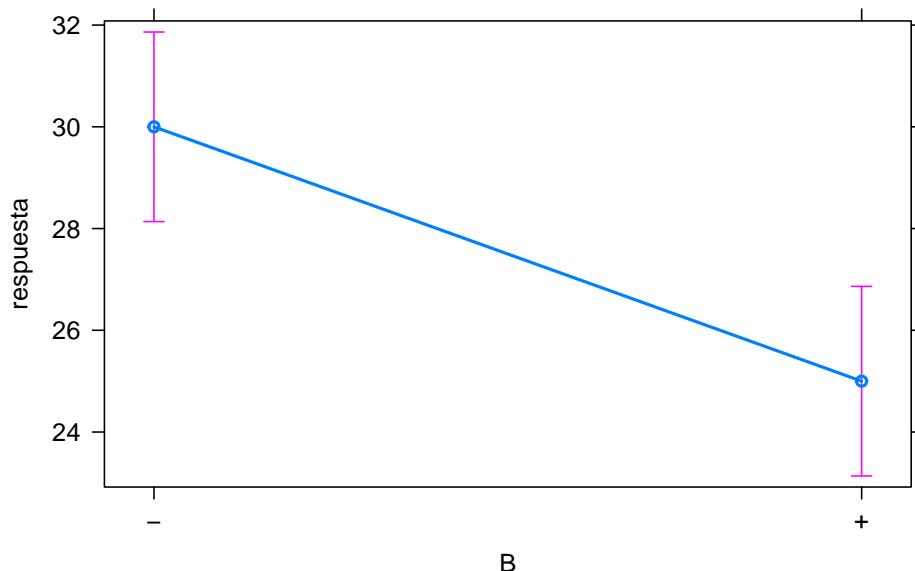
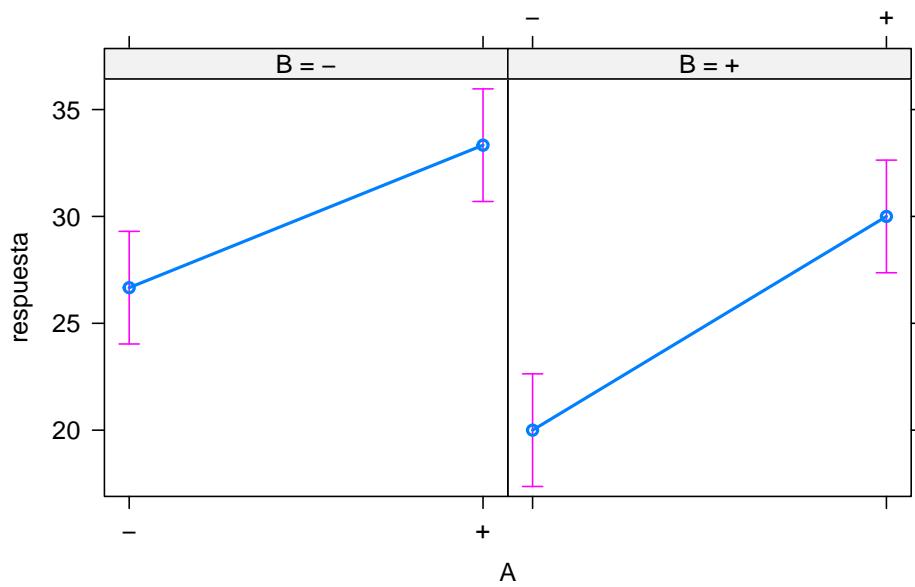
```

```
#>
#> Response: respuesta
#>           Df  Sum Sq Mean Sq F value    Pr(>F)
#> A          1 208.333 208.333 53.1915 8.444e-05 ***
#> B          1  75.000  75.000 19.1489  0.002362 **
#> A:B        1   8.333   8.333  2.1277  0.182776
#> Residuals  8  31.333   3.917
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

library(lattice)
trellis.par.set(background = list(col = "white"))
plot(effect(term = "A", mod = modelof22))
#> NOTE: A is not a high-order term in the model
```

A effect plot

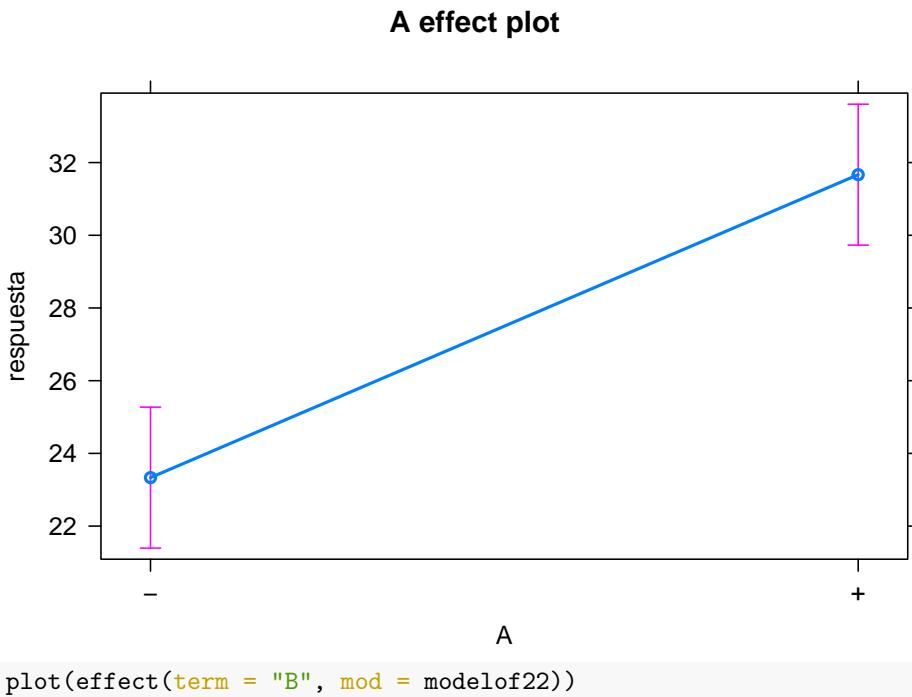
```
plot(effect(term = "B", mod = modelof22))
#> NOTE: B is not a high-order term in the model
```

B effect plot**A*B effect plot**

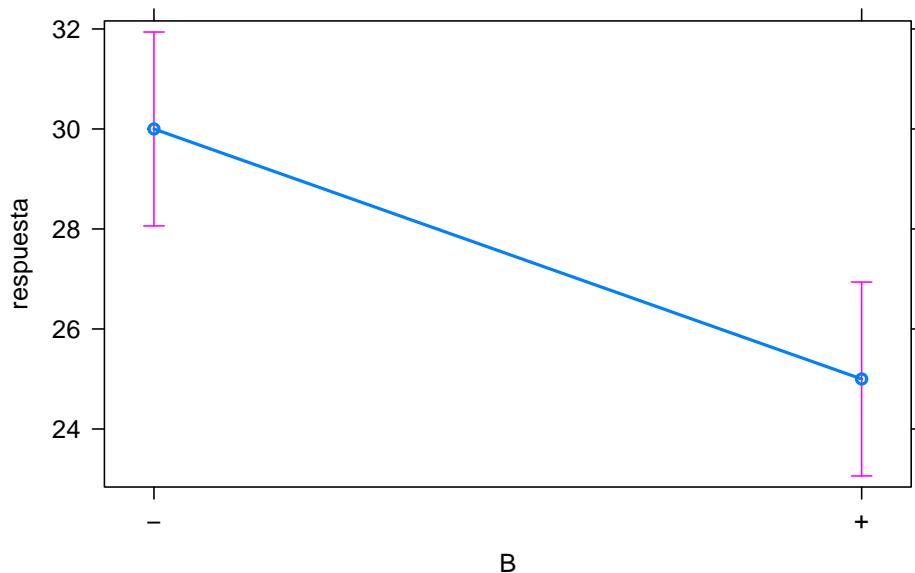
Vemos que los dos efectos principales son significativos, pero no lo es la interacción. Podemos eliminar ese término del modelo para así ganar grados de libertad y tener una mejor estimación del error.



```
modelof22 <- lm(respuesta ~ A + B, data = datosf22)
anova(modelof22)
#> Analysis of Variance Table
#>
#> Response: respuesta
#>           Df  Sum Sq Mean Sq F value    Pr(>F)
#> A          1 208.333 208.333 47.269 7.265e-05 ***
#> B          1  75.000  75.000 17.017  0.002578 **
#> Residuals  9  39.667   4.407
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
plot(effect(term = "A", mod = modelof22))
```



B effect plot



Diseño factorial 2^3

Modelo:

$$y_{ijkl} = \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkl}$$

Datos: - Aleatorizar tratamientos - l réplicas (o no) - Número de experimentos: $l \times 2^3$

Análisis:

- Efectos principales
- Interacciones (más de dos difícil de ver)
- Eliminar no significativas para aumentar precisión

En el siguiente ejemplo, analizamos tres factores, pero omitimos la interacción de orden 3. Después, podríamos eliminar las interacciones menos significativas para quedarnos con el modelo más sencillo.

 datosf23 <- scan(text = "

```

- - - 60
+ - - 72
- + - 54
+ + - 68
- - + 52

```

```

+   -   +   83
-   +   +   45
+   +   +   80
",
  what = list(character(), character(), character(), numeric()),
  sep = "\t")
datosf23 <- as.data.frame(datosf23)
colnames(datosf23) <- c("T", "C", "K", "rendimiento")
kable(dcast(datosf23, T + C + K ~ ., value.var = "rendimiento"))

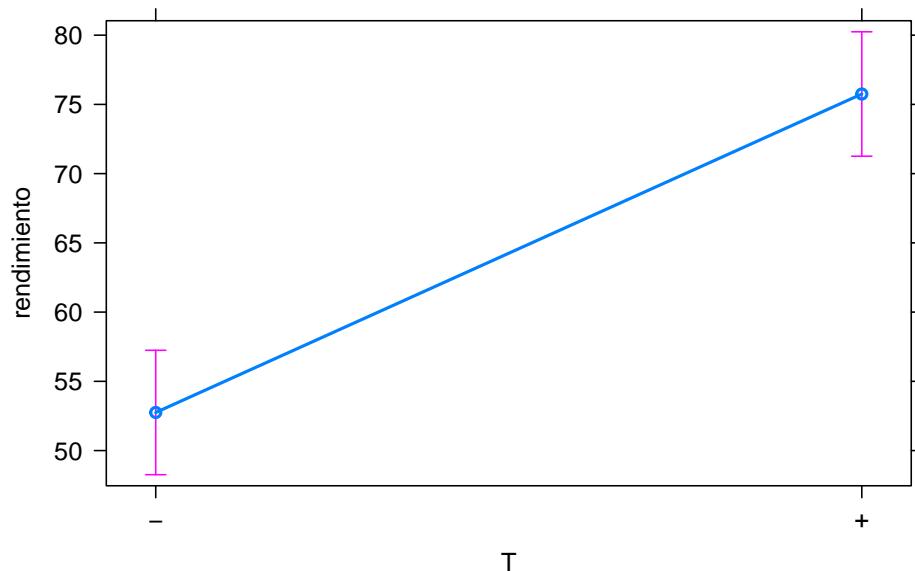
```

T	C	K	.
-	-	-	60
-	-	+	52
-	+	-	54
-	+	+	45
+	-	-	72
+	-	+	83
+	+	-	68
+	+	+	80

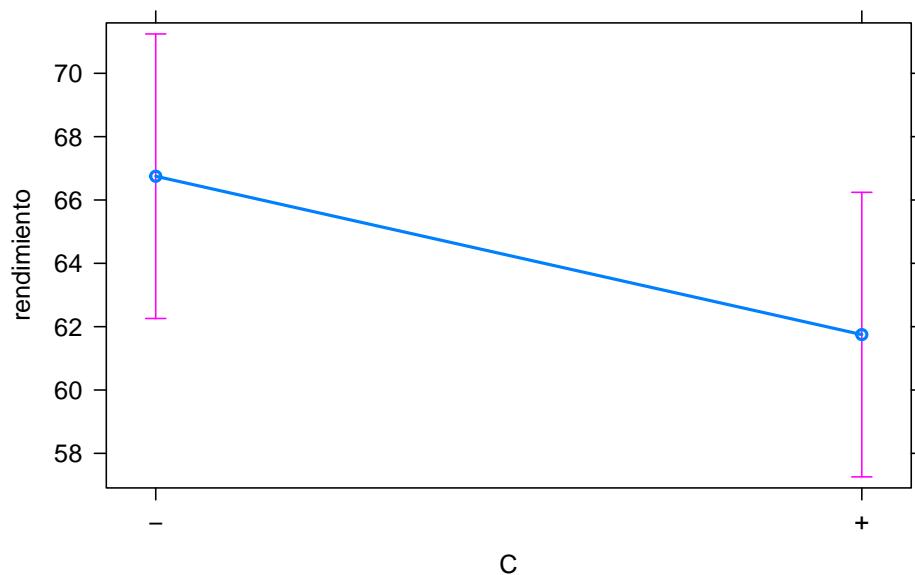
```

modelof23 <- lm(rendimiento ~ T + C + K + T*C + T*K + C*K, data = datosf23)
anova(modelof23)
#> Analysis of Variance Table
#>
#> Response: rendimiento
#>
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> T          1 1058.0 1058.0    2116 0.01384 *
#> C          1   50.0   50.0     100 0.06345 .
#> K          1    4.5    4.5      9 0.20483
#> T:C         1    4.5    4.5      9 0.20483
#> T:K         1  200.0  200.0     400 0.03180 *
#> C:K         1    0.0    0.0      0 1.00000
#> Residuals   1    0.5    0.5
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
plot(effect(term = "T", mod = modelof23))
#> NOTE: T is not a high-order term in the model

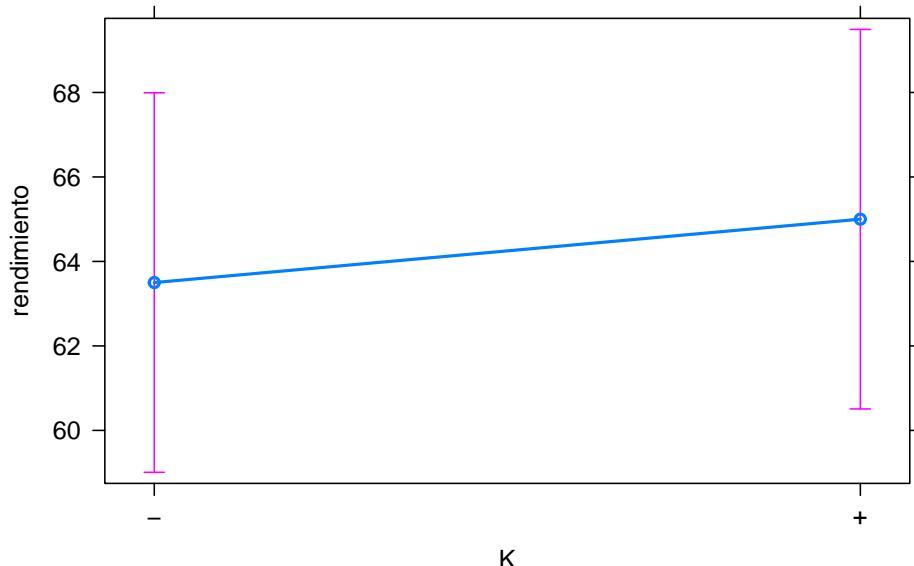
```

T effect plot

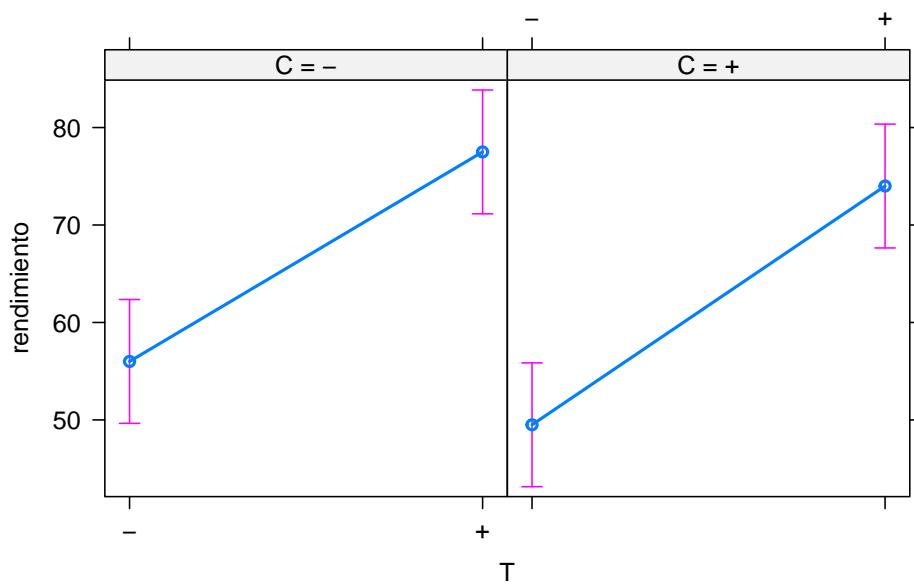
```
plot(effect(term = "C", mod = modelof23))
#> NOTE: C is not a high-order term in the model
```

C effect plot

```
plot(effect(term = "K", mod = modelof23))
#> NOTE: K is not a high-order term in the model
```

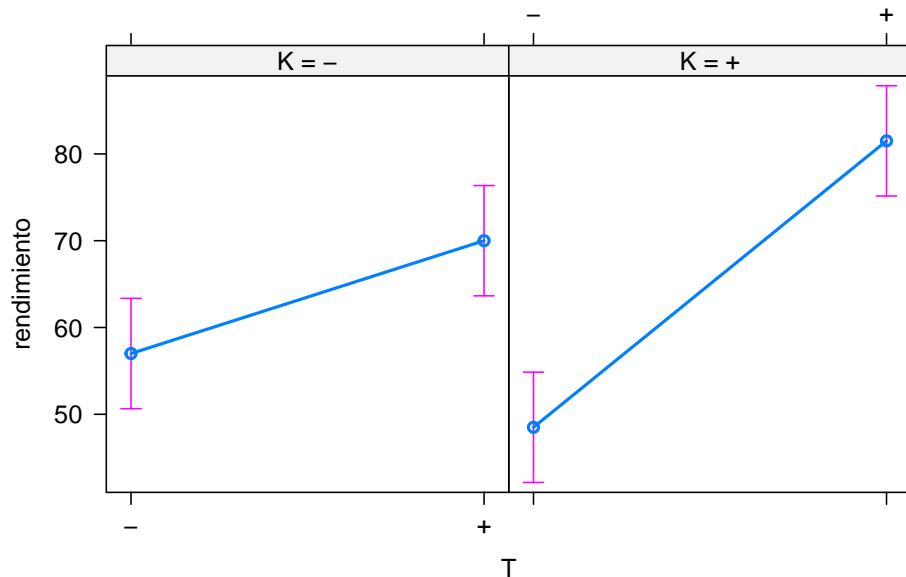
K effect plot

```
plot(effect(term = "T:C", mod = modelof23))
```

T*C effect plot

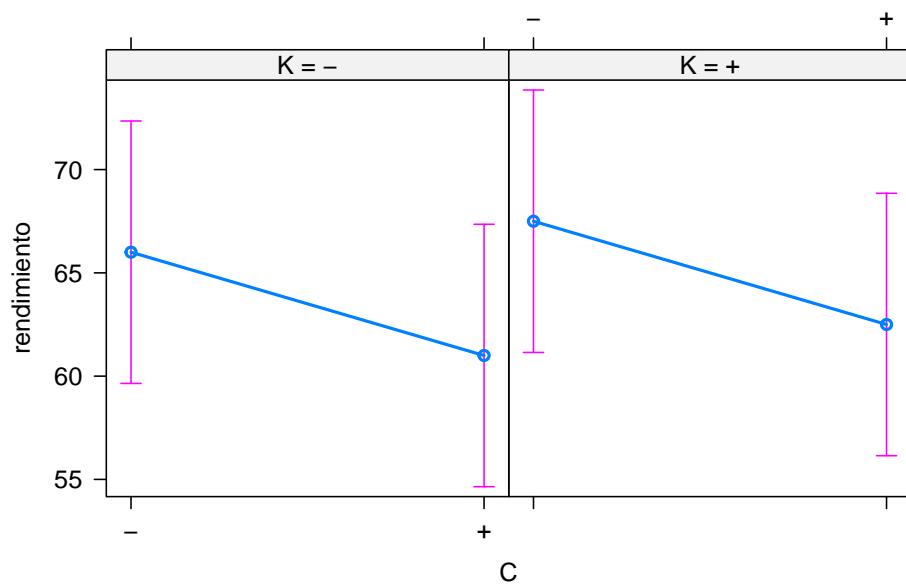
```
plot(effect(term = "T:K", mod = modelof23))
```

T*K effect plot



```
plot(effect(term = "C:K", mod = modelof23))
```

C*K effect plot



Diseño factorial 2^k

Siguiendo la misma estructura que los dos anteriores, con más efectos principales y más interacciones, pero más allá de 3 es muy difícil que se produzcan, y más difícil de interpretar. El número de experimentos necesarios aumenta exponencialmente, y se suelen preferir experimentos fraccionados. Cuando no hay grados de libertad suficientes para realizar contrastes se utilizan herramientas gráficas para seleccionar efectos significativos (Pareto y gráfico normal)

Para la Formación de bloques, se confunden con efectos de interacciones de orden superior, multiplicando los signos y dividiendo en dos bloques

11.5.4. Diseños fraccionales

Los diseños factoriales fraccionales 2^{k-p} utilizan solo una fracción de su equivalente factorial. En estos diseños se confunden los efectos principales con las interacciones de mayor orden. De esta forma, se puede realizar *screening* de muchos factores con pocos experimentos, y una vez eliminados del modelo los efectos no significativos se estima mejor el error.

11.5.5. Diseños avanzados

Existen otros diseños avanzados que no se tratan en este texto, como son:

- Plackett and Burman
- Diseños anidados
- Split-plot
- Medidas repetidas
- Superficie respuesta

Capítulo 12

Modelos de regresión

En preparación.

Regresión lineal simple

Regresión no lineal

Regresión lineal múltiple

Otros modelos*

(GLM, GAM, ...)

Parte IV

Control estadístico de la calidad

Capítulo 13

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 14

Control Estadístico de Procesos

En preparación.

Intro SPC

Gráficos de control

Capacidad y rendimiento

Capítulo 15

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> , <i>teta</i>)
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	Perteneció/perteneciente
\exists	Existe
\Rightarrow	Implica/entonces
∂	Derivada parcial
\approx	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

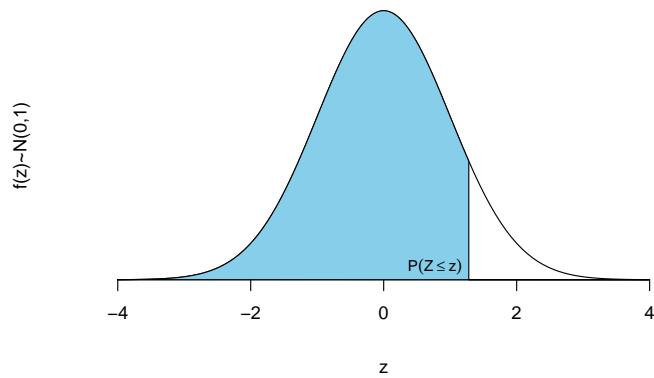
²En este libro se puede utilizar para tomar el entero superior o inferior según el contexto

Apéndice B

Tablas estadísticas

B.1. Distribución normal

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



B.2. Resumen modelos de distribución de probabilidad

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

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 lanzamiento de 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:

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

** Características**

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

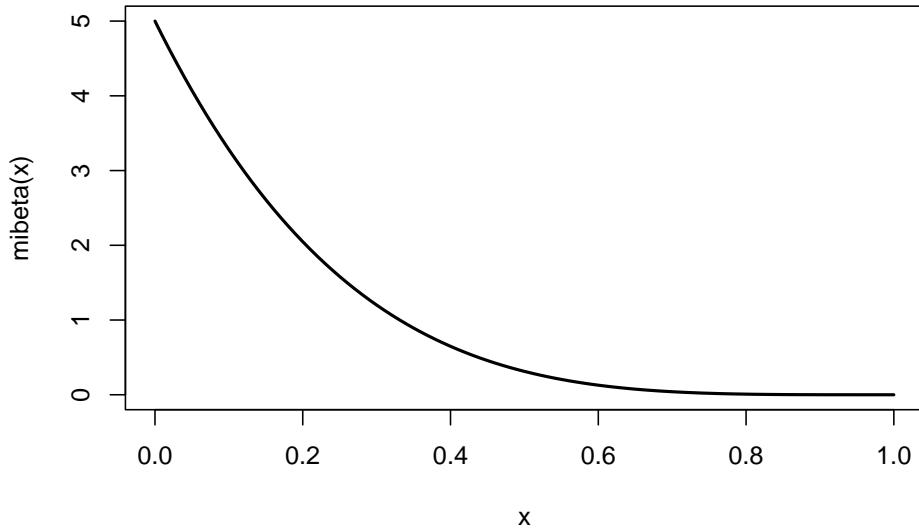
Ejemplo

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

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

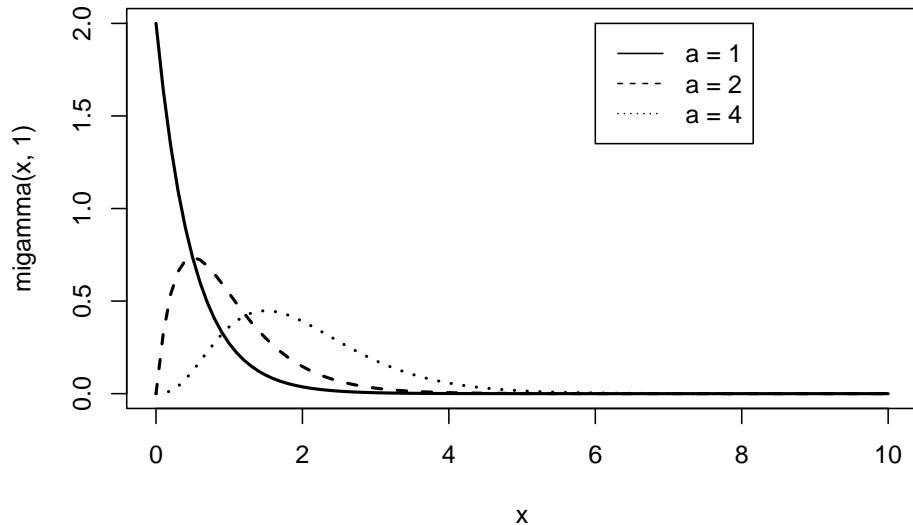
Características

- Esperanza: $E[X] = \frac{a}{b}$
- Varianza: $Var[X] = \frac{a}{b^2}$
- $\int_0^{\infty} x \{ -1 \} e^{-bx} dx$
- 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

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

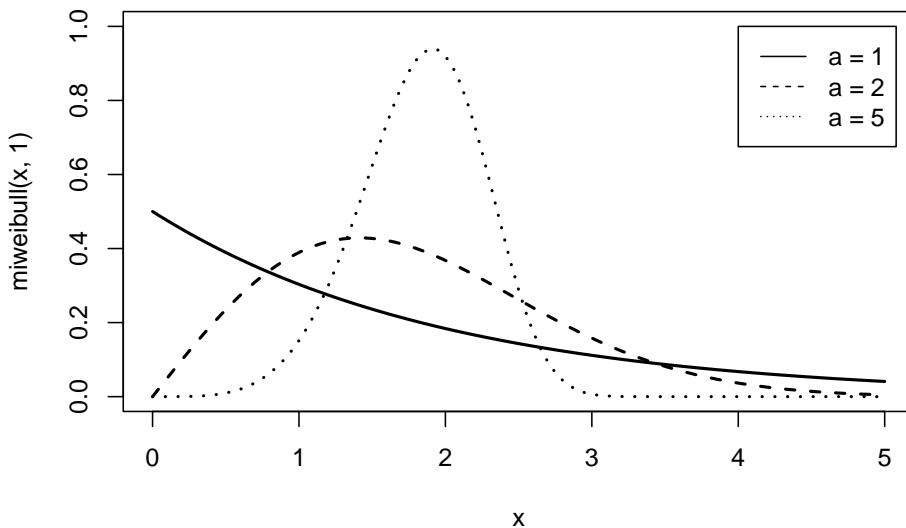
Características

- 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

En 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

- Allen, T. T. (2010). *Introduction to Engineering Statistics and Lean Six Sigma - Statistical Quality Control and Design of Experiments and Systems*. Springer.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*.
- 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.
- Corbalán, F. and Sanz, G. (2010). *La conquista del azar. La teoría de probabilidades*. RBA.
- de Finetti, B. (1992). *Foresight: Its Logical Laws, Its Subjective Sources*, pages 134–174. Springer New York, New York, NY.
- Faraway, J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.
- Feys, J. (2016). Nonparametric tests for the interaction in two-way factorial designs using r. *The R Journal*, 8(1):367–378.
- Giraudoux, P. (2021). *pgirmess: Spatial Analysis and Data Mining for Field Ecologists*. R package version 1.7.1.
- Horikoshi, M. and Tang, Y. (2022). *ggfortify: Data Visualization Tools for Statistical Analysis Results*. R package version 0.4.14.
- ISO (2010). UNE-ISO 16269-4:2010 interpretación estadística de datos – parte 4: Detección y tratamiento de valores atípicos. Norma Internacional.
- Lawson, J. (2015). *Design and Analysis of Experiments with R*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

- 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.
- Matejka, J. and Fitzmaurice, G. (2017). Same Stats, Different Graphs. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1290–1294, New York, NY, USA. ACM.
- Moen, R., Nolan, T., and Provost, L. (2012). *Quality Improvement Through Planned Experimentation 3/E*. McGraw-Hill Education.
- 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.
- Sarasola, J. (2018). Cálculo de la moda para datos agrupados en intervalos.
- 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.
- Wikipedia (2018). Navaja de ockham — wikipedia, la enciclopedia libre. [Internet; descargado 25-enero-2019].
- Wikipedia (2019). Paradoja de simpson — wikipedia, la enciclopedia libre. [Internet; descargado 25-enero-2019].
- Xie, Y. (2021). *bookdown: Authoring Books and Technical Documents with R Markdown*. <https://github.com/rstudio/bookdown>.