Notas de Curso

Elementos de Estadística y Probabilidad

Miguel Nakamura

Centro de Investigación en Matemáticas, A.C.

María Guadalupe Russell

Universidad Autónoma de Sinaloa

2019

Índice general

Capítulo 1

Introducción

1.1. Probabilidad y estadística

En la naturaleza y dada la posición como observador de la misma que tiene el ser humano, existen fenómenos que se llaman *aleatorios*. El que sean aleatorios, significa simplemente que no es posible predecir con entera exactitud sus resultados. Debemos notar que el que un fenómeno sea aleatorio no necesariamente es una propiedad del fenómeno mismo, sino de la posición relativa del observador. Por ejemplo, el fenómeno llamado eclipse solar ha dejado de ser aleatorio porque con leyes de mecánica celeste hoy día se le puede predecir con entera certeza; pero para alguien quien no conoce de mecánica celeste, el eclipse solar es esencialmente un fenómeno aleatorio. Los números aleatorios generados por una computadora en el fondo no son aleatorios, porque se producen con una función recursiva. Pero al no conocer explícitamente dicha función recursiva, sí resultan aleatorios en el sentido de que no podemos predecir la secuencia, y se usan en los programas de cómputo para *simular* fenómenos aleatorios.

Ante la presencia de un fenómeno aleatorio, hay dos actitudes posibles que podríamos tomar:

1. Resignarse («no hay nada que yo pueda hacer al respecto»).

2. Reconocer que la aleatoriedad está presente, enfrentarla, y proceder a cuantificarla.

La primera actitud es pasiva, y de tono conformista, mientras que la segunda es proactiva, realista, y científica. En este curso veremos que la segunda actitud se aborda matemáticamente a través de una abstracción llamada *modelo de probabilidad* $(\Omega, \mathscr{A}, \mathbb{P})$ donde Ω representa los resultados posibles del fenómeno aleatorio, $\mathscr{A}$ representa un sistema de subconjuntos de Ω llamados *eventos* cuyo azar es de interés cuantificar, y $\mathbb{P} : \mathscr{A} \to \mathbb{R}$ es la función que asigna magnitud a cada evento. Se hablará entonces de $\mathbb{P}(A)$, $\forall A \in \mathscr{A}$. La función $\mathbb{P}$ recibirá el nombre de *medida de probabilidad*.

Este modelo de probabilidad realiza todo lo que un observador puede hacer con respecto a un fenómeno aleatorio, es decir, permite cuantificar la incertidumbre a través de la especificación de probabilidades a los eventos. Todo lo demás —es decir, la predicción del resultado específico de un fenómeno aleatorio— sería obra de profetas, adivinos, o divinidades supremas. En el momento de que un fenómeno sea predecible, deja de ser relevante el concepto de modelo de probabilidad (o por lo menos, útil), ya que el fenómeno deja de ser aleatorio por definición.

La teoría de **probabilidad** tiene por objeto estudiar las propiedades del modelo $(\Omega, \mathscr{A}, \mathbb{P})$. Por ejemplo, si se conoce el valor de $\mathbb{P}(A)$ para ciertos tipos de eventos A, el interés puede radicar en deducir o calcular el valor de $\mathbb{P}(B)$ para eventos que son de estructura más compleja que los A. En teoría

de probabilidad, la medida $\mathbb{P}$ es especificada en forma general, y se buscan propiedades matemáticas del objeto $(\Omega, \mathscr{A}, \mathbb{P})$. Como veremos más adelante, la **estadística** más bien tiene que ver con preguntarse cuál $(\Omega, \mathscr{A}, \mathbb{P})$ es la apropiada para una situación práctica dada, cuando se tiene acceso a la observación de una realización del fenómeno aleatorio bajo estudio.

En ocasiones es posible que ante un fenómeno aleatorio dado, el objeto $(\Omega, \mathscr{A}, \mathbb{P})$ sea susceptible de ser especificado desde un principio (Ejemplos: Lanzamiento de una moneda y lanzamiento de un dado perfectamente balanceado). En tales casos la estadística propiamente dicha no es necesaria para investigar la naturaleza de $(\Omega, \mathscr{A}, \mathbb{P})$. Simplemente se procede a utilizar matemáticamente el modelo de probabilidad, usándolo para cuantificar la ocurrencia de los eventos (Ejemplo: Probabilidad de lanzar dos águilas seguidas, *etc*.). Sin embargo, en otras ocasiones, la naturaleza adecuada del objeto $(\Omega, \mathscr{A}, \mathbb{P})$ para un fenómeno aleatorio —y muy particularmente, la medida de probabilidad $\mathbb{P}$—, no es muy clara *a priori*. Es en estos casos en los que la estadística cumple su función, y para ello se basa en el precepto de que es posible obtener observaciones del fenómeno aleatorio para tratar de inferir las propiedades de $\mathbb{P}$. En este sentido, la disciplina llamada estadística tiene inherente el concepto de observaciones del fenómeno aleatorio (datos), mientras que en probabilidad, por su naturaleza, las observaciones mismas ya realizadas no son parte inherente de la teoría.

Si no hay posibilidad de hacer observaciones empíricas del fenómeno aleatorio, y no se conoce $\mathbb{P}$ explícitamente, entonces no puede hacerse estadística, y dotados con teoría de probabilidad únicamente, sólo estaríamos capacitados para hacer aseveraciones del tipo «*si* la medida de probabilidad fuera $\mathbb{P}$, entonces la probabilidad de que ocurra el evento A es ...».

Ejemplo 1.1 Supongamos que una urna contiene 5 bolas, de colores blanco y negro. Supongamos que el evento de interés es «obtener una bola negra al azar». Para explicar la diferencia entre razonamiento de probabilidad y de estadística, pensemos en las siguientes dos situaciones. En la primera situación se sabe que hay 3 bolas blancas y 2 bolas negras. La pregunta que aborda la teoría de probabilidad es ¿cuál es la probabilidad de obtener una bola negra? Para ello, calcula la probabilidad de obtener una bola negra, que es 2/5. En la segunda situación, no se sabe la composición de colores dentro de la urna. Supongamos que al tomar una bola al azar de la urna se obtiene una bola negra. Entonces la pregunta es ¿cuántas bolas negras habrá en la urna dado que observé una bola negra? Esta segunda pregunta es muy diferente. Mientras que la respuesta a la primera pregunta es 2/5 sin lugar a dudas, la respuesta a la segunda pregunta implica incertidumbre, ya que hay varias configuraciones de urnas que pudieran haber dado como resultado haber observado una bola negra. Es por esta razón que en la primera situación se habla de un razonamiento *deductivo*, mientras que en la segunda se habla de un razonamiento *inferencial*.

1.2. Ejemplos de inferencia estadística

1. Una urna contiene 10, 000 bolas, de las cuales algunas son negras y otras son blancas. La pregunta es: ¿Cuántas de las bolas son negras? Existe una manera de averiguar la respuesta a esta pregunta, de tal manera que se pueda ésta contestar absolutamente sin error. Esta manera es contarlas. Otra manera de contestar la pregunta, aunque sea en forma incierta, es mediante la selección al azar de un número de bolas, digamos $n = 10$, y según el resultado, inferir cuál es la com-

posición de la totalidad de la urna. Esto se llama procedimiento de muestreo, y permite contestar la pregunta concediendo cierta incertidumbre, pero analizando sólo una muestra del total de bolas. Al cambiar la urna de pelotas por una población de individuos, obtenemos una situación análoga a la de indagar acerca de alguna característica de los individuos por medio de una encuesta. Al cambiar la urna de pelotas por un lote de producción en una fábrica, obtenemos una situación análoga a la de examinar al azar una muestra de artículos fabricados con el fin de determinar la calidad del lote producido.

2. Una extensión de tierra en la que se sospecha contaminación por metales pesados, se estudia con el objeto de cuantificar la magnitud de la contaminación. Sobre una retícula regular de 10×10 puntos con separación de 50m se realizan mediciones (altamente costosas) de concentración de metales pesados sobre muestras de suelo. La pregunta es: ¿Cuál es la concentración de metales pesados en un punto intermedio entre puntos muestreados? Existe una manera de averiguar la respuesta a esta pregunta, de tal manera que se pueda ésta contestar absolutamente sin error. Esta manera es medir sobre el punto en cuestión. Otra manera de contestar la pregunta, es utilizar las mediciones obtenidas sobre la retícula para inferir el valor que ha de tener en el punto intermedio. La respuesta obtenida por este segundo medio, es susceptible de contener error. La rama de la estadística que aborda este tipo de problemas es conocida como *estadística espacial*.

3. Un medicamento de reciente desarrollo se desea contrastar contra el mejor tratamiento alternativo. La pregunta es: ¿El nuevo medicamento es mejor que el anterior? En este caso puede o no ser practicable

un procedimiento para contestar la pregunta sin error. Por ejemplo, si se trata de una afección dermatológica, puede pensarse en comparar ambos medicamentos con sujetos de prueba, pero si se trata de una enfermedad terminal, no existiría la posibilidad. Aún así, no será practicable experimentar con la población de seres humanos, por lo que necesariamente deberá contemplarse una muestra. La respuesta a la pregunta conlleva de manera obligatoria algún grado de incertidumbre.

4. ¿Cuál será la paridad del Peso Mexicano para el día 1 de enero del año próximo? Una forma de contestar con entera certeza la pregunta anterior es esperarse a que suceda el día 1 de enero en cuestión y observar entonces la paridad. Sin embargo, esta solución no es aceptable para aquellas instancias en las que sea necesario realizar una planeación futura o bien realizar algún tipo de pronóstico para tomar una decisión. En este caso, es intenta contestar la pregunta bajo condiciones de incertidumbre, lo cual se basa en el análisis de observaciones de la paridad, y la determinación de su relación con otras variables económicas y políticas. La observación histórica constituye la fuente de datos, y puede considerarse una observación muestral del fenómeno de interés.

¿Qué tienen en común estos ejemplos?

(i) Que todos de ellos contienen una noción de *incertidumbre*, en el sentido de que la respuesta a la pregunta original por vía de análisis de unas cuantas observaciones del fenómeno conlleva naturalmente la posibilidad de ser imprecisa.

(ii) Que está presente el concepto de *azar*, en el sentido que la recolección de datos da lugar a una observación de un fenómeno aleatorio. Y si se trata de un fenómeno aleatorio, entonces es representable matemáticamente mediante un modelo de probabilidad.

(iii) También tienen en común estos ejemplos, la noción de una población total acerca de la cual es formulada alguna pregunta de interés, y que por circunstancias diversas, la metodología que conduce a la obtención de la respuesta exacta a dicha pregunta formulada es una cuestión imposible de realizar, sea por impedimentos físicos, o económicos.

Una solución basada en el concepto de muestreo, si bien es imprecisa, por lo menos proporciona una solución que es viable, y en muchas ocasiones, no sólo es viable sino que es la única posible. Como veremos a lo largo del curso, la estadística matemática se utiliza para cuantificar la imprecisión en la que se incurre cuando se utiliza un método basado en muestreo u observación incidental de realizaciones de un fenómeno aleatorio, para obtener respuestas.

1.3. Modelación matemática

El objetivo de un modelo matemático es representar alguna faceta de la realidad a través de una abstracción. Un modelo permite manipular artificialmente aspectos de la realidad, con el objeto de obtener respuestas. En una aplicación de matemáticas, existen pues, dos mundos: el de la realidad, y el del modelo matemático que lo representa. En el mundo real típicamente existe formulada alguna pregunta de interés. Si la respuesta a la pregunta puede obtenerse manipulando la realidad, entonces no es útil ni necesario el

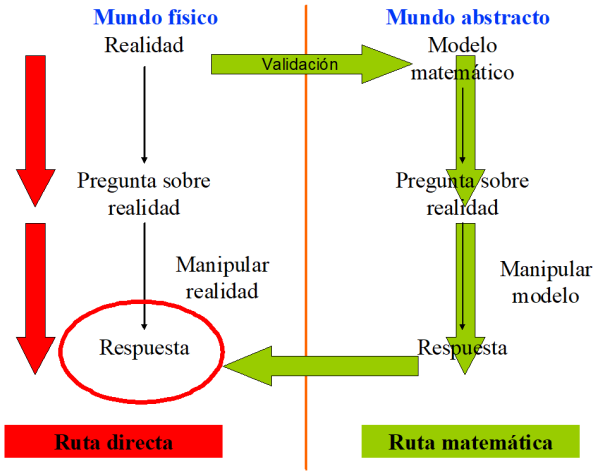

Figura 1.1: Esquema conceptual modelos matemáticos.
Los modelos matemáticos permiten manipular de manera virtual una representación de la realidad para fines de obtener respuestas a preguntas de interés.

concepto de un modelo matemático. Con un modelo matemático, podemos estudiar la realidad con una representación abstracta. En la medida en que el modelo matemático represente adecuadamente a la realidad, las respuestas que obtengamos con el modelo matemático serán también apegadas a la realidad, y por lo tanto, útiles.

El contraste entre el modelo matemático y la realidad recibe el nombre de *validación*. Algunos modelos matemáticos requieren de muy poca validación, porque la identificación entre realidad y modelo es muy clara. Pero otros modelos requieren de validación más cuidadosa y elaborada. En particular, en los modelos estadísticos y probabilísticos, siendo el objeto de estudio el concepto de aleatoriedad, las nociones de validación se concentran en aspectos de aleatoriedad, o variación de un fenómeno aleatorio.

Para obtener una respuesta usando un modelo matemático se recorre la

siguiente cadena de eslabones (ver Figura 1.1): Se representa la realidad con un modelo matemático; se identifica la pregunta de interés en la realidad con una formulación que la representa en el modelo matemático; se manipula el modelo matemático para obtener una respuesta; se identifica la respuesta obtenida con la realidad.

El aspecto de la realidad que es de interés modelar en este curso es el aleatorio. Es decir, lo que nos va a interesar es estudiar modelos matemáticos cuyo objeto es describir la aleatoriedad de un fenómeno. En otras ramas de la matemática, los aspectos de la realidad pueden ser otros, para lo cual se desarrollan modelos específicos para cada fin. Cabe mencionar que un modelo matemático de un fenómeno complejo puede contener a la vez componentes de varias disciplinas matemáticas. Por ejemplo, en un modelo matemático que explique el crecimiento de cierto organismo biológico, puede involucrarse una componente matemática derivada de teoría de ecuaciones diferenciales, y para explicar la variación natural observada en la naturaleza, podría involucrarse también un modelo probabilístico.

1.4. Ejemplos de modelación matemática

1. Un ejercicio típico que se emplea para ilustrar trigonometría es calcular la altura de un árbol (o de un asta bandera) mediante la medición del ángulo que se forma entre la punta del árbol y un punto fijo sobre el suelo, situado a distancia conocida de su base. La realidad es el árbol, la pregunta de interés es su altura, y el modelo matemático consiste de un triángulo rectángulo. La manipulación del modelo matemático consiste en realizar cálculos que aprovechan propiedades matemáticas de los triángulos rectángulos. La identificación entre

realidad y modelos es muy sencilla: uno de los catetos del triángulo es la altura del árbol. Una vez calculada una respuesta (la longitud desconocida del cateto correspondiente usando resultados de trigonometría), una simple identificación transfiere la respuesta obtenida a la realidad. En este caso, la validación es muy directa, pues basta notar que si el suelo es horizontal y el árbol crece siguiendo una vertical, que en efecto se forma un triángulo rectángulo. El hecho de recurrir a un modelo matemático (trigonometría, en este caso), evita la obtención de la respuesta en el ámbito del mundo real, es decir, evita recurrir a la medición directa de la altura del árbol.

2. El sencillo acto de calcular el número de metros cuadrados de loseta para piso que deben adquirirse para instalar en una habitación, constituye un ejemplo de un modelo matemático. En efecto, el modelo matemático es un rectángulo, y la respuesta se obtiene calculando su área. No es necesario manipular a la realidad misma (es decir, la habitación) para obtener la respuesta. Confiamos en la respuesta que da un modelo matemático, porque hacemos de hecho una validación implícita: verificar si la forma de la habitación es rectangular o no.

3. En un proceso industrial de producción, habrá algunas características del producto terminado que dependen de variables durante la fabricación. Por ejemplo, las características metalúrgicas y físicas de un acero dependen de temperaturas, de tipos de procesos, de proporciones de mezcla, y otros factores diversos. Una problema natural en este tipo de proceso industrial es saber los mejores niveles en los que conviene fijar las variables de producción para provocar alguna característica específica en el producto terminado (por ejemplo, máxima resistencia).

Si el número de factores es pequeño, es concebible que por experimentación y un procedimiento de ensayo-error, puede investigarse la respuesta a esta pregunta. Sin embargo, si el número de factores es sustancial, un procedimiento experimental exhaustivo puede bien no ser factible. En este caso un modelo matemático que logre representar la relación entre la característica de interés y los factores de fabricación, sería muy valioso para investigar cuál es la combinación óptima. Este es un ejemplo de un modelo matemático usado para fines de *optimización*.

4. Otro uso común que tiene un modelo matemático es para abordar preguntas del tipo ¿Qué pasaría si ...?. Ejemplos de actualidad y alta complejidad radican en medio ambiente y climatología (¿que pasaría con los casquetes polares si los niveles de bióxido de carbono exceden cierto nivel?), biología (¿qué pasaría con una especie vegetal si sucede un cambio climático en términos de temperatura y precipitación?), o ingeniería (¿qué pasaría si a un edificio lo golpearan rachas de viento de 80 km/hora?). Es claro que en ninguna de estas situaciones puede realizarse físicamente el experimento en escala real. Así, para contestar este tipo de preguntas es necesario recurrir a modelos matemáticos.

1.5. «Regularidad estadística»

¿La aleatoriedad tiene algún aspecto que sea susceptible de modelarse matemáticamente? La respuesta es que sí. Cuando hablamos en sentido coloquial con la palabra *aleatoriedad* o *azar*, puede sugerirse que por tratarse

de algo impredecible, que entonces no hay nada que pueda hacerse matemáticamente. Sin embargo, existe una característica de los fenómenos al azar que se presta a ser considerada con modelos matemáticos. Esta característica se ha tratado de describir en términos generales, como *regularidad estadística*. Para ilustrar el concepto de regularidad estadística, consideremos los siguientes ejemplos.

1. Supongamos que un grupo de 40 alumnos se divide en dos partes iguales. Al primer grupo se le asignará de tarea realizar 200 lanzamientos de una moneda y de registrar en una hoja de papel la serie completa de ocurrencias de «águila» y «sol». Al segundo grupo de alumnos se le asigna de tarea «inventarse» la serie de 200 lanzamientos de una moneda, y registrar asimismo la serie en una hoja de papel. Al día siguiente, el profesor de estadística toma los 40 resultados, y tras breve análisis logra discernir entre quienes lanzaron una moneda de verdad y quienes inventaron los lanzamientos, con una efectividad de cerca de 98 %. (es decir, que el profesor se equivoca aproximadamente en uno o dos alumnos). Este ejemplo demuestra que el azar legítimo tiene alguna peculiaridad que al segundo grupo de alumnos le resulta muy difícil de imitar. Ver Révész (1978) y Schilling (1990).

2. La «Tabla de Galton». Se trata de un dispositivo físico (que muchos museos de divulgación de ciencia exhiben físicamente), consistente en un sistema reticulado de pernos dispuestos de forma triangular, a través del cual se dejan caer pelotas o balines. Las pelotas van descendiendo y cuando chocan contra un perno, toman direcciones al azar hacia la izquierda o la derecha. En la parte inferior hay un sistema de compartimientos que van recolectando las pelotas cuando éstas terminan su

recorrido. Cuando un número grande de pelotas se deja caer por este aparato, los números recolectados en los compartimientos van conformando una distribución, invariablemente de forma acampanada (los museos usan la Tabla de Galton para ilustrar el concepto de distribución normal). El modelo matemático que explica esta distribución predice una distribución binomial y aproximadamente normal, y el hecho que invariablemente se obtiene esta distribución cuando se repite el experimento es la regularidad estadística. Si bien es cierto que es difícil predecir en cuál compartimiento caerá una pelota en particular, algo que por regularidad estadística sí es fácil predecir es la forma que constituirá un gran número de pelotas.[1]

Notemos que los ejemplos anteriores, lo que demuestran es que existe una característica del azar que la delata. Esta característica es la que se presta a ser modelada matemáticamente, y aquella característica del azar que la delata radica en el concepto de probabilidad. Es decir, el azar no es tan azaroso como uno pudiera inicialmente creer. A pesar de que azar significa por definición que no lo puedo predecir, sí existe «algo» del azar que puede cuantificarse, e inclusive predecirse. Este «algo» es lo que se intenta describir por *regularidad estadística*.

El objetivo del presente curso es proporcionar una introducción a dos disciplinas que se relacionan íntimamente: probabilidad y estadística. En programas convencionales, es usual que primero se abarque por completo aquella disciplina llamada probabilidad, y sólo hasta tenerla completamente desarrollada, se contemplan nociones de inferencia estadística. La filosofía

[1] Este es el fenómeno de indagación por encuesta de una población de individuos: si bien es difícil predecir la respuesta de un individuo en particular, con base en una encuesta realizada a un gran número de individuos, sí podemos hacer una predicción acerca del comportamiento global.

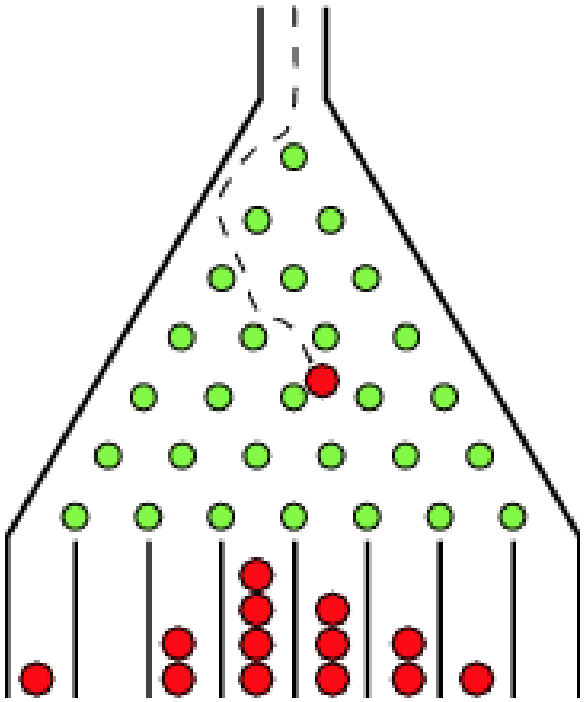

Figura 1.2: La Tabla de Galton
Se trata de un dispositivo físico para realizar un experimento aleatorio basado en pelotas que descienden al azar por un sistema de hileras formadas por pernos. El canal final en el cual una pelota resultará localizada es impredecible, pero cuando se realiza el experimento con un número grande de pelotas se produce un patrón regular y predecible en cuanto a su distribución.

durante el presente curso será distinguir desde un principio la diferencia entre un problema probabilístico y un problema estadístico, así como cubrir algunas nociones básicas de los modelos matemáticos que se emplean para resolver estos dos tipos de problemas.

Es decir, este curso de hecho consiste de una introducción a dos cursos subsecuentes: teoría de probabilidad, y teoría de inferencia estadística. A lo largo del curso se enunciarán algunos resultados matemáticos importantes, aunque no siempre se demostrarán. En los cursos subsecuentes que se mencionan, se profundizará con mayor rigor en los temas expuestos aquí, incluyendo la demostración de varios resultados que requieren del concurso de otros conceptos de matemáticas generales. El énfasis aquí es avanzar en las introducciones de ambos cursos, con el fin de proporcionar una visión integral alrededor de la interrelación que existe entre ambos temas —probabilidad y estadística— así como con otras ramas de la matemática que se cultivarán durante la carrera de licenciatura en matemáticas.

Ejercicios

1.1 Describa ejemplos de fenómenos aleatorios que se suscitan en las siguientes disciplinas: medicina, ingeniería civil, biología, contabilidad, economía, política, física, y meteorología.

1.2 Programe una simulación de lanzamientos de moneda en la computadora. Por medio de repeticiones, obtenga la probabilidad de que en 200 lanzamientos, ocurra por lo menos una corrida de siete soles o siete águilas.

1.3 Consiga un simulador del experimento de la Tabla de Galton. Un ejem-

plo está dado en

`www.math.uah.edu/stat/applets/GaltonBoardExperiment.xhtml`

Realice experimentos diversos cambiando el número de pelotas, así como la probabilidad de un rebote a la derecha en cada perno, para observar el fenómeno de regularidad estadística.

Capítulo 2

Modelos de probabilidad

En este capítulo se definirán los ingredientes de un modelo de probabilidad. En el Capítulo 3 se darán herramientas para especificar ciertos modelos de probabilidad en algunos casos especiales, y en el Capítulo 4 se estudiarán propiedades matemáticas de los modelos de probabilidad en general. En lo que sigue, se presuponen conocidos conceptos elementales y la notación empleada en teoría de conjuntos, como serían unión, intersección, complementos, *etc*.

2.1. Espacio muestral

Definición 2.1 (Experimento) Un *experimento* es un fenómeno que al observarse da lugar a una realización que en esencia es impredecible.

Definición 2.2 (Espacio muestral) Se llama *espacio muestral*, al conjunto formado por todos los resultados posibles de un experimento. Denotaremos al espacio muestral por Ω.

Ejemplo 2.1 Lanzar una moneda. El espacio muestral consiste de dos resultados, que pueden denominarse «águila» y «sol». Por lo tanto, un espacio muestral para representar este experimento es $\Omega = \{\text{A, S}\}$, donde A representa águila y S representa sol.

Ejemplo 2.2 Lanzar un dado. Un espacio muestral es

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

donde cada número representa el número de puntos que muestra la cara superior del dado.

Ejemplo 2.3 Lanzar moneda hasta que aparezca un sol. Un espacio muestral es $\Omega = \{1, 2, 3, 4, 5, 6, \dots\}$. Cada entero representa el número de veces que es necesario lanzar la moneda para que aparezca el sol. También podríamos usar

$$\Omega = \{\text{S, AS, AAS, AAAS}, \dots\},$$

siendo obvio que es más compacta la primera opción. En este ejemplo, el espacio muestral es infinito, numerable. Note que en este espacio muestral existen resultados que son sumamente improbables. Por ejemplo, tener que lanzar la moneda tres millones de veces para obtener el sol, es posible, aunque muy improbable.

Ejemplo 2.4 Distancia recorrida por un automóvil con un litro de gasolina. En este caso, el resultado del experimento es una distancia, y una distancia puede tomar valores sobre un continuo. El espacio muestral puede tomarse

como $\Omega = [0, \infty)$. Éste es un ejemplo de un espacio muestral no numerable.

Ejemplo 2.5 Un mismo experimento puede tener dos o más espacios muestrales. Hablamos de «un» espacio muestral en lugar de «el» espacio muestral, porque la elección de espacio muestral puede no ser única. Por ejemplo, si el experimento es lanzar una moneda, alguien podría proponer el espacio muestral $\Omega = \{\text{A, S, C}\}$, donde C representa caer de canto. La especificación de un espacio muestral, es de hecho, el primer acto de modelación de un fenómeno aleatorio. Esto es, el espacio muestral es una abstracción matemática que hacemos respecto a un fenómeno aleatorio, con fines de sentar una base para modelarlo matemáticamente. Desde el punto de vista de taxonomía matemática, un espacio muestral es simplemente un conjunto en abstracto.

Ejemplo 2.6 Una señal de radio se recibe durante dos segundos. Este experimento consiste de observar una función del tiempo. Los posibles resultados son funciones sobre el intervalo de tiempo $[0, 2]$. El espacio muestral Ω para este experimento es un conjunto de *funciones* sobre $[0, 2]$.

Ejemplo 2.7 El experimento es observar el estado del tiempo del día de mañana. Es espacio muestral es el conjunto de todos las posibles maneras en que pueda suceder el estado del tiempo de mañana. Este conjunto es sumamente complejo, y de hecho, resulta muy difícil describirlo. En él se encuentra una cantidad infinita de posibilidades, y cada una de éstas es difícil de describir con exactitud. Ahora bien, si el experimento consistiera solamente en observar si mañana llueve o no, entonces el espacio muestral

es muy sencillo: $\{\text{llueve},\ \text{no llueve}\}$.

2.2. σ-álgebras

Definición 2.3 (Clase de subconjuntos) Sea Ω un espacio muestral arbitrario. Cualquier conjunto de subconjuntos de Ω recibe el nombre de *clase*. Una clase es entonces un conjunto de subconjuntos. Las clases usualmente se denotan con símbolos caligráficos, como $\mathscr{A}, \mathscr{B}, \mathscr{C}, \mathscr{D}$, *etc*. para enfatizar que son conjuntos cuyos elementos son conjuntos, los cuales son denotados por A, B, C, D, *etc*.

Ejemplo 2.8 Si $\Omega = \{1, 2, 3, 4\}$, las siguientes son ejemplos de clases de subconjuntos de Ω: $\big\{\{1\}, \{1, 2\}, \{3\}\big\}$, $\big\{\{1, 2\}, \{4\}, \varnothing\big\}$, $\{\varnothing\}$, y $\big\{\{1, 2, 3, 4\}\big\}$. Note que $\{\varnothing\} \neq \varnothing$ y que $\{\Omega\} \neq \Omega$.

Definición 2.4 (Complemento) Si Ω es un espacio muestral, y A un subconjunto de Ω, llamamos *complemento* de A al conjunto $\big\{\omega \in \Omega \mid \omega \notin A\big\}$. El complemento de A será denotado por A^c. Aunque la notación no es explícita, se entiende que esta noción se define con respecto al conjunto Ω.

Definición 2.5 (Clase potencia) Sea Ω un espacio muestral. El *conjunto potencia*, o la *clase potencia* de Ω, consiste de todos los subconjuntos de Ω (incluyendo Ω mismo, así como el vacío, $\varnothing$). La clase potencia de Ω se denota por 2^{Ω}.

Definición 2.6 (σ-álgebra) Una clase $\mathscr{A}$ de subconjuntos de Ω se dice ser

una *σ-álgebra*, si se cumplen las siguientes tres condiciones:

(a) $\Omega \in \mathscr{A}$,

(b) $A \in \mathscr{A} \Rightarrow A^c \in \mathscr{A}$, y

(c) $A_i \in \mathscr{A}$, $i = 1, 2, 3, \ldots \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathscr{A}$.

Proposición 2.1 *Si $\mathscr{A}$ es una σ-álgebra, entonces $\varnothing \in \mathscr{A}$.*

Demostración. Ejercicio. □

Ejemplo 2.9 Si $\Omega = \{1, 2, 3, 4\}$, la clase $\big\{\{1\}, \{2\}\big\}$ no es una σ-álgebra, mientras que la clase $\big\{\{1\}, \{2, 3, 4\}, \{1, 2, 3, 4\}, \varnothing\big\}$ sí lo es. En este caso, $2^{\Omega} = \big\{\{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{2, 3, 4\}, \{1, 2, 3, 4\}, \varnothing\big\}$.
Note de paso que la clase potencia contiene $16 = 2^4$ elementos.

Proposición 2.2 *Si $A_1, A_2, \ldots \in \mathscr{A}$, y $\mathscr{A}$ es una σ-álgebra, entonces*

(a) $\cup_{i=1}^{n} A_i \in \mathscr{A}$, $\forall n$,

(b) $\cap_{i=1}^{n} A_i \in \mathscr{A}$, $\forall n$, *y*

(c) $\cap_{i=1}^{\infty} A_i \in \mathscr{A}$.

Demostración. Ejercicio. □

Como consecuencia de esta proposición, se concluye que una σ-álgebra es cerrada bajo todas las operaciones ordinarias finitas y numerables de teoría de conjuntos.

Ejemplo 2.10 En teoría de conjuntos, la diferencia simétrica entre dos conjuntos A y B se define como $A \triangle B = (A \cup B) - (A \cap B)$. Si $A, B \in \mathscr{A}$ y $\mathscr{A}$ es σ-álgebra, entonces $A \triangle B \in \mathscr{A}$.

¿Cuántos elementos contiene una σ-álgebra? La respuesta a esta pregunta es que hay dos extremos: contiene un número finito de elementos, o bien tiene por lo menos la cardinalidad del continuo, como lo asevera el siguiente resultado.

Proposición 2.3 *La cardinalidad de una σ-álgebra es finita, o tiene por lo menos la cardinalidad del continuo.*

Esto quizás explique la razón por la cual el estudio de la teoría de probabilidad se polariza hacia dos extremos contrastantes: se estudia probabilidad sólo sobre σ-álgebras finitas, o bien se estudia probabilidad sobre σ-álgebras infinitas (que necesariamente son entonces no-numerables). Como podría esperarse, el primer caso puede hacerse con relativa facilidad, sin que sea necesario invocar resultados profundos de matemáticas, mientras que para el segundo caso se requiere de herramientas más sofisticadas.

Correspondientemente, existen entonces dos estrategias para abordar el estudio de la teoría de probabilidad. La primera es posponer el estudio for-

mal de teoría de probabilidad por completo, hasta un momento en que se cuente con herramienta matemática necesaria para considerar de una vez los casos más elaborados. La segunda estrategia es estudiar única y exclusivamente probabilidad sobre σ-álgebras finitas. Un problema con la primera estrategia es que se pospone innecesariamente el estudio de probabilidad y estadística, y un problema con la segunda es que la teoría de probabilidad puede aparentar tener mucho menor alcance del que realmente posee.

Por otra parte, abordar el estudio de la materia llamada estadística, también puede realizarse con dos mentalidades extremosas: se estudia estadística sin hacer alusión a teoría de probabilidad, o bien se estudia estadística tomando en cuenta la infraestructura que proporciona la probabilidad. El concepto moderno de estadística está íntimamente ligado al de probabilidad, de modo que un curso de estadística que no tiene pre-requisitos de probabilidad es un concepto obsoleto. Muchos cursos de estadística en la actualidad, especialmente los que se usualmente se imparten a nivel preparatoria, todavía no contienen un panorama que incluye teoría de probabilidad, lo cual puede dotar al estudiante de una visión limitada y equivocada acerca de esta disciplina.

Durante este curso procuraremos hacer algo intermedio. Abordaremos definiciones generales de probabilidad. Los casos de σ-álgebras finitas serán considerados como ejemplos importantes, y mantendremos siempre a la vista las dificultades inherentes en la definición de probabilidad sobre espacios más generales. Con esto, se espera que al arribar a un curso futuro de probabilidad, ya se haya contado con tiempo para madurar muchos de los conceptos generales, a la vez que pueden abordarse los problemas de estadística de una vez.

Definición 2.7 (σ-álgebra generada por una clase) Sea $\mathscr{C}$ una clase de subconjuntos de Ω. Llamamos *σ-álgebra generada por $\mathscr{C}$* a una σ-álgebra, denotada por $\sigma(\mathscr{C})$, que cumple lo siguiente:

(a) $\mathscr{C} \subset \sigma(\mathscr{C})$,

(b) si $\mathscr{A}$ es cualquier otra σ-álgebra tal que $\mathscr{C} \subset \mathscr{A}$, entonces $\sigma(\mathscr{C}) \subset \mathscr{A}$.

Este concepto tiene importancia matemática, porque un modelo de probabilidad se construirá sobre una σ-álgebra generada. El siguiente teorema muestra una manera de construir $\sigma(\mathscr{C})$. También dota de interpretación inmediata a $\sigma(\mathscr{C})$, porque ésta no es más que «la σ-álgebra más pequeña que contiene a $\mathscr{C}$».

Teorema 2.4 *Sea $\mathscr{C}$ una clase de subconjuntos de Ω. Sea*

$$\mathbb{X} = \left\{ \mathscr{S} \mid \mathscr{S} \text{ es } \sigma\text{-álgebra y } \mathscr{C} \subset \mathscr{S} \right\}$$

. Entonces, $\sigma(\mathscr{C}) = \cap_{\mathscr{S} \in \mathbb{X}} \mathscr{S}$.

Definición 2.8 (σ-álgebra trivial) A la clase de subconjuntos $\{\Omega, \varnothing\}$ se le conoce como *σ-álgebra trivial*.

En teoría de probabilidad más avanzada, el concepto de σ-álgebra generada adquirirá aun más relevancia que la que se alcanzará a vislumbrar en este momento. La idea es que para construir medidas de probabilidad, podrá usarse como base una clase $\mathscr{C}$ de subconjuntos de Ω sobre la cual sea

relativamente fácil definir probabilidad, para luego proceder a extender la probabilidad sobre todo $\sigma(\mathscr{C})$.

Por aparte de su relevancia estrictamente matemática, podríamos también justificar intuitivamente el concepto de σ-álgebra generada como sigue: Supongamos que en una aplicación concreta hemos identificado Ω, así como una colección $\mathscr{C}$ de subconjuntos de Ω que es de interés estudiar. Entonces $\sigma(\mathscr{C})$ es la colección de todos los conjuntos que se construyen con operaciones elementales (uniones, intersecciones, complementaciones) sobre los elementos de $\mathscr{C}$, y $\sigma(\mathscr{C})$ es también la más chica de las estructuras matemáticas que contienen a $\mathscr{C}$. Uno podría preguntarse por qué no tomamos siempre la σ-álgebra 2^{Ω}, ya que $\mathscr{C} \subset 2^{\Omega}$. La respuesta es que la σ-álgebra 2^{Ω} es más grande de lo que es de interés, y sería más difícil definir probabilidad sobre su totalidad.

Ejemplo 2.11 Sea $\Omega = \{1, 2, 3, 4, 5\}$. Sea $\mathscr{C} = \big\{\{1\}, \{3\}\big\}$. Entonces

$$\sigma(\mathscr{C}) = \big\{\varnothing, \Omega, \{1\}, \{3\}, \{2, 3, 4, 5\}, \{1, 2, 4, 5\}, \{1, 3\}, \{2, 4, 5\}\big\}.$$

Note que esta σ-álgebra tiene más elementos que la trivial, pero menos que 2^{Ω}.

Ejemplo 2.12 Sea $\Omega = \{1, 2, 3, 4, 5\}$, y $\mathscr{C} = \big\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\big\}$. Entonces $\sigma(\mathscr{C}) = 2^{\Omega}$.

Ejemplo 2.13 Sea $\Omega = \mathbb{R}$, y $\mathscr{C} = \big\{(-\infty, x\,] \mid x \in \mathbb{R}\big\}$. Entonces los siguientes conjuntos pertenecen todos a $\sigma(\mathscr{C})$: $(-5, 10)$, $(-5,\ 10]$, $(7, \infty)$, $\{3\}$, $[4,\ 5]$.

Definición 2.9 (borelianos) Sea $\Omega = \mathbb{R}$, y $\mathscr{C} = \big\{(-\infty, x\,] \mid x \in \mathbb{R}\big\}$. La σ-álgebra de subconjuntos de $\mathbb{R}$ definida por $\mathscr{B} = \sigma(\mathscr{C})$ se conoce como el conjunto de *borelianos*.

Existen subconjuntos de $\mathbb{R}$ que *no* son borelianos, es decir, resulta que $\mathscr{B} \subset 2^{\mathbb{R}}$ y la contención es estricta. Sin embargo, los conjuntos borelianos abarcan intervalos abiertos y cerrados (así como sus uniones numerables e intersecciones numerables), del tipo que son típicamente de interés en aplicaciones (ver Ejercicio 2.14). De hecho, puede demostrarse que la σ-álgebra generada por los conjuntos abiertos de $\mathbb{R}$ también coincide con $\mathscr{B}$.

Definición 2.10 (Evento) Los elementos de una σ-álgebra reciben el nombre de *eventos*.

Definición 2.11 (Evento elemental) Un *evento elemental* es un evento formado por un solo elemento.

Ejemplo 2.14 $\Omega = \{1, 2, 3\}$, y $\mathscr{C} = \big\{\varnothing, \Omega, \{1\}, \{2, 3\}\big\}$. El conjunto $\{1\}$ es evento bajo la σ-álgebra $\mathscr{C}$. El conjunto $\{2\}$ no es evento. El conjunto $\{1\}$ es un evento elemental. El evento $\{2, 3\}$ no es elemental.

Note que la definición de evento depende de cuál sea la σ-álgebra bajo consideración. Esto es, un conjunto $A \subset \Omega$ puede ser evento bajo una σ-álgebra $\mathscr{A}$ pero puede no serlo bajo otra σ-álgebra $\mathscr{B}$. La única situación en la cual *cualquier* subconjunto de Ω sería evento, es cuando la σ-álgebra en consideración fuese 2^{Ω}.

Ejemplo 2.15 En el experimento de lanzar una moneda, consideremos $\Omega = \{\mathrm{A}, \mathrm{S}\}$. Los únicos conjuntos susceptibles de ser eventos son Ω, $\varnothing$, $\{\mathrm{A}\}$, y $\{\mathrm{S}\}$, y las únicas σ-álgebras que pueden sustentar estos conjuntos son la trivial y la potencia.

Definición 2.12 (Ocurrir) Sea $\omega \in \Omega$ el resultado de un experimento. Decimos que un evento A *ocurre*, si $\omega \in A$.

Notemos que un evento A^c ocurre si y sólo si A no ocurre, que el evento $A \cap B$ ocurre si y sólo si ocurre A o B, y que el evento $A \cap B$ ocurre si y sólo si A y B ocurren al mismo tiempo.

Los subconjuntos Ω y $\varnothing$ son siempre eventos, sin importar cuál sea la σ-álgebra subyacente. Se llaman, respectivamente, *evento seguro*, y *evento imposible*. Note que Ω siempre ocurre, y que $\varnothing$ nunca ocurre.

2.3. Medida de probabilidad

Definición 2.13 (Eventos ajenos) Dos eventos A y B se dicen *ajenos*, *disjuntos*, o *mutuamente exclusivos*, si $A \cap B = \varnothing$.

Definición 2.14 (Eventos ajenos a pares) Los eventos en una colección $A_1, A_2, \ldots$ se dicen ser *ajenos a pares* o *mutuamente exclusivos a pares*, si se cumple $A_i \cap A_i = \varnothing$, $\forall i \neq j$.

Definición 2.15 (Medida de probabilidad) Sea $\mathscr{A}$ una σ-álgebra de subconjuntos de Ω. Una *medida de probabilidad* es una función $\mathbb{P}\colon \mathscr{A} \to \mathbb{R}$ que

cumple las siguientes condiciones:

(a) $\mathbb{P}(\Omega) = 1$,

(b) $\mathbb{P}(A) \geq 0$, $\forall A \in \mathscr{A}$, y

(c) Si $A_1, A_2, \ldots$ son ajenos a pares, entonces $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Ejemplo 2.16 Sea $\Omega = \{1, 2, 3\}$, y $\mathscr{A} = \big\{\varnothing, \Omega, \{1\}, \{2, 3\}\big\}$. Ahora definamos la función $\mathbb{P}\colon \mathscr{A} \to \mathbb{R}$ como en la siguiente tabla:

A	$\mathbb{P}(A)$
$\varnothing$	0
Ω	1
$\{1\}$	1/3
$\{2, 3\}$	2/3

La función $\mathbb{P}$ es una medida de probabilidad.

Lo anterior corresponde a una especificación conocida como *definición axiomática* de probabilidad. Una probabilidad puede especificarse como *cualquier* función que cumpla la definición anterior, de manera abstracta e irrespectivamente de algún contexto específico.

También existe una definición, conocida como *definición clásica*, o *frecuentista*, de probabilidad que es la siguiente: Supongamos que un experimento aleatorio específico puede repetirse un número indefinido de veces, digamos N. Sea A un evento, y definamos

$$r_N(A) = \frac{\#\text{ veces que A ocurre}}{N}.$$

Entonces la *probabilidad* de A se define por $\mathbb{P}(A) = \lim_{N\to\infty} r_N(A)$. Más adelante veremos que hay una conexión entre la probabilidad definida axiomáticamente con esta probabilidad clásica, a través de un resultado llamado la Ley de los Grandes Números. Por el momento basta notar que sea lo que sea el número $\mathbb{P}\,(A)$, éste tiene la interpretación de que si N es grande, entonces $\mathbb{P}(A)$ es aproximadamente igual a $r_N(A)$.

Existen algunas propiedades elementales que pueden deducirse de cualquier medida de probabilidad, incluyendo que el requerimiento de una medida de probabilidad en cuanto a uniones numerables de eventos disjuntos a pares, permite transferir a uniones finitas.

Proposición 2.5 *Si $\mathbb{P}$ es una medida de probabilidad sobre $\mathscr{A}$ entonces $\mathbb{P}(\varnothing) = 0$, y si $A_1, A_2, \ldots, A_n$ es una colección finita de eventos de $\mathscr{A}$ disjuntos a pares, entonces $\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i)$.*

Demostración. Notar que $\Omega = \Omega \cup \varnothing \cup \varnothing \cup \cdots$. El lado derecho es una unión numerable de eventos en $\mathscr{A}$, y por ser $\mathbb{P}$ una medida de probabilidad, se cumple

$$\mathbb{P}\big(\Omega \cup \varnothing \cup \varnothing \cup \cdots\big) = \mathbb{P}(\Omega) + \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing) + \cdots .$$

Como $\mathbb{P}(\Omega) = 1$, y $\mathbb{P}(\varnothing) \geq 0$, necesariamente $\mathbb{P}(\varnothing) = 0$. Ahora, notar que $\cup_{i=1}^n A_i = A_1 \cup A_2 \cdots \cup A_n \cup \varnothing \cup \varnothing \cup \cdots$, y que esta unión numerable consiste de eventos disjuntos a pares. Por ser $\mathbb{P}$ medida de probabilidad,

$$\mathbb{P}\big(\cup_{i=1}^n A_i \cup \varnothing \cup \varnothing \cup \cdots\big) = \sum_{i=1}^{n} \mathbb{P}(A_i) + 0 + 0 + \cdots = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

□

2.4. Espacio de probabilidad

Definición 2.16 (Espacio de probabilidad) Llamamos un *espacio de probabilidad*, o *modelo de probabilidad*, a una terna $(\Omega, \mathscr{A}, \mathbb{P})$, donde Ω es un conjunto no vacío, $\mathscr{A}$ es una σ-álgebra de subconjuntos de Ω, y $\mathbb{P}$ es una medida de probabilidad sobre $\mathscr{A}$.

Con esta definición, no hemos especificado algo inexistente. Los modelos de probabilidad son los objetos que estudia la teoría de probabilidad. Por lo pronto, el Ejemplo 2.16 demuestra que existe por lo menos uno.

Ejemplo 2.17 Para un espacio muestral Ω, tómese $\mathscr{A} = \{\Omega, \varnothing\}$, y defina $\mathbb{P}(\Omega) = 1$, $\mathbb{P}(\varnothing) = 0$. Se verifica que $\mathscr{A}$ es σ-álgebra y que $\mathbb{P}$ es una medida de probabilidad sobre $\mathscr{A}$. Este $(\Omega, \mathscr{A}, \mathbb{P})$ se llama el espacio de probabilidad trivial.

Ejemplo 2.18 Sea $\Omega = \{1, 2, 3, 4, 5\}$, y $\mathscr{A} = 2^{\Omega}$. Defina la función $\mathbb{P}(A) = \#(A)/5$, $\forall A \in \mathscr{A}$. Defina la función

$$\mathbb{Q}(A) = \begin{cases} 1 & \text{si } 1 \in A, \\ 0 & \text{si } 1 \notin A. \end{cases}$$

Entonces $(\Omega, \mathscr{A}, \mathbb{P})$ y $(\Omega, \mathscr{A}, \mathbb{Q})$ son ambos espacios de probabilidad. Son espacios de probabilidad distintos porque las medidas $\mathbb{P}$ y $\mathbb{Q}$ son distintas.

Ejemplo 2.19 Sea $\Omega = \{1, 2, 3, \dots\}$, y $\mathscr{A} = 2^{\Omega}$. Ahora sea $B \subset \Omega$ un conjunto finito, con n elementos. Defina la función $\mathbb{P}(A) = \#(B \cap A)/n$. Entonces $(\Omega, \mathscr{A}, \mathbb{P})$ es un espacio de probabilidad.

Ejemplo 2.20 Considere el experimento de observar el número de clientes que ingresan a un comercio en un intervalo fijo de tiempo, y adopte $\Omega = \{0, 1, 2, 3, \dots\}$ y $\mathscr{A} = 2^{\Omega}$ para describirlo. Si para un subconjunto $A \in \mathscr{A}$ definimos $\mathbb{P}(A) = \sum_{i \in A} \exp(-1)/i!$, entonces $(\Omega, \mathscr{A}, \mathbb{P})$ es un espacio de probabilidad.

Finalizamos esta sección ilustrando con un ejemplo que los espacios de probabilidad que se obtienen de considerar espacios muestrales infinitos, introducen detalles finos de complejidad.

Ejemplo 2.21 Considere el experimento de observar al azar una función continua sobre el intervalo $[\,0, 1\,]$. Esta situación surge en problemas de recepción de señales analógicas. El espacio muestral Ω es entonces un conjunto de *funciones*, y los elementos de $\mathscr{A}$ serían entonces conjuntos de *funciones*. Este ejemplo es similar al anterior en el hecho de que Ω es infinito, pero su cardinalidad es la del continuo. A diferencia de dicho ejemplo anterior, en este caso no es fácil definir la medida de probabilidad $\mathbb{P}$, y menos aún sobre toda la clase 2^{Ω}.

Ejercicios

2.1 Encontrar operaciones de teoría de conjuntos entre los eventos A, B y C que denoten que:

(a) solamente ocurre A.

(b) los tres eventos ocurren.

(c) por lo menos dos ocurren.

(d) ocurren dos y no más.

(e) no ocurren más de dos.

(f) ocurren tanto A como B, pero no C.

(g) ocurre por lo menos uno.

(h) ocurre uno, y no más.

(i) no ocurre ninguno.

2.2 Demuestre que 2^{Ω} es una σ-álgebra de subconjuntos de Ω.

2.3 Demuestre la Proposición 2.1.

2.4 Demuestre la Proposición 2.2.

2.5 Demuestre que si Ω tiene n elementos, entonces 2^{Ω} tiene 2^n elementos.

2.6 Verifique que la llamada σ-álgebra trivial, en efecto, es una σ-álgebra.

2.7 Si $\mathscr{A}$ y $\mathscr{B}$ son σ-álgebra de subconjuntos de Ω, demuestre que $\mathscr{A} \cap \mathscr{B}$ también es σ-álgebra de subconjuntos de Ω.

2.8 Sea Λ un conjunto arbitrario de índices. Si $\forall \lambda \in \Lambda$, $\mathscr{A}_\lambda$ es σ-álgebra de subconjuntos de Ω, entonces $\cap_{\lambda \in \Lambda} \mathscr{A}_\lambda$ también es σ-álgebra de subconjuntos de Ω.

2.9 (σ-álgebra inducida por un evento)

(a) Sea $\mathscr{A}$ una σ-álgebra de subconjuntos de Ω. Para un evento arbitrario $A \in \mathscr{A}$ defina la clase siguiente de subconjuntos de Ω :

$$\mathscr{A}_A = \{A \cap B \mid B \in \mathscr{A}\}$$

(es decir, la clase formada por todas las intersecciones de A con los elementos de $\mathscr{A}$). Demuestre que $\mathscr{A}_A$ es σ-álgebra de subconjuntos de A. Note que la propiedad de cerradura para complementación para $\mathscr{A}_A$ debe probarse para complementación respecto a un nuevo espacio muestral A (en lugar de Ω).

(b) Considere $\Omega = \{1, 2, 3, 4, 5, 6, 7\}$, $\mathscr{A} = \big\{\varnothing, \Omega, \{1\}, \{2\}, \{1, 2\}, \{2, 3, 4, 5, 6, 7\}, \{1, 3, 4, 5, 6, 7\}, \{3, 4, 5, 6, 7\}\big\}$, y $A = \{1, 2\}$. Demuestre que $\mathscr{A}$ es σ-álgebra de subconjuntos de Ω y encuentre $\mathscr{A}_A$.

2.10 Demuestre que toda σ-álgebra $\mathscr{A}$ contiene al conjunto vacío, $\varnothing$, y que para toda medida de probabilidad definida sobre $\mathscr{A}$, se cumple $\mathbb{P}(\varnothing) = 0$.

2.11 ¿Puede haber un modelo de probabilidad en el cual el espacio muestral consiste de un solo elemento?

2.12 ¿Puede haber una σ-álgebra que consista de un solo elemento?

2.13 Demuestra que si $\mathscr{A} = \{\Omega, \varnothing\}$, entonces existe sólo una medida de probabilidad posible sobre $\mathscr{A}$.

2.14 Es claro que cualquier intervalo de la forma $(-\infty, x\,]$ pertenece a la σ-álgebra de borelianos, $\mathscr{B}$. También es claro que $\mathbb{R} \in \mathscr{B}$. Demuestre que cada uno de los siguientes tipos de conjuntos también pertenecen a $\mathscr{B}$. En lo que sigue, a, b son números reales con $a < b$. Sugerencia: Exprese cada tipo de conjunto en términos de complementos, uniones numerables, intersecciones numerables de conjuntos que son de la forma $(-\infty, x\,]$ para $x \in \mathbb{R}$ o de otros conjuntos que se sepa son elementos de $\mathscr{B}$.

(a) (a, ∞), un intervalo infinito abierto por la izquierda.

(b) $\{a\}$, un conjunto formado por un solo número real.

(c) $(-\infty, a)$, un intervalo infinito abierto por la derecha.

(d) (a, b), un intervalo finito abierto en los dos extremos.

(e) $(a, b\,]$, un intervalo finito abierto por la izquierda y cerrado por la derecha.

(f) $[\,a, b)$, un intervalo finito cerrado por la izquierda y abierto por la derecha.

(g) $[\,a, b\,]$, un intervalo finito cerrado por los dos extremos.

Capítulo 3

Espacios muestrales finitos y numerables

En este capítulo estudiaremos espacios de probabilidad para un caso muy particular. Se trata de aquel en que el espacio muestral, Ω, es finito o numerable. De esta forma, Ω es de la forma $\{\omega_1, \omega_2, \ldots, \omega_n\}$ o de la forma $\{\omega_1, \omega_2, \ldots\}$. No obstante una patente simplicidad en esta suposición, esta situación abarca una grandísima cantidad de aplicaciones no-elementales de teoría de probabilidad. Incluye juegos de azar del casino, y fenómenos aleatorios en los cuales el resultado consiste de contar números de individuos o incidencias que ocurren al azar. Ejemplos de esto último son el número de individuos que hay en una población de animales después de la temporada de reproducción, el número de pacientes que se curan tras un nuevo tratamiento, o el número de partículas atómicas emitidas por una fuente radiactiva.

Una vez identificado el espacio muestral finito o numerable congruente con el fenómeno bajo estudio, el siguiente paso consiste en especificar la

medida de probabilidad, $\mathbb{P}$. Veremos que esto se puede lograr asignando al i-ésimo punto del espacio muestral un número real p_i, identificado con la probabilidad asociada al resultado i. El conjunto de números reales $\{p_i\}$ debe cumplir las siguientes propiedades:

(i) $p_i \geq 0, \forall\, i.$

(ii) $p_1 + p_2 + \cdots = 1.$

Definición 3.1 (Función indicadora) Sea $A \subset \Omega$ un conjunto arbitrario. Denotamos por $1_A(\cdot)$ a la función $\Omega \to \mathbb{R}$ definida por

$$1_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A, \\ 0 & \text{si } \omega \notin A. \end{cases}$$

Esta función recibe el nombre de *función indicadora del conjunto* A.

3.1. Espacios finitos

El siguiente resultado provee un método común para definir funciones o medidas de probabilidad.

Teorema 3.1 (Caracterización de probabilidad: Espacio finito) *Consideremos $\Omega = \{\omega_1, \ldots, \omega_n\}$ un espacio muestral con un número finito de elementos. Sean $p_1, \ldots, p_n$ números reales tales que $p_i \geq 0,\ i = 1, \ldots, n$ y tales que $\sum_{i=1}^n p_i = 1$. Entonces,*

(a) la función $\mathbb{P} : 2^\Omega \to \mathbb{R}$ definida por $\mathbb{P}(A) = \sum_{i=1}^n 1_A(\omega_i)\, p_i$ es una medida de probabilidad sobre 2^Ω,

(b) $\mathbb{P}\big(\{\omega_i\}\big) = p_i,\ i = 1, \ldots, n$, *y*

(c) si $\mathbb{Q}$ *es una medida de probabilidad tal que* $\mathbb{Q}\big(\{\omega_i\}\big) = p_i,\ i = 1, \ldots, n$, *entonces* $\mathbb{Q}(A) = \mathbb{P}(A),\ \forall A \in 2^{\Omega}$.

Demostración.

De (a) Por definición

$$\mathbb{P}(\Omega) = \sum_{i=1}^{n} 1_{\Omega}(\omega_i)\, p_i = \sum_{i=1}^{n} p_i = 1.$$

El hecho $\mathbb{P}(A) \geq 0$ se sigue de que $p_i \geq 0,\ i = 1, \ldots, n$. Si $A_1,\ A_2, \ldots$ es una sucesión de conjuntos disjuntos a pares, notemos que por ser Ω finito, existe un índice M tal que $A_n = \varnothing,\ \forall n > M$, de tal forma que $\cup_{j=1}^{\infty} A_j = \cup_{j=1}^{M} A_j$. Entonces

$$\mathbb{P}\left(\cup_{j=1}^{\infty} A_j\right) = \mathbb{P}\left(\cup_{j=1}^{M} A_j\right) = \sum_{i=1}^{n} 1_{\cup_{j=1}^{M} A_j}(\omega_i)\, p_i =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{M} 1_{A_j}(\omega_i)\, p_i = \sum_{j=1}^{M}\sum_{i=1}^{n} 1_{A_j}(\omega_i)\, p_i = \sum_{j=1}^{M} \mathbb{P}(A_j) = \sum_{j=1}^{\infty} \mathbb{P}(A_j).$$

Por los resultados anteriores, $\mathbb{P}$ es medida de probabilidad sobre 2^{Ω}.

De (b) $\mathbb{P}(\{\omega_i\}) = \sum_{j=1}^{n} 1_{\{\omega_i\}}(\omega_j)\, p_j = p_i$, ya que por definición de función indicadora $1_{\{\omega_i\}}(\omega_j) = 1$ si $i = j$ y cero en otros casos.

De (c) Para $A \in 2^{\Omega}$, $\mathbb{Q}(A) = \mathbb{Q}\left(\cup_{\omega_i \in A}\{\omega_i\}\right) = \sum_{\omega_i \in A} \mathbb{Q}\left(\{\omega_i\}\right) = \sum_{\omega_i \in A} p_i$ $= \mathbb{P}(A)$.

□

La relevancia de este resultado es la siguiente. Primero, que con la mera especificación de los valores p_i, se define una medida de probabilidad $\mathbb{P}$ sobre *todos* los conjuntos de 2^Ω. Esto constituye un ejemplo de especificación completa de una medida de probabilidad por medio de relativamente pocos ingredientes (en este caso, los números $p_1, \ldots, p_n$). En segundo lugar, que la única medida de probabilidad sobre la σ-álgebra potencia 2^Ω que cumple asignar los valores p_i a los eventos elementales $\{\omega_i\}$ es precisamente, esta medida $\mathbb{P}$ (porque quedó demostrado que no hay otra). Esto significa que los valores $p_1, \ldots, p_n$, además de *especificar* una medida de probabilidad $\mathbb{P}$ (parte (a) del teorema), la *caracteriza* (parte (c) del teorema), en el sentido de que no hay dos medidas de probabilidad diferentes sobre 2^Ω que asignen las mismas probabilidades a los eventos $\{\omega_i\}$.

Notemos que en el caso de ser Ω finita, entonces

$$2^\Omega = \sigma\Big(\big\{\{\omega_1\}, \ldots, \{\omega_n\}\big\}\Big)$$

(ejercicio), es decir, que los eventos $\{\omega_i\}$ generan 2^Ω. El teorema ha mostrado que en este caso particular, basta especificar los valores de una medida de probabilidad $\mathbb{P}$ solamente sobre la clase generadora para definir sin ambigüedad una medida de probabilidad sobre la totalidad de la clase generada. ¿Es esto cierto en lo general? Es decir, si una clase $\mathscr{G}$ genera a una σ-álgebra $\mathscr{A}$, y dos medidas de probabilidad $\mathbb{P}$ y $\mathbb{Q}$ cumplen $\mathbb{P}(G) = \mathbb{Q}(G)$, $\forall G \in \mathscr{G}$, será cierto que $\mathbb{P}(A) = \mathbb{Q}(A)$, $\forall A \in \mathscr{A}$? La respuesta en general, es que no (Ejercicio 3.3). Sin embargo, si la clase $\mathscr{G}$ cumple ciertos requerimientos, la respuesta es que sí (como se vería en un curso de teoría de medida).

Ejemplo 3.1 (Densidad binomial) Si $\Omega = \{0, \ldots, n\}$, una forma de definir

la medida de probabilidad es especificando, para un valor fijo de $p \in [0, 1]$,

$$p_x = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

Esto se llama la distribución *binomial*. Note que se especifica por completo una medida de probabilidad mediante el uso de una fórmula cerrada, que proporciona juegos de valores distintos de probabilidad dependiendo del valor de una sola constante, p. Esto es un ejemplo de un modelo de probabilidad *parametrizado* por el valor de p. Después veremos que este modelo de probabilidad surge en ciertas situaciones en las que se cuentan objetos que cumplen una de dos posibles características de entre n objetos seleccionados al azar. Por ejemplo, el número de artículos defectuosos presentes en un lote de n artículos fabricados, o el número de páginas con errores tipográficos en una obra de n páginas.

3.1.1. Espacios uniformes

Definición 3.2 (Modelo de probabilidad uniforme) Decimos que un modelo de probabilidad con espacio muestral finito $\Omega = \{\omega_1, \ldots, \omega_n\}$ es *uniforme*, si la medida de probabilidad cumple $\mathbb{P}\big(\{\omega_1\}\big) = \mathbb{P}\big(\{\omega_2\}\big) = \cdots = \mathbb{P}\big(\{\omega_n\}\big)$.

Definición 3.3 (Cardinalidad) Sea $B \in 2^\Omega$ un conjunto arbitrario. Denotamos por $\#(B)$ al número de elementos o eventos elementales en el conjunto B, y lo llamamos la *cardinalidad* del conjunto B.

Teorema 3.2 (Cálculo de probabilidades: Modelo uniforme) *En un modelo*

de probabilidad uniforme, $\mathbb{P}\big(\{\omega_i\}\big) = 1/n$*, y*

$$\mathbb{P}(A) = \frac{\#(A)}{\#(\Omega)},$$

para todo $A \in 2^{\Omega}$.

Demostración. Por ser $\mathbb{P}$ una medida de probabilidad, debe cumplir que $1 = \mathbb{P}(\Omega) = \sum_{i=1}^{n} \mathbb{P}\big(\{\omega_i\}\big)$, y por ser este uniforme $\mathbb{P}\big(\{\omega_1\}\big) = \cdots = \mathbb{P}\big(\{\omega_n\}\big)$ de donde se concluye que $\mathbb{P}\big(\{\omega_i\}\big) = 1/n$. Con ello,

$$\mathbb{P}(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_A(\omega_i) = \frac{\#(A)}{n},$$

donde $n = \#(\Omega)$. □

Este resultado es el que se invoca, cuando en el lenguaje coloquial oímos decir «hay 1 oportunidad en 10, 000 de que suceda». El modelo de probabilidad uniforme, a pesar de su simplicidad, tiene muchas aplicaciones, siendo quizás las más populares aquellas que involucran los juegos de azar. Por ejemplo, la ruleta de casino (balanceada) da lugar a un modelo de probabilidad uniforme con 38 resultados posibles, o un lanzamiento de un dado (sin cargar) da lugar a un modelo uniforme con 6 resultados posibles. El lanzamiento de una moneda balanceada da lugar a dos resultados equiprobables, razón por la cual uno, intuitivamente, responde sin titubear que la probabilidad de que al lanzar una moneda caiga águila es 1/2, quizás haciendo uso inconsciente del teorema anterior. Equivalentemente, la definición de un espacio muestral uniforme puede formularse en términos de una negación, es decir, un espacio muestral es uniforme si ninguna de

las probabilidades de los eventos elementales $\{\omega_i\}$ es diferente a las demás. Note que la uniformidad o no uniformidad de una situación dada, es una propiedad que asumimos como modeladores a fin de obtener alguna respuesta. Es decir, si la moneda que se va a lanzar ha sido deformada a golpes, el modelo uniforme no parecería ser adecuado, por lo que no sería sensato adoptarlo para esa situación.

Irónicamente, el resultado anterior es tan intuitivo, que algunas veces se toma equivocadamente como la definición de probabilidad. En efecto, no son pocos los cursos introductorios de probabilidad en los cuales se dedica gran cantidad de tiempo y atención a las llamadas técnicas de conteo, que no son más que formas de calcular $\#(A)$ y $\#(\Omega)$, y el énfasis parecería implicar que el concepto de probabilidad consiste siempre en calcular el cociente de ambas cantidades. Coloquialmente, el resultado del teorema se describe como «número de casos favorables entre número de casos posibles». Es importante señalar que los modelos uniformes no son más que un caso muy particular de los modelos de probabilidad en general que han sido estudiados en el capítulo anterior.

3.2. Espacios numerables

Podemos generalizar el caso expuesto para Ω finito, al caso en que éste sea numerable.

Teorema 3.3 (Caracterización de probabilidad: Espacio numerable) *Sea $\Omega = \{\omega_1, \omega_2 \dots\}$ un espacio muestral numerable. Sean $p_1,\ p_2,\ \dots$ una secuencia de números reales tales que $p_i \geq 0,\ i = 1,\ 2,\ \dots$ y $\sum_{i=1}^{\infty} p_i = 1$. Entonces,*

(a) la función $\mathbb{P}\colon 2^{\Omega} \to \mathbb{R}$ *definida por* $\mathbb{P}(A) = \sum_{i=1}^{\infty} 1_A(\omega_i)\, p_i$ *es una medida de probabilidad sobre* 2^{Ω},

(b) $\mathbb{P}\big(\{\omega_i\}\big) = p_i$, $i = 1, 2, \ldots$ *y*

(c) si $\mathbb{Q}$ *es una medida de probabilidad tal que* $\mathbb{Q}(\{\omega_i\}) = p_i,\ i = 1, 2, \ldots$ *entonces* $\mathbb{Q}(A) = \mathbb{P}(A)$, $\forall A \in 2^{\Omega}$.

Demostración.

De (a) $\mathbb{P}(\Omega) = \sum_{i=1}^{\infty} 1_{\Omega}(\omega_i)\, p_i = \sum_{i=1}^{n} p_i = 1$. El hecho $\mathbb{P}(A) \geq 0$ se sigue de $p_i \geq 0$. Si $A_1, A_2, \ldots$ es una sucesión de conjuntos disjuntos a pares, entonces

$$\mathbb{P}\left(\cup_{j=1}^{\infty} A_j\right) = \sum_{i=1}^{\infty} 1_{\cup_{j=1}^{\infty} A_j}(\omega_i) p_i = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} 1_{A_j}(\omega_i) p_i = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 1_{A_j}(\omega_i) p_i = \sum_{j=1}^{\infty} \mathbb{P}(A_j).$$

De (b) $\mathbb{P}\big(\{\omega_i\}\big) = \sum_{j=1}^{\infty} 1_{\{\omega_i\}}(\omega_j)\, p_j = p_i$.

De (c) Para $A \in 2^{\Omega}$, el cual es un conjunto numerable,

$$\mathbb{Q}(A) = \mathbb{Q}\big(\cup_{\omega_i \in A}\{\omega_i\}\big) = \sum_{\omega_i \in A} \mathbb{Q}\big(\{\omega_i\}\big) = \sum_{\omega_i \in A} p_i = \sum_{i=1}^{\infty} 1_A(\omega_i) p_i = \mathbb{P}(A).$$

□

Observemos que para la demostración de este resultado, se invocan

propiedades de series. Por ejemplo, si los términos de una serie cumplen $|a_i| \leq |b_i|$ y la serie $\sum_{i=1}^{\infty} |b_i|$ es convergente, entonces la serie $\sum_{i=1}^{\infty} |a_i|$ también lo es. Estamos usando esto, al notar que $|1_A(\omega_i)p_i| \leq |p_i|$. Si los términos de una serie cumplen $a_i \geq c$, entonces $\sum_{i=1}^{\infty} a_i \geq c$. En un paso de la demostración de la parte (a) también se está invocando un resultado acerca de series dobles, a saber, que $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}$, lo cual en particular se cumple si las series son de términos no-negativos.

Al igual que en el caso finito, la relevancia de este resultado es que la sucesión de números $p_i \geq 0,\ i = 1,\ 2,\ \ldots$ y tales que $\sum_{i=1}^{\infty} p_i = 1$, caracterizan una medida de probabilidad sobre 2^{Ω}.

Ejemplo 3.2 (Densidad geométrica) Suponga que $\Omega = \{0, 1, 2, \ldots\}$. Para una constante fija $p \in (0, 1]$, defina

$$p_i = p(1-p)^i,\ i = 0,\ 1,\ \ldots .$$

Esta secuencia cumple que $\sum_{i=0}^{\infty} p_i = 1$, y se llama la distribución *geométrica* con parámetro p. Más adelante veremos que esta asignación de probabilidades es útil en problemas que tienen que ver con realizar experimentos en los que pueden ocurrir dos resultados —llamados «éxito» y «fracaso» irrespectivamente de su connotación— y contar el número de «fracasos» que ocurren antes del primer «éxito». Por ejemplo, contar el número de veces que un sistema de cómputo funcione, antes de que ocurra la primera falla (si falla a la primera, entonces el número de veces que sí funcionó es 0).

Ejemplo 3.3 (Densidad Poisson) Suponga que $\Omega = \{0, 1, 2, \ldots\}$. Para una

constante fija $\lambda > 0$, defina

$$p_i = e^{-\lambda}\frac{\lambda^i}{i!},\ i = 0,\ 1,\ \ldots\ .$$

Esta secuencia cumple $\sum_{i=0}^{\infty} p_i = 1$, y se llama la distribución de *Poisson* con parámetro λ. Más adelante veremos que esta asignación de probabilidades surge en problemas que tienen que ver con el conteo del número de incidentes que ocurren en intervalos de tiempo, de longitud, de superficie, o de volumen. Por ejemplo, el número de clientes que arriban a un banco en un intervalo de 10 minutos de duración, el número de grietas en una varilla de acero de longitud L, el número de tormentas que ocurren al azar en cierta zona de superficie 100 km^2, o el número de cardúmenes que ocurren en un volumen de mar.

En este curso introductorio no se abordarán espacios muestrales más generales, por requerirse de herramientas más avanzadas de teoría de medida. En el Capítulo 5 sí se verán algunas maneras de especificar medidas de probabilidad sobre cierta σ-álgebra sobre los números reales (la de Borel, es decir, la σ-álgebra generada por la clase $\{(-\infty,\, x\,] \mid -\infty < x < \infty\}$, a través de instrumentos llamados distribuciones de probabilidad para variables aleatorias. Una extensión posible, para cursos más avanzados es definir medidas de probabilidad sobre $\mathbb{R}^n$, y otra es definirlas sobre espacios muestrales que son *funciones*. Por ejemplo, la observación de una señal de radio sobre un intervalo de tiempo puede pensarse como la observación de una función aleatoria del tiempo, especialmente cuando a la señal original ha sido contaminada por ruido aleatorio debido al medio ambiente y a variaciones eléctricas. En teoría de probabilidad, cuando el objeto aleatorio es

una función, se habla de una gran disciplina llamada *procesos estocásticos*.

Ejercicios

3.1 Verifique las siguientes propiedades de la función indicadora:

(a) $1_{A^c}(\omega) = 1 - 1_A(\omega)$,

(b) Si $A_1,\ A_2,\ \ldots,\ A_n \subset \Omega$, entonces $1_{\cap_{i=1}^n A_i}(\omega) = \prod_{i=1}^n 1_{A_i}(\omega)$, y

(c) Si $A_1,\ A_2,\ \ldots$ es una sucesión de conjuntos disjuntos a pares, entonces $1_{\cup_{i=1}^\infty A_i}(\omega) = \sum_{i=1}^\infty 1_{A_i}(\omega)$.

3.2 Para cada uno de los siguientes experimentos, describa el espacio muestral:

(a) Contar el número de insectos que se encuentran en una planta.

(b) Lanzamiento de una moneda cuatro veces.

(c) Medir el tiempo de vida, en horas, de un componente electrónico.

(d) Registro del peso de una rata a 10 días de haber nacido.

3.3 Muestre que puede haber dos medidas de probabilidad diferentes $\mathbb{P}$ y $\mathbb{Q}$ definidas sobre una σ-álgebra $\mathscr{A}$, que coinciden sobre una clase $\mathscr{G}$ tal que $\mathscr{A} = \sigma(\mathscr{G})$. Sugerencia: Busque un contraejemplo basado en un Ω que tenga un número chico de elementos, digamos 3 o 4.

3.4 Demuestre que si Ω es finita, entonces $2^\Omega = \sigma\Big(\big\{\{\omega_1\}, \ldots, \{\omega_n\}\big\}\Big)$.

3.5 Piense en varios ejemplos reales de experimentos con Ω finito, que sean uniformes y otros que no sean uniformes.

3.6 Muestre que:

(a) Las probabilidades binomiales dadas por $p_x = \binom{n}{x} p^x (1-p)^{n-x}$, en efecto cumplen que $\sum_{x=0}^{n} p_x = 1$, para cualquier $p \in [\,0, 1\,]$.

(b) Las probabilidades geométricas dadas por $p_i = p(1-p)^i$ cumplen que $\sum_{i=0}^{\infty} p_i = 1$, para cualquier $p \in [\,0, 1\,]$.

(c) Las probabilidades Poisson dadas por $p_i = e^{-\lambda}\lambda^i/i!$ también cumplen que $\sum_{i=0}^{\infty} p_i = 1$, para cualquier $\lambda > 0$.

3.7 Considere un experimento en el cual un foco es observado hasta que falle, de forma tal que se reporta el número de horas completas de vida. Asuma que el experimento no puede continuarse indefinidamente. Si se termina a las n horas, ¿Cuál es el espacio muestral?

3.8 Dos equipos juegan el «mejor de siete series». El juego se detiene inmediatamente cuando un equipo ha ganado cuatro juegos de los siete. Conteste las siguientes preguntas.

(a) Describa el espacio muestral para este experimento.

(b) Si los equipos tienen las mismas oportunidades de ganar, ¿Qué probabilidades se pueden asignar a cada punto del espacio muestral?

(c) ¿Cuál es la probabilidad de que exactamente se necesiten 7 juegos para que un equipo sea el ganador?

3.9 ¿Existe una noción de un modelo de probabilidad uniforme cuando el espacio muestral no es finito sino numerable?

3.10 Suponga que la probabilidad de que una persona sea zurda es p. Si se eligen tres personas al azar de manera independiente:

(a) ¿Cuál es la probabilidad de que las tres resulten zurdas?

(b) ¿Cuál es la probabilidad de que las tres resulten derechas?

(c) ¿Cuál es la probabilidad de que por lo menos dos resulten zurdas?

Capítulo 4

Propiedades de probabilidad

En el capítulo anterior, se definió el concepto de un modelo de probabilidad $(\Omega, \mathscr{A}, \mathbb{P})$, donde el espacio muestral Ω representa los posibles resultados de un fenómeno aleatorio, y la σ-álgebra $\mathscr{A}$ es un sistema de eventos que contiene a los eventos que puedan ser de interés en una situación determinada. La utilidad de un modelo de probabilidad es poder cuantificar la probabilidad para eventos que sean relevantes en una situación práctica. Si pudieran enumerarse exhaustivamente los valores de la medida de probabilidad $\mathbb{P}(A)$ para *todos* los eventos que pertenecen a $\mathscr{A}$, entonces estaríamos capacitados para contestar cualquier pregunta en términos de probabilidades de todos los eventos en $\mathscr{A}$. Como vimos, $\mathscr{A}$ puede contener una gran cantidad de eventos, y el modelo de probabilidad se establece en ocasiones mediante la especificación de unas pocas probabilidades (por ejemplo, como lo pudimos hacer en espacios finitos y numerables). Ahora veremos maneras de deducir la medida de probabilidad $\mathbb{P}$ para más y más eventos en $\mathscr{A}$, para casos en los que se conozcan valores de la medida de probabilidad sobre otro juego de eventos. En este capítulo se estudiarán propiedades generales

que tiene cualquier espacio de probabilidad. Constituyen, de hecho, un conjunto de resultados que permiten el *empleo* de un modelo de probabilidad para la obtención de respuestas. Se entiende por una *ley* de probabilidad, algún resultado que establezca cuál es la probabilidad de un evento compuesto por otros mediante operaciones de teoría de conjuntos, en términos de otras probabilidades relacionadas con los eventos componentes.

En todo lo que sigue, se supone arbitrario, pero fijo, un espacio de probabilidad $(\Omega, \mathscr{A}, \mathbb{P})$, y siempre que se hable de experimentos, eventos y de probabilidad, será con respecto a este espacio. Es importante notar que la totalidad de las propiedades aquí mencionadas, se derivan de las tres propiedades que definen una medida de probabilidad *general*. Es decir, los resultados de este capítulo no dependen de que Ω sea finito o que sea numerable.

4.1. Leyes elementales

Teorema 4.1 (Ley del complemento) *Si $\mathbb{P}$ es una medida de probabilidad se tiene que para todo evento A, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*

Demostración. Por ser $\mathbb{P}$ medida de probabilidad y por ser A y A^c dos eventos disjuntos,

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c),$$

lo cual implica que $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. □

Teorema 4.2 (Ley aditiva) *Si $\mathbb{P}$ es medida de probabilidad, para dos eventos cualesquiera A y B, se cumple*

(a) $\mathbb{P}\left(B \cap A^c\right) = \mathbb{P}\left(B\right) - \mathbb{P}\left(A \cap B\right).$

(b) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ *(Ley aditiva).*

(c) Si $A \subset B$ *entonces* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Demostración.

De (a) Para cualesquiera conjuntos A y B se tiene que $B = \{B \cap A\} \cup \{B \cap A^c\}$, unión de eventos disjuntos. Entonces

$$\mathbb{P}\left(B\right) = \mathbb{P}\left(B \cap A\right) + \mathbb{P}\left(B \cap A^c\right),$$

y de aquí que $\mathbb{P}\left(B \cap A^c\right) = \mathbb{P}\left(B\right) - \mathbb{P}\left(B \cap A\right).$

De (b) Note que $A \cup B = A \cup (B \cap A^c)$ es una unión disjunta, y entonces

$$\mathbb{P}\left(A \cup B\right) = \mathbb{P}\big(A \cup \left(B \cap A^c\right)\big) = \mathbb{P}\left(A\right) + \mathbb{P}\left(B \cap A^c\right) = \mathbb{P}\left(A\right) + \mathbb{P}\left(B\right) - \mathbb{P}\left(B \cap A\right).$$

De (c) Si $A \subset B$ se tiene que $A \cap B = A$, y $\mathbb{P}\left(A \cap B\right) = \mathbb{P}\left(A\right)$, y de aquí que

$$0 \leq \mathbb{P}(B \cap A^c) = \mathbb{P}\left(B\right) - \mathbb{P}\left(B \cap A\right) = \mathbb{P}\left(B\right) - \mathbb{P}\left(A\right),$$

de donde $\mathbb{P}\left(A\right) \leq \mathbb{P}\left(B\right).$

□

Para dos eventos A y B que sean ajenos, se cumple $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. Esto puede obtenerse como corolario del teorema anterior, o bien directamente por propiedades de la medida de probabilidad $\mathbb{P}$.

Teorema 4.3 (Ley aditiva, general) *Para una colección finita de eventos, $A_1, \ldots, A_n$ se cumple una generalización de la ley aditiva, es decir,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}\left(A_i\right) - \sum\sum_{i<j} \mathbb{P}\left(A_i \cap A_j\right) + \sum\sum\sum_{i<j<k} \mathbb{P}\left(A_i \cap A_j \cap A_k\right) - \cdots + (-1)^{n+1} \mathbb{P}\left(A_1 \cap A_2 \cap \cdots \cap A_n\right).$$

Ejemplo 4.1 Volaremos todos abordo de un avión bimotor. Cada motor tiene probabilidad 0.01 de fallar.

(a) Suponga que el avión es capaz de volar con un solo motor. Encuentre la probabilidad de un vuelo seguro.

(b) Suponga que el avión requiere de ambos motores para volar. Encuentre la probabilidad de un vuelo seguro.

Con lo visto hasta este punto en este capítulo, ¿hay información suficiente para responder a estas preguntas, o hace falta algo más?

4.2. Probabilidad condicional

A manera de motivar el concepto de probabilidad condicional, consideremos el modelo de probabilidad uniforme que se deriva de la selección de una carta en una baraja estándar de 52 cartas. Supongamos que el evento de interés es el siguiente:

$$A = \text{«seleccionar un corazón»}.$$

Entonces se calcula $\mathbb{P}(A) = 13/52 = 1/4$.

Ahora supongamos que alguien toma una carta al azar, la mira, y les dice que la carta es roja. En términos de eventos, lo que les ha dicho entonces esta persona, es que el evento

$$B = \text{«seleccionar una roja»}$$

acaba de ocurrir. Dada esta información, ¿cuál es ahora la probabilidad de A? Dicha probabilidad se actualiza a ser 1/2 porque de las rojas, la mitad son corazones.

Supongamos, en cambio, que se les informa que ocurrió el evento

$$C = \text{«seleccionar una negra»}.$$

Entonces $\mathbb{P}(A)$ se actualiza a ser 0 porque si la carta es negra, entonces no puede ser corazón.

Si la información fuera que ha ocurrido el evento

$$D = \text{«seleccionar un rey»},$$

la actualización para $\mathbb{P}(A)$ es 1/4, porque hay cuatro reyes de los cuales sólo uno es corazón.

Las probabilidades «actualizadas» se denominan probabilidades *condicionales*. La notación es la siguiente:

$$\mathbb{P}(A \mid B) = 1/2,$$
$$\mathbb{P}(A \mid C) = 0,$$
$$\mathbb{P}(A \mid D) = 1/4,$$

mientras que $\mathbb{P}(A) = 1/4$.

Definición 4.1 (Probabilidad condicional) Si A y B son eventos con $\mathbb{P}(B) > 0$, *la probabilidad condicional de A dado B*, denotada por $\mathbb{P}(A \mid B)$, se define por

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{4.1}$$

Observemos que en el ejemplo anterior acerca de la selección de una carta, se cumple la definición anterior, ya que

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{13}{52}}{\frac{1}{2}} = 1/2,$$
$$\mathbb{P}(A \mid C) = \frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)} = \frac{0}{\frac{1}{2}} = 0,$$

y

$$\mathbb{P}(A \mid D) = \frac{\mathbb{P}(A \cap D)}{\mathbb{P}(D)} = \frac{\frac{1}{52}}{\frac{4}{52}} = 1/4.$$

La probabilidad condicional induce (ver Ejercicio 4.14) un espacio de probabilidad a partir del espacio original $(\Omega, \mathscr{A}, \mathbb{P})$. La interpretación es que si ya ocurrió el evento B, entonces B se convierte en un nuevo espacio muestral, y lo que es ahora relevante es hablar de las probabilidades con respecto a esta restricción.

Teorema 4.4 (Regla de la multiplicación) *Para dos eventos A y B con $\mathbb{P}(B) > 0$, se cumple*

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\ \mathbb{P}(A \mid B). \tag{4.2}$$

Demostración. Inmediata de la Definición (4.1). □

Observe que si $\mathbb{P}(B)$ es conocido, que la ecuación (4.1) es útil si se conoce $\mathbb{P}(A \cap B)$ y se desconoce $\mathbb{P}(A \mid B)$. La ecuación (4.2) es útil si se conoce $\mathbb{P}(A \mid B)$ y se desconoce $\mathbb{P}(A \cap B)$. Ambos casos se presentan en la práctica.

Hacemos las siguientes observaciones:

(a) $\mathbb{P}(A \mid \Omega) = \dfrac{\mathbb{P}(A \cap \Omega)}{\mathbb{P}(\Omega)} = \mathbb{P}(A).$

(b) $\mathbb{P}(\varnothing \mid B) = \dfrac{\mathbb{P}(\varnothing \cap B)}{\mathbb{P}(B)} = \mathbb{P}(\varnothing) = 0.$

(c) $\mathbb{P}(A \mid B)$ no es lo mismo que $\mathbb{P}(B \mid A)$. (verificar)

(d) $\mathbb{P}(\Omega \mid B) = \dfrac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \dfrac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$

(e) $\mathbb{P}(A \mid B) + \mathbb{P}(A^c \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} + \dfrac{\mathbb{P}(A^c \cap B)}{\mathbb{P}(B)} =$

$$\frac{\mathbb{P}\left(A \cap B\right) + \mathbb{P}\left(A^c \cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(B\right)}{\mathbb{P}(B)} = 1.$$

(f) No es cierto que $\mathbb{P}(A \mid B) + \mathbb{P}(A \mid B^c) = 1$. (verificar)

(g) La relación (4.2) es equivalente a $\mathbb{P}(A \cap B) = \mathbb{P}(A)\ \mathbb{P}(B \mid A)$ (siempre y cuando $\mathbb{P}(A) > 0$). (verificar)

Teorema 4.5 (Regla de la multiplicación, generalizada) *Para una colección finita de eventos $A_1, \ldots, A_n$ tales que $\mathbb{P}(\cap_{i=1}^m A_i) > 0$ para $m < n$, se tiene que*

$$\mathbb{P}\big(\cap_{i=1}^n A_i\big) = \mathbb{P}(A_1)\ \mathbb{P}(A_2 \mid A_1)\ \mathbb{P}(A_3 \mid A_1 \cap A_2)\ \mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) \cdots \\ \cdots \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Demostración. Por inducción. Para $n = 2$ resultado claramente vale, pues se cumple la regla de multiplicación anteriormente establecida. Supongamos que es válido el resultado para n, y agreguemos un evento adicional, A_{n+1}. Se cumple entonces

$$\mathbb{P}\big(\cap_{i=1}^{n+1} A_i\big) = \mathbb{P}\Big(A_{n+1} \cap \big(\cap_{i=1}^n A_i\big)\Big) = \\ \mathbb{P}\big(\cap_{i=1}^n A_i\big)\, \mathbb{P}\big(A_{n+1} \mid \cap_{i=1}^n A_i\big)$$

y al incorporar la hipótesis de inducción para el término $\mathbb{P}\big(\cap_{i=1}^n A_i\big)$ se obtiene directamente la cadena de términos multiplicados entre sí que se desea demostrar. □

Observación 4.1 (Árbol de probabilidades) En el dispositivo llamado *árbol de probabilidades*, cada rama representa un posible resultado de un experimento aleatorio (es decir, un elemento del espacio muestral). Es un dispositivo muy útil para representar experimentos, especialmente aquellos que pueden concebirse como formado por etapas de manera secuencial. Ilustraremos el concepto con el siguiente experimento: Se lanza un dado, y luego se lanza un número de monedas igual al número de puntos mostrados por el dado. Se observa el número total de águilas obtenido. El árbol de probabilidades se muestra en la Figura 4.1. Este árbol es sencillo, de dos etapas, pero pueden surgir árboles con un número mayor de etapas, e inclusive, con un número infinito de etapas (Ejercicio 4.2).

Para obtener la probabilidad de una rama, se calcula el producto de todas las probabilidades que se recorren en la rama. Esto a final de cuentas no es más ni menos que una aplicación de la regla de la multiplicación. Note que en cada componente de la rama, la probabilidad que se anota en el árbol entre un nodo y otro, es precisamente, una probabilidad condicional.

Ejemplo 4.2 Aplique el concepto de árbol de probabilidad y la regla de la multiplicación para resolver el siguiente problema. En un laboratorio de investigación, una rata en un laberinto en forma de T tiene dos opciones: dar vuelta a la izquierda y obtener alimento, o dar vuelta a la derecha y recibir una descarga eléctrica. La primera vez la rata elige al azar. Suponga que después de obtener alimento, las probabilidades de izquierda-derecha son 0.6 y 0.4 en la siguiente prueba. Después de una descarga eléctrica, suponga que las probabilidades de izquierda-derecha son 0.8 y 0.2.

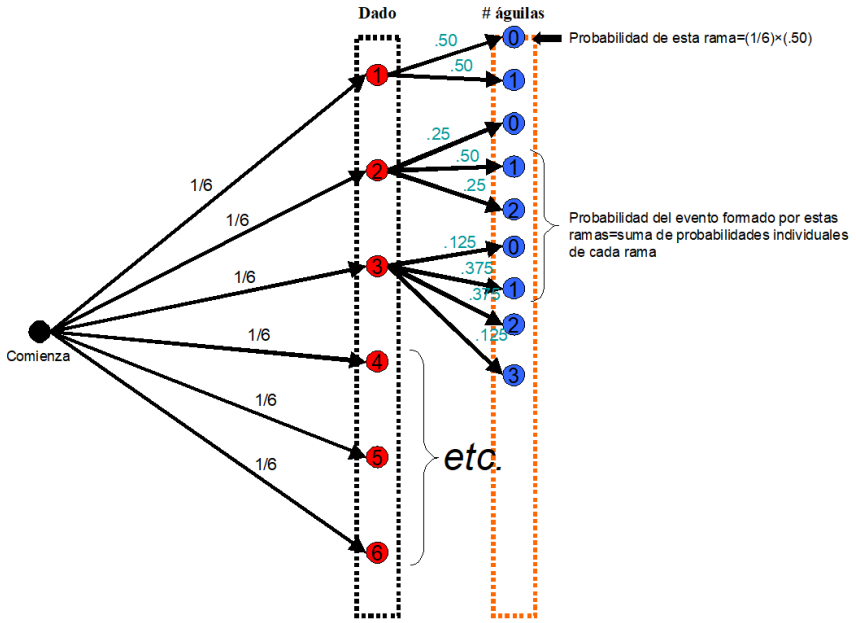

Figura 4.1: Árbol de probabilidades.
Árbol que representa el experimento de lanzar primero un dado y luego n monedas si el número de puntos señalado por el dado es n. La primera etapa del árbol corresponde al lanzamiento del dado, y la segunda etapa al lanzamiento de las monedas. De cada resultado del dado emana un distinto número de ramas, porque los resultados posibles para el número de águilas varían según el resultado del dado. Las ramas del árbol se rotulan con probabilidades condicionales: la probabilidad de que partiendo de un nodo resulte el nodo inmediato.

(a) ¿Cuál es la probabilidad de que la rata vaya a la izquierda en el segundo ensayo?

(b) ¿Cuál es la probabilidad de que la rata vaya a la izquierda en el tercer ensayo?

(c) Dado que dio vuelta a la izquierda en el tercer ensayo, ¿cuál es la probabilidad de que dio vuelta a la izquierda en el primero?

Note que en este ejercicio de probabilidad, se dan por hecho las probabilidades 0.6 y 0.8 (se *suponen* como dadas). Si estas probabilidades no se conocieran, simplemente no podríamos dar respuesta a las preguntas (a), (b), (c). El problema se convertiría entonces en uno de inferencia estadística, y la metodología sería observar algunos ensayos con ratas elegidas al azar en el marco de un experimento, con el fin de determinar las probabilidades anteriores, tomando en cuenta la incertidumbre impuesta por la situación.

El siguiente ejemplo ilustra una situación real en la que la teoría de probabilidad por si sola no puede resolver el problema; será necesario recurrir al concepto de un modelo estadístico, lo cual se hará más adelante.

Ejemplo 4.3 ¿Cuál es la probabilidad de que el número de gises rotos en una caja de 150 gises blancos de la marca Vividel sea a lo más 5? Aquí, si el experimento es observar el número de gises rotos que contiene la caja, entonces $\Omega = \{0, 1, 2, \ldots, 150\}$. Poniendo $p_i = \mathbb{P}\big(\{i\}\big)$ para $i = 0, 1, \ldots, 150$, la respuesta (ver Teorema 3.1) a la pregunta es simplemente $p_0 + p_1 + \cdots + p_5$. (Note, de paso, que *no* es cierto que $\mathbb{P}\big(\{0,\, 1,\, 2,\, 3,\, 4,\, 5\}\big) = 6/151$, porque el espacio muestral *no* es uniforme). Es sensato imaginarse que cada gis

individual tiene una probabilidad p de romperse a la hora del empaque. Por lo tanto, un modelo de probabilidad razonable para describir el número de gises rotos es el modelo binomial, dado por

$$p_i = \binom{150}{i} p^i (1-p)^{150-i}.$$

Pero, ¿cuál es el valor de p? A menos de que el oráculo nos los diga, no sabemos de antemano cuál es su valor. La teoría de probabilidad lo único que puede entonces contestar es: Si p acaso fuera 0.01, entonces la respuesta a la pregunta sería 0.99579, y si p fuera 0.05, entonces la respuesta sería 0.23444. Pero el caso es que si el valor de p no se conoce, entonces hay que plantear un problema de estadística, en el cual se hagan observaciones aleatorias de gises, observar si están rotos o no, y determinar con estos datos el valor plausible de p (con lo cual se determinan valores plausibles de la cantidad $p_0 + p_1 + \cdots + p_5$ a modo de dar respuesta a la pregunta original). Esto, en inferencia estadística recibe el nombre de problema de *estimación paramétrica*, y será objeto de atención en un capítulo futuro.

4.3. Independencia

En el ejemplo anterior relativo a la selección de una carta, observemos que $\mathbb{P}(A) = 1/4$ y que también $\mathbb{P}(A \mid D) = 1/4$. Es decir, en el caso de que la información sea que ocurrió el evento de ser rey, la probabilidad original y la actualizada de ser corazón permanecen exactamente iguales. Esta coincidencia fortuita entre una probabilidad no-condicional con una condicional, da lugar a una característica extraordinariamente importante en teoría de probabilidad y estadística, que se llama *independencia*. La definición formal

de este concepto, es como sigue.

Definición 4.2 (Independencia) Decimos que dos eventos A y B *son independientes*, si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\ \mathbb{P}(B)$.

Notemos que esta definición no solicita ni que $\mathbb{P}(A) > 0$ ni que $\mathbb{P}(B) > 0$. Sin embargo, si $\mathbb{P}(A) > 0$, entonces independencia implica $\mathbb{P}(B \mid A) = \mathbb{P}(B \cap A)/\,\mathbb{P}(A) = \mathbb{P}(B)$, y si $\mathbb{P}(B) > 0$, entonces $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\,\mathbb{P}(B) = \mathbb{P}(A)$. Es decir, la noción de independencia coincide con la interpretación intuitiva ilustrada en el ejemplo de cartas. Notemos que el evento $\varnothing$ es independiente de cualquier otro evento A.

Es muy común que se confundan las nociones de que dos eventos sean *independientes*, y que dos eventos sean *ajenos*. Note que la cualidad de ajenos la da una característica física de los eventos, mientras que la cualidad de independencia es más bien una propiedad de la medida de probabilidad. Intuitivamente, dos eventos *ajenos* significa que no pueden ocurrir al mismo tiempo; *independientes* significa que el hecho de que ocurra uno, no afecta la probabilidad de ocurrencia del otro.

Teorema 4.6 *Si A y B son eventos independientes, entonces*

(a) A y B^c son independientes,

(b) A^c y B son independientes, y

(c) A^c y B^c son independientes.

Demostración.

De (a) (Los demás, se dejan como ejercicio.) Por demostrar, que $\mathbb{P}(A \cap B^c) = \mathbb{P}(A)\,\mathbb{P}(B^c)$. Debido a que $A = (A \cap B) \cup (A \cap B^c)$, donde la unión es disjunta, y dado que $\mathbb{P}$ es medida de probabilidad, se concluye $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$, o bien $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$. Pero por ser A y B independientes, se obtiene $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A)\,\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\,\mathbb{P}(B^c)$ por la ley del complemento.

□

La generalización a una colección de más de dos eventos es la siguiente.

Definición 4.3 (Independencia mutua) Decimos que los eventos en una colección de eventos $A_1, A_2, \ldots$ *son mutuamente independientes*, si se cumple $\mathbb{P}\big(\cap_{i=1}^{m} A_{j_i}\big) = \prod_{i=1}^{m} \mathbb{P}(A_{j_i})$ para todo $m \in \mathbb{N}$ y cualquier colección de m índices distintos $j_1, \ldots, j_m$.

En resumen, la noción de independencia es relevante porque permite la asignación de probabilidades a intersecciones de eventos, con base en una consideración de una noción específica (el que la probabilidad de un evento no se altera por la ocurrencia de otro). Esto es, en ocasiones puede establecerse $\mathbb{P}(A \cap D) = \mathbb{P}(A)\ \mathbb{P}(D)$ por cálculo directo, como se verificó en el ejemplo de las cartas, y en ocasiones se establece $\mathbb{P}(A \cap D) = \mathbb{P}(A)\ \mathbb{P}(D)$ por postulación en un contexto dado.

Ejemplo 4.4 Considere la situación de lanzar una moneda dos veces. Sean

$$A = \text{«águila en el 1er lanzamiento» y}$$

$$B = \text{«águila en el 2do lanzamiento».}$$

Claramente, $\mathbb{P}(A) = \mathbb{P}(B) = 1/2$. La independencia entre A y B, es decir, la especificación de que $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) = (1/2)(1/2) = 1/4$, más que una deducción matemática a través de un teorema, en este caso es una afirmación postulada en el modelo de probabilidad, por razones de contexto. Dicha postulación es sensata en la medida en que se crea que la probabilidad del 2do lanzamiento debe permanecer constante irrespectivamente de lo que ocurra en el 1ro. En este ejemplo, A y B son independientes «por decreto».

Ejemplo 4.5 Imagínese que la moneda usada en el ejemplo anterior fuese una moneda mágica. La primera vez que se lanza tiene probabilidad 1/2 de águila, y la segunda vez, cambia su probabilidad de águila de la siguiente manera: Si la primera vez cayó águila, ahora hace probabilidad de águila 1/4, y si cayó sol, hace probabilidad de águila 3/4. En esta situación *no* sería sensato postular independencia entre A y B, porque la probabilidad de águila en el 2do lanzamiento *sí* cambia dependiendo de cómo haya caído en el 1ro.

Ejemplo 4.6 Un vendedor de billetes de la Lotería Nacional[1] profesa con entusiasmo «¡Compre el esperado, el esperado, la terminación seis!». La im-

[1] Para otro ejercicio acerca de la Lotería, véase el Ejercicio 4.13.

plicación parecería ser que en virtud de que el 6 no ha salido en varios sorteos, que por alguna razón, la probabilidad de terminar en 6 es *mayor* que si el 6 hubiera sucedido recientemente. Esto es una falacia (aunque dota de mucha diversión y pasión al juego de la lotería), porque en términos del concepto de independencia, el que el 6 no haya salido recientemente no altera en lo más mínimo la probabilidad de que se obtenga el 6 en el próximo sorteo. Esto es, los eventos «6 no ha salido en los últimos n sorteos» y «el 6 saldrá en el próximo sorteo» son independientes y que la probabilidad de que ocurra el segundo es exactamente igual, ocurra o no ocurra el primero. Como ejercicio, tomando en cuenta que la lotería tiene 50,000 números, calcule la probabilidad de que el premio mayor termine en 6. Sin embargo, si la lotería fuera «sin reemplazo», es decir, que las bolas numeradas una vez sorteadas no se regresan a la urna antes del siguiente sorteo, estos eventos *no* serían independientes, y el vendedor de lotería tendría razón.

4.4. Regla de Bayes

La regla de Bayes es una ley de probabilidad qué es útil en situaciones para las que se conozcan ciertas probabilidades condicionales.

Definición 4.4 (Partición) Una *partición* de Ω es un conjunto de eventos disjuntos $A_1, \ldots, A_n$ tales que $\mathbb{P}(A_i) > 0$ y $\cup_{i=1}^{n} A_i = \Omega$.

Teorema 4.7 (Ley de la probabilidad total) *Sea $A_1, \ldots, A_n$ una partición*

de Ω y B un evento arbitrario. Entonces,

$$\mathbb{P}(B) = \mathbb{P}(B \mid A_1)\,\mathbb{P}(A_1) + \cdots + \mathbb{P}(B \mid A_n)\,\mathbb{P}(A_n).$$

Demostración.

$$\mathbb{P}(B) \;=\; \mathbb{P}\big(\cup_{i=1}^{n}\,(B \cap A_i)\big) \;=\; \sum_{i=1}^{n} \mathbb{P}(B \cap A_i) \;=\; \sum_{i=1}^{n} \mathbb{P}(B \;\mid\; A_i)\;\mathbb{P}(A_i),$$

por ser $A_1, \ldots, A_n$ una partición de Ω, $\mathbb{P}$ una medida de probabilidad, y por la regla de la multiplicación. □

Corolario 4.8 *Si* $\mathbb{P}(A), \mathbb{P}(A^c) > 0$,

$$\mathbb{P}(B) = \mathbb{P}(B \mid A)\;\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\;\mathbb{P}(A^c).$$

Ejemplo 4.7 En el dispositivo conocido como árbol de probabilidad (Observación 4.1), la ley de la probabilidad total se manifiesta en la práctica de sumar probabilidades sobre todas las ramas que especifican un evento de interés.

Teorema 4.9 (Regla de Bayes) *Sea $A_1, \ldots, A_n$ una partición de Ω y B un evento arbitrario tal que $\mathbb{P}(B) > 0$. Entonces, para toda i se cumple*

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\;\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B \mid A_j)\;\mathbb{P}(A_j)}.$$

Demostración. Por definición de probabilidad condicional, la ley multiplicativa, y la ley de la probabilidad total tenemos que

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A_i)\ \mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B \mid A_j)\ \mathbb{P}(A_j)}.$$

□

La regla de Bayes sirve para «invertir» las probabilidades condicionales, es decir, para calcular $\mathbb{P}(A_i \mid B)$ en términos de $\mathbb{P}(B \mid A_i)$. Es muy común en la práctica que uno pueda especificar o conocer los valores de $\mathbb{P}(B \mid A_i)$ para una partición $A_1, \ldots, A_n$. Un ejemplo típico es el siguiente.

Ejemplo 4.8 La producción de un día en una industria es producida por tres máquinas con porcentajes 20 %, 30 %, y 50 % respectivamente. Suponga que la fracción de artículos defectuosos producida por la máquina uno es 5 %, 3 % por la máquina dos y 1 % por la máquina tres.

(a) ¿Cuál es la fracción de artículos defectuosos en la producción?

(b) Si se elige un artículo al azar de la producción total y se encuentra que es defectuoso, ¿cuál es la probabilidad de que éste provenga de la tercera máquina?

Solución de (a) Sean los eventos

$$B = \text{«artículos defectuosos en la producción»}$$

y A_i = «artículos producidos por la máquina i», $i = 1, 2, 3$. De esta manera, $\mathbb{P}(A_1) = 0.2$, $\mathbb{P}(A_2) = 0.3$, $\mathbb{P}(A_3) = 0.5$, $\mathbb{P}(B \mid A_1) =$

0.05, $\mathbb{P}(B \mid A_2) = 0.03$ y $\mathbb{P}(B \mid A_3) = 0.01$. Sustituyendo valores en la ley de probabilidad total,

$$\mathbb{P}(B) = \mathbb{P}(B \mid A_1)\,\mathbb{P}(A_1) + \mathbb{P}(B \mid A_2)\,\mathbb{P}(A_2) + \mathbb{P}(B \mid A_3)\,\mathbb{P}(A_3) = 0.05 \times 0.2 + 0.03 \times 0.3 + 0.01 \times 0.5 = 0.024,$$

es decir, 2.4 % artículos defectuosos en la producción.

Solución de (b) Se quiere obtener $\mathbb{P}(A_3 \mid B)$. Por la Regla de Bayes,

$$\mathbb{P}(A_3 \mid B) = \frac{\mathbb{P}(B \mid A_3)\ \mathbb{P}(A_3)}{\sum_{j=1}^{3} \mathbb{P}(B \mid A_j)\ \mathbb{P}(A_j)} = \frac{0.01 \times 0.5}{0.024} = 0.208.$$

En cursos más avanzados de probabilidad, se incluyen leyes para eventos más sofisticados, por ejemplo aquellos eventos que se construyen con *secuencias* de eventos. Por ejemplo, si $A_1, A_2, \ldots$ es una sucesión de eventos (no necesariamente ajenos a pares), ¿existe alguna forma de deducir $\mathbb{P}\big(\cup_{i=1}^{\infty} A_i\big)$? Para formalizar esto, lo primero que debe hacerse es definir con precisión lo que significaría el límite de una sucesión de eventos, para luego obtener una ley del siguiente tipo: $\mathbb{P}\big(\lim_{n\to\infty} A_n\big) = \lim_{n\to\infty} \mathbb{P}(A_n)$. Note que en esencia, sigue siendo este resultado una *ley* en el sentido de que da una forma para calcular la probabilidad de un evento como función de otras probabilidades.

Ejercicios

4.1 Complete la demostración del Teorema 4.6.

4.2 También con un árbol de probabilidad pueden conceptualizarse experimentos en los que el número de etapas es infinito. Utilícelo para abordar el siguiente juego de dados, llamado *craps* en inglés. Es un juego clásico de casino de Las Vegas. El jugador lanza dos dados hasta que *gane* o *pierda*. Gana en el primer lanzamiento si obtiene un total de 7 u 11; pierde en el primer lanzamiento si obtiene un total de 2, 3, o 12. Si obtiene cualquier otro total en su primer lanzamiento, ese total recibe el nombre de su *punto*. Luego lanza repetidamente los dados hasta obtener un total de 7 o su punto. Gana si obtiene su punto, y pierde si obtiene un 7. Encuentre la probabilidad de ganar. Establezca que es ligeramente menor a 0.5 (de aquí, la ventaja de la casa).

4.3 Dado $\mathbb{P}(A) = 0.5$, $\mathbb{P}(A \cup B) = 0.6$, encontrar $\mathbb{P}(B)$ si

(a) A y B son eventos disjuntos.

(b) A y B son eventos independientes.

(c) $\mathbb{P}(A \mid B) = 0.4$.

4.4 Busque al menos dos ejemplos que ilustren que dos eventos pueden ser independientes sin ser ajenos, y que pueden ser ajenos sin ser independientes.

4.5 Considere un modelo de probabilidad donde $\Omega = \{a, b, c, d, e\}$, con $\mathbb{P}\big(\{a\}\big) = \mathbb{P}\big(\{b\}\big) = \mathbb{P}\big(\{c\}\big) = 0.1$, $\mathbb{P}(\{d\}) = 0.4$, y $\mathbb{P}(\{e\}) = 0.3$. Considere los eventos $A = \{a, b, c\}$, $B = \{b, c, d\}$, y $C = \{a, c, e\}$.

(a) ¿A y B son independientes? (explique)

(b) ¿A y C son ajenos? (explique)

(c) Calcule $\mathbb{P}(A \mid C)$.

(d) Calcule $\mathbb{P}(A \cup B)$.

(e) Calcule $\mathbb{P}(A \cap C)$.

4.6 Demuestre:

(a) Si A y B son ajenos y C es un evento tal que $\mathbb{P}(C) > 0$, entonces $\mathbb{P}(A \cup B \mid C) = \mathbb{P}(A \mid C) + \mathbb{P}(B \mid C)$.

(b) Si $A \subset B$, entonces $\mathbb{P}(A \mid C) \leq \mathbb{P}(B \mid C)$.

4.7 Demuestre que si A_1, A_2, A_3 son independientes, entonces también lo son A_1^c, A_2^c, A_3^c. ¿Visualiza una demostración general para n eventos independientes en lugar de tres?

4.8 Suponga que 0.001 es la probabilidad de tener tuberculosis (TB). Una prueba clínica de detección de TB tiene las siguientes propiedades: Si la persona sí tiene TB, la prueba lo detecta con probabilidad 0.99. Si no tiene TB, hay probabilidad 0.002 de que la prueba indique que sí tiene TB. Suponga que una persona al azar resulta positivo. ¿Cuál es la probabilidad de que tenga TB?

4.9 Un fabricante de motores tiene en existencia 12 motores, de los cuales 2 son defectuosos. Existen tres estrategias para su empaque:

(a) Los 12 motores en una sola caja.

(b) Dos cajas cada una con un defectuoso contenida en ella.

(c) Dos cajas, una de las cuales contiene los 2 defectuosos.

Suponga que el cliente tiene la política de examinar dos motores al azar si vienen empacados en una sola caja, y un motor de cada caja si vienen empacados en dos cajas. ¿Cuál estrategia de empaque le conviene al fabricante?

4.10 (Pruebas sensoriales en la industria alimenticia) Se investiga la capacidad de que un consumidor detecte la diferencia en sabor entre tres presentaciones de una misma marca de cerveza: Botella de vidrio, lata de aluminio, y barrica a granel. En tres vasos iguales se sirve cerveza de cada tipo, y se le da a un sujeto de prueba en un orden no especificado ni identificado. Se le pedirá al sujeto que con el paladar intente identificar la presentación servida en cada vaso. Se registrará el número de aciertos.

(a) Construya un espacio muestral para este experimento.

(b) Identifique el evento $A =$ «no más de un acierto».

(c) Bajo la suposición de que el sujeto en efecto está adivinando al azar, es decir, que es realmente imperceptible la diferencia, calcule $\mathbb{P}(A)$.

4.11 Se dice que un sistema de componentes *está conectado en paralelo* si para que el sistema funcione, basta con que funcione al menos uno de los componentes. Suponga que se usarán componentes que fallan en forma independiente uno del otro y que la probabilidad de que cada componente funcione es 0.9. ¿Cuántos componentes se requieren para que el sistema funcione con probabilidad mayor o igual que 0.99?

4.12 Explique la paradoja acerca del terrorista en un avión, en la que se razona como sigue: La probabilidad de que dos personas se suban a un avión portando cada uno una bomba, es menor que la probabilidad de que una sola persona lo haga. Por lo tanto, para provocar que mi vuelo sea menos susceptible de un atentado, me subiré yo al avión portando una bomba, con el fin de disminuir la probabilidad de que un terrorista auténtico aborde el mismo avión.

4.13 Un aficionado de corazón de la Lotería Nacional afirmó en alguna ocasión que los números «feos» tales como `11111, 33333, ...` nunca los juega porque esos tienen muy poca probabilidad de salir premiados. En cambio los números «bonitos», afirmó, tales como `13876, 42864, ...` son siempre mejores para jugar. Explique la razón por la que este razonamiento es equivocado, en términos del siguiente enunciado que usa lenguaje de teoría de probabilidad: Considere dos eventos A y B y dos resultados $a \in A$, $b \in B$. Entonces, $\mathbb{P}(A) < \mathbb{P}(B)$ no necesariamente implica que $\mathbb{P}\big(\{a\}\big) < \mathbb{P}\big(\{b\}\big)$. ¿Es cierto que si $\mathbb{P}\big(\{a\}\big) < \mathbb{P}\big(\{b\}\big)$, entonces $\mathbb{P}(A) < \mathbb{P}(B)$?

4.14 (Espacio inducido por probabilidad condicional)

(a) Sea $(\Omega, \mathscr{A}, \mathbb{P})$ un espacio de probabilidad, y fije $A \in \mathscr{A}$ tal que $\mathbb{P}(A) > 0$. Defina, para todo $B \in \mathscr{A}$,

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

y sea $\mathscr{A}_A = \{A \cap B \mid B \in \mathscr{A}\}$. Demuestre que $(A, \mathscr{A}_A, \mathbb{P}_A)$ es un espacio de probabilidad. El espacio $(A, \mathscr{A}_A, \mathbb{P}_A)$ recibe el nombre de *espacio de probabilidad inducido por la probabilidad condicional*. (Note que esto tiene mucha relación con el Ejercicio 2.9.)

(b) Tomando $\Omega = \{1, 2, 3\}$ con $\mathbb{P}\big(\{1\}\big) = 0.3$, $\mathbb{P}\big(\{2\}\big) = 0.3$, $\mathbb{P}\big(\{3\}\big) = 0.4$, $A = \{1, 2\}$ y $\mathscr{A} = 2^{\Omega}$, encuentre explícitamente $\big(A, \mathscr{A}_A, \mathbb{P}_A\big)$.

4.15 El llamado *Problema de Monty Hall* (Ver Morgan et al., 1991) ha sido motivo de innumerables discusiones tanto en canales académicos formales como en medios populares. Versa sobre temas que en el fondo son de probabilidad condicional. Se plantea como sigue.

> En un concurso hay tres puertas cerradas A, B, C. Al participante se le dice que detrás de una de las puertas hay un automóvil, y detrás de las otras dos hay cabras. El concursante elige la puerta A —manteniéndola cerrada—. El anfitrión abre una de las dos puertas restantes y le muestra al concursante una de las cabras. Se le plantea al concursante el siguiente dilema: Conservar la puerta elegida originalmente, o bien cambiarla por la puerta que el anfitrión mantiene cerrada. ¿Qué estrategia conviene al concursante para aumentar la probabilidad de ganar el automóvil, cambiar o no cambiar su elección original?

Muestre que la respuesta es que le conviene cambiar de puerta.

Capítulo 5

Variables aleatorias

En este capítulo se hace una introducción a un importante y útil concepto en teoría de probabilidad: el de *variable aleatoria*. Comencemos con las definiciones matemáticas, para enseguida tratar de explicar la razón por la que es importante y muy útil esta noción.

5.1. Definiciones básicas

Definición 5.1 (Imagen inversa) Sea X una función $\Omega \to \mathbb{R}$, y $A \subset \mathbb{R}$ un subconjunto arbitrario. Llamamos la *imagen inversa de A bajo X*, denotada por $X^{-1}(A)$ al subconjunto de Ω dado por

$$X^{-1}(A) = \big\{\omega \in \Omega \mid X(\omega) \in A\big\}.$$

En palabras, la imagen inversa es el conjunto de resultados que bajo la función X caen en A. Pueden demostrarse de inmediato algunas propiedades de imágenes inversas:

Proposición 5.1 *Sea X una función $\Omega \to \mathbb{R}$, y $A_1, A_2, \ldots$ una sucesión de subconjuntos de $\mathbb{R}$. Entonces*

(a) $X^{-1}\big(\cup_{i=1}^{\infty} A_i\big) = \bigcup_{i=1}^{\infty} X^{-1}(A_i)$.

(b) $X^{-1}\big(\cap_{i=1}^{\infty} A_i\big) = \bigcap_{i=1}^{\infty} X^{-1}(A_i)$.

(c) $X^{-1}\big(A_1^c\big) = [X^{-1}(A_1)]^c$.

(d) Si $A_1, A_2, \ldots$ son disjuntos a pares, entonces $X^{-1}(A_1)$, $X^{-1}(A_2), \ldots$ son disjuntos a pares.

Demostración.

De (a) $\omega \in X^{-1}\big(\cup_{i=1}^{\infty} A_i\big) \iff \cup_{i=1}^{\infty} A_i \iff X(\omega) \in A_i$ para algún $i \iff \omega \in X^{-1}(A_i)$ para algún $i \iff \omega \in \cup_{i=1}^{\infty} X^{-1}(A_i)$.

De (b) $\omega \in X^{-1}\big(\cap_{i=1}^{\infty} A_i\big) \iff X(\omega) \in \cap_{i=1}^{\infty} A_i \iff X(\omega) \in A_i \ \forall i \iff \omega \in X^{-1}(A_i)\, \forall i \iff \omega \in \cap_{i=1}^{\infty} X^{-1}(A_i)$.

De (c) Ejercicio. Note que la notación c de complemento, tiene significado distinto de cada lado de la ecuación; en la izquierda es complemento con respecto a $\mathbb{R}$, mientras que del lado derecho es complemento con respecto a Ω.

De (d) Suponga que hay un ω tal que $\omega \in X^{-1}(A_i) \cap X^{-1}(A_j)$ para $i \neq j$. Entonces $X(\omega) \in A_i \cap A_j$, lo cual no puede suceder porque $A_i \cap A_j = \varnothing$.

□

Notación 5.1 Es conveniente y usual emplear la siguiente notación alternativa para escribir la imagen inversa. En lugar de $X^{-1}\big(\{2\}\big)$, se escribe $X = 2$ o $X \in \{2\}$; en lugar de $X^{-1}\big((-\infty, x\,]\big)$, se escribe $X \leq x$ o $X \in (-\infty, x\,]$; en lugar de $X^{-1}\big((1, 2]\big)$ se escribe $1 < X \leq 2$ o $X \in (1, 2]$.

Definición 5.2 (Variable aleatoria) Sea $(\Omega,\, \mathscr{A},\, \mathbb{P})$ un espacio de probabilidad. Decimos que una función $X\colon \Omega \to \mathbb{R}$ es una *variable aleatoria*, si ésta cumple

$$X^{-1}\big((-\infty, x\,]\big) \in \mathscr{A},\ \forall x \in \mathbb{R}. \tag{5.1}$$

Una variable aleatoria X es entonces una función de valores reales definida sobre el espacio muestral Ω, que cumple cierto requerimiento respecto a las imágenes inversas de todos los conjuntos de la forma $(-\infty, x\,]$. Observe que ser o no variable aleatoria no sólo es una propiedad de la función misma sino también de la σ-álgebra $\mathscr{A}$ (ver el Ejemplo 5.3).

Usualmente se denota a las variables aleatorias por letras mayúsculas, tales como X, Y, Z, *etc*. y por letras minúsculas x, y, z, *etc*., a los valores que toman dichas variables aleatorias.

Ejemplo 5.1 La Tabla 5.1 considera algunos experimentos así como algunas variables aleatorias asociadas a los experimentos. Note que asociado a un mismo experimento aleatorio puede haber más de una variable aleatoria.

Ejemplo 5.2 Considere el experimento de lanzar una moneda balanceada tres veces. Sea X el número de águilas obtenidos de los tres lanzamientos. Denotemos el resultado del lanzamiento de una moneda por la letra A, si

Experimento aleatorio (ω)	**Variable aleatoria ($X(\omega)$)**
Lanzamiento de dos dados	Suma de los números resultantes
Ensayo clínico con 25 pacientes	Número de pacientes que responden favorablemente a un tratamiento
La evolución de una parcela de maíz tras aplicar una cantidad de fertilizante	Producción en toneladas en esa parcela
Observación continua del índice de la bolsa de valores durante un trimestre	El valor máximo del índice alcanzado en el trimestre
Una reacción química de oxidación	El tiempo transcurrido hasta alcanzarse cierto estado de oxidación
Observación de un sector de la bóveda celeste	El número de quásares contabilizados en ese sector
Un huracán en el Atlántico	La velocidad máxima de ráfagas
Un huracán en el Atlántico	La trayectoria sobre la Tierra vista como una esfera
Comportamiento de una red social durante tres días consecutivos	El número de nuevas conexiones realizadas entre sus miembros
Un paquete de Internet viaja entre dos nodos	El tiempo de viaje del paquete
Obtención de una imagen digitalizada mediante una cámara de teléfono celular	El número de pixeles de la imagen que exceden un valor umbral determinado
Perfil de crecimiento de un organismo (planta, animal)	El tamaño máximo adquirido
Perfil de crecimiento de un organismo (planta, animal)	El tiempo necesario para adquirir el tamaño máximo

Tabla 5.1: Ejemplos de funciones en diversos ámbitos, definidas sobre elementos aleatorios ω de un espacio de probabilidad $(\Omega, \mathscr{A}, \mathbb{P})$. El que sean estas funciones variables aleatorias dependerá del requerimiento técnico $X^{-1}(B) \in \mathscr{A}$, $\forall B \in \mathscr{B}$.

éste fue águila, y por la letra S si éste fue sol. El espacio muestral asociado al experimento es

$$\Omega = \big\{\text{AAA, AAS, ASA, SAA, SSA, SAS, ASS, SSS}\big\}.$$

La variable aleatoria X asigna un número a cada punto del espacio muestral, es decir,

$$X(\text{AAA}) = 3,\ X(\text{AAS}) = X(\text{ASA}) = X(\text{SAA}) = 2,$$
$$X(\text{SSA}) = X(\text{SAS}) = X(\text{ASS}) = 1,\ X(\text{SSS}) = 0,$$

y de esta forma el conjunto de valores posibles de X es $\{0, 1, 2, 3\}$.

En cursos más generales sobre teoría de medida, se habla de funciones *medibles*. Si se cuenta con dos conjuntos Ω y Θ, dos σ-álgebras $\mathscr{A}$ y $\mathscr{B}$ (de subconjuntos de Ω y Θ respectivamente), y una función $f\colon \Omega \to \Theta$, se dice que f es $\mathscr{A} - \mathscr{B}$ *medible* si $f^{-1}(B) \in \mathscr{A}$, $\forall B \in \mathscr{B}$. Un resultado, que no se encuentra al alcance del presente curso, es que si una función X cumple la definición anotada (5.1) para una variable aleatoria, entonces se cumple $X^{-1}(B) \in \mathscr{A}$, $\forall B \in \mathscr{B}$. Recuerde que aquí $\mathscr{B}$ representa a los llamados conjuntos borelianos, es decir, la σ-álgebra generada por los subconjuntos de $\mathbb{R}$ de la forma $(-\infty, x\,]$. Con ello, una variable aleatoria no es más que una función $\Omega \to \mathbb{R}$ que es $\mathscr{A} - \mathscr{B}$ medible. Entonces, dando como hecho (sin demostrar por ahora) de que para una variable aleatoria X, se cumple que $X^{-1}(B) \in \mathscr{A}$, $\forall B \in \mathscr{B}$, podemos sin duda hacer la siguiente definición.

Definición 5.3 (Distribución de una v.a.) Si X es una variable aleatoria, la

distribución de X es la función $\mathbb{P}_X : \mathscr{B} \to \mathbb{R}$ definida por

$$\mathbb{P}_X(B) = \mathbb{P}\left(X^{-1}(B)\right).$$

Para ilustrar la conveniencia de la Notación 5.1, escribimos

- $\mathbb{P}(X = 2)$ o $\mathbb{P}(X \in \{2\})$, en lugar de $\mathbb{P}\big(X^{-1}(\{2\})\big)$ o de $\mathbb{P}_X\big(\{2\}\big)$.
- $\mathbb{P}(X \leq x)$ o $\mathbb{P}(X \in (-\infty, x\,])$, en lugar de $\mathbb{P}\big(X^{-1}((1,2\,])\big)$ o de $\mathbb{P}_X(\{2\})$.
- $\mathbb{P}(1 < X \leq 2)$ o $\mathbb{P}\big(X \in (1,2\,]\big)$, en lugar de $\mathbb{P}\big(X^{-1}((1,2\,])\big)$ o de $\mathbb{P}_X\big((1,2\,]\big)$.

El siguiente es un resultado de suma importancia, que tendrá que ver con la futura explicación de por qué razón las variables aleatorias constituyen objetos útiles. Establece que una variable aleatoria induce un espacio de probabilidad. Como consecuencia, éste deberá cumplir todas las propiedades que se estudiaron en el Capítulo 4.

Teorema 5.2 *Sea* $(\Omega, \mathscr{A}, \mathbb{P})$ *un espacio de probabilidad. Si* X *es una variable aleatoria, entonces* $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$ *es un espacio de probabilidad.*

Demostración. Por demostrar, que $\mathbb{P}_X$ es medida de probabilidad sobre $\mathscr{B}$. Debido a que $X^{-1}(\mathbb{R}) = \Omega$, notamos que $\mathbb{P}_X(\mathbb{R}) = \mathbb{P}\big(X^{-1}(\mathbb{R})\big) = \mathbb{P}(\Omega) = 1$, por ser $\mathbb{P}$ una medida de probabilidad. La propiedad $\mathbb{P}_X(B) = \mathbb{P}\big(X^{-1}(B)\big) = \mathbb{P}(X \in B) \geq 0$ se sigue de que $\mathbb{P}(A) \geq 0, \forall A \in \mathscr{A}$ por ser $\mathbb{P}$ medida de probabilidad. Finalmente, si $A_1, A_2, \ldots$ una sucesión de subconjuntos de

$\mathscr{B}$ que son ajenos a pares, entonces

$$\mathbb{P}_X(\cup_{i=1}^{\infty} A_i) = \mathbb{P}\big(X^{-1}(\cup_{i=1}^{\infty} A_i)\big) = \mathbb{P}(\cup_{i=1}^{\infty} X^{-1}(A_i)) = \sum_{i=1}^{\infty} \mathbb{P}\big(X^{-1}(A_i)\big) = \sum_{i=1}^{\infty} \mathbb{P}_X(A_i),$$

pues por ser medida de probabilidad, $\mathbb{P}$ tiene la propiedad de aditividad numerable. □

Ejemplo 5.3 Considere el experimento de lanzar una moneda dos veces. Entonces $\Omega = \{\text{AA, SA, AS, SS}\}$. Sea X la función definida por «número de águilas», es decir, $X(\text{AA}) = 2$, $X(\text{SS}) = 0$, $X(\text{SA}) = X(\text{AS}) = 1$. Si $\mathscr{A} = \{\Omega, \varnothing\}$, entonces X no es variable aleatoria. Si $\mathscr{A} = 2^{\Omega}$, entonces X sí es variable aleatoria. En este caso, suponiendo un espacio uniforme, la distribución de X está dada por

$$\mathbb{P}_X(B) = \frac{1}{4}\,\mathbf{1}_B(0) + \frac{1}{2}\,\mathbf{1}_B(1) + \frac{1}{4}\,\mathbf{1}_B(2).$$

Note que $\mathbb{P}_X\big(\{1\}\big) = \mathbb{P}(X = 1) = 1/2$, $\mathbb{P}_X\big(\{2\}\big) = \mathbb{P}(X = 2) = 1/4$, y $\mathbb{P}_X\big(\{0\}\big) = \mathbb{P}(X = 0) = 1/4$. Note algo importante, que es que el hecho de que una función sea o no variable aleatoria, depende de quién es $\mathscr{A}$.

Ejemplo 5.4 Sea $\Omega = \{p \mid p \text{ es una persona}\}$. Sea $X(p) =$ «estatura», o $X(p) =$ «peso», o $X(p) =$ «edad».

Ejemplo 5.5 Sea

$$\Omega = \{p \mid p \text{ es un patrón climatológico diurno}\}.$$

Variables aleatorias de interés pudieran ser $X(p) =$ «temperatura máxima» o $X(p) =$ «humedad relativa media», o $X(p) =$ «máxima ráfaga de viento».

Ejemplo 5.6 Una máquina produce un tornillo. Sea $X =$ «el diámetro del tornillo». Al un nivel micrométrico, esta variable aleatoria pudiera ser relevante en un contexto de control de calidad o de maquinaria muy precisa.

Ejemplo 5.7 (V.A. constante, o degenerada) Sea $X(\omega) = c$, $\forall \omega \in \Omega$. Entonces X es una v.a. para toda σ-álgebra $\mathscr{A}$, porque

$$X^{-1}\big((-\infty, x\,]\big) = \begin{cases} \Omega & \text{si } x \geq c, \\ \varnothing & \text{si } x < c. \end{cases}$$

Verifique que su distribución se puede escribir mediante la fórmula $\mathbb{P}_X(B) = 1_B(c)$.

Ejemplo 5.8 (V.A. Bernoulli) Sea $\Omega = \{e,\ f\}$, es decir, un experimento con dos resultados posibles, llamados «éxito» y «fracaso». Sea

$$X(\omega) = \begin{cases} 1 & \text{si } \omega = e, \\ 0 & \text{si } \omega = f. \end{cases}$$

Notamos entonces que

$$X^{-1}\big((-\infty, x\,]\big) = \begin{cases} \varnothing & \text{si } x < 0, \\ \{f\} & \text{si } 0 \leq x < 1, \\ \Omega & \text{si } 1 \leq x. \end{cases}$$

Luego, la función X es variable aleatoria si $\mathscr{A} = 2^{\Omega}$. La variable aleatoria

X recibe el nombre de v.a. *Bernoulli*. Si se denota por p a la probabilidad $\mathbb{P}(X = 1)$, entonces la distribución de X se puede calcular mediante la fórmula

$$\mathbb{P}_X(B) = p\ \mathbf{1}_B(1) + (1-p)\ \mathbf{1}_B(0),$$

(verifique) la cual recibe el nombre de *distribución de Bernoulli con parámetro* p. Note que $\mathbb{P}_X\big(\{1\}\big) = \mathbb{P}(X = 1) = p$ y que $\mathbb{P}_X\big(\{0\}\big) = \mathbb{P}(X = 0) = 1 - p$.

Ejemplo 5.9 Dos variables aleatorias distintas pueden tener la misma distribución, aún sobre un mismo espacio muestral. Considere $\Omega = \{1, 2, 3\}$, $\mathscr{A} = 2^\Omega$, y $\mathbb{P}$ definida por $\mathbb{P}\big(\{1\}\big) = \mathbb{P}\big(\{2\}\big) = \big(\mathbb{P}\{3\}\big) = 1/3$ (esto es, el modelo uniforme). Sea X la variable aleatoria definida por $X(1) = 1$, $X(2) = X(3) = 0$. Sea Y la variable aleatoria definida por $Y(1) = Y(2) = 0$, $Y(3) = 1$. Entonces X y Y, a pesar de ser variables aleatorias distintas (como función, son distintas), tienen la misma distribución.

5.2. Motivación del concepto

Como hemos visto, un espacio de probabilidad $(\Omega, \mathscr{A}, \mathbb{P})$ y una variable aleatoria X, juntos dan lugar a un segundo espacio de probabilidad $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$. En general, el espacio muestral Ω puede ser altamente complejo, no sólo en cardinalidad de los posibles resultados, sino en su misma descripción (por ejemplo, ¿cuál es el Ω que describe las distintas formas en que puede suceder el estado del tiempo del día de mañana?).

Como vimos en en Capítulo 3, a menos que Ω sea de estructura relativamente muy sencilla (por ejemplo, finito o numerable), no hay forma

operativa para determinar una medida de probabilidad sobre subconjuntos de Ω. En este capítulo veremos que en cambio, es considerablemente más fácil especificar la probabilidad $\mathbb{P}_X$. Sobre el espacio $(\mathbb{R},\, \mathscr{B},\, \mathbb{P}_X)$ hay diversas herramientas que sirven para especificar distintas medidas de probabilidad $\mathbb{P}_X$. Esto puede ser útil cuando puedan formularse preguntas en términos de alguna variable aleatoria X, o cuando el espacio muestral coincida con $\mathbb{R}$. Entonces, aunque el espacio de probabilidad $(\Omega,\, \mathscr{A},\, \mathbb{P})$ exista como sustento axiomático de la teoría de probabilidad, es posible que los únicos eventos que sean de interés sean los que determina alguna variable aleatoria X, sin importar cuál sea la naturaleza exacta de $(\Omega,\, \mathscr{A},\, \mathbb{P})$. Es decir, en muchas ocasiones, el espacio de probabilidad que es realmente relevante para un problema dado, es el espacio $(\mathbb{R},\, \mathscr{B},\, \mathbb{P}_X)$.

Al adoptar como modelo $(\mathbb{R},\, \mathscr{B},\, \mathbb{P}_X)$ en lugar de $(\Omega,\, \mathscr{A},\, \mathbb{P})$, podremos sólo cuantificar probabilidades de eventos que se describan en términos de X, y no podremos nunca abordar eventos de los contenidos en $\mathscr{A}$ (note que la σ-álgebra $\mathscr{A}$ contiene eventos formulados en términos de elementos de Ω, mientras que la σ-álgebra $\mathscr{B}$ contiene eventos formulados en términos de una variable aleatoria X). En resumen, en muchísimas aplicaciones prácticas de teoría de probabilidad, los eventos de interés pueden parafrasearse en términos de alguna variable aleatoria X. Esto, aunado al hecho de que la especificación de medidas de probabilidad $\mathbb{P}_X$ sobre $\mathscr{B}$ es una tarea relativamente fácil para la cual existen herramientas convenientes, hace que el concepto de variable aleatoria adquiera un papel preponderante.

Ejemplo 5.10 El experimento es observar el estado del tiempo del día de mañana. Madre Naturaleza cuenta con $(\Omega,\, \mathscr{A},\, \mathbb{P})$. Es decir, al seleccionar

un estado del tiempo para el día de mañana, Madre Naturaleza elige un ω con acuerdo a su propio espacio de probabilidad. Supongamos que nosotros, como observadores humanos, por algún motivo nos interesa sólo si llueve o no. Para estudiar la lluvia, lo que es relevante es la variable aleatoria definida por $X(\omega) = 1$ si llueve en el estado ω, y 0 de otra manera. Para responder a la pregunta «¿Cuál es la probabilidad de que llueva mañana?», basta encontrar $\mathbb{P}(X = 1)$. Para Madre Naturaleza, quien opera detrás de bambalinas con un modelo de probabilidad, dicha probabilidad es evidentemente $\mathbb{P}\big(\{\omega \mid X(\omega) = 1\}\big)$, pero para nosotros los observadores, lo que es relevante es que dicha probabilidad es igual a $\mathbb{P}_X\big(\{1\}\big)$. Para contestar a la pregunta de interés, basta entonces especificar la medida de probabilidad $\mathbb{P}_X$, sin tener que conocer o siquiera concebir el modelo $(\Omega, \mathscr{A}, \mathbb{P})$. Es muy difícil observar ω, pero muy fácil observar $X(\omega)$. Notemos, sin embargo, que si un segundo observador preguntara «¿cuál es la probabilidad de que la temperatura máxima no llegue a 23.5?», entonces no podremos contestarle a eso en términos de $\mathbb{P}_X$, debido a que este segundo evento no puede describirse en términos de X.

5.3. Variables aleatorias discretas

Definición 5.4 (Densidad discreta) Sea $S = \{s_1, s_2, \dots\} \subset \mathbb{R}$ un subconjunto a lo más numerable. Una función $f\colon S \to \mathbb{R}$ se llama una *función de densidad discreta* (sobre S) si $f(s_i) \geq 0$, $\forall i$ y $\sum_{i=1}^{\infty} f(s_i) = 1$.

En la expresión anterior, si acaso S es finito, entonces el símbolo $\sum_{i=1}^{\infty}$ se interpreta como una suma finita, por lo que no es necesario invocar el con-

cepto de series. Note la similitud con las condiciones de los Teoremas 3.1 y 3.3, que hablan sobre la especificación de una medida de probabilidad sobre un espacio muestral finito o numerable. También compare los teoremas con la definición siguiente.

Definición 5.5 (V.A. discreta) Decimos que una v.a. X es *discreta*, si existe una función de densidad discreta f_X sobre algún conjunto $S = \{s_1, s_2, \dots\} \subset \mathbb{R}$, tal que

$$\mathbb{P}_X(B) = \sum_{i=1}^{\infty} 1_B(s_i)\, f_X(s_i),\ \forall B \in \mathscr{B}.$$

La función f_X se llama entonces la densidad de la variable aleatoria X.

Una observación importante a propósito de v.a.'s discretas, es que en particular, $\mathbb{P}_X(S) = 1$, lo cual muestra que las discretas «viven» sobre un conjunto a lo más numerable, $f_X(s_i)$ tiene la interpretación de ser igual a $\mathbb{P}_X\big(\{s_i\}\big) = \mathbb{P}(X = s_i)$, y por lo tanto, $0 \leq f_X(s_i) \leq 1$, $\forall i$. Es importante observar que la función $\mathbb{P}_X \colon \mathscr{B} \to \mathbb{R}$ definida como se indica, constituye una medida de probabilidad. Esto es, una función de densidad especifica una medida de probabilidad sobre $\mathscr{B}$. También la caracteriza, en el sentido de que no hay dos medidas de probabilidad sobre $\mathscr{B}$ que asignen los valores $f_X(s_i)$ a los conjuntos $\{s_i\} \in \mathscr{B}$. La demostración no la hacemos aquí, debido a que es muy similar a las demostraciones de los Teoremas 3.1 y 3.3.

Ejemplo 5.11 (Densidad binomial) Considere $S = \{0, 1, \dots, n\}$. Entonces la función f definida por

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

para $x \in S$, es una función de densidad discreta, llamada densidad *binomial*, para cualquier valor fijo de $p \in [0,1]$. Un ejemplo de una variable aleatoria discreta X que posee esta densidad, es el conteo del número de éxitos en un *experimento binomial* (descrito en la Sección 6.1).

Ejemplo 5.12 (Densidad geométrica) Considere $S = \{0, 1, \dots\}$. Entonces la función f definida por $f(x) = p(1-p)^x$ para $x \in S$, es una función de densidad discreta, llamada densidad *geométrica*, para cualquier valor fijo de $p \in (0,1]$. Un ejemplo de una variable aleatoria X discreta que posee esta densidad, es el conteo del número de fracasos observados antes de obtener el primer éxito, en una secuencia de experimentos Bernoulli.

El papel que está jugando una densidad discreta, es proveer un instrumento para *calcular* $\mathbb{P}_X$, a través de la relación $\mathbb{P}_X(B) = \sum_{i=1}^{\infty} 1_B(s_i)\, f_X(s_i)$. Esto es, si especificamos f_X, hemos especificado $\mathbb{P}_X(B)$, para todos los borelianos $B \in \mathscr{B}$. Más aun, note que en el Ejemplo 5.11, de hecho hemos podido especificar $\mathbb{P}_X(B)$ para todos los borelianos, mediante la especificación de *un solo* número, p (suponiendo que el número de ensayos, n se conoce). Este es un ejemplo de una densidad discreta parametrizada.

Existe una sub-clase de variables aleatorias discretas, motivada por situaciones prácticas en las que el fenómeno aleatorio en cuestión consiste de contar números de ocurrencia de diversos incidentes. Ejemplos de esta situación son el número de plantas con daños visibles producidos por una plaga; el número de individuos a favor de un partido político específico; el número de televisores con defectos en su selector de canales en un lote de 100 televisores recién fabricados; el número de clientes que se encuentran

en la fila en un centro de servicio al público, entre las 9 y 10 de la mañana.

Definición 5.6 (V.A. de conteo) Una variable aleatoria discreta que toma valores en $\{0, 1, 2, \dots\}$ recibe el nombre de *variable aleatoria de conteo*.

Notar que una variable aleatoria de conteo puede tener un rango finito (Ejemplo: Conteo de televisores defectuosos en un lote de 100), o puede tener un rango en principio indeterminado (Ejemplo: Número de clientes que acuden a un comercio entre 9 y 10 de la mañana). No hacemos distinción entre estas dos situaciones, e igualmente nos referimos a la variable aleatoria como una de conteo. Arriba, de manera involuntaria, hemos ya visto varios ejemplos de distribuciones para describir variables de conteo: Las distribuciones de Bernoulli, binomial, y geométrica corresponden todas a variables aleatorias de conteo. Estas distribuciones, así como otras que también son de conteo (binomial negativa, Poisson, hipergeométrica, series de potencias, y muchas otras) tienen gran aplicación en la práctica. Serán abordadas algunas con mayor detalle en el Capítulo 6.

Ejemplo 5.13 Si la variable aleatoria X tiene la distribución geométrica, y $Y = X/2,$ entonces X es variable aleatoria de conteo y Y es variable aleatoria discreta, aunque no es de conteo.

Cabe mencionar que en muchas de estas situaciones que dan origen a algún tipo de conteo, existen suposiciones básicas sobre el experimento físico que da lugar al proceso de conteo. Por ejemplo, se mencionó que la distribución binomial surge de contar el número de éxitos en un experi-

mento donde se han realizado n ensayos binarios, cada uno de los cuales tiene probabilidad p de ser éxito. Un detalle más fino es que dichos ensayos deberán ser probabilísticamente independientes entre sí. Esto último es un ejemplo de una suposición tácita que está presente cuando se emplea la distribución binomial para modelar un fenómeno aleatorio. En la práctica, estas suposiciones tácitas deberán establecerse por contexto o por otro tipo de validación.

Ejemplo 5.14 Consideremos las siguientes dos situaciones, que aparentemente corresponden ambas a un proceso físico de conteo de éxitos entre $n = 3$ ensayos.

Situación 1 Se eligen al azar tres individuos y se registra si son hombres o mujeres.

Situación 2 Se lanza una moneda tres veces y se registra si resulta águila o sol.

¿Son válidas las suposiciones de independencia y de misma p en estas situaciones? Un modo en que la suposición de misma p en la Situación 1 puede no ser viable, es si el primer individuo es elegido al azar en un estadio de futbol, el segundo en un club de golf, y el tercero en un salón de belleza. La Situación 2, en cambio es más compatible con la idea de que los ensayos son independientes y que cada uno de ellos tiene la misma p.

5.4. Variables aleatorias continuas

Definición 5.7 (Densidad continua) Una función $f\colon \mathbb{R} \to \mathbb{R}$ se llama una *función de densidad continua* si $f(x) \geq 0$, $\forall x \in \mathbb{R}$, y $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Ejemplo 5.15 (Densidad uniforme) Sea $f(x) = 1_{[0,1]}(x)$. Entonces f es una función de densidad continua. Recibe el nombre de densidad *uniforme* sobre $[\,0, 1\,]$.

Ejemplo 5.16 Sea $g(x) = 1_{(0,1)}(x)$. Entonces g es también una función de densidad continua. Notemos que es igual a f del ejemplo anterior, excepto por el hecho de haber cambiado su valor en dos puntos ($x = 0$ y $x = 1$).

Ejemplo 5.17 (Densidad normal estándar) Sea, para $-\infty < x < \infty$,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

Entonces f es una función de densidad continua. Recibe el nombre de densidad *normal estándar*. Veremos más adelante durante el curso, que esta densidad es de extrema importancia en la teoría de probabilidad y estadística, por diversos motivos.

Ejemplo 5.18 (Densidad exponencial) Sea, para $-\infty < x < \infty$, la función f definida por $f(x) = 1_{(0,\infty)}(x)\, \lambda e^{-\lambda x}$, para un número fijo $\lambda > 0$. Entonces f es una función de densidad continua. Recibe el nombre de densidad *exponencial*.

Note que una densidad continua no necesariamente es continua en el sentido de «función continua». Por ejemplo, la uniforme no es una función continua sobre $\mathbb{R}$, pero sí es una densidad continua; la densidad normal estándar como se anota, sí es una función continua sobre todo $\mathbb{R}$, además de ser una densidad continua.

Definición 5.8 (V.A. continua) Decimos que una v.a. X es *continua*,[1] si existe una función de densidad continua f_X tal que

$$\mathbb{P}_X(B) = \int_{-\infty}^{\infty} 1_B(x)\, f_X(x)\, dx = \int_B f_X(x)\, dx,\ \forall B \in \mathscr{B}.$$

La función f_X se llama entonces la densidad de la variable aleatoria X.

Una observación importante a propósito de variables aleatorias continuas, para contrastar con las discretas, es que nunca podremos encontrar un subconjunto S a lo más numerable, de números reales tales que $\mathbb{P}_X(S) = 1$. La razón es que si S es cualquier conjunto finito o numerable, digamos $S = \{s_1,\, s_2,\, \dots\}$, entonces se tiene que $\mathbb{P}_X(S) = \sum_{i=1}^{\infty} \mathbb{P}_X\big(\{s_i\}\big) = \sum_{i=1}^{\infty} \int_{\{s_i\}} f_X(x)\, dx = \sum_{i=1}^{\infty} 0 = 0$. Esto muestra que las continuas no pueden «vivir» sobre un conjunto a lo más numerable. Note que para una densidad continua, $f_X(x)$ no tiene la interpretación de ser igual a una probabilidad, y por lo tanto, no es necesariamente cierto que $0 \leq f_X(x) \leq 1$. Lo único que sí tiene interpretación probabilística, es el *área* bajo la función f_X sobre un conjunto B, pues esta área es precisamente $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$.

Considere las dos funciones de densidad f y g de los Ejemplos 5.15 y 5.16. Estas densidades estarían definiendo las medidas de probabilidad $\mathbb{P}(B) = \int_B f(x)\, dx$ y $\mathbb{Q}(B) = \int_B g(x)\, dx$, respectivamente. Notemos que por propiedades de integración, que entonces $\mathbb{P}(B) = \mathbb{Q}(B)$, $\forall B \in \mathscr{B}$. Esto es, si dos densidades continuas difieren en a lo más un conjunto numerable de

[1] Estrictamente hablando, el término correcto debería ser *variable aleatoria absolutamente continua con respecto a la medida de Lebesgue*. Sin embargo, para fines del presente curso, le llamaremos simplemente variable aleatoria continua. Existen variables aleatorias que no son ni discretas ni continuas, pero no las consideraremos en este curso. Es un hecho incuestionable el que una grandísima cantidad de situaciones en la práctica se pueden atacar con sólo estos dos tipos de variables aleatorias.

puntos, entonces las medidas de probabilidad a que dan lugar son exactamente las mismas. Lo anterior significa que no hablamos de «la» función de densidad de una variable aleatoria continua, sino de «una» de las posibles funciones de densidad que puede dar lugar a la medida de probabilidad $\mathbb{P}_X$. Es usual que cuando se especifican densidades continuas, que se eligen versiones continuas de las mismas (*continuas* en el sentido de funciones continuas).

Ejemplo 5.19 Si X es una variable aleatoria uniforme sobre $[\,0,1\,]$, y $a, b \in (0,1)$, $a < b$, entonces $\mathbb{P}_X\big((a,b)\big) = b - a$. Note que en esta situación, la probabilidad depende sólo de la *diferencia* entre a y b. Con esta interpretación, una característica distintiva de la densidad uniforme es que la probabilidad de cualquier intervalo contenido en $(0,1)$ de longitud fija tiene la misma probabilidad de ocurrir. No obstante la simplicidad matemática, esta característica de «homogeneidad de probabilidad» o «carencia de preferencia» es una cualidad que en muchos fenómenos aleatorios está presente. Ejemplos son ocurrencia de una descarga eléctrica a lo largo de un minuto de tiempo, el ángulo al cual es emitida una partícula radiactiva, o el ángulo al cual se detiene una ruleta de sorteo. La distribución uniforme es instrumental para la simulación en computadora de otras distribuciones más complejas. Así, es fundamental para la implementación de modernos métodos computacionales en estadística basados en simulación a gran escala en la computadora.

El papel que está jugando una densidad continua, es proveer un instrumento para *calcular* $\mathbb{P}_X$, a través de la relación $\mathbb{P}_X(B) = \int_B f_X(x)\,dx$. Esto es, si especificamos f_X, hemos especificado $\mathbb{P}_X(B)$ para todos los borelianos,

$B \in \mathscr{B}$. Note además, que en el Ejemplo 5.18, que hemos podido especificar $\mathbb{P}_X(B)$ para todos los borelianos, mediante la especificación de *un solo* número, λ. Este es un ejemplo de una densidad continua parametrizada.

5.5. Modelación de distribuciones de v.a.'s

En las secciones anteriores vimos cómo definir una medida de probabilidad $\mathbb{P}_X$ para dos tipos importantes de variables aleatorias, a través del uso de densidades de probabilidad f_X, sean éstas discretas o continuas. En resumen, una vez especificada la densidad, la medida de probabilidad sobre $\mathscr{B}$ queda completamente establecida. Notemos que las propiedades de la medida $\mathbb{P}_X$ se «heredan» de las propiedades de $\mathbb{P}$ (ver demostración del Teorema 5.2), y del hecho de que X fuera variable aleatoria. Esto es, si se parte de un espacio de probabilidad $(\Omega, \mathscr{A}, \mathbb{P})$ y una variable aleatoria X, se encuentra una medida de probabilidad $\mathbb{P}_X$ que hace que $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$ sea también espacio de probabilidad. ¿Qué pasa si proponemos una medida $\mathbb{P}_X$ sobre $\mathscr{B}$, sin conocer $\mathbb{P}$? ¿Será que hay un espacio de probabilidad $(\Omega, \mathscr{A}, \mathbb{P})$ y una variable aleatoria X, que juntos sean las capaces «heredar» las propiedades de probabilidad a $\mathbb{P}_X$? La respuesta es que sí. El resultado que lo establece se llama Teorema de existencia de Kolmogorov. Este se enuncia así:

Teorema 5.3 (Teorema de extensión de Kolmogorov) *Si $\mathbb{Q}$ es una medida de probabilidad sobre $\mathscr{B}$, entonces existe una variable aleatoria X y un espacio de probabilidad $(\Omega, \mathscr{A}, \mathbb{P})$ tal que $\mathbb{Q}$ es la distribución de X.*

La demostración de este resultado se pospone para cursos futuros de

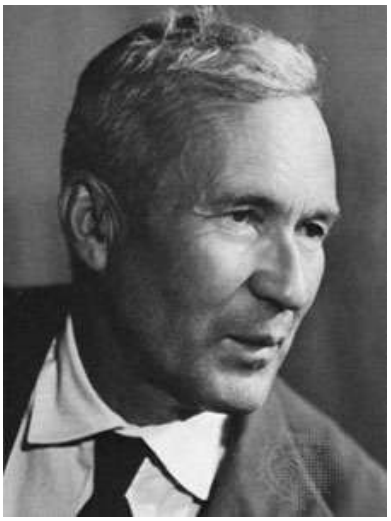

Figura 5.1: Andrei Nikolayevich Kolmogorov (1903–1987)
Fue un matemático ruso notable y laureado por varias contribuciones en topología, mecánica, análisis, y otros. En el contexto particular de teoría de probabilidad, se le reconoce por haber tendido el tratamiento axiomático de probabilidad. Foto: brittanica.com

teoría de probabilidad. Por el momento basta remarcar que la relevancia muy importante de este resultado es que uno puede tomarse la libertad de especificar $\mathbb{P}_X$ aunque uno no conozca una medida $\mathbb{P}$ explícitamente, y que la matemática (es decir, este teorema) nos proporciona «gratis» una estructura teórica que nos otorga licencia para emplear con legitimidad el espacio de probabilidad $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$. Es decir, uno puede recurrir a las propiedades matemáticas del espacio de probabilidad $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$ sin tener que hacer una pausa para determinar si hay o no una estructura que la relaciona con algún Ω, con algún $\mathscr{A}$, con alguna medida $\mathbb{P}$, y mucho menos, que requiera de la definición punto por punto de la función $X\colon \Omega \to \mathbb{R}$. Desde el punto de vista de modelación, recurrimos al Teorema de Kolmogorov al proceder como sigue. Supongamos, por ejemplo, que nos interesa estudiar la variable aleatoria $X =$ temperatura ambiental máxima de un día.

1. Suponemos que los valores de X que somos capaces de observar, son realizaciones de variables aleatorias, cuya distribución es $\mathbb{P}_X$.

2. Propondremos una distribución $\mathbb{P}_X$, para describir el comportamiento probabilístico de X. Por contexto, X ha de ser una variable continua, y por lo tanto, $\mathbb{P}_X$ puede determinarse a través de una densidad continua, f_X. Esta densidad ha de ser alguna que otorga probabilidad en rangos sensatos de temperatura digamos, 0–45 grados (esto es, una densidad que pone gran masa de probabilidad sobre temperaturas en el rango 50–370 no aparentaría ser en primera instancia una densidad sensata desde el punto de vista de modelación).

3. Kolmogorov nos dice que hay algún espacio de probabilidad $(\Omega, \mathscr{A}, \mathbb{P})$ y una variable aleatoria X tal que $\mathbb{P}_X$ es la distribución de X.

4. Supondremos entonces que la naturaleza obedece la estructura probabilística de $(\Omega, \mathscr{A}, \mathbb{P})$, aunque nunca detallemos explícitamente los pormenores de este espacio de probabilidad.

5. Proponemos un modelo estadístico, es decir un conjunto de posibles distribuciones, cuando $\mathbb{P}_X$ no se conoce del todo y la misión es inferirla por medio de observaciones empíricas del fenómeno aleatorio. Las observaciones empíricas de un fenómeno aleatorio, matemáticamente se representan como realizaciones de variables aleatorias $X_1, X_2, \ldots, X_n$, cada una de las cuales tiene distribución $\mathbb{P}_X$.

6. Usamos $\mathbb{P}_X$ para contestar preguntas acerca de probabilidades de eventos fraseados en términos de X, una vez que hayamos determinado cuál es $\mathbb{P}_X$.

En el resto de este capítulo, dando por hecho que para resolver problemas relacionados con variables aleatorias X, basta especificar $\mathbb{P}_X$, comenzaremos a adquirir un lenguaje para describir distribuciones de variables aleatorias X (note que una vez contando con conceptos para describir una medida de probabilidad, quizás se hubiera facilitado en gran manera el Ejercicio 5.12). Consideraremos sólo los casos en que X es discreta o continua, posponiendo la situación má s general para futuros cursos de probabilidad y teoría de medida.

5.6. Momentos

Los momentos constituyen un conjunto de características numéricas que se asocian a $\mathbb{P}_X$, la distribución de probabilidad de una variable aleatoria

X. Después de dar definiciones y ejemplos de su cálculo, pasaremos a interpretar su significado y a ilustrar su utilidad. Adelantamos que habrá cierta similitud entre los momentos, y la razón por la cual es relevante la siguiente frase, digna de una película de policías y ladrones: «... el sujeto medía 1.80m, pesaba alrededor de 75kg, y tenía aproximadamente 28 años de edad ...».

5.6.1. Definiciones

Definición 5.9 (Valor esperado) Sea X una variable aleatoria y $g\colon \mathbb{R} \to \mathbb{R}$ una función arbitraria.[2] El *valor esperado*, o *valor medio*, de $g(X)$ se define por

$$\mathbb{E}\big(g(X)\big) = \begin{cases} \sum_{i=1}^{\infty} g(s_i)\, f_X(s_i) & \text{si } X \text{ es v.a. discreta sobre } s_1, s_2, \dots \\ & \text{con densidad } f_X, \\ \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx & \text{si } X \text{ es v.a. continua con densidad } f_X. \end{cases}.$$

Cuando $g(x) = x$, se obtiene un caso de mucha trascendencia:

Definición 5.10 (Valor esperado de una v.a.) El *valor esperado de* X, o

[2] Estrictamente hablando, no es cualquier función, sino una que sea *medible*. Todas las funciones continuas son medibles, así como las que son continuas a pedazos. Otro resultado que se demuestra en teoría de medida es el siguiente: Si X es v.a. y $g\colon \mathbb{R} \to \mathbb{R}$ es $\mathscr{B} - \mathscr{B}$ medible, entonces la función $\Omega \to \mathbb{R}$ definida por $\omega \to g\big(X(\omega)\big)$ (es decir, la composición $g \circ X$) también es v.a. Luego, esta definición se trata en realidad del concepto de valor esperado de la variable aleatoria $g(X)$.

valor medio de X, o *esperanza de X*, se define por

$$\mathbb{E}(X) = \begin{cases} \sum_{i=1}^{\infty} s_i \, f_X(s_i) & \text{si } X \text{ es v.a. discreta sobre } s_1, s_2, \ldots \\ & \text{con densidad } f_X, \\ \int_{-\infty}^{\infty} x \, f_X(x) \, dx & \text{si } X \text{ es v.a. continua con densidad } f_X. \end{cases}$$

El valor esperado de X se denota alternativamente por μ_X, o por μ, si no hay peligro de ambigüedad.[3]

Luego veremos que la nomenclatura *valor esperado* es un tanto desafortunada,[4] pues no significa que *esperamos* que el valor de X sea μ. De hecho, el valor de μ es posible que sea tal que $\mathbb{P}(X = \mu) = 0$, por lo que sería absurdo «esperar» que la variable aleatoria X tome el valor μ.

Cuando $g(x) = x^n$ en lo anterior, se obtiene otro caso especial:

Definición 5.11 (Momentos de una v.a.) Para $k = 1, 2, \ldots$ la cantidad $\mathbb{E}(X^k)$ se llama el *k-ésimo momento de la v.a. X*. Se denota por μ_k.

Note que la media de X coincide con el primer momento, es decir, $\mu_X = \mu_1$.

[3] En teoría de medida, llegarán a ver que la definición general de valor esperado no es más que cierta integral, $\mathbb{E}(X) = \int_\Omega X(\omega) \, d\,\mathbb{P}(\omega)$, pero en este momento no poseemos herramienta suficiente para explicar lo que ella significa. En el caso de que X sea discreta o continua, resulta que esta integral es exactamente igual a lo que se indica en la presente definición, por lo que esto constituye a final de cuentas una forma de *calcular* dicha integral. De paso, mencionaremos que desde el punto de vista de teoría de medida, lo que aquí hemos denominado simplemente *densidades* de probabilidad (discretas y continuas), son realmente objetos con un nombre muy rimbombante: *derivadas de Radon-Nykodym con respecto a la medida de conteo y la medida de Lebesgue*.

[4] Proviene de la traducción del inglés, *expected value*, y el sentido en el que se usa este sustantivo, tiene más bien que ver quizás con la Desigualdad de Chebychev, que veremos un poco más adelante.

Cuando $g(x) = (x - \mu_1)^k$, se obtiene otro caso especial:

Definición 5.12 (Momentos centrales de una v.a.) Para $k = 1, 2, \ldots$ la cantidad

$$\mathbb{E}\big((X - \mu_1)^k\big)$$

se llama el *k-ésimo momento central de la v.a. X*. Se denota por μ'_k.

El primer momento central es igual a 0 (verifique), y además se puede notar que los momentos centrales siempre pueden escribirse en términos de los momentos (no centrales), haciendo uso de la propiedad de linealidad del valor esperado (Ejercicio 5.13). Por ejemplo,

$$\mu'_3 = \mathbb{E}\left((X - \mu_1)^3\right) = \mathbb{E}\left(X^3 - 3X^2\mu_1 + 3X(\mu_1)^2 - (\mu_1)^3\right) =$$
$$\mathbb{E}(X^3) - 3\mu_1\,\mathbb{E}(X^2) + 3(\mu_1)^2\,\mathbb{E}(X) - (\mu_1)^3 = \mu_3 - 3\mu_1\mu_2 + 3(\mu_1)^3 - (\mu_1)^3.$$

Definición 5.13 (Varianza de una v.a.) El segundo momento central, μ'_2, recibe el nombre de *varianza de X*, y se denota por σ^2_X, $\mathbb{V}(X)$, o por σ^2, si no hay peligro de ambigüedad.

Definición 5.14 (Desviación estándar de una v.a.) La raíz cuadrada positiva de la varianza, se llama la *desviación estándar*. Se denota por σ_X, $\mathrm{SD}(X)$, o por σ, si no hay peligro de ambigüedad.

Existen también algunas definiciones de otras cantidades que son funciones de los momentos anteriores. Dos ejemplos comunes son los siguientes.

Definición 5.15 (Coeficiente de simetría de una v.a.) Definimos el *coeficiente de simetría*[5] *de una v.a.*, denotado por γ, como μ'_3/σ^3, es decir

$$\gamma = \frac{\mathbb{E}\big((X-\mu)^3\big)}{\sigma^3}.$$

Definición 5.16 (Coeficiente de kurtosis de una v.a.) Definimos el *coeficiente de kurtosis*[6] *de una v.a.*, denotado por κ, como $\mu'_4/\sigma^4 - 3$, es decir

$$\kappa = \frac{\mathbb{E}\big((X-\mu)^4\big)}{\sigma^4} - 3.$$

5.6.2. Interpretaciones

μ_X, **media** La media de X es una descripción de la localización central de la distribución de probabilidad. Puede tomar cualquier valor real. Nos indica entonces una *posición* de la distribución de probabilidad.

σ^2, **varianza** La varianza de X es una descripción de la *dispersión* de la distribución de probabilidad. Toma valores no-negativos. Hay que notar que la varianza no es más que el valor esperado de $(X-\mu_X)^2$, y que entre más distancia tienda a haber entre X y μ_X, mayor será la varianza. Varianza cero es un caso extremo, que corresponde a la varianza de una variable aleatoria constante.

γ, **coeficiente de simetría** Toma cualquier valor real. Si una densidad es simétrica, entonces el valor del coeficiente es cero. Un valor negativo indica asimetría hacia la izquierda (valores mayores de probabilidad a la izquierda) y un valor positivo indica asimetría hacia la derecha

[5] En inglés, *skewness*.
[6] En inglés, *kurtosis*, o en algunas ocasiones *excess*.

(valores mayores de probabilidad a la derecha).

κ, **coeficiente de kurtosis** Toma cualquier valor real. Pretende ser indicativa de la forma del pico de una densidad alrededor de su centro. Un valor negativo indica una densidad aplanada en el centro, mientras que un valor positivo indica un pico más afilado. Un valor de cero indica que el pico es similar al de una distribución normal estándar.

En teoría de probabilidad puede mostrarse que la sucesión infinita de momentos $\mu_1,\ \mu_2,\ \mu_3,\ \ldots$, salvo en casos muy excepcionales, determina por completo la distribución de probabilidad de una v.a. Por esto los momentos son importantes instrumentos en desarrollos teóricos en probabilidad y estadística.[7] Dos variables aleatorias con distribuciones distintas, en general no tienen los mismos momentos. Pero si coinciden entre sí en los primeros momentos, digamos del primero al cuarto o quinto, estas dos distribuciones son cualitativamente bastante parecidas. Con esta interpretación, los momentos son características descriptivas de una distribución de probabilidad sobre $\mathbb{R}$. Con respecto a la analogía anunciada, relativa al escenario de policías y ladrones, la comparación es la siguiente.

- Para dar una descripción *gráfica* de un individuo, podemos recurrir a una fotografía. Para dar una descripción *gráfica* de una distribución de probabilidad, podemos recurrir a una gráfica de su función de densidad.

[7] No será tema de estudio del presente curso, pero aquí se puede mencionar de paso que otro de los instrumentos poderosos que existen en teoría de probabilidad para especificar una medida de probabilidad $\mathbb{P}_X$ (aparte del empleo de funciones de densidad, como hemos visto), es mediante el uso de cierta función llamada *función generadora de momentos* (fgm). La fgm proporciona un modo de especificar momentos, y por lo tanto, distribuciones de probabilidad. Un ejemplo: La fgm de la densidad exponencial (Ejemplo 5.18), es $m(t) = \lambda/(\lambda - t)$, y puede verificarse que el primer momento es $d\,m(t)/dt|_{t=0}$, que el segundo momento es $d^2\,m(t)/dt^2|_{t=0}$, y así sucesivamente.

- Aunque no tengamos una fotografía, podemos pretender *describir* a un individuo con otros métodos, porque podemos hacer alusión a una infinidad de características físicas: Su edad, su estatura, su complexión, su peso, el color de su piel, el color de su cabello, el tipo de cabello, la longitud de su cabello, el color de sus ojos, *etc*. Para *describir* una distribución de probabilidad, podemos hacer alusión a una infinidad de características: μ_1, μ_2, μ_3, $\ldots$

- Entre más grande sea la *lista de características físicas* a las que uno haga alusión acerca de un individuo, más precisa será la descripción del mismo. Podemos llegar hasta el detalle de longitud de las pestañas, circunferencia del dedo gordo del pie derecho, diámetro del iris del ojo izquierdo, longitud de la uña del dedo índice, y aun más, si se quisiera. Entre más grande sea la *lista de momentos* a las que uno haga alusión acerca de una distribución de probabilidad, más precisa será la descripción de la misma. Podemos llegar hasta el detalle de μ_{122}, μ_{123}, μ_{124}, y aun más, si se quisiera.

- Con sólo unas pocas descripciones físicas, *por ejemplo, edad, estatura, peso, y color de piel*, podemos llegar a una muy buena idea de cómo es un individuo. Prueba de ello es que estas son de las primeras características que se le preguntan a los testigos en la escena de un delito. Con sólo unos pocos momentos, *por ejemplo*, μ_X, σ^2, γ, *y* κ, podemos llegar a una muy buena idea de cómo es una distribución de probabilidad. Prueba de ello es que estos momentos son de los primeros que se reportan en los paquetes computacionales de estadística, y aun en muchas calculadoras de bolsillo.

- La longitud de las pestañas, circunferencia del dedo gordo del pie derecho, diámetro del iris del ojo izquierdo, y longitud de la uña del dedo índice, sin duda *me dicen algo acerca del individuo, pero no mucho*. Los momentos μ_{122}, μ_{123}, μ_{124}, sin duda *me dicen algo acerca de la distribución de probabilidad, pero no mucho*.

- ¿Para obtener una descripción general de un individuo, qué preferiría: Le proporcionaran peso y estatura, o le proporcionaran diámetro del iris y circunferencia del dedo gordo? ¿Para obtener una descripción general de una distribución de probabilidad, qué preferiría: Le proporcionaran μ_X y σ_X^2, o le proporcionaran μ_{872} y μ_{4125}?

- Una descripción de un individuo basada *solamente en su estatura*, es muy pobre, por no decir raquítica. Una descripción de una distribución de probabilidad basada *solamente en su valor esperado*, es muy pobre, por no decir raquítica.[8]

5.7. Algunas interpretaciones probabilísticas de algunos momentos

Al margen de las interpretaciones físicas mencionadas anteriormente para los momentos, existen en teoría de probabilidad numerosos resultados acerca de propiedades analíticas de los momentos. Mencionamos en esta sección algunos de los más importantes en cursos introductorios.

[8]Una muy, muy buena pregunta es entonces: ¿Por qué en la práctica la gente hace precisamente esto? Es decir, ¿por qué, cuando confrontados ante una variable aleatoria, oímos mucho decir en noticias, revistas, medios publicitarios, y conversaciones cotidianas, cosas tales como «el promedio de longevidad del mexicano es hoy 72 años» o «la temperatura máxima para mañana será de un promedio de 27 grados» o «la vida nominal de un foco incandescente de 40 Watts es 1000 horas»?

5.7.1. La Desigualdad de Chebychev

Un resultado que involucra a la media y a la desviación estándar de una variable aleatoria con una aseveración en términos de una probabilidad, es el siguiente. Note la generalidad al respecto del teorema, ya que no se involucran mayores detalles acerca de $\mathbb{P}_X$ excepto por los valores de su media y su desviación estándar. A final de cuentas es un resultado que habla de una propiedad muy general que tiene cualquier $\mathbb{P}_X$.

Teorema 5.4 (Desigualdad de Chebychev) *Si X es una variable aleatoria con media μ_X y desviación estándar $\sigma_X < \infty$, entonces*

$$\mathbb{P}\big(|X - \mu_X| \geq r\,\sigma_X\big) \leq \frac{1}{r^2},\ \forall r > 0.$$

Equivalentemente,

$$\mathbb{P}\big(|X - \mu_X| < r\,\sigma_X\big) \geq 1 - \frac{1}{r^2},\ \forall r > 0.$$

Demostración. Supongamos que X es una v.a. continua con densidad f_X.

Entonces

$$\sigma_X^2 = E\big(X-\mu_X\big)^2 = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,dx =$$
$$\int_{\{x|(x-\mu_X)^2\geq r^2\sigma_X^2\}} (x-\mu_X)^2 f_X(x)\,dx + \int_{\{x|(x-\mu_X)^2<r^2\sigma_X^2\}} (x-\mu_X)^2 f_X(x)\,dx \geq$$
$$\int_{\{x|(x-\mu_X)^2\geq r^2\sigma_X^2\}} (x-\mu_X)^2 f_X(x) \;\geq$$
$$\int_{\{x|(x-\mu_X)^2\geq r^2\sigma_X^2\}} r^2\sigma_X^2\, f_X(x)\,dx =$$
$$r^2\sigma_X^2\,\mathbb{P}\big((X-\mu_X)^2\geq r^2\sigma_X^2\big) = r^2\sigma_X^2\,\mathbb{P}\big(|X-\mu_X|\geq r\sigma_X\big).$$

Al dividir la desigualdad $\sigma_X^2 \geq r^2\sigma_X^2 P\big(|X-\mu_X|\geq r\sigma_X\big)$ entre $r^2\sigma_X^2$, se obtiene la Desigualdad de Chebychev. Si X es discreta la demostración es similar, reemplazando integrales por sumas finitas o series. □

Note que la interpretación que da este resultado es que la probabilidad de encontrar X lejos de μ_X es pequeña. Esto quizás dota de significado al empleo de la palabra «esperado» en el término *valor esperado*. En particular, tomando $r=2$, encontramos que $\mathbb{P}\big(|X-\mu_X|<2\,\sigma_X\big)\geq 1-1/2^2 = 3/4$, es decir, que para cualquier variable aleatoria, por lo menos 75 % de la masa de probabilidad de X se encuentra concentrada sobre el intervalo $\mu_X \pm 2\,\sigma_X$. De paso, también se confirma una interpretación de σ_X como medida de dispersión, pues si este valor crece, entonces el intervalo mencionado es de longitud mayor.

Ejemplo 5.20 (Longevidad de un foco incandescente) El empaque (Figura 5.2) de un foco de 20W afirma que la «Vida Útil Promedio» del mismo es de 8000 horas. Supongamos que esto se interpreta como un modo velado

de decir que siendo X la longevidad de un foco seleccionado al azar, que entonces 8000 es μ_X. Es decir, que el empaque no nos dice más que el primer momento de X. Sería poco realista que el fabricante —al concederle a los consumidores cierto conocimiento de teoría de probabilidad— incluyera en su empaque información para determinar la densidad f_X, lo cual sería sumamente interesante. La finalidad sería dotarnos de información para poder nosotros calcular, por ejemplo, la probabilidad de que el foco prescriba antes de alcanzar 5000 horas como $\mathbb{P}(X < 5000) = \int_0^{5000} f_X(x)\,dx$, la probabilidad de que dure más de 1500 horas como $\mathbb{P}(X > 1500) = \int_{1500}^{\infty} f_X(x)\,dx$, o bien cualquier otra probabilidad relacionada con X. Sin embargo, supongamos que el fabricante nos hubiera facilitado, si no la densidad f_X, por lo menos un momento de segundo orden, o que nos hubiera informado además que la desviación estándar es $\sigma_X = 240$ horas. Podemos imaginarnos que la leyenda en el empaque diría algo así como «Vida Útil Promedio: 8000 horas; Desviación Vida Útil Promedio: 240 horas». En tal caso, por lo menos podríamos obtener algo de información a través de la Desigualdad de Chebychev:

$$P\big(|X - 8000| < 3 \times 240\big) \geq 1 - 1/3^2,$$

es decir, $\mathbb{P}\big(7280 < X < 8720\big) \geq 0.8888$. Este ejemplo muestra que con sólo dos momentos, podemos comenzar a hacer alguna aseveración acerca de la distribución de X, cosa que no hubiera podido hacerse[9] si sólo contáramos

[9] Si nos atrevemos a suponer que la longevidad de un foco sigue la densidad exponencial (Ejemplo 5.18), y si nos dicen que $\mu_X = 8000$, entonces la densidad de hecho estaría dada por

$$f_X(x) = \mathbf{1}_{(0,\infty)}(x)\frac{1}{8000}e^{-x/8000}.$$

Esto a su vez determinaría que σ_X fuese en realidad 8000, en lugar del valor de 240 que se empleó hipotéticamente en el ejemplo. De paso, esta situación ilustra que el fabricante presuntamente habrá utilizado algún procedimiento de inferencia estadística para determinar el valor 8000 horas que publica en su empaque.

con μ_X.

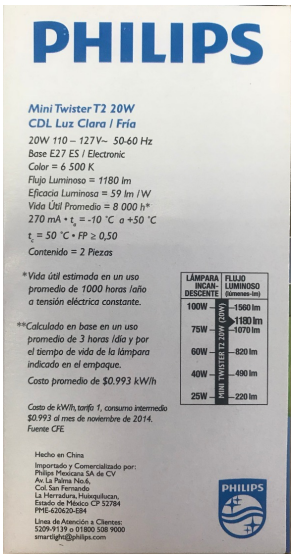

Figura 5.2: Longevidad de un foco
Datos típicamente presentados en el empaque de un foco. Es evidente que tanto el fabricante como el consumidor conciben que la longevidad es una variable aleatoria. Hay una apología basada en las condiciones bajo las cuales se llegó a la cifra de 8000 horas, por si acaso la duración resulta ser *menor*. Sin embargo, la información que provee acerca de la distribución de probabilidad de la longevidad no abarca más que el primer momento.

5.7.2. La Ley de los Grandes Números

El siguiente resultado es de mucha importancia en la teoría de probabilidad y estadística, y provee de una interpretación probabilística al valor esperado de una variable aleatoria. Es un ejemplo de un resultado límite en teoría de probabilidad, en el sentido de que se incorpora en la hipótesis una sucesión de variables aleatorias. Desde el punto de vista de la materia que nos ocupa —probabilidad y estadística— la consideración de sucesiones de

variables aleatorias adquiere relevancia porque las observaciones empíricas de un fenómeno aleatorio (es decir, los datos), se pueden concebir matemáticamente como una sucesión de v.a.'s. También es un primer resultado que establece una propiedad acerca de un *promedio* de variables aleatorias.

Antes debe formularse lo que significa el que dos o más variables aleatorias sean independientes, lo cual se fundamenta en el concepto de independencia mutua para eventos (ver Definición 4.3). Aunque la definición formal de independencia parece a primera vista ser muy aparatosa, el significado de que variables aleatorias sean independientes es muy simple: Que la ocurrencia de una o varias de ellas no afecta la probabilidad de ocurrencia de ninguna de las demás. Por ejemplo, si X y Y son dos variables aleatorias independientes, entonces $\mathbb{P}\big(X > 5 \mid Y \leq 4\big) = \mathbb{P}(X > 5)$, o más generalmente, $\mathbb{P}\big(X \in A \mid Y \in B\big) = \mathbb{P}(X \in A)$. La definición siguiente la incluimos sólo para ilustrar cómo se matematiza este concepto. Durante este curso, la independencia será más bien una suposición que atañe a la modelación, más que un concepto que tenga que demostrarse. Es decir, si el contexto de un problema indica que los valores que tomen algunas variables *no* influyen sobre las probabilidades de eventos relacionados con las demás, las variables son independientes «por decreto», y entonces por suposición de modelación, cumplirían la siguiente definición.

Definición 5.17 (Independencia entre variables aleatorias) Se dice que las variables aleatorias en una sucesión $X_1, X_2, \ldots$ son *independientes*, si para toda sucesión de números reales $x_1, x_2, \ldots$, los eventos $X_1 \leq x_1, X_2 \leq x_2, \ldots$ son independientes. Es decir, para toda colección finita de m índices

diferentes $j_1, j_2, \ldots, j_m$, se cumple

$$\mathbb{P}\big(X_{j_1} \leq x_{j_1}, X_{j_2} \leq x_{j_2}, \ldots, X_{j_m} \leq x_{j_m}\big) = \\ \mathbb{P}(X_{j_1} \leq x_{j_1})\ \mathbb{P}(X_{j_2} \leq x_{j_2}) \cdots \mathbb{P}(X_{j_m} \leq x_{j_m}).$$

Teorema 5.5 (Ley de los Grandes Números) *Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias independientes, cada una con la misma densidad*[10] *f_X. Suponga que para una función*[11] *$g\colon \mathbb{R} \to \mathbb{R}$, se cumple $\mathbb{V}\big(g(X)\big) < \infty$. Entonces, $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}\big(g(X)\big)\right| > \epsilon\right) = 0.$$

En lo anterior, el símbolo $\lim$ es como en cálculo (se trata del límite de una sucesión de probabilidades), y el símbolo $\mathbb{E}\big(g(X)\big)$ es como en la Definición 5.9. La condición $\mathbb{V}\big(g(X)\big) < \infty$ es una condición técnica que es requerida para poder demostrar el teorema. Veremos la demostración después de comentar el resultado; en ella aparecerá el uso de la Desigualdad de Chebychev. Lo importante por ahora es la interpretación del resultado: Lo que dice es que el promedio de variables aleatorias independientes converge al valor esperado, pues no importa cuán tan pequeña sea la constante ϵ, la probabilidad se va a cero. Se utiliza la siguiente notación abreviada para

[10] A una sucesión $X_1, X_2, \ldots$ de variables aleatorias independientes e idénticamente distribuidas se le llama sucesión *i.i.d.* (también en inglés, de *independent and identically distributed*).

[11] Medible.

denotar la convergencia señalada en la LGN:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{\mathrm{P}} \mathbb{E}\big(g(X)\big), \tag{5.2}$$

concepto que en teoría de probabilidad se llama *convergencia en probabilidad*. Notemos que, con acuerdo en la notación que hemos establecido, que $\big|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}\big(g(X)\big)\big| > \epsilon$ es el evento

$$\left\{\omega \in \Omega \mid \Big|\frac{1}{n}\sum_{i=1}^{n} g\big(X_i(\omega)\big) - \mathbb{E}\big(g(X)\big)\Big| > \epsilon\right\}.$$

Demostración. El valor esperado de $n^{-1}\sum_{i=1}^{n} g(X_i)$ es $\mathbb{E}\big(g(X)\big)$ y su varianza es $\mathbb{V}\big(g(X)\big)/n$ (Ejercicio 5.20). Sea $\epsilon > 0$. Tomando $r = \epsilon\sqrt{n}/\sqrt{\mathbb{V}\big(g(X)\big)}$ en la Desigualdad de Chebychev (Teorema 5.4), obtenemos

$$\begin{aligned}\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}\big(g(X)\big)\Big| > \epsilon\Big) = \\ \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}\big(g(X)\big)\Big| > \frac{\epsilon\sqrt{n}}{\sqrt{\mathbb{V}\big(g(X)\big)}}\sqrt{\frac{\mathbb{V}\big(g(X)\big)}{n}}\Big) \leq \\ \left(\frac{\epsilon\sqrt{n}}{\sqrt{\mathbb{V}\big(g(X)\big)}}\right)^{-2} = \frac{\mathbb{V}\big(g(X)\big)}{\epsilon^2 n},\end{aligned}$$

lo cual converge a cero cuando $n \to \infty$ irrespectivamente de los valores de ϵ y $\mathbb{V}\big(g(X)\big)$. □

Tomando varios casos particulares para g, se obtiene la siguiente colección de ejemplos. En todos ellos, supondremos que se trata de una sucesión $X_1, X_2, \ldots$ de variables aleatorias independientes, cada una con la misma densidad f_X.

Ejemplo 5.21 Si $g(x) = x$, la LGN dice que

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{P}} \mu_X.$$

Notemos que este resultado dota de inmediato de otra interpretación al valor esperado $\mu_X = \mathbb{E}(X)$, pues éste coincide con el valor al cual converge un promedio infinito de variables aleatorias.

Ejemplo 5.22 Más generalmente, si $g(x) = x^k$, la LGN dice que

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{\text{P}} \mathbb{E}\left(X^k\right).$$

Ejemplo 5.23 Si A es un conjunto boreliano, y si $g(x) = 1_A(x)$, entonces la LGN dice que

$$\frac{1}{n}\sum_{i=1}^{n} 1_A\left(X_i\right) \xrightarrow{\text{P}} \mathbb{E}\big(1_A(X)\big).$$

Más aun, notemos que (si f_X es densidad continua)

$$\mathbb{E}\big(1_A(X)\big) = \int_{-\infty}^{\infty} 1_A(x)\, f_X(x)\, dx = \mathbb{P}_X(A) = \mathbb{P}(X \in A)$$

(ver Definición 5.8). Por otra parte,

$$\frac{1}{n}\sum_{i=1}^{n} 1_A(X_i) = \frac{\text{número de veces que } X_i \in A}{n}.$$

Es decir, lo que LGN está diciendo en realidad es

$$\frac{\text{número de veces que } X_i \in A}{n} \xrightarrow{\text{P}} \mathbb{P}(X \in A),$$

lo cual no es otra cosa más que la definición clásica o frecuentista de probabilidad.

Ejemplo 5.24 Como caso particular del anterior, si se toma $g(x) = 1_{(-\infty,y]}(x)$, entonces obtenemos

$$\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,y]}(X_i) \xrightarrow{\text{P}} \mathbb{P}\big(X \leq y\big).$$

Definición 5.18 (Función de distribución) En teoría de probabilidad, a $\mathbb{P}(X \leq y)$ visto como función de y, se le llama *función de distribución acumulada* (fda), o simplemente *función de distribución*, de la variable aleatoria X. Se le denota por $F_X(y)$. Note que si X tiene densidad f_X, entonces $F_X(y) = \int_{-\infty}^{y} f_X(x)\,dx$ en el caso continuo[12] y $F_X(y) = \sum_{x\leq y} f_X(x)$ en el caso discreto. Correspondientemente, a la cantidad aleatoria $\frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,y]}(X_i)$ se le llama *función de distribución empírica (fde)*. Note que la fde no es más que la proporción de veces que un dato no excede el valor y.

Note que una cosa es la *distribución* de X (denotada por $\mathbb{P}_X$, una medida de probabilidad sobre $\mathscr{B}$) y otra cosa distinta es la *función de distribución* de X (una función de $\mathbb{R}$ a $\mathbb{R}$). La relación que hay entre ambos conceptos es $\mathbb{P}_X\big((-\infty, x\,]\big) = F_X(x)$, $\forall x \in \mathbb{R}$.

Ejemplo 5.25 Considere el experimento de lanzamiento de tres monedas, y sea X el número total de águilas observadas. La fda de la variable aleatoria

[12] Note además, que en este caso, debido al teorema fundamental del cálculo, se cumple $f_X(x) = \frac{d}{dx}F_X(x)$.

X es:

$$F_X(x) = \begin{cases} 0 & \text{si } -\infty < x < 0, \\ 1/8 & \text{si } 0 \leq x < 1, \\ 1/2 & \text{si } 1 \leq x < 2, \\ 7/8 & \text{si } 2 \leq x < 3, \\ 1 & \text{si } 3 \leq x < \infty. \end{cases}.$$

En general, si X es una variable aleatoria discreta, entonces la fda será una función escalonada no decreciente con saltos en x de magnitud $\mathbb{P}(X = x)$.

Ejemplo 5.26 Si X tiene densidad exponencial $f_X(x) = \lambda e^{-\lambda x}$, entonces $F_X(y) = 1 - e^{-\lambda y}$ (use cálculo para verificar).

Ejemplo 5.27 Si X es una variable aleatoria discreta sobre S, con densidad f_X, tomando $g(x) = 1_{\{s\}}(x)$ para $s \in S$, la LGN nos dice

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{s\}}(X_i) \xrightarrow{\text{P}} \mathbb{P}\big(X = s\big) = f_X(s).$$

Ejemplo 5.28 Si X es una variable aleatoria continua, y $a, b \in \mathbb{R}$ son tales que $a < b$, entonces la LGN nos dice

$$\frac{1}{n}\sum_{i=1}^{n} 1_{(a,b)}(X_i) \xrightarrow{\text{P}} \mathbb{P}\big(X \in (a,b)\big) = \int_a^b f_X(x)\,dx.$$

Más adelante durante el curso, al tocar el tema de estadística descriptiva, veremos que $\frac{1}{n}\sum_{i=1}^{n} 1_{\{s\}}(X_i)$ y $\frac{1}{n}\sum_{i=1}^{n} 1_{(a,b)}(X_i)$ se llaman *histogramas*, y que la graficación de estos objetos como función de s o de (a, b), proporciona un instrumento útil para determinar la naturaleza de la distribución de una

variable aleatoria X.

Vale la pena tocar un ejemplo donde la conclusión de la LGN no se cumple, a manera de entender la relevancia de las suposiciones que intervienen. La observación importante que debe hacerse es que no *cualquier* promedio de variables aleatorias converge al valor esperado de ellas:

Ejemplo 5.29 Considere X_1 una variable aleatoria con distribución uniforme en $(0, 1)$. Ahora tómese $X_2 = X_1$, $X_3 = X_1$, $X_4 = X_1$, *etc*. Es claro que todas las variables tienen la misma distribución uniforme sobre $(0, 1)$, con media $1/2$, y con varianza $1/12 < \infty$, pero *no* son independientes (¿por qué?). Entonces se obtendría $\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(nX_1) = X_1$. Tomemos $\epsilon = 1/4$, usemos la densidad uniforme para calcular una probabilidad, y notaremos entonces que $\mathbb{P}\Big(\big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\big| > \epsilon\Big) = \mathbb{P}\big(|X_1 - 1/2| > 1/4\big) = 1/2$, y que esta probabilidad, siendo constante, *no* converge a 0.

Ejemplo 5.30 (Integración de Monte Carlo) La LGN justifica cierta técnica típicamente vista en métodos numéricos, para integrar una función. Un ejemplo sencillo es el siguiente. Sea $g\colon (0, 1) \to \mathbb{R}$ una función integrable dada, y supongamos que interesa calcular numéricamente el valor de $\int_0^1 g(x)\,dx$. Sea X una v.a. con densidad f_X que es uniforme en $(0, 1)$. Entonces $\mathbb{E}(g(X)) = \int_0^1 g(x)\, f_X(x)\,dx = \int_0^1 g(x)\,dx$. Con ello, si en la computadora se genera una sucesión $X_1, X_2, \ldots, X_n$ i.i.d. de uniformes, la LGN dice que $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$ resultará ser aproximadamente igual a $\int_0^1 g(x)\,dx$.

5.8. Momentos muestrales

Definición 5.19 (Momentos muestrales) Si $X_1, X_2, \ldots, X_n$ es una colección de variables aleatorias, y $k \in \mathbb{N}$, a la variable aleatoria $\frac{1}{n}\sum_{i=1}^{n} X_i^k$ se le llama *k-ésimo momento muestral*. Se denota por $\hat{\mu}_k$. La variable aleatoria $\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_1)^k$ se llama k-ésimo momento central muestral, y se denota por $\hat{\mu}'_k$.

Notemos que un momento μ_k es un número (igual a $\mathbb{E}(X^k)$), mientras que un momento muestral, $\hat{\mu}_k$, es una variable aleatoria. Notemos además que lo que dice la LGN (Ejemplo 5.22) es entonces, que si $X_1, X_2, \ldots, X_n$ son independientes y todos con distribución $\mathbb{P}_X$, que los momentos muestrales convergen en probabilidad a los momentos, es decir,

$$\hat{\mu}_k \xrightarrow{\mathrm{P}} \mu_k, \, \forall k \in \mathbb{N}. \tag{5.3}$$

Otra observación importante, especialmente para fines de estadística matemática, es que mientras que los momentos μ_k sí dependen de $\mathbb{P}_X$, los momentos muestrales *no* dependen de $\mathbb{P}_X$. Dependen sólo de los datos, $X_1, X_2, \ldots, X_n$, y una vez observados los datos se puede calcular $\hat{\mu}_k$.

Definición 5.20 (Media muestral) El primer momento muestral, recibe el nombre de *media muestral*. Se le denota indistintamente por $\hat{\mu}_1$, por $\hat{\mu}_X$ o $\bar{X}$. Note que el primer momento muestral no es otra cosa que la llamada *media aritmética* de los datos $X_1, X_2, \ldots, X_n$.

Definición 5.21 (Varianza muestral) El segundo momento central mues-

tral recibe el nombre de *varianza muestral*, denotada por $\hat{\sigma}^2$, o en ocasiones por s^2, o por S^2. Es decir,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Definición 5.22 (Desviación estándar muestral) La *desviación estándar muestral*, denotada por $\hat{\sigma}$, o en ocasiones por s, o por S, se define como la raíz cuadrada positiva de la varianza muestral.

5.9. Función generadora de momentos

Si en la Definición 5.9 se toma otro caso particular para la función $g(x)$ se obtiene una noción de suma importancia y utilidad en teoría de probabilidad. En efecto, considere $g(x) = \exp(tx)$, para $t \in \mathbb{R}$.

Definición 5.23 (Función generadora de momentos) La función $m_X : \mathbb{R} \to \mathbb{R}$ definida por $m_X(t) = \mathbb{E}\big(\exp(tX)\big)$ recibe el nombre de *función generadora de momentos* de la variable aleatoria X.

Ejercicios

5.1 Sea X una variable aleatoria discreta con densidad dada por $f_X(0) = 1/4$, $f_X(1) = 1/2$, $f_X(2) = 1/4$. Encuentre μ, σ^2, σ, μ_2, y μ_3'.

5.2 Sea X una variable aleatoria continua con densidad dada por $f_X(x) = 1_{(0,1)}(x)$. Encuentre μ, σ^2, σ, μ_2, y μ_3'.

5.3 Si $X_1, \ldots, X_n$ es una muestra i.i.d. con densidad $f_X(x)$, explica si cada una de las siguientes aseveraciones es verdadera o falsa:

(a) El k-ésimo momento de X_1 tiene valor esperado constante dado por $\int_{-\infty}^{\infty} u^k f_X(u)\, du$.

(b) El segundo momento central de X_1 es igual a

$$\int_{-\infty}^{\infty} \left(x - \int_{-\infty}^{\infty} u\, f_X(u)\, du \right)^2 dx.$$

(c) $\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2) - \mathbb{V}(X_1 X_2)$.

(d) Si el primer momento de X_1 es 0, entonces el segundo momento de X_1 coincide con el segundo momento central.

(e) $\mathbb{E}(X_1 + X_2 X_2) = \int_{-\infty}^{\infty} u f_X(u)\, du \left\{ 1 + \int_{-\infty}^{\infty} u f_X(u)\, du \right\}$.

(f) $\mathbb{V}(\prod_{i=1}^{n} X_i) = [\mathbb{V}(X_1)]^n$.

(g) $\mathbb{V}[\mathbb{E}(X_1)] = \mathbb{E}[\mathbb{V}(X_1)]$.

5.4 Considere la variable aleatoria X con distribución uniforme sobre $(0, 1)$. Verifique explícitamente que la Desigualdad de Chebychev para $r = 3$ se cumple, mediante el cálculo de $\mathbb{P}\big(|X - \mu_X| \geq 3\,\sigma_X\big)$ y su comparación con la cota $1/3^2$. Note que como parte del ejercicio, antes deben calcularse los valores de μ_X y σ_X.

5.5 Considere la variable aleatoria X con distribución de Poisson con parámetro $\lambda = 1$, es decir, con densidad dada por $f_X(x) = e^{-1}/x!$ para $x = 0, 1, \ldots$. Verifique explícitamente que la Desigualdad de Chebychev

para $r = 4$ se cumple, mediante el cálculo de $\mathbb{P}\big(|X - \mu_X| \geq 4\,\sigma_X\big)$ y su comparación con la cota $1/4^2$. Note que como parte del ejercicio, antes deben calcularse los valores de μ_X y σ_X.

5.6 Verifique que la función de distribución de una variable aleatoria está dada por $F_X(x) = \begin{cases} \sum_{t \leq x} f\,(t) & \text{si } X \text{ es variable aleatoria discreta,} \\ \int_{-\infty}^{x} f\,(t)\;dt & \text{si } X \text{ es variable aleatoria continua.} \end{cases}$

5.7 Muestre que la varianza de una v.a. puede calcularse como $\mu_2 - \mu_1^2$, es decir, $\mathbb{V}(X) = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2$.

5.8 Muestre que $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2$. Compare con el Ejercicio 5.7.

5.9 Demuestre, usando la LGN, que la media muestral converge a la media, y que la varianza muestral converge a la varianza.

5.10 (Integración de Monte Carlo) Todos los lenguajes de programación de alto nivel, por ejemplo C, incluyen una función que proporciona números aleatorios. La función típicamente se llama `RND`, `RANDOM`, `runif`, *etc.*, y de ordinario generan v.a.'s independientes con distribución uniforme sobre $(0, 1)$. Adopte su función integrable favorita sobre $(0, 1)$, para la cual sí pueda calcular teóricamente la integral, por ejemplo $g(x) = e^x$, o $g(x) = \sin(x)$. Escriba un programa de computadora que genere n números aleatorios, y que calcule $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$, para verificar entonces que este valor se acerca a $\int_0^1 g(x)\,dx$. La máxima funcionalidad de un mé todo numérico se aprecia cuando la integral es difícil de obtener teóricamente, por ejemplo cuando $g(x) = x^2 e^{\sin(x)} \log(x + 1)$.

5.11 *Verificación empírica de la LGN, y de la Desigualdad de Chebychev, usando un programa de cómputo*.

(a) Escriba un programa que genere n números aleatorios con distribución uniforme en $(0, 1)$, y que calcule el promedio muestral y la varianza muestral. Verifique, ejecutando varias veces el programa, que aunque en cada simulación se generan distintos números aleatorios (cuide que cada vez que se corra, que varíe la semilla del generador de números aleatorios), que siempre sucederá que la media muestral se acerca a $1/2$ y que la varianza muestral se acerca a $1/12$. Ejecute el programa para diversos valores ascendentes de n, y observe lo que ocurre a medida que estos aumentan.

(b) Haga lo mismo que en el inciso (a), con el siguiente cambio: En lugar de generar uniformes en $(0, 1)$, genere variables con distribución normal estándar (suponiendo que el lenguaje de programación que usted maneja también tiene un generador de normales). Verifique que la media muestral se acerca a 0 y que la varianza muestral se acerca a 1.

(c) Genere n números aleatorios con distribución uniforme sobre $(0, 1)$, y calcule la proporción de veces que dichos números se alejan del valor $1/2$ en una distancia mayor que $1/\sqrt{12}$. Compare con la cota predicha por la Desigualdad de Chebychev.

5.12 Describa de manera cualitativa, el tipo de densidad de probabilidad que podría proponerse para las siguientes situaciones (diga por ejemplo, si es continua o discreta, cuál sería su soporte, su forma genérica, *etc*.):

(a) el tiempo que tarda un cliente en pasar por una caja en una tienda de autoservicio.

(b) el número de automóviles que pasan por una caseta de cobro entre las 9 y 10 de la mañana.

(c) el número de veces que enciende un foco antes de fundirse.

(d) el tiempo que dura un foco encendido antes de fundirse.

5.13 Muestre que el operador valor esperado es *lineal*, es decir, que $\mathbb{E}(a + bX) = a + b\,\mathbb{E}(X)$, para constantes $a, b \in \mathbb{R}$.

5.14 Muestre que el valor esperado de una v.a. constante, es la misma constante (tome $g(x) = c$). Muestre que la varianza de una v.a. constante es cero.

5.15 Considere un modelo de probabilidad donde $\Omega = \{a, b, c, d, e\}$, con $\mathbb{P}\big(\{a\}\big) = \mathbb{P}\big(\{b\}\big) = \mathbb{P}\big(\{c\}\big) = 0.1$, $\mathbb{P}\big(\{d\}\big) = 0.4$, y $\mathbb{P}\big(\{e\}\big) = 0.3$. Defina la variable aleatoria X como $X(a) = 0$, $X(b) = 2$, $X(c) = 3$, $X(d) = 4$, y $X(e) = 5$.

(a) Encuentre y grafique la función de densidad de X, $f_X(x)$.

(b) Calcule el valor esperado de X.

(c) Calcule la varianza de X.

(d) Calcule el tercer momento de X.

5.16 Considere la función

$$f(x) = \frac{1}{2}x\,\mathbf{1}_{(0,1]}(x) + \frac{1}{2}\,\mathbf{1}_{(1,2]}(x) - \frac{1}{2}(x-3)\,\mathbf{1}_{(2,3]}(x).$$

(a) Verifique que $f(x)$ es una función de densidad para una variable aleatoria continua, X.

(b) Calcule $\mathbb{P}\big(X < \frac{1}{2}\big)$.

(c) Calcule $\mathbb{P}\big(X > \frac{3}{2}\big)$.

(d) Calcule $\mathbb{P}\big(1 < X < 2\big)$.

5.17 Suponga que se sabe que 10 % de una cierta población de individuos es zurda. Se examinará una muestra independiente de 50 individuos de esta población. Sea X el número de zurdos en esta muestra.

(a) Diga cuál es la distribución de X.

(b) Encuentre los valores de $\mathbb{E}(X)$ y $\mathbb{V}(X)$.

(c) Calcule la probabilidad de encontrar en la muestra a lo más 2 zurdos.

5.18 Opine sobre si cada una de las situaciones referidas en la Tabla 5.1 corresponden a una variable aleatoria discreta, continua, o ninguna de las dos.

5.19 Encuentre la falacia en la siguiente "demostración" de que la varianza de X es cero:

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 = \mathbb{E}(XX) - \big(\mathbb{E}(X)\big)^2 = \\ \mathbb{E}(X)\,\mathbb{E}(X) - \big(\mathbb{E}(X)\big)^2 = 0.$$

5.20 Si $X_1, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas con media μ y varianza σ^2, y $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, demuestre que $\mathbb{E}(\bar{X}_n) = \mu$ y $\mathbb{V}(\bar{X}_n) = \sigma^2/n$.

Capítulo 6

Familias de distribuciones y modelos estadísticos

En este capítulo haremos un breve recuento de algunas distribuciones de probabilidad, que son importantes en un curso introductorio tal como el que nos ocupa. Cabe mencionar que existen tomos enteros[1] que versan sobre distribuciones de probabilidad, y que estos pueden llegar a un grado de especialización muy, muy específico. Por ejemplo, distribuciones discretas de conteo, distribuciones multivariadas, distribuciones para extremos, distribuciones de «colas pesadas», distribuciones que surgen como límites de promedios de variables aleatorias, y distribuciones discretas, por mencionar algunas. El grado de especialización es tal, que también existen tomos enteros (libros[2] acerca de *una sola* distribución. La razón por la que es posible llegar a tal grado de especialización, es que en torno a una

[1] Algunos ejemplos disponibles en la biblioteca son Patil et al. (1984), Johnson et al. (1994), Johnson et al. (1992), y Fang et al. (2017).

[2] Ejemplos disponibles en la biblioteca son Tong (1990), Consul (1989), Bowman & Shenton (1988), y Chhikara & Folks (1989).

sola distribución de probabilidad, existen muchísimas cuestiones relativas a la misma. Por ejemplo, su *génesis* (es decir, la justificación de dónde y por qué surge como distribución de probabilidad), sus propiedades matemáticas (incluyendo momentos, aproximaciones, relaciones con otras distribuciones, funciones generadoras de momentos, *etc*.), extensiones a vectores aleatorios, métodos de simulación numérica, estimación de parámetros, *etc*. En el WWW existen páginas dedicadas a las distribuciones de probabilidad. Explore, por ejemplo, `http://www.anesi.com/poisson.htm` o `http://www.stat.berkeley.edu/users/stark/Java/ProbCalc.htm`.

Hemos visto que para determinar una distribución de probabilidad $\mathbb{P}_X$ para una variable aleatoria X, uno de los recursos que existen es el empleo de funciones de densidad —sean de tipo discreto o continuo—. Naturalmente, entonces será usual que para describir una la distribución $\mathbb{P}_X$ de una variable aleatoria X, que se anote la forma que tiene la densidad de probabilidad. Es importante recordar que lo que está permitiendo a final de cuentas una densidad de probabilidad, es el cálculo de $\mathbb{P}_X(B)$ para cualquier conjunto de Borel, B, a manera de hacer viable la utilización del espacio de probabilidad $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$.

Es posible, como en todos los casos que se abarcan en este capítulo, que se hable de «una» densidad cuando en realidad se trata de varias. Por ejemplo de «la densidad Poisson» cuando en realidad de lo que se trata es de una *familia* de densidades, dada por $e^{-\lambda}\lambda^x/x!$. Esto es, hay una densidad de Poisson para cada valor $\lambda > 0$. Correspondientemente, para cada valor de λ, hay una $\mathbb{P}_X$ distinta. Hablamos entonces de *familias paramétricas*. El parámetro puede ser unidimensional, como en la Poisson, bidimensional, como en la familia normal, o de mayor dimensión aún.

En lo que sigue, el formato que se procurará seguir es: La definición de

la densidad, la indicación de los valores posibles de los parámetros, notación específica para la familia, la media y la varianza de la distribución, y comentarios generales acerca de situaciones típicas en las que surge y se utiliza cada distribución.

Notación 6.1 Cuando una variable aleatoria X tenga una distribución $\mathbb{P}_X$ definida a través de una función de densidad f_X, escribimos $X \sim f_X$, que se lee «X tiene densidad f_X» o en corto, «X es f_X». Más aún, si la densidad f_X pertenece a alguna familia paramétrica como las que veremos a continuación, es posible que exista una notación específica para denotar a la familia. Por ejemplo, la familia Poisson tiene densidades dadas por $e^{-\lambda}\lambda^x/x!$, y en tal caso escribiremos $X \sim \text{Poisson}(\lambda)$, o $X \sim \text{Poisson}(4)$, si se trata de una variable aleatoria Poisson específica (aquella que corresponde a $\lambda = 4$). También se puede emplear $X \sim \mathbb{P}_X$ directamente, no habiendo ambigüedad porque una densidad f_X determina $\mathbb{P}_X$ y viceversa. Por la misma razón, también podemos emplear $X \sim F_X$, donde F_X es una función de distribución, o incluso $X \sim m_X$, donde m_X es una función generatriz de momentos.

La razón por la que se estudian de manera específica las distribuciones de probabilidad, es que uno desea armarse de un arsenal de modelos probabilísticos para aplicar a una situación surgida en la práctica. En el arte de modelación de fenómenos aleatorios, es necesario estar familiarizados con una gran variedad de modelos de probabilidad, con el objeto de poder proponer modelos matemáticos sensatos que describan la aleatoriedad de un fenómeno. Para poder postular un modelo, es naturalmente necesario conocer las propiedades teórico-matemáticas de las distribuciones. Por otra

parte, es importante señalar que por muy grande que sea el arsenal de distribuciones que conozcamos a fondo, nunca podremos agotar la totalidad de las distribuciones de probabilidad posibles, por lo que es necesario contar también con conocimiento matemático sólido para profundizar en alguna distribución previamente desconocida para nosotros, así como para verificar si un modelo de probabilidad parece ser sustentado o no por datos en una situación específica. Por ello, las bases que se obtienen en el estudio de teoría de probabilidad son también relevantes para las aplicaciones. En la Sección 6.4 veremos algunas herramientas —que son ciencia, en el sentido de ser resultados matemáticos— auxiliares en el arte de modelación matemática de fenómenos aleatorios.[3]

[3]Lo que se dice en el párrafo anterior, tiene una analogía muy directa con el arte de la aplicación de la medicina, y la formación profesional que recibe un médico durante su carrera. El médico aprende, en las aulas de una facultad de medicina, la etiología (causas de las enfermedades) y la prognosis (la evolución de los síntomas) de una gran cantidad de enfermedades. Ante un paciente con ciertos síntomas, el médico hace una evaluación y propone un cuadro que explica los síntomas, lo cual se llama el *arte* de realizar el diagnóstico de la enfermedad. Una vez hecho el diagnóstico correcto, el tratamiento es otra cuestión, totalmente diferente, y más parecida a ser ciencia (aunque puede haber una componente de arte en esto también). Es bien posible que los síntomas no los conozca previamente el médico. Para ello, el médico también recibe instrucción en fisiología, en anatomía, en histología, en patología, en bioquímica y otras áreas de conocimiento médico-teórico que le permiten relacionar los síntomas con otras enfermedades. Para ser un médico competente, el médico debe ser capaz de cultivar un arte, con el auxilio de la ciencia (conocimientos teóricos, tecnología para análisis clínicos, imágenes computarizadas, *etc*.). El diagnóstico es definitivamente un arte. Si fuera ciencia, no existirían los médicos, porque existiría en su lugar un software automatizado al que uno pudiera alimentar síntomas para obtener un diagnóstico, y por ende, un tratamiento. La modelación matemática también es un arte. Si fuera ciencia, existiría un gran teorema que nos diría en todo momento y para cualquier situación en la práctica, cuál es el modelo matemático apropiado para resolver todo problema.

Es muy interesante notar que un médico no aprende a ejercer su arte exclusivamente en el aula, sino que durante su formación profesional reside un año entero en un hospital atendiendo casos —es decir, tratando pacientes— bajo supervisión de médicos experimentados. En analogía, la formación ideal de un matemático aplicado, debería incluir en algún momento la incorporación de numerosos casos de estudio, lo cual infortunadamente está lejos de ser estándar en carreras de matemáticas. Sería difícil sostener que un estudiante de medicina puede aprender el arte de curar pacientes, mediante el aprendizaje exclusivo de anatomía, descripciones de metodología quirúrgica, y farmacología. Para la disciplina de estadística aplicada, la situación análoga sería pretender que un estudiante aprenda a resolver problemas estadísticos planteados por otros, enseñándole teoría de probabilidad, inferencia

estadística, y numerosos métodos estadísticos, exclusivamente.

Imagínense que un médico aprendiera acerca de úlceras de la siguiente manera. En un libro de texto, y sentado en un aula de clase, lee un libro de texto acerca de lo que son las úlceras, sus caracterizaciones, sus causas, sus efectos, los tratamientos adecuados, y la sintomatología que se asocia con ellas. Hacia el final del capítulo, viene en el libro de texto un ejercicio que dice:

> «Suponga que un paciente se queja de síntomas de ardor estomacal, que una radiografía muestra ...(términos médicos), que un análisis (términos médicos)... resulta positivo en ... (más términos médicos). Indique el tratamiento que es indicado para este caso.»

Claramente, es un ejercicio que contribuye poco a la formación de un médico en su preparación para la vida profesional. En la vida real, llegará a su consultorio un paciente con un problema de salud; no habrá personaje alguno que le diga «Suponga...», y tampoco un personaje que le diga cuáles pruebas clínicas específicas ensayar. El médico tendrá que desarrollar una capacidad para hacer suposiciones por sí mismo, validarlas por diversos medios técnicos confirmatorios, y luego hacer una integración de su conocimiento médico para resolver el problema de su paciente. Esta habilidad va más allá de la medicina.

Similarmente, una educación en matemáticas, en la que un capítulo se estudie, por ejemplo, regresión lineal simple, y que al final se incluya un problema del siguiente estilo:

> «Suponga que X es el tiempo de secado de un proceso industrial y que Y es la dureza del material terminado. Ajuste un modelo de regresión lineal a los siguientes datos. Construya un intervalo de confianza 95 % para el valor esperado de la dureza obtenida tras 30 minutos de secado.»,

no contribuye más que de manera ínfima a la formación de un matemático en el arte de modelación matemática. Simplemente ejercita un aspecto mecanicista de una técnica estadística; no constituye una aplicación cierta de la estadística matemática. En la vida real, llegará a su despacho un cliente con un problema matemático; no habrá personaje alguno que le diga «Suponga» y tampoco un personaje que le diga cuál modelo matemático ensayar. El matemático tendrá que desarrollar una capacidad para hacer suposiciones por sí mismo, validarlas por diversos medios técnicos confirmatorios, y luego hacer una integración de su conocimiento matemático para resolver el problema de su cliente. Esta habilidad va más allá de la matemática.

En resumen, la aplicación de las matemáticas es un *arte*, y es muy difícil enseñar un arte en la demarcación del aula de clases; debe complementarse con una exposición intensa e intencional, a una variedad de problemas reales, así como a los dueños de los mismos. En el CIMAT, en fechas recientes, se ha formalizado el concepto de un *laboratorio*. Existen actualmente tres: de estadística, de matemáticas aplicadas, y de computación. El objetivo es proporcionar instancias que concentran y resuelven problemas de matemáticas provenientes del exterior, y que eventualmente, los alumnos puedan exponerse a ellos, y participar en su solución bajo supervisión.

Lo anterior lo anotamos como visión muy personal de lo que debería ser una educación integral en matemáticas, más aun cuando uno de los perfiles terminales de la carrera sea la posibilidad de ser contratado —como matemático— en diversas instituciones (es decir, para *ejercer* matemáticas aplicadas). La razón por la cual durante el presente curso introductorio hemos enfatizado también aspectos no-matemáticos (es decir, relacionados con el arte de modelación) es que pienso que es útil estar expuestos a ellos desde temprano para

6.1. Distribuciones discretas

6.1.1. Distribución uniforme (discreta)

Definición Si $S = \{x_1, \ldots, x_n\}$ es un conjunto finito de $\mathbb{R}$, decimos que X tiene distribución uniforme sobre S si $f_X(x_i) = 1/n$, $\forall i$.

Parámetros Los parámetros son los números $x_1, \ldots, x_n$, y n.

Notación $X \sim \text{Unif}(x_1, \ldots, x_n)$.

Momentos $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2$.

Comentarios Surge en experimentos de urnas, ruletas, loterías, *etc*. En el caso particular $S = \{1, 2, 3, \ldots, N\}$ entonces hay un solo parámetro, N (Ver Figura 6.1). Las expresiones para momentos tienen gran similitud con las definiciones de primer momento muestral (Definición 5.20) y varianza muestral (Ejercicio 5.9). Sin embargo, los conceptos son radicalmente diferentes: allá se trata de funciones de variables aleatorias X_i mientras que aquí se trata de funciones de constantes conocidas x_i.

---

entender el papel que juegan los modelos matemáticos en la solución de problemas. Estoy convencido que esto dota de una visión más amplia para el estudio posterior de temas matemáticos, aunque éstos pertenezcan al ámbito de matemáticas puras. Esto es particularmente cierto en el área de probabilidad y estadística, pues es muy fácil adquirir la visión de que la probabilidad y estadística es un conjunto grande de *recetas*, que no tienen conexiones entre sí. Por ejemplo, muchos de ustedes han manifestado conocer lo que son los histogramas, y cómo se construyen. Pero estoy casi seguro que pocos me podrían explicar para qué sirven los histogramas, cual es la teoría que los sustenta y cómo pueden ser éstos instrumentales en la solución de un problema estadístico/probabilístico.

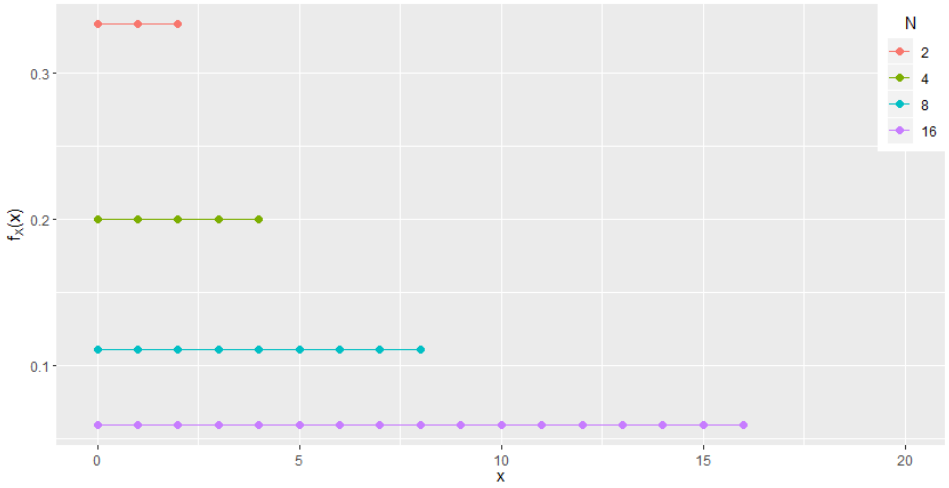

Figura 6.1: Densidades uniformes discretas $\mathrm{Unif}(0, \ldots, N)$ para distintos valores de N.

6.1.2. Distribución binomial

Definición

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

para $x = 0, 1, 2, \ldots, n$.

Parámetros $n = 1, 2, \ldots$; $0 \leq p \leq 1$.

Notación $X \sim \mathrm{Bin}(n, p)$.

Momentos $\mu = np$, $\sigma^2 = np(1-p)$.

Comentarios Surge de contar en número de éxitos en un *experimento binomial*, que consiste en n repeticiones independientes e idénticamente distribuidas de un experimento Bernoulli, es decir, aquel en el cual hay sólo dos resultados posibles, llamados éxito y fracaso, y siendo p la probabilidad (común) de obtener un éxito (ver Ejemplo 5.8). Cuan-

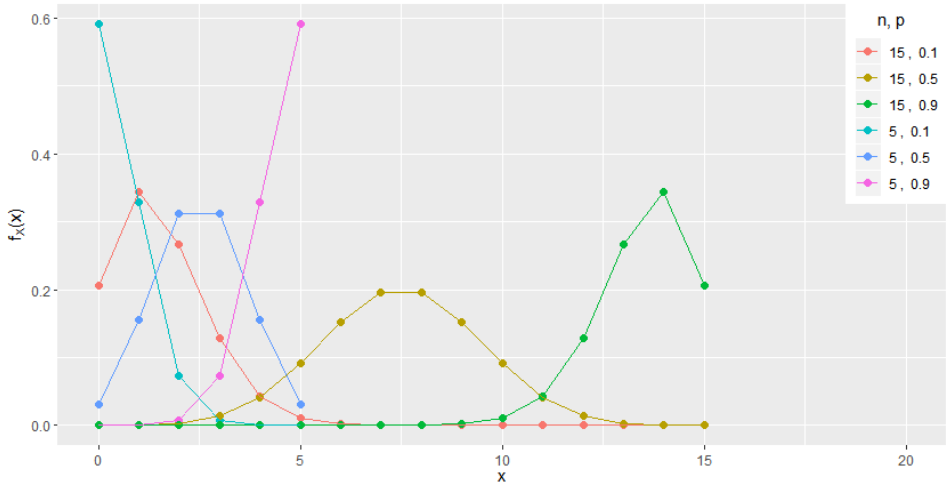

Figura 6.2: Densidades binomiales $\text{Bin}(n, p)$ para distintos valores de los parámetros n, p.

do $n = 1$, la distribución resultante se llama distribución de Bernoulli, y se denota $X \sim \text{Ber}(p)$. Note entonces que si $X \sim \text{Ber}(p)$, $\mathbb{E}(X) = p$ y $\mathbb{V}(X) = p(1-p)$. Note que para la distribución binomial, siempre se tiene $\sigma^2 \leq \mu$.

6.1.3. Distribución geométrica

Definición

$$f_X(x) = p(1-p)^x$$

para $x = 0, 1, 2, \ldots$.

Parámetros $0 < p \leq 1$.

Notación $X \sim \text{Geom}(p)$.

Momentos $\mu = \dfrac{1-p}{p}$, $\sigma^2 = \dfrac{1-p}{p^2}$.

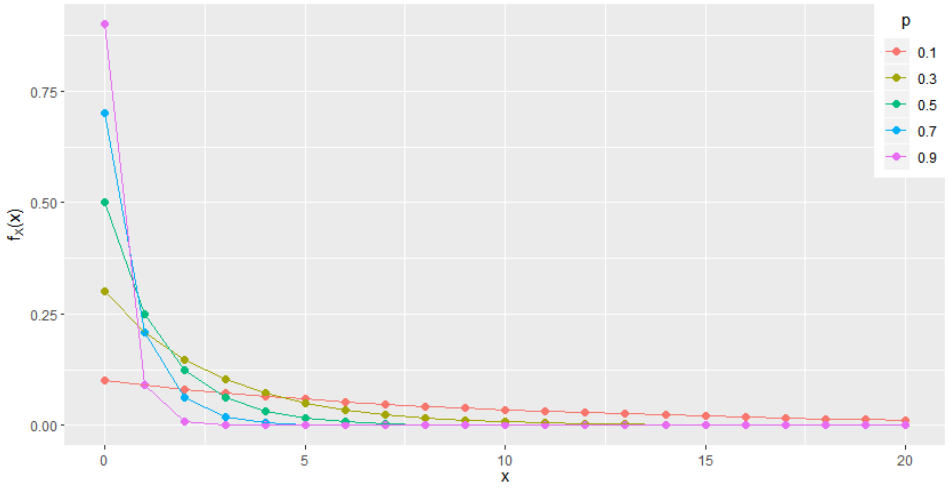

Figura 6.3: Densidades geométricas $\mathrm{Geom}(p)$ para distintos valores del parámetro p.

Comentarios Surge de contar el número de *fracasos* antes de obtener el primer éxito en un experimento binomial, donde p es la probabilidad de obtener un éxito (si se obtiene un éxito en el primer ensayo, el número de fracasos es cero). Note que para la distribución geométrica, siempre se tiene $\mu \leq \sigma^2$. La distribución geométrica posee una propiedad análoga a la que tendrá en breve la distribución exponencial en cuanto a carencia de memoria, en un contexto discreto (Ejercicio 6.9). En algunos textos y paquetes de cómputo, se le llama distribución geométrica a la densidad $f_X(x) = p(1-p)^{x-1}$ para $x = 1, 2, \ldots$, lo cual corresponde a contar el número de *ensayos* antes de obtener el primer éxito en un experimento binomial. Esto puede causar alguna confusión.

Figura 6.4: Siméon Denis Poisson (1781–1840)
El matemático francés, contemporáneo, discípulo y amistad entrañable de Lagrange y Laplace, hizo contribuciones en astronomía, física, teoría de integración, y teoría de probabilidad. Foto: brittanica.com

6.1.4. Distribución de Poisson

Definición

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

para $x = 0, 1, 2, \ldots$.

Parámetros $\lambda > 0$. El parámetro λ en ocasiones se llama la intensidad de la distribución de Poisson.

Notación $X \sim \text{Poisson}(\lambda)$, o $X \sim \text{P}(\lambda)$.

Momentos $\mu = \lambda$, $\sigma^2 = \lambda$.

Comentarios Surge en situaciones en las que se cuenta el número de eventos que suceden en intervalos en la línea o en el tiempo, en la superficie, o en el espacio, que tienen ciertas características, como son:

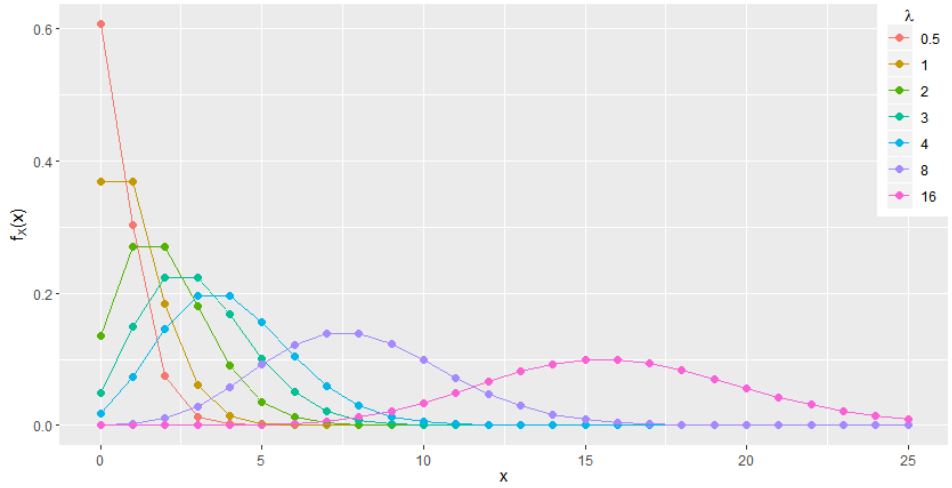

Figura 6.5: Densidades de $\text{Poisson}(\lambda)$ para distintos valores del parámetro λ.

- Que se está contando el número de eventos que suceden en un área (o intervalo de tiempo, o volumen) determinada.
- Que la probabilidad de que suceda un evento sobre un área muy pequeña, es también muy pequeña. En este sentido, en ocasiones se asocia a la densidad de Poisson con el estudio de «eventos raros».
- Que en un mismo lugar (o en el mismo tiempo), no pueden suceder más de uno solo de los eventos que se están contando.
- Que si se duplica el tamaño de la superficie (intervalo de tiempo, *etc*.), entonces se duplica la probabilidad de registrar ahí un evento.

La densidad de Poisson tiene múltiples aplicaciones en comunicaciones (número de llamadas que se reciben en un conmutador telefónico en un intervalo de tiempo), resistencia de materiales (número de fisuras que aparecen al azar en tramos, superficies, o volúmenes de materiales), astrono-

mía (número de supernovas en un sector del universo), medicina (número de afectados por una epidemia en un radio determinado), radioactividad (número de partículas captadas por un contador Geiger), meteorología (número de celdas de tormentas eléctricas sobre un área determinada), *etc*.

Note que para una distribución de Poisson, siempre se tiene que $\mu = \sigma^2$. Otra propiedad importante de esta distribución es que si $X_1 \sim \text{Poisson}(\lambda_1)$ y $X_2 \sim \text{Poisson}(\lambda_2)$ siendo X_1 y X_2 independientes, entonces $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

6.2. Distribuciones continuas

6.2.1. Distribución uniforme (continua)

Definición

$$f_X(x) = \frac{1}{b-a}\, 1_{(a,b)}(x)$$

para $x \in \mathbb{R}$.

Parámetros $-\infty < a < b < \infty$.

Notación $X \sim \text{U}(a, b)$.

Momentos $\mu = \dfrac{a+b}{2}$, $\sigma^2 = \dfrac{(b-a)^2}{12}$.

Comentarios Cuando $a = 0$ y $b = 1$, la distribución resultante se llama *uniforme estándar,* y es la que de ordinario se simula en los lenguajes de programación con funciones de generación de números aleatorios. Una razón por la que la distribución uniforme estándar es muy importante, es porque a partir de ella se pueden simular números aleatorios con cualquier otra distribución.

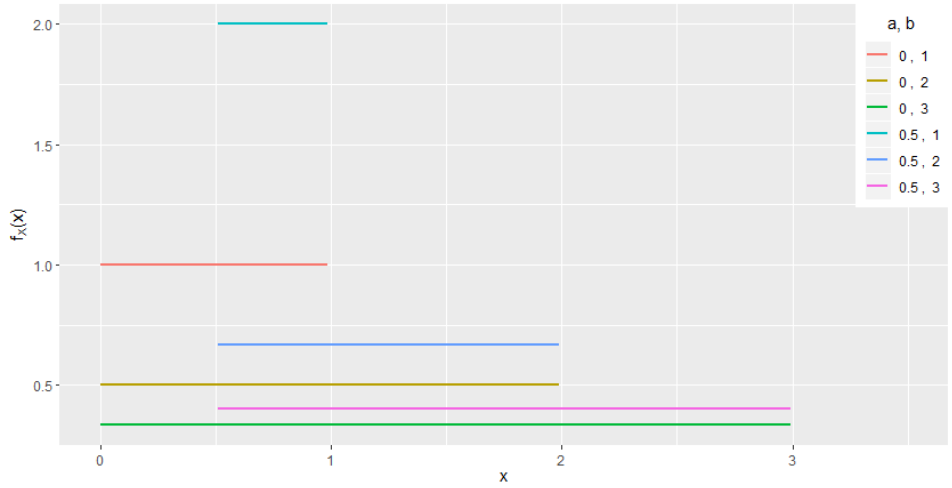

Figura 6.6: Densidades uniformes $\mathrm{U}(a, b)$ para distintos valores de los parámetros a y b. A diferencia de otras densidades continuas que serán abordadas en esta sección, el soporte de la densidad uniforme es finito, dado por $[a, b]$. Si $X \sim U(a, b)$, lo que esto significa es que $\mathbb{P}(X < a) = \mathbb{P}(X > b) = 0$.

6.2.2. Distribución exponencial

Definición

$$f_X(x) = \lambda e^{-\lambda x} \, 1_{(0,\infty)}(x)$$

para $x \in \mathbb{R}$.

Parámetros $\lambda > 0$.

Notación $X \sim \exp(\lambda)$.

Momentos $\mu = \dfrac{1}{\lambda}$, $\sigma^2 = \dfrac{1}{\lambda^2}$.

Comentarios Surge en problemas de tiempos de falla, confiabilidad, y supervivencia, es decir, en el análisis de tiempos hasta que ocurra una falla (donde «falla» puede ser que se descomponga una máquina, que recaiga o muera un paciente, que se sufra un accidente que requie-

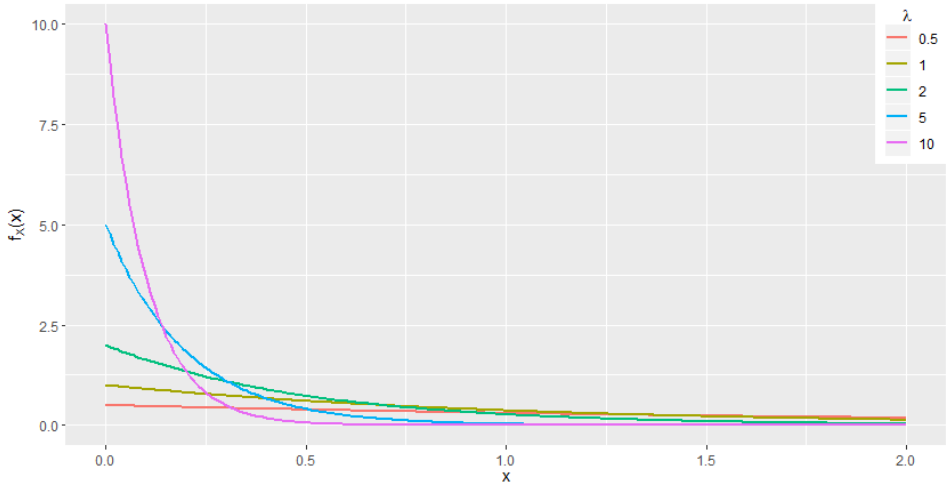

Figura 6.7: Densidades exponenciales $\exp(\lambda)$ para distintos valores del parámetro λ.

ra cobertura por parte de una compañía aseguradora, que arribe un cliente a un negocio, *etc.*). Para esta densidad, la función de distribución es $F_X(x) = 1 - e^{-\lambda x}$. Una peculiaridad de esta distribución es que cumple la propiedad de *carencia de memoria*, que es $\mathbb{P}(X > x + h \mid X > x) = \mathbb{P}(X > h)$ (Ejercicio 6.8). Note además que la distribución exponencial cumple $\mu = \sigma$.

6.2.3. Distribución normal

Definición

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

para $x \in \mathbb{R}$.

Parámetros $-\infty < \mu < \infty$, $\sigma > 0$.

Notación $X \sim \mathcal{N}(\mu, \sigma^2)$. Precaución: Algunos textos usan $\mathcal{N}(\mu, \sigma)$, por lo

que si vemos escrito $\mathcal{N}(2,\, 4)$, debemos aclarar si 4 corresponde con σ o con σ^2.

Momentos Los símbolos usados en la parametrización anterior no son incidentales. Resulta que la media es precisamente μ, y que la desviación estándar es σ, es decir, que la varianza es σ^2.

Comentarios Cuando $\mu = 0$ y $\sigma = 1$, la densidad resultante se llama *normal estándar*. La distribución normal es protagonista en un teorema importante de teoría de probabilidad y estadística, que se llama el Teorema del Límite Central (TLC), que se verá en un capítulo siguiente. El TLC hace que la distribución normal adquiera importancia por sí sola, pero surge también de manera natural en muchos problemas, como por ejemplo, errores de medición. La función de distribución normal estándar, denotada de manera especial por Φ, está dada por $\Phi(x) = \left(1/\sqrt{2\pi}\right) \int_{-\infty}^{x} \exp(-y^2/2)\, dy$. Esta integral no puede resolverse explícitamente, por lo cual deben emplearse métodos numéricos o tablas de probabilidades normales. La distribución normal tiene numerosas propiedades teóricas, entre las que se encuentran las siguientes:

- Si $X \sim \mathcal{N}(\mu,\, \sigma^2)$, entonces $a + bX \sim \mathcal{N}(a + b\mu,\, \sigma^2 b^2)$.
- $$\text{Si } X \sim \mathcal{N}(\mu,\, \sigma^2), \text{ entonces } \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,\, 1). \tag{6.1}$$
- La densidad normal es de forma acampanada, simétrica alrededor de μ.
- La densidad normal tiene puntos de inflexión en $\mu \pm \sigma$, y un máximo en μ.

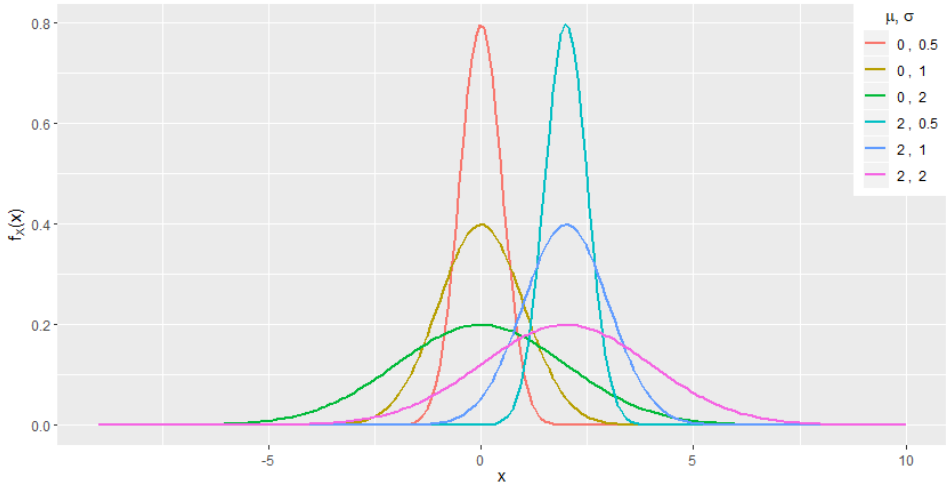

Figura 6.8: Densidades normales o gaussianas $\mathcal{N}(\mu,\, \sigma^2)$ para distintos valores de los parámetros μ, σ.

- Si $X \sim \mathcal{N}(\mu,\, \sigma^2)$, entonces $\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9544$ (aproximadamente 95 %), y $\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9974$ (aproximadamente 99 %). Recuerde que la Desigualdad de Chebychev garantiza que dichas probabilidades son por lo menos 0.7500 y 0.8888, respectivamente.

Uso de tablas La densidad normal no tiene una forma cerrada, explícita, para la integral $\int_a^b f_X(x)\, dx$. Debido a que la distribución normal es de suma importancia, casi cualquier libro de texto incluiría entre sus apéndices por lo menos una *Tabla de Distribución Normal Estándar*. Se trata de valores tabulados de la función

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\, dy$$

sobre una retícula de valores de x. Es decir, se trata de la función de

distribución de una variable aleatoria normal estándar, Z. Esta tabla es útil para *cualquier* densidad normal, debido a la propiedad (6.2.3). Por ejemplo, si $X \sim \mathcal{N}(\mu, \sigma^2)$, y se desea hallar el valor de $\mathbb{P}(a \leq X \leq b)$, entonces la tabla para Φ es útil notando que

$$\mathbb{P}\big(a \leq X \leq b\big) = \mathbb{P}\Big(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\Big) = \\ \mathbb{P}\Big(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\Big) = \Phi\Big(\frac{b-\mu}{\sigma}\Big) - \Phi\Big(\frac{a-\mu}{\sigma}\Big).$$

Por supuesto, que con los avances tecnológicos, el uso de tablas impresas ha caído en desuso, porque los lenguajes de programación cuentan con funciones que calculan numéricamente muchas distribuciones así como sus inversas, y existen también calculadoras *online*. Sin embargo es interesante percibir el valor histórico que tuvo el empleo de tablas publicadas (logarítmicas, trigonométricas, probabilísticas) en el desarrollo de la ciencia.

6.2.4. Distribución Gumbel

Definición La función de distribución es

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy = \exp\Big(-\exp\big(-\frac{x-\alpha}{\beta}\big)\Big),$$

lo cual da lugar a que la función de densidad es

$$f_X(x) = \frac{1}{\beta}\exp\left(-\frac{1}{\beta}\left(x-\alpha+\beta\exp\Big(-\frac{x-\alpha}{\beta}\Big)\right)\right)$$

para $x \in \mathbb{R}$.

Parámetros $-\infty < \alpha < \infty$, $\beta > 0$.

Notación $X \sim \text{Gumbel}(\alpha, \beta)$.

Momentos $\mu = \alpha + \beta\gamma,$ donde $\gamma = 0.577216$, $\sigma^2 = \dfrac{\pi^2\beta^2}{6}$.

Comentarios Ésta es una de las distribuciones que surgen en el estudio de extremos, que tiene aplicación en problemas donde el concepto relevante es el de una variable aleatoria que es un máximo. Por ejemplo, niveles máximos de mareas, velocidades máximas de viento, precipitaciones máximas, *etc*. La *teoría de extremos* es una rama de la probabilidad y la estadística que tiene que ver con las propiedades matemáticas de dichas variables aleatorias. En esta teoría se demuestra que (asintóticamente), sólo existen tres distribuciones capaces de modelar extremos, y la Gumbel es una de ellas. Las otras dos se llaman Frechet y Weibull.

6.2.5. Distribución log-normal

Definición

$$f_X(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\ln\, x - \mu)^2}{2\sigma^2}\right\} 1_{(0,\infty)}(x)$$

para $x \in \mathbb{R}$.

Parámetros $-\infty < \mu < \infty$, $\sigma > 0$.

Notación $X \sim \ln\mathcal{N}(\mu,\, \sigma^2)$.

Momentos $\mathbb{E}(X) = \exp\left(\mu + \sigma^2/2\right)$, $\mathbb{V}(X) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2)$.

Comentarios El nombre log-normal proviene de la siguiente propiedad: Si $X \sim \ln\mathcal{N}(\mu,\, \sigma^2)$, entonces $\ln(X) \sim \mathcal{N}(\mu,\, \sigma^2)$. Esta distribución surge

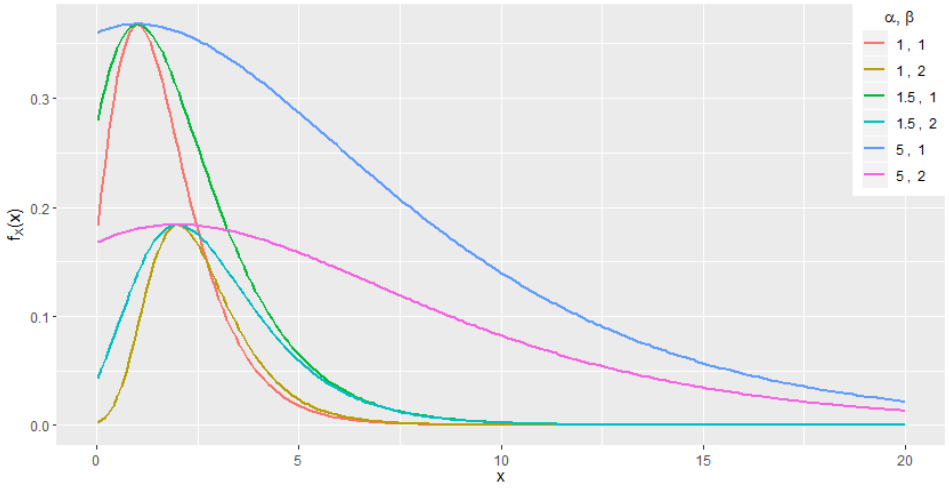

Figura 6.9: Densidades $\text{Gumbel}(\alpha, \beta)$ para distintos valores de los parámetros α, β.

en problemas de mediciones de concentraciones de alguna substancia contaminante, así como para describir tamaños (por ejemplo, tamaños de piedrecillas en el lecho de un río, meteoritos, alturas de árboles, *etc.*).

6.2.6. Distribución Weibull

Definición

$$f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(x/\lambda\right)^k\right) 1_{(0,\infty)}(x)$$

para $x \in \mathbb{R}$.

Parámetros $k > 0$, $\lambda > 0$.

Notación $X \sim \text{Weibull}(k, \lambda)$.

Momentos $\mathbb{E}(X) = \lambda\Gamma(1 + 1/k)$, $\mathbb{V}(X) = \lambda^2\Gamma(1 + 2/k) - \mu^2$.

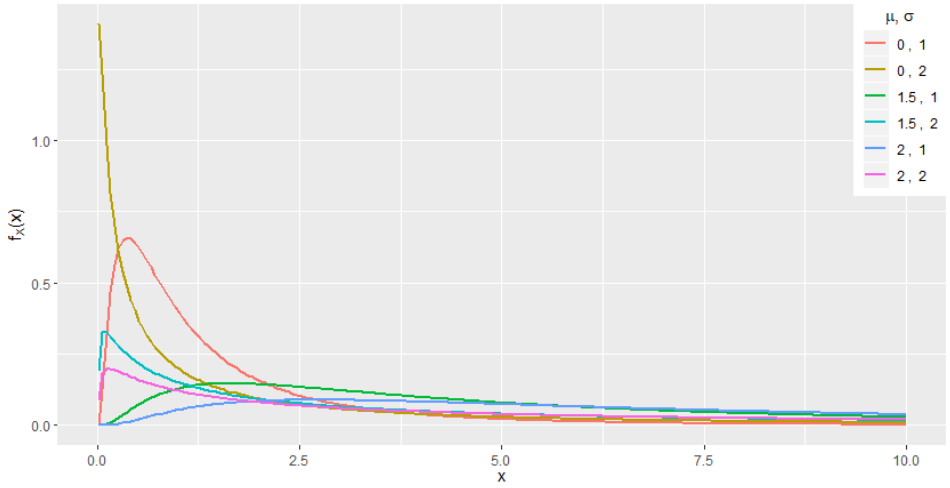

Figura 6.10: Densidades log-normales $\ln \mathcal{N}(\mu,\, \sigma^2)$ para distintos valores de los parámetros μ, σ.

Comentarios En las expresiones para momentos, Γ es la función gamma definida por $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, dt$. Dicha función cumple que al evaluarla en un número natural n, se cumple $\Gamma(n) = (n-1)!$. El parámetro k se llama de forma, y el parámetro λ se llama de escala. Cuando $k = 1$, se obtiene la distribución exponencial. Esta distribución se utiliza en teoría de supervivencia y confiabilidad, es decir, para modelar situaciones donde la variable aleatoria de interés es un tiempo de espera. También se utiliza en teoría de valores extremos. Ha tenido aplicaciones en muy diversas disciplinas, tales como ingeniería, meteorología, actuaría, hidrología, y física.

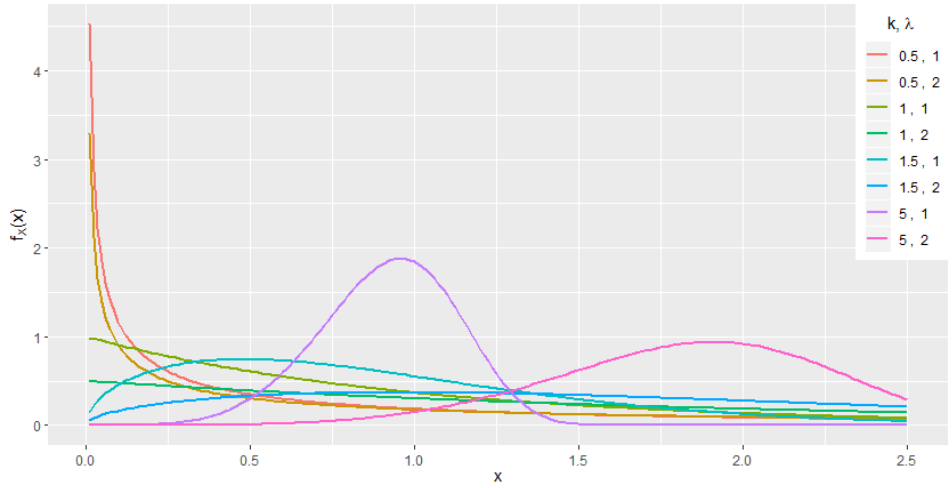

Figura 6.11: Densidades $\mathrm{Weibull}(k, \lambda)$ para distintos valores de los parámetros k, λ. En el contexto de esta distribución, k recibe el nombre de *parámetro de forma*, y λ *parámetro de dispersión* por razones evidentes. Cuando $k = 1$, se reduce a una distribución exponencial.

6.3. Modelos estadísticos

Es esta sección comenzaremos a preparar el camino hacia la realización de inferencia estadística. Supongamos que estamos estudiando una variable aleatoria X. El espacio muestral es claramente $\mathbb{R}$, y la σ-álgebra puede tomarse como los borelianos, $\mathscr{B}$. Supongamos que nos interesa hacer afirmaciones acerca de alguna probabilidad, por ejemplo $\mathbb{P}(X > 50)$, es decir, afirmaciones acerca de la distribución $\mathbb{P}_X$. Si $\mathbb{P}_X$ fuese conocida, no hay nada que hacer, ya que simplemente podría calcularse $\mathbb{P}_X\big((50, \infty)\big)$ de primera intención. Sin embargo, si $\mathbb{P}_X$ se desconoce, entonces habría que establecer $\mathbb{P}_X$ mediante la exploración de una muestra de observaciones de valores aleatorios de X, digamos $X_1, X_2, \ldots, X_n$. En esta sección supondremos que los n valores empíricos de una variable aleatoria ya han sido observados. Veremos algunas herramientas útiles para comenzar a dilucidar cuál puede

ser la naturaleza de $\mathbb{P}_X$. Este será el primer paso en la modelación probabilística de X. El siguiente paso, y el más difícil, será la cuantificación acerca de la incertidumbre que uno acarrea en torno a las afirmaciones que hagamos en relación a $\mathbb{P}_X$ que resulta por no conocer $\mathbb{P}_X$ con certeza, y por ende tampoco podemos hacer la afirmación con entera certeza.

Una primera noción a reconocer con la intención de discernir $\mathbb{P}_X$ es la de un modelo estadístico. Veremos que su definición matemática pudiera parecer ser inocua, pero que la motivación detrás de este concepto tiene todo que ver con el objetivo de inferir algo acerca de una $\mathbb{P}_X$ que es desconocida.

Definición 6.1 (Modelo estadístico) Un *modelo estadístico* es una colección de medidas de probabilidad sobre $\mathscr{B}$. Como concepto de teoría de conjuntos, si $\mathscr{P}$ es el universo de todas las distribuciones posibles para X, un modelo estadístico es meramente un subconjunto $\mathscr{M} \subset \mathscr{P}$.

Note que $\mathscr{M}$ es un conjunto de medidas de probabilidad, y que este conjunto induce un conjunto de modelos de probabilidad dado por $\{(\mathbb{R}, \mathscr{B}, \mathbb{P}) \mid \mathbb{P} \in \mathscr{M}\}$. Los elementos de este último conjunto se diferencian por la medida $\mathbb{P}$. Es usual entonces, por abuso de lenguaje, decir que $\mathscr{M}$ es el modelo estadístico, en lugar de que $\{(\Omega, \mathscr{A}, \mathbb{P}) \mid \mathbb{P} \in \mathscr{M}\}$ lo sea.

Los miembros del modelo estadístico quedan determinados por medidas de probabilidad $\mathbb{P}$ sobre $\mathbb{R}$. Como hemos visto, una medida de probabilidad sobre $\mathbb{R}$ no es más que lo que llamamos la *distribución* de una variable aleatoria X (recuerde el Teorema de Extensión de Kolmogorov, 5.3). Más aun, vimos que la distribución de X a su vez se puede determinar a través de una función de densidad, o de una función de distribución acumulada, o a tra-

vés de otros medios (como funciones generadoras de momentos y funciones características, que no hemos visto con detalle durante el presente curso). Correspondientemente, existe un segundo abuso de notación y de lenguaje, cuando en lugar de explicitar las medidas $\mathbb{P}$, se explicitan en su lugar subconjuntos de densidades, f, o funciones de distribución, F (se considerarán ejemplos de esto más adelante).

Un modelo estadístico $\mathscr{M}$ puede consistir de un solo elemento, por ejemplo, si $\mathscr{M}$ consiste sólo de la medida de probabilidad $\mathbb{P}$ que corresponde a $\mathrm{Bin}(5,\, 0.25)$. Entonces el modelo estadístico contiene uno y sólo un modelo de probabilidad, $(\mathbb{R}, \mathscr{A}, \mathbb{P})$. Note que conceptualmente hay diferencia entre $\mathbb{P}$ (una medida de probabilidad) y $\{\mathbb{P}\}$ (un modelo estadístico que contiene un solo elemento). Un modelo estadístico, en el otro extremo de complejidad, puede consistir de *todas* las medidas de probabilidad sobre $\mathscr{B}$, es decir, $\mathscr{M} = \mathscr{P}$.

Si bien la definición matemática de modelo estadístico es sumamente elemental, puede reflexionarse ampliamente sobre el papel que juega $\mathscr{M}$ en la tarea de realizar inferencia estadística. El modelo estadístico $\mathscr{M}$ es el punto de partida, y se debe elegir de tal manera que contenga a la distribución desconocida, $\mathbb{P}_X$. Representa entonces una restricción del conjunto $\mathscr{P}$ sobre el cual se considerarán candidatos para $\mathbb{P}_X$. En este tema es relevante entonces la siguiente noción:

Definición 6.2 (Error de especificación) Si $\mathbb{P}_X$ es la distribución (presuntamente desconocida) de los datos, y $\mathbb{P}_X \notin \mathscr{M}$, decimos que en el proceso de modelación se ha cometido *error de especificación*. Decimos que un modelo estadístico $\mathscr{M}$ está *correctamente especificado*, si $\mathbb{P}_X \in \mathscr{M}$.

Notar que $\mathscr{M} = \mathscr{P}$ y $\mathscr{M} = \{\mathbb{P}\}$, donde $\mathbb{P}$ es una única distribución de probabilidad dada, son elecciones válidas y extremas de modelo estadístico. Es claro que si se eligiera $\mathscr{M} = \mathscr{P}$, entonces *nunca* se cometería error de especificación, mientras que si $\mathscr{M} = \{\mathbb{P}\}$, entonces *siempre* se cometería error de especificación —a menos de que $\mathbb{P} = \mathbb{P}_X$, una elección demasiado fortuita dado que el problema de inferencia estadística está planteado precisamente porque *no* conocemos $\mathbb{P}_X$—. Es evidente que entonces hay alguna razón por la cual uno no siempre realiza la elección $\mathscr{M} = \mathscr{P}$, y que alguna ventaja implica reducir el modelo estadístico a un conjunto menor. La razón radica en que entre más grande sea $\mathscr{M}$, más incierta sería la conclusión inferida a partir de datos, ya que $\mathbb{P}_X$ tendría más posibilidades. Pero entre más chico se elige $\mathscr{M}$, más posible es cometer error de especificación. Este «toma y daca» entre imprecisión de inferencia y error de especificación ilustra que la idea a conseguir en la práctica es entonces: Elegir $\mathscr{M}$ lo más chico posible sin que se cometa error de especificación.[4] Así, el modelo estadístico tiene la interpretación de recoger todo lo que uno pueda especificar acerca de la medida de probabilidad $\mathbb{P}_X$, para luego utilizar los datos para realizar

[4]Mediante otra analogía, podemos explicar de mejor manera este concepto de «toma y daca» entre dificultad para inferencia y error de especificación: Un policía ministerial está investigando un delito. Se trata de un proceso inherentemente inferencial, pues con base en pistas (llamémosle datos, X) él está tratando de encontrar una explicación a los hechos observados. Su objetivo primordial es encontrar al culpable (llamémosle $\mathbb{P}_X$), y para ello lo primero que hace es restringir su búsqueda a una subcolección de sospechosos (llamémosle $\mathscr{M}$). Por ejemplo, si testigos de los hechos declaran que el delincuente era varón, de unos 25 años de edad, y de complexión robusta, esta descripción automáticamente restringe el conjunto de sospechos posibles y descarta a otros (mujeres, y personas fuera de ese rango de edad). Su labor se restringe a buscar entre sospechosos que cumplan esas características. Es claro que si su conjunto de sospechosos es demasiado amplio, su labor de investigación es correspondientemente más difícil. Pero si su conjunto de sospechosos es demasiado pequeño, el policía ministerial corre gran riesgo de cometer error de especificación. Un error de especificación claro se daría si el delincuente es adolescente y él restringe su búsqueda a adultos mayores. Así, la investigación óptima procede cuando el conjunto de sospechosos es lo más chico posible sin cometer error de especificación. La elección $\mathscr{M} = \mathscr{P}$ corresponde en esta analogía a que la investigación abarca todos los habitantes del país.

inferencia dentro de un subconjunto específico.

Si se comete error de especificación, es decir, si se estipula un modelo $\mathscr{P}$ tal que $\mathbb{P}_X \notin \mathscr{P}$, y a pesar de ello se procede a realizar análisis y conclusiones acerca de $\mathbb{P}_X$ en términos de los datos $X_1, \ldots, X_n$, los resultados generalmente no serán válidos. La teoría estadística estudia modelos estadísticos, particularmente los procedimientos para analizarlos correctamente. Se llama inferencia estadística al acto de concluir algo acerca de $\mathbb{P}_X$ con base en piezas de información $X_1, \ldots, X_n$. Si en alguna situación, la medida $\mathbb{P}_X$ sí fuese conocida, no es entonces relevante el concepto de un modelo estadístico. Formular entonces algún modelo estadístico, hablar de errores de especificación, y de inferencia acerca de $\mathbb{P}_X$, sería por demás una actividad meramente académica (aunque posible, dados datos $X_1, \ldots, X_n$). La estadística es relevante cuando alguna característica de $\mathbb{P}_X$ no se conoce por completo.

Ejemplo 6.1 Supongamos que se está estudiando la variable aleatoria X, donde X es el número de clientes que ingresan a una tienda departamental en una hora. Si bien la distribución exacta de X puede no conocerse del todo, sí hay algunas características generales que podemos afirmar respecto a $\mathbb{P}_X$. Por ejemplo, sabemos sin lugar a dudas que X toma valores enteros positivos, y que se trata de una variable aleatoria discreta. Así, excluimos de inmediato la colección entera de distribuciones continuas para X, y aun para las discretas, se descartan aquellas cuyo soporte incluya también enteros negativos. Esto en sí mismo constituye ya una restricción del conjunto $\mathscr{P}$. Una restricción aun mayor, es considerar que X es una distribución de Poisson, conclusión a la que podría llegarse si el contexto dicta que

los axiomas de la Sección 6.1.4 son posibles. En tal caso, podríamos decir $X \sim \text{Poisson}(\lambda)$, y entonces el modelo estadístico sería $\mathscr{M} = \{\mathbb{P}_\lambda \mid \lambda > 0\}$, donde $\mathbb{P}_\lambda$ representa la distribución de Poisson con parámetro λ.

El ejemplo anterior toca de manera directa la siguiente importante noción:

Definición 6.3 (Modelo estadístico paramétrico) Si X es una variable aleatoria con una densidad (discreta o continua) de probabilidad perteneciente a alguna familia paramétrica, denotaremos por $\theta \in \mathbb{R}^d$, $d \geq 1$ al parámetro, y usaremos la notación genérica $f(x;\theta)$ para denotar a la densidad de X que depende de θ, para $\theta \in \Theta \subset \mathbb{R}^d$. Si convenimos en identificar una distribución con una densidad de probabilidad, entonces el subconjunto $\mathscr{M} = \{f(x;\theta) \mid \theta \in \Theta\}$ recibe el nombre de *modelo estadístico paramétrico*, y Θ se llama el *espacio paramétrico*. Cuando no es posible representar $\mathscr{M}$ de esta manera, el modelo recibe el nombre de *no-paramétrico*.

Ejemplo 6.2 Si $X \sim \text{Bin}(n,p)$, entonces $\theta = (n,p)$ y $f(x;\theta) = \binom{n}{x}p^x(1-p)^{n-x}$, y $\Theta = \mathbb{N} \times [0,1]$. Podemos escribir entonces un modelo estadístico como $\mathscr{M} = \{f(x;\theta) \mid \theta \in \Theta\}$. Note que dependiendo de la situación, el parámetro θ puede ser unidimensional, o multidimensional.

Un modelo paramétrico es entonces un subconjunto de $\mathscr{P}$ que es representable mediante un sistema de rótulos (θ), cuya variación a lo largo de un conjunto (Θ), genera $\mathscr{M}$ (una especie de *curva* en $\mathscr{P}$ generada por el parámetro θ). Un modelo no-paramétrico sería un subconjunto que no es posible

representar de esta manera. Se trataría de un subconjunto que por su complejidad, no admite de una representación simplificada, y en este sentido un no-paramétrico es más «grande» que un paramétrico. Conclusión: Los modelos no-paramétricos surgen cuando uno quiere protegerse contra el que un modelo paramétrico cometa error de especificación. Pero a cambio, un modelo no-paramétrico introduciría mayor incertidumbre en la inferencia. Un modelo estadístico bien puede ser un subconjunto de medidas de probabilidad que no pueda describirse por una familia paramétrica, y en estos casos se habla de *estadística no-paramétrica*. Lo que este término en general significa es técnicas para analizar datos que no requieren de suponer una forma genérica específica (tal como un modelo paramétrico) sino que son válidos bajo condiciones más generales.

Ejemplo 6.3 Consideremos la familia de distribuciones dada por las distribuciones de Poisson, es decir, aquellas cuya densidad (discreta) es $f(x;\lambda) = e^{-\lambda}\lambda^x/x!$. Sea $\mathbb{P}_\lambda$ la medida de probabilidad sobre $\mathbb{R}$ dada por $\mathbb{P}_\lambda(B) = \sum_{x\in B} f(x;\lambda)$. Estas densidades definen un conjunto de modelos de probabilidad cuando se les mira con el siguiente enfoque: $\big\{(\mathbb{R},\, \mathscr{A},\, \mathbb{P}_\lambda) \mid \lambda > 0\big\}$. Dicho conjunto, abusando de notación, se escribe también como $\{\mathbb{P}_\lambda \mid \lambda > 0\}$, o más aun, como $\{f(x;\lambda) \mid \lambda > 0\}$. Lo importante es denotar que un modelo estadístico consiste de un *conjunto* de medidas de probabilidad. Este modelo estadístico es paramétrico, con $d = 1$, $\theta = \lambda$ y $\Theta = (0, \infty)$.

Ejemplo 6.4 Otro modelo estadístico es $\big\{\, \mathbb{P}_\lambda \mid \lambda \in \{1, 2\}\big\}$. Este modelo contiene solamente dos elementos: Los modelos de probabilidad que corresponden a distribuciones de Poisson para dos valores de λ (1 y 2). También

es un modelo paramétrico, con $d = 1$, $\theta = \lambda$ y $\Theta = \{1, 2\}$.

Ejemplo 6.5 Otros modelos estadísticos paramétricos son (usando ya la notación convenida) $\mathscr{M}_1 = \{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma > 0\}$ y $\mathscr{M}_2 = \{\mathcal{N}(0, \sigma^2) \mid \sigma > 0\}$. Note que en este ejemplo, se tiene que $\mathscr{M}_2 \subset \mathscr{M}_1$. En este sentido se puede hablar de modelos estadísticos «más grandes» que otros, de contención de un modelo estadístico en otro, o de un modelo anidado en otro. De hecho, los modelos estadísticos se pueden ilustrar gráficamente a través de diagramas de Venn en el universo de las medidas de probabilidad.

Ejemplo 6.6 El modelo descrito por $\{f(x) \mid f$ es una densidad continua, simétrica alrededor del origen$\}$, es un modelo estadístico no-paramétrico. La razón es que no es posible barrer la totalidad de este conjunto de densidades mediante el barrido de un parámetro de dimensión finita.

Ejemplo 6.7 (Encuesta de opinión) Suponga que se seleccionará al azar una muestra de n individuos para preguntarles su opinión respecto a algún punto específico. Suponga que la respuesta es binaria, que pueden dar una respuesta favorable (1) o desfavorable (0). La muestra es $X_1, \ldots, X_n$, donde por contexto se puede establecer que $X_i \sim \text{Ber}(p)$, para alguna $p \in [0, 1]$. Podemos entonces establecer un modelo estadístico para la variable aleatoria X, dado por el conjunto de todas las distribuciones Bernoulli. En este caso, el parámetro p tiene la interpretación de ser la proporción de individuos en la población muestreada que tienen opinión favorable. El valor de este parámetro es desconocido —por eso se piensa en una encuesta—. Si se conociera el valor de p, la encuesta es una actividad ociosa, porque se

sabría de entrada que $X_i \sim \mathrm{Ber}(p)$. ¿Qué puede decir acerca de error de especificación en este caso?

Ejemplo 6.8 (Problema de medición) Suponga que se está midiendo una característica física de algún objeto con un instrumento sensible y preciso. Al repetir el proceso de medición, no se obtienen exactamente las mismas lecturas. La razón es que el instrumento es sensible a ligerísimas variaciones del medio ambiente (humedad, temperatura, *etc*.) que influyen en el proceso de la medición misma, aunque el objeto que se esté midiendo sea siempre el mismo. Las mediciones son observaciones $X_1, \ldots, X_n$ de una variable aleatoria X. Es variable aleatoria simplemente porque antes de realizar la medición, no podemos predecir cuál será la lectura correspondiente. Supongamos que se adopta el modelo estadístico dado por $\{\mathcal{N}(\mu, \sigma^2)\}$ para fines de determinar la distribución de X. En este caso, el parámetro μ tiene la interpretación de ser el valor esperado de X, lo cual puede concebirse como la magnitud física desconocida que uno pretende medir, siempre y cuando el instrumento no introduzca un sesgo. El parámetro σ tiene la interpretación de ser la variabilidad de las mediciones, lo cual es una característica del instrumento de medición más que del objeto que se pretende medir. El parámetro μ (y posiblemente también σ) es desconocido. Si se conociera μ, entonces no tendría necesidad de medir —por eso se está recurriendo al instrumento de medición—. ¿Qué puede decir acerca de error de especificación en este caso?

6.4. Exploración de datos y ajuste de distribuciones

6.4.1. Notas sobre el proceso de modelación

La primera pregunta natural que alguien pudiera formular acerca de este proceso, es cómo saber cuál modelo estadístico específico considerar para representar el fenómeno X. En general, no existe un procedimiento universal que logre el objetivo de identificar correctamente tal $\mathscr{M}$. Su determinación resulta de aplicar varios criterios por parte del modelador de datos, que toman en cuenta el contexto del problema, algunas consideraciones teóricas, experiencia previa, retroalimentación, y hasta instintos (ver Ejemplo 6.1). Cualquier modelo estadístico (de hecho, cualquier modelo matemático) que pretenda representar una situación dada será a final de cuentas una propuesta por parte del modelador, y el grado de aproximación que logre el modelo podría ser adecuado para un objetivo e inadecuado para otro.

Con relación a modelos estadísticos, motivados por el hecho de que una distribución de probabilidad $\mathbb{P}_X$ no se conoce plenamente, a continuación se describe una clasificación ilustrativa por medio de ejemplos, de situaciones en las cuales existen diversos grados de «desconocimiento» de $\mathbb{P}_X$, y por lo tanto, diversos grados de especificación de $\mathscr{M}$:

1. *Modelo estadístico finito.* Aunque no muy común, esta sería una situación en la cual se conociera por algún motivo que $\mathbb{P}_X$ radica en una colección finita de distribuciones. Un ejemplo es una fábrica que cuenta con dos máquinas para producir tornillos de diámetro X, una que produce artículos bajo una distribución $\mathbb{P}_1$ (conocida) y la otra bajo distribución $\mathbb{P}_2$ (conocida). Si se obtiene una muestra de X y no se conoce cuál máquina produjo X, lo que sí se sabría es que

$\mathbb{P}_X \in \mathscr{M} = \{\mathbb{P}_1, \mathbb{P}_2\}$.

2. *Familia de distribuciones.* La familia sí es conocida; el valor específico del parámetro no lo es. Esta situación sí es muy común. Ocurre cuando por algún motivo, sí se conoce la clase de densidades que son factibles (ver Ejemplo 6.7). Hay dos motivos principales por los cuales uno podría conocer dicha clase:

   a) Por contexto directo. Un ejemplo es un experimento binario, en el cual uno elige al azar una persona y X es un indicador de masculino-femenino o de derecho-zurdo. En esta situación claramente una distribución para X es Bernoulli, pero el parámetro $p = \mathbb{P}(X = 1)$ no es conocido. Surgen muchos otros ejemplos de esta situación en problemas de muestreo o de conteo. Las distribuciones geométrica, binomial, y uniforme típicamente ocurren bajo este contexto.

   b) Por contexto y una consideración teórica. Un ejemplo es alguna situación que por contexto pudiera dar lugar a un experimento Poisson (número de clientes que entran a un comercio, o número de fallas de maquinaria en un intervalo de tiempo). En este caso se puede adoptar tentativamente la familia Poisson por razones teóricas (si alguien no conociera los axiomas que dan lugar a una densidad de Poisson no podría dar con esta argumentación). Un segundo ejemplo es adoptar una familia de distribuciones de extremos si el contexto de los datos es observar máximos o mínimos de alguna variable aleatoria (si alguien no conociera teoría de extremos, tampoco podría dar con esta propuesta). Esta situación también abarca las situaciones en las que se pudiera involucrar a

un probabilista teórico para fines de estudiar las propiedades teóricas del mecanismo que produce datos. Por ejemplo, en finanzas hay procesos aleatorios cuyas propiedades distribucionales pueden calcularse usando suposiciones elementales.

3. *Modelación empírica.* La familia no es conocida, pero se adopta de una familia paramétrica (obviamente, el valor del parámetro posible también se desconoce). Esta situación también es frecuente en la práctica. Cuando el análisis del contexto no permite proponer familias específicas, la opción es recurrir a modelos estadísticos basados únicamente en consideraciones empíricas. Es decir, el argumento para defenderlos radica en que un proceso de validación produce afinidad entre datos observados y modelo propuesto. Un ejemplo típico es cuando se adopta un polinomio en análisis de regresión; el polinomio puede no tener significado físico, pero de cualquier forma se adopta para fines de analizar si los datos son explicables con ese polinomio.

4. *Modelos no-paramétricos.* La familia no es conocida, y no hay afinidad con algún modelo paramétrico (ni siquiera empíricamente). Esta situación representa el grado de desconocimiento acerca de la distribución de X más extremo. Un ejemplo ilustrativo es el caso en que X es el error cometido por un instrumento de medición. Quizás no hay condiciones para suponer un modelo paramétrico específico (tal como la familia normal) pero sí las hay para suponer densidades *simétricas*. Las densidades simétricas son tan numerosas y generales, que no admiten una representación paramétrica.

6.4.2. Histogramas

Los histogramas son instrumentos gráficos útiles en la exploración de posibles densidades para una variable aleatoria X. Consideraremos los dos casos para X, discreta y continua:

Caso discreto Si $X_1, X_2, \ldots, X_n$ son realizaciones de una variable aleatoria discreta, entonces existe un conjunto $S = \{x_1, \ldots, x_m\}$ sobre el que estas realizaciones toman valores ($m \leq n$). Por la LGN (ver Ejemplo 5.27), la cantidad

$$f_n(x) = \frac{\#\text{ de } X_i \text{ que fueron iguales a } x}{n}$$

converge a $\mathbb{P}(X = x) = f_X(x)$, sin importar cuál sea la densidad de X ni que la conozcamos o no. El histograma es simplemente la gráfica de $f_n(x)$ como función de $x \in S$. El histograma puede verse entonces como una gráfica aproximada de la función de densidad de X, f_X. Note que el histograma no depende de ninguna forma postulada para la densidad de X. Sólo depende de las observaciones $X_1, X_2, \ldots, X_n$. Es decir, para dibujar un histograma, no importa cuál sea la densidad de X; ésta no interviene en su construcción, y por lo tanto no tiene relevancia el que no la conozcamos.

Caso continuo Si $X_1, X_2, \ldots, X_n$ son realizaciones de una variable aleatoria continua, entonces la idea anterior no es practicable porque para una continua,

$$\frac{\#\text{ de } X_i \text{ que fueron iguales a } x}{n}$$

siempre convergerá a 0 para cualquier valor de x. Sin embargo, si en

lugar de puntos aislados pensamos en intervalos, la situación es diferente. Supongamos que las observaciones suceden sobre algún intervalo $I = (a, b]$. Sean $a = a_0 < a_1 < \cdots < a_k = b$ números que determinan una partición del intervalo I, en el siguiente sentido: $I = \cup_{i=1}^{k}(a_{i-1}, a_i\,]$. La partición es *regular*, si $a_i - a_{i-1}$ es constante, es decir, que la partición consiste de sub-intervalos de la misma longitud. Aunque puede construirse un histograma para una partición irregular, lo más usual es que se haga sobre una partición regular. Por la LGN (ver Ejemplo 5.28), si $a < b$ la cantidad

$$f_n(a, b) = \frac{\#\text{ de } X_i \text{ que cayeron en el intervalo } (a, b\,]}{n}$$

converge a $\mathbb{P}(X \in (a, b\,]) = \int_a^b f_X(x)\,dx$, sin importar cuál sea la densidad de X ni que la conozcamos o no. El histograma es una gráfica de una función escalonada sobre el intervalo I, dibujada de tal modo que el *área* de un rectángulo sobre el sub-intervalo $(a_{i-1}, a_i\,]$ es igual a $f_n(a_{i-1}, a_i)$. El histograma puede verse entonces como una gráfica aproximada de la función de densidad de X, f_X, en el sentido de abarcar *áreas* similares sobre los sub-intervalos $(a_{i-1}, a_i\,]$.

Note, similarmente, que el histograma en el caso continuo no depende de ninguna forma postulada para la densidad de X. Sólo depende de las observaciones $X_1, X_2, \ldots, X_n$. Sí depende de algunas elecciones arbitrarias, que son, el número de subintervalos, k, y la partición sobre el intervalo I. Si se acuerda que la partición sea regular, entonces, lo que queda por establecer en la práctica, es sólo determinar el valor de k. Lo que queremos es que el histograma nos sea informativo en cuanto a la posible naturaleza de $\mathbb{P}_X$. Si k es demasiado peque-

ño, el histograma tendrá un aspecto demasiado burdo para ser informativo, y si k es demasiado grande, entonces el histograma tendrá un aspecto demasiado rugoso como para recoger aspectos relevantes (muchos subintervalos estarían vacíos, por lo que f_n será 0). Con la experiencia, se adquiere un don para determinar el número k. Entre mayor sea n, mayor puede y debe ser k. Una regla empírica es tomar $k = \log_2(n)$, aunque no debe tornarse este aspecto de selección de k como un aspecto de vida o muerte (esta regla empírica prescribe que para $n = 1000$ se tomen alrededor de $k = 10$ clases). Para lo que queremos emplear un histograma, poca diferencia cualitativa habrá entre hacer un histograma con 10 clases o con 12 clases.

En resumen, debido a la LGN, los histogramas proporcionan una primera fotografía acerca de la distribución $\mathbb{P}_X$. Su construcción y la convergencia que los avala no depende de cuál sea $\mathbb{P}_X$. Evidentemente, si n no es muy grande, por ejemplo $n = 5$, el concepto de histograma debe tomarse con muchísima reserva, es decir, que no podemos otorgarle la misma significancia que cuando n fuera 200 o 1000. Podemos sin duda construir un histograma para $n = 5$, pero la pregunta es, ¿tiene sentido? Los paquetes de cómputo sin duda, si se les pide, graficarán con bombos y platillos histogramas para $n = 5$, pero nuestro rol como analistas de información deberá tomar éstos con la debida reserva.

6.4.3. Estimación por el método de momentos

Ahora, a diferencia de la sección anterior, supondremos que $X_1, X_2, \ldots, X_n$ tienen una densidad descrita por algún modelo paramétrico, $f(x; \theta)$. Es decir, supondremos que conocemos que la densidad pertenece a la familia,

y que lo único que nos hace falta por discernir es el valor particular del parámetro θ. Un ejemplo sencillo que da lugar a esta situación es la siguiente. Suponga que se lanzará una moneda, y que X es la variable aleatoria que toma el valor 1 si cae águila y 0 si cae sol.

Caso moneda balanceada En este caso, podemos sin duda postular que $X \sim \text{Ber}(1/2)$, ya que el ser balanceada significa que $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = 1/2$, lo cual no es más que la densidad Bernoulli con parámetro 1/2. En este caso se conoce entonces la distribución de X. No es necesario invocar conceptos de observación empírica. El modelo de probabilidad $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$ está completo.

Caso moneda deforme Supongamos que la moneda ha sido distorsionada, por lo que la suposición de balanceada ha dejado de poderse sostener. Sabríamos entonces que $\mathbb{P}(X = 1)$ toma algún valor, digamos p, y que $\mathbb{P}(X = 0)$ sería $1 - p$. Sin embargo, el valor específico de p no necesariamente es $1/2$. Lo que estamos diciendo, en otras palabras, es $X \sim \text{Ber}(p)$, donde $p \in [0, 1]$. Un número n de lanzamientos experimentales de dicha moneda son variables aleatorias $X_1, X_2, \ldots, X_n \sim \text{Ber}(p)$. En este caso, conocemos que la densidad de X pertenece a la familia Bernoulli, y lo único que nos hace falta por discernir es el valor particular del parámetro p. Para acabar de especificar el modelo de probabilidad $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$, hace falta entonces especificar el valor de p.

El problema en general se llama *problema de estimación paramétrica*, y será el tema concreto del Capítulo 8. Con base en datos $X_1, X_2, \ldots, X_n \sim f(x; \theta)$ deben encontrarse valores plausibles de θ. Llamaremos un *estimador*

puntual de θ a una función $\hat{\theta}(X_1, X_2, \ldots, X_n)$ de los datos $X_1, X_2, \ldots, X_n$ que tiene por objeto aproximarse al valor desconocido de θ (en algún sentido), sin importar cuál sea éste. El método de momentos es una de las técnicas de uso general que pueden utilizarse para este fin. Intuitivamente, consiste en igualar momentos muestrales con momentos teóricos.

Supongamos que θ es de dimensión d. Consideremos los momentos muestrales $\hat{\mu}_1, \ldots, \hat{\mu}_d$. También pueden usarse momentos muestrales centrados, o alguna combinación de centrados y no-centrados. Ahora consideremos, dado que la familia $f(x;\theta)$ está dada, las contrapartes teóricas de la distribución, es decir, $\mu_1, \ldots, \mu_T$. Notemos que las cantidades $\mu_1, \ldots, \mu_d$ serán todas función del parámetro θ, porque todas ellas son función de $f(x;\theta)$. Ahora establezcamos el siguiente sistema de d ecuaciones con d incógnitas:

$$\begin{aligned} \hat{\mu}_1 &= \mu_1 \\ \hat{\mu}_2 &= \mu_2 \\ &\vdots \\ \hat{\mu}_d &= \mu_d. \end{aligned}$$

A la solución de este sistema, la llamaremos $\hat{\theta}$, y recibe el nombre de estimador por el método de momentos. Es importante notar que la solución $\hat{\theta}$ depende de los datos $X_1, X_2, \ldots, X_n$; depende de haber supuesto que la familia es $f(x;\theta)$, pero no presupone conocer el valor de θ.

Cuando $d = 1$, lo más usual es utilizar $\hat{\mu}_1$ y μ_1 (es decir, medias). Cuando $d = 2$, lo más usual o conveniente es usar $\hat{\mu}_1$, $\hat{\mu}_2'$ y μ_1, μ_2' (es decir, medias y varianzas). A continuación, algunos ejemplos.

Ejemplo 6.9 (Densidad Bernoulli) En este caso, $X \sim \text{Ber}(p)$, se tiene $d = 1$. Los datos son una colección $X_1, X_2, \ldots, X_n$ de ceros y unos al azar. La media muestral es $\hat{\mu}_1 = n^{-1} \sum_{i=1}^n X_i$ y la media teórica es $\mu_1 = \mathbb{E}(X) = p$ (recuerde propiedades de la densidad Bernoulli). Haciendo el sistema (de una sola ecuación) $\hat{\mu}_1 = \mu_1$ se obtiene directamente $\hat{p} = n^{-1} \sum_{i=1}^n X_i$. Observe que el estimador puntual de p por el método de momentos no es más que la proporción de veces que la variable X toma el valor 1.

Ejemplo 6.10 (Densidad de Poisson) Aquí también se tiene $d = 1$. Para la familia Poisson, el valor esperado es λ. Siendo el primer momento muestral $n^{-1} \sum_{i=1}^n X_i$, obtenemos de inmediato $\hat{\lambda} = n^{-1} \sum_{i=1}^n X_i$.

Ejemplo 6.11 (Densidad exponencial) Nuevamente, $d = 1$. La media de una exponencial es $1/\lambda$. Igualando esto con la media muestral, obtenemos la ecuación $1/\lambda = n^{-1} \sum_{i=1}^n X_i$, de donde se concluye $\hat{\lambda} = n/\sum_{i=1}^n X_i$.

Ejemplo 6.12 (Densidad geométrica) El valor esperado de una geométrica con parámetro p es $(1-p)/p$. Igualando con el primer momento muestral, obtenemos $(1-p)/p = n^{-1} \sum_{i=1}^n X_i$, de donde resolviendo por p obtenemos $\hat{p} = (1 + n^{-1} \sum_{i=1}^n X_i)^{-1}$.

Ejemplo 6.13 (Densidad normal) Aquí, $d = 2$. El primer momento teórico es μ, y la varianza es σ^2. El primer momento muestral es $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ y la varianza muestral es $S^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Haciendo el sistema de dos ecuaciones $\mu = \bar{X}$ y $\sigma^2 = S^2$, y resolviendo, encontramos que $\hat{\mu} = \bar{X}$ y que $\hat{\sigma} = S$.

Ejemplo 6.14 (Distribución log-normal) Recuerde que para esta distribución, $\mathbb{E}(X) = \exp(\mu + \sigma^2/2)$, $\mathbb{V}(X) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2)$. Hay dos parámetros ($d = 2$). Tomando media y varianza muestral, se plantea el siguiente sistema de dos ecuaciones:

$$\begin{aligned}\bar{X} &= \exp(\mu + \frac{1}{2}\sigma^2),\\ S^2 &= \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2).\end{aligned}$$

La solución a este sistema de ecuaciones proporciona

$$\hat{\mu} = -\frac{1}{2}\ln\frac{S^2 + \bar{X}^2}{\bar{X}^2} + \ln\bar{X},$$

y

$$\hat{\sigma} = \sqrt{\ln\frac{S^2 + \bar{X}^2}{\bar{X}^2}}.$$

Observe en todos los ejemplos que las soluciones al sistema de ecuaciones planteado por el método de momentos, dependen de los datos $X_1, X_2, \ldots, X_n$.

La razón por la cual el método de momentos proporciona un método justificable matemáticamente, radica en la LGN. En efecto, la LGN nos afirma (ver Notación (5.2) y convergencia (5.3)) que no importa cuál sea la densidad $f(x;\theta)$ ni cuál sea el valor particular del parámetro θ, que $\hat{\mu}_k \xrightarrow{\mathrm{P}} \mu_k$. Esto es lo que provoca que las soluciones al sistema de ecuaciones converjan a θ, es decir, $\hat{\theta} \xrightarrow{\mathrm{P}} \theta$, $\forall\theta$. Note que para que se cumpla esta convergencia, no es necesario conocer el valor particular de θ. Teniendo a la mano datos $X_1, X_2, \ldots, X_n$, uno puede calcular $\hat{\theta}$, y la LGN entonces lo que me dice a

final de cuentas es que a medida que crece n, que el valor de $\hat{\theta}$ se parecerá a θ.

6.4.4. Estimación por el método de máxima verosimilitud

(Este concepto no le he transcrito aún a las presentes notas. Por favor basarse en notas tomadas en clase y/o ayudantía para este tema.)

Definición 6.4 (Densidad ajustada) Si bajo la suposición $X_1, X_2, \ldots, X_n \sim f(x;\theta)$, $\hat{\theta}$ es el estimador por algún método (por ejemplo, momentos o máxima verosimilitud), llamamos a $f(x;\hat{\theta})$ la densidad ajustada por los datos.

Note que si en efecto, $X_1, X_2, \ldots, X_n \sim f(x;\theta)$, entonces la densidad ajustada $f(x;\hat{\theta})$ convergerá a $f(x;\theta)$. Pero si en realidad, la densidad de X es una, y para el método de momentos considero una alternativa diferente, entonces el método de momentos no necesariamente produciría que la densidad ajustada se parezca a la densidad verdadera de X. Con un ejemplo, ilustraremos esto explícitamente.

Ejemplo 6.15 (Error de especificación de un modelo) Supongamos que en la realidad $X_1, X_2, \ldots, X_n \sim \mathrm{P}(5)$, pero que al no saber esto cometemos un error: suponer que $X_1, X_2, \ldots, X_n \sim \mathrm{Geom}(p)$. Es decir, las v.a.'s en la realidad tienen densidad dada por $e^{-5}5^x/x!$ mientras que estamos suponiendo que su densidad es de la forma $p(1-p)^x$. Aplicamos el método de momentos tomando en cuenta esta familia geométrica, y encontramos (ver Ejemplo 6.12) que el estimador para el parámetro p es $\hat{p} = \{1 + n^{-1}\sum_{i=1}^{n} X_i\}^{-1}$. Por

otra parte, por la LGN, sabemos que $n^{-1}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} 5$, porque el valor esperado de las X_i en la realidad es 5, con lo que $\hat{p} \xrightarrow{\mathrm{P}} \{1+5\}^{-1} = 1/6$. Con esto, nuestra densidad ajustada cumple $\hat{p}(1-\hat{p})^x \xrightarrow{\mathrm{P}} (1/6)(5/6)^x$, lo cual como función de x no se parece a la densidad real, $e^{-5}5^x/x!$.

Note que si no hubiéramos cometido el error de especificación del modelo, es decir, que la suposición hubiera sido correcta en cuanto a que $X_1, X_2, \ldots, X_n \sim \mathrm{P}(\lambda)$ para alguna λ, entonces nuestro proceder con el método de momentos hubiera sido calcular el estimador (ver Ejemplo 6.10) como $\hat{\lambda} = n^{-1}\sum_{i=1}^{n} X_i$. En este caso, también por la LGN, $\hat{\lambda} \xrightarrow{\mathrm{P}} 5$, y la densidad ajustada $e^{-\hat{\lambda}}\hat{\lambda}^x/x!$ sí hubiera convergido a la densidad real, $e^{-5}5^x/x!$.

6.4.5. Comparación gráfica de histogramas con densidades ajustadas

Repasemos los dos conceptos importantes vistos en las secciones anteriores, para después pasar a ponerlas a trabajar juntas en la tarea de modelación probabilística.

Primero, contamos con un instrumento gráfico llamado histograma, que cumple aproximarse a la verdadera densidad de X, cualquiera que ésta sea. No depende de suponer nada acerca de la naturaleza de la verdadera densidad de X. El histograma siempre converge a la densidad real, sin depender de ninguna suposición al respecto de ella.

Segundo, un dispositivo llamado estimación por el método de momentos, que produce una densidad ajustada *bajo la suposición específica* de una familia particular, $f(x;\theta)$. Cuando la suposición específica es correcta, entonces la densidad ajustada será muy parecida a la densidad verdadera de X. Cuando la suposición específica es incorrecta, entonces la densidad ajus-

tada no necesariamente es parecida a la densidad verdadera de X (ver Ejemplo 6.15).

Lo anterior sugiere el siguiente dispositivo, que se basa en un sencillo argumento de tipo lógico. Supongamos que se observa una v.a. X cuya función de densidad desconocemos. Supongamos que se postula o se sospecha que una familia $f(x;\theta)$ específica es útil para explicar la variación de la v.a. X. Supongamos que se cuenta con una muestra $X_1, X_2, \ldots, X_n$ de realizaciones empíricas de X, y que se realizan las siguientes dos actividades: Se construye un histograma, y se calcula una densidad ajustada por el método de momentos basada en la familia $f(x;\theta)$. Ahora se comparan el histograma, y la densidad ajustada.

Si estas dos gráficas tienden a coincidir, entonces la interpretación es que la familia postulada está siendo efectiva para explicar la densidad de X, es decir, que no hay causa para sospechar que hemos cometido error de especificación. La lógica es: Dado que densidad ajustada se parece a histograma, entonces densidad ajustada se parece a densidad real (porque histograma siempre se parece a densidad real), entonces no hay error flagrante de especificación (porque si lo hubiera, difícilmente se hubiera dado el parecido entre histograma y densidad ajustada).

Por otra parte, si histograma y densidad ajustada resultaran ser visiblemente diferentes, entonces la interpretación sería que la postulación de la familia $f(x;\theta)$ debe ser errónea. La lógica es: Si no fuera errónea, entonces me hubieran salido bastante parecidas.

Hemos abordado una técnica exploratoria para examinar la naturaleza de la distribución de probabilidad de una variable aleatoria, a través del análisis de realizaciones empíricas de la misma. Lo anterior lo hemos hecho por ahora a un nivel gráfico-intuitivo, y usando un estimador puntual

específico, relativamente fácil de calcular y entender, basado en el método de momentos. Cabe mencionar que el problema de determinar si un juego de datos es o no explicado por una familia $f(x;\theta)$ puede abordarse desde un punto de vista analítico formal, por medio de herramientas de estadística matemática. Existe una disciplina entera (dentro de estadística matemática) conocida con el nombre de *bondad de ajuste*, que tiene por objetivo confrontar una densidad de probabilidad con un juego de datos. Por otra parte, en la llamada *teoría de estimación paramétrica*, se abordan muchos otros principios generales y métodos para realizar estimación, entre los cuales el método de momentos es sólo un caso particular. Finalmente, podemos señalar que en otra rama llamada *inferencia no-paramétrica*, se estudian métodos de análisis para casos en los que no puede prescribirse una familia paramétrica $f(x;\theta)$ para modelar la situación.

Ejercicios

6.1 Verifique las expresiones anotadas arriba para $\mathbb{E}(X)$ y $\mathbb{V}(X)$, para las siguientes distribuciones: Uniforme discreta, geométrica, binomial, Poisson, uniforme continua, y exponencial.

6.2 Encuentre estimadores por el método de momentos para los parámetros de todas las densidades (discretas y continuas) que se anotaron en este capítulo. Note que algunas ya se cubrieron a manera de ejemplos; en este caso, sólo verifíquelos.

6.3 (Verificación empírica en la computadora, de estimación por el método de momentos) Genere una colección de n datos $X_1, X_2, \ldots, X_n$, de alguna familia $f(x;\theta)$. Ahora calcule el estimador correspondiente $\hat{\theta}$, y ve-

rifique que $\hat{\theta} \approx \theta$. Repita el experimento para diversos valores de n y de θ.

6.4 (Paquetería estadística de cómputo) En este momento del curso se considera adecuado familiarizarse con el empleo de paquetería comercial de estadística. La Facultad de Matemáticas cuenta con uno, llamado Statistica. Cabe señalar que el software para estadística constituye una industria pujante y altamente competida, y que Statistica representa tan solo uno de los ejemplos que existen en el mercado. Su interfase es típica para aplicaciones de Windows, siendo relativamente fácil adiestrarse en sus funciones. La siguiente es una lista de habilidades que se consideran deseables manejar en la presente etapa del curso:

(a) La creación de una base de datos en formato Statistica (*.sta) (Menú File→New Data).

(b) La importación de datos en otro formato (por ejemplo *.txt) para ser convertido a una base de datos en formato Statistica (menú File→ Import Data).

(c) El cálculo de momentos muestrales para una lista de datos (módulo Basic Statistics→Descriptive Statistics).

(d) La construcción de histogramas dada una lista de datos (menú Graphs→ Stats 2D Graphs→ Histograms).

(e) La superposición de histogramas con densidades ajustadas para diversos modelos (menú Graphs→Stats 2D Graphs→Histograms→Fit type).

(f) Crear y guardar gráficas, para fines de impresión, preparación de materiales escritos y presentaciones.

6.5 Utilice una calculadora (o tabla) de distribución normal para encontrar las siguientes cuatro probabilidades:

(a) $\mathbb{P}(X > 7)$ si $X \sim \mathcal{N}(5,\, 4^2)$.

(b) $\mathbb{P}(-1 < X < 3)$ si $X \sim \mathcal{N}(1,\, 1)$.

(c) $\mathbb{P}(X < 0.5)$ si $X \sim \mathcal{N}(0,\, 1)$.

(d) $\mathbb{P}\big(|X - 2| > 2\big)$ si $X \sim \mathcal{N}(1,\, 1)$.

6.6 Suponga que la estatura de seres humanos en cierta población tiene distribución normal con media 1.60m y desviación estándar 0.07m.

(a) Suponga que se construye un marco para una puerta, cuyo claro es de 1.70m. Si se elige al azar a una persona de esta población, ¿cuál es la probabilidad de que por su estatura ésta pueda pasar por debajo?

(b) ¿De qué altura debe construirse el marco de una puerta para que pase por debajo de ella 95 % de la población?

6.7 Considere $X \sim \mathrm{Poisson}(\lambda)$. Encuentre:

(a) El valor esperado de la variable aleatoria $2X$.

(b) La varianza de $2X$.

(c) La densidad de $2X$, es decir, $f_{2X}(x)$.

6.8 A partir de la definición de probabilidad condicional, demuestre la propiedad de carencia de memoria que posee la distribución exponencial:

$$\mathbb{P}(X > x + h \mid X > x) = \mathbb{P}(X > h).$$

6.9 Muestre que la distribución geométrica posee una propiedad de carencia de memoria similar a la que posee la distribución exponencial, pero que aplica para tiempos de espera discretos. Coloquialmente esto significa que si se tiene la intención de repetir un experimento Bernoulli hasta obtener el primer éxito, entonces dado que el primer éxito todavía no ha ocurrido, la probabilidad condicional del número adicional de ensayos necesarios no depende del número de ensayos ya realizados.

Capítulo 7

Planteamiento de inferencia estadística

Suponga que uno se enfrenta a un fenómeno aleatorio, cuantificable a través de una variable aleatoria X. Hemos visto que la distribución de X, $\mathbb{P}_X$, es el ingrediente necesario para conformar un modelo de probabilidad $(\mathbb{R}, \mathscr{A}, \mathbb{P}_X)$, y que hay circunstancias bajo las cuales no se conoce $\mathbb{P}_X$. En este caso es necesario formular un modelo estadístico. Un modelo estadístico constituirá la infraestructura matemática para hacer conclusiones válidas acerca de la naturaleza desconocida de $\mathbb{P}_X$, utilizando para ello el registro de observaciones empíricas del fenómeno aleatorio X.

Comenzaremos por notar que si se piensa con sumo cuidado, el desconocimiento acerca de $\mathbb{P}_X$ puede dar origen a diversas preguntas formuladas por posibles actores confrontados con una situación de análisis de datos. Las diferencias —algunas sutiles— tienen que ver con que el interés ulterior puede radicar en alguna característica concreta de la distribución, y no toda la distribución propiamente dicha. Ello dará lugar a identificar dos grandes

tipos de problemas estadísticos que surgen en la práctica.

Se concluirá el capítulo con algunas nociones técnicas relevantes para inferencia estadística: funciones de una muestra que se llaman *estadísticas*, y la distribución teórica de ellas, conocida como *distribución muestral*. Se mencionarán importantes teoremas de teoría de probabilidad cuyo objetivo es proporcionar ciertas aproximaciones útiles. Todos estos conceptos figurarán de manera muy importante en los Capítulos 8 y 9.

7.1. Problemas estadísticos: Estimación y pruebas de hipótesis

En este curso abarcaremos dos tipos de problemas estadísticos llamados estimación y prueba de hipótesis. Esto no significa que no existen muchos otros problemas estadísticos. Sin embargo, con nociones de estos dos problemas se adquiere capacidad para abordar un numeroso conjunto de situaciones que se dan en la práctica, incluyendo encuestas, aplicaciones en control estadístico de calidad en procesos industriales, verificación de hipótesis en ciencias experimentales, y muchos más.

Supongamos que estamos ante un fenómeno aleatorio cuya distribución no se conoce, y que se tiene acceso a observaciones $X_1, \ldots, X_n$ en forma de variables aleatorias, que no son más que realizaciones empíricas del fenómeno. Supongamos también que tras la aplicación de técnicas de estadística descriptiva (es decir, cálculo de momentos muestrales, graficación de histogramas, densidades, densidades ajustadas, y otros similares) que se ha vislumbrado un modelo estadístico paramétrico $\{f(x;\theta) \mid \theta \in \Theta\}$, donde $f(x;\theta)$ es una densidad y θ es un parámetro, cuyo valor se desconoce. Re-

cuerde que si θ se conociera, no habría necesidad de formular un modelo estadístico. La razón por la cual los modelos estadísticos son relevantes se debe precisamente a que el valor de θ no se conoce. En virtud de este desconocimiento acerca del valor de θ, hay dos tipos de preguntas que pueden formularse de inmediato. En lo que sigue, sea $\Theta_0 \subset \Theta$, un subconjunto propio del espacio paramétrico.

Tipo 1 de pregunta ¿Cuál es el valor de θ?

Tipo 2 de pregunta ¿Θ_0 contiene el valor desconocido de θ, sí o no?

Cuál de las dos preguntas estadísticas pueda ser relevante en una situación dada, depende del contexto y de la pregunta de interés formulada en el ámbito de la aplicación concreta. Veremos ejemplos más adelante. Cuando la pregunta relevante sea de Tipo 1, hablamos de un problema de *estimación* (paramétrica) y cuando la pregunta relevante es de Tipo 2, hablamos de un problema de *prueba de hipótesis*. Luego veremos que el conjunto Θ_0 recibirá el nombre de hipótesis nula. En la nomenclatura del Capítulo 6, Θ_0 no es más que un modelo estadístico.

En un principio puede parecer que si se contesta la pregunta de Tipo 1, que automáticamente se contesta la pregunta del Tipo 2. Sin embargo, este razonamiento lógico no procede, y se debe a que cualquier respuesta que podemos dar con respecto a la pregunta de Tipo 1 estará dotada necesariamente de incertidumbre. Por esta razón, basar una respuesta a la pregunta de Tipo 2 con base a una respuesta incierta a la pregunta de Tipo 1, no constituye un procedimiento lógico. Por otra parte, intuitivamente se desprende que la pregunta del Tipo 2 es más sencilla de contestar que la pregunta de Tipo 1, por lo que las nociones de precisión en cada caso no son equivalentes, ni cualitativa ni cuantitativamente hablando. Por ejemplo, suponga que

un ser extraterrestre está interesado en determinar si las monedas terrestres son legales (es decir, que tienen probabilidad de «águila» igual a 1/2) o no, y para ello lanza 50 veces una moneda. Encuentra que ocurren 28 «águilas», calcula que un estimador para la probabilidad de águila es $\hat{p} = 28/50 = 0.56$. Luego concluye, debido a que $0.56 \neq 0.50$, que la moneda no ha de ser legal. Esto es un error, y se debe a que el extraterrestre confundió las nociones de estimación con prueba de hipótesis.[1]

Ejemplo 7.1 Retomar el Ejemplo 6.7, en el que un modelo estadístico para describir los resultados de una encuesta está dado por $\{\mathrm{Ber}(p) \mid p \in [0,1]\}$. Recordemos que el parámetro p tiene la interpretación de ser la proporción de la población que posee opinión favorable acerca de un punto específico. Supongamos que el interesado en la encuesta pregunta «¿Cuál es la proporción de individuos que prefieren mi producto?» Se trataría entonces de un problema de estimación. Pero si el interesado en la encuesta pregunta «¿Domino el mercado, sí o no?» entonces se trataría de un problema de prueba de hipótesis, con $\Theta_0 = (1/2, 1]$.

Ejemplo 7.2 Se nos presenta una moneda y se nos invita a jugar un juego de azar. Se nos permite lanzar la moneda unas cuantas veces con el objeto de determinar si es legal o no. Los lanzamientos son variables aleatorias $\mathrm{Ber}(p)$, por lo que el modelo estadístico es también $\{\mathrm{Ber}(p) \mid p \in [0,1]\}$. La pregunta de interés aquí no es conocer el valor de p, sino determinar si es

[1] Nos burlamos del extraterrestre, pero hay muchas instancias terrestres de razonamientos similares. Por ejemplo, hemos todos oído hablar de casos parecidos al siguiente: Se realizó una encuesta, en la cual 56 % favorece al candidato A, y 44 % favorece al candidato B. Por lo tanto, A tiene mayores votos que B. Este razonamiento es tan errado como lo sería que el extraterrestre concluyera que las águilas son más numerosas que los soles porque en la muestra le salieron más águilas que soles.

plausible o no es plausible que $p = 1/2$. Se trata entonces de una prueba de hipótesis, con $\Theta_0 = \{1/2\}$.

Ejemplo 7.3 En un experimento sobre parapsicología, se estudia a una persona que dice ser dotado de telepatía. Se elaboran 5 cartas con diversos dibujos, se van eligiendo al azar y se le solicita a la persona, situada en una habitación aislada, que nos adivine la carta que salió. Se repite el experimento n veces, y se registra si hubo o no acierto en cada realización. Un modelo estadístico es otra vez $\{\mathrm{Ber}(p) \mid p \in [0,1]\}$, donde un éxito es acertar la carta, y p es la probabilidad de acierto para el sujeto de prueba. En ausencia de poderes telepáticos, estaríamos de acuerdo en que p toma el valor 1/5, pero si existe telepatía legítima, entonces $p > 1/5$. La pregunta científica de interés es «¿El sujeto me está adivinando al azar, o posee telepatía?». Esto constituye entonces un problema de prueba de hipótesis, con $\Theta_0 = \{1/5\}$. Aquí no es relevante el valor preciso de p, sino conocer si excede o no excede el valor 1/5 que se obtendría bajo la carencia de poderes telepáticos.

Ejemplo 7.4 En inspección de lotes producidos en una fábrica o en la inspección de materia prima, es usual que se defina un límite de tolerancia de defectuosos, por ejemplo, se dice que la proporción de defectos no deberá exceder 0.05. Tras un esquema de muestreo para inspección de artículos, si p es la proporción (desconocida) de defectuosos, la pregunta relevante es «¿p excede o no excede 0.05?». Se trata de un problema de prueba de hipótesis.

Ejemplo 7.5 Suponga que una máquina en una línea de producción se puede calibrar a que produzca artículos con una tamaño medio determinado. Suponga que las especificaciones de la línea de producción son que la medida nominal de las piezas deberá ser 2.5cm. Se hace una inspección de algunos artículos seleccionados al azar, y se les mide su tamaño. Sea μ la media de las piezas que está produciendo la máquina en este momento. La pregunta relevante es «La máquina se encuentra correctamente calibrada, o no?». Se trata de una prueba de hipótesis, pues la pregunta relevante en términos del parámetro es «¿$\mu = 2.5,$ o $\mu \neq 2.5$?». La respuesta a esta pregunta determinará si se para la línea de producción con el fin de recalibrar la máquina.

Ejemplo 7.6 Un fabricante de automóviles desea conocer el número medio μ de kilómetros por litro que proporciona cierto motor, porque existe una norma que establece que este valor deberá publicarse. Se trata de un problema de estimación de μ.

Ejemplo 7.7 Un cliente que compró un automóvil de la marca del ejemplo anterior, nota que una calcomanía pegada sobre el parabrisas dice que su auto da 10.7 kilómetros por litro. Sin embargo, el cliente sospecha que su automóvil no da tanto, y desea documentar su caso con análisis de datos. Hace un registro de consumo de gasolina y de kilómetros recorridos. Para el cliente, el problema es uno de prueba de hipótesis, ya que su pregunta es en realidad «¿$\mu = 10.7$ o $\mu < 10.7$?» Su pregunta, en principio, no es «¿Cuál es el valor de μ?». Su pregunta tiene más bien que ver con conocer si datos obtenidos de su automóvil sugieren que la especificación del fabricante es

correcta o no.

Ejemplo 7.8 Se sabe que la supervivencia media de pacientes de cáncer bajo cierto tratamiento preestablecido es 18.7 meses. Una nueva droga se está desarrollando, y para ser liberada al mercado es necesario saber si la nueva droga es mejor que la droga anterior o no lo es. Se trata entonces de un experimento (llamado en este contexto un ensayo clínico) que tiene por objeto determinar si la supervivencia media de la nueva droga es mayor que 18.7 meses o no lo es. Se trata de una prueba de hipótesis.

Definición 7.1 (Problema de estimación) Considere un modelo estadístico paramétrico $\{f(x;\theta) \mid \theta \in \Theta\}$, donde $f(x;\theta)$ es una densidad y θ es un parámetro. El problema es de estimación cuando el valor de θ se desconoce y el objetivo es es discernir los valores plausibles que éste puede tomar con base en la información contenida en una muestra.

Definición 7.2 (Problema de prueba de hipótesis) Sea $\mathbb{P}_0$ el modelo de probabilidad (desconocido) que corresponde al fenómeno aleatorio de interés. Sea $\mathscr{M}$ un modelo estadístico correctamente especificado, es decir que cumple $\mathbb{P}_0 \in \mathscr{M}$. Sean $\mathscr{P}$ y $\mathscr{Q}$ dos modelos estadísticos tales que $\mathscr{P}, \mathscr{Q} \subset \mathscr{M}$, $\mathscr{P} \cup \mathscr{Q} = \mathscr{M}$, y $\mathscr{P} \cap \mathscr{Q} = \phi$. Decimos que el problema es de prueba de hipótesis cuando el objetivo es discernir entre $\mathbb{P}_0 \in \mathscr{P}$ y $\mathbb{P}_0 \in \mathscr{Q}$, con base en la información contenida en una muestra.

El problema llamado bondad de ajuste, en el que la pregunta es si los datos provienen o no de una distribución específica, también es una prueba

de hipótesis en el siguiente sentido. Por ejemplo, considere la pregunta que plantea si un fenómeno sigue la distribución normal o si sigue alguna otra. Poniendo $\mathscr{P} = \{\text{densidades normales}\}$ y $\mathscr{Q} = \{\text{densidades no normales}\}$, se ve que se trata en realidad de una prueba de hipótesis.

Cuando los modelos $\mathscr{M}$, $\mathscr{P}$ y $\mathscr{Q}$ se encuentran en un contexto paramétrico, éstos como hemos visto, se pueden describir en términos de valores de parámetros. Por ejemplo, si $\mathscr{M}$ es el modelo de todas las densidades de Poisson, podemos escribir $\mathscr{M} = (0, \infty)$, $\mathscr{P} = (0, 5)$, y $\mathscr{Q} = [5, \infty)$. En este caso, es usual y conveniente usar la notación Θ, Θ_0, y Θ_1 en lugar de $\mathscr{M}$, $\mathscr{P}$ y $\mathscr{Q}$.

Un ejemplo de un problema estadístico importante que no está incluido en lo anterior, es el llamado problema de predicción, en el cual la pregunta es cuál es el valor que adquirirá una variable aleatoria en un momento futuro. Por ejemplo, predecir la inflación para el año próximo. Este no es ni prueba de hipótesis ni estimación de un parámetro.

7.2. Estadística: Mitos y realidades

No pretendemos que los objetivos que se señalarán enseguida sean exhaustivos, ni que constituyan una definición formal de estadística. Es una tarea difícil (si no imposible) determinar qué constituye materia de estudio de una disciplina y qué no lo constituye. Lo siguiente pretende dar un sabor de boca acerca de cuáles son los problemas a los que se dedica la estadística en general.

Suponga que se tiene un fenómeno aleatorio. No se conoce la medida de probabilidad que lo rige, pero se cuenta con la posibilidad de acceder a observaciones empíricas del mismo (es decir, datos). Note con ello que hay por

lo menos dos ingredientes importantes: La noción de aleatoriedad, y la noción de datos. Note además que el no conocer una medida de probabilidad o tener acceso sólo a algunas observaciones de un fenómeno aleatorio no constituyen defectos de la matemática ni de la estadística; son simplemente condiciones impuestas por la realidad en la que vivimos. La disciplina llamada *estadística*, en todas sus ramas, hoy día abarca metodología, principios y conceptos, que tienen como objetivos:

1. Resumir y describir información relevante contenida en datos (estadística descriptiva).

2. Explorar y discernir relaciones estructurales que pueden existir en un fenómeno aleatorio (análisis exploratorio de datos).

3. Explorar y discernir modelos estadísticos apropiados, así como validar modelos estadísticos, en el sentido de confrontar datos con un modelo dado (selección de modelos y bondad de ajuste).

4. Dotar de medios para abordar las preguntas de interés que surjan respecto al fenómeno aleatorio (modelación estadística).

5. Utilizar la información contenida en los datos para extraer de ellos conclusiones relevantes y válidas acerca del fenómeno aleatorio bajo consideración. Esto incluye alguna forma de determinar la incertidumbre en la que se incurre debido al hecho de que la conclusión se extrae a partir de observación parcial del fenómeno aleatorio, así como interpretar con corrección los resultados que se deriven de análisis de datos (inferencia estadística).

6. Determinar cuántos datos, dónde, cuándo y cómo seleccionarlos, co-

mo función del objetivo específico que se pretende lograr con ellos, y de manera óptima con respecto a la cantidad de información relevante que los datos son capaces de proporcionar (diseño muestral y diseño experimental).

7. Desarrollar nuevos modelos estadísticos, así como los métodos apropiados para realizar análisis con ellos, elaborar criterios y nociones útiles y relevantes en problemas estadísticos (investigación en estadística matemática).

8. Desarrollar metodología numérica y gráfica para realizar los puntos anteriores (investigación en cómputo estadístico).

Cuando para todo lo anterior se utilizan modelos matemáticos, se habla de *estadística matemática*. Es importante notar, contrario a la concepción popular de estadística, que la estadística tiene relevancia aun *antes* de obtener los datos. Esto es, es muy común la creencia de que la estadística tiene que ver sólo con el análisis de datos, y que por tanto, su rol comienza una vez que se hayan realizado y concluido los procesos correspondientes de recolección de datos. Esto es, la estadística matemática puede contribuir también de manera muy importante en la fase de planeación y formulación de un estudio en el que se recabará información por medio de muestreo.

Otra concepción popular de la palabra *estadística* es que se equipara a gráficas, tablas, números, porcentajes, y diagramas. Estos conceptos tienen que ver básicamente con el punto llamado estadística descriptiva en el panorama anterior. Cabe notar que la relación de estadística con modelos de probabilidad se pierde totalmente si se le concibe de esta manera, siendo que la definición misma que hemos señalado para un modelo estadístico conlleva en su parte constitutiva modelos de probabilidad.

Otra interpretación errónea de estadística tiene que ver con el hecho de que con métodos de estadística descriptiva se pretenden objetivos de inferencia estadística, o por lo menos el público general así lo percibe. Por ejemplo, en el periódico de hoy (*El Correo de Hoy*, 3 de mayo de 2000) hay en la primera plana un encabezado que dice textualmente: «A 60 días de las elecciones para gobernador del estado, el 51.4 por ciento de los ciudadanos votarían por Juan Carlos Romero Hicks, candidato del PAN, cuyo más cercano competidor, Juan Ignacio Torres Landa, del PRI, tiene una preferencia electoral de 32.3 por ciento». No es difícil encontrar ejemplos de este tipo en medios electrónicos e impresos cotidianos, en los que se menciona que el resultado es producto de un «estudio estadístico». Si se lee con atención la cita anterior, se trata de una afirmación con relación a una población mayor (el padrón electoral estatal, entero), con base en el análisis de una muestra (de 800 individuos, según el diario). ¿Será *absolutamente cierto* que el 51.4 % de los ciudadanos hoy votarían por el PAN? La respuesta es que no exactamente, no obstante la redacción empleada para construir el encabezado. Para conocer con absoluta certeza cuál es dicho porcentaje, habría que haber encuestado a *todos* los ciudadanos, ejercicio que obviamente no es posible realizar. Para dar el salto hacia inferencia estadística más formal, sería necesario conocer acerca de cómo cuantificar incertidumbre. Esto tiene que ver con conocer acerca de probabilidad. Obviamente, la cultura popular no está capacitada para entenderlo, y la realidad es que tampoco los medios están capacitados para explicarlo al público, dando por resultado afirmaciones vagas tales como la ejemplificada.

Existe una reputación estereotípica que se ha formado con relación a la estadística. Por ejemplo, existen chistes como «Hay tres tipos de mentiras: Las buenas, las malas, y las *estadísticas*», y libros tales como «How to Lie

with Statistics» (Huff & Geis, 1993, Cómo mentir con estadística). Lo anterior sugiere, de manera indebida, que la estadística como disciplina es dolosa por naturaleza. La estadística sí tiene que ver con extraer conclusiones bajo condiciones de incertidumbre. Sin embargo, dicha incertidumbre no es introducida por la estadística misma como disciplina, sino por un contexto impuesto por una situación real. Por ejemplo, el que el porcentaje 51.4 arriba mencionado contenga incertidumbre (posibilidad de error), no es consecuencia de un método estadístico. Dicha incertidumbre es consecuencia de que bajo la imposibilidad de encuestar al padrón entero, se vislumbra como una de las únicas alternativas viables de ser implementadas, la obtención de información por medio de una encuesta. Es el hecho de aceptar el proceso de encuesta lo que introduce incertidumbre, no la estadística misma. Sin embargo, note que la asociación de estadística con mentiras se da porque se argumenta que es la estadística como disciplina la que no proporciona respuestas con 100 % de certeza.

Los modelos estadísticos son modelos matemáticos, y como cualquier modelo matemático, la corrección de las respuestas que se obtengan dependen de que los modelos sean adecuados y válidos. Por ejemplo, si para calcular la altura de un árbol situado sobre un terreno plano con pendiente, se resuelve un triángulo rectángulo, la respuesta será errónea. Esto de ninguna manera significa que la trigonometría haya mentido. Un libro titulado «Cómo mentir con trigonometría» adquiriría una interpretación absurda, que curiosamente no adquiere de la misma manera el título «Cómo mentir con estadística». En este tenor, retomamos una frase que alguna vez escuchamos (G.P. Patil), que nos parece muy pertinente: «Es muy fácil mentir con estadística. Pero es mucho más fácil, mentir sin ella».

La estadística propiamente interpretada en su concepción más amplia,

procede a cuantificar la incertidumbre en la que se incurre. Es decir, la estadística no rehuye a la incertidumbre; por el contrario, la reconoce de entrada como una realidad, la enfrenta directamente y procura cuantificarla.

También se ha oído decir que «la estadística no es una ciencia exacta, porque las respuestas que da son aproximadas». Esto también es un error muy grave de concepción entre lo que significa dar respuestas, ser científico, ser exacto, y ser aproximado. Dicha afirmación denota que no se ha entendido lo que sí hace la estadística y lo que no hace. Ante un escenario provisto de incertidumbre, es imposible producir una respuesta con entera certeza. Pero esto de ninguna manera significa que no puede tomarse una actitud científica al respecto, o que no pueda invocarse una ciencia exacta para ello (matemáticas).

Efectivamente, la palabra aproximación puede tener cabida, en la medida en que cualquier modelo matemático representa o aproxima una realidad utilizando conceptos abstractos. Por otra parte, una respuesta incierta no significa lo mismo que una respuesta aproximada. En este sentido, tras un arreglo de palabras, podríamos decir que la estadística versa sobre «modelos matemáticos que aproximan los aspectos aleatorios de la realidad, y que tienen por objeto cuantificar de manera exacta la incertidumbre en que se incurre cuando se dan respuestas», lo cual modifica sustancialmente la afirmación anterior.

7.3. Estadísticas

Sean $X_1, \ldots, X_n$ variables aleatorias, que representan el proceso de obtención de información (datos), por medio de la observación de realizaciones empíricas de un fenómeno aleatorio.

Definición 7.3 (Estadística) Una función $T\colon \mathbb{R}^n \to \mathbb{R}$ evaluada en $(X_1, \ldots, X_n)$ se llama una estadística.

Note que una estadística es una variable aleatoria porque los datos son variables aleatorias, y que no depende de parámetros desconocidos. Depende sólo de los datos; disponibles los datos, el valor de una estadística se puede calcular.

Ejemplo 7.9 $T(x_1, \ldots, x_n) = n^{-1} \sum_{i=1}^{n} x_i$ da lugar a una estadística, que antes habíamos llamado media muestral, $\bar{X}_n$.

Ejemplo 7.10 Los momentos muestrales son estadísticas.

Ejemplo 7.11 $T(X_1, \ldots, X_n) = X_1$ es una estadística.

Ejemplo 7.12 $T(X_1, \ldots, X_n) = \text{máx}(X_1, \ldots, X_n)$ y $T(X_1, \ldots, X_n) = \text{mín}(X_1, \ldots, X_n)$ son estadísticas.

Ejemplo 7.13 Sea $x \in \mathbb{R}$, y defina

$$T(x_1, \ldots, x_n) = n^{-1} \sum_{i=1}^{n} 1_{(-\infty, x]}(x_i).$$

Esto da lugar a una estadística que antes habíamos llamado la función de distribución empírica en el punto x (ver Definición 5.18).

Ejemplo 7.14 $T(X_1, \ldots, X_n) = (X_i - \mu)/\sigma$ no es estadística (a menos de

que los valores de μ y σ se conozcan), pero

$$T(X_1, \ldots, X_n) = \frac{X_i - \bar{X}_n}{\sqrt{S_n}}$$

sí lo es (S_n es la desviación estándar muestral).

7.4. Distribuciones muestrales

Una estadística T es una variable aleatoria. Por lo tanto, tiene sentido hablar de la distribución de T, es decir, de la medida de probabilidad $\mathbb{P}_T(B) = \mathbb{P}(T \in B)$. También tiene entonces sentido hablar de momentos de T, es decir, media de T, varianza de T, *etc*.

Definición 7.4 (Distribución muestral) La distribución de una estadística T recibe el nombre de *distribución muestral* de T.

La distribución de T depende, entre otras cosas, de cuál sea la distribución de los datos $X_1, \ldots, X_n$, así como de n. En particular, si la distribución de los datos depende de algún parámetro, la distribución muestral de T en general dependerá del parámetro. Una estadística no depende del parámetro, pero la distribución muestral de la estadística sí puede depender de parámetros.

Ejemplo 7.15 Sean X_1, X_2 independientes, cada una con distribución discreta dada por la densidad $f(x) = 1/3$ para $x = 1, 2, 3$. Sea $T(X_1, X_2) = X_1 + X_2$. Notemos que T es una variable aleatoria discreta, con soporte $\{2, 3, 4, 5, 6\}$. La densidad de T, es decir, la distribución muestral de T, está

dada por

$$
\begin{aligned}
f_T(2) &= \mathbb{P}(X_1 = 1, X_2 = 1) = (1/3)(1/3) = 1/9, \\
f_T(3) &= \mathbb{P}(X_1 = 1, X_2 = 2) + \mathbb{P}(X_1 = 2, X_2 = 1) = \\
&\quad (1/3)(1/3) + (1/3)(1/3) = 2/9, \\
f_T(4) &= \mathbb{P}(X_1 = 1, X_2 = 3) + \mathbb{P}(X_1 = 2, X_2 = 2) + \\
&\quad \mathbb{P}(X_1 = 3, X_2 = 1) = \\
&\quad (1/3)(1/3) + (1/3)(1/3) + (1/3)(1/3) = 3/9, \\
f_T(5) &= \mathbb{P}(X_1 = 2, X_2 = 3) + \mathbb{P}(X_1 = 3, X_2 = 2) = \\
&\quad (1/3)(1/3) + (1/3)(1/3) = 2/9,
\end{aligned}
$$

y

$$f_T(6) = \mathbb{P}(X_1 = 3, X_2 = 3) = (1/3)(1/3) = 1/9.$$

El valor esperado de T es entonces

$$\mathbb{E}(T) = 2(1/9) + 3(2/9) + 4(3/9) + 5(2/9) + 6(1/9) = 4.$$

7.4.1. Distribución muestral de $\bar{X}_n$

Veremos en esta sección algunos resultados importantes acerca de la distribución muestral de la estadística $\bar{X}_n$. Los siguientes resultados dan todos alguna propiedad acerca de la distribución de la variable aleatoria $\bar{X}_n$, con distintas suposiciones acerca de la distribución de los ingredientes $X_1, \ldots, X_n$.

Teorema 7.1 *Si $X_1, \ldots, X_n$ son variables aleatorias cada una con media μ y varianza σ^2, entonces $\mathbb{E}(\bar{X}_n) = \mu$ y $\mathbb{V}(\bar{X}_n) = \sigma^2/n$.*

Esto (Ejercicio 5.20) dice que la media del promedio coincide con la media de los ingredientes. Note además que un corolario de esto es $\mathbb{V}(\bar{X}_n) \to 0$ cuando $n \to \infty$, es decir, que a medida que aumenta n, la distribución de $\bar{X}_n$ concentra probabilidad cada vez más y más en torno a μ, lo cual no es más que lo que concluye la Ley de los Grandes Números (Teorema 5.5).

Teorema 7.2 *Si $X_1, \ldots, X_n$ son variables aleatorias i.i.d. cada una con distribución $\mathcal{N}(\mu, \sigma^2)$, entonces $\bar{X}_n$ tiene distribución $\mathcal{N}\big(\mu, \sigma^2/n\big)$.*

Esto dice que el promedio de normales es también normal. La conclusión del Teorema 7.1 se replica, al notar que si $\bar{X}_n \sim \mathcal{N}\big(\mu, \sigma^2/n\big)$, entonces $\mathbb{E}(\bar{X}_n) = \mu$ y $\mathbb{V}(\bar{X}_n) = \sigma^2/n$. Note que el Teorema 7.2 tiene hipótesis más fuertes que el Teorema 7.1. Note además que no es cierto en lo general que el promedio permanece cerrado en cuanto a su distribución genérica; véase por ejemplo el Ejercicio 7.5, en el que los ingredientes son Bernoulli, pero el promedio no es Bernoulli (ni siquiera es binomial).

Recuerde una propiedad de la distribución normal, que dice que si $X \sim \mathcal{N}(\mu, \sigma^2)$, entonces $a + bX \sim \mathcal{N}\big(a + b\mu, \sigma^2 b^2\big)$. Aplicando este resultado, se observa que si $\bar{X}_n \sim \mathcal{N}\big(\mu, \sigma^2/n\big)$, entonces $(\bar{X}_n - \mu)/(\sigma/\sqrt{n}) = \sqrt{n}(\bar{X}_n - \mu)/\sigma \sim \mathcal{N}(0, 1)$. A la cantidad $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ se le llama la *estandarización* de $\bar{X}_n$. Note que estandarizar significa restar la media y dividir entre la desviación estándar. Note entonces que la conclusión del teorema

es equivalente a decir $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ tiene distribución $\mathcal{N}(0,\, 1)$.

Teorema 7.3 (Teorema Central del Límite) *Si $X_1, \ldots, X_n$ son variables aleatorias i.i.d. con media μ y varianza $\sigma^2 < \infty$, y n es grande, entonces la distribución de $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ es aproximadamente $\mathcal{N}(0,\, 1)$. Más precisamente, cuando $n \to \infty$, $\forall x$ se cumple*

$$P\big((\bar{X}_n - \mu)/\sigma \leq x\big) \to \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du.$$

Este Teorema Central del Límite (TCL) es de suma importancia en teoría estadística. Note que en las hipótesis no se pide que las variables sean normales. De hecho, no hay suposición acerca de la distribución de las variables excepto por su media y su varianza; la conclusión es válida si las variables son exponenciales, Poisson, binomiales, normales, uniformes, o con cualquier otra distribución (continua o discreta). Esto dice que el promedio de muchas variables i.i.d. es aproximadamente normal, sin importar su distribución original. Si las variables de entrada fueran normales, el Teorema 7.2 lo que dice es que en el Teorema 7.3 se verifica aceptando que «es exactamente» es un caso particular de «es aproximadamente».

Note que decir «la distribución de $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ es aproximadamente $\mathcal{N}(0,\, 1)$» es equivalente a decir «la distribución de $\bar{X}_n$ es aproximadamente $\mathcal{N}(\mu,\, \sigma^2/n)$».

Observación 7.1 Resumimos teoremas acerca de la distribución de $\bar{X}_n$ de la siguiente manera: Siempre se cumple $\mathbb{E}(\bar{X}_n) = \mu$ y $\mathbb{V}(\bar{X}_n) = \sigma^2/n$. Más

aun, si las X_i's son normales, entonces $\bar{X}_n$ es exactamente $\mathcal{N}(\mu,\, \sigma^2/n)$, y si las X_i 's no son normales y n es grande, entonces $\bar{X}_n$ es aproximadamente $\mathcal{N}\big(\mu,\, \sigma^2/n\big)$.

Un corolario importante del TCL es que automáticamente se obtiene un resultado acerca de la distribución muestral de otra estadística, $\sum_{i=1}^n X_i$. Se sigue de notar que $\sum_{i=1}^n X_i = n\bar{X}_n$.

Teorema 7.4 *Si $X_1, \ldots, X_n$ son variables aleatorias i.i.d. con media μ y varianza $\sigma^2 < \infty$, y n es grande, entonces la distribución de $\sum_{i=1}^n X_i$ es aproximadamente $\mathcal{N}\big(n\mu,\, n\sigma^2\big)$.*

Ejemplo 7.16 Suponga que $X_1, \ldots, X_n$ son variables aleatorias i.i.d. cada una con distribución $\mathrm{U}(0, 1)$. Si n es grande, entonces $\bar{X}_n$ es aproximadamente $\mathcal{N}\big(1/2,\, 1/(12n)\big)$. Si n no es grande, entonces —por lo menos con lo teoremas vistos aquí— no podemos decir nada acerca de la distribución de $\bar{X}_n$, excepto por el hecho de que ésta es una distribución que cumple $\mathbb{E}(\bar{X}_n) = 1/2$ y $\mathbb{V}(\bar{X}_n) = 1/(12n)$, cualquiera que sea el valor de n.

Ejemplo 7.17 Suponga que $X_1, \ldots, X_n$ son variables aleatorias i.i.d. cada una con distribución $\mathrm{P}(5)$, con n grande. ¿Cuál es la probabilidad de que el promedio de las n variables diste en más que 1 unidad de su media? Como $\mathbb{E}(\bar{X}_n) = 5$, lo que se busca es $\mathbb{P}(\big|\bar{X}_n - 5\big| > 1) = 1 - \mathbb{P}(4 \leq \bar{X}_n \leq 6)$. Pero por el TCL, $\mathbb{P}(4 \leq \bar{X}_n \leq 6) \approx \mathbb{P}\big(\sqrt{n}(4-5)/\sqrt{5} \leq Z \leq \sqrt{n}(6-5)/\sqrt{5}\big) = \mathbb{P}\big(-\sqrt{n}/\sqrt{5} \leq Z \leq \sqrt{n}/\sqrt{5}\big) = \Phi(\sqrt{n}/\sqrt{5}) - \Phi(-\sqrt{n}/\sqrt{5})$. Como

ejercicio, encuentre valores (usando calculadoras de normal estándar) de la probabilidad que se pide para $n = 30, 50, 100$.

Recordemos que con LGN, ya sabíamos que la probabilidad de $\bar{X}_n$ se concentra alrededor de μ. Una observación importantísima obtenida del ejemplo anterior es: Con el TCL, si n es grande, ahora puede calcularse la probabilidad de que $\bar{X}_n$ resulte estar en una vecindad de μ. Es decir, podemos ahora cuantificar la posibilidad de que $\bar{X}_n$ caiga cerca de μ.

Ejemplo 7.18 Suponga que una compañía aérea sabe que las piezas de equipaje que portan los pasajeros tienen un peso que tiene media μ y desviación estándar σ. Suponga que para un vuelo, se documentan 123 piezas de equipaje. Si usted ha viajado en avión, notará que durante el proceso de documentación, se presta suma atención al peso del equipaje. Sean $X_1, \ldots, X_{123}$ los correspondientes pesos en kilogramos de las piezas de equipaje. Para calcular parámetros de vuelo, los pilotos requieren de conocer el peso de la carga, y en particular, saber si no excede en peso cierto máximo L. Es decir, es relevante saber si el peso total de equipaje, $\sum_{i=1}^{123} X_i$, excede la cantidad L. El TCL sirve para calcular la probabilidad de que lo anterior ocurra, de la siguiente manera: Sea Y una variable aleatoria con distribución $\mathcal{N}\big(123\mu,\, 123\sigma^2\big)$ y Z una variable normal estándar. Entonces $\mathbb{P}\big(\sum_{i=1}^{123} X_i > L\big) \approx \mathbb{P}(Y > L) = \mathbb{P}\big((Y - 123\mu)/\sqrt{123}\sigma > (L - 123\mu)/\sqrt{123}\sigma\big) = 1 - \mathbb{P}\big(Z \leq (L - 123\mu)/\sqrt{123}\sigma\big) = 1 - \Phi\big((L - 123\mu)/\sqrt{123}\sigma\big)$. Note que con el TCL, no obstante el que no se hayan pesado las maletas y que no se explicite la distribución de las X_i's, se puede determinar el riesgo de exceder el límite L

usando una probabilidad normal.[2]

Ejercicios

7.1 Determine si para abordar las siguientes situaciones, se debe plantear una prueba de hipótesis o un problema de estimación.

(a) Un ingeniero de producción desea determinar si un nuevo adhesivo tiene mejores propiedades de adherencia que el adhesivo que actualmente se usa. El nuevo adhesivo es más caro que el anterior, por lo que el ingeniero no desea recomendarlo al menos que su superioridad sea exhibida con evidencia experimental. Se harán pruebas de adherencia sobre varias muestras de material.

(b) Una secretaría de estado desea determinar si el índice de desempleo hoy varía con respecto al desempleo del trimestre anterior, que fue de 6 %.

(c) La misma secretaría de estado desea determinar el desempleo de hoy.

(d) Una teoría genética predice que en ciertos cruces de plantas, que la distribución de cierto rasgo específico deberá ocurrir en sus tres variantes con proporciones 0.50, 0.25, 0.25. Se hace un experimento genético para contrastar los datos obtenidos con la teoría genética, y determinar si la evidencia empírica contradice o no a la teoría.

[2] Hoy día, con tecnología de básculas electrónicas durante la documentación, literalmente se pesan todas y cada una de las piezas de equipaje. Sin embargo, este procedimiento probabilístico alguna era utilizado por practicidad. Sin embargo, aún a la fecha, a los pasajeros no se nos pesa en el momento del abordaje. En enero de 2003 ocurrió un accidente aéreo (Air Midwest Flight 5481) y la investigación posterior mostró que la causa fue un error de estimación para el peso de los pasajeros abordo de la aeronave.

(e) Se desea realizar un experimento para determinar la probabilidad de que cierto tipo de microorganismo sea inhibido por un antibiótico dado.

(f) Se desea contabilizar el número de plantas de agave que existen en la zona de denominación de origen tequila, con base en muestreo aleatorio de los campos de cultivo.

(g) Se desea determinar la vida media de cierto tipo de foco incandescente.

(h) Se desea determinar el número de kilogramos de brócoli que típicamente transporta un tipo determinado de camión de carga.

(i) Se desea estudiar el número medio de clientes que arriban a un comercio en cierto horario, con el fin de determinar si es necesario contratar un nuevo empleado o no.

7.2 Sean $X_1,\, X_2$ independientes, cada una con distribución discreta dada por la densidad $f(1) = 1/5$, $f(2) = 2/5$, y $f(3) = 2/5$. Sea $T(X_1, X_2) = (X_1 + X_2)/2$. Encuentre la distribución muestral de T. Encuentre también el valor esperado de T, y compare éste con el valor esperado de X_1 y X_2.

7.3 Sean X_1, X_2 independientes, cada una con distribución discreta dada por la densidad $f(1) = 2/5$, $f(2) = 1/5$, y $f(3) = 2/5$. Sea $T(X_1, X_2) = (X_1 + X_2)/2$. Encuentre la distribución muestral de T. Encuentre el valor esperado de T. Note que este caso es idéntico al Problema 7.2, con la diferencia de que la distribución de los datos ha cambiado.

7.4 Sean X_1, X_2 independientes, cada una con distribución discreta dada

por la densidad $f(1) = 1/5$, $f(2) = 2/5$, y $f(3) = 2/5$. Sea $T(X_1, X_2) = X_1 X_2$. Encuentre la distribución muestral de T y el valor esperado de T. Note que este caso es idéntico al Problema 7.2, con la diferencia de que la estadística ha cambiado.

7.5 Sean X_1, X_2, X_3 independientes, cada una con distribución $\mathrm{Ber}(p)$. Sea $T(X_1, X_2, X_3) = (X_1 + X_2 + X_3)/3$. Encuentre la distribución muestral de T (en este caso, la distribución muestral dependerá de p). Calcule $\mathbb{E}(T)$.

Capítulo 8

Estimación paramétrica

En este capítulo consideraremos los problemas de inferencia estadística llamados estimación paramétrica y prueba de hipótesis paramétrica. Supongamos entonces que se cuenta una muestra aleatoria $X_1, X_2, \ldots, X_n$ y algún modelo estadístico $\{f(x;\theta) \mid \theta \in \Theta\}$, donde el parámetro θ puede ser un vector. El valor de θ es desconocido. Por el momento supondremos que el modelo $\{f(x;\theta)\}$ está dado. Constituyen problemas estadísticos de muy distinta índole los que versan con preguntas tales como «¿Cuál es el modelo?», «¿El modelo $\{f(x;\theta)\}$ es correctamente especificado?», «¿Cuál modelo es mejor, $\{f(x;\theta)\}$ o $\{g(x;\eta)\}$», o «¿Cuál es el valor numérico del parámetro θ?».

Los conceptos que desarrollaremos aquí, siendo éste un curso introductorio, se identifican con ciertas actitudes que uno puede tomar ante la incertidumbre y ante los problemas de inferencia, denominados comúnmente como métodos *frecuentistas*, y en ocasiones, *clásicos*. Es importante conocer estos conceptos simplemente por el hecho de que son los que con más fuerza están radicados en las aplicaciones, y con los que los usuarios se encuentran

más familiarizados. Esto de ninguna manera indica que sean los únicos conceptos que existen, ni que sean correctos e indicados en todos los problemas de inferencia. Es también importante señalar que en la medida en que se profundiza en el estudio de estadística matemática para fines de inferencia (a nivel maestría y doctorado), que existen varios otros procedimientos y conceptos que no tenemos tiempo de tocar aquí. Quizás la noción más importante que se debe adquirir aquí es el de *pensar estadísticamente*, más que esperar empaparse de un conjunto de métodos específicos. Aquí, *pensar estadísticamente* significa que se entiendan algunas particularidades que introduce la variación y la aleatoriedad en un modelo matemático, entender que la incertidumbre acerca de una inferencia puede abordarse rigurosamente, y entender el papel que juegan en lo anterior los modelos de probabilidad.

8.1. Estimación

Como vimos, el problema de estimación es inferir valores plausibles del parámetro θ. Dos palabras deben analizarse con detenimiento aquí. Según el diccionario de español, *plausible* tiene las siguientes dos acepciones: 1) digno, merecedor de aplauso, y 2) atendible, admisible, recomendable. El sentido de estimación en estadística tiene que ver más con la segunda de estas acepciones. Por otra parte, *inferir*, en el diccionario de español, dice que tiene las siguientes acepciones: 1) sacar consecuencia o deducir una cosa de otra, 2) llevar consigo, ocasionar, conducir a un resultado, y 3) tratándose de ofensas, agravios, heridas, *etc*., hacerlos o causarlos.

Lamentablemente, estas acepciones son un tanto defectuosas para denotar lo que queremos significar en inferencia estadística. La acepción que más se parece para *plausible* es la segunda, y para *inferir* es la primera, aunque

esta última tiene un defecto grave en que se usa a su vez la palabra *deducir*. Inferir y deducir son dos acciones que estrictamente hablando, y sobre todo en un entorno de ciencia, no son sinónimos, sino más bien antónimos. En el diccionario de inglés, *inferir* significa derivar como conclusión a partir de hechos. Inferir también se relaciona estrechamente con la palabra *inducir*, que significa extraer conclusiones generales a partir de instancias particulares. En contraste, *deducir* significa derivar una conclusión a partir de lógica y de premisas universales y generales. En este sentido, la matemática es una ciencia deductiva, pero el problema mismo que la estadística procura resolver es de naturaleza inherentemente inductiva (los datos son los hechos, instancias particulares, y a partir de ellos debe extraerse una conclusión). Por otra parte, en inglés, un significado de *plausible* es «que aparece digno de ser creído».

Así, una explicación más precisa acerca de cuál es el problema de estimación en estadística podría ser: Obtener una conclusión a la luz de datos, que apunte en la dirección de señalar los valores del parámetro θ que puedan ser considerados dignos de credibilidad. De paso, y aprovechando la breve discusión anterior, podríamos mencionar que la ciencia como tal contiene entre sus métodos tanto componentes deductivas como deductivas. La estadística apoya la parte inductiva (descubrir algo nuevo de la naturaleza), y la estadística matemática lo hace con base en herramientas deductivas (matemáticas).

Hablaremos enseguida de dos actitudes diferentes que uno podría tomar para abordar el problema de estimación y producir una respuesta: estimación puntual, y estimación por intervalos.

8.2. Estimación puntual de un parámetro

Esta actitud consiste de producir un único valor para dar un solo valor plausible de θ. De aquí, el adjetivo *puntual* que se usa para denotarlo. El instrumento que se usa es una estadística $T(X_1, X_2, \ldots, X_n)$ para producir dicho valor único, que en este contexto recibe el nombre de un *estimador puntual* para θ. Recuerde que una estadística, por definición, depende únicamente de los datos, o sea que una vez seleccionada la muestra, el valor de la estadística puede calcularse por completo. Cuando se trata de estimación puntual, se utiliza la notación $\hat{\theta}_n$ para denotar el valor del estimador puntual, $\hat{\theta}_n = T(X_1, X_2, \ldots, X_n)$. Al igual que lo comentado con respecto a las estadísticas, $\hat{\theta}_n$ puede significar dos cosas diferentes dependiendo del contexto. Si se escribe $\hat{\theta}_n = \bar{X}_n$, se trata de una variable aleatoria (la media muestral), pero si se escribe $\hat{\theta}_n = 15.78$ significa que una estadística ya fue evaluada sobre los valores resultantes en la muestra, y que por lo tanto se trata de un número real. Este número se llama entonces *estimación puntual*. Algunos textos recurren a una notación que explicita lo anterior; se utiliza $x_1, \ldots, x_n$ para denotar valores numéricos ya observados de las variables aleatorias $X_1, \ldots, X_n$.

Ejemplo 8.1 Suponga $X_1, X_2, \ldots, X_n \sim \text{Ber}(p)$. Una estadística que puede ser usada como estimador puntual del parámetro p es $\hat{p}_n = n^{-1} \sum_{i=1}^{n} X_i$. Note que este estimador existe de manera conceptual aún antes de efectuar la muestra. Puedo hablar del estimador $\hat{p}_n$ como variable aleatoria. Supongamos que tras realizar el muestreo para $n = 10$, se obtienen 4 éxitos. Entonces $\hat{p}_n$ toma el valor $4/10$, y escribimos de igual manera $\hat{p}_n = 0.40$, como número real ($\hat{p}_n$ sería llamada la estimación).

Ejemplo 8.2 Un estimador puntual para el parámetro λ de una densidad de Poisson es $\hat{\lambda}_n = n^{-1} \sum_{i=1}^{n} X_i$.

Ejemplo 8.3 En general, el método de momentos es un ejemplo de una técnica para producir estimadores puntuales de los parámetros de un modelo estadístico, por lo que todos los ejemplos cubiertos en el método de momentos constituyen ejemplos de estimación puntual.

Puede haber más de un estimador puntual para un parámetro. Ante esto, el asunto es ahora cuantificar la incertidumbre o calidad de $\hat{\theta}_n$ como estimador de θ. ¿Cómo podemos valorar la cercanía que tras obtención de datos, resultará tener el valor calculado de $\hat{\theta}_n$? Es decir, ¿qué tan cerca estará $\hat{\theta}_n$ del valor (desconocido) de θ?, o equivalentemente, ¿cuál es la magnitud de $\left|\hat{\theta}_n - \theta\right|$? Es obvio que nunca podremos decir cuál es el valor numérico de $\left|\hat{\theta}_n - \theta\right|$, pues con una muestra podemos conocer una realización de $\hat{\theta}_n$ pero no conocemos θ. Sería entonces un razonamiento circular pretender calcular el valor de $\left|\hat{\theta}_n - \theta\right|$. Luego de aceptar que no podemos dilucidar cuál exactamente es el valor numérico de $\left|\hat{\theta}_n - \theta\right|$, es decir, que su valor es aleatorio y desconocido, podemos recurrir al concepto de probabilidad. En efecto, si se fija $\epsilon > 0$, un concepto útil para cuantificar la magnitud de $\left|\hat{\theta}_n - \theta\right|$ es la cantidad[1] $\mathbb{P}\big(\left|\hat{\theta}_n - \theta\right| > \epsilon\big)$.

Ahora bien, si $\mathbb{E}(\hat{\theta}_n) = \theta$, entonces la Desigualdad de Chebychev (5.4) sostiene que

$$\mathbb{P}\big(|\hat{\theta}_n - \theta| < r\sigma_{\hat{\theta}_n}\big) \geq 1 - 1/r^2, \, \forall r > 0,$$

[1]Note que el concepto de probabilidad ha sido invocado dos veces durante el proceso. Primero, durante la definición de un modelo estadístico para describir un fenómeno aleatorio, y luego, la probabilidad sirve para describir también la incertidumbre asociada con una inferencia.

donde $\sigma_{\hat{\theta}_n}$ es la desviación estándar de la variable aleatoria $\hat{\theta}_n$. Poniendo $r = \epsilon/\sigma_{\hat{\theta}_n}$, la Desigualdad de Chebychev concluye

$$\mathbb{P}\big(|\hat{\theta}_n - \theta| < \epsilon\big) \geq 1 - \sigma^2_{\hat{\theta}_n}/\epsilon^2.$$

Note que el lado derecho de esta desigualdad es una constante, que se podría calcular aunque uno no conociera θ, siempre y cuando uno sí pudiera calcular $\sigma_{\hat{\theta}_n}$. También puede concluirse por la Desigualdad de Chebychev, que

$$\mathbb{P}\big(|\hat{\theta}_n - \theta| < 2\sigma_{\hat{\theta}_n}\big) \geq 1 - 1/2^2 = 0.75, \text{ y}$$
$$\mathbb{P}\big(|\hat{\theta}_n - \theta| < 3\sigma_{\hat{\theta}_n}\big) \geq 1 - 1/3^2 = 0.89.$$

De cualquier forma, lo anterior sugiere que el número $\sigma_{\hat{\theta}_n}$ puede ser instrumental en la especificación de la precisión de $\hat{\theta}_n$ como estimador puntual de θ, y de hecho merece una definición especial:

Definición 8.1 (Error estándar) Se llama error estándar del estimador puntual a la desviación estándar de la estadística $\hat{\theta}_n$, es decir, $\sigma_{\hat{\theta}_n}$. Se le denota también por $\mathrm{ES}(\hat{\theta}_n)$ o por $\mathrm{SE}(\hat{\theta}_n)$ (del inglés, *Standard Error*).

Ejemplo 8.4 Suponga $X_1, X_2, \ldots, X_n \sim \mathrm{Ber}(p)$. Sabemos entonces, por propiedades de la densidad Bernoulli, que $\mathbb{E}(X_i) = p$ y que $\mathbb{V}(X_i) = p(1-p)$. Sea $\hat{p}_n = n^{-1}\sum_{i=1}^n X_i$. Por propiedades de la distribución muestral de la media muestral, sabemos entonces que $\forall p$ se cumple $\mathbb{E}(\hat{p}_n) = p$ y que $\mathbb{V}(\hat{p}_n) = p(1-p)/n$. El error estándar de $\hat{p}_n$ es entonces $\sigma_{\hat{p}_n} = \sqrt{p(1-p)/n}$.

Ejemplo 8.5 Suponga $X_1, X_2, \ldots, X_n \sim \mathrm{P}(\lambda)$. Sabemos entonces, por propiedades de la densidad Poisson, que $\mathbb{E}(X_i) = \lambda$ y que $\mathbb{V}(X_i) = \lambda$. Sea $\hat{\lambda}_n = n^{-1} \sum_{i=1}^{n} X_i$. Por propiedades de la distribución muestral de la media muestral, sabemos entonces que $\forall \lambda$ se cumple $\mathbb{E}(\hat{\lambda}_n) = \lambda$ y que $\mathbb{V}(\hat{\lambda}_n) = \lambda/n$. El error estándar de $\hat{\lambda}_n$ es entonces $\sigma_{\hat{\lambda}_n} = \sqrt{\lambda/n}$.

Note que en estos ejemplos, el error estándar decrece con n. En términos de precisión y de las desigualdades arriba anotadas, esto significa que entre más datos se tengan, la precisión mejora. Note que en general, el error estándar puede depender de θ, el cual no se conoce. Esto introduce un problema de circularidad en el argumento anterior, pues para cuantificar la precisión de $\hat{\theta}_n$, puede usarse $\sigma_{\hat{\theta}_n}$, pero para calcular $\sigma_{\hat{\theta}_n}$ se necesitaría θ. Por lo tanto, se motiva la siguiente definición.

Definición 8.2 (Error estándar estimado) Si $\sigma_{\hat{\theta}_n} = \sigma_{\hat{\theta}_n}(\theta)$ es el error estándar de un estimador puntual $\hat{\theta}_n$, llamamos a $\hat{\sigma}_{\hat{\theta}_n} = \sigma_{\hat{\theta}_n}(\hat{\theta}_n)$ el error estándar estimado de $\hat{\theta}_n$.

Note que el error estándar estimado es una estadística, pues depende de $\hat{\theta}_n$ y no depende ya de θ. Es usual que en aplicaciones y en paquetes computacionales de estadística, que se omite el adjetivo *estimado* y se habla de error estándar como si éste fuera $\hat{\sigma}_{\hat{\theta}_n}$.

El modo correcto de reportar un resultado usando estimación puntual, es anotando el valor numérico de $\hat{\theta}_n$ así como el del error estándar asociado. Es muy común que en aplicaciones en la práctica se utilicen estimadores puntuales, y que se reporten valores de $\hat{\theta}_n$ sin hacer referencia alguna a un

error estándar. Un estimador puntual desprovisto de una estimación de su error estándar es dramáticamente insuficiente para realizar inferencia estadística. Dos estimaciones puntuales pueden ser ambas de valor numérico 10.55, pero el error estándar de uno de ellos podría ser diez veces menor que el segundo. Obviamente, se trataría de dos situaciones muy diferentes de estimación, a pesar de que ambas llegaran a compartir el valor puntual 10.55.

Ejemplo 8.6 Suponga que $n = 50$, y que se trata de un modelo Poisson. Suponga que utilizando datos de la muestra, se calcula $\hat{\lambda}_n = 2.39$. El modo correcto de reportar un resultado de estimación puntual es $\hat{\lambda}_n = 2.39$, con error estándar $\sqrt{\hat{\lambda}_n/n} = \sqrt{2.39/50} = 0.22$. En publicaciones científicas, es usual que lo anterior se reporte así: $2.39(0.22)$. No debe confundirse con la notación 2.39 ± 0.22, que más bien, como veremos, se reserva para estimación por intervalos. La interpretación que tiene el número 0.22 se sigue de las observaciones anteriores.

8.2.1. Propiedades de estimadores puntuales

En teoría de estimación puntual, es usual que se definan propiedades deseables que debe poseer una estadística $\hat{\theta}_n$ que pretende ser usada como estimador puntual de un parámetro θ. Ejemplos de estas propiedades son las siguientes:

Definición 8.3 (Estimador insesgado) Si un estimador puntual cumple $\mathbb{E}(\hat{\theta}_n) = \theta$, $\forall \theta \in \Theta, \forall n$, el estimador se dice *insesgado* para θ.

Definición 8.4 (Estimador consistente) Si el error estándar de un estimador puntual $\hat{\theta}_n$ converge a cero cuando $n \to \infty$, decimos que el estimador es *consistente*.[2]

Ejemplo 8.7 El estimador $\hat{\lambda}_n = n^{-1} \sum_{i=1}^{n} X_i$ es insesgado y consistente para el parámetro de Poisson.

Ejemplo 8.8 El estimador $\hat{p}_n = n^{-1} \sum_{i=1}^{n} X_i$ es insesgado y consistente para el parámetro de la densidad de Bernoulli.

Ejemplo 8.9 La varianza muestral definida como

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

no es un estimador insesgado del parámetro σ^2 porque puede demostrarse que $\mathbb{E}(S_n^2) = ((n-1)/n)\,\sigma^2$, aunque sí es consistente. Un estimador insesgado de la varianza σ^2 estaría dado por

$$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Note que S_n^2 no es insesgado, pero sí es *asintóticamente insesgado*, en el sentido de que $\mathbb{E}(S_n^2) \to \sigma^2$ cuando $n \to \infty$. Algunas calculadoras de bolsillo que tienen funciones estadísticas, cuentan con dos teclas distintas para calcular varianza muestral con denominador n o con denominador $n-1$.

[2]Existen otras nociones de consistencia. Un ejemplo es que se debe cumplir $\lim_{n\to\infty} \mathbb{P}\big(\big|\hat{\theta}_n - \theta\big| > \epsilon\big) = 0$, $\forall \epsilon > 0$, lo cual es lo mismo que decir que hay convergencia en probabilidad. Esto lo habíamos denotado por $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.

En la teoría de estimación puntual, existe un capítulo llamado estimación insesgada de varianza mínima, que tiene por objeto la identificación de estadísticas para ser usadas como estimadores puntuales de un parámetro, y que cumplan minimizar su varianza (es decir, su error estándar). Un resultado matemático ilustrativo en este sentido es el siguiente: Para observaciones de una densidad de Poisson, *no existe* un estimador insesgado cuya varianza sea menor que λ/n. Es decir, la media muestral es óptima en este sentido de estimación insesgada. También existe un resultado importante llamado la Cota de Cramer-Rao, que matemáticamente establece cuál es la varianza más chica posible que puede adquirir en teoría un estimador insesgado. Esto es un ejemplo de cómo la estadística matemática sirve para optimizar, en el sentido de encontrar la mejor forma de realizar una inferencia.

8.3. Estimación por intervalos

Esta es una actitud distinta que puede tomarse para atacar el problema de estimación. En lugar de ofrecer como valor plausible de θ un único valor (una estimación puntual), se procura ofrecer un *conjunto* de valores. Desde varios puntos de vista, esto parece ser una actitud más sensata ante el desconocimiento del valor de θ. Cuando la dimensión de θ es uno, este conjunto de valores se reduce a un *intervalo*, y de aquí se desprende el término de estimación por intervalos. Procedemos primero a definir qué se entiende por un conjunto de confianza, antes de pretender buscarlos para casos específicos.

Definición 8.5 (Conjunto de confianza) Sea θ un parámetro k-dimensional, es decir, $\Theta \subset \mathbb{R}^k$. Sea $C(X_1, X_2, \ldots, X_n)$ un subconjunto de $\mathbb{R}^k$ tal

que es aleatorio en el sentido de que depende de las variables aleatorias $X_1, X_2, \ldots, X_n$. Decimos que el conjunto C es un conjunto de confianza $1-\alpha$ si se cumple

$$\mathbb{P}\big(\theta \in C(X_1, X_2, \ldots, X_n)\big) = 1-\alpha, \, \forall \theta.$$

Note que lo que es aleatorio es el conjunto, y que α es un número entre cero y uno, que coincide por la ley del complemento con la probabilidad $\mathbb{P}\big(\theta \notin C(X_1, X_2, \ldots, X_n)\big)$. Note que la confianza, $1-\alpha$, no es más que la probabilidad de que el conjunto C cubra a θ. Es común parafrasear lo anterior como «$1-\alpha$ es la probabilidad de que el parámetro caiga en el conjunto C», pero esto no es correcto. El parámetro θ es un valor fijo, que no «cae» o deja de caer en sitio alguno. Debemos hablar en términos de que el conjunto C cubra a θ, no de que θ caiga en C. El parámetro no decide ni tiene oportunidad de «caer o no caer» en C, porque θ no se mueve; el conjunto C sí tiene movimiento y por tanto la oportunidad de cubrir o no cubrir a θ.

En la práctica, uno establece cuál es la confianza que se desea producir. Típicamente se elige un valor alto para $1-\alpha$, y se procede a buscar un conjunto aleatorio C que cumpla la propiedad de cobertura. Una vez habiendo observado los datos y habiendo calculado con ellos el conjunto C y que éste ha dejado de ser aleatorio, no tiene sentido preguntarse por cuál es la probabilidad del evento $\theta \in C$, porque en esta aseveración ya no hay nada aleatorio. Por esto, si C ya es un conjunto fijo —ya calculado con los datos— decimos «tenemos confianza $1-\alpha$ de que C contenga a θ», en lugar de decir «la probabilidad de que θ esté en C es $1-a$».

Durante el curso que nos ocupa, no abordaremos más que el caso $k=1$.

En este caso de $\Theta \subset \mathbb{R}$, el concepto de conjunto de confianza podemos restringirlo a un intervalo como se indica a continuación.

Definición 8.6 (Intervalo de confianza) Sean $L(X_1, X_2, \ldots, X_n)$ y $U(X_1, X_2, \ldots, X_n)$ dos estadísticas tales que $L \leq U$. Decimos que (L, U) es un intervalo de confianza $1 - \alpha$ si se cumple

$$\mathbb{P}\big(L(X_1, X_2, \ldots, X_n) \leq \theta \leq U(X_1, X_2, \ldots, X_n)\big) = 1 - \alpha, \, \forall \theta.$$

Note que lo que es aleatorio son los extremos del intervalo, es decir, el intervalo mismo —que juega el rol del conjunto C en lo anterior— es aleatorio. La confianza, $1 - \alpha$, es la probabilidad de que el intervalo cubra a θ. No es la probabilidad de que θ «caiga» en el intervalo.

Es muy importante remarcar que en la definición de un intervalo aparece un calificador: $\forall \theta$. Esto es, sin importar cuál sea el valor de θ, la propiedad de cobertura se debe cumplir. En particular, aunque uno no conozca θ, la probabilidad de cobertura se cumple si (L, U) es un intervalo de confianza para θ. Por esta razón, el intervalo (L, U) es útil y relevante para un problema de estimación, que se origina ante una situación de desconocimiento de θ. No sería útil el concepto de intervalo de confianza si su propiedad de cobertura fuese válida sólo para *algunos* valores de θ, simplemente porque no habría modo de discernir si el valor desconocido de θ se encuentra entre los favorecidos.

El problema matemático en la práctica es entonces, determinar las dos estadísticas L y U tales que juntas cumplan la propiedad de cobertura que debe poseer un intervalo de confianza. Esto conlleva alguna manera de establecer teóricamente que la cobertura coincide con un valor predeterminado

$1-\alpha$.

¿Por qué razón no insistir en que la confianza sea siempre uno ($\alpha = 0$)? Si se toma $L = -\infty$ y $U = \infty$, es evidente que se cumple $\mathbb{P}\big(L(X_1, X_2, \ldots, X_n) \leq \theta \leq U(X_1, X_2, \ldots, X_n)\big) = 1$. Sin embargo, el intervalo $(-\infty, \infty)$ es tan ancho que no resulta ser informativo. Debe sacrificarse algo de confianza con el fin de que el intervalo se vuelva informativo. Por ejemplo, ¿cuál de las siguientes dos aseveraciones resulta ser más útil?

- Estar 100 % confiado en que $-\infty < \theta < \infty$.
- Estar 98 % confiado en que $4.6 \leq \theta \leq 6.8$.

El haber sacrificado un poco de confianza se traduce en la obtención de fronteras más informativas, es decir, más precisas. Existe entonces una negociación entre dos conceptos: confianza y precisión. Confianza alta se asocia con precisión baja. Precisión alta se asocia con confianza baja. No pueden tenerse ambas, precisión, y confianza, altas a la vez. En el caso de intervalos de confianza, la confianza se asocia con la probabilidad de cobertura, $1-\alpha$, mientras que la precisión se asocia con la longitud del intervalo (L, U).

La interpretación que tiene un intervalo de confianza es que de todos los intervalos construidos con muestras $X_1, X_2, \ldots, X_n$, una proporción $1-\alpha$ de ellos contendrán a θ y una proporción α de ellos fallarán. En la práctica, no se construyen muchos intervalos sino uno solo, ya que se observa una sola muestra. La interpretación es entonces que el intervalo único que se construyó usando la muestra dada, tiene «confianza» $1-\alpha$ de cubrir a θ. Nunca sabremos si este intervalo en particular realmente cubrió o no cubrió θ, pues para saberlo tendríamos que conocer θ. Pero matemáticamente tendríamos la certeza de que habiendo de hecho seleccionado al azar uno de todos los intervalos posibles, que se ha cubierto θ con probabilidad $1-\alpha$. Note el

concepto de repetibilidad que en necesario invocar para dotar de interpretación a un intervalo de confianza. Es por esta razón que los intervalos de confianza son conceptos llamados *frecuentistas*, en estadística matemática.

Notación 8.1 (Cuantiles de una normal) Sea $\alpha \in (0,1)$, y $\Phi(x)$ la función de distribución normal estándar. Denotamos por z_α al número que cumple $1 - \Phi(z_\alpha) = \alpha$. Esto es, si Z es una variable normal estándar, z_α cumple $\mathbb{P}(Z > z_\alpha) = \alpha$, o bien $\Phi(z_\alpha) = 1 - \alpha$.

8.3.1. Intervalos para la media de una distribución normal

Para plantear un primer ejemplo de un intervalo de confianza, supongamos que el modelo estadístico es $\{\mathcal{N}(\mu, \sigma^2) \mid -\infty < \mu < \infty\}$, donde el parámetro σ se supone conocido, de tal manera que el parámetro de interés es la media, μ. El siguiente resultado establece un intervalo de confianza para μ. Debe entenderse bien la demostración, pues en la demostración radica la base para muchos de los resultados subsecuentes.

Teorema 8.1 *Considere $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, donde σ es una constante conocida. Sean*

$$L(X_1, X_2, \ldots, X_n) = \bar{X}_n - z_{\alpha/2}\sigma/\sqrt{n},$$

y

$$U(X_1, X_2, \ldots, X_n) = \bar{X}_n + z_{\alpha/2}\sigma/\sqrt{n}.$$

Entonces (L, U) *es un intervalo de confianza* $1-\alpha$ *para* μ*.*

Demostración. Por demostrar, que $\mathbb{P}(L \leq \mu \leq U) = 1-\alpha, \forall \mu$. Por propiedades de la distribución muestral de $\bar{X}_n$ (Teorema 7.2), sabemos que $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$, y por propiedades de la distribución normal (ver relación 6.2.3), sabemos que

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Con esto, $\forall \mu$ se cumple entonces

$$\begin{aligned}
\mathbb{P}(L \leq \mu \leq U) = \mathbb{P}\big(\bar{X}_n - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\sigma/\sqrt{n}\big) = \\
\mathbb{P}\big(-z_{\alpha/2}\sigma/\sqrt{n} \leq \mu - \bar{X}_n \leq z_{\alpha/2}\sigma/\sqrt{n}\big) = \\
\mathbb{P}\big(-z_{\alpha/2}\sigma/\sqrt{n} \leq \bar{X}_n - \mu \leq z_{\alpha/2}\sigma/\sqrt{n}\big) = \\
\mathbb{P}\big(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\big) = \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = \\
1 - \frac{\alpha}{2} - \Big(\frac{\alpha}{2}\Big) = 1 - \alpha.
\end{aligned}$$

□

Hacemos varias observaciones con relación a este resultado:

- Se puede controlar la confianza $1-\alpha$, a través de la constante $z_{\alpha/2}$.
- Si la confianza crece, es decir, α disminuye, entonces $z_{\alpha/2}$ crece, con lo que se hace más ancho el intervalo (para obtener mayor confianza, se debe pagar el precio correspondiente en precisión).
- Si n crece, entonces el ancho del intervalo decrece (si hay más datos,

la precisión aumenta).

- Si σ crece, entonces el ancho del intervalo crece (si el fenómeno es muy variable, incide desfavorablemente sobre la precisión).

Dado que $\bar{X}_n$ es un estimador puntual de μ, y que el error estándar de $\bar{X}_n$ es precisamente $\sigma/\sqrt{n}$, note que el intervalo de confianza $1-\alpha$ que acabamos de comprobar, a final de cuentas es de la siguiente forma:

$$\text{(estimador puntual)} \pm z_{\alpha/2}\text{(error estándar)}.$$

Más aún, note que en este caso la distribución del estimador puntual es normal, lo cual se asocia con el hecho de que en el teorema aparezcan las cantidades $z_{\alpha/2}$. Después veremos que esto es una forma de intervalos que será común a muchas situaciones. El hecho de que haya aparecido en la formulación de un intervalo de confianza un estimador puntual es meramente incidental. Estimación puntual y estimación por intervalos son actitudes muy distintas que se toman para abordar el problema de estimación. Sin embargo, esta coincidencia muestra que estudiar estimadores puntuales puede redituar después para la construcción de intervalos de confianza.

En la demostración del Teorema 8.1, los siguientes hechos han sido cruciales:

(i) la distribución del estimador puntual ($\bar{X}_n$) es normal, y

(ii) el error estándar del estimador puntual ($\sigma/\sqrt{n}$)se conoce.

Note que el hecho de que las observaciones originales hubieran sido normales, no fue relevante excepto para concluir que $\bar{X}_n$ también es normal. Esto sugiere que si las condiciones (i) y (ii) se sustituyen por

(i') la distribución del estimador puntual es *aproximadamente* normal, y

(ii') el error estándar del estimador puntual se conoce *aproximadamente*,

que entonces podríamos construir intervalos que aproximadamente cumplen el requerimiento de cobertura. En efecto, de teoría de probabilidad conocemos resultados que provocan las condiciones (i') y (ii') en el caso en que n sea grande: El Teorema Central del Límite, y la Ley de los Grandes Números.

Por ejemplo, el siguiente resultado se obtiene al aproximar el error estándar de $\bar{X}_n$ con un estimador consistente, $S_n/\sqrt{n}$.

Teorema 8.2 *Considere $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, donde σ no necesariamente es conocida. Sean $L(X_1, X_2, \ldots, X_n) = \bar{X}_n + z_{\alpha/2} S_n/\sqrt{n}$ y $U(X_1, X_2, \ldots, X_n) = \bar{X}_n - z_{\alpha/2} S_n/\sqrt{n}$, donde $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Entonces, si n es grande, (L, U) es aproximadamente un intervalo de confianza $1-\alpha$ para μ.*

Demostración. Si n es grande, entonces por la LGN (ver Ejemplo 5.22), $S_n \approx \sigma$, con lo que la demostración anterior se cumple con aproximaciones en lugar de igualdades. □

Ejemplo 8.10 Se toman $n = 150$ observaciones de una distribución normal cuya media es desconocida. Suponga que se hacen cálculos y se encuentra que $\bar{X}_n = 4.9173$ y $S_n = 0.9567$. Un intervalo de confianza $90\,\%$ es entonces

$$\bar{X}_n \pm z_{0.10/2} S_n/\sqrt{n},$$

es decir,

$$4.9173 \pm 1.645(0.9567/\sqrt{150}),$$

$$4.9173 \pm 0.1285.$$

El intervalo de confianza $90\,\%$ para μ es $(4.7888,\ 5.0458)$.

8.3.2. Intervalos de Wald

El resultado anterior, se puede generalizar como sigue.

Teorema 8.3 (Intervalos de Wald) *Sea $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$ un estimador puntual del parámetro θ tal que la distribución de $\hat{\theta}_n$ es aproximadamente normal, y sea $\hat{\sigma}_{\hat{\theta}_n}$ una estimación consistente del error estándar de $\hat{\theta}_n$. Entonces, si n es grande,*

$$\hat{\theta}_n \pm z_{\alpha/2}\hat{\sigma}_{\hat{\theta}_n}$$

es aproximadamente un intervalo de confianza $1-\alpha$ para θ.

Demostración. Si n es grande, entonces por ser $\hat{\sigma}_{\hat{\theta}_n}$ consistente, $\hat{\sigma}_{\hat{\theta}_n} \approx \sigma_{\hat{\theta}_n}$, con lo que la misma demostración anterior se cumple con aproximaciones en lugar de igualdades. □

En los intervalos que son de la forma $\hat{\theta}_n \pm z_{\alpha/2}\hat{\sigma}_{\hat{\theta}_n}$, como los intervalos de Wald, es usual que a la cantidad $z_{\alpha/2}\hat{\sigma}_{\hat{\theta}_n}$ se le denomine *error de estimación,* o *margen de error*. Estos intervalos son simétricos alrededor de su punto central, $\hat{\theta}_n$.

8.3.3. Intervalos para medias: Muestras grandes

A continuación desarrollamos un ejemplo de aplicación de los intervalos de Wald, en el cual el TCL es el resultado que proporciona un estimador puntual aproximadamente normal.

Teorema 8.4 (Intervalo para medias: Muestra grande) *Sea $X_1, X_2, \ldots, X_n$ una muestra i.i.d. de una distribución con media μ. Si n es grande, y $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, entonces $\bar{X}_n \pm z_{\alpha/2} S_n/\sqrt{n}$ es aproximadamente un intervalo de confianza $1-\alpha$ para μ.*

Demostración. Por el TCL y porque n es grande, $\bar{X}_n$ es aproximadamente normal, y $S_n/\sqrt{n}$ es un estimador consistente del error estándar de $\bar{X}_n$. □

Ejemplo 8.11 Tomemos una muestra de $n = 100$ variables aleatorias. Se calculan con los datos, $\bar{X}_n = 0.4888$ y $S_n = 0.2945$. Un intervalo de confianza $99\,\%$ es entonces

$$\bar{X}_n \pm z_{0.01/2} S_n/\sqrt{n},$$

es decir,

$$0.4888 \pm 2.576(0.2945/\sqrt{100}), \text{ o bien}$$

$$0.4888 \pm 0.0759.$$

El intervalo para μ es $(0.4129,\ 0.5647)$, y la confianza es $99\,\%$.[3]

[3] De hecho, para elaborar este ejemplo, se generaron 100 números aleatorios con distribución uniforme en $(0, 1)$, por lo que la media verdadera es $\mu = 0.5$. En este caso sabríamos que el intervalo calculado sí fue exitoso en contener a μ, pero sólo podemos verificarlo porque *sabemos* que la media verdadera es $\mu = 0.5$. Es importante entender aquí, que de haberse

8.3.4. Intervalos para proporciones: Muestras grandes

Hay un caso particular de los intervalos de Wald que es de importancia extrema. Se trata del caso de estimación de una proporción, p, con base en observaciones $X_1, X_2, \ldots, X_n \sim \mathrm{Ber}(p)$, cuando n es grande.

Teorema 8.5 (Intervalo para proporción: Muestra grande) *Sea $X_1, X_2, \ldots, X_n$ una muestra i.i.d. de una distribución Bernoulli con parámetro p. Si n es grande, y $\hat{p}_n = \bar{X}_n$, entonces*

$$\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

es aproximadamente un intervalo de confianza $1-\alpha$ para p.

Demostración. Por el TCL y porque n es grande, $\hat{p}_n$ es aproximadamente normal, y $\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$ es un estimador consistente del error estándar de $\hat{p}_n$, el cual es $\sqrt{\frac{p(1-p)}{n}}$. □

Ejemplo 8.12 Suponga que se han encuestado con muestreo aleatorio i.i.d. a $n = 850$ electores para determinar si votarán por un candidato determinado. Sea p la proporción de votos que ese candidato ganará. Suponga que en la muestra, se encontraron 322 votos favorables al candidato. Un intervalo de confianza del $95\,\%$ está entonces dado por

$$\hat{p}_n \pm z_{0.05/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}},$$

repetido este experimento digamos 1000 veces, encontraríamos que aproximadamente 10 de los casos no contendrán al valor 0.5.

donde $\hat{p}_n = 322/850 = 0.379$, es decir

$$0.379 \pm 1.96\sqrt{\frac{0.379(1-0.379)}{850}}.$$

El intervalo para p es entonces 0.379 ± 0.033, o bien $(0.346, 0.412)$. El margen de error es 0.033, y la confianza es $95\,\%$.

Note que no es cierto que el margen de error sea igual $1 -$ confianza. Anoto a continuación una cita tomada del periódico del día 30 de mayo de 2000 (*El Correo de Hoy*), que denota que aún entre los directivos de las empresas encuestadoras, hay una tremenda confusión entre lo que significan los conceptos estadísticos:

> Eduardo Jones, subdirector de Estudios Políticos y Sociales del Centro de Estudios de Opinión, acompañado de Carlos Jiménez, investigador del Área de Estudios Políticos y Sociales, dieron a conocer los resultados y aseguraron que el margen de error de su sondeo es de apenas el 5 por ciento, por lo que tiene una confiabilidad del 95 por ciento.

8.3.5. Problemas de diseño

Un aspecto importante para considerar es que con conocimiento de estas fórmulas para construir intervalos de estimación, que uno puede realizar *diseño*. Una pregunta típica de diseño es ¿De qué tamaño debe ser la muestra para estimar p? Note que se trata de diseño porque es algo que se está considerando aun antes de obtener la muestra. Para contestar esta pregunta, una de las alternativas que existen es especificar con antelación dos cosas: la confianza requerida, y el margen de error, e, que sería aceptable. Con

estos dos números casi podría determinarse el valor necesario de n, porque el margen de error es $e = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$, de donde se obtiene

$$n = (z_{\alpha/2})^2 \frac{p(1-p)}{e^2}.$$

Para soslayar el hecho de que esta expresión depende del valor desconocido de p, existen tres alternativas:

(I) Utilizar como valor de p algún valor que se juzgue acertado o aproximadamente correcto antes de muestrear. Por ejemplo, puede ocurrir que recientemente se hizo una encuesta similar y se encontró entonces que $\hat{p}_n$ fue cercano a 0.30, por lo que quizás pueda suponerse que el valor actual de p es algo de ese orden de magnitud.

(II) Realizar muestreo previo de manera económica, con fines de determinación aproximada de p, quizás con medios diferentes de los que se emplearán en la encuesta definitiva (por ejemplo, haciendo una encuesta telefónica expedita, antes de una labor de entrevistas personales). En estudios de muestreo, un estudio preliminar obtenido para fines de diseño se llama una *muestra piloto,* y es muy común recurrir a ella.

(III) Tomar una actitud conservadora, y poner $p = 1/2$ en la fórmula anterior. Se trata de una actitud conservadora porque la cantidad $p(1-p)$ se maximiza en $p = 1/2$, de modo que el tamaño de muestra prescrito por esta elección es mayor que la que se requeriría si acaso p fuera otro valor. Ésta puede ser una solución aceptable, si no hay consideraciones de tipo económico en la obtención de una muestra de tamaño n. Por ejemplo, para estimar un valor de p en la vecindad de 0.40 se

requiere una muestra 2.6 veces mayor que la necesaria para estimar un valor de p en la vecindad de 0.10.

Similarmente, para el caso de estimación de una media μ, el margen de error es $e = z_{\alpha/2}\sigma/\sqrt{n}$, de donde se obtiene

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Para determinar el tamaño de muestra necesario para estimar μ con margen de error e y confianza $1-\alpha$, es necesario establecer el valor de σ, para lo cual puede recurrirse a las primeras dos estrategias mencionadas para estimación de proporciones (para este caso no existe un análogo a la estrategia conservadora).

8.3.6. Funciones de estimadores puntuales

Para emplear el Teorema 8.3 y poder construir intervalos de confianza de Wald, es necesario contar con un estimador puntual que sea aproximadamente normal. El siguiente resultado permite considerar estimadores puntuales que son función de estadísticas aproximadamente normales, habilitando con ello la construcción de intervalos de Wald.

Teorema 8.6 (Método delta) *Suponga que la distribución muestral de $\hat{\theta}_n$ es aproximadamente $\mathcal{N}(\theta, \sigma^2_{\hat{\theta}_n})$, y $\sigma_{\hat{\theta}_n} \approx 0$. Suponga que $g\colon \mathbb{R} \to \mathbb{R}$ es una función derivable tal que $g'(\theta) \neq 0$. Entonces $g(\hat{\theta}_n)$ es aproximadamente $\mathcal{N}\big(g(\theta), |g'(\theta)|^2\ \sigma^2_{\hat{\theta}_n}\big)$.*

Demostración. No la veremos, pero mencionaremos que intervienen en la demostración el Teorema de Taylor (de cálculo), el TCL y la LGN. □

Note que la condición $\sigma_{\hat{\theta}_n} \approx 0$, se cumple, en particular, si $\hat{\theta}_n$ es un estimador consistente de θ. Note además que el teorema está dando una expresión aproximada para $\sigma_{g(\hat{\theta}_n)}$, dada por $|g'(\theta)|\ \sigma_{\hat{\theta}_n}$. Más aun, si interesa estimar $g(\theta)$ usando el estimador puntual $g(\hat{\theta}_n)$, el método delta proporciona no sólo el error estándar sino que también asegura la distribución normal que es requerida para invocar los intervalos de Wald, que serían entonces de la forma

$$g(\hat{\theta}_n) \pm z_{\alpha/2}\hat{\sigma}_{g(\hat{\theta}_n)}.$$

Ejemplo 8.13 Para la distribución geométrica con parámetro p, vimos (Ejemplo 6.12) el estimador por el método de momentos es $\hat{p}_n = 1/\left(\bar{X}_n + 1\right) = g(\bar{X}_n)$, donde $g(x) = 1/\left(x+1\right)$. Note que $g'(x) = -1/(x+1)^2$. Sabemos, por el TCL, que $\bar{X}_n$ es aproximadamente $\mathcal{N}\big((1-p)/p,\, (1-p)/np^2\big)$. Notamos además, que si n es grande, entonces $(1-p)/np^2 \approx 0$. Luego, por el teorema, se concluye que $g(\bar{X}_n)$ es aproximadamente

$$\mathcal{N}\Big(\big(\frac{1-p}{p}+1\big)^{-1},\big(\frac{1-p}{p}+1\big)^{-4}\frac{1-p}{np^2}\Big) = \mathcal{N}\big(p,\, p^2\frac{1-p}{n}\big).$$

Por lo tanto, si n es grande, el intervalo de Wald para el parámetro p estaría dado aproximadamente por

$$\hat{p}_n \pm z_{\alpha/2}\hat{p}_n\sqrt{\frac{1-\hat{p}_n}{n}}.$$

Ejemplo 8.14 Suponga que es de interés estimar la probabilidad $\mathbb{P}(X > x)$

para una variable aleatoria exponencial, donde x es un número fijo. Sabemos entonces que $\mathbb{P}(X > x) = \exp(-\lambda x)$, donde λ es el parámetro de una exponencial. Por el método de momentos, un estimador de λ está dado por $1/\bar{X}_n$. Un estimador natural para $\mathbb{P}(X > x)$ está dado por $\exp(-\hat{\lambda}_n x) = \exp(-x/\bar{X}_n) = g_x(\bar{X}_n)$, donde $g_x(u) = \exp(-x/u)$ y $g'_x(u) = \left(x/u^2\right)\exp(-x/u)$. Ahora bien, por el TCL sabemos que $\bar{X}_n$ es aproximadamente $\mathcal{N}\left(1/\lambda,\, 1/\left(n\lambda^2\right)\right)$. Como $1/\left(n\lambda^2\right) \approx 0$ para n grande, el teorema concluye que $g_x(\bar{X}_n)$ es aproximadamente

$$\mathcal{N}\Big(g_x\left(\frac{1}{\lambda}\right),\, |g'_x\left(\frac{1}{\lambda}\right)|^2 \frac{1}{n\lambda^2}\Big) = \mathcal{N}\Big(\exp(-\lambda x),\, \lambda^2 x^2 \exp(-2\lambda x)/n\Big).$$

El intervalo de confianza de Wald es aproximadamente

$$\exp(-\hat{\lambda}_n x) \pm z_{\alpha/2}\hat{\lambda}_n x \exp(-\hat{\lambda}_n x)\sqrt{\frac{1}{n}},$$

donde $\hat{\lambda}_n = 1/\bar{X}_n$.

Hemos visto intervalos de confianza para muestras grandes, incluyendo el importante caso de estimación de proporciones. Esto es lo que abarcaremos en el presente curso introductorio. Cabe mencionar que existen muchas situaciones que no hemos cubierto con lo anterior, que tendrían que ser objeto de cursos adicionales de estadística matemática. Entre ellas, debe tomarse en cuenta que:

Observación 8.1 Los materiales anteriormente expuestos para abordar el tema de distribuciones muestrales han sido seleccionados para una situación

introductoria y común. Sin embargo, vale la pena mencionar que emergen contextos que no se amoldan a la situación $X_1, \ldots, X_n$ i.i.d., y para las cuales la teoría estadística tendría que responder con métodos y teoría *ad hoc*. Por mencionar sólo algunas de las dificultades:

- No siempre n es grande.
- No siempre $\hat{\theta}_n$ es aproximadamente normal, aunque n sí sea grande.
- No siempre $\dim(\theta) = 1$. Aun más, puede ser que $\dim(\theta)$ en efecto sea de dimensión infinita, por ejemplo, cuando el dato observado consiste de una función.
- No siempre $X_1, \ldots, X_n$ son i.i.d. Existen fenómenos que producen correlación entre observaciones sucesivas. Ejemplos radican en observación de series de tiempo.
- No siempre se conoce la distribución muestral de $\hat{\theta}_n$. Teóricamente puede ser intratable encontrar analíticamente la distribución muestral. Este tipo de problemas motiva recurrir a la computadora para fines de implementar diversos algoritmos diseñados para aproximar la distribución requerida.
- No siempre se cuenta con un modelo paramétrico como los que se han asumido en todo el capítulo. Ello da lugar a una disciplina de estadística matemática llamada *estadística no-paramétrica*, que versa sobre teoría y técnicas para realizar estimación formal sin la presunción de que el modelo estadístico consiste de una familia concreta.

Ejercicios

8.1 Considere cada uno de los estimadores por el método de momentos para cada una de las familias de distribuciones cubiertas en la Sección 6.4.3 y determine si los estimadores resultantes, vistos como estimadores puntuales, son insesgados y consistentes.

8.2 Use una calculadora (o tabla) de distribución normal estándar para comprobar que $z_{0.005} = 2.576,\ z_{0.01} = 2.326,\ z_{0.025} = 1.960,\ z_{0.05} = 1.645,\ z_{0.10} = 1.282,$ y $z_{0.25} = 0.674$. (Los valores 2.58, 1.96, 1.64 son célebres por ser frecuentemente socorridos.)

8.3 Demuestre que $\Phi(-z_\alpha) = \alpha$.

8.4 Verifique la siguiente interpretación: Si el intervalo de confianza es exitoso en cubrir a θ, entonces la distancia entre $\hat{\theta}_n$ y θ es a lo más el margen de error.

8.5 ¿Por qué razón un intervalo simétrico alrededor del estimador puntual en el Teorema 8.1? Sean β y γ dos constantes positivas tales que $\beta + \gamma = \alpha$. Note que el último paso de la demostración también se hubiera obtenido si se reemplaza $z_{\alpha/2}$ por z_β, y $-z_{\alpha/2}$ por $-z_\gamma$. Demuestre que la elección $\beta = \gamma = \alpha/2$ produce que la longitud del intervalo de confianza obtenido es mínima.

8.6 En un experimento psicológico, los individuos reaccionan A o B. El experimentador desea estimar $p =$ «proporción de gente que reacciona como A». ¿Cuántos sujetos de prueba debe incluir para estimar p con confianza $90\,\%$ con un margen de error de $4\,\%$, si

(a) sabe que p es alrededor de 0.2, y

(b) no tiene idea acerca de p?

8.7 Un investigador sabe que en una población, $\sigma = 18$. Desea estimar μ con confianza $95\,\%$ y error de estimación 2.5. ¿Qué tamaño de muestra debe emplear? ¿Si σ fuera la mitad, en cuanto se reduce el tamaño de muestra?

8.8 Suponga que $\hat{\lambda}_n = 32.86$, y que $n = 150$. Encuentre un intervalo de $95\,\%$ de confianza para la probabilidad $\mathbb{P}(X > 40)$.

8.9 Considere el modelo $\mathcal{N}(\mu, \sigma^2)$. Suponga que es de interés el valor de $k\mu^2$, donde k es una constante positiva dada. Use el método delta para encontrar un intervalo de confianza para $k\mu^2$, con base en el estimador puntual para μ dado por $\hat{\mu} = \bar{X}_n$.[4]

[4]Un ejemplo de esta situación es la siguiente: Por leyes de mecánica, se sabe que la distancia recorrida (d, en metros) en caída libre después de t segundos es $d = 9.81t^2/2$. Suponga entonces que se tienen observaciones $X_1, \ldots, X_n$ tales que son observaciones de $\mathcal{N}(t, \sigma)$. Luego, el ejercicio proporciona un intervalo de confianza para la distancia recorrida.

Capítulo 9

Pruebas de hipótesis paramétricas

Recordemos que el problema de prueba de hipótesis surge de preguntarse si el modelo de probabilidad que rige al fenómeno de interés se encuentra o no se encuentra en un conjunto preestablecido de modelos. Consideraremos el caso de hipótesis paramétricas. Sea $\{f(x;\theta) \mid \theta \in \Theta\}$ el modelo estadístico bajo consideración, y sea θ_0 el valor del parámetro (desconocido) que corresponde al fenómeno aleatorio de interés. Suponemos que el modelo estadístico está correctamente especificado, es decir que cumple $\theta_0 \in \Theta$. Sean Θ_0 y Θ_1 dos subconjuntos del espacio paramétrico (es decir, dos modelos estadísticos) tales que $\Theta_0, \Theta_1 \subset \Theta$, $\Theta_0 \cap \Theta_1 = \varnothing$. El problema es discernir la plausibilidad de Θ_0 y Θ_1, con base en información contenida en una muestra aleatoria.

9.1. Pruebas de hipótesis: Analogía directa con un juicio penal

Pretendemos aquí ilustrar conceptos básicos de pruebas de hipótesis, recurriendo a analogías que no son de estadística ni de matemáticas. En efecto, en un juicio penal en el que hay un acusado hay dos hipótesis bajo consideración:

$$H_0 : \text{inocente}$$

$$H_1 : \text{no inocente.}$$

Es un hecho de que existe incertidumbre acerca de cuál sea la verdad, por lo que un juicio debe procurar reducir la incertidumbre al máximo para emitir una resolución. El objetivo del juicio es determinar si la evidencia es suficiente para rechazar la inocencia del acusado. En este proceso existen dos posibles errores que el juez puede cometer:

$$\text{Error de Tipo I}: \text{ rechazar inocencia cuando se es inocente,}$$

$$\text{Error de Tipo II}: \text{ no rechazar inocencia cuando no se es inocente.}$$

Es importante señalar que por razones éticas, es muy claro que cometer error de Tipo I es mucho más grave que cometer error de Tipo II.

Una «prueba de la hipótesis» lo constituye el juicio. En este contexto legal, se examina evidencia, se realiza el juicio, y se da una sentencia (rechazar o no rechazar H_0).

En el contexto estadístico, se examinan datos, se realiza una prueba, y se realiza una conclusión (rechazar o no rechazar H_0).

¿Cuál es la posibilidad de cometer error del Tipo I? En el juicio, no se puede cuantificar, aunque por ser el más grave de los dos tipos de error, se espera que esta posibilidad sea remota. En la prueba estadística, se cuantifica con un concepto llamado *nivel*, y se denota por $\alpha = \mathbb{P}($ cometer error del Tipo I). La probabilidad de cometer error del Tipo II se denota por $\beta = \mathbb{P}($ cometer error del Tipo II).

¿Por qué no hacer ambos tipos de error arbitrariamente pequeños, digamos $\alpha = 0.00001$ y $\beta = 0.00000000001$? ¡Porque si uno disminuye, el otro necesariamente crece! Veamos un ejemplo en el contexto legal. Supongamos que hay dos jueces, que actúan de manera distinta:

Juez #1: Rechaza inocencia	$\Leftrightarrow$	Existe un video que muestra al acusado cometiendo el crimen
Juez #2: Rechaza inocencia	$\Leftrightarrow$	No existe coartada (testigo que declare que el acusado no estuvo presente en la escena del delito)

Es muy raro que si alguien es inocente, exista un video de su persona cometiendo el delito. Por lo tanto, el Juez #1 comete poco error del Tipo I. Pero, como existen muchísimos delitos que no están capturados en video, entonces comete muchísimo error de Tipo II.

En cambio, el Juez #2 comete poco error del Tipo II, porque si alguien es culpable será muy raro que exista un testigo que pueda dar la coartada. Pero, comete muchísimo error del Tipo I porque muchos inocentes no podrán producir coartada, por ejemplo si estaban en su casa durmiendo a solas. En estadística, los jueces serán *estadísticas de prueba*, y las sentencias se basarán en una regla muy clara que determine si H_0 debe o no rechazarse a la luz de los datos.

No es posible hacer arbitrariamente pequeños ambos tipos de error. Por lo tanto en estadística matemática se adopta la siguiente filosofía de operación:

1. Fijar α de antemano.

2. Minimizar β.

El usuario de una prueba estadística estipula entonces α, la probabilidad tolerable del peor de ambos tipos de error, y una vez controlado éste, procede matemáticamente a buscar pruebas que minimicen β. Note que esto hace que el intercambio de roles entre H_0 y H_1 no dé lugar a una situación simétrica, es decir, si se intercambian las hipótesis, no rechazar H_0 no sería del todo equivalente a rechazar H_1. La razón es que al intercambiar los roles, se invierten los errores I y II, o sea que la filosofía de operación arriba descrita, dictaría fijar β, para luego minimizar α. Equivalentemente, actuaría como si el más grave de los dos tipos de error fuera el de Tipo II.

Note que jamás se usa el verbo «aceptar». La prueba «rechaza» o «no rechaza» H_0 pero no se habla de «aceptar». El no rechazar H_0 no significa que se acepta H_0, ni el rechazar H_0 significa que se acepta H_1. El no rechazar H_0 simplemente significa que no hay evidencia en los datos en contra de H_0. Rechazar H_0 significa que hay evidencia en su contra, pero no constituye demostración ni argumento incuestionable de que H_1 sea cierta. La hipótesis H_0 representa el estado de creencia actual, y H_1 representa la alternativa con la que se contrasta H_0. En el juicio se dice «el acusado es inocente hasta que no se encuentre evidencia de lo contrario», y no se dice «el acusado es culpable mientras no demuestre su inocencia».

Otro punto muy importante que merece notar es el siguiente. Si fuera

tajantemente prohibido cometer error de Tipo I en un juicio, simple y sencillamente no habría juicios porque el único modo de garantizar que nunca se cometa error de Tipo I es no realizar juicio alguno. La sociedad y el sistema jurídico toleran un (pequeño) margen de tolerancia para cometer error de Tipo I. Igualmente, en estadística, si se insistiera en que $\alpha = 0$, las pruebas de hipótesis nunca rechazarían H_0, provocando un correspondiente valor de β inaceptablemente alto.

Note además que no se habla simplemente de «rechazar H_0» a secas, sino de «rechazar H_0 a favor de H_1», es decir, es posible que una H_0 se rechace cuando la alternativa es H_1 pero que no se rechace cuando la alternativa es H_1'.

Las pruebas de hipótesis se emplean en ciencia para evaluar evidencia experimental que sustente alguna hipótesis científica, H. En tal caso, hay dos posibilidades: colocar la hipótesis H en la nula, H_0, o colocarla en la alternativa, H_1. Si $H_0 = H$, entonces el sustento de H debe interpretarse como no rechazar H_0, y si $H_1 = H$, entonces el sustento a favor de H es rechazar H_0 a favor de H_1. Sin embargo, para sustentar una hipótesis, desde el punto de vista lógico es mucho más fuerte rechazar H_0 a favor de H que dejar de rechazar H a favor de una alternativa. La explicación es que si se pone $H_0 = H$, siempre quedará la duda en que si la razón por la cual H_0 haya podido dejarse de rechazar, es porque la alternativa que se le puso enfrente es un contrincante débil. En contraste, si H_0 representa el estado actual de la naturaleza, y la evidencia empírica muestra que H_0 debe rechazarse a favor de H, esto constituye una base mucho más sólida a favor de H. En resumen, cuando el objetivo es sustentar una hipótesis con datos experimentales, ésta debe colocarse en la alternativa, siendo un rechazo de H_0 un argumento de mayor contundencia a favor de H.

Esta diferencia de fuerza entre «no rechazar» y «rechazar» también puede ilustrarse con campeonatos de boxeo. Si H_0 es el campeón del mundo (el estado actual de la naturaleza), y H_1 es el retador, ocurre un hecho muy espectacular el día en que H_1 vence a H_0. En este caso sucede que se baja del trono al campeón mundial para ser ocupado por un nuevo pugilista, quien ahora se ostenta como el campeón mundial. En cambio, si en una pelea, el campeón mundial H_0 defiende su cetro y gana, esto constituye un argumento más débil para su carácter de campeón mundial, pues su victoria puede deberse simplemente a que su retador, H_1 era un contrincante débil. Es decir, es más contundente bajar del trono al campeón (rechazar H_0) que ganar una defensa del cetro (dejar de rechazar H_0).

A propósito de encuentros de boxeo, notemos también que aun en el caso de que H_0 sea vencido por H_1, que esto no *demuestra* que H_1 es el mejor peleador del mundo. Simplemente denota, que a partir de ahora, mientras no se demuestre lo contrario, lo es. En ciencia, hay instancias de este razonamiento. Por ejemplo, la mecánica de Newton fue alguna vez H_0, el estado actual del conocimiento, hasta que emergió un retador llamado Einstein, quien con una H_1 llamada teoría de relatividad, desbancó a la mecánica de Newton (para velocidades cercanas a la de la luz). Hoy día, la teoría de relatividad ocupa el lugar el H_0. El hecho de haber desbancado a la mecánica de Newton no *demuestra* que la teoría de relatividad es cierta. Es posible que en un futuro se descubran fenómenos que no son explicados por la teoría de relatividad, que se elabore una nueva teoría, que se contraste experimentalmente, y que se concluya rechazar relatividad a favor de la nueva teoría. Hoy día, no sabemos si existe o no esta teoría nueva, por lo que la actitud científica es que la relatividad sigue valiendo mientras no se demuestre lo contrario, es decir, que es el actual campeón del mundo (o estado actual del

conocimiento).

9.2. Definiciones básicas

Hipótesis Es una suposición acerca de los valores de los parámetros. Como se trata de una suposición fraseada en términos de los parámetros, matemáticamente una hipótesis es un subconjunto de valores del espacio paramétrico. Ejemplos de hipótesis fraseadas son «$\mu = 0$», «$\mu \leq 2$», y «$p > 1/2$», y en términos de subconjuntos se trata de $\{0\}$, $(-\infty, 2]$, y $(1/2, 1\,]$, respectivamente. Las hipótesis deben tener existencia previa, aun antes de observar datos. Los valores del parámetro formulados en las hipótesis deben de tener un significado especial que los distinga de los demás valores. Por ejemplo, $p = 1/2$ en el lanzamiento de una moneda tiene el significado especial de ser honesta, y cualquier otro valor carece de un significado tan distintivo. De hecho, la toma de datos se concibe precisamente para confrontar una hipótesis preestablecida con la realidad. Un razonamiento a la inversa, es decir, examinar alguna estructura o patrón en los datos y luego formular una hipótesis estadística, resulta ilógico e indebido, pues de hecho puede dar lugar a la inducción de conclusiones de manera circular.[1]

[1]Lo anterior es muy importante señalarlo. Es muy común cometer el error de formular una hipótesis conforme los resultados de muestreo, quizás porque se confunde el problema de estimación con el problema de prueba de hipótesis. Si tras realizar muestreo se obtiene un valor estimado puntual para la media igual a 2.5, resulta improcedente formularse la hipótesis $\mu = 2.5$. Este sería un ejemplo de una hipótesis formulada con base en una muestra. Las hipótesis son suposiciones acerca de los parámetros que tienen algún significado científico relevante, al margen de los datos que se hubieran podido obtener. Por ejemplo, la hipótesis $\mu = 2.5$ podría ser relevante si la constante 2.5 describiera la media de calibración nominal que debería tener una máquina industrial (ver Ejemplo 7.5). En ocasiones, las hipótesis planteadas pueden ser muestreos disfrazados indebidamente de hipótesis. En el Ejemplo 7.8, no es claro en la formulación si la constante 18.7 que aparece en la hipótesis, es o no es a su vez producto de un problema de estimación previamente resuelto. Si es así, el problema

Hipótesis nula (H_0) Es la hipótesis que se desea probar, la que se supone cierta hasta que los datos sugieran lo contrario. Se puede interpretar como el estado actual de la naturaleza, o la suposición válida de entrada (ver ejemplos más adelante). Se representa por un subconjunto del espacio paramétrico Θ, digamos Θ_0.

Hipótesis alternativa (H_1) Es la hipótesis a favor de la cual se rechaza H_0 si es que la evidencia lo sugiere así. Se representa por un subconjunto del espacio paramétrico, digamos $\Theta_1 \subset \Theta$, tal que $\Theta_0 \cap \Theta_1 = \varnothing$. No necesariamente se requiere que $\Theta_0 \cup \Theta_1 = \Theta$, es decir, no necesariamente se tiene que $\Theta_1 = \Theta_0^c$ (tomando el complemento respecto a Θ).

Hipótesis simple Es una hipótesis que como subconjunto de Θ, se compone de un solo elemento.

Hipótesis compuesta Es una hipótesis que como subconjunto de Θ, se compone de más de un elemento.

Notación 9.1 Una vez concebidas tanto H_0 como H_1 se recurre a la notación H_0: (especificación) *vs.* H_1: (especificación), donde las especificaciones como ciertos subconjuntos se parafrasean en términos de alguna característica de una distribución de probabilidad. Un ejemplo

sigue siendo de prueba de hipótesis. Pero la hipótesis se debería formular en términos de la *diferencia* entre dos medias (la del tratamiento original y la del tratamiento nuevo), en lugar de formularse en términos de la segunda media únicamente. En el Ejercicio 5.12 hay una situación similar, en la que concierne a la secretaría de estado que intenta determinar si el desempleo ha cambiado con respecto al valor de 6 % que se obtuvo en el trimestre anterior. Si el 6 % fue producto de estimación estadística, es un error formular la hipótesis como $p = 0.06$. Si el 6 % fue producto, por ejemplo, de haber estudiado exhaustivamente una población para determinar el desempleo, entonces la hipótesis $p = 0.06$ adquiere un significado completamente distinto.

sigue.

Ejemplo 9.1 Supongamos que μ es el parámetro que denota la media de una familia de distribuciones, y que $\mu \in \Theta = \mathbb{R}$. El que la media toma el valor cero es un ejemplo de una hipótesis legítima. Se plantea entonces $H_0 : \mu = 0$ *vs.* $H_1 : \mu \neq 0$. La nula es simple, y la alternativa es compuesta. Los subconjuntos de Θ implícitamente considerados son $\Theta_0 = \{0\}$ y $\Theta_1 = (-\infty, 0) \cup (0, \infty)$.

Ejemplo 9.2 $H_0 : p = 1/2$ *vs.* $H_1 : p > 1/2$. La nula es simple, y la alternativa es compuesta.

Ejemplo 9.3 $H_0 : p = 1/2$ *vs.* $H_1 : p = 1/4$ (no siempre se cumple que H_0 y H_1 sean complementarias). Nula y alternativa son simples.

Ejemplo 9.4 Una moneda, ¿es o no balanceada? Mientras no me digan lo contrario, y mientras ésta no se vea obviamente deformada, se supone que sí. Entonces $H_0 : p = 1/2$. En este sentido, H_0 es el estado actual de la naturaleza.

Prueba de hipótesis Una prueba de hipótesis consiste de examinar evidencia en forma de datos, para dar lugar a una de dos resoluciones posibles: rechazar H_0 a favor de H_1, o no rechazar H_0. Bajo este planteamiento, hay dos tipos de error que se pueden cometer:

Error de Tipo I Se comete cuando se resuelve rechazar H_0 a favor de H_1 siendo que H_0 es cierta.

Error de Tipo II Se comete cuando se resuelve no rechazar H_0 a favor de H_1 cuando H_1 es cierta.

Observación 9.1 La definición de errores de Tipo I y II no es simétrica, en el sentido de que si se invierten los roles de las hipótesis, entonces se invierten los tipos de error. La situación asimétrica se da porque en general, la teoría estadística considera las hipótesis de tal forma que el error de Tipo I es más grave que el error de Tipo II.

Ejemplo 9.5 ¿Un medicamento es mejor que el medicamento actual? Mientras no se demuestre lo contrario, la suposición de entrada es que el medicamento es a lo más igual que el actual. Note además otro punto: cometer error de Tipo I en este contexto implica lanzar al mercado un medicamento siendo que en la realidad no es mejor, mientras que error de Tipo II implica dejar de lanzar al mercado un mejor medicamento. Desde el punto de vista de relevancia para la salud, aquí es más grave el error de Tipo I.

Ejemplo 9.6 En un juicio, se dice que uno es inocente hasta que se demuestre lo contrario. (Se trata de la llamada presunción de inocencia.) La inocencia es una hipótesis nula. Es la suposición que se supone válida de entrada. La culpabilidad es la hipótesis alternativa. Como vimos, el error de Tipo I es más grave que el error de Tipo II.

Prueba estadística de una hipótesis Consiste de dos cosas:

i) Elegir una estadística $T = T(X_1, \ldots, X_n)$ llamada la *estadística*

de prueba, y

ii) Determinar un subconjunto de valores posibles de T, llamado la *región crítica*, C, de la prueba.

La regla a utilizar es: Rechazar H_0 si y sólo si $T \in C$. Cuando $T \in C$ se dice que la prueba es significativa, y cuando $T \notin C$ se dice que la prueba es *no significativa*. Note que una prueba estadística queda determinada por dos cosas: La estadística de prueba, T, y la región crítica, C. La región crítica no depende de la muestra $X_1, \ldots, X_n$, lo cual quiere decir que aun antes de tomar la muestra, la región crítica tiene existencia propia. Los datos intervienen para tomar o no la resolución de rechazar H_0, lo cual se realiza con la región crítica, al comparar el valor de T con el conjunto C. En resumen, una prueba de hipótesis consta de dos ingredientes, y se caracteriza por la pareja (T, C).

Ejemplo 9.7 Suponga $X_1, \ldots, X_{20} \sim \mathrm{Ber}(p)$. ¿Una moneda es legal? Las hipótesis son $H_0 : p = 1/2$ *vs.* $H_1 : p \neq 1/2$. Suponga que se adopta como estadística de prueba a $T = \sum_{i=1}^{n} X_i$. ¿Cuál es una región crítica razonable? Parecería ser razonable

$$C = \{0, 1, 2, 3, 4, 5\} \cup \{15, 16, 17, 18, 19, 20\}.$$

Note que si H_1 hubiera sido $H_1 : p > 1/2$, entonces la región crítica sensata hubiera sido entonces $C = \{15, 16, 17, 18, 19, 20\}$. Es decir, el conjunto C depende de H_1.

Función de potencia La distribución muestral de T depende de θ, y por lo tanto la probabilidad $\mathbb{P}(T \in C)$ también depende de θ. La función $\pi\colon \Theta \to \mathbb{R}$ definida por $\pi(\theta) = \mathbb{P}(T \in C)$ se llama *función de potencia* de la prueba (T, C).

Tamaño y nivel de una prueba El *tamaño* de una prueba define como la máxima probabilidad de cometer error del Tipo I. El *tamaño* está dado por $\sup_{\theta\in\theta_0} \pi(\theta)$. Por otra parte, decimos que una prueba tiene *nivel* α si su tamaño es a lo más α.

Potencia de una prueba La potencia en una alternativa $\theta \in \Theta_1$ se define como $\pi(\theta)$. En palabras, la potencia es la probabilidad de no cometer error del Tipo II. Se aplica esta nomenclatura únicamente a valores de $\theta \in \Theta_1$, porque no puede cometerse error de Tipo II en el caso $\theta \in \Theta_0$.

La teoría estadística asume la siguiente actitud para determinar una prueba (T, C). Primero, presupone que el valor de un nivel α está predeterminado. Este valor representa la cantidad de error de Tipo I que se está en la disposición de cometer. Acto seguido, se determina la estadística T y la región crítica C de tal manera que la prueba resultante tiene nivel α y tal que la probabilidad de error de Tipo II se minimiza (o equivalentemente, la potencia se maximiza).

Ejemplo 9.8 Suponga que un laboratorio desarrolla una droga para aumentar la concepción de varones. La hipótesis nula es $H_0\colon p = 1/2$, donde p es la probabilidad de concebir un varón, y la hipótesis alternativa es $H_1\colon p > 1/2$. Se toma una muestra de n sujetos de prueba a los que se les administra la droga, y se registra si concibieron o no un varón. La muestra es entonces

$X_1, \ldots, X_n$, donde cada $X_i \sim \text{Ber}(p)$. Sea

$$T_n = \sum_{i=1}^{n} X_i.$$

Es natural interpretar valores grandes de T_n como evidencia en contra de H_0 y a favor de H_1. Suponga que $n = 5$ (para poder hacer cuentas fácilmente) y que se resuelve rechazar H_0 si y sólo si $T_n \in \{4, 5\}$. El conjunto $C = \{4, 5\}$ recibe el nombre de *región crítica*, y la variable aleatoria T_n recibe el nombre de *estadística de prueba*. ¿Cuál es el nivel de esta prueba? ¿Cuál es la probabilidad de cometer error del Tipo II, bajo la suposición $p = 0.7$?

Ejemplo 9.9 En el ejemplo anterior, considere una región crítica de la forma $C = [c, n]$, utilizando la misma estadística de prueba y el mismo juego de hipótesis nula y alternativa. ¿Cuál es el valor de c que debe emplearse a manera de que el nivel de la prueba sea a lo más 0.15 ?

Ejemplo 9.10 ¿Cuál es la gráfica de la función de potencia π que corresponde al ejemplo anterior?

9.3. Pruebas de hipótesis para muestras grandes

A continuación se anotarán, a manera de formulario, pruebas de hipótesis (es decir, estadísticas de prueba y regiones críticas) para pruebas de hipótesis acerca de medias y proporciones. El enfoque será presentar de entrada las pruebas, sin derivarlas aquí a partir de teoría estadística. Posteriormente, haremos algunos comentarios acerca las propiedades matemáticas que tienen estas pruebas. En la formulación de hipótesis, interviene en

lo general una constante que define las hipótesis. Por ejemplo, las hipótesis $H_0\colon \mu = 2.5$, $H_1\colon \mu > 2.5$, $H_0\colon \mu \leq 5$, *etc*. para medias, y las hipótesis $H_0\colon p = 0.50$, $H_1\colon p > 0.25$, $H_0\colon p \leq 0.75$, *etc*. para proporciones. En lo sucesivo, usaremos la notación μ_0 y p_0 para denotar a constantes que definen hipótesis. Por ejemplo, las hipótesis anteriores pueden ser denotadas genéricamente como $H_0\colon \mu = \mu_0$, $H_1\colon \mu > \mu_0$, $H_0\colon \mu \leq \mu_0$, $H_0\colon p = p_0$, $H_1\colon p > p_0$, $H_0\colon p \leq p_0$, *etc*.

Estadísticas de prueba Como antes, $X_1, \ldots, X_n$ denota la muestra aleatoria. Recordamos que $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, $S_n = \sqrt{n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$, y que $\hat{p} = \frac{\#\text{ éxitos}}{n}$ cuando se trata de muestreo Bernoulli (en este caso, $\hat{p} = \bar{X}_n$). Las estadísticas de prueba serán:

$$T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}, \text{ para pruebas relativas a una media, y}$$
$$T_n = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \text{ para pruebas relativas a una proporción.}$$

Regiones críticas Las siguientes tablas resumen la forma de la región crítica, C, dependiendo de la forma que tienen las hipótesis nulas y alternativas, de tal manera que el nivel de las pruebas que resultan, es la constante α :

H_0	H_1	región crítica, C
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert T_n\rvert > z_{\alpha/2}$
	$\mu > \mu_0$	$T_n > z_\alpha$
	$\mu < \mu_0$	$T_n < -z_\alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	$T_n > z_\alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	$T_n < -z_\alpha$

H_0	H_1	región crítica, C
$p = p_0$	$p \neq p_0$	$\|T_n\| > z_{\alpha/2}$
	$p > p_0$	$T_n > z_\alpha$
	$p < p_0$	$T_n < -z_\alpha$
$p \leq p_0$	$p > p_0$	$T_n > z_\alpha$
$p \geq p_0$	$p < p_0$	$T_n < -z_\alpha$

Note que el valor de α es seleccionable a través de las constantes z_α o $z_{\alpha/2}$, dependiendo del caso. Los puntos de corte que definen a la región crítica se llaman *puntos críticos*. La prueba cuya región crítica es $|T_n| > z_{\alpha/2}$ recibe el nombre de *prueba de dos colas*, y las demás, *pruebas de una sola cola*. Note que el ser de una cola o de dos colas, es función de la hipótesis alternativa. Note además que si una prueba rechaza a nivel α, y $\alpha' > \alpha$, entonces la prueba también rechaza a nivel α'.

9.4. p-valores

La teoría estadística sobre la que se basa el discurso para pruebas de hipótesis, que recibe la denominación de Teoría de Neyman-Pearson, establece que dada una muestra y un nivel, α, que el resultado de la prueba es binario, en el sentido de concluir «rechazar H_0» o «no rechazar H_0». Esto es criticable. Supongamos que el valor crítico de una prueba fuera 1.96, la región crítica de la forma $T_n > 1.96$, y datos en dos situaciones diferentes dieran lugar a valores de la estadística de prueba $T_n = 2.35$ y $T_n = 7.29$. La actitud de Neyman-Pearson diría simplemente en ambos casos, «rechazar H_0», siendo que es intuitivamente claro que ambas situaciones son diferentes en alguna cualidad. En el segundo caso, se rechaza con mayor fuerza

que en el primero, y al decir sólo «rechazar H_0» no involucramos esta fuerza de la evidencia en contra de H_0. El concepto de p-valor tiene por objeto cuantificar la fuerza con la que se rechaza una hipótesis nula. Se describe a través de una probabilidad. Tiene la interpretación de ser la probabilidad de haber observado un valor «más extremoso» de la estadística de prueba que ya se observó, o bien, la probabilidad de haber rechazado H_0 sólo por azar. De esta forma, un p-valor grande denota que la evidencia en contra de H_0 es débil, y un p-valor chico denota que los datos contienen mucha evidencia en contra de H_0. En este sentido de p-valores, podríamos no hablar de *pruebas de hipótesis*, sino de *pruebas de significancia*, donde la cuantificación del concepto abstracto de significancia es el p-valor.

Sea t_n el valor ya observado y calculado de la estadística de prueba T_n. La definición de p-valores se resume en la siguiente tabla. La tercera columna, en términos de la distribución normal estándar, obedece al caso que nos ocupa de muestras grandes, en la cual la distribución relevante es la normal:

región crítica, C	p-valor	p-valor en términos de Φ
$\|T_n\| > z_{\alpha/2}$	$\mathbb{P}(\|T_n\| > \|t_n\|)$	$1 - \Phi(\|t_n\|) + \Phi(-\|t_n\|)$
$T_n > z_\alpha$	$\mathbb{P}(T_n > t_n)$	$1 - \Phi(t_n)$
$T_n < -z_\alpha$	$\mathbb{P}(T_n < t_n)$	$\Phi(t_n)$

Ejemplo 9.11 Cuando un mago nos adivina la carta, hemos hecho inconscientemente el uso de una prueba de hipótesis estadística, incluyendo el concepto de p-valor. Para el mago, hay dos hipótesis, H_0: el mago adivina por mero azar, y H_1: el mago tiene poderes que le ayudaron a adivinar. Se realiza un experimento, se observa el resultado, y el público se asombra de que el mago haya adivinado la carta. ¿Por qué razón el público se asombra,

es decir, rechaza H_0 en lugar de no rechazar H_0 y concluir que fue simple suerte que favoreció al mago? Porque el p-valor asociado es $1/52$, lo cual el público considera suficientemente chico como para interpretar que el hecho de que se haya adivinado la carta constituye evidencia experimental a favor de H_1.

Ejemplo 9.12 Si el mago nos adivina la carta, cuando el mazo consiste de sólo 3 cartas en lugar de 52, no nos asombraríamos tanto, porque el p-valor es $1/3$, lo cual es grande como para concluir que H_0 debe rechazarse contundentemente.

Los ejemplos anteriores indican que para fines de entretenimiento, un p-valor de 0.0192 es suficientemente chico como para rechazar H_0. Pero dependiendo del contexto, es posible que este valor no sea admisible. Por ejemplo, en un ensayo clínico para determinar la bondad de un tratamiento médico, un p-valor de 0.001 puede aún considerarse como grande.

Ejercicios

1. Considere cada uno de los estimadores por el método de momentos para cada una de las familias de distribuciones cubiertas en la Sección 6.4.3 y determine si los estimadores resultantes, vistos como estimadores puntuales, son insesgados y consistentes.

2. Demuestre que $\Phi(-z_\alpha) = \alpha$.

3. Verifique la siguiente interpretación: Si el intervalo de confianza es exitoso en cubrir a θ, entonces la distancia entre $\hat{\theta}$ y θ es a lo más el

margen de error.

4. En un experimento psicológico, los individuos reaccionan A o B. El experimentador desea estimar $p =$ proporción de gente que reacciona como A. ¿Cuántos sujetos de prueba debe incluir para estimar p con confianza $90\,\%$ con un margen de error de $4\,\%$, si a) sabe que p es alrededor de 0.2, y b) no tiene idea acerca de p ?

5. Un investigador sabe que en una población, $\sigma = 18$. Desea estimar μ con confianza $95\,\%$ y error de estimación 2.5. ¿Qué tamaño de muestra debe emplear? ¿Si σ fuera la mitad, en cuanto se reduce el tamaño de muestra?

6. Suponga que $\hat{\lambda} = 32.86$, y que $n = 150$. Encuentre un intervalo de $95\,\%$ de confianza para la probabilidad $\mathbb{P}(X > 40)$.

7. Considere el modelo $\mathcal{N}(\mu, \sigma^2)$. Suponga que es de interés el valor de $k\mu^2$, donde k es una constante positiva, dada. Use el método delta para encontrar un intervalo de confianza para $k\mu^2$, con base en el estimador puntual para μ, $\hat{\mu} = \bar{X}_n$.[2]

8. Determine el significado geométrico de p-valor, considerando casos de pruebas de dos colas y de una cola (izquierda y derecha).

9. Demuestre que una hipótesis se rechaza a nivel α si y sólo si el p-valor correspondiente es menor que α.

10. Demuestre que el p-valor es el nivel más chico al que se hubiera rechazado la hipótesis nula con los datos dados.

[2]Un ejemplo de esta situación es la siguiente: Por leyes de mecánica, se sabe que la distancia recorrida (d, en metros) en caída libre después de t segundos es $d = 9.81t^2/2$. Suponga entonces que se tienen observaciones $X_1, \ldots, X_n$ tales que son observaciones de $\mathcal{N}(t, \sigma^2)$. Luego, el ejercicio proporciona un intervalo de confianza para la distancia recorrida.

Referencias

Bowman, K. & Shenton, L. (1988). *Properties of Estimators for the Gamma Distribution*. New York and Basel: Marcel Dekker.

Chhikara, R. & Folks, J. (1989). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. New York and Basel: Marcel Dekker.

Consul, P. (1989). *Generalized Poisson Distributions: Properties and Applications*. New York: Marcel Dekker.

Fang, K., Kotz, S., & Ng, K. (2017). *Symmetric Multivariate and Related Distributions*. CRC Press.

Huff, D. & Geis, I. (1993). *How to Lie with Statistics*. New York: W.W. Norton.

Johnson, N., Kotz, S., & Balakrishnan, N. (1994). *Continuos Univariate Distributions*. John Wiley.

Johnson, N., Kotz, S., & Kemp, A. (1992). *Univariate Discrete Distributions*. John Wiley.

Morgan, J. P., Chaganty, N. R., Dahiya, R. C., & Doviak, M. J. (1991). Let's make a deal: The player's dilemma. *The American Statistician*, *45*(4), 284–287.

Patil, G., Boswell, M., & Joshi, S. (1984). *Dictionary and Classified Bibliography of Statistical Distributions in Scientific Work*. International Cooperative Publishing House.

Révész, P. (1978). Strong theorems on coin tossing. *Proceedings of the International Congress of Mathematicians, Helsinki*, 749–754.

Schilling, M. (1990). The longest run of heads. *The College Mathematics Journal*, *21*(3), 196–207.

Tong, Y. (1990). *The Multivariate Normal Distribution*. New York: Springer-Verlag.