Simulación e inferencia estadística

por

José Ramón Berrendero

Resumen. El progreso de los ordenadores ha permitido el desarrollo de metodologías nuevas y útiles para el análisis estadístico. En este artículo se presentan tres técnicas de estadística computacional que descansan de forma esencial en la capacidad de llevar a cabo cálculos intensivos: los contrastes de permutaciones, las técnicas de remuestreo bootstrap y los métodos de simulación basados en cadenas de Markov. Cada una de ellas motiva a su vez el interés de nuevos resultados matemáticos que sirven para determinar con precisión las condiciones bajo las que pueden aplicarse.

1. Introducción

> *It is, in fact, the coupling of the subtleties of the human brain with rapid and reliable calculations, both arithmetical and logical, by the modern computer that has estimulated the development of experimental mathematics. This development will enable us to achieve Olympian heights.*
> (Nicholas Metropolis, [17], pag. 130)

A principios del siglo XX, el concienzudo William Sealy Gosset escribió en tarjetas de cartulina la estatura y la longitud del dedo corazón de la mano izquierda de 3000 criminales, que había tomado de unos datos publicados en la revista *Biometrika*. Mezcló con cuidado todas las tarjetas y las dividió en grupos de 4. Para cada grupo y cada una de las dos variables, calculó la diferencia entre la media $\bar{x}$ de los 4 datos y la media μ de los datos de las 3000 tarjetas, dividida por la desviación típica s de los 4 datos. Con ello obtuvo, para cada variable, una muestra de 750 realizaciones del estadístico $t = (\bar{x} - \mu)/s$ para muestras de tamaño 4. Posteriormente utilizó estas realizaciones con el fin de confirmar los cálculos teóricos que había llevado a cabo para determinar la distribución de t. Gosset, un empleado de la empresa cervezera Guinness, publicó sus resultados bajo el seudónimo *Student* en un famoso artículo de 1908 ([29]). Al parecer Guinness había prohibido a sus empleados publicar sus descubrimientos después de un caso en el que otro investigador de la empresa había revelado secretos comerciales. Gosset adujo que sus resultados matemáticos no tendrían utilidad para la competencia y la empresa le permitió publicarlos, pero bajo seudónimo para evitar suspicacias entre sus compañeros.

Este es uno de los primeros ejemplos conocidos en que una simulación se utilizó para guiar la investigación y confirmar un resultado teórico en inferencia estadística. Por *simulación* entendemos el uso de realizaciones artificiales de variables

aleatorias procedentes de una distribución probabilística dada con el fin de resolver aproximadamente o ilustrar algún problema científico. Dado que la distribución t de Student (nombre con el que denominamos hoy la distribución de $\sqrt{n-1}(\bar{x}-\mu)/s$) corresponde a datos procedentes de poblaciones normales, lo que Gosset estaba haciendo era aproximar la distribución normal por la de los datos de las 3000 tarjetas y luego extraer muestras de tamaño 4 con el fin de obtener realizaciones de t bajo normalidad.

Desde los tiempos de Gosset, las aplicaciones de las técnicas de simulación a problemas de inferencia estadística no han parado de crecer, especialmente desde que todos tenemos un ordenador sobre nuestra mesa y no precisamos escribir miles de números en tarjetas de cartulina para obtener realizaciones de las variables aleatorias que nos interesan. De hecho, fue la invención de los primeros grandes ordenadores la que impulsó las técnicas de simulación (también llamadas de Montecarlo) como método general para la resolución de problemas científicos. Y fue en el campo de la física, en relación con investigaciones encaminadas a conseguir armamento nuclear. La idea surgió al analizar unos cálculos realizados por uno de los primeros ordenadores electrónicos, ENIAC, sobre datos de un problema de difusión de neutrones. La primera exposición formal de cómo se podrían aplicar los métodos se atribuye a John von Neumann en una carta a Robert Richtmyer fechada el 11 de marzo de 1947. Metropolis y Ulam, en un artículo publicado en 1949 ([19]), fueron los primeros que usaron la expresión *Métodos de Montecarlo* referida a estas técnicas. Una exposición fascinante del desarrollo de los métodos de simulación en esta época se puede encontrar en [17].

Algunas aplicaciones de la simulación en inferencia estadística son bastante inmediatas. Si estamos interesados en algún aspecto de la distribución de un estadístico (por ejemplo t en el caso de Gosset) podemos generar artificialmente miles de copias o realizaciones del mismo bajo distintas distribuciones de los datos y usar los resultados para establecer conjeturas, o valorar hasta qué punto un resultado asintótico es útil para aproximar el comportamiento del estadístico en muestras finitas. Otra aplicación obvia es la de comparar empíricamente el comportamiento de dos estimadores alternativos (por ejemplo, la media y la mediana) bajo distintos escenarios para decidir cuál es mejor.

El propósito de este artículo es presentar algunas aplicaciones algo más sutiles. En ellas, la simulación de variables aleatorias no solo se usa para evaluar un procedimiento de inferencia, sino que forma parte de la propia metodología. Esto se traduce en que, en las aplicaciones que vamos a revisar aquí, sin simulación no sería posible llevar a cabo la inferencia salvo en casos muy particulares. Hay algo paradójico en el hecho de que para hacer inferencia en una situación incierta sea útil añadir, generando artificialmente variables aleatorias, aún más incertidumbre. El principio básico que subyace en los ejemplos que vamos a presentar es aprovechar las regularidades del azar, tal y como se expresan en las leyes de los grandes números y los teoremas centrales del límite, para aproximar cantidades deterministas, normalmente integrales, mediante los valores resultantes de un experimento aleatorio. Introduciremos métodos en los que se aplican diferentes variantes de esta idea y que han desempeñado un papel relevante en la actividad investigadora de las últimas

décadas en estadística.

El análisis de ejemplos que muestran la interacción entre los métodos de simulación y los de inferencia puede resultar revelador para profundizar en la relación entre estadística y matemáticas. En la segunda mitad del siglo XX, la rápida evolución de las herramientas computacionales ha impulsado el desarrollo de metodologías estadísticas basadas en cálculo intensivo, en cuya aplicación el análisis matemático pasa a un segundo plano. Pero, por otra parte, las técnicas basadas en computación intensiva han motivado el interés por nuevos resultados matemáticos que permiten delimitar las condiciones precisas bajo las cuales se pueden aplicar, por lo que los efectos finales sobre el papel que desempeñan las matemáticas en estadística son ambiguos. Los tres métodos que vamos a presentar aquí (contrastes de permutaciones, bootstrap y simulación basada en cadenas de Markov) pueden ayudar a valorar estos efectos.

Las tres próximas secciones están dedicadas a cada uno de los tres métodos: en la sección 2 introducimos los contrastes de permutaciones, dedicamos la sección 3 a los métodos bootstrap y la sección 4 a los métodos de simulación basados en cadenas de Markov. La sección 5 incluye algunos comentarios finales.

2. Contrastes de permutaciones

2.1. La idea básica

Los puntos de la figura 1 corresponden a las notas obtenidas por 30 estudiantes (15 alumnos y 15 alumnas) en un examen de estadística realizado en la Universidad Autónoma de Madrid. Se puede observar que la nota media de los alumnos ($\bar{x} = 4{,}6$) es inferior a la de las alumnas ($\bar{y} = 5{,}9$) de forma que la diferencia es $d_0 = \bar{x} - \bar{y} = -1{,}33$ puntos. Por otra parte, hay bastante solapamiento entre las notas de ambos grupos así que es natural preguntarse si la diferencia en las medias es significativa, es decir, si se debe a que la población de alumnos tiene media más baja que la población de alumnas o si, por el contrario, ambas poblaciones son iguales y la diferencia observada se debe a razones puramente aleatorias. El procedimiento clásico para responder a esta pregunta es llevar a cabo un contraste de la hipótesis nula de que las dos medias son iguales. Recordemos que la hipótesis nula es aquella que representa el *statu quo*, es decir, la que estamos dispuestos a admitir a menos que tengamos una fuerte evidencia empírica contra ella. La distribución de referencia de este contraste clásico es la t de Student (aquella que Gosset dedujo con tanto trabajo) y es válida si las dos poblaciones son normales. En esta sección vamos a describir un procedimiento alternativo, el contraste de permutaciones, que no requiere ninguna hipótesis paramétrica sobre las distribuciones. En la práctica este procedimiento requiere, como veremos, llevar a cabo una sencilla simulación.

La idea básica se remonta a trabajos de Eden, Yates, Fisher y Pitman publicados en los años 30 del siglo XX ([8], [11], [12] y [21]). El razonamiento es el siguiente: bajo la hipótesis nula de que las dos poblaciones son iguales, es indiferente de cuál de ellas procede cada una de las 30 observaciones disponibles. Así, la muestra observada es solo una de las $\binom{30}{15}$ posibles particiones de los 30 datos en dos grupos de 15 y cada

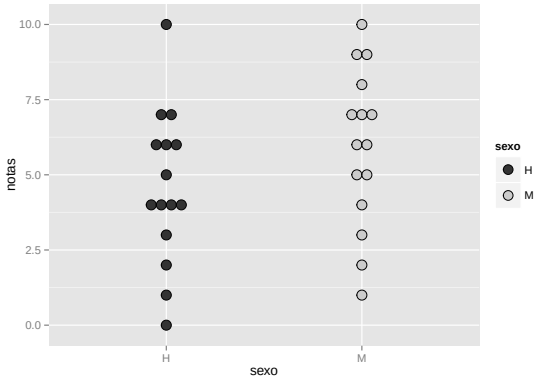

Figura 1: Notas obtenidas en un examen por un grupo de 15 hombres y otro de 15 mujeres.

una de estas particiones es probabilísticamente indistinguible de las otras. Si ahora calculamos la diferencia de medias para cada una de estas particiones obtenemos la *distribución de permutaciones* de la diferencia. Esta distribución constituye una referencia adecuada para decidir si el valor d_0 obtenido es coherente o no con la hipótesis nula. Calculamos la proporción de particiones para las que la diferencia de las medias es menor o igual que d_0. Esta proporción (la probabilidad, bajo la hipótesis nula y usando la distribución de permutaciones, de obtener un valor igual o aún más negativo que d_0) es el llamado p-valor del contraste. Si el p-valor es muy pequeño (típicamente menor que 0,05) afirmaremos que la diferencia es significativa. La razón es que, si no hubiera diferencias significativas entre las poblaciones, la partición de las 30 notas en los dos grupos correspondientes a hombres y a mujeres no tendría nada de especial respecto a cualquier otra en la que hombres y mujeres aparecen mezclados y, en consecuencia, el p-valor tendría una distribución uniforme.

Más formalmente, sea $x_1, \ldots, x_m$ la muestra de la población 1, sea $y_1, \ldots, y_n$ la de la población 2 y sea $d_0 = \bar{x} - \bar{y}$ la diferencia de medias observada. Las dos muestras se unen para formar una muestra combinada $z_1, \ldots, z_{m+n}$. Bajo la hipótesis nula de que las dos poblaciones son iguales, las $\binom{m+n}{n}$ particiones de la muestra combinada tienen la misma verosimilitud. Sea d_k la diferencia de medias observada para la $k-$ésima partición. El p-valor del contraste viene dado por:

$$p = \frac{\sum_{k=1}^{\binom{m+n}{n}} \mathbb{I}_{\{d_k \leq d_0\}}}{\binom{m+n}{n}}, \tag{1}$$

donde $\mathbb{I}_A$ denota la función indicatriz que vale 1 en A y 0 en el complementario de A. Si $p \leq \alpha$, donde α es un nivel de significación prefijado, entonces se rechaza la hipótesis nula de que las poblaciones son iguales. Es bastante sencillo comprobar que, si las dos poblaciones son iguales, la probabilidad de rechazar la igualdad de medias siendo cierta es aproximadamente α. Es exactamente α si elegimos como nivel de significación cualquiera de los valores posibles de p, es decir, $\alpha = c/\binom{m+n}{n}$, para

algún $c \in \{1, 2, \ldots, \binom{m+n}{n}\}$ (nótese que la distribución de permutaciones es discreta, aunque en la práctica este hecho es relevante solo para muestras muy pequeñas).

2.2. La simulación lo hace posible

Eden y Yates, y también Fisher, propusieron este método como posibilidad teórica, casi como experimento mental para justificar los procedimientos basados en la hipótesis de normalidad, pues pensaban que era bastante natural aplicarlo para llevar a cabo el contraste de comparación de medias. Sin embargo, los medios de cálculo de la época lo hacían inviable salvo para muestras muy pequeñas. En palabras de Fisher (el énfasis es mío), "the statistician does not carry out this very simple and very tedious process, but *his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method*" ([12], pag. 59).

Incluso con los medios de cálculo actuales el método no es factible para muestras de tamaño moderado. En el ejemplo de las notas que estamos analizando hay que considerar más de 150 millones de particiones para calcular p definido en (1). Las técnicas de simulación son un medio bastante inmediato de resolver el problema. Para aproximar el valor exacto de p la idea es seleccionar aleatoriamente un número grande (pero manejable) B de particiones y estimar el p-valor mediante

$$\hat{p} = \frac{\sum_{k=1}^{B} \mathbb{I}_{\{d_k \leq d_0\}}}{B}.$$

La idea de muestrear de la distribución de permutaciones fue utilizada inicialmente ya en [8] y propuesta explícitamente en 1957 por Dwass ([7]).

Para tener una idea de la precisión con la que $\hat{p}$ aproxima p en función de B obsérvese que, condicionando a los datos originales, la variable aleatoria $B\,\hat{p}$ tiene distribución binomial $\mathrm{Bin}(B, p)$. Por lo tanto,

$$\mathrm{E}(\hat{p}|z) = p, \quad \mathrm{Var}(\hat{p}|z) = \frac{p(1-p)}{B}.$$

Como consecuencia del teorema central del límite, un intervalo de confianza para p de nivel aproximado 95 % es $(\hat{p} - 2\hat{\sigma}_B;\ \hat{p} + 2\hat{\sigma}_B)$, donde $\hat{\sigma}_B^2 = \hat{p}(1-\hat{p})/B$. Resulta notable que un valor B varios órdenes de magnitud menor que el número total de particiones basta para alcanzar una precisión suficiente, de manera que $\hat{p}$ se puede calcular en pocos segundos con cualquier ordenador estándar.

Volviendo al ejemplo de las notas, hemos muestreado 5000 permutaciones del conjunto total de permutaciones de la muestra combinada z, y hemos calculado las correspondientes 5000 diferencias de medias (usando los m primeros valores de la permutación para representar el papel de $x_1, \ldots, x_m$, y los n últimos el de $y_1, \ldots, y_n$). Los resultados se han representado en el histograma de la figura 2. La línea vertical está situada en d_0. Vemos que no es un valor demasiado extremo si lo comparamos con la distribución de permutaciones. De hecho, se tiene $\hat{p} = 0{,}075$, con un intervalo de confianza (0,068; 0,083) a nivel 95 %. Con estos datos, no es posible afirmar que

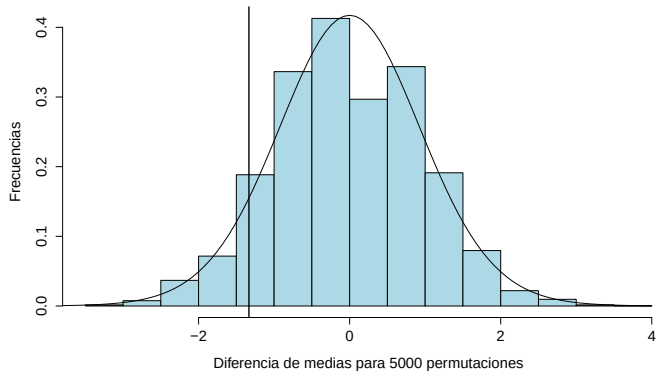

Figura 2: Distribución de permutaciones de la diferencia de medias (basada en 5000 permutaciones de los 30 datos originales). La curva es la función de densidad de la distribución t de Student con 38 grados de libertad, convenientemente normalizada.

la diferencia observada es significativa, a no ser que estemos dispuestos a asumir niveles de significación bastante altos, superiores al 7,5 %.

Como curiosidad, si se aplica el contraste clásico basado en la distribución t de Student (suponiendo varianzas poblacionales iguales) el p-valor resultante es 0,085, bastante próximo al del contraste de permutaciones. La función de densidad de la distribución de referencia del contraste clásico (una t de Student con $n+m-2=28$ grados de libertad, convenientemente normalizada) es la curva representada en la figura 2. Como vemos, es muy similar a la distribución de permutaciones, así pues es lógico que los p-valores de ambos contrastes sean parecidos.

Exactamente el mismo procedimiento que hemos descrito se puede utilizar para contrastar la hipótesis nula de que las varianzas son iguales en las dos poblaciones. En este caso se usa el cociente entre las dos varianzas muestrales, $r_0 = 0{,}945$, para comparar. Muestreamos 5000 permutaciones y calculamos la correspondiente distribución. Los resultados se han representado en la figura 3. La línea vertical corresponde al valor r_0.

En este caso se rechaza la hipótesis nula si r_0 está suficientemente lejos de 1. Por lo tanto, el p-valor es la proporción de permutaciones cuyo cociente de varianzas dista de 1 más que lo que dista r_0,

$$\hat{p} = \frac{\sum_{k=1}^{5000} \mathbb{I}_{\{|r_k-1|>|r_0-1|\}}}{5000} \approx 0{,}907.$$

El contraste clásico para comparar dos varianzas usa como referencia la distribución F de Fisher-Snedecor (con 14 grados de libertad en el numerador y en el denominador, $F_{14,14}$, para los tamaños muestrales del ejemplo) que es la representada en la figura. El correspondiente p-valor es 0,917, próximo al del contraste de permutaciones. En ambos casos no puede rechazarse que las varianzas sean iguales.

Una ventaja clara del uso de permutaciones es su generalidad. La aplicación no depende de suponer que las poblaciones tienen distribución normal y se puede adap-

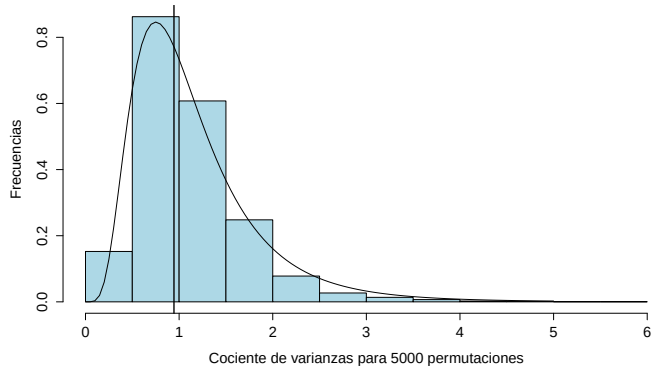

Figura 3: Distribución de permutaciones del cociente de varianzas (basada en 5000 permutaciones de los 30 datos originales). La curva es la función de densidad de la distribución $F_{14,14}$.

tar, como hemos visto, a diferentes hipótesis nulas. Por su parte, los procedimientos clásicos requieren normalidad de las poblaciones y para cada hipótesis nula se necesita deducir matemáticamente la distribución de referencia adecuada (en el ejemplo, t de Student para comparar medias y F de Fisher-Snedecor para comparar varianzas).

2.3. Comportamiento asintótico

Resulta natural preguntarse cuál es el comportamiento del contraste de permutaciones cuando las medias poblacionales son iguales (es decir, la hipótesis nula de igualdad de las medias en la que realmente estamos interesados es cierta) pero las poblaciones no son iguales. ¿Sigue siendo el nivel de significación real igual a α, al menos asintóticamente? Romano en 1990 (véase el teorema 3.1 en [23]) demostró que, si las poblaciones tienen la misma varianza o se verifica $n/(n+m) \to 1/2$, cuando $m, n \to \infty$ (es decir, los tamaños de las dos muestras son parecidos) entonces el contraste es asintóticamente válido, en el sentido de que su nivel de significación asintótico es α. Similares propiedades verifica el contraste basado en la t de Student. Así pues, dado que en nuestro ejemplo de notas las dos muestras tienen 15 datos, no es extraño que el contraste de permutaciones y el clásico den lugar a p-valores similares.

El nivel de significación real podría ser bastante diferente de α en el caso en que las varianzas poblacionales fuesen diferentes y simultáneamente las muestras estuvieran desequilibradas. Sin embargo, muy recientemente, Chung y Romano han introducido una estandarización de la diferencia de medias que resuelve este problema, es decir, proporciona un contraste asintóticamente válido incluso en este caso (véase el teorema 2.2 y la observación 2.1 de [5]).

Los resultados que hemos mencionado en este apartado son buenos ejemplos de que, si bien por una parte los métodos de computación intensiva permiten llevar a cabo inferencias sin apenas usar ningún tipo de análisis matemático, por otra parte

generan interés por nuevos resultados asintóticos cuya obtención requiere el uso de técnicas matemáticas no triviales.

3. BOOTSTRAP

3.1. EL MUNDO BOOTSTRAP

Un problema central en estadística es aproximar la distribución en el muestreo de un estadístico, esto es, conocer cómo se distribuyen los valores que toma al muestrear repetidamente de la población. Esta distribución permite valorar la precisión de un estimador y es la base de los principales procedimientos de inferencia. Aproximar empíricamente la distribución en el muestreo del estadístico z es lo que se proponía Gosset al trabajar con sus tarjetas. Con el fin de abordar este problema de una forma muy flexible, Bradley Efron ([9]) introdujo en 1979 los métodos bootstrap, en los que se combinan técnicas de simulación con un principio general (el principio *plug-in*) según el cual cualquier cantidad desconocida que dependa de una distribución F se puede aproximar reemplazando F por un estimador adecuado $\widehat{F}$ obtenido a partir de los datos. La palabra *bootstrap* alude a una de las aventuras del Barón de Münchausen, escritas en el siglo XVIII por R. E. Raspe, según la cual el Barón cayó a las aguas de un profundo lago y consiguió salir tirando de los cordones de sus botas (de donde procede la expresión en inglés *to pull oneself up by one's own bootstrap*).

Para fijar ideas consideremos la siguiente situación: $\bar{X}_n$ es la media muestral de n observaciones independientes e idénticamente distribuidas (i.i.d.) de una distribución F y queremos estimar la función de distribución del estadístico $\sqrt{n}(\bar{X}_n - \mu)$ evaluada en el punto x, es decir,

$$H_n(x) = \mathrm{P}_F\{\sqrt{n}(\bar{X}_n - \mu) \leq x\},$$

donde $\mu = \mathrm{E}_F(X)$ y el subíndice F señala la distribución de los datos bajo la que se calcula la probabilidad o la esperanza.

Supongamos que estamos dispuestos a admitir que la distribución F es, por ejemplo, exponencial de media $\mu = 1/\lambda$. Esto significa que $F = F_\lambda$, donde $F_\lambda(x) = 1 - e^{-\lambda x}$ si $x > 0$, y $F_\lambda(x) = 0$ en caso contrario. Si λ fuese conocido, podríamos generar un número grande de muestras de tamaño n de una distribución exponencial de media $\mu = 1/\lambda$, calcular para cada una de ellas el estadístico que nos interesa $\sqrt{n}(\bar{X} - \mu)$, y usar la proporción de veces que el resultado es menor o igual que x para aproximar $H_n(x)$. Como en la práctica no se conoce el parámetro, aplicamos el principio *plug-in* y reemplazamos F_λ por su estimador más natural en este caso que es $\widehat{F} = F_{\hat{\lambda}}$, donde $\hat{\lambda} = 1/\bar{X}_n$ es el estimador de máxima verosimilitud (y de momentos) de λ. La aplicación de esta sustitución lleva al estimador

$$\widehat{H}_n(x) = \mathrm{P}_{F_{\hat{\lambda}}}\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x\},$$

donde $\bar{X}_n^*$ denota la media de n observaciones i.i.d. de $F_{\hat{\lambda}}$. Nótese que $\mathrm{E}_{F_{\hat{\lambda}}}(X) = 1/\hat{\lambda} = \bar{X}_n$. Tanto esta esperanza como la probabilidad anterior han de entender-

se condicionadas a la muestra original $X_1, \ldots, X_n$. El estimador $\widehat{H}_n(x)$ se llama *estimador bootstrap paramétrico* de $H_n(x)$.

En la práctica, es necesario poder calcular $\widehat{H}_n(x)$ de forma efectiva. Obtener una expresión no es totalmente inmediato, pero al sustituir F por $F_{\hat{\lambda}}$ pasamos del duro mundo real, en el que disponemos de una única muestra de tamaño n, al mundo bootstrap en el que la distribución de la que proceden los datos es totalmente conocida y disponemos de todos los datos que nuestra capacidad de cálculo permita. Es en este punto cuando la simulación desempeña su papel. Utilizando el ordenador es posible simular muestras $X_1^{*b}, \ldots, X_n^{*b}$ de $F_{\hat{\lambda}}$, donde $b = 1, \ldots, B$ y B es un número muy grande. Para cada una de estas muestras artificiales podemos calcular su media $\bar{X}_n^{*b}$. El valor de $\widehat{H}_n(x)$ se puede entonces aproximar de la siguiente forma:

$$\widehat{H}_n(x) \approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}_{\{\sqrt{n}(\bar{X}_n^{*b} - \bar{X}_n) \leq x\}}.$$

En este ejemplo tan sencillo no hace falta aproximar por simulación. Se puede demostrar que $\widehat{H}_n(x) = G_n(x + \sqrt{n}\bar{X}_n)$, donde G_n corresponde a la función de distribución gamma con parámetro de escala $\sqrt{n}\bar{X}_n$ y parámetro de forma n. Sin embargo, la aproximación por simulación es la que confiere al método su generalidad, ya que permite al usuario despreocuparse de si es posible o no derivar una expresión cerrada de $\widehat{H}_n(x)$.

3.2. Bootstrap no paramétrico

El método que hemos descrito en el apartado anterior no es del todo satisfactorio, pues se basa en hipótesis paramétricas sobre la distribución F que pueden no ser razonables para los datos disponibles. La versión más utilizada del bootstrap, que vamos a describir en este apartado, no requiere ninguna hipótesis sobre la distribución.

Si no se hace ninguna hipótesis sobre F, el estimador más natural para aplicar el principio *plug-in* es la función de distribución empírica de la muestra, F_n, que asigna probabilidad $1/n$ a cada uno de los datos muestrales. En la figura 4 se ha representado la función de distribución exponencial F_λ, para $\lambda = 1$, junto con las funciones de distribución empíricas correspondientes a cuatro muestras de diferentes tamaños $n = 20, 50, 100, 500$ procedentes de F_λ. Se observa que la aproximación va mejorando a medida que aumenta el tamaño muestral. El teorema de Glivenko-Cantelli garantiza la convergencia uniforme casi segura de F_n a F.

Para muestras de tamaño moderado, es esperable que no haya mucha diferencia entre calcular probabilidades y esperanzas bajo F y bajo F_n puesto que ambas distribuciones se parecen, lo que conduce al *estimador bootstrap no paramétrico* (o simplemente *estimador bootstrap*) de $H_n(x)$,

$$\widehat{H}_n^*(x) = \mathrm{P}_{F_n}\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x\}.$$

Al igual que en el apartado anterior, no importa si no somos capaces de obtener una expresión cerrada de $\widehat{H}_n^*(x)$, ya que siempre es posible aproximar su valor a

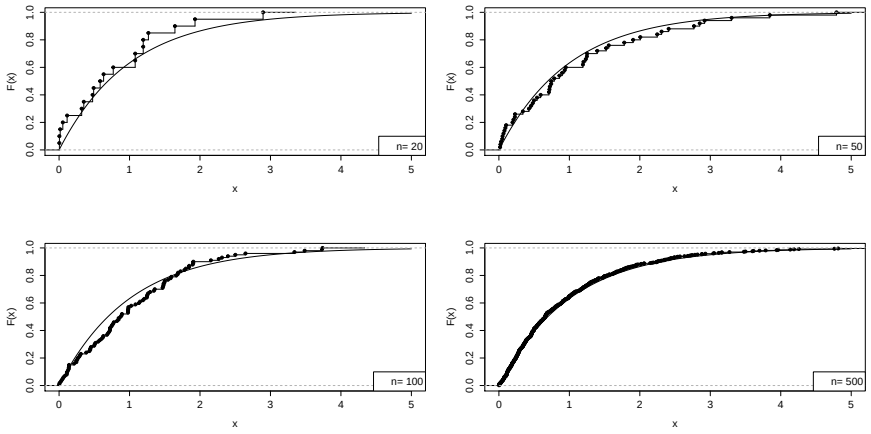

Figura 4: Aproximación de la función de distribución empírica a la función de distribución para muestras exponenciales de distintos tamaños.

partir de un número B grande de muestras artificiales de F_n. Ahora, como F_n asigna probabilidad $1/n$ a cada observación de la muestra original, para obtener las muestras simuladas basta sortear con reemplazamiento entre los datos originales $X_1, \ldots, X_n$. Es decir, las muestras artificiales y la muestra original contienen los mismos valores con la única diferencia de que en las muestras artificiales pueden estar repetidos. Las muestras $X_1^*, \ldots, X_n^*$ reciben el nombre de *muestras bootstrap o remuestras.*

En la figura 5 se han representado los resultados de aplicar este método a cuatro muestras de tamaño n (para $n = 20, 50, 100, 500$) de una distribución exponencial de media $\mu = 1$. De cada una de las cuatro muestras se obtuvieron $B = 10^4$ remuestras. Los histogramas de la figura representan los valores $\sqrt{n}(\bar{X}_n^{*b} - \bar{X}_n)$, para $b = 1, \ldots, 10^4$, en cada caso. Como se dijo más arriba, en esta situación se puede calcular la distribución exacta del estadístico cuya densidad es la curva representada en la figura. Vemos que incluso para $n = 20$ la aproximación del histograma a la densidad exacta es razonable y que si basamos nuestras inferencias sobre μ en las aproximaciones bootstrap es posible que obtengamos buenos resultados. Estas inferencias son totalmente no paramétricas en el sentido de que no requieren ningún conocimiento sobre la distribución de los datos. Tampoco necesitan ningún desarrollo teórico especial, únicamente la capacidad de simular los valores $\sqrt{n}(\bar{X}_n^{*b} - \bar{X}_n)$ extrayendo remuestras de la muestra original.

El procedimiento que hemos descrito puede llevarse a cabo de forma totalmente análoga para un estadístico arbitrario, T_n, de manera que el bootstrap permite estimar la distribución de estadísticos complejos sin hacer ninguna hipótesis sobre la distribución de la que proceden los datos y sin necesidad de ningún cálculo teórico específico.

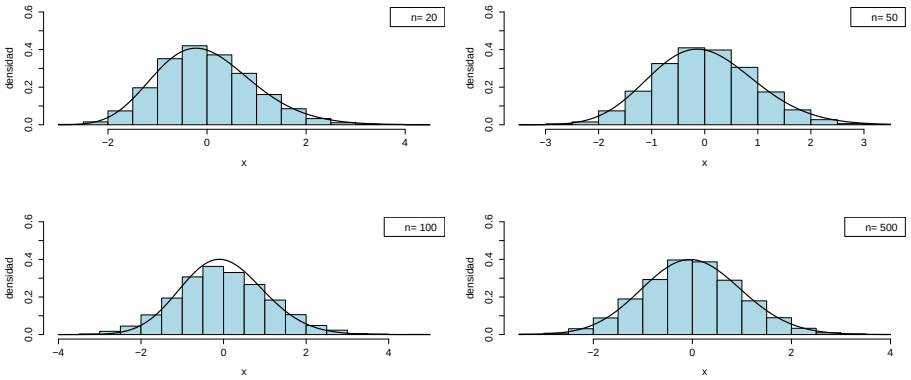

Figura 5: Aproximaciones bootstrap a la distribución de $\sqrt{n}(\bar{X}_n - \mu)$ para cuatro muestras de una distribución exponencial de media 1.

3.3. Intervalos de confianza

Una vez que hemos estimado la distribución en el muestreo de $\sqrt{n}(\bar{X}_n - \mu)$, podemos usar la estimación para deducir intervalos de confianza para μ. Existe una literatura muy amplia sobre el cálculo de intervalos de confianza mediante bootstrap (véanse, por ejemplo, los capítulos 12, 13, 14 y 22 de [10]). Aquí vamos a revisar brevemente uno de los métodos más directos, conocido en la literatura como *método híbrido*.

Para seguir con el ejemplo anterior, si la distribución $H_n(x)$ fuese conocida, entonces se podría obtener un intervalo de confianza para μ de nivel exacto $1-\alpha$ despejando μ en la ecuación siguiente:

$$1-\alpha = \mathrm{P}_F\{H_n^{-1}(\alpha/2) \leq \sqrt{n}(\bar{X}_n - \mu) \leq H_n^{-1}(1-\alpha/2)\},$$

donde $H_n^{-1}(\alpha) = \text{ínf}\{x : H_n(x) \geq \alpha\}$. El intervalo de confianza correspondiente sería

$$\big(\bar{X}_n - n^{-1/2}H_n^{-1}(1-\alpha/2),\ \bar{X}_n - n^{-1/2}H_n^{-1}(\alpha/2)\big). \tag{2}$$

Dado que H_n no es conocida, resulta natural reemplazarla por el estimador bootstrap $\widehat{H}_n^*$. En la práctica, esto requiere ordenar todos los valores simulados $\sqrt{n}(\bar{X}_n^{*b} - \bar{X}_n)$, seleccionar los percentiles que dejan una proporción de valores $\alpha/2$ a su izquierda y a su derecha, y utilizar estos valores en (2) en lugar de $H_n^{-1}(\alpha/2)$ y $H_n^{-1}(1-\alpha/2)$, respectivamente.

En la figura 6 se representa el histograma de los $B = 10^4$ valores bootstrap $\sqrt{n}(\bar{X}_n^{*b} - \bar{X}_n)$ para una muestra de tamaño $n = 50$ de una distribución exponencial de media $\mu = 1$, junto con la densidad exacta del estadístico $\sqrt{n}(\bar{X}_n - \mu)$. Los valores $H_n^{-1}(\alpha/2)$ y $H_n^{-1}(1-\alpha/2)$ de (2) se señalan con las líneas verticales discontinuas mientras que las aproximaciones bootstrap corresponden a las líneas verticales continuas. El intervalo bootstrap obtenido es (0,88; 1,46) que, para esta muestra concreta, contiene al verdadero valor del parámetro $\mu = 1$.

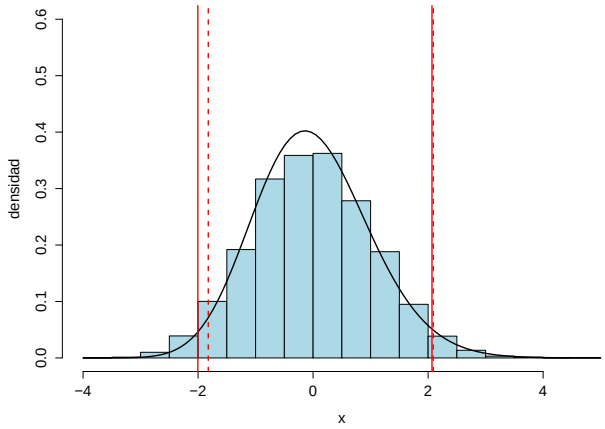

Figura 6: Valores utilizados para construir el intervalo de confianza bootstrap (líneas verticales continuas) y para el intervalo de confianza exacto (líneas verticales discontinuas).

Existen otros métodos más refinados de obtener intervalos de confianza mediante bootstrap. Por ejemplo, se puede demostrar (véase el capítulo 22 de [10]) que es preferible basar el intervalo bootstrap en una versión *studentizada* (en el sentido de ser análoga al estadístico considerado por Gosset) $\sqrt{n}(\bar{X}_n - \mu)/S_n$, donde S_n es la desviación típica muestral.

3.4. Consistencia

El proceso que lleva a la estimación bootstrap final de $H_n(x)$ involucra dos aproximaciones distintas:

$$H_n(x) \approx \widehat{H}_n^*(x) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}_{\{\sqrt{n}(\bar{X}_n^{*b} - \bar{X}_n) \leq x\}}.$$

La ley fuerte de los grandes números (aplicada a observaciones de la distribución F_n) garantiza que si $B \to \infty$ el último término converge al segundo. En esta aproximación todo ocurre dentro del mundo bootstrap y un valor $B = 10^4$ como el usado en la figura 5 suele ser suficiente para la mayoría de los propósitos. Sin embargo, la aproximación entre los dos primeros términos cuando $n \to \infty$ requiere trabajo teórico adicional. Es lo que se llama establecer la consistencia del bootstrap. En términos más intuitivos, hay que demostrar que al aumentar el tamaño muestral el mundo bootstrap se parece lo suficiente al mundo real.

Poco tiempo después del famoso artículo original de Efron se publicaron los primeros resultados de consistencia del bootstrap ([2] y [26]). Por ejemplo, el siguiente resultado, en el que se demuestra la consistencia fuerte del bootstrap respecto a la norma del supremo $\|\cdot\|_\infty$ para la distribución de la media, es el teorema 1A de [26]:

Teorema 1. *Si* $E_F(X^2) < \infty$, *entonces* $\lim_{n\to\infty} \|\widehat{H}_n^* - H_n\|_\infty \to 0$, *con probabilidad igual a 1.*

En [2] también se obtienen resultados de consistencia del bootstrap para funciones suaves de la media, cuantiles y U-estadísticos. Desde estos resultados iniciales, la teoría ha evolucionado mucho en distintas direcciones. Existen resultados para estadísticos procedentes de funcionales diferenciables ([22]) o resultados muy generales para procesos empíricos ([14] y [25]). En otra dirección, en [15] se estudia la velocidad a la que convergen los estimadores bootstrap utilizando desarrollos de Edgeworth. Un resumen bastante informativo de resultados de consistencia del bootstrap se puede encontrar en el capítulo 29 de [6].

Debido a su flexibilidad y facilidad de uso, los métodos bootstrap reemplazan en ocasiones a otros métodos asintóticos basados en complicados teoremas centrales del límite. Aunque muchas veces los resultados para muestras finitas son aceptables, no hay que olvidar que la justificación teórica del bootstrap es, en último término, también asintótica. Necesitamos información muestral suficiente para que el mundo bootstrap se parezca al mundo real. Si no es así el Barón de Münchausen no podrá salir del lago tirando de los cordones de sus botas.

4. Simulación basada en cadenas de Markov

4.1. Cadenas de Markov

En algunos problemas estadísticos, especialmente en inferencia bayesiana, aparece la necesidad de calcular momentos de vectores aleatorios con distribuciones no estándar, incluso no especificadas totalmente. En estos casos, los procedimientos basados en cadenas de Markov que vamos a describir en esta sección son especialmente útiles. El objetivo de estos métodos es diseñar una cadena de Markov cuya distribución estacionaria sea aquella de la que nos interesa simular valores. Si la cadena es ergódica, lo que significa que se cumple la ley fuerte de los grandes números, podemos aproximar los momentos de la distribución estacionaria mediante promedios adecuados de los valores de la cadena.

En el apartado siguiente describiremos el algoritmo de Metropolis-Hastings, una receta casi universal para diseñar la cadena que nos interesa. Pero antes introduciremos algunos conceptos básicos sobre cadenas. Para evitar en lo posible problemas técnicos nos ceñiremos a cadenas que toman valores en un conjunto finito, el espacio de estados $S = \{1, 2, \ldots, k\}$. Una sucesión de variables aleatorias $\{X_t : t = 1, 2, \ldots\}$ que toman valores en S es una cadena de Markov (homogénea) si para todo $i, j, i_0, i_1, \ldots, i_{t-1} \in S$,

$$\mathrm{P}\{X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \ldots, X_0 = i_0\} = \mathrm{P}\{X_{t+1} = j | X_t = i\} \equiv p_{ij}.$$

Es decir, el estado de una cadena en el futuro (etapa $t+1$) depende únicamente de su estado presente (etapa t) pero no de la trayectoria que ha conducido al estado presente. Las probabilidades p_{ij} se llaman *probabilidades de transición*. Cuando p_{ij} no depende de t, la cadena es homogénea.

Una distribución de probabilidad $\pi = (\pi_1, \ldots, \pi_k)$ sobre S es *estacionaria* para la cadena de Markov con probabilidades de transición p_{ij} si

$$\sum_{i=1}^{k} \pi_i p_{ij} = \pi_j, \quad \text{para todo } j \in S. \tag{3}$$

La fórmula de la probabilidad total permite interpretar (3) de la siguiente forma: si en cierta etapa la cadena se encuentra en cada estado j con probabilidad π_j, entonces en la etapa siguiente (y por lo tanto en todas las etapas futuras) también se encontrará en cada estado j con probabilidad π_j. Una condición suficiente para que se verifique (3) es la *condición de equilibrio detallado*:

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \text{para todo } i, j \in S. \tag{4}$$

La parte izquierda de (4) representa el flujo de probabilidad del estado i al j, mientras que la parte derecha es el flujo de probabilidad del estado j al i. Si en (4) sumamos en i, obtenemos (3) ya que $\sum_{i=1}^{k} p_{ji} = 1$, para todo $j \in S$.

4.2. El algoritmo de Metropolis-Hastings

Dada una distribución π sobre S, queremos definir una cadena cuya distribución estacionaria sea π. Para ello, la idea es diseñar un mecanismo tal que la cadena visite cada estado la proporción de veces que sea precisa. Si, por ejemplo, π asigna el doble de probabilidad a un estado que a otro, la cadena debe visitar a largo plazo el primer estado el doble de veces que el segundo.

Si en el instante t la cadena se encuentra en el estado i, ¿como se decide el estado en $t+1$ de manera que se cumpla la condición anterior? Supongamos que se propone un nuevo estado j de acuerdo con una distribución q_{ij} tal que $q_{ij} > 0$, para todo $i, j \in S$. Si tuviéramos la fortuna de que se cumpliera (4), con q_{ij} en el lugar de p_{ij}, ya habríamos encontrado la cadena que buscábamos. Normalmente no tendremos tanta suerte y existirá un par de estados $i \neq j$ para los que, por ejemplo,

$$\pi_i q_{ij} > \pi_j q_{ji}. \tag{5}$$

Intuitivamente, lo que significa (5) es que con las probabilidades de transición q_{ij} la cadena pasa de estar en i a estar en j más frecuentemente de lo que debería para que π sea su distribución estacionaria. Con el fin de corregir este desequilibrio, la idea es aceptar el estado j solo con cierta probabilidad $a_{ij} < 1$, de forma que si el estado propuesto j no se acepta, la cadena permanece en i. Recíprocamente, definimos $a_{ji} = 1$, el mayor valor posible, puesto que las transiciones de j a i son menos de las requeridas. Así pues, para pasar de i a j se debe proponer el valor j (lo que ocurre con probabilidad q_{ij}) y además este valor se debe aceptar (lo que ocurre con probabilidad a_{ij}). En consecuencia, las probabilidades de transición de la cadena así generada son $p_{ij} = q_{ij} a_{ij}$. Para deducir cuál debe ser el valor de a_{ij} imponemos la condición (4):

$$\pi_i q_{ij} a_{ij} = \pi_j q_{ji} a_{ji} = \pi_j q_{ji},$$

de donde obtenemos

$$a_{ij} = \frac{\pi_j q_{ji}}{\pi_i q_{ij}}.$$

Si la desigualdad en (5) fuese la contraria, entonces definiríamos $a_{ij} = 1$ y deduciríamos a_{ji} de forma totalmente análoga. En general, si $i \neq j$,

$$a_{ij} = \text{mín}\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}. \tag{6}$$

Recapitulamos a continuación el algoritmo de Metropolis-Hastings para generar una cadena con distribución estacionaria π. Se selecciona un valor inicial arbitrario $x_0 \in S$ para comenzar la cadena. Suponiendo que en la etapa t la cadena se encuentra en el estado i ($X_t = i$), tenemos que seguir los siguientes pasos para obtener X_{t+1}:

1. Generar una propuesta de nuevo estado a partir de la distribución q_{ij}. Sea y el valor resultante.
2. Definir
$$X_{t+1} = \begin{cases} y, & \text{con probabilidad } a_{ij}, \\ i, & \text{con probabilidad } 1 - a_{ij}, \end{cases}$$
donde a_{ij} viene dada por (6).

La utilidad principal del algoritmo es la de aproximar el valor esperado (bajo la distribución estacionaria) de una función de los estados de la cadena, $\mathrm{E}_\pi[h(X)]$, mediante los correspondientes promedios $n^{-1}\sum_{t=0}^{n-1} h(X_t)$. Las cadenas para las cuales los promedios convergen con probabilidad 1 al valor esperado se llaman *fuertemente ergódicas*. Bajo la hipótesis $q_{ij} > 0$ para todo $i, j \in S$ la cadena generada por el algoritmo de Metropolis-Hastings es irreducible (es decir, todos los estados están comunicados) y por lo tanto se verifica

$$\mathrm{P}\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} h(X_t) = \mathrm{E}_\pi[h(X)]\right\} = 1,$$

para cualquier función acotada $h : S \to \mathbb{R}$ (véase, por ejemplo, [20], pág. 53).

Aunque hemos considerado únicamente variables discretas con soporte finito, se puede aplicar el algoritmo para variables continuas de forma totalmente análoga. Si queremos generar observaciones de una variable aleatoria con función de densidad $f(x)$ con soporte $S \subset \mathbb{R}^d$, entonces el algoritmo consiste en elegir un valor inicial cualquiera $x_0 \in S$ y, suponiendo que en la etapa t la cadena se encuentra en el estado x_t ($X_t = x_t$), seguir los pasos siguientes para obtener X_{t+1}:

1. Generar una propuesta de nuevo estado a partir de la distribución definida por la función de densidad $q(y|x_t)$. Sea y el valor resultante.
2. Definir
$$X_{t+1} = \begin{cases} y, & \text{con probabilidad } a(x_t, y), \\ x_t, & \text{con probabilidad } 1 - a(x_t, y), \end{cases}$$
donde
$$a(x, y) = \text{mín}\left\{\frac{f(x)q(x|y)}{f(y)q(y|x)}, 1\right\}.$$

En este caso continuo el algoritmo es igualmente simple pero la teoría para demostrar la convergencia a la distribución estacionaria f es más complicada pues es necesario considerar cadenas de Markov con espacio de estados $S \subset \mathbb{R}^d$ (véase, por ejemplo, la sección 7.3.2 de [24]).

Una propiedad del algoritmo que lo hace especialmente útil es que solo depende de la densidad objetivo f a través de los cocientes $f(x)/f(y)$. Por ello, en lugar de trabajar con f directamente basta considerar cualquier función $\tilde{f}$ proporcional a f.

4.3. Regreso a Metropolis

La versión original del algoritmo fue propuesta por Metropolis y sus coautores en 1953 ([18]). La generalización que hemos descrito aquí fue propuesta por Hastings en 1970 ([16]). En la versión original del algoritmo solo se consideraban propuestas simétricas tales que $q(x|y) = q(y|x)$. En este caso, la probabilidad de aceptar el valor propuesto se reduce a

$$a(x, y) = \text{mín}\left\{\frac{f(y)}{f(x)}, 1\right\}.$$

Esta versión más simple del algoritmo resulta bastante intuitiva. El valor propuesto se acepta siempre que corresponda a un valor de mayor densidad de probabilidad que el valor actual (es decir, si $f(y) > f(x)$). En caso contrario, se acepta únicamente con probabilidad $f(y)/f(x)$. Un caso particular de esta situación es el *camino aleatorio de Metropolis*, que corresponde al caso en que la distribución para generar los valores propuestos es de la forma $q(y|x) = g(x - y)$, para una función par g. Por ejemplo, dado $X_t = x_t$, si los valores propuestos son de la forma $Y = x_t + \epsilon$, donde ϵ tiene distribución normal con vector de medias 0 y matriz de covarianzas Σ, tenemos un camino aleatorio de Metropolis.

En el caso simétrico que acabamos de describir, la probabilidad de aceptar los valores propuestos no depende de la distribución que utilizamos para generarlos. Sin embargo, diferentes distribuciones darán lugar a diferentes rangos de valores para las propuestas, lo que a su vez influye decisivamente en la proporción de valores que son finalmente aceptados. En este sentido, aunque casi cualquier distribución es válida para generar propuestas, una buena elección es básica para mejorar la eficiencia del algoritmo. En el ejemplo de camino aleatorio anterior, esto significa elegir la matriz de covarianzas Σ de manera que la dirección en la que se mueve la cadena y la magnitud de los incrementos sean los más adecuados.

Vamos a aplicar el camino aleatorio de Metropolis para generar una cadena de Markov cuya distribución estacionaria no es estándar. Concretamente, dados $\theta_1, \ldots, \theta_p \in \mathbb{R}^d$ consideramos la función de densidad definida para $x \in \mathbb{R}^d$, $x \notin \{\theta_1, \ldots, \theta_p\}$, mediante

$$f(x|\theta_1, \ldots, \theta_p) = c \exp\left\{-\frac{\|x\|^2}{2}\right\} \prod_{i=1}^{p} \exp\left\{-\frac{1}{\|x - \theta_i\|^2}\right\}, \tag{7}$$

donde c es la constante necesaria para que $f(x|\theta_1, \ldots, \theta_p)$ sea una función de densidad. La densidad anterior es producto de $p + 1$ factores, el primero de los cuales

corresponde salvo constantes a la densidad normal multivariante con vector de medias 0 y matriz de covarianzas la identidad. Los p factores restantes perturban la distribución normal introduciendo unos puntos $\theta_1, \ldots, \theta_p$ alrededor de los cuales la densidad de probabilidad es muy baja. La constante c por la que hay que multiplicar este producto de factores para obtener una función de densidad es en general desconocida.

Hemos generado 10000 puntos que siguen aproximadamente esta distribución para $p = d = 2$, $\theta_1 = (1, -1)$ y $\theta_2 = (-1, 1)$ usando el camino aleatorio de Metropolis. Las observaciones propuestas se han generado de acuerdo con la distribución normal con vector de medias el origen y matriz de covarianzas 0,02 veces la matriz identidad. Los puntos resultantes se han representado en la figura 7.

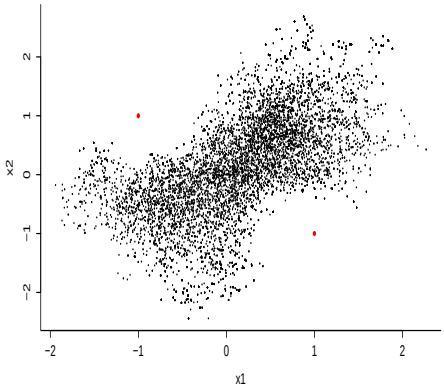

Figura 7: Diez mil puntos (generados mediante el camino aleatorio de Metropolis) de la distribución definida por la función (7) para $p = d = 2$, $\theta_1 = (1, -1)$ y $\theta_2 = (-1, 1)$. Los puntos más grandes corresponden a θ_1 y θ_2, que definen los centros de las zonas con baja densidad de probabilidad.

4.4. Aplicación en inferencia bayesiana

En general, los métodos de simulación basados en cadenas de Markov son especialmente útiles en problemas de inferencia bayesiana. Desde el punto de vista bayesiano se supone que el parámetro θ es un vector aleatorio que toma valores en un espacio paramétrico $\Theta \subset \mathbb{R}^d$ de acuerdo con una distribución de probabilidad a priori $\pi(\theta)$ que refleja el conocimiento del parámetro antes de observar los datos. Con el fin de recabar más información sobre el parámetro, se recoge una muestra $X = (X_1, \ldots, X_n)$ de observaciones con densidad (o función de probabilidad) $f(x|\theta)$. Entonces, de acuerdo con la fórmula de Bayes, la distribución a posteriori de θ dada la muestra es

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{\int_\Theta f(X|\theta)\pi(\theta)d\theta}. \tag{8}$$

Los estimadores bayesianos se basan en la distribución a posteriori del parámetro dada la muestra. Por ejemplo, puede usarse la media o la mediana de la distribución a posteriori $\pi(\theta|X)$ para estimar θ. En general necesitamos calcular, para una función adecuada $h(\theta)$, el valor de

$$\mathrm{E}[h(\theta)|X] = \int_{\Theta} h(\theta)\pi(\theta|X)d\theta. \tag{9}$$

El valor de $\mathrm{E}[h(\theta)|X]$ es en muchos casos imposible de calcular analíticamente e incluso puede ser muy difícil de aproximar usando los métodos tradicionales de integración numérica. La razón es que evaluar la integral en (9) requiere aproximar la integral en el denominador de (8), cuya dimensión coincide con la del estimador que queremos calcular. Sin embargo, si aplicamos el algoritmo de Metropolis-Hastings, podemos generar un número grande B de valores $\theta_1^*, \ldots, \theta_B^*$ cuya distribución aproximada es $\pi(\theta|X)$ y, a su vez, usar la aproximación

$$\mathrm{E}[h(\theta)|X] \approx \frac{1}{B}\sum_{b=1}^{B} h(\theta_b^*).$$

Como hemos visto en el ejemplo del apartado anterior, para generar $\theta_1^*, \ldots, \theta_B^*$ basta saber que $\pi(\theta|X)$ es proporcional a $f(X|\theta)\pi(\theta)$ y no es necesario evaluar la integral del denominador de (8).

5. Comentarios finales

Gosset no fue el primero en aproximar la distribución de un estadístico complicado mediante muestras generadas artificialmente. Algunos autores mencionan como antecedente el famoso experimento de Buffon en el siglo XVIII para aproximar el valor de π lanzando una aguja sobre un sistema de líneas paralelas. Stigler, por su parte ([27]), presenta tres ejemplos publicados en el último cuarto del siglo XIX y directamente relacionados con problemas de inferencia estadística. En cada ejemplo se utiliza un mecanismo de aleatorización diferente. Erastus L. de Forest empleó en 1876 un conjunto de 100 cartas para generar observaciones del estadístico

$$\frac{1}{m}\sum_{i=1}^{m} \log\left(\frac{\bar{v}_i}{\bar{v}_i'}\right),$$

con $\bar{v}_i = n^{-1}\sum_{i=1}^{n} v_{ij}$ y $\bar{v}_i' = n^{-1}\sum_{i=1}^{n} v_{ij}'$, donde v_{ij} y v_{ij}' corresponden a los valores absolutos de variables aleatorias independientes con distribución normal estándar. George H. Darwin, uno de los diez hijos de Charles Darwin, que trabajó en el estudio de las mareas y en problemas de meteorología y fue conferenciante invitado en el Congreso Internacional de Matemáticos de 1908, construyó en 1877 una ruleta para generar observaciones del valor absoluto de una normal. Posteriormente utilizaba una moneda para asignar aleatoriamente el signo. Finalmente, Francis Galton –un primo de Darwin cuyas contribuciones en estadística son muy importantes– diseñó unos

dados especiales en 1890 con el fin de generar observaciones con distribución normal ([13]). De acuerdo con Stigler, los dados de Galton son tal vez el dispositivo más antiguo que aún se conserva (en la *Galton collection*, University College London) para este propósito.

La selección de temas que se han tratado en este artículo no es en absoluto exhaustiva y responde a los gustos del autor. Muchas otras técnicas de estadística computacional descansan en la posibilidad de llevar a cabo cálculos intensivos. Para terminar, vamos a mencionar algunas de ellas indicando alguna referencia que pueda resultar útil al lector interesado.

En la sección 3 hemos discutido el uso del bootstrap para estimar la distribución de un estadístico. En problemas de clasificación se ha propuesto utilizar el promedio de los clasificadores obtenidos al aplicar un procedimento (por ejemplo, un árbol de clasificación) a muestras bootstrap de los datos de entrenamiento. Esta técnica se conoce como *bagging* (**b**ootstrap **agg**regat**ing**) y fue propuesta por Breiman en 1996 ([3]). Algunos ejemplos muy sencillos de los efectos del bagging se pueden encontrar en [1].

Los métodos de validación cruzada se suelen utilizar para estimar el error de predicción de un modelo estadístico. La idea básica consiste en dividir los datos disponibles en K partes y utilizar sucesivamente $K-1$ de ellas para estimar los parámetros del modelo y la restante para validarlo. Finalmente los K errores de predicción obtenidos se promedian para dar lugar a un estimador del error de predicción único. El valor ideal de K depende de consideraciones sobre el sesgo y la varianza del estimador obtenido. Un valor muy utilizado es $K = n$, donde n es el tamaño muestral, lo que equivale a dejar en cada etapa un único dato fuera y utilizar los $n-1$ restantes para estimar el modelo. Hay cientos de artículos relacionados con este tema, pero una referencia clásica es [28].

El algoritmo de Metropolis-Hastings no es el único método de simulación basado en generar una cadena de Markov apropiada. Tal vez ni siquiera sea el más usado, ya que el muestreo de Gibbs (*Gibbs sampling*) es también muy conocido. Este método se utiliza para generar observaciones de una distribución multivariante a partir de observaciones de las distribuciones condicionadas univariantes. Una referencia que contiene las ideas básicas sobre este importante algoritmo es [4] y otra con información más detallada es [24].

Referencias

[1] J.R. Berrendero, The bagged median and the bragged mean. *The American Statistician* **61** (2007), 325-330.

[2] P.J. Bickel y D.A. Freedman, Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** (1981), 1196–1217.

[3] L. Breiman, Bagging predictors. *Machine learning* **24** (1996), 123-140.

[4] G. Casella y E.I. George, Explaining the Gibbs sampler. *The American Statistician* **46**, (1992), 167-174.

[5] E. CHUNG Y J.P. ROMANO. Exact and asymptotically robust permutation tests. *The Annals of Statistics* **41** (2013), 484-507.

[6] A. DASGUPTA, *Asymptotic theory of statistics and probability*, Springer, Nueva York, 2008.

[7] M. DWASS, Modified randomization tests for nonparametric hypotheses, *Ann. Math. Statist.* **28** (1957), 181-187.

[8] T. EDEN Y F. YATES, On the validity of Fisher's z test when applied to an actual example of non-normal data, *Journal of Agricultural Science*, **23** (1933), 6-17.

[9] B. EFRON, Bootstrap methods: another look at the Jackknife, *Ann. Statist.* **7** (1979), 1–26.

[10] B. EFRON Y R. TIBSHIRANI, *An Introduction to the Bootstrap*, Chapman & Hall, Nueva York, 1993.

[11] R.A. FISHER, *The design of experiments*, Oliver and Boyd, Edimburgo, 1935.

[12] R.A. FISHER, The Coefficient of Racial Likeness and the Future of Craniometry, *The Journal of the Royal Anthropological Institute of Great Britain and Ireland* **66** (1936), 57-63.

[13] F. GALTON, Dice for statistical experiments. *Nature* **42**, (1890), 13-14.

[14] E. GINÉ Y J. ZINN, Bootstrapping general empirical measures, *Ann. Probab.* **18**, (1990), 851–869.

[15] P. HALL, *The bootstrap and Edgeworth expansion*, Springer, Nueva York, 1992.

[16] W.K.HASTINGS, Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, (1970), 97-109.

[17] N. METROPOLIS, The beginning of the Monte Carlo method, *Los Alamos Science*, Special issue (1987), 125–130.

[18] N. METROPOLIS, A.W. ROSENBLUTH, M.N. ROSENBLUTH, A.H. TELLER Y E. TELLER. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, (1953), 1087-1092.

[19] N. METROPOLIS Y S. ULAM, The Monte Carlo method, *J. Amer. Statist. Assoc.* **44** (1949), 335–341.

[20] J.R. NORRIS, *Markov chains.* Cambridge University Press, Cambridge, 1998.

[21] E.G.J. PITMAN, Significance tests which may be applied to samples from any populations, *J. Roy. Statist. Soc. Suppl.* **4** (1937), 119-130.

[22] W.C. PARR, The bootstrap: some large sample theory and connections with robustness, *Statist. Probab. Lett.*, **3**, 1985.

[23] J.P. ROMANO, On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.* **85** (1990), 686-692.

[24] C.P. ROBERT. Y G. CASELLA *Monte Carlo statistical methods.* Springer, Nueva York, 1999.

[25] G.R. SHORACK Y J.A. WELLNER, *Empirical processes with applications to statistics*, John Wiley & Sons, Nueva York, 1986.

[26] K. Singh, On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** (1981), 1187–1195.

[27] S.M. Stigler, Stochastic simulation in the nineteenth century. *Statistical Science* **6** (1991), 89-97.

[28] M. Stone, Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B* **36** (1974), 111-147.

[29] Student, The probable error of a mean, *Biometrika* **6** (1908), 1–25.

José Ramón Berrendero, Dpto. de Matemáticas, Universidad Autónoma de Madrid
Correo electrónico: joser.berrendero@uam.es
Página web: http://www.uam.es/joser.berrendero