EST-46115: Modelación Bayesiana

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Integración Monte Carlo.
Objetivo: Estudiar integración numérica en el contexto probabilistico. Estudiar, en particular, el método Monte Carlo y entender sus bondades y limitaciones en el contexto de inferencia Bayesiana.
Lectura recomendada: Sección 6.1 de Johnson et al. [1]. Una lectura mas técnica sobre reglas de cuadratura se puede encontrar en la sección 3.1 de Reich and Cotter [2]. Y una buena referencia (técnica) sobre el método Monte Carlo lo encuentran en Sanz-Alonso et al. [3].

1. INTRODUCCIÓN

En inferencia bayesiana lo que queremos es poder resolver

$$\mathbb{E}[f] = \int_{\Theta} f(\theta)\, \pi(\theta|y)\, \mathrm{d}\theta\,. \tag{1}$$

Lo que necesitamos es resolver integrales con respecto a la distribución de interés.

- La pregunta clave (I) es: ¿qué distribución?
- La pregunta clave (II) es: ¿con qué método numérico resuelvo la integral?
- La pregunta clave (III) es: ¿y si no hay método numérico?

2. ¿POR QUÉ INTEGRAR?

Consideremos de interés estimar la proporción de volumen de la hiper-esfera contenida en un hiper-cubo unitario conforme aumenta la dimensión del problema.

```
distancia_euclideana ← function(u) sqrt(sum(u * u));

experimento ← function(ndim){
  nsamples ← 1e5;
  y ← matrix(runif(nsamples * ndim, -0.5, 0.5), nsamples, ndim);
  mean(apply(y, 1, distancia_euclideana) < 0.5)
}
```

```
puntos.grafica ← tibble(dims = c(1, 5, 10, 25, 50, 100)) ▷
  mutate(points = map(dims, function(dim){
    tibble(x = seq(0, 15, length.out = 1000)) ▷
      mutate(y = chi.pdf(x, dim))
  }), dimensions = factor(dims))
```

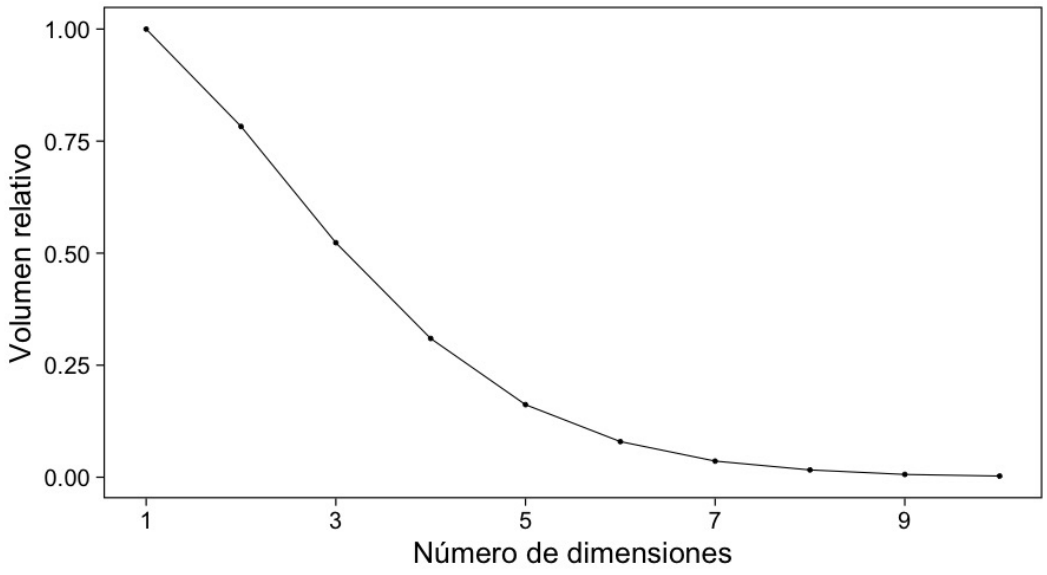

FIGURA 1. *Evolución del volumen relativo de la hiper-esfera circunscrita dentro del hiper-cubo unitario.*

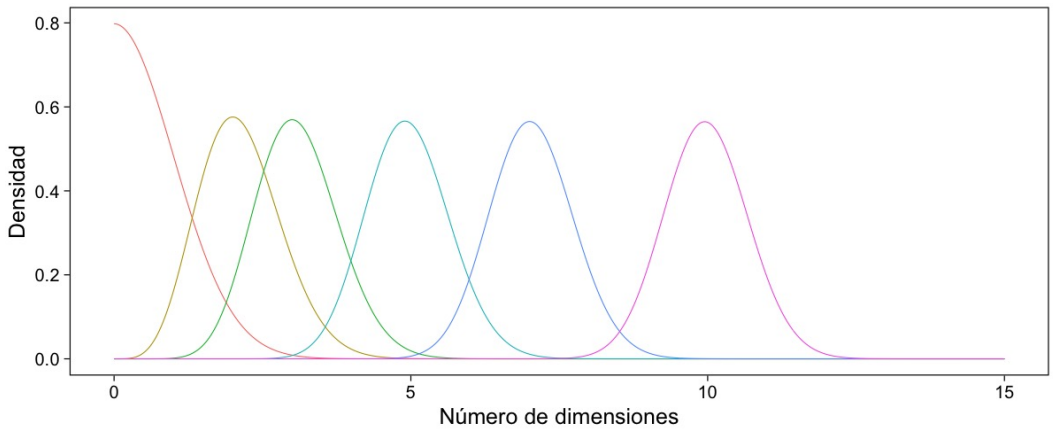

FIGURA 2. *Densidad de los vecinos de la moda para una Normal multivariada estándar* $\mathsf{N}(0, I_p)$.

Otro detalle interesante de altas dimensiones es la poca intuición probabilística que tenemos de estos espacios y de lo que es una muestra típica de una distribución.

Por ejemplo, para $X \sim \mathsf{N}(0, 1)$ estamos acostumbrados a asociar la moda como el valor de mas alta densidad. Lo cual es un error terrible en varias dimensiones.

Consideremos un análisis analítico. Por ejemplo, sabemos que si $X \sim \mathsf{N}(0, I_p)$, entonces tenemos que

$$\sum_{i=1}^{p} X_i^2 \sim \chi_p^2 \,. \tag{2}$$

Gráfiquemos el valor central de estas variables aleatorias y sus percentiles del 2.5 % y 97.5 %. Lo que observamos es que una `muestra típica` no se comporta como el promedio de nuestra distirbución, *aka* el individuo promedio no es tan común.

Lo que sucede se conoce como el fenómeno de `concentración de medida` donde los puntos de más alta densidad no corresponden a los puntos de mayor volumen (probabilidad).

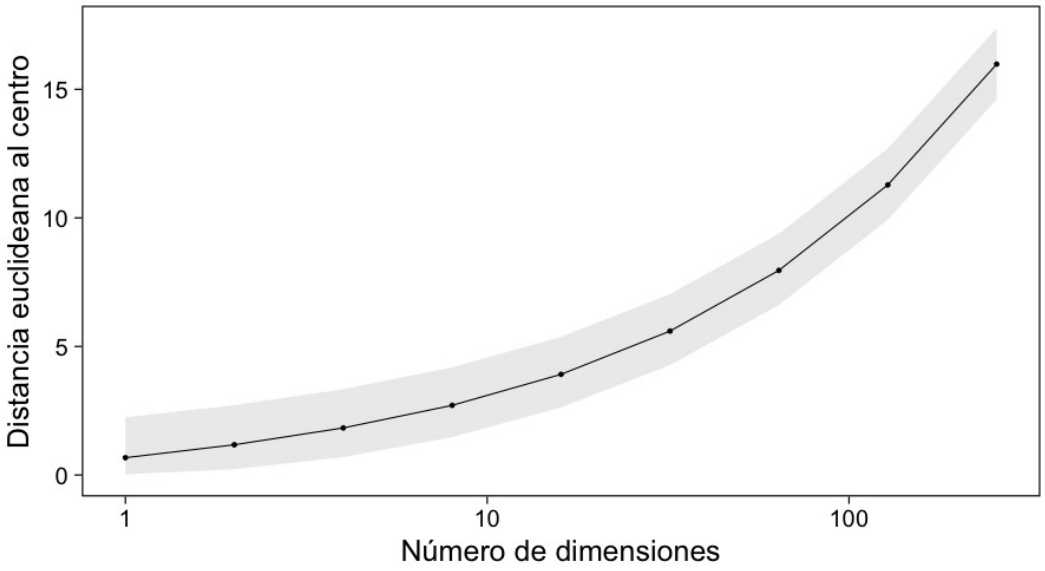

FIGURA 3. *Distancia euclideana de puntos aleatorios de una Gaussiana multivariada al centro de la distribución. Esto ilustra que aunque el centro es el comportamiento promedio, los puntos típicos de una Gaussiana se encuentran lejos.*

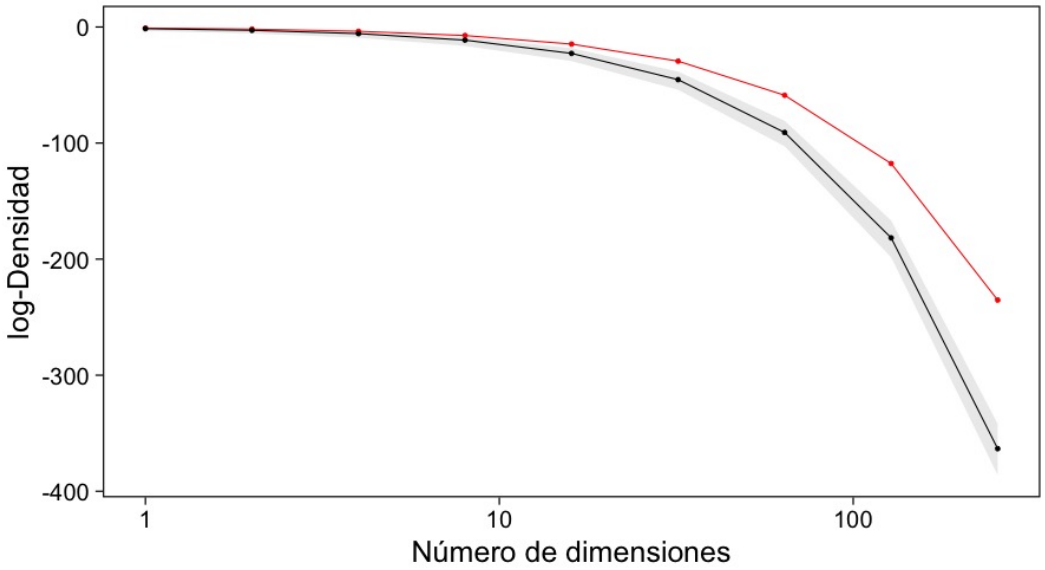

FIGURA 4. *En rojo la log-densidad de la moda de una Gaussiana multivariada. En negro la log-densidad de muestras aleatorias de una Gaussiana multivariada. Esto muestra que los elementos con mayor densidad no corresponden a vecindades de mayor volumen.*

Esto explica por qué no queremos realizar la aproximación

$$\pi(f) \approx f(\theta^\star)\,, \qquad \text{donde} \qquad \theta^\star = \arg\max_{\theta\in\Theta} \pi(\theta)\,. \tag{3}$$

3. INTEGRACIÓN NUMÉRICA

Recordemos la definición de integrales Riemann:

$$\int_a^b f(x)\mathrm{d}x\,.$$

La aproximación utilizando una malla de N puntos sería:

$$\sum_{n=1}^{N} f(u_n)\Delta u_n.$$

El método útil cuando las integrales se realizan cuando tenemos pocos parámetros. Es decir, $\theta \in \mathbb{R}^p$ con p pequeña.

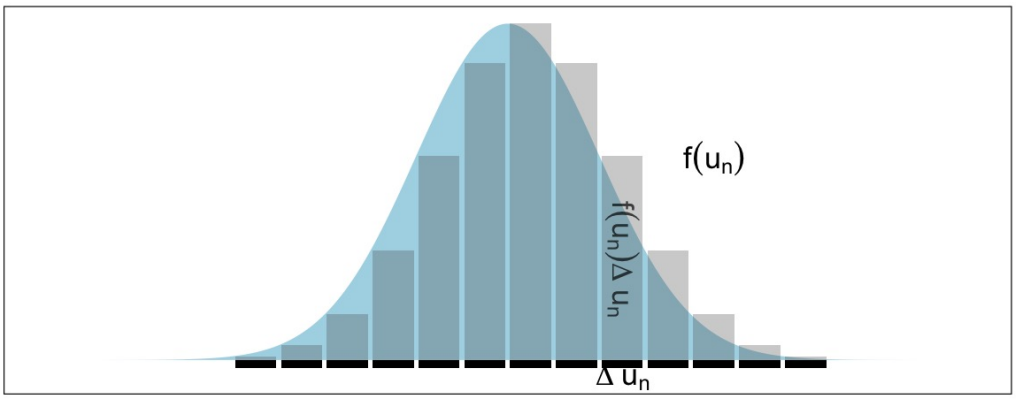

FIGURA 5. *Integral por medio de discretización con* $N = 11$.

3.1. Ejemplo: Proporción

Supongamos que $p(S_n = k|\theta) \propto \theta^k(1-\theta)^{n-k}$ cuando observamos k éxitos en n pruebas independientes. Supongamos que nuestra inicial es $p(\theta) = 2\theta$ (`checa que es una densidad`).

```
crear_log_post ← function(n, k){
  function(theta){
    verosim ← k * log(theta) + (n - k) * log(1 - theta)
    inicial ← log(theta)
    verosim + inicial
  }
}
```

```
# observamos 3 exitos en 4 pruebas:
log_post ← crear_log_post(4, 3)
prob_post ← function(x) { exp(log_post(x))}
```

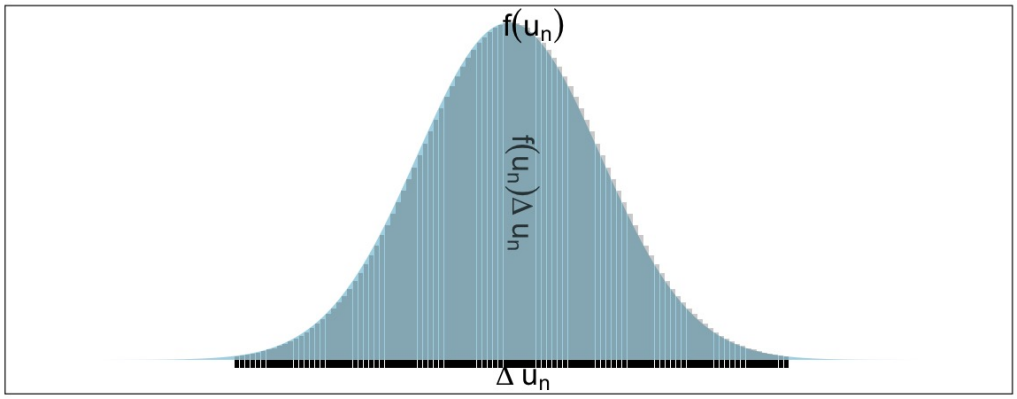

FIGURA 6. *Integral por medio de una malla fina,* $N = 101$.

```
# integramos numericamente
p_x ← integrate(prob_post, lower = 0, upper = 1, subdivisions = 100L)
p_x
```

```
0.03333 with absolute error < 3.7e-16
```

Y ahora podemos calcular la media posterior:

$$\mathbb{E}[\theta|S_n] = \int_\Theta \theta\, \pi(\theta|S_n)\, \mathrm{d}\theta\,. \tag{4}$$

```
media_funcion ← function(theta){
  theta * prob_post(theta) / p_x$value
}
integral_media ← integrate(media_funcion,
                           lower = 0, upper = 1,
                           subdivisions = 100L)
media_post ← integral_media$value
c(Numerico = media_post, Analitico = 5/(2+5))
```

```
Numerico Analitico
  0.7143    0.7143
```

3.2. Más de un parámetro

Consideramos ahora un espacio con $\theta \in \mathbb{R}^p$. Si conservamos N puntos por cada dimensión, ¿cuántos puntos en la malla necesitaríamos? Lo que tenemos son recursos computacionales limitados y hay que buscar hacer el mejor uso de ellos. En el ejemplo, hay zonas donde no habrá contribución en la integral.

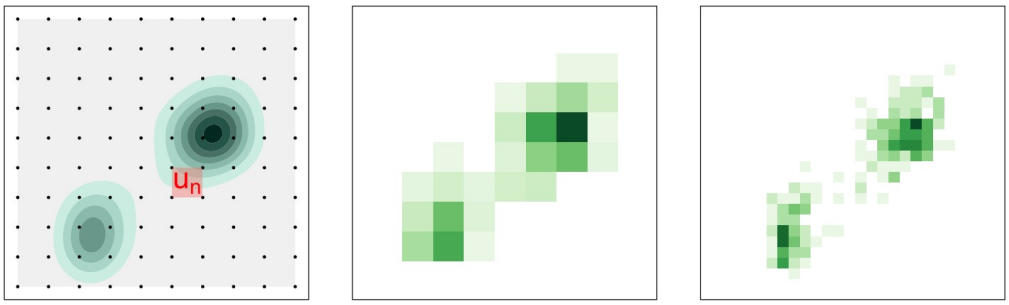

FIGURA 7. *Integral por método de malla.*

3.3. Reglas de cuadratura

Por el momento hemos escogido aproximar las integrales por medio de una aproximación con una `malla uniforme`. Sin embargo, se pueden utilizar aproximaciones

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{n=1}^{N} f(\xi_n)\,\omega_n\,.$$

Estas aproximaciones usualmente se realizan para integrales en intervalos cerrados $[a, b]$. La regla de cuadratura determina los pesos ω_n y los centros ξ_n pues se escogen de acuerdo a `ciertos criterios de convergencia`.

Por ejemplo, se consideran polinomios que aproximen con cierto grado de precisión el integrando. Los pesos y los centros se escogen de acuerdo a la familia de polinomios. Pues para cada familia se tienen identificadas las mallas que optimizan la aproximación. Ver sección 3.1 de Reich and Cotter [2].

4. INTEGRACIÓN MONTE CARLO

$$\pi(f) = \mathbb{E}_\pi[f] = \int f(x)\pi(x)\mathrm{d}x\,,$$

$$\hat{\pi}_N^{\mathsf{MC}}(f) = \frac{1}{N}\sum_{n=1}^{N} f(x^{(n)}), \qquad \text{donde } x^{(n)} \overset{\text{iid}}{\sim} \pi, \qquad \text{con } n = 1, \ldots, N\,,$$

$$\pi(f) \approx \hat{\pi}_N^{\mathsf{MC}}(f)\,.$$

4.1. Ejemplo: Dardos

Consideremos el experimento de lanzar dardos uniformemente en un cuadrado de tamaño 2, el cual contiene un circulo de radio 1.

Si escogemos N suficientemente grande entonces nuestro promedio converge a la integral. En Fig. 9 se muestra para cada n en el eje horizontal cómo cambia nuestra estimación $\hat{\pi}_n^{\mathsf{MC}}(f)$.

También podemos en replicar el experimento unas M veces y observar cómo cambiaría nuestra estimación con distintas semillas. Por ejemplo, podemos replicar el experimento 10 veces. En `R` y `python` lo usual es utilizar `arreglos multidimensionales` para poder guardar muestras bajo distintas replicas.

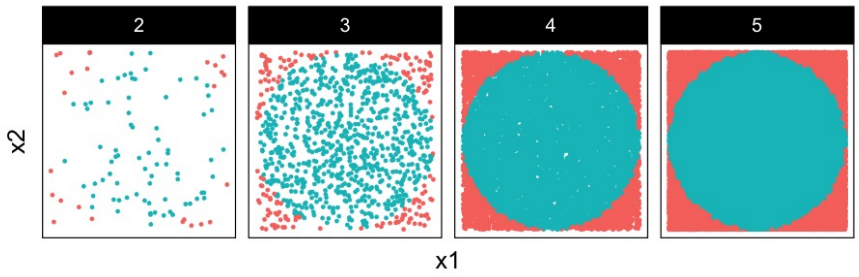

FIGURA 8. *Integración Monte Carlo para aproximar* π.

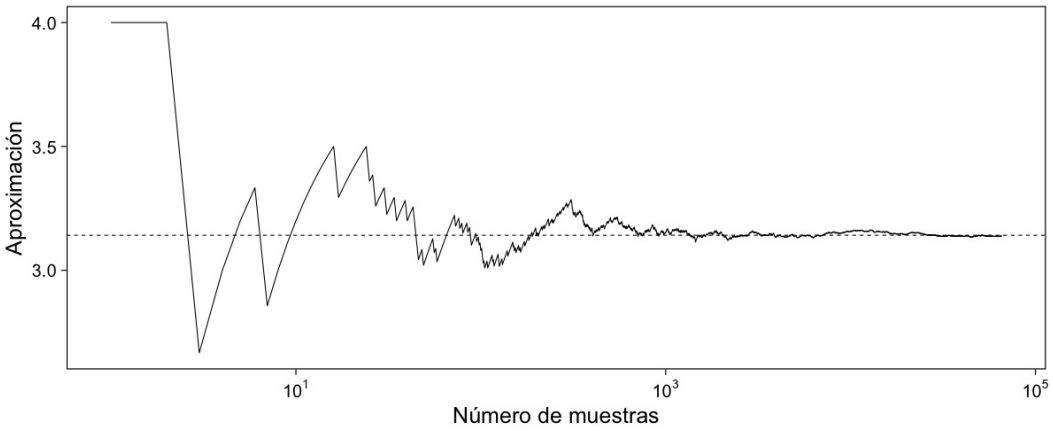

FIGURA 9. *Estimación* $\hat{\pi}_N^{MC}(f)$ *con* $N \to \infty$.

```
set.seed(108)
nsamples ← 10**4; nexp ← 100
U ← runif(nexp * 2 * nsamples)
U ← array(U, dim = c(nexp, 2, nsamples))
apply(U[1:5,,], 1, str)
```

```
 num [1:2, 1:10000] 0.455 0.164 0.529 0.415 0.474 ...
 num [1:2, 1:10000] 0.404 0.9282 0.0883 0.3126 0.4386 ...
 num [1:2, 1:10000] 0.351 0.449 0.369 0.814 0.695 ...
 num [1:2, 1:10000] 0.664 0.627 0.185 0.882 0.113 ...
 num [1:2, 1:10000] 0.4635 0.0115 0.0117 0.5086 0.9384 ...
NULL
```

```
resultados ← apply(U, 1, function(x){
  dardos ← apply(x**2, 2, sum)
  exitos ← ifelse(dardos ≤ 1, 1, 0)
  prop   ← cummean(exitos)
  4 * prop
})
```

Lo cual nos permite realizar distintos escenarios posibles.

Bajo ciertas consideraciones teóricas podemos esperar un buen comportamiento de nuestro estimador de la integral. E incluso podríamos (si el número de simulaciones lo permite)

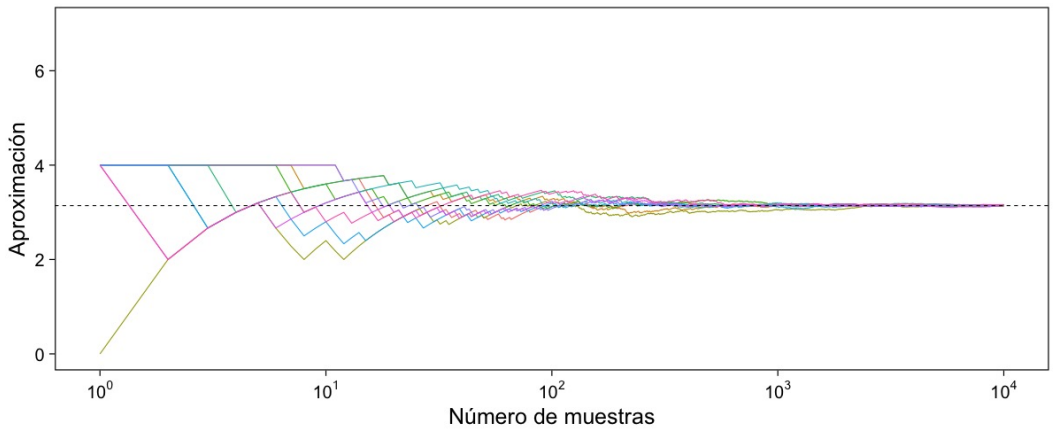

FIGURA 10. *Réplica de las trayectorias de diversas realizaciones de la aproximación de la integral.*

aproximar dicho comportamiento utilizando distribuciones asintóticas, (TLC).

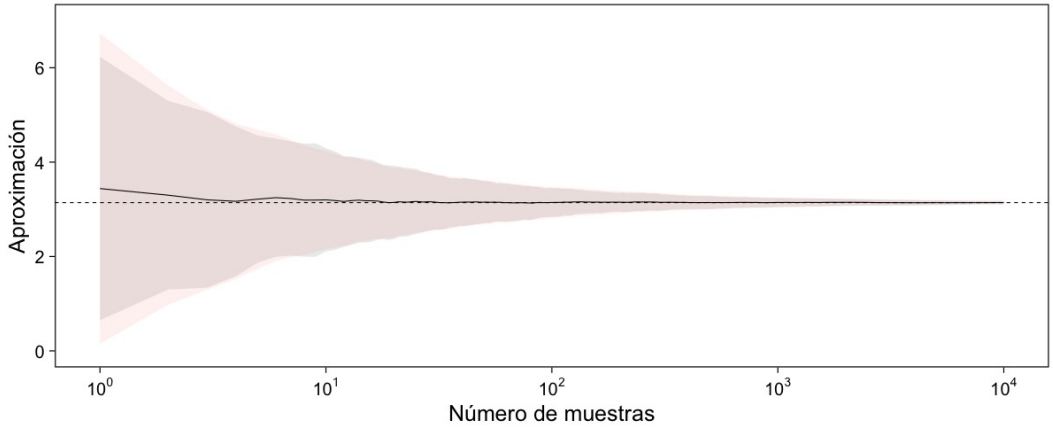

FIGURA 11. *Comportamiento promedio e intervalos de confianza.*

Podemos explicar la reducción de los intervalos de confianza por medio de la varianza de la estimación de la integral en las distintas réplicas que tenemos. Mas adelante explicaremos de dónde viene la línea punteada. Vemos cómo, aunque captura bien la reducción en varianza, puede sub- o sobre-estimarla.

4.2. Propiedades

A continuación enunciaremos algunas propiedades clave del método Monte Carlo. Poco a poco las iremos explicando y en particular discutiremos algunas de ellas.

4.2.1. `Teorema` *[Error Monte Carlo]* Sea $f : \mathbb{R}^p \to \mathbb{R}$ cualquier función bien comportada[†]. Entonces, el estimador Monte Carlo es **insesgado**. Es decir, se satisface

$$\mathbb{E}\left[\hat{\pi}_N^{\mathsf{MC}}(f) - \pi(f)\right] = 0, \tag{5}$$

para cualquier N. Usualmente estudiamos el error en un escenario pesimista donde medimos el **error cuadrático medio** en el peor escenario

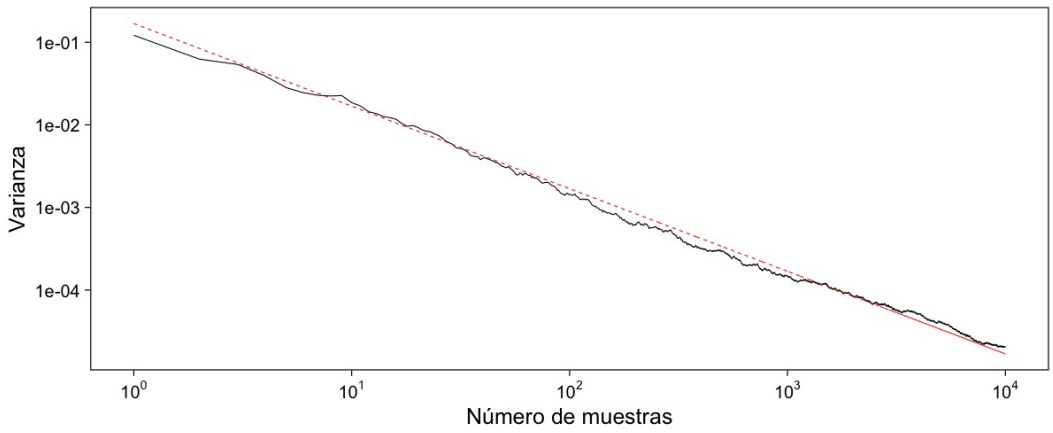

FIGURA 12. *Comportamiento promedio e intervalos de confianza.*

$$\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left(\hat{\pi}_N^{\mathsf{MC}}(f) - \pi(f)\right)^2\right] \leq \frac{1}{N}.$$

Esta desigualdad nos muestra una de las propiedades que usualmente se celebran de los métodos Monte Carlo. La integral y nuestra aproximación de ella por medio de simulaciones tiene un error acotado proporcionalmente por el número de simulaciones.

En particular, la varianza del estimador (**error estándar**) satisface la igualdad

$$\mathsf{ee}^2\left(\hat{\pi}_N^{\mathsf{MC}}(f)\right) = \frac{\mathbb{V}_\pi(f)}{N}.$$

Esta igualdad, aunque consistente con nuestra desigualdad anterior, nos dice algo mas. El error de nuestra aproximación **depende** de la varianza de f bajo la distribución π.

4.2.2. `Teorema` *[TLC para estimadores Monte Carlo]* Sea f una función **bien comportada** [††], entonces bajo una N suficientemente grande tenemos

$$\sqrt{N}\left(\hat{\pi}_N^{\mathsf{MC}}(f) - \pi(f)\right) \sim \mathsf{N}\left(0, \mathbb{V}_\pi(f)\right). \tag{6}$$

4.2.3. `Nota:` El estimador Monte Carlo del que hablamos, $\hat{\pi}_N^{\mathsf{MC}}(f)$, es una estimación con una `muestra finita de simulaciones`. En ese sentido podemos pensar que tenemos un *mapeo* de muestras a estimador

$$(x^{(1)}, \ldots, x^{(N)}) \mapsto \hat{\pi}_N^{\mathsf{MC}}(f), \tag{7}$$

con $x^{(i)} \overset{\mathsf{iid}}{\sim} \pi$.

De lo cual es natural pensar: ¿y si hubiéramos observado otro conjunto de simulaciones? Nuestro proceso de estimación es el mismo pero la muestra puede cambiar.

En este sentido nos preguntamos por el `comportamiento promedio` bajo distintas muestras observadas

$$\mathbb{E}[\hat{\pi}_N^{\mathsf{MC}}(f)] = \mathbb{E}_{x_1,\ldots,x_N}[\hat{\pi}_N^{\mathsf{MC}}(f)]. \tag{8}$$

De la misma manera nos podemos preguntar sobre la `dispersión alrededor de dicho promedio` (varianza)

$$\mathbb{V}[\hat{\pi}_N^{\mathsf{MC}}(f)] = \mathbb{V}_{x_1,\ldots,x_N}[\hat{\pi}_N^{\mathsf{MC}}(f)]\,. \tag{9}$$

Al ser un ejercicio de `estimación` la desviación estándar del estimador recibe el nombre de `error estándar`. Lo cual denotamos por

$$\mathsf{ee}[\hat{\pi}_N^{\mathsf{MC}}(f)] = \Big(\mathbb{V}[\hat{\pi}_N^{\mathsf{MC}}(f)]\Big)^{1/2} = \left(\frac{\mathbb{V}_\pi(f)}{N}\right)^{1/2}. \tag{10}$$

4.2.4. ***Nota:*** Para algunos estimadores la fórmula del error estándar se puede obtener de manera analítica. Para otro tipo, tenemos que utilizar propiedades asintóticas (p.e. cota de Cramer-Rao).

Hay casos en los que no existe una fórmula asintótica o resultado analítico, pero podemos usar simulación [`8)`] para cuantificar dicha dispersión.

4.2.5. ***Nota:*** Hay situaciones en las que la distribución normal asintótica no tiene sentido. Para este tipo de situaciones también veremos cómo podemos utilizar simulación para cuantificar dicha dispersión.

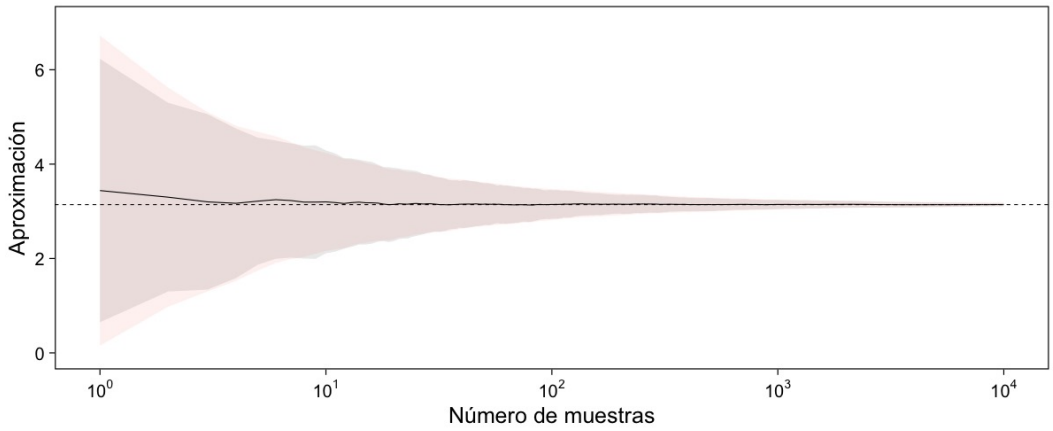

FIGURA 13. *Comportamiento promedio e intervalos de confianza con aproximación asintótica.*

4.3. Ejemplo: Proporciones

Consideramos la estimación de una proporción θ, tenemos como inicial $p(\theta) \propto \theta$, que es una $\mathsf{Beta}(2,1)$. Si observamos 3 éxitos en 4 pruebas, entonces sabemos que la posterior es $p(\theta|x) \propto \theta^4(1-\theta)$, que es una $\mathsf{Beta}(5,2)$. Si queremos calcular la media y el segundo momento posterior para θ, en teoría necesitamos calcular

$$\mu_1 = \int_0^1 \theta\; p(\theta|X=3)\,\mathrm{d}\theta, \qquad \mu_2 = \int_0^1 \theta^2\; p(\theta|X=3)\,\mathrm{d}\theta\,. \tag{11}$$

Utilizando el `método Monte Carlo`:

```
theta ← rbeta(10000, 5, 2)
media_post ← mean(theta)
momento_2_post ← mean(theta^2)
c(mu_1 = media_post, mu_2 = momento_2_post)
```

```
  mu_1   mu_2
0.7131 0.5343
```

Incluso, podemos calcular cosas mas *exóticas* como

$$P(e^{\theta} > 2|x)\,. \tag{12}$$

```
mean(exp(theta) > 2)
```

```
[1] 0.5918
```

4.4. Ejemplo: Sabores de helados

Supongamos que probamos el nivel de gusto para 4 sabores distintos de una paleta. Usamos 4 muestras de aproximadamente 50 personas diferentes para cada sabor, y cada uno evalúa si le gustó mucho o no. Obtenemos los siguientes resultados:

```
# A tibble: 4 × 4
  sabor          n gusto prop_gust
  <chr>      <dbl> <dbl>     <dbl>
1 fresa         50    36     0.72
2 limon         45    35     0.778
3 mango         51    42     0.824
4 guanabana     50    29     0.58
```

Listing 1. *Resultados de las encuestas.*

Usaremos como inicial $\mathsf{Beta}(2,1)$ (pues hemos obervado cierto sesgo de cortesía en la calificación de sabores, y no es tan probable tener valores muy bajos) para todos los sabores, es decir $p(\theta_i)$ es la funcion de densidad de una $\mathsf{Beta}(2,1)$. La inicial conjunta la definimos entonces, usando `independencia inicial`, como

$$p(\theta_1,\theta_2,\theta_3,\theta_4) = p(\theta_1)p(\theta_2)p(\theta_3)p(\theta_4)\,.$$

Pues inicialmente establecemos que ningún parámetro da información sobre otro: saber que mango es muy gustado no nos dice nada acerca del gusto por fresa. Bajo este supuesto, y el supuesto adicional de que las muestras de cada sabor son independientes, podemos mostrar que las `posteriores son independientes`:

$$p(\theta_1,\theta_2,\theta_3,\theta_4|k_1,k_2,k_3,k_4) = p(\theta_4|k_1)p(\theta_4|k_2)p(\theta_4|k_3)p(\theta_4|k_4)$$

```
# A tibble: 4 × 7
  sabor          n gusto prop_gust a_post b_post media_post
  <chr>      <dbl> <dbl>     <dbl>  <dbl>  <dbl>      <dbl>
1 fresa         50    36     0.72      38     15      0.717
2 limon         45    35     0.778     37     11      0.771
3 mango         51    42     0.824     44     10      0.815
4 guanabana     50    29     0.58      31     22      0.585
```

Listing 2. *Resultado de inferencia Bayesiana.*

Podemos hacer preguntas interesantes como: ¿cuál es la probabilidad de que mango sea el sabor preferido? Para contestar esta pregunta podemos utilizar simulación y responder por medio de un procedimiento Monte Carlo.

```
## Generamos muestras de la posterior
paletas ← datos ▷
  mutate(alpha = a_post, beta = b_post) ▷
  nest(params.posterior = c(alpha, beta)) ▷
  mutate(muestras.posterior = map(params.posterior, modelo_beta)) ▷
  select(sabor, muestras.posterior)
```

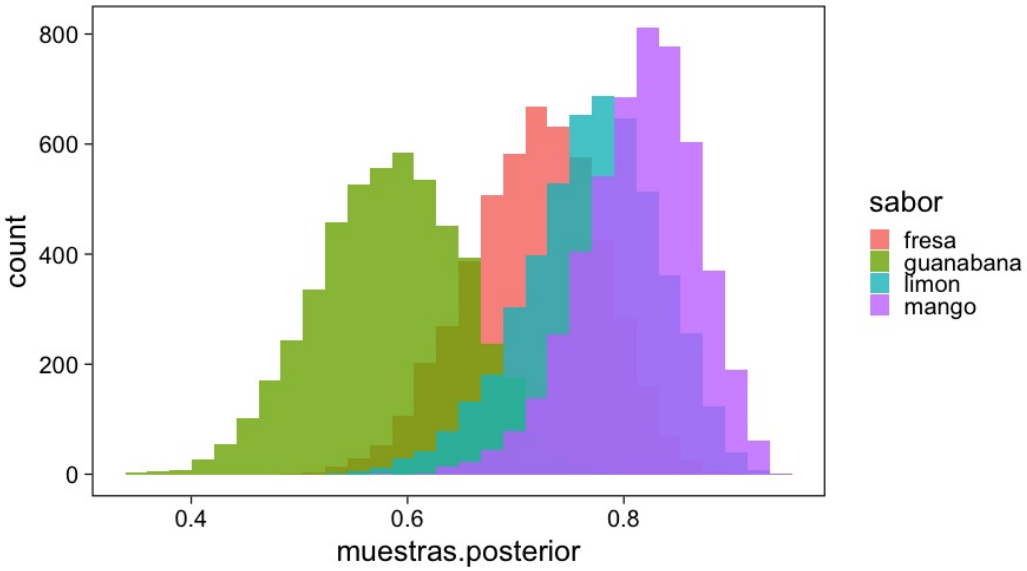

Figura 14. *Histogramas de la distribución predictiva marginal para cada θ_j.*

```
## Utilizamos el metodo Monte Carlo para aproximar la integral.
paletas ▷
  unnest(muestras.posterior) ▷
  mutate(id = rep(seq(1, 5000), 4)) ▷ group_by(id) ▷
  summarise(favorito = sabor[which.max(muestras.posterior)]) ▷
  group_by(favorito) ▷ tally() ▷
  mutate(prop = n/sum(n))
```

```
# A tibble: 4 × 3
  favorito      n   prop
  <chr>     <int>  <dbl>
1 fresa       291 0.0582
2 guanabana     2 0.0004
3 limon      1331 0.266
4 mango      3376 0.675
```

Listing 3. *Aproximación Monte Carlo.*

Escencialmente estamos preguntándonos sobre calcular la integral:

$$\mathbb{P}(\text{mango sea preferido}) = \int_\Theta f(\theta_1, \ldots, \theta_4)\, p(\theta_1, \ldots, \theta_4 | X_1, \ldots, X_n) d\theta\,, \tag{13}$$

donde $f(\theta_1, \ldots, \theta_4) = \mathbb{I}_{[\theta_4 \geq \theta_j, j \neq 4]}(\theta_1, \ldots, \theta_4)$.

4.5. Tarea: Sabores de helados

- ¿Cuál es la probabilidad a priori de que cada sabor sea el preferido?
- Con los datos de arriba, calcula la probabilidad de que la gente prefiera el sabor de mango sobre limón.

5. EXTENSIONES: MUESTREO POR IMPORTANCIA

Incluso cuando tenemos una integral **complicada** podemos `relajar` el problema de integración. De tal forma que podemos `sustituir`

$$\int f(x)\pi(x)\mathrm{d}x = \int f(x)\frac{\pi(x)}{\rho(x)}\,\rho(x)\mathrm{d}x = \int f(x)\, w(x)\, \rho(x)\, \mathrm{d}x\,,$$

donde ρ es una densidad de una variable aleatoria `adecuada`.

Esto nos permite utilizar lo que sabemos de las propiedades del método Monte Carlo para resolver la integral

$$\pi(f) = \int f(x)\pi(x)\mathrm{d}x = \int f(x) w(x)\, \rho(x)\, \mathrm{d}x =: \rho(fw)\,,$$

por medio de una aproximación

$$\pi(f) \approx \sum_{n=1}^{N} \bar{w}^{(n)} f(x^{(n)}), \qquad x^{(n)} \overset{\text{iid}}{\sim} \rho\,. \tag{14}$$

Al estimador le llamamos el estimador por importancia y lo denotamos por

$$\pi_N^{\mathsf{IS}}(f) = \sum_{n=1}^{N} \bar{w}^{(n)} f(x^{(n)}), \qquad \bar{w}^{(n)} = \frac{w(x^{(n)})}{\sum_{m=1}^{N} w(x^{(m)})}\,. \tag{15}$$

5.1. Propiedades: muestreo por importancia

Lamentablemente, utilizar muestreo por importancia `impacta la calidad de la estimación` (medida, por ejemplo, en términos del **peor error cuadrático medio cometido**). El impacto es un factor que incorpora la *diferencia* entre la distribución `objetivo` –para integrales de la forma $\int f(x)\mathrm{d}x$, implica la distribución uniforme– y la distribución `sustituto`. Puedes leer más de esto (aunque a un nivel mas técnico) en la sección 5 de las notas de Sanz-Alonso et al. [3].

5.2. Ejemplo

El análisis del error en la sección anterior habla en del error cuadrático medio en el peor escenario posible bajo una familia de funciones de prueba (resumen). El ejemplo anterior muestra el error Monte Carlo cometido con respecto a una función resumen $f(\theta) = \theta$ con la cual, vemos, se reduce la varianza. Esto no contradice lo anterior pues para esta función resumen nuestra distribución instrumental satisface el criterio de reducción de varianza. En general, lo complicado es encontrar dicha distribución que podamos usar en la estimación Monte Carlo.

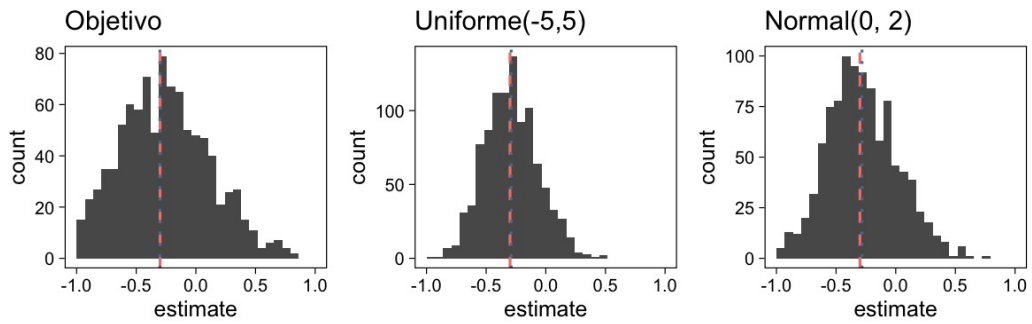

FIGURA 15. *Muestreo por importancia utilizando distintas distribuciones instrumentales. Distribución* bootstrap *de* π_N^{IS} *con* $B = 10,000$ *y* $n = 100$.

REFERENCIAS

[1] A. Johnson, M. Ott, and M. Dogucu. *Bayes Rules! An Introduction to Applied Bayesian Modeling*. 2021.

[2] S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, Cambridge, 2015. ISBN 978-1-107-06939-8 978-1-107-66391-6.

[3] D. Sanz-Alonso, A. M. Stuart, and A. Taeb. Inverse Problems and Data Assimilation. *arXiv:1810.06191 [stat]*, jul 2019.