FACULTAD DE CIENCIAS

Evaluación de p-valores extremadamente bajos por técnicas de muestreo por importancia

(Evaluating extremely low p-values by Importance sampling techniques)

Trabajo fin de grado para acceder al

GRADO EN MATEMÁTICAS

Realizado por: Jon López Fernández

Dirigido por: Francisco Matorras Weinig

Convocatoria de Junio, curso 2021/22

Agradecimientos

Me gustaría agradecer a mi tutor Francisco por darme la oportunidad y confiar en mi cuando me paseaba por los despachos de los profesores buscando alguien que me dirigiera, supo animarme cuando lo necesitaba y me dejo el espacio necesario para asi poder trabajar juntos, lo logramos.

Por otro lado me gustaría agradecer a mis padres, que gracias a ellos he tenido la oportunidad de poder especializarme en algo que de verdad me interesa depués de varios años de desilusión con el grado, y por darme su apoyo tanto mental como económico para poder venir a vivir aquí y disfrutar de este maravilloso sitio. Han hecho que dejara de lado todas mis inseguridades de lado y empezara a vivir, aprender y sobre todo disfrutar de cada cosa que hago.

No me quiero dejar en el olvido tampoco a mi hermana Bea, que a pesar de llevar tanto tiempo separados, la he sentido más cerca que nunca. En este año tan complicado para mí en lo personal por muchas razones, ha sabido siempre sacarme una sonrisa y entenderme, es una persona fundamental en mi vida y que ojala no me falte nunca.

Por ultimo quería agradecer a mis compañeros, tanto de la UC como los de la UPV los cuales me han apoyado, me han visto crecer, tanto en lo personal como en lo intelectual, pero sobre todo, por haberme soportado durante estos 4 largos años. Os estaré eternamente agradecido.

Resumen

En muchos campos de la ciencia es importante el cálculo preciso de p-valores de test estadísticos, generalmente asociados a contrastes de hipótesis. A menudo nos encontramos con que no se está seguro de la corrección de aplicar métodos asintóticos y se utilizan métodos de MonteCarlo. Sin embargo, en algunas aplicaciones los p-valores que se quieren calcular, hacen inviable este método en modo práctico.

Se ha estudiado la aplicación a este caso del llamado muestreo por importancia que teóricamente debería resolver este problema, aunque su aplicación práctica no es evidente.

En este trabajo, presentamos una propuesta de elección de funciones de muestreo y exploramos su aplicación, tanto desde el punto de vista teórico como con simulaciones de algunos ejemplos típicos con R.

Se comprueba que el método da resultados excelentes en algunos casos, permitiendo calcular p-valores muy pequeños con muestreos reducidos, aunque por otra parte se comprueba que el método no es universal.

Abstract

In many scientific fields the precise calculation of p-values of statistical tests, usually associated with hypothesis testing, is important. Often we find that we are not sure of the correctness of applying asymptotic methods that's why MonteCarlo methods are used. However, in some applications the calculus of the p-values make this method unfeasible in practical terms.

The application of the importance sampling method has been studied, which theoretically should solve this problem, although its practical application is not evident.

In this paper, we present a proposal for the choice of sampling functions and explore their application, from a theoretical point of view and with simulations in R of some typical examples.

It is shown that the method gives excellent results in some cases, allowing the calculation of very small p-values with reduced samples, although it is also shown that the method is not universal.

Índice general

1. Introducción

La estadística se ocupa fundamentalmente de la sistematización, recogida, ordenación y representación de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de hacer previsiones sobre los mismos, tomar decisiones u obtener conclusiones.

La física es la ciencia natural que se encarga de reconocer y estudiar a modo general el funcionamiento de los componentes principales del universo. Entre ellos la materia, el espacio, tiempo, la energía y las interacciones fundamentales que ocurren entre ellas.

Vamos a tratar de combinar estas dos ciencias para así poder dar solución al problema descrito anteriormente. Observar como la estadística es una herramienta fundamental a la hora de reconocer y evaluar todos los experimentos, descubrimientos, investigaciones, etc. que se hacen en el mundo de la física. Es aquí donde surge la idea de física estadística; La física estadística es una rama de la física que evolucionó a partir de una base de la mecánica estadística, que utiliza métodos de la teoría de probabilidad y la estadística, y en particular las herramientas matemáticas para tratar con grandes poblaciones y aproximaciones, para resolver problemas físicos.

A pesar de que hay muchos problemas de física estadística que pueden resolverse analíticamente mediante aproximaciones y expansiones, la mayoría de las investigaciones actuales utilizan la gran potencia de procesamiento de las computadoras modernas para simular o aproximar soluciones.

Un enfoque común para los problemas estadísticos es usar una simulación de Monte Carlo para obtener información sobre las propiedades de un sistema complejo.

La simulación de Montecarlo es un método enfocado en la resolución de problemas de carácter matemático a través de un modelo estadístico que consiste en generar posibles escenarios resultantes de una serie de datos iniciales. Este método trata de simular un escenario real y sus distintas posibilidades, permitiendo al usuario realizar una predicción del comportamiento de las variables según las estimaciones obtenidas con el método.

El Problema del Método de Monte Carlo es cuando llegamos a una variación de 2sigma donde ya no es válido el método, es entonces donde aparece nuestro metodo, el metodo de muestreo por importancia.

En estadística, el muestreo por importancia es una técnica general para estimar las propiedades de una determinada distribución, disponiendo únicamente de muestras generadas a partir de una distribución diferente a la de interés. El método fue introducido por primera vez por Gerald Goertzel, Herman Kahn y Theodore E. Harris en 1949, y está relacionado con el muestreo paraguas en la física computacional. Dependiendo de la aplicación, el término puede referirse al proceso de muestreo de esta distribución alternativa, al proceso de inferencia o a ambos.2.3

En la física de altas energías, utilizamos las pruebas estadísticas basadas en la verosimilitud, para el descubrimiento de nuevos fenómenos y para la construcción de intervalos de confianza de los parámetros del modelo.

Se derivan fórmulas explícitas para las distribuciones asintóticas de los estadísticos de las pruebas utilizando los resultados de Wilks y Wald. Uno de los resultados más conocidos sobre la máxima verosimilitud es que, el estadístico de la prueba de la razón de verosimilitud, tiene una distribución asintótica chi-cuadrado χ^2.

En los experimentos de física de partículas se suelen buscar procesos que se han predicho pero que aún no se han visto. La importancia estadística de una señal observada puede cuantificarse mediante un p-valor. Es útil para caracterizar la sensibilidad de un experimento dado de la significación esperada que se obtendría para una variedad de hipótesis de señal.

Encontrar tanto la significación para un conjunto de datos específico como la significación esperada puede implicar cálculos de Monte Carlo que son computacionalmente costosos.

El trabajo describe el formalismo de una búsqueda como prueba estadística y se definen con precisión los conceptos de significación estadística y sensibilidad. Varios estadísticos de prueba basados en la razón de verosimilitud. Utilizamos el Importance Sampling para hallar el estadístico de prueba y, a partir de ello, hallar los p-valores y las cantidades relacionadas para una muestra de datos dada. A continuación desarrollaremos el metodo de muestreo por importancia, de que trata y en que consta.

2. Contexto

2.1. Conceptos básicos

Se comienza haciendo un repaso de conceptos vistos en la asignatura "INFERENCIA ESTADÍSTICA". En esta sección se estudirán definiciones básicas que se manejarán a lo largo del resto del documento, como las de esperanza matemática, contraste de hipótesis o p-valor. Estos conceptos y resultados han sido tomados de la referencia [2]

Definición 2.1.1 *(σ-álgebra) Una familia de subconjuntos de X, representada por Σ, es una σ-álgebra sobre X cuando se cumplen las siguientes propiedades:*

1. *El conjunto vacío está en $\Sigma : \emptyset \in \Sigma$.*
2. *Si E está en Σ, también está su complementario $E^c = X \setminus E$.*
3. *Si $E_1, E_2, E_3, ...$ es una sucesión de elementos de Σ, entonces la unión (numerable) de todos ellos también está en Σ.*

Definición 2.1.2 *Un espacio de probabilidad es la terna $(\Omega, \mathcal{F}, \mathrm{P})$ donde el conjunto Ω es llamado espacio muestral y es el conjunto de los posibles resultados del experimento, $\mathcal{F}$ es una σ-álgebra de subconjuntos de Ω que satisface*

1. *$\Omega \in \mathcal{F}$.*
2. *Si $A \in \mathcal{F}$ entonces $A^c \in \mathcal{F}$.*
3. *Si $A_1, A_2, \ldots, A_n \in \mathcal{F}$ entonces $A_1 \cup A_2 \cup \cdots \cup A_n \in \mathcal{F}$.*

Definición 2.1.3 *Una variable aleatoria (v.a.) X es una función real definida en el espacio de probabilidad $(\Omega, \mathcal{A}, P)$, asociado a un experimento aleatorio.*

$$X : \Omega \to R$$

La noción básica de la teoría de la probabilidad es la de un experimento aleatorio o random experiment: un experimento cuyo resultado no puede determinarse de antemano.

Definición 2.1.4 *(Probabilidad) Una probabilidad P es una regla que asigna un número $0 \leq P(A) \leq 1$ a un evento A, tal que $P(\Omega) = 1$, y para cualquier secuencia $A_1, A_2, ...$ de eventos disjuntos*

$$P(\cup_i A_i) = \textstyle\sum_i P(A_i)$$

Definición 2.1.5 *Dada una variable aleatoria X, su función de distribución, $F_X(x)$*

$$F_X(x) = \mathrm{Prob}(X \leq x)$$

Definición 2.1.6 *Esperanza matemática*

- *Caso discreto*
  Para una variable aleatoria discreta X con función de probabilidad $\mathrm{P}[X = x_i]$ con $i = 1, 2, \ldots, n$ la esperanza se define como

$$\mathrm{E}[X] = \sum_{i=1}^{n} x_i \, \mathrm{P}[X = x_i]$$

- *Caso continuo*
  Para una variable aleatoria continua X con función de densidad $f_X(x)$ el valor esperado se define como la integral de Riemann

$$\mathrm{E}[X] = \int_R x f_X(x) dx$$

Definición 2.1.7 *Sea $g : R \longrightarrow R$ una función (continua) y $X : \Omega \longrightarrow R$ una variable aleatoria, podemos componer las dos funciones para obtener una nueva variable aleatoria $Y = g(X)$.*

$$\begin{array}{rccc} Y = g(X): & \Omega & \longrightarrow & R \\ & \omega & \longmapsto & g(X(\omega)) \end{array}$$

Se define la esperanza de $g(X)$ como,

- *Caso discreto*
  Sea X una v.a discreta, con función de probabilidad $\mathrm{P}[X = x_i]$ con $i = 1, 2, \ldots, n$ la esperanza se define como

$$\mathrm{E}[g(X)] = \sum_{i=1}^{n} g(x_i) \, \mathrm{P}[X = x_i]$$

- *Caso continuo*
  Sea X una v.a continua, con función de densidad $f_X(x)$ el valor esperado se define como

$$\mathrm{E}[g(X)] = \int_R g(x) f_X(x) dx$$

Definición 2.1.8 *Sea X una variable aleatoria con media $\mu = \mathrm{E}(X)$, se define la varianza de la variable aleatoria X, denotada por $\mathrm{Var}(X)$, σ_X^2 o simplemente σ^2 como*

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - \mathrm{E}^2[X]$$

2.1.1. Métodos para obtener estimadores

Método de máxima verosimilitud

Definición 2.1.9 *Dada una muestra aleatoria simple $(X_1, ..., X_n)$ se define la función de verosimilitud en $(x_1, x_2, ..., x_n)$ como $L(x_1, ..., x_n; \theta) = g(x_1, ..., x_n; \theta)$*

- *En el caso continuo,*
  $L(x_1, ..., x_n; \theta) = f(x_1; \theta)...f(x_n; \theta)$ *donde* $f(x_n; \theta)$ *es la función de densidad que describe a la variable aleatoria X, de la cual se desconoce algún parámetro.*
- *En el caso discreto,*
  $L(x_1, ..., x_n; \theta) = p(X_1 = x_1; \theta)...p(X_n = x_n; \theta)$ *donde* $p(x; \theta)$ *es la ley de probabilidad que describe a la variable aleatoria X, de la cual se desconoce algún parámetro.*

El método de máxima verosimilitud consiste en obtener el estimador $\theta^*(X_1, ..., X_n)$ (o estimadores θ_j^*, j = 1,2,..,k) maximizando la función de verosimilitud respecto de θ (o de θ_j, j=1,2,...,k).

$$máx_\theta L(x_1, ..., x_n; \theta)$$

Se despeja el parámetro desconocido θ(o los parámetros desconocidos θ_j y se tiene el estimador $\theta^* = \theta^*(X_1, ..., X_n)$ o estimadores $\theta_j^* = \theta_j^*(X_1, ..., X_n)$ para j=1,2,...,k).

En algunos casos, vamos a maximixar $LnL(x_1, ..., x_n; \theta)$, ya que la función Ln, logaritmo neperiano, es una trasformación monótona creciente y se conservan los máximos. Nos hara así funciones más fáciles de maximzar.

2.1.2. Contraste de Hipótesis

Introducción

Uno de los métodos más usados de investigación es el método de la contraste de hipótesis, en muchos casos la hipótesis se concentra en contrastar el valor de un parámetro. También hay otros tipos de contrastes llamados contrastes no paramétricos que son más generales, con los cuales no trabajaremos.

Para plantear un problema de contraste de hipótesis se han de determinar dos hipótesis, y a partir de los datos de una muestra:

- Decidir si una determinada hipótesis sobre una población es rechazable o no.
- Medir el error con el que se rechaza o no se rechaza dicha hipótesis.

Sea X una variable aleatoria cuya función de distribución $F(x, \theta_1, ..., \theta_k)$ depende de uno o de varios parámetros desconocidos $\theta_1, ..., \theta_k$.

El contaste de Hipótesis es el procedimientos para aceptar o rechazar una hipótesis que se emite acerca de un parámetro u otra característica de la población.

Pasos para realizar un contraste o test,

1. Identificar la hipótesis nula H_0 y la hipótesis alternativas H_1
   - H_0 es la hipótesis verdadera por defecto.
   - H_1 es la hipótesis que se acepta si se rechaza la hipótesis nula.
2. Construir un estadístico de prueba $T(X_1, ..., X_n)$ relacionado con el parámetro desconocido θ, a través de un estimador que discrimine entre H_0 y H_1.

$$\begin{cases} S_0 = & \{(x_1, ..., x_n) \in S_0 | T(X_1, ..., X_n) \text{ Cumple unas condiciones}\}. \\ H_1 = & \{(x_1, ..., x_n) \notin S_0 | T(X_1, ..., X_n) \text{ No cumple las condiciones}\} = S_0^c. \end{cases}$$

3. Se saca una muestra aleatoria de tamaño n, sea $x_1, ..., x_n$ (datos concretos) y se realiza el contraste o test,

$$\begin{cases} Si\ (x_1, ..., x_n) \in & S_0 \Rightarrow \text{se acepta } H_0 (\text{no se rechaza } H_0). \\ Si\ (x_1, ..., x_n) \in & S_1 \Rightarrow \text{Se rechaza } H_0. \end{cases}$$

Tipos de Contrastes

Hay dos tipos de hipótesis; Por un lado están las hipótesis **no paramétricas**, las cuales hacen referencia a una carácteristica de la población. Por otro lado, y en los cuales nosotros nos centraremos están los contrastes de hipótesis **paramétricos**, los cuales se utilizan para estudiar si una determinada afirmación acerca de un parámetro poblacional es confirmada o invalidada por los datos de una muestra extraída de dicha población.

- Contraste bilateral

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_1 \end{cases}$$

- Contraste simple

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_1 \end{cases}$$

- Contrastes unilaterales,
  - Contraste unilateral por la derecha

$$\begin{cases} H_0 : & \theta \leq \theta_0 \\ H_1 : & \theta > \theta_1 \end{cases}$$

  - Contraste unilateral por la izquierda

$$\begin{cases} H_0 : & \theta \geq \theta_0 \\ H_1 : & \theta < \theta_1 \end{cases}$$

Cálculo del p-valor

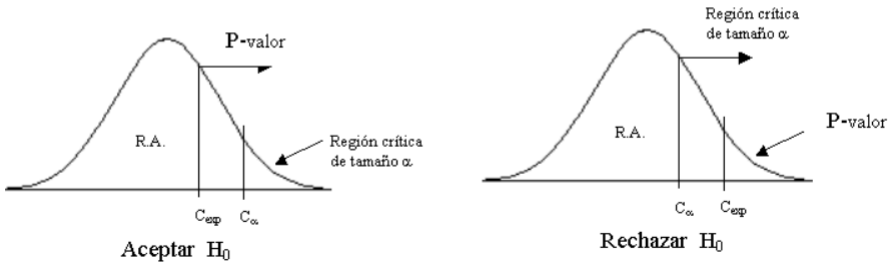
Definición 2.1.10 *Se llama p-valor o α_c crítico al mayor de los α tal que $T(X_1, ..., X_n) \in S_0$.*

$$p\text{-}valor = \alpha_c = max\{\alpha | T(X_1, ..., X_n) \in S_0\}$$

Es el límite para juzgar un resultado como estadísticamente significativo.

Si $\alpha_c < \alpha \Longrightarrow T(X_1, ..., X_n) \notin S_0$, siendo S_0 la región de aceptación, se considera que el resultado es estadísticamente significativo y por tanto se rechaza la hipótesis nula.

$$\boxed{\text{Si } \alpha < P \Rightarrow \text{Aceptar } H_0 \quad ; \quad \text{Si } \alpha \geq P \Rightarrow \text{Rechazar } H_0}$$



Intuitivamente, el p-valor se define como la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta. Si p cumple que es menor que el nivel de significación dado este se considera como resultado estadísticamente significativo $\Longrightarrow$ permite rechaza la hipótesis nula.

Test de la razón de verosimilitud

Sea una variable aleatoria X con función de distribución $F(x; \theta_1, ..., \theta_m)$ que depende de varios parámetros desconocidos. Se quiere efectuar el contraste general,

$$\begin{cases} H_0 : & (\theta_1, ..., \theta_m) \in W \\ H_1 : & (\theta_1, ..., \theta_m) \in W^c \end{cases}$$

La función de verosimilitud es $L(x_1, ..., x_n; \theta_1, ..., \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, ..., \theta_m)$ donde f es la función de densidad.

Sea L(W) el máximo de L en W y L(Ω) el máximo de L en $\Omega = W \cup W^c$ todo el espacio paramétrico.

Se llama razón de verosimilitud a, $\lambda = \frac{L(W)}{L(\Omega)}$

Como $L(W) \leq L(\Omega) \Longrightarrow 0 \leq \lambda \leq 1$.

Se rechaza H_0 si $\lambda < c$ para $0 < c \leq 1$, es decir, $C = \{(x_1, ..., x_n) | \lambda < c\}$.

Fijado α se determina c, $\alpha = sup_{(\theta_1,...,\theta_m)\in W} p(\lambda < c)$

Condiciones previas para poder aplicar el Teorema de Wilks; es necesario que el valor óptimo no esté en el extremo y que las hipótesis estén "anidadas", es decir, que H_0 esté contenida en H_1, que podamos conseguir H_0 dando valores a los parámetros de H_1. Si se cumplen las anteriores condiciones estamos en disposición de aplicar el siguiente Teorema;

Teorema 2.1.1 *Teorema de Wilks, Sea* $\{X_n\}$ *una muestra con función de distribución desconocida, si* $n \longrightarrow \infty \Rightarrow$ *la distribución del estadístico de prueba* $-2 \cdot \log(\wedge)$ *se aproxima asintóticamente a una distribución* $\approx \chi^2$ *bajo la hipótesis nula* H_0.

En este caso, $\wedge$, denota la razón de verosimilitud, la distribución χ^2 tiene grados de libertad iguales a la diferencia en dimensionalidad de Θ y Θ_0, dónde Θ es el espacio de parámetros completo y Θ_0 es el subconjunto del espacio de parámetros asociado con H_0.
Información extraída del árticulo [8] y la siguiente página web [17].

En general, se define la razón de verosimilitud como el cociente de las verosimilitudes de dos hipótesis. Para nuestro problema en concreto, definimos la razón de verosimilitud como

Definición 2.1.11 *La razón entre la posibilidad de observar un resultado extraordinario en una población en cuestión versus la posibilidad de no verlo. Para un ejemplo* $x_i (i = 1, 2, ..., n)$ *de tamaño n con función de densidad f(x,) se define como:*

$$\lambda = \frac{\prod_{i=1}^{n} f(x_i, \theta_0)}{\prod_{i=1}^{n} f(x_i, \theta_a)} = \frac{L(\theta_0)}{L(\theta a)}$$

.

Luego para, $k > 0$ *fijado el test de razón de verosimilitud entre la hipótesis nula* $H_0 : \theta_0$ *y la hipótesis alternativa* $H_1 : \theta_a$.

$$\begin{cases} Para \quad \lambda > k \quad se\ acepta \quad H_0 \\ Para \quad \lambda < k \quad se\ rechaza \quad H_0 \\ Para \quad \lambda = k \quad se\ toma\ cualquiera\ de\ las\ dos \end{cases}$$

Más información en [3].

2.2. Simulación de Montecarlo

La simulación de Montecarlo es un método enfocado en la resolución de problemas de carácter matemático a través de un modelo estadístico que consiste en generar posibles escenarios resultantes de una serie de datos iniciales.

Este método trata de simular un escenario real y sus distintas posibilidades, permitiendo al usuario realizar una predicción del comportamiento de las variables según las estimaciones obtenidas con el método.

Como ya sabemos, este método se basa en simular posibles escenarios, y lo hace generando números completamente aleatorios. Ahora bien, es importante entender que la simulación de Montecarlo es útil cuando generamos una gran cantidad de escenarios.

Para ello se han creado programas informáticos especializados que generan todos los números siguiendo distribuciónes específicas que representa las variables aleatorias que podrían darse en escenarios reales. [13]

2.2.1. Problemas del Método de Montecarlo

Nuestro problema va a requerir el cálculo de p-valores extremadamente bajos, en torno a 10^{-6} o menores, donde la simulación puede ser compleja, por lo que el cálulo por el muestreo normal se haría muy largo, impracticable en muchos casos.

Es por ello, por lo que No utilizamos el método de Montecarlo directo, ya que es muy ineficiente para el caso que nosotros estamos analizando. Tendríamos que realizar muchísimas iteraciones y creaciones de números aleatorios para que tan solo unos poco de ellos se encontraran donde nosotros queremos, es por eso, por lo que mediante el método de muestreo por importancia vamos a crear valores justamente en el lugar donde nosotros queremos obtener información y mediante la asignación de unos respectivos pesos(los cuales tenemos que calcular para cada punto) daremos más o menos importancia a estos según el lugar en el que nos encontremos. Creando así información justo donde nosotros queremos y desechando la información restante.

En nuestro caso en concreto el cálculo del p-valor va a estar definido sobre q, nuestro test estadístico, y el p-valor no se calcula (en general) de una forma sencilla. Normalmente hay dos formas.

- La primera es utilizar la convergencia asintótica prevista por el teorema de Wilks, que dice que cuando se cumplen ciertas condiciones q tiende a una distribución de χ^2 con un número de grados de libertad igual a la diferencia entre los del modelo de H_1 y el de H_0.
- La alternativa es hacerlo por muestreo, mediante pesudoexperimentos. Fijado un número c cualquiera, se basa en repetir por muestreo el experimento según una ley de distribución de la hipótesis nula, y calcular la proporción de ellos que son mayores que c.

Otro de los problemas asociados a la simulación de Montecarlo se plantea cuando estamos ante un "descubrimiento", cuando la hipótesis nula no es cierta. Uno se pregunta entonces ¿me puedo creer Wilks a un nivel de 10^{-7} para este caso? Porque normalmente generar del orden de 10^8 simulaciones es intratable, porque cada una de ellas implica cálculos complicados, normalmente ajustes.

La computación de tal cantidad de experimentos es demasiado costosa en tiempo, es por ello por lo que nosotros queremos llegar a resultados más óptimos en casos tan extremos. La idea es utilizar una muestra reducida, de tan solo unos pocos experimentos,y poder sacar las mismas conclusiones. Es aquí donde encontramos el problema del Metodo de Montecarlo y el porqué de analizar el método de muestreo por importancia.

A continuación vamos a ver más gráficamente, mediante un ejemplo simple, los problemas que tiene el método de Montecarlo directo y como utilizando el método de muestreo por importancia podemos darle solución.

Ejemplo

Un ejemplo básico para entender las limitaciones del método de Montecarlo para estimar valores es el siguiente.

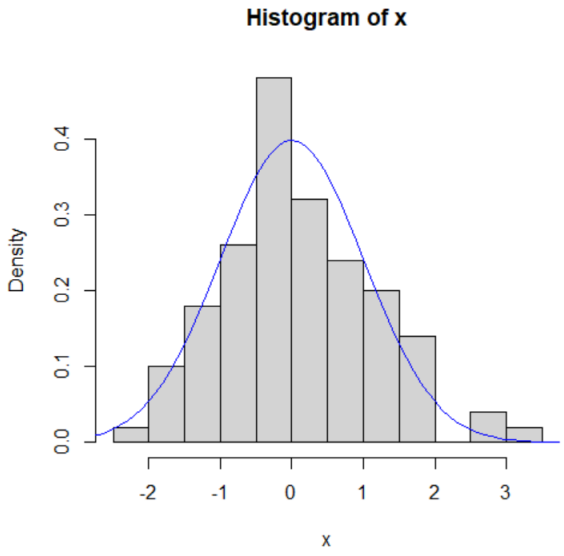
```
> x=rnorm(100,0,1)
> mean(x>3)
[1] 0.01
> pnorm(3,0,1, lower.tail=FALSE)
[1] 0.001349898
```

Figura 2.1: Ejemplo básico

Buscamos calcular el p-valor de una distrución normal a partir de un c fijado. Utilizamos una muestra reducida de 100 puntos tomados aleatoriamente de una distribución normal mediante el método de Montecarlo simple. Para calcular el p-valor, contamos cuántos de esos puntos son mayores que el punto c fijado.

Comenzamos fijando c=3. Figura 2.1

Por un lado sabemos que el área que deja una distribución normal;$N(0,1)$ a la derecha de c es igual a 0,001349898. Por otro lado, utilizando el método de Montecarlo como método de muestreo, para una muestra de 100 valores aleatorios de una normal: $N(0,1)$, hemos obtenido un valor de 0.01.



En este caso tan simple, lo que observamos es que para una muestra tan pequeña, las diferencias son significativas, por lo que no sería tan bueno. Luego el método de Monte Carlo directo para un número tan reducido de n no es preciso, y es por eso por lo que introducimos el método de muestreo por importancia.

Veremos luego este mismo ejemplo con el método de muestreo por importancia, y veremos como con el mismo número de observaciones obtendremos un resultado mucho mejor.

2.3. Muestreo por importancia

El muestreo de importancia es una técnica de reducción de la varianza que se utiliza en el método de Montecarlo. La idea detrás del muestreo de importancia es que ciertos valores de las variables aleatorias de entrada en una simulación tienen más impacto en el parámetro que se estima que otros. Si estos valores "importantes"se enfatizan mediante el muestreo con mayor frecuencia, entonces la varianza se puede reducir. Por tanto, la metodología básica en el muestreo de importancia es elegir una distribución que "fomente"los valores importantes. Este uso de distribuciones "sesgadas"dará como resultado un estimador sesgado si se aplica directamente en la simulación. Sin embargo, los resultados de la simulación se ponderan para corregir el uso de la distribución sesgada, y esto asegura que el nuevo estimador de muestreo de importancia sea insesgado. El peso viene dado por la razón de verosimilitud , es decir, la derivada Radon-Nikodym de la verdadera distribución subyacente con respecto a la distribución de simulación sesgada.

La cuestión fundamental en la implementación de la simulación de muestreo de importancia es la elección de la distribución sesgada que fomenta las regiones importantes de las variables de entrada. Elegir o diseñar una buena distribución sesgada es el "arte"del muestreo de importancia. Las recompensas por una buena distribución pueden ser enormes ahorros de tiempo de ejecución; la penalización por una mala distribución puede ser tiempos de ejecución más largos que para una simulación general de Monte Carlo sin muestreo de importancia. [16]

Comenzaremos con un ejemplo sencillo para obtener así una idea intuitiva y visual del método, para luego poder explicarlo de manera más técnica.

Caso general

Supongamos que $f : R^d \to R$ y que se requiere evaluar la integral

$$I = \int_{\Theta} f(\mu)\, d\mu.$$

Claramente I también puede escribirse como

$$I = \int_{\Theta} \left\{ \frac{f(\mu)}{g(\mu)} \right\} g(\mu)\, d\mu,$$

donde $g(\theta)$ es una función de densidad de probabilidad sobre Θ. En otras palabras, $I = E_g(\frac{f(\mu)}{g(\mu)})$. La distribución $g(\mu)$ se conoce como la distribución de muestreo por importancia y generalmente se elige de manera que sea fácil de simular. Debería valer para cualquier g que represente una probabilidad, siempre que $g(\mu)$ no sea 0 donde $f(\mu) \neq 0$.

Si ahora generamos una muestra $\mu_1, \ldots, \mu_N$ de $g(\mu)$ entonces podemos aproximar la integral I a través del estimador insesgado

$$\hat{I}_{MI} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\mu_i)}{g(\mu_i)}.$$

La varianza de este estimador está dada por

$$\text{Var}(\hat{I}_{MI}) = \frac{1}{N} \text{Var}_g(\frac{f(\mu)}{g(\mu)}) = \frac{1}{N}\{E_g(\frac{f(\mu)^2}{g(\mu)^2}) - I^2\}.$$

La precisión de $\hat{I}_{MI}$ depende tanto del tamanõ de muestra, N, como de la distribución de muestreo por importancia, $s(\mu)$. De hecho, si $s(\theta)$ se elige proporcional a $f(\mu)$ entonces Var $(\hat{I}_{MI}) = 0$ sin importar el tamaño de muestra.

En la práctica dicha elección no es posible, pero esta idea sugiere que $s(\mu)$ debe tener una forma similar a la de $f(\mu)$, excepto tal vez en regiones donde los valores de $f(\mu)$ sean despreciables.

Tenemos entonces que, para una integral I dada, existe una infinidad de estimadores insesgados, en principio con precisiones distintas. Un aspecto importante del método de Monte Carlo se refiere al diseño de técnicas de reducción de varianza para dichos estimadores. Una de las técnicas más sencillas consiste precisamente en elegir una distribución de muestreo por importancia adecuada. Generalmente se requiere que $g(\mu)$ satisfaga las siguientes condiciones:

- Debe ser fácil de simular;
- Debe tener una forma similar a la de $f(\theta)$, la función que se desea integrar;
- Debe tener las colas más pesadas que $f(\theta)$, pues de otra forma la varianza de $\hat{I}_{MI}$ podría llegar a ser muy grande o incluso infinita.

Está información ha sido obtenida de la referencia [15]

Descripción detallada

En nuestro problema, aplicamos el muestreo por importancia en el caso particular del cálculo del p-valor.

Procedemos ahora con la explicación más detallada del método. El muestreo por importancia es un método de Montecarlo la técnica más fundamental de la reducción de la varianza.

Sea,

$$l = E_f[H(X)] = \int H(X) \cdot f(X)dx$$

Sea g otra función de densidad que domina a H_f, si $g(x) = 0 \Rightarrow H(x) \cdot f(x) = 0$. Usando la función de densidad g podemos definir l como,

$$l = \int H(X) \cdot \frac{f(X)}{g(X)} \cdot g(X) = E_g[H(X) \cdot \frac{f(X)}{g(X)}]$$

A esta función de densidad le llamamos "Importance Sampling"density, proposal density or instrumental density.

Si $X_1, ..., X_N$ es un muestreo aleatorio de g, con X_i v.a.i.i.d con función de densidad g, luego una estimación de l es,

$$\hat{l} = \frac{1}{N} \sum_{k=1}^{N} H(x_k) \cdot \frac{f(x_k)}{g(x_k)}$$

$W(X) = \frac{f(X)}{g(X)} \longrightarrow$ Razón de verosimilitud o peso. Es por eso también que $\hat{l}$ es llamado **estimador de la razón de verosimilitud**.

En el caso particular en el que no hay cambio de medida, f = g $\Rightarrow$ W = 1, y el estimador de la razón de verosimilitud se limita a ser el mismo.

Método de minimización de la Varianza

Queremos minimizar la varianza de $\hat{l}$ con respecto de g

$$\text{mín}_g \quad Var_g \cdot (H(X) \cdot \frac{f(X)}{g(X)})$$

La solución es

$$g^* = \frac{|H(X)| \cdot f(X)}{\int |H(X)| \cdot f(X)dx} \tag{2.1}$$

En el caso particular en el que H(X) $\geq$ 0, entonces

$$g^* = \frac{H(X) \cdot f(X)}{l} \tag{2.2}$$

Comprobamos calculando la varianza,

$$Var_{g^*}(\hat{l}) = Var_{g^*}(H(X) \cdot W(X)) = Var_{g^*}(l) = 0 \quad \#$$

Es importante darse cuenta de que, aunque es un estimador insesgado (aquel cuya esperanza matemática coincide con el valor del parámetro que se desea estimar) para cualquier pdf g que domine H· f, no todas esas pdfs son apropiadas. Una de las principales reglas para elegir una buena pdf de muestreo de importancia es que el estimador debe tener una varianza finita. Esto es equivalente a

$$E_g[H^2(X)\frac{f^2(X)}{g^2(X)}] = E_f[H^2(X)\frac{f(X)}{g(X)}] < \infty \tag{2.3}$$

Esto sugiere que g no debería tener una"cola más ligera"que f y la relación de probabilidad, f/g debería estar acotada.

En general, la implementación de la densidad óptima del muestreo de importancia g* según 2.1 y 2.2 es problemática. La principal dificultad reside en el hecho de que para derivar g*(x) es necesario conocer l. Pero l es precisamente la cantidad que queremos estimar a partir de la simulación.

En el caso en el que la pdf f pertenece a alguna familia de distribución paramétrica, se elige la distribución de muestreo de importancia de la misma familia. En particular, supongamos que $f(\cdot) = f(\cdot; v)$ pertenece a la familia

$$f(\cdot; v, v \in V)$$

entonces el problema de encontrar una densidad de muestreo de importancia óptima en esta clase se reduce al siguiente problema de minimización paramétrica

$$min_{v \in V} Var_v(H(X)W(X; u, v)), \tag{2.4}$$

Donde W(X;u,v) =f(X;u)/f(X;v). Llamaremos al vector v el vector de parámetros de referencia. Como bajo $f(\cdot; v)$ la expectativa $l = E_v[H(X)W(X; u, v)]$ es constante, la solución óptima de 2.4 coincide con la de

$$min_{v \in V} V(v), \tag{2.5}$$

donde,

$$V(v) = E_v[H^2(X)W^2(X; u, v)] = E_u[H^2(X)W(X; u, v)]. \tag{2.6}$$

Ref [1]

Vamos a transformar la minimización funcional en una paramétrica, que es más sencilla. Incluso se puede hacer a partir de una muestra, como más tarde veremos.

Nuestro caso particular

Supongamos que tenemos una variable aleatoria $\vec{x}$, cada una de las componentes serán variables i.i.d que siguen la misma ley. Vamos a suponer que nuestra hipótesis alternativa sigue una ley $\rho(\vec{x}|\mu)$ y la nula $\rho(\vec{x}|\mu = 0)$, μ es el parámetro que ajustaremos de nuestros datos y que si lo mandamos a cero nos da nuestro H_0.

Tenemos nuestro test estadístico

$$\boxed{q(\vec{x}) = -2log(\frac{\rho(x|\vec{\mu}=0)}{\rho(x|\vec{\mu}_{best}}))}$$

μ_{best} es el que maximiza el cociente, el resultado del ajuste. Llamamos q_0 al estimador que corresponde a nuestra muestra de referencia $\vec{x}_0$, nuestros "datos", $q_0 = q(\vec{x}_0)$.

Queremos calcular el p-valor de esta medida, o sea si muestreamos de $\rho(\vec{x}, \mu = 0)$, cuántas veces q será mayor que q_0. Entonces, podemos plantear nuestro problema de cálculo del p-valor, como calcular el valor esperado del estimador $\theta(q(\vec{x}), q_0)$, siendo θ la función salto, 1 para valores positivos, 0 para negativos.

$$E[\theta(q(\vec{x}) - q_0)]$$

Luego volviendo atrás $\theta(q(\vec{x}) - q_0)$ sería nuestro H(X). Por lo que la función g* sería en este caso la propia función inicial f(X) para los valores que cumplen $q(\vec{x}) > q_0$ y cero en los demás.

2.4. Planteamiento del problema

Tenemos lo que podemos llamar un análisis o un experimento, que consiste en la recolección de distintas medidas/observaciones llamadas evento o suceso, (un muestreo en lenguaje de estadística). Cada observación corresponde a varias variables, pero normalmente se condensan en una. Es decir, trabajaremos con una muestra $\vec{x_i}$ (de tamaño fijo), que queremos contrastar con un modelo de física "conocida" (normalmente nuestra hipótesis nula H_0) y un modelo de física alternativa (H_1).

Hay una ley de probabilidad que corresponde a todo el experimento. Aunque no es necesario se pueden introducir correlaciones, casi siempre a través de nuisances(ruido). Pero en este caso, vamos a suponer que son independientes e idénticamente distribuidas (i.i.d), de modo que la función de densidad de probabilidad (pdf) del experimento es,

$$P(\vec{x}) = \prod_i \rho(\vec{x}_i)$$

En muchos casos se hacen histogramas, se cuentan cuántos casos han caído entre x y x+Δx, y se construye la ley de probabilidad ρ' a partir de estos casos.

Ejemplo

Si suponemos que cada canal del histograma sigue una ley de poisson, con una media dada por nuestro modelo.

$$P(\vec{n}) = \prod_i \frac{1}{n_i!}\lambda_i^{n_i} * e^{-\lambda_i}$$

y los λ_i vienen dados por el modelo, son función de nuestros POI (parameters of interest). En el primer caso se suele llamar "unbinned" y al segundo "binned".

Si es **unbinned**, sabremos escribir analíticamente $\rho(x|\vec{\mu})$, siendo $\vec{\mu}$ los parámetros de nuestro modelo, sea el de H_0 o el de H_1. Normalmente el modelo de H_0 está contenido en H_1 y se puede construir fijando alguno de los parámetros de H_1 (típicamente a 0), a esto se le llama que está "nested". Por ejemplo, H_0 puede ser una distribución exponencial E, y el H_1 la exponencial con una normal N superpuesta, (1-α)*E+α*N(μ,σ). H_0 es H_1 cuando $\alpha = 0$.

Para los **binned**, los λ, el valor esperado en cada bin, se producen por simulación, en función de los parámetros del modelo, casi siempre H_0 no tiene parámetros libres y H_1 es la superposición de ese H_0 con una muestra simulada, multiplicada por lo que se llama una "signal strength", o sea sacas λ_i de simulación y para H_1; $\lambda_i' = \lambda_i + \delta\mu_i$ donde los μ se sacan de otra simulación.

En resumen tendremos un modelo para la pdf, que potencialmente depende de algunos parámetros y otro modelo alternativo que además de estos parámetros contie-

ne uno o más (usaremos solo uno) que caracteriza cuánto de "nueva física" hay. Casi siempre se usa como test estadístico un likelihood ratio, cociente de las verosimilitudes que mejor ajustan a tus datos.

Para nuestro caso **unbinned** sería,

$$q = \sum_i [log(\rho_0(x_i|\mu = 0)) - log(\rho_1(x_i|\mu))]$$

Para el **binned**

$$q = \sum_i [n_i\lambda_i - \lambda_i - n_i\lambda_i' - \lambda_i']$$

Lo que buscamos es el p-valor. Antes lo hemos definido de manera genérica, pero a lo largo de todo el trabajo vamos a definir el p-valor como la probabilidad de que dada la hipótesis H_0, tengamos un valor de q igual o mayor que el que vemos en datos, con la idea de poder rechazar esta hipótesis (y por tanto poder decir que ha habido un descubrimiento, porque el modelo "conocido" no representa a nuestros datos). Es decir, se define el p-valor como:

$$p - valor = pv = \int_{(q(x)>qo)} \rho(x|\mu)dx = \int_\Omega \theta(q(x) - q_0)\rho(x|\mu)dx = E[\theta(q(x) - q_0)]$$

$$\text{Donde } \theta(q(x) > q_0) = \begin{cases} 1 & si \quad q(x) - q_0 \geq 0 \\ 0 & en \quad caso \quad contrario \end{cases}$$

Se decide si rechazamos o no la hipótesis nula en base a un CL prefijado. El problema se plantea cuando estamos ante un "descubrimiento", cuando la hipótesis nula no es cierta. Por razones históricas, en el campo de la fisica, para poder hablar de descubrimiento, tienes que tener un p-value equivalente al menos a 5σ. En la física de altas energías (también llamada física de partículas) el valor estándar requerido para poder anunciar un "descubrimiento" es p=0,0000003 a lo que llamamos 5σ. [12]

¿Cómo se calcula este p-value?

Tengamos en cuenta que es sobre q, sobre nuestro test estadístico, que no se calcula (en general) de una forma sencilla. Normalmente hay dos formas.

- La primera es utilizar la convergencia asintótica prevista por el teorema de Wilks, que dice que cuando se cumplen ciertas condiciones (Que N tienda a infinito y que los modelos sean nested) este q tiende a una distribución de χ^2 con un número de grados de libertad igual a la diferencia entre los del modelo de H_1 y el de H_0.
- La alternativa es hacerlo por muestreo, pesudoexperimentos. Repites por muestreo tu experimento según tu ley, binned o unbinned.

Uno se plantea entonces ¿me puedo creer Wilks a un nivel de 10^{-7} para este caso? Porque normalmente generar del orden de 10^8 simulaciones es intratable, porque cada una de ellas implica cálculos complicados, normalmente ajustes...

Aquí es donde entra nuestra idea de hacer el importance sampling, produciríamos pseudoexperimentos de acuerdo con otra distribución (cuál, es el quid de la cuestión) y sumaríamos los pesos de aquellos cuya q sea mayor que q_0, la que da nuestra muestra real. O sea,

1. Calculamos el likelihood ratio para nuestro experimento, q_0.
2. Realizamos N pseudoexperimentos según otra pdf, calculamos w_j y q_j.
3. Calculamos nuestro estimador, en este caso es el p-valor, que es simplemente "contar los pesos" de aquellos con $q > q_0$.

2.5. Elección de la función de probabilidad de muestreo

En nuestro caso esta definición de la función g no sirve, porque l es justamente lo que queremos calcular. Como no podemos calcularlo, la idea intuitiva que proponemos es coger justamente como función de distribución de muestreo, aquella que mejor se ajuste a nuestra observación, es decir, estimar el parámetro que mejor se ajuste a H_1. Coger como familia H_1 y escoger el parámetro que mejor se ajusta y luego aplicar el muestreo por importancia para ese parámetro obteniendo. Con la intención de obtener un estimador óptimo.

La clave está en escoger la pdf alternativa que nos permita calcular este p-valor con menor varianza. Dado que andamos buscando demostrar que nuestros datos son muy extremos y queremos ver los $q > q_0$, parece lógico usar una pdf que se parezca a esos datos. Ya que tenemos esta familia, parece razonable usar el alfa que mejor se ajusta a los datos. Esto se soporta en los argumentos que vamos a ver a continuación.

Objetivo: Demostrar que nuestra idea de coger el parámetro de muestreo como ese que mejor se ajusta a los datos hace que el método sea ÓPTIMO y que siempre FUNCIONA.

2.5.1. Problema de elección de g(x)

A lo largo del trabajo, vamos a tratar con problemas con un único parámetro desconocido, en este caso el parámetro desconocido es μ.
Para analizar los problemas que surgen a la hora de seleccionar g(x), vamos a tomar como ejemplo el caso particular $\mu = 0$, nuestra hipótesis nula H_0.
Sea $\{\vec{x}\}$ una variable aleatoria en la que cada componente es independiente e identicamente distribuida que siguen la misma ley. Suponemos nuestro contraste de hipótesis de la siguiente manera:

$$\begin{cases} H_0 & = \rho(\vec{x}|\mu = 0) \\ H_1 & = \rho(\vec{x}|\mu) \end{cases}$$

μ es el parámetro que ajustaremos de nuestros datos y
que si lo mandamos a cero nos da nuestro H_0.

En este caso nuestro problema consiste en primero encontrar el mejor μ que ajusta nuestra muestra, a este le llamaremos μ_{best} y será el que maximice la verosimilitud, y luego, estudiar el estadístico que compara la H_1(con este μ que hemos encontrado) y $H_0(\mu = 0)$:

$$\boxed{q(\vec{x}) = -2log(\frac{\rho(\vec{x}|\mu=0)}{\rho(\vec{x}|\mu=\mu_{best})})}$$

*Recordemos que en la mayoría de los casos las x son independientes, entonces ρ se factoriza y tenemos suma de logaritmos.

Llamaremos q_0 al estimador que corresponde a nuestra muestra de referencia $\vec{x}_0$, es decir nuestros "datos": $q_0 = q(\vec{x}_0)$.

Recordemos que nuestro objetivo es calcular el p-valor de esta medida, o sea, si muestreamos de nuestro H_0, es decir, $\rho(\vec{x}|\mu = 0)$, ¿Cuántas veces será q mayor que q_0?. Entonces, podemos plantear nuestro problema de calcular el p-valor como calcular el valor esperado del estimador $\Theta(q(\vec{x}) - q_0)$, donde Θ es la siguiente función:

$$\Theta = \begin{cases} 1 & \text{si } q(\vec{x}) - q_0 > 0 \\ 0 & \text{en caso contrario} \end{cases}$$

Calculando la esperanza del estimador obtenemos lo siguiente:

$$\begin{aligned} E[pv] &= \int \Theta(q(\vec{x}) - q_0)\rho(\vec{x}|\mu = 0)dx \\ &= \int \Theta(q(\vec{x}) - q_0)\frac{\rho(\vec{x}|\mu = 0)}{\rho(\vec{x}|\mu)}\rho(\vec{x}|\mu)dx \\ &= \int \Theta(q(\vec{x}) - q_0)W(\vec{x}|\mu)\rho(\vec{x}|\mu)dx \\ &= E_u[\Theta W_u] \end{aligned} \tag{2.7}$$

Aquí es donde aparece el importance sampling, en vez de hacer el valor esperado sobre la muestra inicial, lo hacemos sobre otra muestra, aplicando un peso.

De acuerdo con lo explicado en la sección (2.3), vamos a tomar la familia de funciones dada por H_1, es decir, $\rho(\vec{x}|\mu)$. Vamos a demostrar como coger $\mu = \mu_{best}$ es una buena opción. Vamos allá con la demostración.

La varianza del pv calculado con una muestra $\rho(\vec{x}|\mu)$, será:

$$Var[pv] = \quad E_\mu[(\Theta W_\mu)^2] \quad - \underbrace{(E_\mu[\Theta W_\mu])^2}_{\text{No depende de } \mu \text{ (2.1)}} \tag{2.8}$$

Por lo que nos basta encontrar el valor que minimice el primer término $E_\mu[(\Theta W_u)^2]$

$$\begin{aligned}
E_u[(\Theta W_\mu)^2] &= \int \Theta^2 W_\mu^2 \rho(\vec{x}|\mu) dx \\
&= \int \Theta^2 \frac{\rho(\vec{x}|\mu=0)^2}{\rho(\vec{x}|\mu)^2} \frac{\rho(\vec{x}|\mu)}{\rho(\vec{x}|\mu=\mu_{best})} \rho(\vec{x}|\mu=\mu_{best}) dx \\
&= \int \Theta^{\not 2} \frac{\rho(\vec{x}|\mu=0)^2}{\rho(\vec{x}|\mu)\cancel{\rho(\vec{x}|\mu)}} \frac{\cancel{\rho(\vec{x}|\mu)}}{\rho(\vec{x}|\mu=\mu_{best})} \rho(\vec{x}|\mu=\mu_{best}) dx \\
&= \int \Theta \frac{\rho(\vec{x}|\mu=0)}{\rho(\vec{x}|\mu)} \frac{\rho(\vec{x}|\mu=0)}{\rho(\vec{x}|\mu=\mu_{best})} \rho(\vec{x}|\mu=\mu_{best}) dx \\
&= \int \Theta W(x|\mu) W(\vec{x}|\mu=\mu_{best}) \rho(x|\mu=\mu_{best}) dx
\end{aligned} \tag{2.9}$$

Es decir, vamos a definir la función $C(\mu)$ como:

$$C(\mu) = Var_\mu(pv) + pv^2 = \int \Theta W(x|\mu) W(\vec{x}|\mu=\mu_{best}) \rho(x|\mu=\mu_{best}) dx$$

Vamos a intentar resolverlo de dos maneras;por un lado, podemos calcular la estimación de $Var_\mu(pv) + pv^2$, por muestreo y, aunque sea un poco "pobre" restringir nuestro muestreo a una única muestra, nuestro x_0. Entonces

$$C(\mu) \approx \Theta(q(x_0 - q_0)) W(\vec{x}_0|\mu) W(\vec{x}_0|\mu=\mu_{best}) = W(\vec{x}_0|\mu) W(\vec{x}_0|\mu=\mu_{best})$$

Será mínimo cuando $W(\vec{x}|\mu)$ sea mínimo. Esto sucederá cuando $\rho(\vec{x}|\mu)$ sea máximo, lo que sabemos que pasa en $\mu = \mu_{best}$. Por lo tanto, nos dice que dentro de esta familia de funciones, la mejor, es decir, la que nos da una varianza menor, corresponde exactamente con $\mu = \mu_{best}$.

Podemos incluso saber cuanto de mejor es la varianza, o mas bien C, si muestreamos con μ_{best}, en vez de muestrear con H_0

$$\frac{C(0)}{C(\mu)} \approx \frac{W(\vec{x}_0|\mu=0)W(\vec{x}_0|\mu=\mu_{best})}{W(\vec{x}_0|\mu=\mu_{best})^2} = \frac{W(\vec{x}_0|\mu=0)}{W(\vec{x}_0|\mu=\mu_{best})}$$

W, para $\mu = 0$ es 1 (por ser el cociente de densidades de probabilidad comparado con H_0), entonces:

$$\frac{C(0)}{C(\mu)} \approx \frac{1}{W(\vec{x}_0|\mu=\mu_{best})}$$

Con la definición de W

$$\frac{C(0)}{C(\mu)} \approx \frac{\rho(\vec{x}_0|\mu=\mu_{best})}{\rho(\vec{x}_0|\mu=0)}$$

Y con la definición de q_0

$* \; q(\vec{x}_0) = -2log(\frac{\rho(\vec{x}_0|\mu=0)}{\rho(\vec{x}|\mu=\mu_{best})}) \Rightarrow \frac{q(\vec{x}_0)}{2} = log((\frac{\rho(\vec{x}_0|\mu=0)}{\rho(\vec{x}|\mu=\mu_{best})})^{-1}) \Rightarrow \frac{q(\vec{x}_0)}{2} = log(\frac{\rho(\vec{x}|\mu=\mu_{best})}{\rho(\vec{x}_0|\mu=0}))$ $\Rightarrow e^{\frac{q_0}{2}} = \frac{\rho(\vec{x}|\mu=\mu_{best})}{\rho(\vec{x}_0|\mu=0})$

Luego;

$$\boxed{\frac{C(0)}{C(\mu)} \approx e^{q_0/2}}$$

Este método es interesante para casos de muy bajo p-value, o sea alto q_0, ya que reduciría la varianza en un factor muy grande.

Punto débil de la demsotración: reemplazamos la integral por la estimación con un único dato. En muchos casos, no me parece que sea un problema, porque nuestro "dato" es en realidad una muestra con varios datos de la misma distribución.

Vamos a intentar demostrarlo ahora mediante la integral. Lo que buscamos es el μ que nos minimice la varianza, esta será mínima cuando $\frac{\partial C(\mu)}{\partial \mu} = 0$. Veamos entonces para que valores de μ se hace cero la derivada anterior:

$$\begin{aligned} \frac{\partial C(\mu)}{\partial \mu} &= \frac{\partial Var(pv)}{\partial \mu} + \cancel{\frac{\partial pv^2}{\partial \mu}} \\ &= \int \Theta \frac{\partial W(\vec{x}\ \mu)}{\partial \mu} W(\vec{x}|\mu=\mu_{best})\rho(\vec{x}|\mu=\mu_{best})dx \end{aligned} \tag{2.10}$$

Nosotros sabemos que $\frac{\partial W(\vec{x}_0|\mu)}{\partial \mu}|_{\mu=\mu_{best}} = 0$ pero sabemos que esto es para x_0 y no

para cualquier x.

Básicamente queremos el promedio sobre muestras de datos con $\mu = \mu_{best}$

$$\frac{\partial C(\mu)}{\partial \mu} = E[\Theta \frac{\partial W(\vec{x}|\mu)}{\partial \mu} W(\vec{x}|\mu = \mu_{best})] \tag{2.11}$$

Y es por ello que consideramos razonable escoger $\mu = \mu_{best}$ para realizar el muestreo por importancia.

2.5.2. Limitaciones de la propuesta

Sabemos que lo anterior no es una demostración propiamente dicha, ya que podemos encontrar ciertas limitaciones en las que no sería válida.

Veamos ahora para que casos el método no es válido o por que no sirve la pseudodemostración en esos casos:

$$pv = \int_{q(\vec{x})>q_0} \rho(\vec{x}; \mu = 0)dx = \int_{q(\vec{x})>q_0} \frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu)} \cdot \rho(\vec{x};\mu)dx$$

Si calculamos por tanto la varianza

$$\int_{q(\vec{x})>q_0} (\frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu})^2 \cdot \rho(\vec{x};\mu)dx = \int_{q(\vec{x})>q_0} \frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu)} \cdot \rho(\vec{x};\mu = 0)dx$$

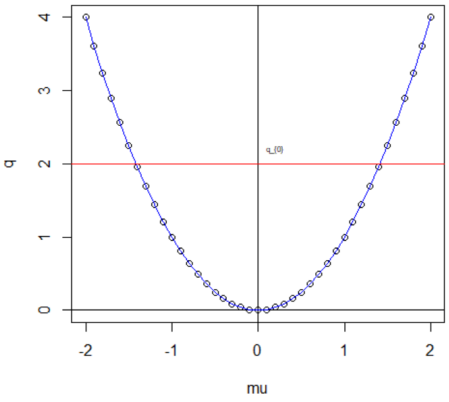


Figura 2.2: q en función de $\mu \in R$

Los problemas de la idea anterior comienzan en estos casos en los que obtenemos el mismo valor de q para diferentes valores de μ muy diferentes. En estas situaciones mediante la utilización del metodo de muestreo por importancia estamos haciendo un barrido de aquellos μ positivos por lo que nos estamos olvidando de toda la parte negativa.

En la imagen, se observa como en este caso la la solución al problema sería simplemente multiplicar el resultado por dos, pero puede darse el caso de que la distribución no sea simétrica por lo que en estos casos el factor por el que multiplicaríamos no sería exactamente 2, por lo que para estos casos tendríamos que precisar más en la solución.

A continuación veremos un ejemplo en el que nuestro parámetro desconocido, en ese caso, lo llamamos α, solo puede tomar valores positivos por lo que no tendríamos este problema y la demostración entonces sí, sería válida.

En este caso en concreto q en función de α es una aplicación monótona y creciente, evitando así los problemas anteriores y obteniendo una solución precisa.

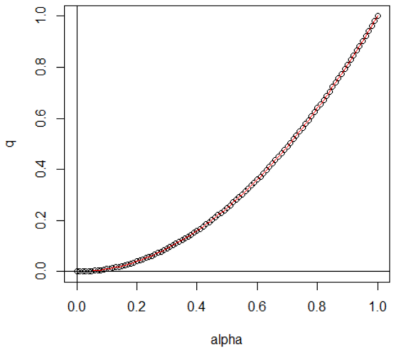


Figura 2.3: α toma valores en R^+

3. Aplicación del método

3.1. Aplicación a ejemplos sencillos

Como hemos dicho antes, el propósito de la implementación de este método en la estimación de los p-valores es para poder determinar en esos casos extremos en los que no sabemos si el teorema de Wilks es válido o no. Para ello empezamos desarrollado código en R.

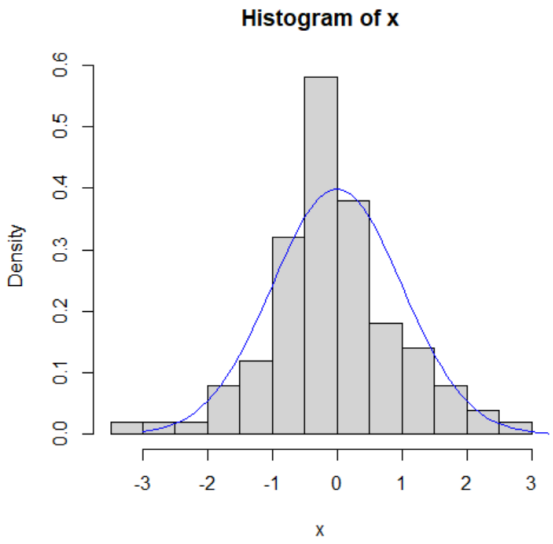


Figura 3.1: Ejemplo N(0,1)

Primero comenzamos con ejemplos sencillos para así poder entender el método. Volvemos al ejemplo implementado para ver los errores que cometía el metodo de Monte Carlo simple.

Ejemplo 3.1.1 *Cogemos el ejemplo básico en la que los datos siguen una distribución Normal: N(0,1). Es un ejemplo en el que sabemos la solución. Por un lado obtenemos el valor mediante la integral conocida y por otro lado obtenemos la estimación mediante el método de muestreo escogido. Tomamos μ como el parámetro desconocido. Luego la hipótesis nula H_0 es que los datos siguen una distribución Normal N(0,1) y la alternativa $H_1 : N(0, \mu)$.*

$$\begin{cases} H_0 : & N(0,1) \\ H_1 : & N(\mu,1) \end{cases}$$

En la imagen 3.1 podemos observar como para c=3 fijo, utilizando el método de Monte Carlo simple, lanzando 100 datos cuanto es el p-valor de nuestro experimento. Sabemos cuánto vale el p-valor para valores mayores que 3 en una distribución normal utilizando la integral, por lo que resulta fácil comparar con la estimación del p-valor obtenido.

En este caso lo que observamos es que para una muestra tan pequeña (n=100), las diferencias son significativas, por lo que no sería tan bueno. Por lo que el método de Monte Carlo directo para un número tan reducido de n no es preciso, y es por eso por lo que introducimos el método de muestreo por importancia. En este caso ninguno de los puntos cumple $x_i > 3$ por lo que necesitaríamos una muestra del

```
> x=rnorm(100,0,1)
> mean(x>3)
[1] 0
> pnorm(3,0,1, lower.tail=FALSE)
[1] 0.001349898
```

orden de mínimo 10^3 para poder tener algún dato, aunque la estimación sería muy pobre. Necesitaríamos número de muestra de alrededor de 10^4 para poder obtener una buena estimación.

Antes de pasar al muestreo por importancia es interesante ver la dispersión de los datos generados con el método de Montecarlo simple, para así poder luego compararlo con el muestreo por importancia.

Haciendo muestreos aleatorios de n=100, conseguimos valores muy dispersos. Sabiendo que el p-valor de una normal N(0,1) para c=3 es **0.001349898** , repetimos este experimento 100 veces para ver si es casualidad o se da que sigue siempre el mismo patrón. Obtenemos resultados dispersos alrededor de 0. La media de los experimentos comete un error alrededor de 10^{-3}.

Vamos a comenzar a introducir el muestreo por importancia,

```
# MUESTREO POR IMPORTANCIA (CON PESOS)

h=c()
for (k in 1:100){
      y=rnorm(100,3,1)
      z=y[y>3]
      w=dnorm(z,0,1)/dnorm(z,3,1)
      #print(sum(w)/100)
      h[k]=sum(w)/100
}
```

En la Imagen anterior podemos ver la implementación del método de muestreo por importancia básico. Nos encontramos en el caso anterior en el que el c=3 fijo y queremos calcular el p-valor para los $x_i > 3$. Por lo tanto siguiendo nuestra conjetura, para obtener el estimador óptimo y de menor varianza, tenemos que muestrear a partir de una Normal: N(3,1). Seguimos los siguientes pasos:

1. Se muestrean 100 datos de una Normal: N(3,1).
2. Nos quedamos con aquellos > 3.
3. Calculamos los pesos $w_i = \frac{\rho(z,0,1)}{\rho(z,3,1)}$
4. Sumamos los pesos y calculamos la media.
5. Repetiremos el mismo proceso 100 veces.

Mediante el anterior código, con tan solo una muestra de 100 datos obtenemos una estimación del p-valor de 0,001368093 $\approx$ 0,001349898. Por lo que con el mismo número de datos que utilizando el método de Montecarlo, obtenemos una estimación bastante más precisa.

Vamos a ver a continuación una comparativa de los datos obtenidos de los mustreos simple y por importancia. Lo vamos a ver mediante el diagrama de cajas, en donde se ve claramente mayor concentración de los datos cuando utilizamos el **Importance Sampling**.

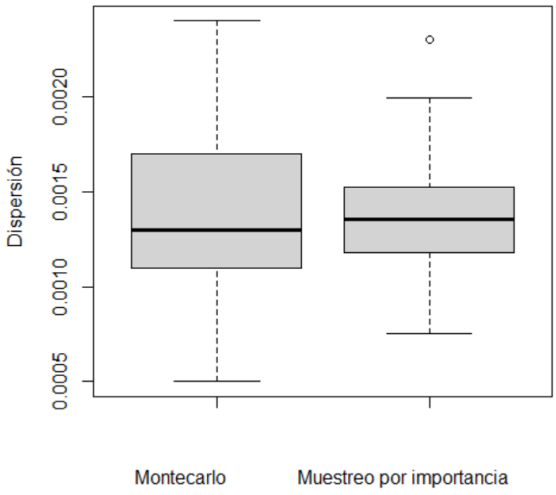


Figura 3.2: Comparación de la dispersión

Los diagramas de cajas y bigotes, son una representación gráfica que permite resumir las características principales de los datos (posición, dispersión, asimetría, ...) e identificar la presencia de valores atípicos. La caja de un boxplot comienza en el primer cuartil (25 %) y termina en el tercero (75 %). Por lo tanto, la caja representa el 50 % de los datos centrales, con una línea dentro que representa la mediana.

En el caso particular de la figura 3.2 estamos comparando el muestreo simple (izquierda) con el muestreo por importancia(derecha), en el cual podemos observar claramente como los datos se encuentran mucho más dispersos en el muestreo simple, es decir, al crear más datos en la zona en la que nosotros estamos interesados, generamos una mayor concentración de los datos. Obteniendo así una mejor estimación del p-valor.

Para un estudio más detallado, vamos a realizar un barrido para distintas funciones de muestreo con importancia, para distintos valores de μ, para ver hasta que punto es correcta en este caso nuestra conjetura. Luego estamos buscando para que valor de μ se obtienen los mejores resultados, es decir, aquel en el que el error cometido sea mínimo. Tratamos de demostrar que nuestra conjetura es cierta, o que al menos es válida. Nuestra suposición es que el resultado óptimo se dará en $\mu = c$ justamente. Es lo que queremos probar mediante simulaciones.

Continuamos en el caso N(0,1). Hacemos varios experimentos para así ver la variabilidad de ese mínimo. Vamos a probar el resultado para c = 1,2,3 y 4, para poder dar así veracidad a los resultados.

La simulación consiste en;

1. Fijar un c cualquiera.
2. Dar una lista de posibles valores de μ.
3. se fija un valor de μ y se repite el proceso arriba descrito:
   a) Se muestrean 100 datos de una Normal: N(μ,1).
   b) Nos quedamos con aquellos > 3.
   c) Calculamos los pesos $w_i = \frac{\rho(z,0,1)}{\rho(z,\mu,1)}$
   d) Sumamos los pesos y calculamos la media.
   e) Repetiremos el mismo proceso 100 veces.
   f) Una vez obtenidos los datos calculamos la media de los 100 experimentos y ya tenemos nuestra estimación para cada valor de μ.
   g) Calculamos el sesgo, el error cometido y la desviación típica de los experimentos.

El **sesgo** lo calculamos como el la diferencia entre el valor ideal y la estimación obtenida, lo cual nos da una imagen representativa de la dispersión de las estimaciones. Por su parte, el cálculo del **error cometido**, se calcula igual que el sesgo pero en valor absoluto, para hacer énfasis en cuanto de diferente es. Por último calculamos la **desviación típica** de las estimaciones utilizando la función sd().

Caso básico de una Normal N(0,1) con c=1 fijo.

Calculamos el p-valor de una Normal: N(0,1) para valores mayores que c=1. Lo podemos calcular mediante la integral o mediante la función de R pnorm(1,0,1, lower.tail=F) = 0.1586553.

El Error cometido lo hemos calculado como la diferencia entre él valor real y el valor de la muestra en valor absoluto, como podemos ver en el gráfico 3.3. Por otro lado compararemos la diferencia sin el valor absoluto, en la cual se podrá observar más claramente como se dispersan los puntos a lo largo de toda la muestra, podemos verlo en el gráfico 3.4.

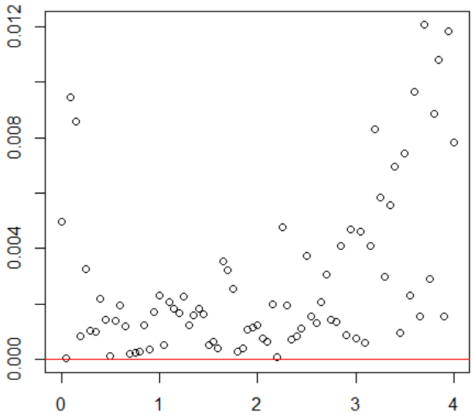


Figura 3.3: Desviación del valor ideal por muestreo por importancia a partir de N(μ,1)

En el gráfico 3.3 está representado el error cometido, donde se puede ver que hay una zona de valores de μ para los cuales el error es mínimo. Es cierto que en este caso en concreto de c=1, los resultados están muy dispersos, por lo que no podemos sacar muchas conclusiones, pero si podemos observar cómo a medida que nos alejamos de $\mu' = 1$ los valores están bastante más dispersos y cometen un error mayor.

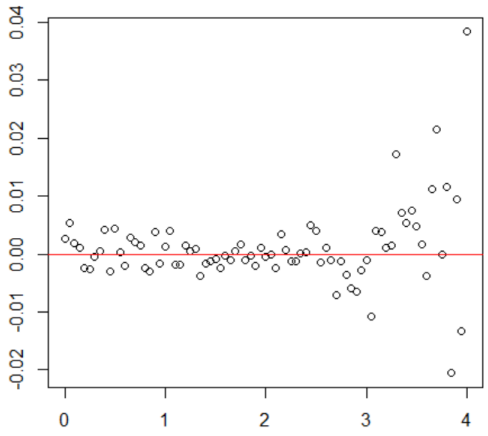
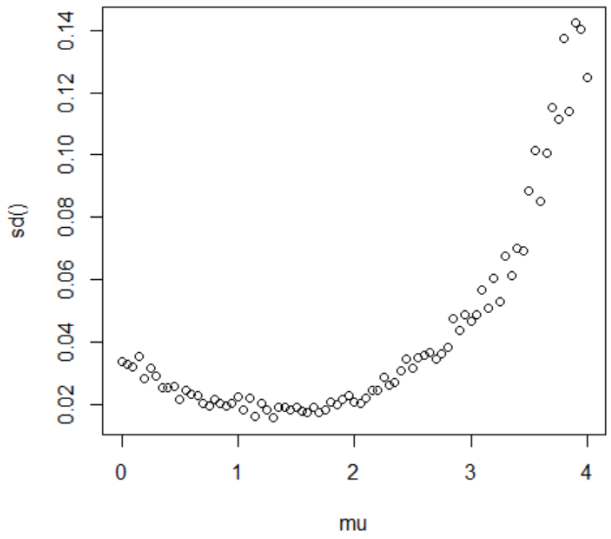


Figura 3.4: Diferencia entre el valor ideal y las estimaciones obtenidas a partir de N(μ,1)

En la figura 3.4 se ve que globalmente no está sesgado, aunque se intuye que para valores de μ en torno a 4 hay mucha dispersión. Al igual que en la grafica 3.3, observamos como a medida que nos alejamos de los valores de μ próximos a c=1, los errores son cada vez más grandes, por lo que la conjetura de elegir el μ_{best} como el observado en datos nos lleva a pensar que efectivamente es el que andamos buscando.
Podríamos concluir que no hay sesgo, pero sí que aumenta la dispersión. Por ello lo miramos en la siguiente prueba.

Hemos visto antes como los datos estan más concentrados en el importance sampling que en el método de Monte Carlo directo, y estamos buscando el estimador que minimice la varianza. Veámoslo gráficamente. Para ver mejor la dispersión de los datos a medida que nos movemos por los valores de μ' utilizaremos la función sd() de R. Obtenemos así el siguiente resultado,



Vemos como a medida que nos alejamos de $\mu' = 1$ los valores están cada vez más dispersos. Es interesante como podemos restringir el intervalo óptimo de los posibles valores de μ', y es que en torno a $\mu' = 1$ es cuando los datos están más próximos y por

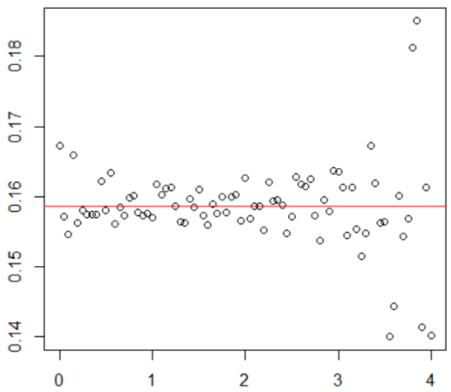
tanto la estimación es mejor.
Es cierto que vemos que el mínimo está siempre un poco a la derecha de nuestro punto c. Por ello vamos a tratar de darle veracidad a lo visto en c=1, graficando para diferentes valores de c, pudiendo así comparar y verificar que efectivamente el mejor μ' está en torno a c, o que al menos coger $\mu' = c$ es óptimo.

* Vemos que para este caso muestrear bajo $\mu' = c$, no es el valor del parámetro para el cual obtenemos la menor varianza como habíamos supuesto, pero a pesar de no ser el valor óptimo, podemos probar que está muy cerca en cuanto a desviación típica del óptimo. Es por ello que creemos que seguir muestreando con $\mu' = c$ es realmente bueno y obtenemos una gran estimación.

En los gráficos anteriores hemos visto como hay un intervalo de valores de μ' para los cuales el resultado parece óptimo, sin poder elegir un valor en concreto. Por ello vamos a intentar acotar el rango de valores de μ'.

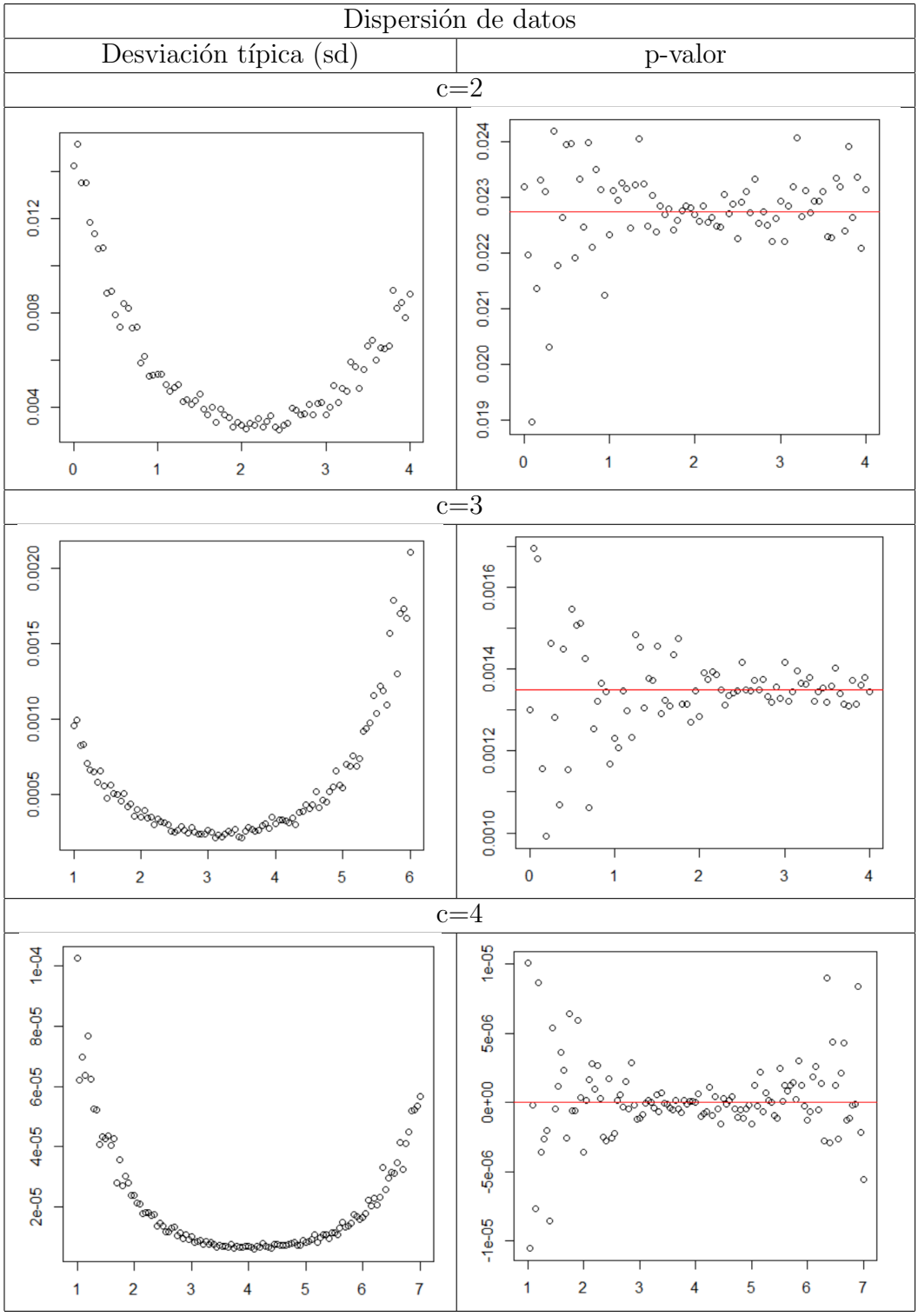
Como sabemos, mediante la función de R **pnorm** podemos calcular el valor del p-valor fijado un punto c cualquiera de una distribución Normal. Veamos para que valores de μ' se da la mejor estimación. Lo podemos ver mediante la siguiente gráfica en la cual hemos calculado el p-valor en función de diferentes μ' mientras que en rojo hemos trazado el valor exacto del pnorm para cada c fijado.

Caso c=1:



Podemos ver como a medida que muestreamos μ con valores más cercanos a c la estimación es mucho mejor, y a medida que nos alejamos de esta, el p-valor es bastante menos preciso, hasta llegar a $\mu' = 4$ que nos alejamos bastante de lo que estamos buscando.

Esto mismo se ve aún más claro para valores más grandes de c. Tenemos aquí algún ejemplo que nos ayuda una vez más a verificar que el μ' optimo que estamos buscando es justamente el valor de c, o un valor muy próximo a c. Mediante la siguiente tabla intentamos graficar tanto la desviación típica como el error cometido para c=2,3,4, para comprobar efectivamente que tomar como $\mu_{best} = c$ es al menos próximo al óptimo.



Cuadro 3.1: Resultados de muestreo por importancia para distintos valores de μ'. Las gráficas de la izquierda representan, dado un c fijo, la desviación típica de las estimaciones para distintos valores de μ'. Las gráficas de la derecha representan la comparación del valor ideal(en rojo) con la estimación obtenida por muestreo por importancia para distintos valores de μ'. Todo ello para distintos valores de c=2,3 y 4.

Interpretación:

- **Caso c=2:** Observamos una dispersión más pronunciada en el intervalo [0,1] y también al alejarnos de c, en [3,4]. Lo cual implica que nuestro estimador óptimo se encontrará en el intervalo complementario. Estamos intentando demostrar que el estimador de varianza mínima que mejor se ajusta a nuestros datos es justamente lo que vemos en los datos, lo cual se verifica en el gráfico de la desviación típica donde vemos que los datos están más concentrados cuando $\mu \approx c$.

  Calculamos las aproximaciones del p-valor y las comparamos con el valor ideal, con el valor obtenido en la observación. Aquí podemos ver con bastante claridad como las estimaciones son cada vez mejores cuando μ' se acerca a c y cuando nos alejamos de este cometemos un error mayor.

- **Caso c=3:** Al igual que en el caso anterior, vemos que el intervalo óptimo esta en torno a c. Una vez visto que el mínimo no está exactamente en c, buscamos entonces probar que muestrear con $\mu = c$ realmente funciona y que la estimación es buena. Observando la imagen, le quitamos importancia a a que el mínimo no este exactamente en c, ya que podemos ver que las diferencias son escasas. En el gráfico de la derecha podemos ver que las estimaciones del p-valor son muy buenas cuando nos acercamos a 3 tanto por la izquierda como por la derecha.
- **Caso c=4:** Podemos observar en los dos gráficos como tanto la desviación típica, como en el de la comparación de la estimación de los p-valores, que los errores cometidos son mayores a medida que nos alejamos de c, es decir los datos son más precisos para µ' cada vez más cercano a c. Lo que nos da una idea de que nuestro µbest se encontrara entorno a ese mismo valor. Al igual que en todos los casos anteriores verifica nuestra suposición inicial.

  * En este caso podemos ver claramente los beneficios de nuestra elección, y es que en la primera gráfica se ve cómo podemos obtener los mismos resultados muestreando con una décima parte de datos de $\mu' = 4$ de los que obtendriamos muestreando para $\mu' = 1$.

3.1.1. Observaciones

De esta prueba podemos obtener las siguientes conclusiones

- La conclusión más interesante que sacamos de los ejemplos anteriores es que con un muestreo de pocos elementos, somos capaces de calcular p-valores muy pequeños, todo lo pequeños que queramos.
- Hemos visto la gran diferencia entre el método de Montecarlo simple y el método de muestreo por importancia. En el primero observamos como prácticamente la mayoría de los datos que muestreamos son inservibles, mientras que al utilizar el muestreo por importancia muestreamos tan solo en los lugares que queremos, donde estamos interesados, obteniendo así una dispersión de los datos mucho menor y una mejor estimación del p-valor.
- Por otro lado, nuestra suposición era que el μ' que mejor se ajusta a los datos, aquel que nos producía una menor dispersión era c o en torno a c. Hemos visto que el mínimo no se da exactamente en ese punto, de hecho se puede demostrar en algunos casos, que efectivamente no está en c. Pero sí hemos visto que se encuentra entorno a c, y que las diferencias entre estos son muy pequeñas.
- Hemos hecho pruebas para distribuciones normales con distintos valores de σ. Los datos obtenidos mediante las simulaciones se volvían cada vez más dispersos y las estimaciones menos precisas a medida que el valor de σ aumentaba, por lo que obteníamos resultados menos concluyentes. Hemos visto que la situación en la que mejor se comportan los datos es para $\sigma = 1$. Obtenemos las mejores estimaciones de los datos, cometiendo menor error.
- En cambio, si vamos cambiando los valores de μ de nuestra hipótesis nula, observamos como los datos se comportarán de igual manera y las conclusiones serán las mismas.
- El método es muy efectivo en estos casos tan sencillos. Veamos ahora que pasa para casos más reales.

4. Aplicación a casos reales

Después de haber visto cómo se comporta el método para un ejemplo sencillo, vamos a comenzar a aplicar el método de muestreo por importancia a un caso más práctico, un modelo un poco más próximo a la realidad manteniendo la simplicidad para ver si los resultados son igual de interesantes.

4.1. Presentación del problema

En esta sección vamos a tratar de presentar el problema al que queremos dar solución. Suponemos que unos datos siguen ciertas distribuciones de probabilidad ρ_0 conocida. Pero tenemos una muestra a la cual llamaremos observación, que parece que no se ajusta a esa distribución, si no que parece que es la superposición de dos funciones de distribución, bien de la misma familia o de diferentes,$(1-\alpha)\rho_0 + \alpha\rho_1$, en la que el parámetro α es desconocido. Es un caso muy habitual en ciencia, en la que se busca una señal desconocida, sobre un fondo conocido.

Las funciones ρ_0 y ρ_1 pueden depender de varios parámetros, pero en nuestro caso los vamos a fijar, dejando un único parámetro libre, α.

Tratamos de ver si esa muestra, esa observación, ha sido una casualidad fruto del muestreo o si hay evidencias de que la hipótesis nula es falsa y por tanto no siguen la función de distribución ρ_0 que en principio tendría que seguir.

En este caso la aplicación del método de muestreo por importancia para estimar el p-valor si va a ser válida ya que nos encontramos en el caso en el que α solo puede tomar valores positivos, y más aún, solo puede tomar valores en el intervalo [0,1].

¿Por que? La razón a lo mencionado anteriormente es que nos encontramos ante una función de distribución cuya integral en todo el espacio tiene que valer 1, y por tanto $\alpha \in [0, 1]$ necesariamente.

4.2. Planteamiento caso general

Tendremos un modelo para la pdf(probability distribution function), que depende de algunos parámetros y otro modelo alternativo que además de estos parámetros

contiene uno más, que caracteriza cuánto de "nueva física" hay. Nos quedamos en el caso en el que H_0 no tiene parámetros libres y H_1 tan solo 1, por lo que si pudieramos aplicar Wilks, nos encontraríamos ante una distribución χ^2 con 1 grado de libertad.

Utilizaremos como test estadístico un likelihood ratio, que es el cociente de las verosimilitudes que mejor ajustan a nuestros datos.

$$\begin{cases} H_0: & \rho_0(x|\mu_0) \\ H_1: & (1-\alpha)\rho_0(x|\mu_0) + \alpha\rho_1(x|\mu_1) \end{cases}$$

Donde ρ_0, ρ_1, μ_0 y μ_1 son conocidos. Si llamamos ρ' a la función de distribución de la hipótesis alternativa, se define el likelihood ratio para nuestro caso unbinned, como,

$$q(\vec{x}) = -2log(\tfrac{\rho(\vec{x}|\alpha=0)}{\rho'(\vec{x}|\alpha)})$$

Lo que buscamos es el p-value, o sea la probabilidad de que dada la hipótesis H_0, tengamos un valor de q igual o mayor que el que vemos en datos, con la idea de poder rechazar esta hipótesis (y por tanto poder decir que ha habido un descubrimiento, porque el modelo "conocido" no representa a nuestros datos).

Recordemos, que para poder hablar de descubrimiento, tenemos que tener un p-value equivalente al menos a 5-σ.

Como nos encontramos en el caso en el que el parámetro libre α solo puede tomar valores en [0,1], el método nos dice que independientemente de la función con la que muestreemos nos dará una buena solución. Nos encontramos en unos de esos casos en los que no se puede aplicar Wilks ya que rompe una de las condiciones, el valor óptimo del parámetro libre α se encontrará muchas veces en el extremo, $\alpha = 0$.

Nuestra idea es muestrear justamente con la función que mejor se ajuste a la observación y ver que efectivamente obtenemos una solución óptima.

4.3. Planteamiento caso específico

Una vez visto el planteamiento del caso general, vamos a intentar centrarnos en un caso específico para así poder aplicar el método y obtener resultados relevantes.

Como bien hemos dicho antes nuestra Hipótesis nula va a ser que la muestra sigue una distribución conocida sin parámetros libres, en este caso $\rho_0 = N(0,1)$ y la Hipótesis alternativa es que la muestra sigue la función de distribución de una superposición de dos Normales $(1-\alpha)N(0,1)+\alpha\cdot N(2,1)$ en la que tenemos un parámetro desconocido α.

$$\begin{cases} H_0: & N(0,1) \\ H_1: & (1-\alpha)N(0,1) + \alpha \cdot N(2,1) \end{cases}$$

4.3.1. Procedimiento

- Pasamos a un "experimento", cada repetición será un muestreo de M elementos, que repetiremos N veces para calcular el p-valor.
- Cada uno de estos M números sigue una normal,

$$\begin{cases} H_0: & N(0,1) \\ H_1: & (1-\alpha)N(\mu_1,1) + \alpha N(\mu_2,1) \end{cases}$$

- Un experimento consistiría en sacar M números según una de estas leyes, obtener el α_{best} y calcular el q.
  1. Para muestreo normal usaríamos H_0, N(0,1), para muestreo por importancia probaremos con $(1-\alpha)N(0,1) + \alpha N(2,1)$ para varios α.
  2. Obtenemos el α_{best} utilizando la función de R llamada bmle2. [5]. Ponemos límites a α, obligandola a estar entre [0,1], para que cumpla ser una pdf.
  3. Calculamos q con la fórmula de $-2 \cdot log(\frac{\rho(\vec{x}|\alpha=0)}{\rho(\vec{x}|\alpha=\alpha_{best})})$.

  Procedimiento de muestreo

  Pasos a seguir:

1. Calculamos el likelihood ratio para nuestro experimento, q_0.
2. Realizamos N pseudoexperimentos según otra pdf.
3. Volvemos a realizar el ajuste, α_{best}, para cada pseudoexperimento.
4. Calculamos w_j definido como;
$$w_j = \frac{\rho_0(\vec{x}_j,\alpha=0)}{\rho'(\vec{x}_j,\alpha)}$$
5. Calculamos q_j como
$$q_j = -2 \cdot log(\frac{\rho(\vec{x}_j|\alpha=0)}{\rho(\vec{x}_j|\alpha=\alpha_{best})})$$
6. Calculamos nuestro estimador, como;
$$\frac{1}{N}\sum_{q_j>q_0} w_j$$
en este caso es el p-valor, que es simplemente "contar los pesos" de aquellos con $q_j > q_0$ entre el número de ejemplos.

4.3.2. Implementación

Paso 1

Creamos el primer experimento, lo que será nuestro observación. En el cual iremos creando la muestra de manera aleatoria en función de N(0,1) y de N(2,1), para un α fijado.
Tomamos M=N=100, y creamos nuestro experimento para $\alpha = 0,1$ fijo.
*M=número de datos, N= número de iteraciones.

Paso 2

Buscamos α_{best} que mejor se ajusta a nuestra observación. Para ello utilizamos la función mle2, del paquete de R ("bbmle") [5].
Lo que hace la función **mle2** es, una vez escrita la función de verosimilitud (likelihood) y dados los datos de la muestra, te calcula el α que mejor se ajusta a estos, acotando los posibles valores de α al intervalo cerrado [0,1].

Paso 3

Una vez obtenido el α_{best}, calculamos el qobs como,

$$q_{obs} = -2(\sum[log(\rho(\vec{x}|\alpha = \alpha_{best}) - \rho(\vec{x}|\alpha = 0))])$$

Por lo tanto tenemos todos los datos referentes a nuestra observación. Tenemos que ver si lo que hemos visto en la muestra se da siempre o ha sido una fluctuación. ¿Como probamos eso? utilizamos el método de muestreo por importancia.

Paso 4

Comenzamos con el muestreo por importancia:

Vamos a tomar una muestra de N= 100 iteraciones. Queremos poder rechazar la hipótesis nula y poder concluir que efectivamente ha habido un descubrimiento. Para ello, estamos buscando el α que mejor se ajusta a los datos.

Tomamos como posibles valores de α el intervalo cerrado [0,1]. Para cada alpha vamos a hacer un muestreo de M=100 datos.

Comenzamos muestreando para los distintos valores de α, y mediante la función mle2, al igual que cuando calculamos el de la observación, buscamos el α_{best}, el α que mejor se ajusta a estos datos. Calculamos la función de verosimilitud para cada α y calculamos los pesos asociados de la siguiente manera:

$$w_i = \prod \frac{\rho(\vec{x}\ \alpha=0)}{\rho(\vec{x}|\alpha)}$$

Paso 5

Una vez calculados los pesos hacemos la media de todos ellos para los que $q > q_{obs}$ obteniendo así el p-valor para cada valor de α.

Paso 6

Por último para el cálculo de la desviación típica tenemos varias formas de calcularlo. En este caso vamos a utilizar la media de la raíz cuadrada de la suma de los pesos al cuadrado, es decir,

$$sqrt(sum(w[q > q_{obs}]))/N$$

para cada valor de α.

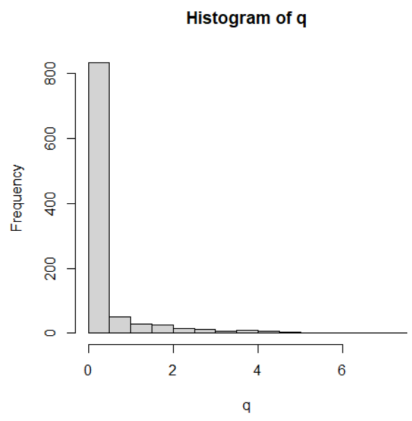
4.4. Aplicación

Nuestras Hipótesis son,

$$\begin{cases} H_0: & N(0,1) \\ H_1: & (1-\alpha)N(0,1) + \alpha \cdot N(2,1) \end{cases}$$

4.4.1. Resultados:

Comenzamos calculando el p-valor para el método de MonteCarlo directo, para luego poder así compararlo con el método de muestreo por importancia.



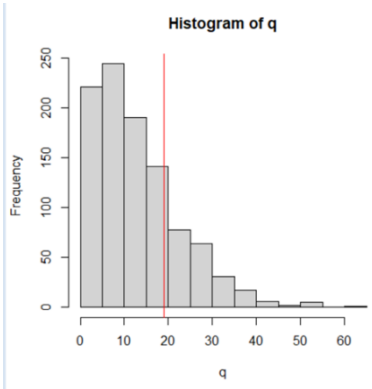
Partimos de una observación, estos siguen una distribución (0.9)N(0,1) + (0.1)N(2,1) con α = 0,1. Entonces ajustamos mediante mle2 y obtenemos el α_{best} con la cual obtenemos el q_{obs} = 19,43111.
Muestreamos N=1000 iteraciones, con M=100 datos cada uno, de una distribución N(0,1) nuestra hipótesis nula y calculamos el q, para poder así contar $q > q_{obs}$ y calcular el p-valor.

Pero como podemos ver en el histograma de la izquierda, aun muestreando un número considerable para M, vemos como no tenemos ningún

$q > q_{obs}$. Es por ello que en estos casos tendríamos que hacer M más grande para poder tener datos donde nosotros estamos interesados, y es justo aquí donde está el problema del método de Montecarlo, la gran carga computacional que supone el método en caso de querer sacar conclusiones relevantes. No para estos casos en concreto donde la implementación es sencilla si no en casos más reales en el que si supondría un problema. Es aquí donde cobra sentido el método de muestreo por importancia.

Método muestreo por importancia:

Haciendo el análisis anterior tmabién para N=1000, pero muestreando sobre $(1-\alpha_{best})$N(0,1) + (α_{best})N(2,1) obtenemos el siguiente histograma, en el cual podemos observar claramente los beneficios de la aplicación del Importance sampling, obteniendo así muchos datos en el lugar que estamos interesados.



Como bien hemos dicho antes, se puede observar claramente las diferencias respecto al gráfico anterior. Donde muestreamos justamente alrededor de $q_{obs} = 19{,}43111$. Mientras que utilizando el método de Montecarlo para M=1000 no obteníamos ningún dato.

Hemos analizado el mismo caso que en el muestreo directo de Montecarlo en el que muestreamos a partir del alpha visto en la observación $\alpha = 0{,}1$. Cogemos una cifra inferior de datos como N=100, y no deja de sorprendernos como cogiendo $\alpha = \alpha_{best}$ somos capaces de obtener un p-valor de $4{,}334061 \cdot 10^{-06}$.

Estamos calculando probabilidades de 10^{-7}, luego si utilizásemos el método de Montecarlo como método para estimar el p-valor, necesitaríamos datos del orden de 10^7. A esto nos referimos cuando hablamos de coste computacional, mientras que con el muestreo por importancia somos capaces de calcular con una muestra de tan solo 100 datos, utilizando Montecarlo necesitaríamos decenas de millones de datos para poder estimarlo.

En estos casos cuando hacemos el ajuste mediante el mle2, muchos de los α son 0, y por ello no podemos aplicar Wilks. En el caso en el que pudiéramos aplicar Wilks, diríamos que los datos siguen una distribución χ^2 con 1 grado de diferencia. Por lo tanto calculando usando la función pchisq(19.43111,df=1,lower.tail=F)= 1.042943e-05.

Vemos que queda del mismo orden de magnitud, aproximadamente el doble. La diferencia no sabemos si es por la no validez de Wilks o por nuestro método.

4.4.2. Estudio detallado del muestreo por importancia

Queremos probar que nuestra conjetura es cierta. Buscamos demostrar que lo mejor es muestrear entorno a α_{obs} o que por lo menos es válido y está muy cerca del valor óptimo, como veíamos en los casos básicos.

Vamos a seguir el mismo procedimiento que en el ejemplo de las Gaussianas. Comenzaremos calculando el p-valor para luego calcular la desviación típica.

En la gráfica 4.1 vemos como para distintos valores de α obtenemos distintos p-valores, la vamos a utilizar simplemente como comprobación de que las cosas funcionan bien y nos sirve para ver la relación entre la desviación del método y el p-valor para cada α. Es relevante la zona en la que los p-valores no son muy pequeños. Puede darse que los valores sean cero o muy cercanos a cero por dos principales motivos,

- No hay datos. Luego $q_{obs} > q_j$ para todo j.
- A pesar de haber datos mayores que q_{obs}, su peso es tan tan pequeño que es como sumar ceros.

Es por ello, por lo que tiene sentido mirar más en torno a 0.1, nuestro α_{obs}, por arriba y por abajo.

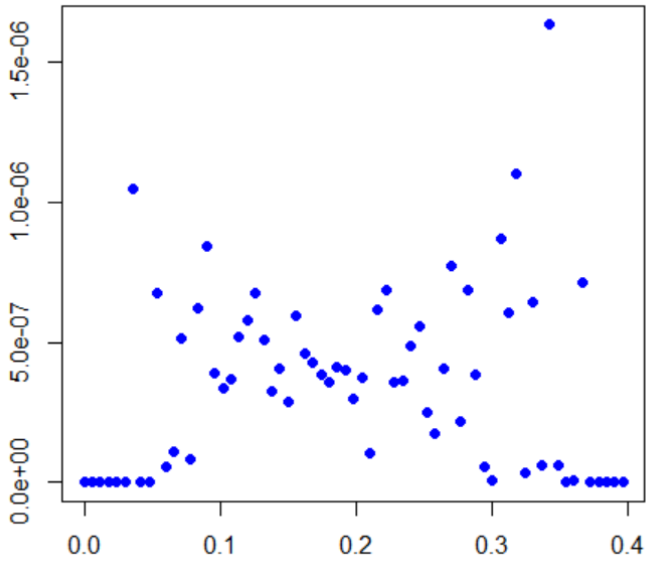
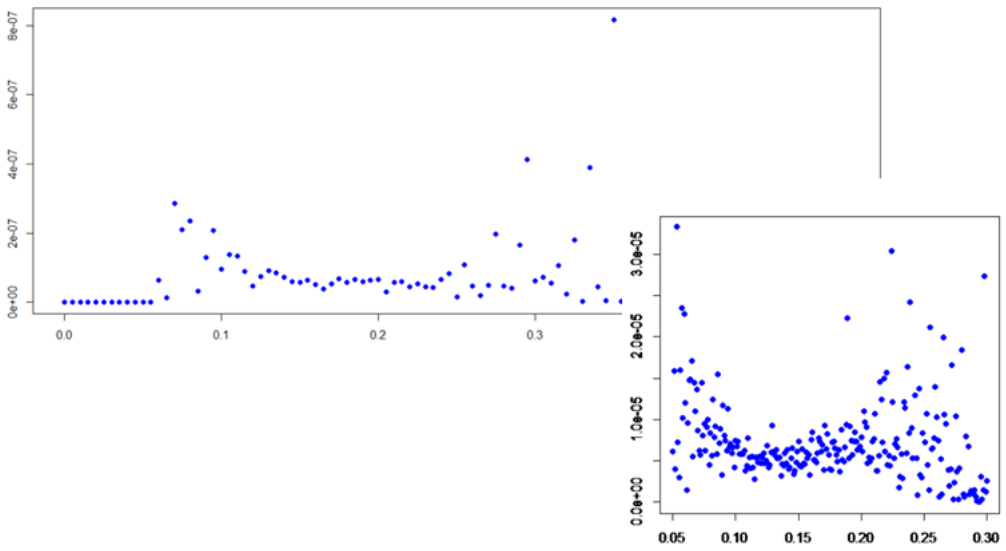


Figura 4.1: p-valores obtenidos mediante muestreo por importancia para distintos valores del parámetro α.

Una desviación estándar alta indica que los datos se extienden sobre un rango de valores más amplio. Lo esperado, ya que hemos visto en la gráfica anterior que para esos valores de α es para los cuales más puntos obtenemos de $q > q_{obs}$. En cambio esta gráfica es la que nos dice la potencia del método. Buscamos la que minimiza la dispersión, claro siempre que el p-valor tenga sentido (si da siempre cero la dispersión es cero).



En ambas gráficas podemos ver la misma representación de la desviación típica, en la primera vemos de manera más general, sin acotar el rango de α mientras que en la segunda hemos acotado los valores de α a aquellos distintos de cero y hemos muestreado más valores.

Ahora sale muy clara la dispersión y vemos que se comporta justo como esperábamos que lo hiciera. El mínimo sale un poco más allá de 0.1, pero está cerca. Es verdad que si muestreamos con $\alpha = \alpha_{best}$ no conseguimos el resultado más óptimo pero se puede probar que está muy cerca en cuanto a desviación del óptimo.

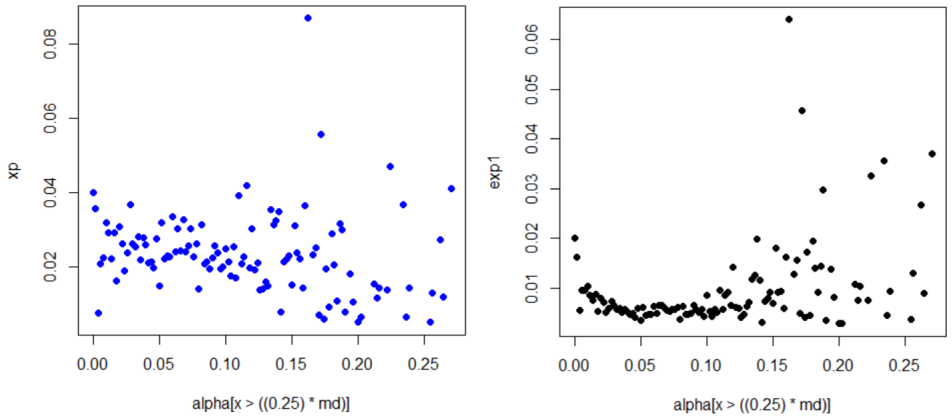
En cuanto a la estimación de la desviación típica, vemos como los datos se concentran más cuando utilizamos valores de α cercanos(a la derecha) a 0.1, nuestro α_{best}. Como habíamos visto en la gaussiana, al inicio de las pruebas.

Hemos visto ya como el parámetro de varianza mínima no se encuentra exactamente en α_{best}, pero sabemos que en un intervalo donde este está contenido, obtenemos valores muy cerca en cuanto a desviación del óptimo, por ello, tratamos de quedarnos solamente con los valores de α que nos interesan. Mediante una líneas de código en R hemos implementado una función mediante la cuál encontramos el intervalo óptimo. Anexo: archivo= intervalooptimo.R

Dado que hay una zona amplia de valores con dispersión baja y que estos a su vez muestran cierta dispersión, planteamos un criterio para definir un intervalo de valores de la α de muestreo que proporcionarían un resultado parecido y próximo al óptimo.

- Llamamos x a la los p-valores obtenidos y ex a las desviaciones típicas obtenidas en la simulación.
- Una vez definidas las variables, comenzamos quitando todos los valores iguales a cero, ya que no nos interesan.
- A continuación calculamos la mediana de todos estos valores, elegimos la mediana ya que buscamos el estadistico que menor variación tiene.
- Por último nos vamos a quedar con aquellos por encima de 0,2 por el valor de la mediana. Hemos elegido, 0.2 ya que si cogiésemos un valor más pequeño se cuelan valores en los que no estamos interesados.

Graficando las anteriores líneas de código obtenemos los siguientes resultados para $\alpha = 0{,}1$;



Donde obtenemos el intervalo [0.000 0.288], rango que incluye a α_{best}, de acuerdo con nuestra conjetura.

Vayamos ahora a valores de α más exagerados, podremos ver asi más claramente lo que sucede y aparecerán zonas de ceros al principio y al final de la gráfica, que nos ayudarán a identificar más claramente el intervalo óptimo mediante las líneas de código anteriores.

Lo resumimos en la tabla 4.1, en donde podemos ver como para todo α_0, encontramos un intervalo óptimo en el que este está contenido. Por lo que en la linea argumental que estábamos siguiendo, podemos considerar nuestra suposición como válida.

Tabla de intervalos óptimos en función de α_0	
α_0	Intervalo óptimo
0.2	[0.001 , 0.322]
0.4	[0.308 , 0.649]
0.6	[0.417 , 0.798]

Cuadro 4.1: Intervalos de valores de la α de muestreo que proporcionarían un resultado parecido y próximo al óptimo, para distintos valores de α_0.

4.5. Resumen/Observaciones

Una vez implementado el muestreo por importancia a casos reales;

- Hemos descrito e implementado el procedimiento para aplicar el muestreo por importancia a casos reales.
- Hemos visto que en casos reales como el de la doble Gaussiana, donde para una muestra de 100 valores somos capaces de estimar el p-valor. Mientras que si utilizásemos el método de Montecarlo directo necesitaríamos alrededor de decenas de millones de experimentos para poder hacer una buena estimación.
- Hemos visto también que para muestras de 100 experimentos, el método de muestreo por importancia cálcula con precisión p-valores del orden de 10^{-7}. En realidad, podemos obtener p-valores extremadamente bajos, tan bajos como queramos.
- Somos capaces de mediante un algoritmo de búsqueda acotar el intervalo óptimo de α. Y hemos visto como para cualquier valor de α_0 este siempre se encuentra en el intervalo óptimo.
- Al igual que en el ejemplo sencillo de la Gaussiana hemos visto que nuestra conjetura de que el estimador de varianza mínima no es exactamente $\alpha = \alpha_{obs}$ pero sí que está muy cerca.

5. Conclusiones

En este trabajo hemos estudiado la aplicación del muestreo por importancia a la estimación de p-valores muy bajos para contraste de hipótesis, proponiendo una solución que minimiza el coste computacional. Se ha intentado abordar tanto desde el punto de vista teórico, donde se dan argumentos que soportan la propuesta como sobre ejemplos programados en R.

Después de hacer un análisis básico pero exahustivo del método, somos capaces de concluir que;

- Aún siendo el método de Montecarlo uno de los métodos más utilizados en investigación, hemos probado como para casos extremos en los que tenemos una muestra reducida de datos el método no es eficiente. Mientras que el muestreo por importancia sí lo es.
- En relación con lo anterior, quizas la conclusión más importante que hemos sacado es que con un número reducido de datos podemos obtener p-valores extremadamente bajos, en realidad, tan bajos como queramos.
- Uno de los grandes motivos de la implementación del método era para aquellos casos en los que no sabemos si es aplicable el teorema de Wilks. Se ha visto cómo podemos mejorar los resultados del muestreo directo, reduciendo órdenes de magnitud del tamaño de la muestra.
- Hemos visto para que casos la elección de la función de probabilidad de muestreo es óptima, siempre que el likelihood ratio dependa monotonamente del parámetro.
- Una vez hemos restringido el método a problemas en los que hay un único parametro que solo puede tomar valores positivos. Tomando $\alpha = \alpha_{obs}$ somos capaces de calcular p-valores tan pequeños como queramos.
- A pesar de al principio creer que lo óptimo, o lo mejor era muestrear con $\alpha = \alpha_{obs}$, hemos visto que exactamente no es el parámetro que produce menor varianza, por lo que hemos tratado de demostrar, reduciendo a intervalos de valores de la α de muestreo que proporcionarían un resultado parecido y próximo al óptimo como para cualquier valor de α_0, α_{best} siempre se encuentra dentro del intervalo.
- Podemos concluir que es un método prometedor aunque sería interesante investigar su extensión a los casos donde hemos comprobado que su aplicación no es válida.

6. Bibliografía

[1] Libro: Simulations and the Monte Carlo Method by: Revra y Rubistein Dirk.P.Kroese (SECOND EDITION), 2007.

[2] Libro: INFERENCIA ESTADÍSTICA/ 3º del Grado en Matemáticas. Gloria Pérez Sainz de Rozas. Curso 2020/21.

[3] Libro: Mathematical Statistics with Applications in R Book • Third Edition • 2020.

[4] Libro: SIMULATION AND THE MONTE CARLO METHOD. Third Edition. Reuven Y. RubinsteinTechnion and Dirk P. Kroese. University of Queensland.

[5] Ben Bolker and R Development Core Team (2021). bbmle: Tools for General Maximum Likelihood Estimation. R package version 1.0.24. https://CRAN.R-project.org/package=bbmle

[6] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[7] Asymptotic formulae for likelihood-based tests of new physics. Glen Cowan1, Kyle Cranmer2, Eilam Gross3, Ofer Vitells3. 2007. The European physical journal C.

[8] Stat 8112 Lecture Notes: The Wilks, Wald, and Rao Tests. Charles J. Geyer. September 26, 2020.

[9] Introducing Monte Carlo Methods with R; Christian Robert, George Casella.

[10] Gentle, J.E. (2003). Random number generation and Monte Carlo methods. Springer-Verlag.

[11] Apuntes Inferencia estadística, Universidad Carlos III Madrid, Inferencia Bayesiana, Montecarlo. http://halweb.uc3m.es/esp/Personal/personas/mwiper/docencia/Spanish/Inferencia%20Bayesiana/Practicals/montecarlo.html#:~:text=La%20muestra%20de%20importancia%20es,...%2CwN.

[12] https://www.noticiasdelcosmos.com/2021/05/que-significa-5-sigma.html#:~:text=La%20f%C3%ADsica%20de%20altas%20energ%C3%ADas,se%20le%20llama%205%20Sigmas.

[13] https://www.sdelsol.com/glosario/simulacion-de-montecarlo/#:~:text=La%20simulaci%C3%B3n%20de%20Montecarlo%20es,una%20serie%20de%20datos%20iniciales.

[14] https://en.wikipedia.org/wiki/Importance_sampling

[15] http://www.dpye.iimas.unam.mx/eduardo/MCB/node20.html

[16] https://hmong.es/wiki/Importance_sampling

[17] https://hmong.es/wiki/Wilks'_theorem