Universidad de Cantabria

Facultad
de
Ciencias

Aplicabilidad de técnicas de muestreo de importancia para contrastes de hipótesis para "descubrimientos" en física de partículas.

(Applicability of importance sampling techniques for hypothesis testing in "discoveries" in particle physics)

Trabajo de Fin de Grado
para acceder al

GRADO EN MATEMÁTICAS

Autora: Elena López Fuente.

Director: Francisco Matorras Weinig.

Junio-2024

Resumen

En el análisis estadístico, obtener p-valores bajos de manera precisa es crucial para la toma de decisiones basada en contrastes de hipótesis, aunque este cálculo puede ser complejo y costoso en términos computacionales. Este trabajo aborda este desafío implementando un método de muestreo por importancia, el cual, es una técnica de simulación que mejora la eficiencia al centrarse en las áreas más relevantes del espacio de probabilidad. Un aspecto clave de este método es la selección de la función auxiliar para pesar el muestreo. Proponemos su elección en base a la hipótesis alternativa (H_1) del contraste, en particular, aquella que maximiza la verosimilitud.

Para validar este método, se han utilizado ejemplos sencillos analíticamente y simulaciones en R. También se ha aplicado el método de entropía cruzada para encontrar los parámetros óptimos del muestreo por importancia. Además, se ha comprobado la eficacia del método en distribuciones normales, exponenciales y de Poisson, mostrando una mejora significativa en la precisión del cálculo del p-valor y una reducción notoria de la varianza en comparación con métodos tradicionales. También se ha explorado la aplicación del método a contrastes de señal sobre ruido utilizando histogramas, demostrando que el método es efectivo en este contexto.

En conclusión, el método no solo se aplica a situaciones estándar, sino que también se adapta y utiliza en una amplia gama de problemas estadísticos, proporcionando una herramienta sólida y versátil para el cálculo de p-valores bajos.

Palabras clave: contraste de hipótesis, muestreo por importancia, máxima verosimilitud, p-valor, entropía cruzada.

Abstract

In statistical analysis, obtaining low p-values accurately is crucial for decision-making based on hypothesis testing, although this calculation can be complex and computationally expensive. This paper addresses this challenge by implementing an importance sampling method which is a simulation technique that increases efficiency by focusing on the most relevant areas of the probability space. A key aspect of this method is the selection of the auxiliary function for sampling weighting. We propose its selection based on the alternative hypothesis (H_1) of the hypothesis test, in particular the one that maximizes likelihood.

To validate this method, we have employed simple analytical examples and simulations in R. The cross-entropy method has also been applied to find optimal parameters for importance sampling. Furthermore, the efficacy of the method has been tested on normal, exponential, and Poisson distributions, demonstrating a significant improvement in the accuracy of p-value calculation and a noticeable reduction in variance compared to traditional methods. Additionally, the method has been explored for its application to signal-to-noise contrasts using histograms, showing effectiveness in this context as well.

In conclusion, the method not only applies to standard situations but also adapts and is used in a wide range of statistical problems, providing a robust and versatile tool for calculating low p-values.

Keywords: hypothesis testing, importance sampling, maximum likelihood, p-value, cross-entropy.

Índice general

Introducción

Este trabajo busca dar respuesta a un problema estadístico que se plantea en la investigación experimental en física de partículas. Parte de la presentación de resultados implica un contraste de hipótesis y los descubrimientos requieren descartar la hipótesis nula con p-valores extremadamente bajos. Además, los métodos utilizados se basan en aproximaciones que no siempre son ciertas, en el documento exploramos la viabilidad de calcularlo en base a muestreo por importancia.

El muestreo por importancia nos ayuda a calcular el p-valor muestreando por una función de distribución con más datos extremos. El problema del método radica en que no proporciona una guía clara sobre qué función escoger.
El objetivo es buscar una distribución de probabilidad de la que muestrear, que mejore un muestreo directo. Nuestra propuesta es utilizar la misma familia de H_1 y, en particular, la que mejor se ajusta a los datos observados. Para apoyar nuestra conjetura, nos basamos en ejemplos sencillos y tratamos de generalizarlo.

A lo largo del trabajo se desarrolla una propuesta de elección de funciones de muestreo y se analiza su aplicación, tanto desde el punto de vista teórico como con simulaciones con R.

Buscaremos este óptimo con el método de entropía cruzada. También aplicaremos en las simulaciones el método de minimización de la varianza. A su vez, haremos hincapié en cuáles son las ventajas de aplicar el muestreo por importancia frente a no aplicarlo. Por otro lado, veremos que el método del muestreo por importancia tiene ciertas limitaciones, las estudiaremos y trataremos de buscar alguna propuesta alternativa para abordar estas excepciones.

En el documento estudiaremos primero distribuciones sencillas de una variable, de las que conocemos la solución, normal, exponencial, Poisson para pasar a continuación a un caso más realista en base a histogramas, más similar a los experimentos que motivan este estudio.

Como sabemos, enunciar un teorema no es algo sencillo por lo que nuestra intención es conocer las hipótesis necesarias bajo las cuales se verifica nuestra conjetura y considerar las posibles excepciones que se puedan generar.

En el trabajo mostramos que en ciertos escenarios, el método arroja resultados altamente eficientes y precisos, permitiendo el cálculo de p-valores incluso con muestreos reducidos. Sin embargo, comprobaremos que el método no es universalmente aplicable.

Capítulo 1

Conceptos básicos

Queremos realizar un contraste de hipótesis en física de partículas, donde se tienen distribuciones normalmente discretizadas en forma de histogramas, que se contrastan con distintos modelos. Nuestro interés se centra en los casos en los que se observa una discrepancia significativa entre la observación y el modelo probabilístico esperado. Mediante el contraste de hipótesis, buscamos determinar si dicho modelo puede ser descartado.
En este área, por convenio, se considera que un modelo puede ser desechado, y, por tanto, se puede anunciar la existencia de "*nueva física*", cuando el p-valor correspondiente a la hipótesis es menor al equivalente a "*5 sigmas*" de una distribución normal. Generalmente, este p-valor se calcula utilizando el estadístico de razón de verosimilitud (likelihood ratio), basado en aproximaciones que se describirán más adelante. Debido a la complejidad de las simulaciones involucradas y el valor tan bajo del p-valor se vuelve inviable la estimación mediante muestreo. En este trabajo, estudiamos una alternativa para obtener dicho p-valor utilizando el muestreo por importancia.

En el desarrollo del documento consideraremos con un abuso de la notación x como vector salvo que se indique lo contrario.

Se comienza haciendo un repaso de conceptos vistos en las siguientes asignaturas: Estadística Básica, Cálculo de Probabilidades e Inferencia Estadística. [3], [4] y [5]. En primer lugar, recordaremos las definiciones de algunos conceptos básicos que serán de gran utilidad para el desarrollo del documento.

Definición 1.1 Probabilidad
Dado el espacio medible (Ω, σ)*, se llama probabilidad a cualquier aplicación real* P*, definida en* σ*, tal que*

P.1 : $P(\Omega) = 1$.

P.2 : Para todo $A \in \sigma$*, se cumple que* $P[A] \geq 0$.

P.3 : Si $\{A_n\}_n$ *es una sucesión de conjuntos disjuntos dos a dos contenida en* σ*, entonces* $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$.

Definición 1.2 Variable aleatoria
Dados dos espacios medibles (Ω, σ) *y* (Ω^*, σ^*)*, se llama variable aleatoria a cualquier aplicación*

$$X: \quad \Omega \longrightarrow \Omega^*,$$

tal que $X^{-1} \in \sigma$ *para todo* $B \in \sigma^*$. *Se dice que* X *es una variable aleatoria real si* $\Omega^* = \mathbb{R}$ *y* $\sigma^* = \beta$.

Definición 1.3 Función de densidad
Se dice que ρ_P *es la función de densidad de la distribución* P *si se cumple que*

$$F_P(x) = \int_{-\infty}^{x} \rho_P \, dt.$$

Definición 1.4 Función de distribución
Sea P *una probabilidad definida en* β. *Se llama función de distribución de* P *a la aplicación* F_P *definida por*

$$\begin{array}{rccl} F_P(x): & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & F_P(x) = P((-\infty, x]). \end{array}$$

Definición 1.5 Esperanza
Si X *es una variable aleatoria real, tal que existe su función de densidad* ρ_X. *Podemos definir la esperanza matemática de* X *como la cantidad:*

$$E[X] = \int t\rho_X(t) \, dt,$$

siempre que esta expresión tenga sentido.

Definición 1.6 Varianza
Sea X *una variable aleatoria tal que* $E[|X|^2] < \infty$. *Se define su varianza como:*

$$Var(X) = \sigma^2 = E[(X - E[X])]^2 = E[X^2] - E^2[X].$$

Definición 1.7 Contraste de hipótesis
El contraste de hipótesis o prueba de hipótesis es un procedimiento estadístico utilizado para tomar decisiones sobre una población a partir de una muestra de datos. Este procedimiento evalúa la validez de una afirmación específica (la hipótesis nula, H_0*) contra una alternativa (*H_1*) basada en los datos disponibles. Los pasos fundamentales del contraste de hipótesis son:*

1. *Formulación de hipótesis: nula y alternativa.*
2. *Selección del nivel de significación,* α*, la probabilidad de rechazar la hipótesis nula cuando en realidad es verdadera.*
3. *Elección de la prueba del estadístico. Para ello, se considera la naturaleza de los datos y las hipótesis.*
4. *Cálculo del estadístico, valor numérico que se calcula a partir de los datos de la muestra y se utiliza para tomar decisiones sobre la hipótesis nula.*
5. *Determinación del p-valor.*
6. *Toma de decisión, comparar el p-valor con el nivel de significación.*

Definición 1.8 P-valor
El p-valor, $\mathfrak{p}$, es una medida de probabilidad que varía en un rango de 0 a 1 y cuantifica la evidencia en contra de una hipótesis nula, es decir, el p-valor representa la probabilidad de obtener resultados igualmente extremos o más extremos que los observados, asumiendo que la hipótesis nula es verdadera. Se compara el p-valor con un nivel de significancia, α, para tomar decisiones sobre rechazar o no la hipótesis nula en un análisis estadístico.

$$Si\ \alpha < \mathfrak{p} \Longrightarrow Aceptar\ H_0.$$
$$Si\ \alpha \geq \mathfrak{p} \Longrightarrow Rechazar\ H_0.$$

En el desarrollo del trabajo se considera la siguiente notación: $\mathfrak{p} = \int_{x_0}^{\infty} \rho_X(t)\, dt$, siendo ρ_X la función de densidad de la variable aleatoria X.

Definición 1.9 Método de la máxima verosimilitud
Supondremos que $\Theta \subset \mathbb{R}$ y que para cada $\theta \in \Theta$ la distribución de probabilidad $P_{X_1}^{\Theta}$ es discreta o bien admite una función de densidad. Dada la muestra aleatoria simple $x_1, ..., x_n$ que suponemos fija, la función de verosimilitud (likelihood) es la aplicación:

$$\begin{array}{rccc} \mathcal{L}: & \Theta & \longrightarrow & \mathbb{R} \\ & \theta & \longmapsto & \mathcal{L}(\theta) = \prod_{i=1}^{n} \rho_\theta(x_i), \end{array}$$

donde, con un abuso de notación, hemos representado con $\rho_\theta(x_i)$ a $P_\theta[X_i = x]$ si la distribución es discreta o, como es habitual, a la función de densidad cuando ésta existe.
Esta función determina la verosimilitud del parámetro a la vista de la muestra obtenida.
Nótese que la función de verosimilitud depende de la muestra. Por ello, sería más correcto denotarla como $\mathcal{L}_{(x_1,...,x_n)}(\theta)$, pero no se suele hacer.
El estimador máximo verosímil, E.M.V. (maximum likelihood estimator M.L.E) de θ es cualquier valor del parámetro en el que se alcance el máximo de la función de verosimilitud, si tal valor existe. Es decir, el E.M.V. $\hat{\theta}_n = \hat{\theta}_n(x_1, ..., x_n)$ es cualquier valor que satisfaga:

$$\mathcal{L}_{(x_1,...,x_n)}(\hat{\theta}_n) = \sup_{\theta \in \Theta} \mathcal{L}_{(x_1,...,x_n)}(\theta).$$

Siempre que la función de verosimilitud sea derivable, este máximo se busca anulando la derivada. Se denomina ecuación de verosimilitud a

$$\frac{\partial}{\partial \theta} \log \mathcal{L}(\theta) = 0.$$

En el documento se considera la notación μ_{MV} para indicar cuál es el mejor μ que se ajusta a los datos observados y se considera este valor como la hipótesis alternativa para los desarrollos.

Definición 1.10 Cociente de verosimilitud
Es el cociente entre la verosimilitud bajo la hipótesis nula y la verosimilitud bajo la hipótesis alternativa, $\frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu_{MV})}$.
En términos simples, si el cociente de verosimilitud es alto, significa que los datos observados son más probables bajo la hipótesis alternativa que considerando la nula, lo que sugiere que la hipótesis alternativa es más probable. Por el contrario, si el cociente de verosimilitud toma un valor pequeño, sugiere que los datos observados son más probables bajo la hipótesis nula. Denotamos por, $\wedge$, la razón de verosimilitud.

Definición 1.11 Sesgo de un estimador
Sean $\hat{\theta}$ un estimador y θ el valor real del parámetro que se está estimando. El sesgo de $\hat{\theta}$ en $\theta \in \Theta$ es la diferencia entre la esperanza matemática del estimador y el valor del parámetro, es decir, $Bias(\hat{\theta}) = E[\hat{\theta}] - \theta$.

- *Si $Bias(\hat{\theta}) = 0$ para todo $\theta \in \Theta$, es decir, $E[\hat{\theta}] = \theta$ para todo $\theta \in \Theta$, el estimador es insesgado.*
- *Si $Bias(\hat{\theta}) \neq 0$, el estimador es sesgado.*

Observación 1.12 *Condiciones necesarias para aplicar el Teorema de Wilks.*
Se debe verificar que:

- *El valor óptimo no esté en los extremos.*
- *Las hipótesis estén "anidadas", es decir, que H_0 esté contenida en H_1, de otra forma, que H_0 sea un caso particular de H_1.*

Teorema 1.13 Teorema de Wilks *(Extraído de [7] y [12]).*
Sea $\{X_n\}$ una muestra con función de distribución desconocida, si $n \to \infty$, entonces la distribución del estadístico de prueba $-2\log(\wedge)$ 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, donde Θ es el espacio de parámetros completo y Θ_0 es el subconjunto del espacio de parámetros asociado a H_0.

Corolario 1.14 *El Teorema de Wilks establece que, bajo ciertas condiciones, la distribución asintótica del estadístico del cociente de verosimilitud sigue una distribución χ^2.*

Observación 1.15 *El Teorema de Wilks nos permite calcular el p-valor y de esta forma realizar el contraste de hipótesis.*

Observación 1.16 *Si H_0 y H_1 corresponden a verosimilitudes con m y n parámetros libres respectivamente, el número de grados de libertad es exactamente $n-m$. En particular, si H_0 no depende de parámetros y H_1 depende de un parámetro, entonces los grados de libertad, df, son 1.*

Consideramos los siguientes resultados del artículo de: "Tests of Statistical hypotheses concerning several parameters when the number of observations is large " para poder enunciar el Teorema de Wald [11].

Proposición 1.17 *Existen $k-r$ funciones $\xi^{r+1}(\theta), ..., \xi^k(\theta)$ tales que satisfacen las siguientes condiciones:*

- *La transformación que lleva el punto θ en el punto ξ con las coordenadas $\xi^1(\theta), ..., \xi^k(\theta)$ es una transformación topológica de Ω en sí misma.*
- *Las derivadas parciales de primer y segundo orden de $\xi^1(\theta), ..., \xi^k(\theta)$ son uniformemente continuas y acotadas por θ.*
- *El límite inferior máximo del valor absoluto del jacobiano $\frac{\partial(\xi^1,...,\xi^k)}{\partial(\theta^1,...,\theta^k)}$ es positivo.*

Observación 1.18 *Sean $u_1, ..., u_r$, r variables independientes que siguen una distribución normal y varianza igual a 1. Se denota la función de distribución de U^2 por $F_r(\lambda^2, t)$, esto es, $P[(U^2 < t)] = F_r(\lambda^2, t)$, siendo $\lambda^2 = \mu_1^2, ..., \mu_r^2$.*

Observación 1.19 *Es fácil ver que $\lim_{n\to\infty}\{P[\bar{Q}_n < t|\theta] - P[Q_n < t|\theta]\} = 0$ uniformemente en θ y t, donde $\lambda_n^2(\theta) = n\sum\sum \xi^p(\theta)\xi^q(\theta)\bar{c_{pq}}^*(\theta)$.*

Teorema 1.20 Teorema de Wald
Sean $F_r(\lambda^2, t)$ la función de distribución definida en la Observación 1.18 y $\lambda_n(\omega, E_n)$ el estadístico de la razón de verosimilitud para probar la hipótesis $\xi^1(\theta), ..., \xi^k(\theta) = 0$. Se considera, además, la expresión definida en la Observación 1.19. Entonces, asumiendo la Proposición 1.17, tenemos $\lim_{n\to\infty}\{P[-2\log(\lambda_n(\omega, E_n)) < t|\theta] - F_r[\lambda_n^2(\theta), t]\} = 0$ uniformemente en t y θ. Si la hipótesis que se quiere probar es cierta, esto es, si θ es un punto de ω, $\lambda_n^2(\theta) = 0$ y, por lo tanto, el límite de la distribución de $-2\log(\lambda_n(\omega, E_n))$ es la distribución χ^2 de r grados de libertad.

1.1. Muestreo por importancia

Para la elaboración de este apartado se ha consultado esencialmente [10].
El muestreo por importancia, también conocido como muestreo ponderado o muestreo estratificado, es una técnica utilizada en estadística y métodos numéricos para mejorar la estimación de propiedades de una distribución de probabilidad o calcular valores difíciles de obtener directamente mediante simulación. La idea para mejorar la eficiencia de las estimaciones es, en lugar de seleccionar aleatoriamente observaciones según la ley de probabilidad que corresponde, asignar probabilidades de selección ponderadas a cada elemento de la población, dando más peso a ciertos subconjuntos que son de mayor interés o importancia. Esto puede mejorar la precisión de las estimaciones, especialmente cuando hay heterogeneidad en la población y se desea asegurar que ciertos grupos estén bien representados en la muestra.
Particularizando en nuestro trabajo, el muestreo por importancia se basa en la idea de que ciertos valores de las variables aleatorias en una simulación tienen más impacto en el parámetro que se estima que otros. Si estos valores más relevantes se enfatizan mediante el muestreo con mayor frecuencia, 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, $W(\vec{x}; \mu) = \frac{f(\vec{x};\mu=0)}{g(\vec{x};\mu)}$. 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 principales ventajas de 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 sin muestreo de importancia.
El método del muestreo por importancia es una de las técnicas más eficientes para minimizar la varianza.
Sea

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

donde H es la variable que se quiere estudiar y f es la función de densidad de X. Por simplificar, denotamos con un subíndice f la esperanza para indicar que se toma respecto a f.
Sea g otra función de densidad tal que Hf está dominada por g. Es decir, si $g(x) = 0 \Longrightarrow H(x)f(x) = 0$. Teniendo en cuenta la función de densidad $g(x)$, podemos representar l como:

$$l = \int H(X)\frac{f(x)}{g(x)}g(x)\,dx = E_g\left[H(x)\frac{f(X)}{g(X)}\right],$$

donde el subíndice g significa que la esperanza se toma con respecto a g. Esta densidad se denomina densidad de muestreo de importancia.

Observación 1.21 *Se puede comprobar de forma trivial que si $g(x) \neq 0$ entonces, $\frac{f}{g} \cdot g = f$. Por lo que obtendríamos el mismo resultado independientemente de g.*

Considerando lo anterior, si $x_1, ...x_n$ es una muestra aleatoria de g, es decir, $x_1, ...x_n$ son variables aleatorias independientes e igualmente distribuidas que tienen como función de densidad g, entonces

$$\hat{l} = \frac{1}{N}\sum_{i=1}^{N} H(x_k)\frac{f(X_k)}{g(X_k)} \tag{1.1}$$

es un estimador insesgado de l y se llama estimador de muestreo de importancia. A la siguiente proporción se denomina razón de verosimilitud:

$$W(x) = \frac{f(x)}{g(x)}.$$

En el caso particular donde no hay cambio de medida, es decir, $g = f$, se tiene que $W = 1$.

1.1.1. El método de minimización de la varianza

Debido a que la elección de la función de densidad del muestreo de importancia g está estrechamente relacionada con la varianza del estimador $\hat{l}$, consideramos el problema de minimizar la varianza de $\hat{l}$ con respecto a g, es decir,

$$min_g Var_g\left(H(X)\frac{f(X)}{g(X)}\right).$$

La solución al problema anterior es:

$$g^*(x) = \frac{|H(X)|f(X)}{\int |H(X)|f(X)\,dx}. \tag{1.2}$$

Asumiremos que $H(x) \geq 0$, luego se tiene que:

$$g^*(x) = \frac{H(X)f(X)}{\int H(X)f(X)\,dx} \tag{1.3}$$

y

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

Es importante darse cuenta de que aunque $\hat{l}$ es un estimador insesgado para cualquier función de probabilidad g que domine a $H(x)f(x)$, no todas esas funciones de probabilidad son apropiadas. Es necesario para elegir una buena función de probabilidad de muestreo

de importancia que el estimador definido en la ecuación (1.1) tenga una varianza finita, en otras palabras,

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

De esto se concluye que g no debería tener una "*cola más ligera*" que f y que, preferiblemente, el cociente de probabilidad $\frac{f}{g}$ debe estar acotado. En general, la implementación de la función densidad de muestreo de importancia óptima g^* según (1.2) y (1.3) es problemática. La principal dificultad radica en el hecho de que para derivar $g^*(x)$ es necesario conocer l. Pero a su vez, l es precisamente la cantidad que queremos estimar a partir de la simulación.

Como normalmente no podemos calcular g^*, renunciamos a encontrar $g*$ que minimiza funcionalmente la varianza y, buscamos el parámetro óptimo que minimiza la varianza en una familia de funciones. Entonces, otra propuesta que se hace es buscar el parámetro óptimo que minimiza la varianza en una familia de funciones. En particular, supongamos que $g(\cdot) = g(\cdot; u)$ pertenece a la familia:

$$\mathcal{F} = \{g(\cdot; v), v \in \varphi\}.$$

Entonces el problema de encontrar una función de densidad de muestreo de importancia óptima se reduce al siguiente problema de minimización paramétrica:

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

donde $W(X;u,v) = \frac{g(X;u)}{g(X;v)}$. Llamaremos vector de parámetros de referencia al vector v. Dado que bajo $g(\cdot; v)$ la esperanza $l = E_v[H(X)W(X;u,v)]$ es constante, la solución óptima del problema anterior coincide con la de:

$$min_{v\in\varphi} V(v),$$

donde $V(v) = E_v[H^2(X)W^2(X;u;v)] = E_u[H^2(X)W(X;u;v)]$.

También se ha propuesto otra alternativa basada en la entropía cruzada.

Definición 1.22 Entropía cruzada de Kullback-Leibler (Kullback-Leibler Cross-Entropy) *Extraida de [10].*
La entropía cruzada de Kullback-Leibler entre dos funciones de densidad continuas g y h se define como

$$\mathcal{D}(g,h) = E_g\left[\log\frac{g(X)}{h(X)}\right] = \int g(x)\log\frac{g(x)}{h(x)}\,dx = \int g(x)\log g(x)\,dx - \int g(x)\log h(x)\,dx.$$

1.1.2. Método de la entropía cruzada ($\mathcal{D}_{CE}$)

El método de entropía cruzada se emplea para encontrar la distribución de probabilidad que minimiza la divergencia de Kullback-Leibler (KL) entre la distribución objetivo y la distribución propuesta, es decir, mide la discrepancia entre la distribución que se toma como hipótesis nula y la que se considera como alternativa.
La parte teórica que sigue se muestra para una variable de una dimensión $x \in \mathbb{R}$, pero se podría generalizar para una variable de n dimensiones, $x \in \mathbb{R}^n$.
El objetivo es minimizar la entropía cruzada, es decir, la distancia entre las dos funciones de distribución, con el objetivo de mejorar la calidad de las predicciones del modelo. Esto

es equivalente a maximizar respecto a la variable de la hipótesis nula, μ_{H_0}. Es una medida útil porque penaliza más fuertemente las predicciones incorrectas con probabilidades altas. Se define de la siguiente forma:

$$\mathcal{D}_{CE} = \int H(x) f(x;\mu_0) \log f(x;\mu_1)\, dx = E_{\mu_0}[H(X) \log f(X;\mu_1)],$$

siendo $f(x;\mu_0)$ la distribución considerada en la hipótesis nula, $f(x;\mu_1)$ en la alternativa y

$$H(x) = \begin{cases} 1 & \text{si } x > 0, \\ 0 & \text{si } x \leq 0. \end{cases} \tag{1.4}$$

Tenemos que encontrar la función de densidad h tal que la distancia entre el óptimo del muestreo por importancia de la función de densidad g^* en (1.2) y h sea mínima. Por tanto, el resultado obtenido por el método de minimizar la varianza y el obtenido por el método de Cross Entropy son equivalentes. Teniendo en cuenta la definición de entropía cruzada de Kullback-Leibler (1.1.2), como el primer término no depende de μ_1, minimizar la distancia de Kullback-Leibler entre g^* y $f(x;\mu_1)$ es equivalente a maximizar respecto de μ_1,

$$\int H(x) f(x;\mu_0) \log f(x;\mu_1)\, dx = E_{\mu_0}[H(X) \log f(x;\mu_1)].$$

Además, para nuestro caso tiene la ventaja que $\log(f(x,\mu_1))$ tiene cierta similitud con la definición de máxima verosimilitud si probamos con la H_1, por lo que tendrá cierta similitud encontrar el óptimo del método de la máxima verosimilitud y el de la entropía cruzada.

En el trabajo se considera la notación μ_{IS} para indicar cuál es el mejor μ que se ajusta a los datos observados considerando el método del muestreo por importancia.
En las líneas precedentes se ha explicado en qué consiste y cómo se calcula el método de muestreo por importancia. Recordamos lo que queremos calcular: $\int_{x>x_0} f(x)\, dx$ y, es aquí, donde nos apoyaremos en el muestreo por importancia para el cálculo.

1.2. Aplicación al cálculo de p-valores

Queremos utilizar este método para el cálculo del p-valor, $\mathfrak{p}$, considerando $H(x)$ la función definida anteriormente.
Se trata de un método muy ineficiente si estudiamos casos improbables porque la mayoría se pierden, de forma sencilla necesitamos muestrear un cierto de número de veces: $\frac{1}{\mathfrak{p}}$.
Intuitivamente sería conveniente basarnos en una distribución que esté poblada en un entorno de x_0.
La propuesta es muestrear con una familia relacionada con las hipótesis consideradas en el contraste: H_0 y H_1. En los casos sencillos, con una paramétrica relacionada con $f(x)$, por ejemplo $N(\mu, 1)$ para la distribución normal y, en los casos más complicados, usaremos H_1 y los parámetros correspondientes a x_0.

Otro punto interesante a estudiar es la comparación analítica de las varianzas entre no aplicar el muestreo por importancia y aplicarlo. De modo teórico se obtienen los siguientes resultados:
Recordamos la definición del p-valor como la probabilidad de que ocurra un suceso más extremo que otro que se ha fijado con anterioridad, es decir,

$$\mathfrak{p} = E_f[H(x - x_0)] = \int H(x - x_0) f(x)\, dx = \int_{x>x_0} f(x)\, dx,$$

donde $H(x - x_0)$ es la función a trozos definida previamente en el trabajo expuesto.
Consideramos los siguientes puntos.
Por un lado, aplicando la definición de esperanza y suponiendo que existe la función de densidad $f(x)$, sabemos que:

$$E_f[H(X)] = \int H(x)f(x)\,dx.$$

Por otro lado, aplicando el muestreo por importancia se tiene que:

$$E_f[H(X)] = \int H(x)f(x)\,dx = \int H(x)\frac{f(x)}{g(x)}g(x)\,dx = \int H(x)W(x)g(x)\,dx = E_g[H(x)W(x)].$$

Hallamos la varianza de ambas expresiones:

$$Var_f(H(X)) \overset{\text{Def. 1.6}}{=} E_f[H^2(x)] - E_f^2[H(x)] \overset{H(x)=H^2(x)}{=} E_f[H(x)] - E_f^2[H(x)] \overset{\text{Def. 1.8}}{=} \mathfrak{p} - \mathfrak{p}^2.$$

Como se considera q_0 grande y el p-valor, $\mathfrak{p}$, toma un valor pequeño, podemos prescindir del segundo término, $Var_f(H(X)) \approx \mathfrak{p}$.

$$\begin{aligned} Var_g(H(X)W(X)) &\overset{\text{Def. 1.6}}{=} E_g[H^2(X)W^2(X)] - E_g^2[H(X)W(X)] \\ &= E_g[H^2(X)W^2(X)] - \left(\int H(x)W(X)g(x)\,dx\right)^2 \\ &= E_g[H^2(X)W^2(X)] - \left(\int H(x)\frac{f(x)}{g(x)}g(x)\,dx\right)^2 \\ &= E_g[H^2(X)W^2(X)] - \left(\int H(x)f(x)\,dx\right)^2 \\ &\overset{\text{Def. 1.8}}{=} E_g[H^2(X)W^2(X)] - \mathfrak{p}^2. \end{aligned}$$

Por el mismo razonamiento, $Var_g(H(X)W(X)) \approx E_g[H^2(X)W^2(X)]$.
Nos basaremos en estos resultados para estudiar cuanto se mejora al aplicar el método de muestreo por importancia frente a no aplicarlo.

Capítulo 2

Distribución normal

Vamos a estudiar la distribución normal. Es cierto que conocemos de antemano los resultados, pero es un buen ejemplo para comprobar que los métodos y los resultados obtenidos están en consonancia con lo esperado.
En particular, para el problema que queremos estudiar, consideramos la cola superior. Queremos calcular el p-valor de x_0, variable aleatoria independiente e igualmente distribuida que sigue una distribución $N(0,1)$, es el dato observado sobre el que queremos estudiar qué tan probable / improbable es que se dé dicha observación. Vamos a probar, siguiendo la propuesta del muestreo con una distribución gaussiana centrada en μ, es decir, $N(\mu,1)$. Para los desarrollos definiremos $n \in \mathbb{N}$ como la dimensión de la variable aleatoria X. Resolver este problema es equivalente a considerar $\bar{x} \sim N\left(\mu, \frac{1}{\sqrt{n}}\right)$.

Demostración.
Veamos que ambos problemas son equivalentes.
Sabemos por definición que $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ y $x_i \sim N(\mu,1)$, $\forall i \in \{1,...,n\}$.
Recordamos las propiedades siguientes:
Sean X, Y variables aleatorias independientes, se tiene que:

(I) $E[X+Y] = E[X] + E[Y]$.

(II) $Var(X+Y) = Var(X) + Var(Y)$.

Considerando lo anterior:

- $E\left[\sum_{i=1}^{n} x_i\right] = E[x_1 + ... + x_n] \overset{(I)}{=} E[x_1] + ... + E[x_n] \overset{E[x_i]=\mu, \forall i \in \{1,...,n\}}{=} n\mu.$
- $Var\left(\sum_{i=1}^{n} x_i\right) = Var(x_1 + ... + x_n) \overset{(II)}{=} Var(x_1) + ... + Var(x_n) \overset{Var(x_i)=1, \forall i \in \{1,...,n\}}{=} n.$
- $\sigma\left(\sum_{i=1}^{n} x_i\right) = \sqrt{n}.$

Luego, se tiene que $\sum_{i=1}^{n} x_i \sim N(n\mu, \sqrt{n})$.
Calculamos la media y desviación típica de $\bar{x}$.

- $E[\bar{x}] = E\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}n\mu.$
- $Var(\bar{x}) = Var\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sqrt{n}^2 = \frac{1}{n}.$
- $\sigma(\bar{x}) = \frac{1}{\sqrt{n}}.$

Por tanto, $\bar{x} \sim N\left(\mu, \frac{1}{\sqrt{n}}\right)$ y queda probado que los problemas son equivalentes. □

2.1. Aplicación método de la máxima verosimilitud

Recordamos que la hipótesis nula es que los datos observados siguen una distribución $N(0,1)$ y la alternativa es que siguen la ley $N(\mu,1)$. Queremos calcular el μ óptimo considerando el método de la máxima verosimilitud.

Teorema 2.1 *Sea X una variable aleatoria que sigue una distribución normal. El valor que satisface: $\mu_{MV} = max\{\log(\mathcal{L})\}$ es*

$$\mu_{MV} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{2.1}$$

Demostración.

$$\log(\mathcal{L}) = \log(\rho(\vec{x};\mu)) = \log\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{\frac{-(x_i-\mu)^2}{2}}\right) = \sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{2\pi}}e^{\frac{-(x_i-\mu)^2}{2}}\right)$$
$$= -\frac{1}{2n}\sum_{i=1}^{n}(x_i-\mu)^2 + \sum_{i=1}^{n}\log\left(\sqrt{2\pi}\right).$$

La expresión anterior es continua y derivable, hallamos el máximo

$$0 = \frac{\partial\log(\mathcal{L})}{\partial\mu} = -\frac{1}{2n}\sum_{i=1}^{n}2(x_i-\mu)(-1) \Longrightarrow \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu) = 0 \Longrightarrow \sum_{i=1}^{n}(x_i-\mu) = 0$$
$$\Longrightarrow \sum_{i=1}^{n}x_i - n\mu = 0 \Longrightarrow \mu = \frac{\sum_{i=1}^{n}x_i}{n} = \bar{x}.$$

Vamos a asegurarnos de que el candidato obtenido es máximo.

$$\frac{\partial^2\log(\mathcal{L})}{\partial\mu^2} = -1 < 0.$$

Luego,

$$\mu_{MV} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

es un máximo de verosimilitud. □

2.2. Cálculo del cociente de verosimilitud

Comparamos la hipótesis nula con la hipótesis alternativa. Para ello, recordamos la función de densidad de la distribución normal: $\rho(\vec{x};\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$. Como hemos supuesto que la variable aleatoria es independiente y $x_i \sim N(\mu, 1)\ \forall i \in \{1, ..., n\}$, se considera $\rho(\vec{x};\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2}}$.
Tomando $\mu = \mu_{MV}$ y aplicando logaritmos tenemos:

$$\log(\mathcal{L}(\mu_{MV})) = \log(\rho(\vec{x};\mu=\mu_{MV})) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x_i-\mu_{MV})^2}{2}}\right) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}} e^{\frac{-(x_i-\mu_{MV})^2}{2}}\right). \tag{2.2}$$

Por otro lado, tomamos $\mu = 0$ y volviendo a aplicar logaritmos obtenemos:

$$\log(\mathcal{L}(0)) = \log(\rho(x;\mu=0)) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-x_i^2}{2}}\right) = \sum_{i=1}^{n} \left(-\log(\sqrt{2\pi}) - \frac{x_i^2}{2}\right). \tag{2.3}$$

Considerando lo anterior, calculamos el estadístico:

$$\begin{aligned}
q(\vec{x}) &= -2\log\left(\frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu=\mu_{MV})}\right) \\
&= -2\log\left(\frac{\prod_{i=1}^{n}\rho(x_i;\mu=0)}{\prod_{i=1}^{n}\rho(x_i;\mu=\mu_{MV})}\right) \\
&= -2\left[\log(\prod_{i=1}^{n}\rho(x_i;\mu=0)) - \log(\prod_{i=1}^{n}\rho(x_i;\mu=\mu_{MV}))\right] \\
&= -2\left[\sum_{i=1}^{n}\log\rho(x_i;\mu=0) - \sum_{i=1}^{n}\log\rho(x_i;\mu=\mu_{MV})\right] \overset{(2.2)\text{ y }(2.3)}{=} \\
&= -2\left[\sum_{i=1}^{n}\left(-\log(\sqrt{2\pi}) - \frac{x_i^2}{2}\right) - \left(\sum_{i=1}^{n}\left(-\log(\sqrt{2\pi}) - \frac{(x_i-\mu_{MV})^2}{2}\right)\right)\right] \\
&= -2\left[-n\log(\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}x_i^2 + n\log(\sqrt{2\pi}) + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_{MV})^2\right] \\
&= -2\left[-\frac{1}{2}\sum_{i=1}^{n}x_i^2 + \frac{1}{2}\mu_{MV}^2 n + \frac{1}{2}\sum_{i=1}^{n}(x_i^2 - 2x_i\mu_{MV})\right] \\
&= -n\mu_{MV}^2 + \sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}x_i^2 + 2\mu_{MV}\sum_{i=1}^{n}x_i \\
&= -n\mu_{MV}^2 + 2\mu_{MV}\sum_{i=1}^{n}x_i \overset{(2.1)}{=} \\
&= -n\mu_{MV}^2 + 2n\mu_{MV}\mu_{MV} \\
&= n\mu_{MV}^2.
\end{aligned}$$

Teorema 2.2 *Sea el cociente de verosimilitud $q(\vec{x}) = n\mu_{MV}^2$ entonces sigue una distribución χ^2 de un grado de libertad en H_0.*

Demostración.
Sabemos por (2.1) que $\mu_{MV} = \bar{x}$. Además, $\bar{x} \sim N(\mu, \frac{1}{\sqrt{n}})$. Entonces, $\mu_{MV} \sim N(\mu, \frac{1}{\sqrt{n}})$. Por tanto, $q(\vec{x}) = n\mu_{MV}^2$ tiene como distribución el cuadrado de una normal, es decir, χ^2 de un grado de libertad. □

Lema 2.3 *Si una variable aleatoria tiene como distribución el cuadrado de una normal, es equivalente a decir que sigue una distribución χ^2 de un grado de libertad.*

Demostración.
Recordamos la función de densidad: $\rho(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{\frac{-x}{2}}, x \in (0, \infty)$.
Sea $n \in \mathbb{N}^+$,

$$\Gamma\left(\frac{n}{2}\right) = \begin{cases} (\frac{n}{2}-1)! & \text{si n es par,} \\ \frac{(n-1)!}{2^{n-1}(\frac{n}{2}-\frac{1}{2})!}\sqrt{\pi} & \text{si n es impar.} \end{cases}$$

En nuestro caso, tomamos n=1,

$$\Gamma\left(\tfrac{1}{2}\right) = \tfrac{(1-1)!}{2^{1-1}(\frac{1}{2}-\frac{1}{2})!}\sqrt{\pi} = \tfrac{0!}{2^0 0!}\sqrt{\pi} = \sqrt{\pi},$$
$$\rho(x) = \tfrac{1}{2^{\frac{1}{2}}\sqrt{\pi}} x^{\frac{1}{2}-1} e^{\frac{-x}{2}}.$$

Por otro lado, teníamos que $x \sim N(0,1)$.
Definimos una nueva variable aleatoria, Y, tal que $Y = X^2$. Queremos encontrar la función de densidad de la función Y.

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) \overset{\text{X distrib. sim. resp. del 0}}{=} P(-\sqrt{y} \leq x \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx.$$

Derivando respecto de y y aplicando la Regla de la Cadena:

$$f_y(y) = \tfrac{d}{dy}\int_{-\sqrt{y}}^{\sqrt{y}} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = \tfrac{1}{\sqrt{2\pi}}\tfrac{d}{dy}\int_{-\sqrt{y}}^{\sqrt{y}} e^{-\frac{x^2}{2}}\, dx = \tfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tfrac{1}{2\sqrt{x}}\Big|_{-\sqrt{y}}^{\sqrt{y}} =$$
$$\tfrac{1}{\sqrt{2\pi}}\left[e^{-\frac{\sqrt{y}^2}{2}}\tfrac{1}{2\sqrt{y}} - e^{-\frac{(-\sqrt{y})^2}{2}}\tfrac{1}{2(-\sqrt{y})}\right] = \tfrac{2}{\sqrt{2\pi y}} e^{-\frac{y}{2}}.$$

Luego, Y sigue una distribución χ^2 de 1 grado de libertad y queda probado el lema. □

2.3. Cálculo del p-valor por muestreo por importancia

Vamos a calcular la probabilidad de la cola superior, $\mathfrak{p} = \int_c^\infty \rho(x;0)\, dx$, para $c \longrightarrow \infty$. Probamos con funciones que siguen una distribución $N(\mu, 1)$. En este caso no es necesario el contraste de hipótesis ya que podemos considerarlo directamente sobre la variable x.

Nos basamos en el método de "*Cross Entropy*" para buscar el μ óptimo. Recordamos que:

$$\mathcal{D}_{CE} = \int H(x)\rho(x;\mu)\log\rho(x;\mu)\, dx = E_{\mu=0}[H(x)\log\rho(x;\mu)].$$

Como $x \sim N(\mu, 1)$; $\rho(x;\mu) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2}}$, reiteramos que:

$$H(x) = \begin{cases} 0 & \text{si } x < c, \\ 1 & \text{si } x \geq c. \end{cases}$$

Consideramos la cola superior y denotamos por $\Phi(c) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ la función de distribución de la normal. Aplicamos el método de la entropía cruzada para la cola superior:

$$\begin{aligned}
\mathcal{D}_{CE} &= \int_c^\infty 1 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \log\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}\right) dx \\
&= \int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left[-\log\sqrt{2\pi} - \frac{(x-\mu)^2}{2}\right] dx \\
&= -\log\sqrt{2\pi}(1-\Phi(c)) - \frac{1}{2}\int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} (x-\mu)^2\, dx \\
&= -\log\sqrt{2\pi}(1-\Phi(c)) - \frac{1}{2}\int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} (x^2 - 2x\mu + \mu^2)\, dx.
\end{aligned}$$

Queremos maximizar respecto a μ. Teniendo en cuenta que la expresión anterior es continua y derivable, $0 = \frac{\partial \mathcal{D}_{CE}}{\partial \mu}$

$$\begin{aligned}
0 &= \frac{\partial \mathcal{D}_{CE}}{\partial \mu} \\
&= -\frac{1}{2}\int_c^\infty \frac{1}{\sqrt{2\pi}}(-2x + 2\mu) e^{-\frac{x^2}{2}} \\
&\Longrightarrow 0 = \int_c^\infty \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} - \mu \int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\
&\Longrightarrow \int_c^\infty \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} = \mu \int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\
&\Longrightarrow \mu = \frac{-\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\Big|_c^\infty}{1-\Phi(c)} = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{c^2}{2}}}{1-\Phi(c)} \overset{Corolario\ 2{,}5}{\approx} c \\
&\Longrightarrow \mu_{IS} \approx c.
\end{aligned}$$

Recordamos nuestra hipótesis, $H_1 : \mu = c$. Luego, el resultado obtenido,

$$\mu_{IS} \approx c. \tag{2.4}$$

encaja con la hipótesis alternativa considerada. Utilizaremos este resultado para probar el caso general.

Teorema 2.4 *Si $x \to \infty$,*

$$\int_x^\infty e^{-\frac{t^2}{2}}\, dt \approx e^{-\frac{x^2}{2}} \left[\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - ...\right].$$

Aproximación extraída de [8].

Demostración.
Vamos a probar que se verifica la aproximación anterior. Considerando $u = \frac{1}{t}$ y desarrollando por el método de integración por partes tenemos que:

$$\int_x^\infty e^{-\frac{t^2}{2}}\,dt = -\frac{1}{t}e^{-\frac{t^2}{2}}\Big|_x^\infty - \int_x^\infty \frac{1}{t^2}e^{-\frac{t^2}{2}}\,dt = \frac{1}{x}e^{-\frac{x^2}{2}} - \int_x^\infty \frac{1}{t^2}e^{-\frac{t^2}{2}}\,dt.$$

Podríamos seguir desarrollando por partes la integral obtenida y aplicar inducción pero, solamente nos interesa el primer término. Falta demostrar que $-\int_x^\infty \frac{1}{t^2}e^{-\frac{t^2}{2}}\,dt.$ tiende a 0 y que tiende más rápido a 0 que $\frac{1}{x}e^{-\frac{x^2}{2}}$. Desarrollando tenemos que:

$$\int_x^\infty \frac{1}{t^2}e^{-\frac{t^2}{2}}\,dt = \int_x^\infty \frac{t}{t^3}e^{-\frac{t^2}{2}}\,dt \leq \frac{1}{x^3}\int_x^\infty te^{-\frac{t^2}{2}}\,dt = \frac{1}{x^3}e^{-\frac{t^2}{2}} = \frac{1}{x^2}\frac{1}{x}e^{-\frac{t^2}{2}}.$$

Entonces,

$$\int_x^\infty e^{-\frac{t^2}{2}}\,dt = \frac{1}{x}e^{-\frac{t^2}{2}} - \frac{1}{x^2}\frac{1}{x}e^{-\frac{t^2}{2}}.$$

Luego el segundo término tiende más rápido a 0 que el primero ya que $x \longrightarrow \infty$.
Además,

$$0 < \int_x^\infty \frac{1}{t^2}e^{-\frac{t^2}{2}}\,dt.$$

porque $\frac{1}{t^2} > 0$ y la exponencial siempre es positiva.
Por tanto, queda probada la aproximación y justificado que nos quedemos con el primer término. □

Corolario 2.5 *Considerando que la aproximación es cierta y* $x \to \infty$*,*

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

Demostración.
El resultado anterior se tiene de forma directa por el Teorema 2.4. □

2.4. Cálculo del p-valor para la cola inferior por muestreo por importancia

De forma análoga aplicamos el método de entropía cruzada para la cola inferior.

$$\begin{aligned}
\mathcal{D}_{CE} &= \int_{-\infty}^{-c} 1 \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \log\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}}\right) dx \\
&= \int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\left[-\log\sqrt{2\pi} - \frac{(x-\mu)^2}{2}\right] dx \\
&= \int_{-\infty}^{-c} -\frac{1}{\sqrt{2\pi}}\log\sqrt{2\pi}e^{-\frac{x^2}{2}} - \int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\frac{(x-\mu)^2}{2}\,dx \\
&= -\log\sqrt{2\pi}\ \ \Phi(-c) - \frac{1}{2}\int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}(x-\mu)^2\,dx \\
&= -\log\sqrt{2\pi}\ \ \Phi(-c) - \frac{1}{2}\int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}(x^2 - 2x\mu + \mu^2)\,dx.
\end{aligned}$$

Lema 2.6 *Sea F una función de distribución simétrica y $c \in \mathbb{R}$ entonces se tiene que $F(-c) = 1 - F(c)$.*

Demostración.
Suponemos: $F(-c) = 1 - F(c)$.
Procedemos de forma directa:

$$F(-c) = 1 - F(c) \longrightarrow F(-c) + F(c) = 1 \longrightarrow P((-\infty, -c)) + P((-\infty, c))$$
$$= P((-\infty, -c)) + P((c, \infty)) + P((-c, c)) \overset{x \sim \mathcal{N}(0,\sigma)}{=} P((-\infty, -c)) + P((c, \infty)) + P((-c, c))$$
$$= P(\mathbb{R}) = 1.$$

Luego, queda probado el Lema 2.6. □

Queremos maximizar respecto a μ. Teniendo en cuenta que la expresión anterior es continua y derivable, $0 = \frac{\partial \mathcal{D}_{CE}}{\partial \mu}$

$$\begin{aligned}
0 &= \frac{\partial \mathcal{D}_{CE}}{\partial \mu} \\
&= -\frac{1}{2} \int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}} (-2x + 2\mu) e^{-\frac{x^2}{2}} \\
&\Longrightarrow \int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} = \mu \int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\
&\Longrightarrow \mu_{IS} = \frac{-\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \Big|_{-\infty}^{-c}}{\Phi(-c)} = \frac{\frac{-1}{\sqrt{2\pi}} e^{-\frac{c^2}{2}}}{\Phi(-c)} \overset{\text{Lema 2.6}}{=} \frac{\frac{-1}{\sqrt{2\pi}} e^{-\frac{c^2}{2}}}{1 - \Phi(c)} \approx -c.
\end{aligned}$$

La última aproximación se obtiene como resultado del Teorema 2.4 y, como era de esperar, hemos obtenido el resultado que buscábamos.

2.5. Cálculo del p-valor considerando la doble cola.

Sabemos que en el caso de la distribución normal es trivial ya que las colas son simétricas. El interés está en abordarlo como si no lo fuesen para ilustrar un problema que plantea el método.
Queremos calcular: $P(|x| \geq c) = P(x > c) + P(x < -c)$.
Aplicando el método de la entropía cruzada:

$$\begin{aligned}
\mathcal{D}_{CE} &= \int H(x) f(x; \mu_0) \log f(x; \mu_1)\, dx \\
&= \int f(x; \mu = 0) \log \left(\frac{f(x; \mu = 0)}{f(x; \mu)} \right) dx \\
&= \int_{|x| \geq c} f(x; \mu = 0) \log \left(-\frac{1}{2} q(x) \right) dx \\
&= -\frac{1}{2} \int_{|x| \geq c} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} (-\mu^2 + 2\mu x)\, dx.
\end{aligned}$$

De nuevo, considerando que la expresión precedente es continua y derivable, maximizamos respecto de μ:

$$
\begin{aligned}
0 &= \frac{\partial \mathcal{D}_{CE}}{\partial \mu} \\
&= -\frac{1}{2\sqrt{2\pi}} \int_{|x|\geq c} e^{\frac{-x^2}{2}} (-2\mu + 2x)\, dx \\
&= -\frac{1}{2\sqrt{2\pi}} \left[\int_{-\infty}^{-c} e^{\frac{-x^2}{2}} (-2\mu + 2x)\, dx + \int_{c}^{\infty} e^{\frac{-x^2}{2}} (-2\mu + 2x)\, dx \right] \\
&= \int_{-\infty}^{-c} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \mu + \int_{c}^{t} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \mu - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-c} x e^{\frac{-x^2}{2}}\, dx - \frac{1}{\sqrt{2\pi}} \int_{c}^{\infty} x e^{\frac{-x^2}{2}}\, dx \\
&= \mu\Phi(-c) + \mu(1 - \Phi(c)) + \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \Big|_{-\infty}^{-c} + \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \Big|_{c}^{\infty} \\
&= \mu\Phi(-c) + \mu(1 - \Phi(c)) + \frac{1}{\sqrt{2\pi}} e^{\frac{-c^2}{2}} - \frac{1}{\sqrt{2\pi}} e^{\frac{-c^2}{2}} \\
&= \mu\Phi(-c) + \mu(1 - \Phi(c)) \overset{\text{Lema 2.6}}{=} \\
&= 2\mu\Phi(-c) = 2\mu(1 - \Phi(c)) \Longrightarrow \mu_{IS} = 0.
\end{aligned}
$$

De lo anterior se concluye que, la función elegida para hacer muestreo por importancia no es óptima ya que nuestra hipótesis nula era $H_0 : \mu = 0$, y, considerando las dos colas, obtenemos el mismo resultado que no pesando, es decir, el resultado obtenido al aplicar el método no mejora.

2.6. Observaciones

Es crucial tener presente el objetivo del trabajo: establecer las hipótesis y condiciones que deben ser verificadas para formular un teorema. Reconocemos que esto no es una tarea sencilla. Por lo tanto, también prestaremos atención a los puntos donde sabemos que nuestra conjetura no es aplicable.

Hemos visto que al considerar ambas colas, el resultado obtenido al aplicar muestreo por importancia considerando la función escogida, no mejora en comparación con no pesar. Podríamos probar con otras funciones para pesar, pero nuestro interés está en considerar funciones de la familia y, así, poder proceder de este modo en otros casos más complejos.

Otra posible forma de solventar este problema, para calcular el p-valor considerando ambos extremos, sería considerar la suma de dos distribuciones gaussianas centradas en $-c$ y c simultáneamente. Esto son posibles propuestas para resolver el problema que se presenta, se dejan planteadas.

Se puede concluir que en el caso de la gaussiana si se estudian las dos colas y extendido a casos genéricos como veremos más adelante, donde μ puede tomar tanto valores positivos como negativos, en estos casos, la propuesta que estamos considerando nosotros de muestreo por importancia no mejora los resultados.

A continuación se comenta otra posible alternativa más detallada. Recordamos el problema que estamos estudiando. Dados unos datos observados que siguen una distribución normal y c es el estadístico observado. Queremos calcular $P(|x| \geq c)$, hallar el p-valor de c.
Considerando los resultados obtenidos en el trabajo, $P(|x| \geq c) = P(x > c) + P(x < -c)$

por lo que es necesario considerar ambas colas a la vez. Como se ha probado, estudiarlas de forma independiente no supone ningún inconveniente.
Si la distribución a estudiar está centrada en $\mu = 0$, el problema se soluciona de forma inmediata considerando una de las colas y un factor 2. Y, en caso de no estar centrada en el 0, tipificamos la función y sobre esta expresión auxiliar actuaríamos de forma análoga.
Por lo que se puede concluir que, en la distribución normal, esta limitación comentada se puede resolver de forma sencilla pero obviamente es necesario tenerlo en cuenta ya que en otras distribuciones no simétricas afectará pero, no es inmediato establecer el valor de este factor.

2.7. ¿Cuánto mejoramos al aplicar el método?

Basándonos en los resultados del Capítulo 1 y particularizando para el caso de la distribución normal:

$$f \sim N(0,1); f(x, \mu = 0) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \textit{ sin aplicar el muestreo por importancia.}$$
$$g \sim N(\mu,1); g(x, \mu) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2}}, \textit{ aplicándolo.}$$

Fijamos $c \in \mathbb{R}$,

$$H_c(x) = \begin{cases} 0 & \text{si } x < c, \\ 1 & \text{si } x \geq c. \end{cases}$$

Considerando las hipótesis anteriores, se tiene:

$$\begin{aligned} Var_g(H(X)W(X)) \approx E_g[H^2(X)W^2(X)] &= \int H(X)\frac{f(x)}{g(x)} f(x)\, dx \\ &= \lim_{t\to\infty} \int_c^t \frac{\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2}}} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx \\ &= \lim_{t\to\infty} \int_c^t \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2}}{e^{\frac{-(x-\mu)^2}{2}}}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \lim_{t\to\infty} \int_c^t e^{-x^2 + \frac{(x-\mu)^2}{2}}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \lim_{t\to\infty} \int_c^t e^{-x^2 + \frac{x^2 - 2\mu x + \mu^2}{2}}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \lim_{t\to\infty} \int_c^t e^{\frac{-x^2}{2} - \mu x + \frac{\mu^2}{2}}\, dx \\ &= \frac{e^{\frac{\mu^2}{2}}}{\sqrt{2\pi}} \lim_{t\to\infty} \int_c^t e^{\frac{-x^2}{2} - \mu x}\, dx \end{aligned}$$

$$= \frac{e^{\frac{\mu^2}{2}}}{\sqrt{2\pi}} \lim_{t\to\infty} \int_c^t e^{\frac{\mu^2}{2} - \left(\frac{x+\mu}{\sqrt{2}}\right)^2} dx$$

$$\overset{\text{Cambio de variable: } u = \frac{x+\mu}{\sqrt{2}}}{=} \frac{e^{\frac{\mu^2}{2}}\sqrt{2}}{\sqrt{2\pi}} \lim_{t\to\infty} \int_{\frac{c+\mu}{\sqrt{2}}}^t e^{\frac{\mu^2}{2} - u^2} du$$

$$= \frac{e^{\frac{\mu^2}{2}}}{\sqrt{\pi}} \lim_{t\to\infty} \int_{\frac{c+\mu}{\sqrt{2}}}^t e^{\frac{\mu^2}{2} - \frac{2u^2}{2}} du$$

$$= \frac{e^{\mu^2}}{\sqrt{\pi}} \lim_{t\to\infty} \int_{\frac{c+\mu}{\sqrt{2}}}^t e^{\frac{-2u^2}{2}} du$$

$$\overset{\text{Cambio de variable: } w = \sqrt{2}u}{=} \frac{e^{\mu^2}}{\sqrt{2\pi}} \lim_{t\to\infty} \int_{c+\mu}^t e^{\frac{-w^2}{2}} dw$$

$$\overset{\text{Def. 1.4}}{=} e^{\mu^2}(1 - \Phi(c+\mu)).$$

Queremos representar la comparativa de las varianzas respecto de μ. Para ello consideramos el logaritmo decimal del cociente y, de esta forma, se aprecian mejor los resultados que queremos analizar.
La programación realizada se encuentra adjunta en el archivo " Comparación varianzas distrib. normal" y se basa en la siguiente idea. Realizamos un barrido en c y tomamos $\mu = c$, que recordamos que es nuestra hipótesis. Posteriormente representamos el logaritmo decimal del cociente de las varianzas respecto de c. La gráfica resultante se muestra en la Figura 2.1.

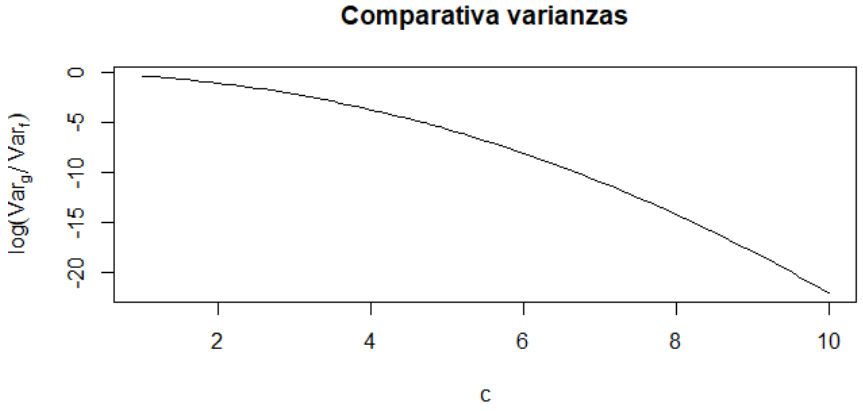

Figura 2.1: Comparativa de las varianzas aplicando y sin aplicar el muestreo por importancia respecto de c, logaritmo decimal del cociente de la varianza con respecto a la no pesada.

Como era de esperar, en $\mu = 0$, $Var_f = Var_g \Longrightarrow log\left(\frac{Var_g}{Var_f}\right) = 0$.
Además, en esta gráfica se observa como para valores pequeños de c no se gana mucho pero para valores grandes de c mejoramos considerablemente. Para 5 sigmas, mejoramos la varianza 5 órdenes de magnitud, es decir, reducimos en esos 5 órdenes de magnitud el número de datos a muestrear.

2.8. Cálculo por muestreo

En este caso sabemos cuáles van a ser los resultados numéricos ya que lo hemos demostrado. Sin embargo, lo planteamos a modo de ejemplo para casos en los que no se puede demostrar. Para ello, como hemos comentado previamente, nos basaremos en la herramienta de R. Recordamos el resultado demostrado $\mu_{IS} \approx c$, si $c \longrightarrow \infty$.

En primer lugar, generamos datos que siguen una distribución normal, muestreamos dichos datos y vemos a ver cuantos de ellos son superiores a q_0, siendo q_0 el estadístico de los datos observados. Calculamos el p-valor del muestreo como el cociente de la suma de los pesos tales que superan a q_0 y el número de veces que se hace el experimento. Y, hallamos la varianza como el cociente del cuadrado de la suma de los pesos que son superiores a q_0 entre el número de veces que se muestrea.
El p-valor del muestreo y la varianza, quedarán en función de μ y se representan gráficamente respecto a este parámetro.
Para realizar esta comparativa vamos a fijar el parámetro c, recordamos que $c = \sqrt{q_0}$, realizamos un barrido sobre μ. Esto se ha programado en el archivo adjunto: "Distribución normal" los cuales están comentados y, en el anexo se hace referencia al código programado (A.3.1).

En la Figura 2.2, representamos el p-valor experimental respecto de μ y, además, mostraremos con una recta azul el p-valor teórico que utilizaremos como valor de referencia para comparar el p-valor obtenido a partir del método de muestreo. Nos apoyamos en la función *pnorm* de R para obtener el p-valor teórico.

En la Figura 2.3, se representa el logaritmo en base 10 de la varianza respecto de μ, se muestra el logaritmo para que la función resultante sea cuadrática y poder observar de una forma más clara el mínimo. A su vez, se representa con una recta vertical el valor $\mu = c$ que debe encontrarse en un entorno del mínimo de la función como hemos visto de forma analítica.

Se han considerado distintos valores para q_0 para poder sacar conclusiones más generalizadas. Sabemos que mostrar solamente una representación gráfica, sin reflejar un desarrollo analítico no sirve como demostración pero, en este caso, ya lo habíamos demostrado de forma analítica y la representación numérica tiene como fin apoyar la demostración citada anteriormente en el trabajo.

En cuanto a la representación, hemos prescindido del dato más anómalo para que el resultado se viese mucho más claro y la escala fuese más precisa. Está comentado en el propio código y también se menciona en el apéndice.

Se han hecho diferentes pruebas variando los diferentes parámetros; se adjuntan algunas de ellas en el apéndice (A.4.1). En el documento mostraremos uno de los muchos casos numéricos obtenidos a partir de las simulaciones. Para detallar el análisis se han considerado todos las representaciones producidas y así, poder generalizar en mayor medida las conclusiones y cometer menos errores particularizando estas observaciones. A continuación se muestran dos gráficas representativas:

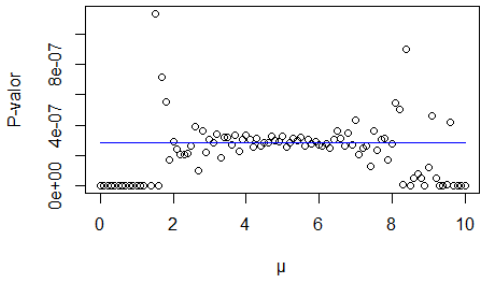

Figura 2.2: P-valor en función de μ, tomando q_0=25 y tamaño de la muestra 1000.

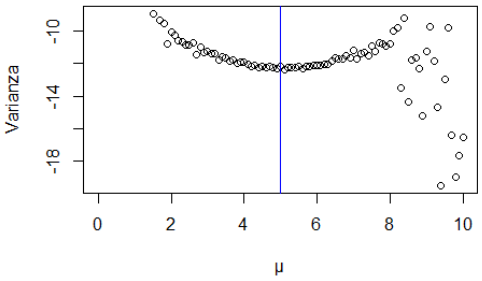

Figura 2.3: Variación del logaritmo de la varianza en función de μ tomando q_0=25 y tamaño de la muestra 1000.

Veamos el análisis de las gráficas. En la Figura 2.2 se observa que toma el valor 0 para los primeros valores de μ, $\mu \in (0, 2)$. Luego oscila respecto al p-valor para un intervalo en el que se encuentra $\mu = c$ y al encontrarnos fuera del intervalo toma valores muy próximos a 0. En conclusión, observamos que, hay un rango, los extremos, en el que el método con el muestreo elegido falla. También se observa que la conjetura de tomar como $\mu_{IS} \approx c$ es razonable.

En la Figura 2.3 se observa que el valor mínimo se tiene en un entorno de $\mu = c$ y se aprecia que la varianza no varía mucho en valores cercanos a este punto. Por ejemplo, si tomásemos $\mu = 2$, obtendríamos unos resultados bastante buenos. Esta representación tiene como objetivo ver la dispersión de los datos a medida que variamos el valor de μ. De hecho, se observa que según nos alejamos de $\mu = c$, los puntos representados están cada vez más dispersos y, en consecuencia, podemos considerar como intervalo óptimo de los posibles valores de μ, donde se tiene mínima varianza, en un entorno de $\mu = c$ ya que los datos están más próximos entre sí resultando así tener una aproximación más precisa.

Observación 2.7 *Se pueden concluir los siguientes resultados:*

- *En base a los resultados obtenidos, se deduce que a partir de un muestreo de pocos datos, se pueden calcular p-valores considerablemente pequeños.*
- *El método de muestreo por importancia tiene como ventaja que al muestrear solo la región que nos interesa, obtenemos una estimación más precisa y una menor dispersión de los datos observados.*
- *La suposición de que μ_{IS} se encuentra en un entorno de c es cierta, se ha demostrado analíticamente y se ha apoyado de forma gráfica.*
- *El método de muestreo por importancia es muy efectivo cuando los datos observados siguen una distribución normal.*

Capítulo 3

Distribución exponencial

En el capítulo anterior hemos probado que en el caso de la normal nuestro método funciona y si se parece a ésta no nos sorprende que funcione. Y, en consecuencia, nuestro interés se encuentra en estudiar casos que se alejen de la normal. Consideraremos otro caso que es suficientemente sencillo para tener solución analítica del muestreo por importancia, cálculo del p-valor de forma analítica y, además, se trata de un ejemplo suficientemente diferente de una normal, siendo una distribución muy asimétrica.

Vamos a generalizar los resultados en la mayor medida posible por lo que partimos de las siguientes hipótesis. Sean X una variable aleatoria independiente e igualmente distribuida que sigue una distribución exponencial y $n \in \mathbb{N}$, la dimensión de la variable aleatoria. Recordamos la función de densidad de la distribución exponencial: $\rho(\vec{x};\alpha) = \prod_{i=1}^{n} \alpha e^{-\alpha x_i}$, $\alpha > 0$. En este caso definimos nuestro contraste de hipótesis como:

$$H_0 : (\alpha = 1) : e^{-x}$$
$$H_1 : (\alpha \neq 1) : \alpha e^{-\alpha x}$$

Si quisiéramos ver la similitud con la hipótesis nula que se suele considerar: $H_0 : \mu = 0$, bastaría hacer un cambio de variable: $\alpha = 1 + \mu$.

3.1. Aplicamos el método de la máxima verosimilitud

Pretendemos calcular el p-valor de x_0 considerando este un valor *"grande"* por lo que consideramos la cola superior. En este caso, bastaría con hallar $\int_{x_0}^{\infty} \rho(x;\alpha = 1)$ y, este valor es conocido, $\frac{1}{\alpha}e^{-\alpha x}$. A pesar de ello, vamos a calcularlo de otro modo y así comprobaremos que los resultados que se exponen con nuestro método están en consonancia con el resultado esperado.
Nuestra conjetura es que si no sabemos con qué muestrear vamos a considerar lo que mejor se ajusta a H_1, es decir, con el α obtenido a partir del método de la máxima verosimilitud. En otras palabras, dado x cuál es el mejor α que se ajusta a nuestros datos, para ello aplicamos dicho método. $\mathcal{L} = \prod_{i=1}^{n} \alpha e^{-\alpha x_i}$:

$$\log(\mathcal{L}) = \log\left(\prod_{i=1}^{n} \alpha e^{-\alpha x_i}\right) = \sum_{i=1}^{n} \log(\alpha e^{-\alpha x_i}) = \sum_{i=1}^{n} [\log\alpha - \alpha x_i] = n\log\alpha - \alpha \sum_{i=1}^{n} x_i:$$

Sabemos que es derivable con $\alpha > 0$, hallamos el máximo

$$0 = \frac{\partial \log(\mathcal{L})}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} x_i \Longrightarrow \frac{n}{\alpha} = \sum_{i=1}^{n} x_i \Longrightarrow \alpha = \frac{n}{\sum\limits_{i=1}^{n} x_i} = \frac{1}{\bar{x}}:$$

Vamos a asegurarnos de que el candidato obtenido es máximo.

$$\frac{\partial^2 \log(\mathcal{L})}{\partial \alpha^2} = \frac{-n}{\alpha^2} < 0.$$

Luego,

$$\alpha_{MV} = \frac{1}{\bar{x}} = \frac{n}{\sum\limits_{i=1}^{n} x_i}. \tag{3.1}$$

es un máximo de verosimilitud.
Como $\alpha_{MV} = 1 + \mu_{MV} = \frac{1}{\bar{x}} \Longrightarrow \mu_{MV} = \frac{1}{\bar{x}} - 1$.
Hallamos el cociente de verosimilitud de forma genérica:

$$\begin{aligned}
q_\alpha(\vec{x}) &= -2 \log \left(\frac{\rho(\vec{x}; \alpha = 1)}{\rho(\vec{x}; \alpha)} \right) \\
&= -2 \log \left(\frac{\prod\limits_{i=1}^{n} \rho(x_i; \alpha = 1)}{\prod\limits_{i=1}^{n} \rho(x_i; \alpha)} \right) \\
&= -2 \left[\log \left(\prod_{i=1}^{n} \rho(x_i; \alpha = 1) \right) - \log \left(\prod_{i=1}^{n} \rho(x_i; \alpha) \right) \right] \\
&= -2 \left[\sum_{i=1}^{n} \log \rho(x_i; \alpha = 1) - \sum_{i=1}^{n} \log \rho(x_i; \alpha) \right] \overset{\text{Def.1.3}}{=} \\
&= -2 \left[\sum_{i=1}^{n} \log \left(e^{-x_i} \right) - \sum_{i=1}^{n} \log \left(\alpha e^{-\alpha x_i} \right) \right] \\
&= -2 \left[\sum_{i=1}^{n} (-x_i) - \sum_{i=1}^{n} \left[\log \alpha + \log \left(e^{-\alpha x_i} \right) \right] \right] \\
&= -2 \left[-\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log \alpha + \sum_{i=1}^{n} \alpha x_i \right] \\
&= -2 \left[-\sum_{i=1}^{n} x_i - n \log \alpha + \alpha \sum_{i=1}^{n} x_i \right] \\
&= 2 \sum_{i=1}^{n} x_i + 2n \log \alpha - 2\alpha \sum_{i=1}^{n} x_i.
\end{aligned}$$

Particularizando para α_{MV}.

$$\begin{aligned} q_{\alpha_{MV}}(\vec{x}) &= 2\sum_{i=1}^{n} x_i + 2n\log\alpha_{MV} - 2\alpha_{MV}\sum_{i=1}^{n} x_i \overset{(3.1)}{=} \\ &= 2\sum_{i=1}^{n} x_i + 2n\log\left(\frac{1}{\overline{x}}\right) - 2\frac{n}{\sum_{i=1}^{n} x_i}\sum_{i=1}^{n} x_i \\ &= 2\sum_{i=1}^{n} x_i - 2n\log\left(\overline{x}\right) - 2n. \end{aligned}$$

Como hemos comentado anteriormente, nuestro interés es considerar una única medida ya que en otro caso tendería a una normal y esto se limita al caso del capítulo anterior.
Hallamos el cociente de verosimilitud considerando solamente una muestra:

$$q_\alpha(x) = -2\log(\mathcal{L}) = -2\log\left(\frac{e^{-x}}{\alpha e^{-\alpha x}}\right) = 2x + 2\log(\alpha) - 2\alpha x.$$

Recordamos que $\alpha_{MV} = \frac{1}{\overline{x}}$ y, como estamos considerando únicamente una muestra entonces $\alpha_{MV} = \frac{1}{x}$.
Veamos numéricamente que ocurre aplicando el programa R en la distribución exponencial. Se puede proceder de dos formas, teniendo en cuenta que:

$$q_0 = 2x + 2\log\left(\frac{1}{x}\right) - 2\frac{1}{x}x = 2x - 2\log(x) - 2.$$

Si se trabaja analíticamente sobre q_0, la resolución lleva al uso de funciones de Lambert, que se explican en el anexo (A.2). De forma alternativa, se puede trabajar directamente con x, si separamos la cola superior y la inferior. Haremos aquí el desarrollo para la cola superior.

3.2. Cálculo del p-valor por muestreo por importancia

Vamos a muestrear por importancia con funciones de la forma: $\alpha e^{-\alpha x}$, $\alpha > 0$. Encontramos el valor del μ óptimo por entropía cruzada:

$$\begin{aligned} \mathcal{D}_{CE} &= \int \rho(x;1)\log\left(\frac{\rho(x;1)}{\rho(x;\alpha)}\right)\,dx = \int_{x>x_0} e^{-x}\log\left(\frac{e^{-x}}{\alpha e^{-\alpha x}}\right)\,dx \\ &= \int_{x>x_0} e^{-x}\left(\frac{q_\alpha(x)}{-2}\right)\,dx = -\frac{1}{2}\int_{x_0}^{\infty} e^{-x}\left(2x - 2\log\alpha - 2\alpha x\right)\,dx \end{aligned}$$

Se tiene que la expresión obtenida al aplicar la entropía cruzada es continua y derivable. Queremos maximizar respecto a $\alpha : 0 = \frac{\partial \mathcal{D}_{CE}}{\partial \alpha}$

$$0 = \frac{\partial D_{CE}}{\partial \alpha} \Longrightarrow 0 = -\frac{1}{2}\int_{x_0}^{\infty} e^{-x}\left(\frac{2}{\alpha} - 2x\right) dx$$

$$\Longrightarrow 0 = \int_{x_0}^{\infty} e^{-x}\left(2x - \frac{2}{\alpha}\right) dx = \begin{bmatrix} u = 2x \Rightarrow du = 2 \\ dv = e^{-x} \Rightarrow v = -e^{-x} \end{bmatrix}$$

$$= -2xe^{-x}\Big|_{x_0}^{\infty} + 2\int_{x_0}^{\infty} e^{-x}\, dx + \frac{2}{\alpha}e^{-x}\Big|_{x_0}^{\infty}$$

$$= 2x_0e^{-x_0} + 2e^{-x_0} - \frac{2}{\alpha}e^{-x_0} \Longrightarrow 0 = 2x_0e^{-x_0}\alpha + 2e^{-x_0}\alpha - 2e^{-x_0}$$

$$\Longrightarrow \alpha\left(2x_0e^{-x_0} + 2e^{-x_0}\right) = 2e^{-x_0} \Longrightarrow \alpha_{IS} = \frac{2e^{-x_0}}{2e^{-x_0}\left(x_0+1\right)} = \frac{1}{x_0+1} \approx \frac{1}{x_0}.$$

La última aproximación se tiene porque estamos estudiando p-valores pequeños, es decir, $x_0 \longrightarrow \infty$ (si consideramos la cola superior).
A partir de $\alpha_{IS} \approx \frac{1}{x_0}$ podemos decir que nuestra conjetura es cierta.

3.3. Cálculo por muestreo

Se pretende representar al igual que se hizo con la distribución normal el p-valor experimental y el logaritmo en base 10 de la varianza respecto de μ. Consideramos el logaritmo ya que la variación es de órdenes de magnitud. Se han hecho diferentes simulaciones variando los parámetros (A.4.2). Además, también se comenta en el apéndice el código seguido (A.3.2). Las simulaciones consisten en hacer un barrido en μ según la función del muestreo por importancia, en cada caso muestreamos N, calculamos el p-valor y la varianza de la estimación y, finalmente, representamos. A continuación, se muestra un ejemplo gráfico.

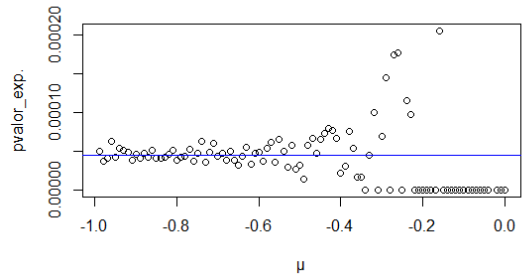

Figura 3.1: P-valor en función de μ.

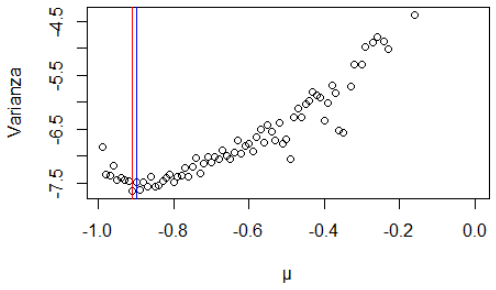

Figura 3.2: Variación del logaritmo decimal de la varianza en función de μ.

En el análisis numérico, se examinan todas las representaciones obtenidas para garantizar una reducción significativa de los errores típicos que surgen al considerar parámetros fijos.

En las ilustraciones mostradas se han tomado como parámetros: x_0=10 y 1000 como tamaño de la muestra.

En la Figura 3.1 se ha representado el p-valor experimental respecto de μ y como se hizo con la distribución normal, mostraremos con una recta azul el p-valor teórico que utilizaremos como valor de referencia para comparar el p-valor obtenido a partir del método de muestreo.

En la Figura 3.2, se representa el logaritmo decimal de la varianza respecto de μ. Además, se representa en color azul el valor obtenido a partir del método de la máxima verosimilitud, $\mu_{MV} = -1 + \frac{1}{x_0}$ y, en rojo, el obtenido por la entropía cruzada, $\mu_{IS} = -1 + \frac{1}{x_0+1}$. Podemos observar en la Figura 3.2 cómo estos valores coinciden con el entorno del valor mínimo del logaritmo de la varianza. Además, vemos que la varianza mejora en 7.5 órdenes de magnitud.

Como se realizó en la normal, se consideran distintos valores para x_0 y así poder obtener conclusiones más generalizadas. Entendemos que presentar únicamente una representación gráfica sin respaldarla con un análisis detallado no es suficiente como demostración. Sin embargo, en este caso, ya hemos demostrado el resultado de manera analítica. La representación numérica, en este contexto, tiene como objetivo complementar la demostración previamente realizada en el trabajo.

En cuanto al análisis de las gráficas. En primer lugar, como se trata de la distribución exponencial y $\alpha > 0$, como hemos considerado $\alpha = 1 + \mu$, se tiene que $\mu > -1$.

Para la distribución exponencial al igual que para la normal, las gráficas apoyan las demostraciones analíticas correspondientes a cada una de las distribuciones. Sin embargo, serán de gran utilidad las gráficas para aquellos casos donde el desarrollo analítico no sea tan sencillo como, por ejemplo, es el caso de los histogramas que se estudia a continuación.

3.4. Conclusiones

Vemos que la conjetura funciona para la cola superior y, además, supone una mejora sustancial. Sin embargo, como en el caso de la distribución normal, se ha visto que cuando se estudian las dos colas, el muestreo con este tipo de funciones no mejora la solución sin peso. En este caso, además, las colas son asimétricas.

Capítulo 4

Distribución de Poisson

Otro ejemplo típico que además nos sirve para introducir el problema real al que queremos llegar es la distribución de Poisson. Esta distribución sigue una probabilidad discreta, que describe el número de veces que ocurre un evento durante un intervalo específico; el cual puede ser de tiempo, distancia, área, volumen, entre otros. Algunos ejemplos que siguen esta distribución: el número de coches que pasan a través de un cierto punto en una carretera; la cantidad de llamadas telefónicas en una central telefónica por minuto. De forma general, se puede entender como ejemplos típicos de esta distribución aquellos experimentos que supongan un contaje, la probabilidad no sea muy grande, no cambien en cada medida y sean independientes.

Bien es cierto que la distribución de Poisson no tiene mucho interés ya que con la herramienta de R podemos hallar el valor de p-valor que buscamos. Sin embargo, se detalla este ejemplo ya que proporciona información desde el punto de vista del análisis matemático por las propiedades que se pueden aplicar en los diferentes desarrollados expuestos a continuación debido a que se trata de una probabilidad discreta y, en este caso, la probabilidad se calcula a partir de sumatorios y no de integrales por no ser la función de densidad, en el caso que estamos estudiando, continua.

Vamos a aplicar el método de Poisson que sigue una ley de probabilidad:

$$P(n;\lambda) = \frac{\lambda^n}{n!}e^{-\lambda},$$

donde λ es el parámetro de tasa que representa la tasa promedio de eventos por unidad de tiempo y n es el número de eventos.
Además, por relacionarlo con el caso real de los histogramas, normalmente λ mide la diferencia entre la hipótesis dada por un fondo (H_0) y la hipótesis que sobre ese fondo presupone que hay una señal caracterizada por un parámetro μ. Por lo que consideramos:

$$\lambda = \begin{cases} b & \text{si } H_0 : \mu = 0, \\ \mu s + b & \text{si } H_1 : \mu \neq 0. \end{cases}$$

Nuestro objetivo es encontrar el valor de λ que mejor se ajusta a los datos observados. Para ello, buscamos el valor de λ que optimiza el método de entropía cruzada, a este valor lo denominamos en consonancia de los ejemplos anteriores, λ_{IS}.

4.1. Cálculo del p-valor por muestreo por importancia

El objetivo de aplicar este método es obtener el valor óptimo, λ, que se ajusta a los datos. Consideramos n_0 como el valor correspondiente a x_0 en los desarrollos de los capítulos precedentes, es decir, el valor fijado de referencia sobre el que queremos calcular el p-valor que equivaldría a la observación. Basándonos en el método de la entropía cruzada y tomando $\lambda = \mu s + b$ para simplificar los cálculos.

$$\begin{aligned}
\mathcal{D}_{CE} &= \int \rho(x;\mu=0)\log\rho(x;\mu)\,dx \overset{\text{Distrib. discreta (Prob. Poisson)}}{=} \\
&= \sum_{n=n_0}^{\infty} P(n;\mu=0)\log(P(n;\mu)) \\
&= \sum_{n=n_0}^{\infty} \frac{b^n}{n!}e^{-b}\left[\log\left(\frac{\lambda^n}{n!}e^{-\lambda}\right)\right] \\
&= \sum_{n=n_0}^{\infty} \frac{b^n}{n!}e^{-b}\left[n\log(\lambda)+\log(e^{-\lambda})-\log(n!)\right] \\
&= \sum_{n=n_0}^{\infty} \frac{b^n}{n!}e^{-b}\left[n\log(\lambda)-\lambda-\log(n!)\right] \\
&= \log(\lambda)\sum_{n=n_0}^{\infty}\frac{b^n e^{-b}}{n!}n - \lambda\sum_{n=n_0}^{\infty}\frac{b^n e^{-b}}{n!} - \log(n!)\sum_{n=n_0}^{\infty}\frac{b^n e^{-b}}{n!}.
\end{aligned}$$

Se tiene que el último término es constante y para facilitar el desarrollo y que no sea tan engorroso, denotamos:

$$\alpha = \sum_{n=n_0}^{\infty}\frac{b^n e^{-b}}{n!}n.$$
$$\beta = \sum_{n=n_0}^{\infty}\frac{b^n e^{-b}}{n!}.$$

Es fácil ver que α y β no dependen de λ. Teniendo en cuenta la notación precedente se tiene:

$$\mathcal{D}_{CE} = \alpha\log(\lambda) - \lambda\beta + cte.$$

Hallamos el máximo.

$$\frac{\partial\mathcal{D}_{CE}}{\partial\lambda} = 0 \Longrightarrow \frac{\alpha}{\lambda} - \beta = 0 \Longrightarrow \lambda_{IS} = \frac{\alpha}{\beta}.$$

Comprobamos que se trata de un máximo.

$\frac{\partial^2\log(\mathcal{D}_{CE})}{\partial\lambda^2} = \frac{-\alpha}{\lambda^2} < 0$, porque $\alpha > 0$ y $\lambda^2 > 0$.

Luego, $\lambda_{IS} = \frac{\alpha}{\beta}$, es un máximo.

Continuamos con nuestros cálculos, partimos de α y desarrollamos para dejarlo en función de β,

$$\alpha = \sum_{n=n_0}^{\infty} \frac{b^n e^{-b}}{n!} n = \sum_{n=n_0}^{\infty} \frac{b^{n-1} e^{-b} b}{(n-1)!} = b \sum_{n=n_0}^{\infty} \frac{b^{n-1} e^{-b} b}{(n-1)!} = b \left(\frac{b^{n_0-1} e^{-b}}{(n_0-1)!} + \sum_{n=n_0}^{\infty} \frac{b^n e^{-b}}{n!} \right) = \frac{b^{n_0} e^{-b}}{(n_0-1)!} + b\beta.$$

Teniendo en cuenta lo anterior, se tiene que:

$$\lambda_{IS} = \frac{\alpha}{\beta} = \frac{\frac{b^{n_0} e^{-b}}{(n_0-1)!} + b\beta}{\beta} = \frac{b^{n_0} e^{-b}}{(n_0-1)!\beta} + b = \frac{n_0 b^{n_0} e^{-b}}{n_0!\beta} + b \leq n_0 + b.$$

Por otro lado,

$$\alpha = \sum_{n=n_0}^{\infty} \frac{b^n e^{-b}}{n!} n > n_0 \sum_{n=n_0}^{\infty} \frac{b^n e^{-b}}{n!} = n_0\beta \Longrightarrow \alpha \geq n_0\beta \Longrightarrow \frac{\alpha}{\beta} \geq n_0 \Longrightarrow \lambda_{IS} \geq n_0.$$

Luego, tenemos que:

$$\left.\begin{array}{c} \lambda_{IS} \geq n_0 \\ \wedge \\ \lambda_{IS} \leq n_0 + b \end{array}\right\} \Longrightarrow \lambda_{IS} \in [n_0, n_0 + b] \Longrightarrow \lambda_{IS} \approx n_0.$$

La justificación de la implicación anterior se tiene porque estamos considerando una probabilidad muy pequeña y, en consecuencia, n_0 considerablemente grande.

La conjetura que estamos tratando de probar es aplicar el muestreo por importancia para calcular p-valores considerando como función para pesar una expresión de la familia de la distribución que siguen los datos observados.
Nuestra hipótesis es considerar como hipótesis alternativa que los datos observados siguen como función para pesar en el muestreo por importancia la que mejor se ajusta a los datos. En este caso, n_0, debido a que solamente tenemos un dato, el valor que mejor se ajusta a la media es exactamente el valor observado. En conclusión, cuando n_0 es *"muy grande"*, es compatible con nuestra conjetura.

Por tanto, se puede concluir que, en este caso, también funciona y mejoraría extremadamente el resultado por muestreo.

Capítulo 5

Histogramas

En primer lugar, veamos de qué forma vamos a interpretar y, por tanto, a estudiar los histogramas. En este contexto, se trata de presentar una forma compacta y sencilla de representar una muestra, potencialmente grande. Para ello, se mide una cantidad, generalmente continua con su ley de probabilidad y se cuentan cuántos casos se dan para distintos rangos de esta variable. Esto es lo que consideraremos como el vector asociado a los canales. Es un vector de m coordenadas donde la i-ésima coordenada del vector corresponde con el i-ésimo canal del histograma, $\forall i \in \{1, ..., m\}$. El objetivo de esta discretización es reducir considerablemente la dimensión de los datos observados y simplificar el problema que nos aborda.

Si m es estrictamente grande, cada uno de los datos de este nuevo vector sigue una distribución de Poisson independiente. Esta ley se ha estudiado en el capítulo anterior y, como se comentó, se trata de un caso particular del histograma en el que solamente se tiene un canal.
De nuevo, considerando que m es grande, podemos considerar que los canales son independientes entre sí por lo que para hallar la probabilidad del histograma consideraremos el producto de las probabilidades de los distintos canales. Es de este modo, como se trabaja en el campo de la física de partículas.

Particularizando al problema de señal ruido como en el caso de Poisson, la variable, n_i, discretizada para el i-ésimo canal sigue la distribución definida a continuación,

$$n_i \sim P(\lambda_i) = \frac{\lambda_i^{n_i}}{n_i!} e^{-\lambda_i}; \quad \lambda_i = \begin{cases} b_i & \text{para } H_0 \\ b_i + \mu s_i & \text{para } H_1 \end{cases}$$

siendo μ común a todos los canales, b_i y s_i los valores esperados de fondo y señal, prefijados por el modelo en cada caso. Por construcción, $b_i > 0$ y $s_i > 0$, $\forall i \in \{1, ..., m\}$.

5.1. Aplicamos el método de la máxima verosimilitud

Calculamos el estadístico que mide cómo de distintas son la verosimilitud de H_0 y la verosimilitud de la hipótesis alternativa. Nuestro interés es encontrar un caso que sea muy poco probable, que sean muy distintos los datos observados respecto de la hipótesis nula. Calculamos el máximo de verosimilitud asociada a la hipótesis alternativa. Para ello, primero

desarrollamos el logaritmo de la verosimilitud de H_1

$$\begin{aligned}\log(\mathcal{L}) = \log(\rho(n;\lambda_{H_1})) &= \log\left(\prod_{i=1}^{m}\rho(n;\lambda_{H_1})\right)\\ &= \sum_{i=1}^{m}\log\rho(n;\lambda_{H_1}(i)) = \sum_{i=1}^{m}\log\left(\frac{(b_i+\mu s_i)^{n_i}}{n_i!}e^{-(b_i+\mu s_i)}\right)\\ &= \sum_{i=1}^{m}\Big[\log((b_i+\mu s_i)^{n_i}) + \log(e^{-(b_i+\mu s_i)}) - \log(n_i!)\Big]\\ &= \sum_{i=1}^{m}[n_i\log(b_i+\mu s_i) - b_i - \mu s_i - \log(n_i!)]\end{aligned}$$

Dado que el argumento del logaritmo es positivo, y teniendo en cuenta que la expresión anterior es continua y derivable, para hallar el máximo de verosimilitud, $0 = \frac{\partial\log(\mathcal{L})}{\partial\mu}$

$$0 = \frac{\log(\partial\mathcal{L})}{\partial\mu} \Longrightarrow 0 = \sum_{i=1}^{m}\left[\frac{n_i s_i}{b_i+\mu s_i} - s_i\right] \Longrightarrow \sum_{i=1}^{m}\frac{n_i s_i}{b_i+\mu s_i} = \sum_{i=1}^{m} s_i.$$

La expresión anterior no se puede resolver analíticamente; se podría resolver con aproximaciones para μ pequeño considerando el desarrollo de Taylor de primer orden de $\frac{1}{1+y} \approx 1-y$, pero esto no tiene un gran interés. Lo vamos a abordar de forma numérica.
Comprobamos que se trata de un máximo.

$$\frac{\partial^2\log(\mathcal{L})}{\partial\mu^2} = \sum_{i=1}^{m}\frac{-n_i s_i^2}{(b_i+\mu s_i)^2} < 0.$$

Luego, es un máximo de verosimilitud.

5.2. Cálculo por muestreo

Vamos a estudiar numéricamente con R la aplicación al caso de los histogramas.
En primer lugar, definimos el modelo, como hemos comentado, suponemos que los canales son independientes entre sí. Recordemos que en el experimento se consideran los parámetros s_i y b_i constantes y se toma el mismo μ para todas las entradas del histograma.

Sabemos que si b tendiese a infinito, la distribución de cada canal tendería a la distribución normal y la combinación de los canales independientes seguirá una distribución normal también. En consecuencia, cumpliría el teorema de Wilks. Nuestro interés es probar que este método funciona y también en los casos en los que Wilks falla (entendiendo que no se cumple cuando no se verifican las hipótesis del teorema, concretamente cuando el parámetro está acotado o el tamaño de la muestra no es suficientemente grande para aplicar las aproximaciones asintóticas).

Aunque en principio μ puede tomar cualquier valor, dependiendo de si se trata de una señal por exceso o por defecto respecto a H_0. Nosotros vamos a restringirnos al caso $\mu \geq 0$. Esto nos aleja de las hipótesis del teorema de Wilks (Teorema 1.13). Sin embargo, no supone ningún inconveniente ya que podemos definir nuestro método bajo las condiciones que nos interese.

La limitación de estudiar simultáneamente las dos colas en la normal tiene cierta similitud con estudiar aquellos casos en los que la señal no sobrepasa la hipótesis nula, se mantiene a la baja o, por el contrario, a la alza por lo que solo consideraremos la posible presencia de una señal positiva, esta situación engloba muchas situaciones reales.
En nuestro problema, como hemos comentado, nos vamos a centrar en excesos por arriba, es decir, fluctuaciones a la alza, $\mu \geq 0$.

Para asegurarnos de que μ siempre toma un valor positivo o nulo, configuramos el código que ajusta un modelo utilizando la estimación de máxima verosimilitud y restringiendo μ a valores no negativos. Esto se puede ver en el código adjunto en el apéndice (A.3.3). En consecuencia, al prescindir de la cola inferior de la probabilidad, como hemos visto anteriormente en la documentación de este trabajo para la distribución de Gauss, siendo esta una distribución simétrica, es necesario dividir la probabilidad obtenida por un factor ya que hay casos que no estamos considerando y tienen cierta probabilidad.

Hemos programado dos ejemplos en R donde el segundo de ellos se trata de un caso límite. Para el primer ejemplo, se ha realizado una simulación para calcular el p-valor. Cada simulación supone muestrear un vector aleatorio, donde cada una de sus componentes sigue una ley de Poisson. Suponemos el fondo como una recta de pendiente 1 y la señal, plana concentrada en 10 canales. Hemos considerado $q_0 = 16$, $m = 50$ y tamaño de la muestra: 1000.

En cada repetición, generamos los datos considerando la función de distribución de Poisson definida por el fondo y la señal determinados y, además, ajustamos el valor de μ óptimo según la aplicación numérica $mle2$ de R del método de máxima verosimilitud descrito anteriormente. [1]
Seguidamente, calculamos el p-valor, en este caso, para $q_0 = 16$. Para ello, aplicamos el muestreo por importancia para distintos valores de μ. Para cada μ, se hace el muestreo para las distintas funciones $\rho(\vec{x}, \mu)$, correspondientes a la probabilidad de H_1 para cada μ prefijado. Para cada uno, obtenemos 1000 muestras aleatorias según ese μ, ajustamos y calculamos el valor de μ de máxima verosimilitud, el estadístico, q, y hallamos el peso, W. Con esto, calculamos el p-valor

$$\mathfrak{p} = \frac{\sum_{i=1,\, q_i > q_0}^{n_{rep}} W_i}{n_{rep}},$$

donde n_{rep} indica el número de repeticiones y, además, calculamos su varianza,

$$Var(\mathfrak{p}) = \frac{\sum_{i=1,\, q_i > q_0}^{n_{rep}} W_i^2}{n_{rep}^2}.$$

Los resultados se muestran en la Figura 5.1 y Figura 5.2.

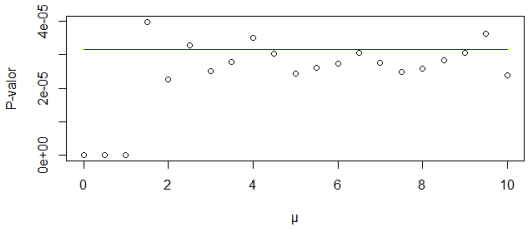

Figura 5.1: P-valor estimado en función de μ y Wilks.

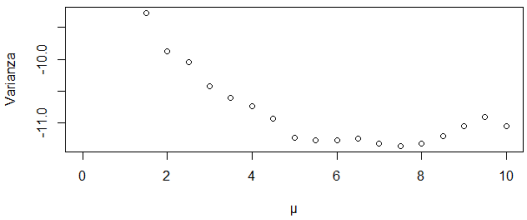

Figura 5.2: Variación del logaritmo decimal de la varianza en función de μ.

En la Figura 5.1 se compara el p-valor obtenido con la estimación corregida de Wilks, como en el apartado anterior, corrigiendo por un factor dos por estar considerando la cola superior. Vemos que el valor obtenido se asemeja al esperado a Wilks, aunque quizá un poco más bajo, para un gran rango de valores de μ y, que la varianza se hace mínima para valores de μ comprendidos entre 6 y 8, aproximadamente, se muestra en la Figura 5.2. Además, se observa que la varianza mejora en 11 órdenes de magnitud respecto a no pesar. Es importante mencionar que somos capaces de hacer unas estimaciones razonablemente precisas de p-valores que son del orden de 10^{-5}, con solo 1000 muestras.

Para comparar con nuestra conjetura, mostramos la Figura 5.3 donde se muestra el valor promedio de q, en función del valor de μ de la muestra generada.

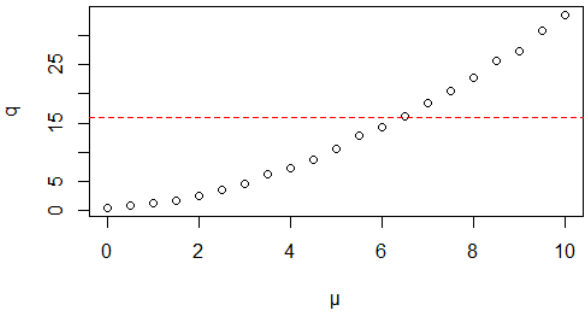

Figura 5.3: Representamos el valor promedio de q respecto de μ

De la Figura 5.3, obtenemos que $q_0 = 16$, correspondería a un μ_0 en torno a 6.5, que encaja perfectamente con la observación del mínimo en la Figura 5.2.

A continuación se expone un segundo ejemplo, donde se considera $m = 2$, con valores bajos en cada canal, del orden de 10 y, 1000 muestras. Veamos que bajo estas hipótesis no se cumple el asintótico. Es cierto que se trata de un caso poco realista pero el interés está en ver que Wilks falla mientras que el método de muestreo por importancia da resultados muy precisos.

Siguiendo el mismo razonamiento que en el ejemplo anterior, representamos en la Figura 5.4 el p-valor obtenido. En azul, se muestra el p-valor con la estimación corregida de

Wilks, mientras que en rojo se indica el p-valor obtenido a partir de una simulación sin pesos con un millón de muestras.

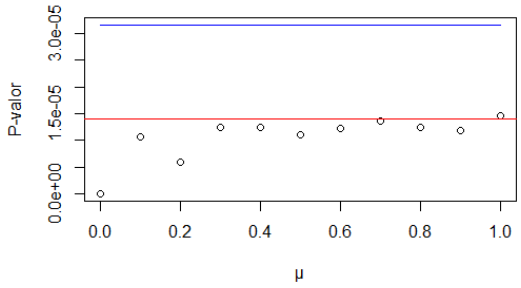

Figura 5.4: P-valor estimado en función de μ y Wilks.

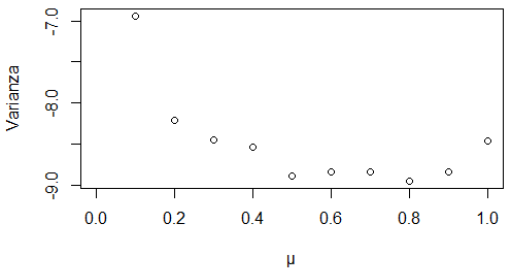

Figura 5.5: Variación del logaritmo decimal de la varianza en función de μ.

En la Figura 5.4 se observa claramente que los resultados obtenidos con el método de muestreo por importancia concuerdan con el p-valor resultante de la simulación sin pesos, pero no con lo esperado según la corrección de Wilks. A partir de este ejemplo, parece que nuestro método puede proporcionar resultados eficientes incluso en situaciones donde el método de Wilks no lo hace. Sin embargo, se requiere de más investigación y pruebas adicionales para confirmar que esta observación se aplica de manera general

Para comparar con nuestra conjetura, mostramos la Figura 5.6, donde se muestra el valor promedio de q, en función del valor de μ de la muestra generada.

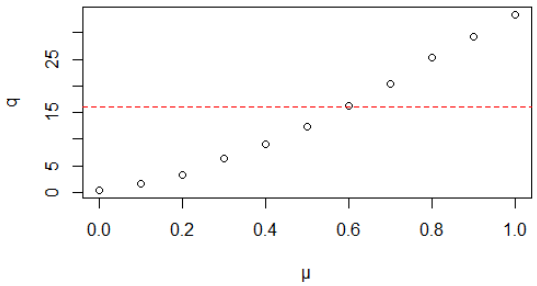

Figura 5.6: Representamos el valor promedio de q respecto de μ

A partir de la Figura 5.6 , obtenemos que $q_0 = 16$, correspondería a un $\mu_0 \approx 0{,}6$, que está en consonancia con el mínimo de la Figura 5.5. La varianza mejora en 9 órdenes de magnitud en comparación a no pesar.

En conclusión, con 1000 muestras hemos obtenido valores más precisos que empleando un millón sin pesos. Y, además, se ha probado que nuestro método es eficiente también en los casos en los que Wilks no es eficaz.

Capítulo 6

Aproximación al caso general

Recordamos que nuestro objetivo es encontrar un método para seleccionar $g(x)$ para aplicar el método de muestreo por importancia. La conjetura es que $g(x)$ sea de la familia de H_1, en particular, aquel que mejor ajusta a los datos observados sobre los que queremos estimar su p-valor.

Hemos analizado bajo qué hipótesis podríamos considerar una de las funciones de la familia H_1 como una función de ponderación que mejore los resultados obtenidos. Para ello, ha sido necesario suponer algunas aproximaciones por lo que no podemos considerar el resultado expuesto como una demostración para el caso general. Sin embargo, sí que es válida para un caso bastante general. Y, como se muestra a continuación, tenemos argumentos que nos hacen creer que se puede extender a más casos.

Suponemos las siguientes hipótesis, sea $\vec{x} \in \mathbb{R}^n$ una variable aleatoria, donde cada componente son variables independientes e igualmente distribuidas que siguen la misma distribución, $\hat{\rho}(x;\mu)$. Entonces, $\rho(\vec{x};\mu) = \prod_{i=1}^{n} \hat{\rho}(x_i;\mu)$.
Definimos nuestro contraste de hipótesis del siguiente modo:

$$H_0 : \rho(\vec{x};\mu = 0),$$
$$H_1 : \rho(\vec{x};\mu > 0),$$

donde μ es el parámetro que ajustaremos de los datos observados. En general, $H_1 : \mu \neq 0$ aunque, al igual que en el caso anterior, hemos comprobado que nuestro método funciona si consideramos la cola superior. Por lo tanto, tomaremos como hipótesis alternativa: $\mu > 0$.

Vamos a abordar el problema de un único parámetro para simplificar los cálculos.
La idea que seguimos para la demostración es:

1. Queremos encontrar el mejor μ que se ajusta a nuestra muestra, μ_{MV}, que es el que maximiza la verosimilitud, es decir, el resultado del ajuste.

2. Estudiar el estadístico que compara H_1 (tomando $\mu = \mu_{MV}$) y H_0 ($\mu = 0$)

Calculamos el estadístico:

$$q(\vec{x}) = -2\log\left(\frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu=\mu_{MV})}\right) = -2\log\left(\frac{\prod_{i=1}^{n}\hat{\rho}(x_i;\mu=0)}{\prod_{i=1}^{n}\hat{\rho}(x_i;\mu=\mu_{MV})}\right) = -2\sum_{i=1}^{n}\log\left(\frac{\hat{\rho}(x_i;0)}{\hat{\rho}(x_i;\mu_{MV})}\right) \tag{6.1}$$

Denotamos por q_0 al estimador que corresponde a la muestra de referencia que llamaremos $\vec{x_0}$, es decir, a nuestros datos $q_0 = q(\vec{x_0})$, los cuales tendrán un $\mu_{MV} = \mu_0$.
Queremos calcular el *p-valor* de q_0, hallar la probabilidad de que q sea mayor que q_0.

Podemos plantear nuestro problema del cálculo del *p-valor* como calcular el valor esperado del estimador $H(q(\vec{x}) - q_0)$ siendo H la función definida en capítulos anteriores (1.4), la función vale 1 para valores positivos y 0 en otro caso.

En los desarrollos que siguen, vamos a considerar la notación simplificada de forma que dx indica que integramos a todas las coordenadas de $\vec{x}$, $dx = dx_1 dx_2 ... dx_n$.

Calculamos la esperanza del estimador. Recordemos que la notación de E_μ representa la esperanza calculada sobre una densidad de probabilidad $\rho(\vec{x}, \mu)$.

$$\begin{aligned}
E_{\mu=0}[H] &= \int_{-\infty}^{\infty} H(q(\vec{x}) - q_0) \cdot \rho(\vec{x};\mu=0)\, dx \\
&= \int_{-\infty}^{\infty} H(q(\vec{x}) - q_0) \cdot \frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu)} \cdot \rho(\vec{x};\mu)\, dx \\
&= \int_{-\infty}^{\infty} H(q(\vec{x}) - q_0) \cdot W(\vec{x};\mu) \cdot \rho(\vec{x};\mu)\, dx \\
&= E_\mu[H \cdot W_\mu].
\end{aligned}$$

En este punto es donde aparece el concepto de muestreo por importancia ("*importance sampling*"). En lugar de calcular el valor esperado sobre la muestra inicial, lo hacemos sobre otra muestra auxiliar aplicando un peso con el fin de simplificar los cálculos.
Observamos que la definición del peso: $W(\vec{x};\mu) = \frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu)}$ es muy similar a la del $q(\vec{x})$, sin tener en cuenta el logaritmo ni el factor -2. Además, en el peso μ es una variable, mientras que en la definición del estadístico, μ_{MV} es una función de x.

De aquí, la conjetura que queremos probar es que, tomando $\mu = \mu_0$, el resultado es óptimo o al menos claramente mejor que no pesar.

Según el libro: "*Simulation and the Monte Carlo Method*" [10], una técnica de encontrar un muestreo por importancia es tomar una familia de funciones de probabilidad y escoger para el muestreo la que da menor varianza en la estimación. Vamos a tomar la familia de funciones dadas por H_1, es decir, $\rho(\vec{x};\mu)$ y queremos encontrar el mejor valor de μ. Como hemos comentado, queremos probar que tomar $\mu = \mu_0$ es una buena opción.

Para optimizar, usaremos el método de la Entropía Cruzada (1.1.2) en el que se propone escoger el elemento de la familia que minimiza la entropía cruzada, tal y como hemos

explicado en el capítulo 1.

$$\begin{aligned}
\mathcal{D}_{CE} &= E_{\mu=0}\left[H(q(\vec{x})-q_0)\cdot\log\left(\frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu)}\right)\right] \\
&= \int_{-\infty}^{\infty} H(q(\vec{x})-q_0)\cdot\rho(\vec{x};\mu=0)\cdot\log\left(\frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu)}\right)dx \\
&= \int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)\cdot\log\left(\frac{\rho(\vec{x};\mu=0)\cdot\rho(\vec{x};\mu=\mu_{MV})}{\rho(\vec{x};\mu)\cdot\rho(\vec{x};\mu=\mu_{MV})}\right)dx \\
&= \int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)\cdot\log\left(\frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu=\mu_{MV})}\cdot\frac{\rho(\vec{x};\mu=\mu_{MV})}{\rho(\vec{x};\mu)}\right)dx \\
&= \int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)\cdot\left[\log\left(\frac{\rho(\vec{x};\mu=0)}{\rho(\vec{x};\mu=\mu_{MV})}\right)+\log\left(\frac{\rho(\vec{x};\mu=\mu_{MV})}{\rho(\vec{x};\mu)}\right)\right]dx \\
&\overset{(6.1)}{=} \int_{q(\vec{x})>q_0}^{\infty} \rho(\vec{x};\mu=0)\cdot\left(\frac{q(\vec{x})}{-2}+\frac{q(\vec{x};\mu)}{2}\right)dx \\
&= -\frac{1}{2}\int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)\cdot(q(\vec{x})-q(\vec{x};\mu))\,dx.
\end{aligned}$$

En la integral aparecen dos términos, el primero no depende de μ, luego no afecta a la hora de maximizar la expresión. Por lo tanto, maximizar $\mathcal{D}_{CE}$ respecto de μ es equivalente a maximizar: $\Delta_{CE} = \int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)\cdot q(\vec{x};\mu)\,dx$.

En los desarrollos que se encuentran en [2] basados en el Teorema de Wald (1.20), tenemos que: $q(\vec{x}) = \frac{(\mu-\mu_{MV})^2}{\sigma^2} + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$, siendo N el tamaño de la muestra.

De modo asintótico, si N $\longrightarrow \infty$, $\mu_{MV} \sim N(\mu = 0, \sigma)$ porque estamos muestreando de $\rho(\vec{x}, \mu = 0)$.

Por otra parte, si N es suficientemente grande y con unas condiciones de regularidad, según el teorema de Wilks, $q(\vec{x}, \mu)$ se comporta como χ^2 de un grado de libertad ya que estamos suponiendo solo un parámetro, μ.

Suponemos que $\mu_{MV} \sim N(0, \sigma)$, entonces $\mu - \mu_{MV} \sim N(\mu, \sigma)$ y, en consecuencia, $\frac{\mu-\mu_{MV}}{\sigma} \sim N\left(\frac{\mu}{\sigma}, 1\right)$. Se tiene de forma directa a partir de las propiedades de la esperanza y de la varianza.

Sea $\eta = \frac{\mu-\mu_{MV}}{\sigma} \sim N\left(\frac{\mu}{\sigma}, 1\right)$, tenemos que: $q(\vec{x}, \mu) = \eta^2$. Recordar que nos estamos centrando en valores de $\mu > 0$. Además, el objeto es calcular p-valores pequeños, lo cual implica valores de μ grandes respecto de σ. En esas condiciones y conociendo el comportamiento de una distribución normal, la probabilidad de que η sea negativa es muy pequeña, por lo que podríamos considerar $\eta = +\sqrt{q}$, despreciando los valores negativos. Y, en este caso, tendríamos:

$$\Delta_{CE} = \int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)\cdot q(\vec{x},\mu)\,dx = \int_{\eta^2>q_0} e^{-\frac{\eta^2}{2}}\cdot\eta^2\,d\eta \overset{\text{cola superior}}{\approx} \int_{\eta>\sqrt{q_0}} e^{-\frac{\eta^2}{2}}\cdot\eta^2\,d\eta.$$

La igualdad entre la primera y la segunda integral se tiene considerando el cambio de variable de $\vec{x}$ a η y utilizando las propiedades anteriores que nos indican la ley de probabilidad

de η.
Podemos ver que este problema es equivalente al desarrollado en la distribución normal, por lo que aprovechamos el resultado de (2.4), es decir, el valor óptimo coincide con el valor del límite de la integral. En nuestro caso, $\frac{\mu_0}{\sigma} = \sqrt{q_0}$, despejando, $\mu_{IS} = \sigma\sqrt{q_0}$. Si tomamos valores grandes de μ, $\sigma\sqrt{q_0} \approx \mu_0$. Por lo que queda probada la conjetura en estas condiciones.

Como hemos comentado, hay que tener en cuenta que hemos considerado las siguientes aproximaciones para justificar la conjetura.

- N $\longrightarrow \infty$.
- Hemos tomado solamente la cola superior.
- Hemos supuesto que $\frac{\mu}{\sigma} \longrightarrow \infty$.

Aunque no es una demostración universal, es una justificación válida para un caso bastante amplio.

6.1. Enfoque alternativo

Además, tenemos ciertos argumentos que nos hacen creer que se puede extender a más casos. Consideramos que $q(\vec{x};\mu)$, que es la función que maximizamos para obtener μ_{MV}, es derivable. Luego, considerando un entorno del valor óptimo y teniendo en cuenta el desarrollo de Taylor, se tiene la siguiente aproximación:

$$q(\vec{x};\mu) \approx q(\vec{x};\mu_{MV}) + \frac{1}{2}q''(\vec{x};\mu_{MV})(\mu - \mu_{MV}(\vec{x}))^2$$

Es importante considerar que μ_{MV} depende de la muestra observada, pero no de μ. Además, q y q'', donde q'' es la derivada segunda de q respecto de μ, están evaluados para un valor concreto de $\mu = \mu_{MV}$, y no dependen de μ.
Calculamos el máximo de Δ_{CE}.

$$\begin{aligned}
0 =& \frac{\partial \Delta_{CE}}{\partial \mu} \approx \int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)\frac{\partial q_\mu(\vec{x};\mu)}{\partial \mu}\,dx \\
=& \int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)q''_\mu(\vec{x};\mu_{MV})(\mu - \mu_{MV}(\vec{x}))\,dx \\
=& \int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)q''_\mu(\vec{x};\mu_{MV})\mu\,dx - \int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)q''_\mu(\vec{x};\mu_{MV})\mu_{MV}(\vec{x})\,dx \Longrightarrow
\end{aligned}$$

$$\mu_{IS} = \frac{\int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)q''_\mu(\vec{x};\mu_{MV})\mu_{MV}(\vec{x})\,dx}{\int_{q(\vec{x})>q_0} \rho(\vec{x};\mu=0)q''_\mu(\vec{x};\mu_{MV})\,dx}.$$

Observación 6.1 *Se pueden concluir:*

- $q''(\vec{x},\mu_{MV})$ *nos indica la curvatura en el mínimo, si consideramos formas cuadráticas, la distribución normal o formas asintóticas, se tiene que* $q''(\vec{x},\mu_{MV})$ *es una constante aunque en general depende de* x*. Si no dependiese de* x*, sería un factor que multiplica a ambas integrales y, en este caso, obtendríamos* $\mu_{IS} = E[\mu_{MV}]_{q>q_0}$*, o sea, que es el promedio de los* μ_{MV} *que verifican* $q > q_0$*. Si tenemos solamente cola superior y la*

distribución de μ no es muy ancha, $\mu_0 = E[\mu_{MV}]$, evaluamos la media por el único valor que tenemos que es μ_0 y, obtendríamos que $\mu_{IS} = \mu_0$, como conjeturamos.
En cambio, si tomamos ambas colas, existirán algunos valores de μ positivos y otros negativos, por lo que la media quedaría en torno a cero, no estaríamos haciendo muestreo por importancia y desde luego no valdría la conjetura.

6.2. Resultados principales

Por lo tanto, tenemos una demostración válida para un caso bastante general y contamos con varios argumentos que sugieren su posible extensión a otros casos.

A partir de lo anterior, hemos presentado varios argumentos teóricos que, bajo ciertas condiciones, nos permiten comprobar que la aplicabilidad del método de muestreo por importancia es más eficiente que el muestreo directo. Utilizando en dicho método como función de ponderación una de la familia de H_1 y, en particular aquella que mejor se ajusta a los datos observados, se obtienen mejores resultados.

Capítulo 7

Conclusiones

En el análisis estadístico, el cálculo preciso de p-valores bajos es fundamental en la toma de decisiones basada en contrastes de hipótesis. Sin embargo, este cálculo puede ser complejo y costoso computacionalmente. Este trabajo aborda este problema mediante la implementación de un método de muestreo por importancia, que permite estimar eficientemente p-valores bajos.

Como hemos comentado, el muestreo por importancia es una técnica de simulación que mejora la eficiencia del muestreo al enfocarse en las áreas más relevantes del espacio de probabilidad. Esta técnica es particularmente útil cuando se trata de eventos raros, como es el caso de p-valores bajos.

Otro punto de relevancia en el trabajo es la selección de la función auxiliar para pesar, es decir, para realizar el muestreo por importancia. Es crucial elegir una expresión adecuada que facilite el muestreo de forma eficiente. En este documento, se propone seleccionar esta función auxiliar de entre una familia de funciones basadas en la hipótesis alternativa (H_1) del contraste de hipótesis, en particular la función que maximiza la verosimilitud.

Para validar el método propuesto, se han estudiado ejemplos sencillos tanto desde un punto de vista analítico como mediante simulaciones en R. Además, se ha utilizado el método de la entropía cruzada para buscar el valor óptimo de los parámetros del muestreo por importancia.

Se ha comprobado la validez del método propuesto aplicándolo a distribuciones comunes como la normal, exponencial y de Poisson. Los resultados indican una mejora significativa en la precisión del cálculo del p-valor, reduciendo la varianza en órdenes de magnitud en comparación con métodos tradicionales. También se ha explorado la aplicación del método a contrastes de señal sobre ruido utilizando histogramas. Mediante simulaciones, se ha comprobado la viabilidad del método en este contexto, mostrando que también puede ser aplicado eficazmente en análisis de datos basados en histogramas para distinguir señales significativas del ruido.

Además de los ejemplos específicos, se ha demostrado la validez de la propuesta en casos más generales. Esto implica que el método no solo es aplicable a situaciones estándar, sino que también puede ser adaptado y utilizado en una variedad más amplia de problemas estadísticos, ofreciendo una herramienta versátil y sólida para el cálculo de p-valores bajos.

Desde el punto de vista personal, además, he desarrollado aptitudes en la herramienta de R, probando y asegurándome de que los resultados obtenidos tenían sentido y experimentando nuevas funciones que han sido necesarias y útiles para obtener los resultados expuestos. Además, he puesto en práctica los diferentes conceptos y teoremas referentes a la estadística que he ido adquiriendo durante la carrera. Sin olvidarme de toda la parte esencial del análisis necesario para llevar a cabo los diferentes desarrollos.

Apéndice A

A.1. Distribuciones más comunes, propiedades elementales y su función de densidad.

A.1.1. Distribución normal.

La distribución normal, también conocida como la distribución gaussiana, es una de las distribuciones de probabilidad más importantes en estadística y probabilidad. Es utilizada ampliamente en diversos campos debido a sus propiedades matemáticas y su capacidad para modelar una amplia variedad de fenómenos naturales y artificiales.

Esta distribución presenta la característica de ser simétrica en torno a su media donde alcanza su máximo y expone una forma acampanada. A medida que nos alejamos de la media, la densidad de probabilidad disminuye, pero nunca alcanza cero.

La distribución normal está determinada por la media (μ) y la desviación estándar (σ). Se utiliza la notación $\mathcal{N}(\mu, \sigma^2)$, donde σ^2 se corresponde con la varianza.

Por simplicidad, en el desarrollo del trabajo, se considera, para indicar la función de distribución de la normal, la notación: $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx$

A.1.2. Distribución exponencial.

La distribución exponencial es una de las distribuciones de probabilidad continuas con mayor relevancia en estadística y probabilidad. Se utiliza comúnmente para modelar el tiempo hasta que ocurre un evento en un proceso de decaimiento aleatorio. Por ejemplo, el tiempo entre llegadas de clientes a un sistema de servicio, el tiempo entre fallos de un equipo técnico, o el tiempo que tarda un usuario en realizar una acción en un sitio web pueden seguir una distribución exponencial.

Su función de densidad es $f(x; \alpha) = \alpha e^{-\alpha x}$, donde x es la variable aleatoria y $\alpha > 0$ es el parámetro de tasa que controla la tasa media de eventos por unidad de tiempo.

La media y la varianza de la distribución exponencial están dadas por: $\frac{1}{\alpha}$ y $\frac{1}{\alpha^2}$ respectivamente.

A.1.3. Distribución de Poisson.

La distribución de Poisson es una distribución de probabilidad discreta que modela el número de eventos que ocurren en un intervalo de tiempo o en un espacio específico, cuando estos eventos ocurren de manera independiente y a una tasa promedio constante. Particularmente como procesos de conteo algunos ejemplos donde se aplica para computar son: el número de llamadas que se reciben en un período de tiempo concreto, el número de errores en un texto escrito, o incluso el número de accidentes en una intersección durante un determinado de tiempo.

Su función de probabilidad es $P(\lambda) = \frac{e^{-\lambda}\lambda^n}{n!}$, donde λ es el parámetro de tasa que representa la tasa promedio de eventos por unidad de tiempo o espacio, y n es el número de eventos.

La media y la varianza de la distribución de Poisson son iguales al parámetro de tasa, λ.

A.2. Función de Lambert

Para la elaboración de este apartado se ha consultado [6]
La función de Lambert, también conocida como función Omega o log producto. Se le denotada como W(x) y es la inversa de $f(W) = We^W$, es decir, W(x) satisface $W(x)e^{W(x)} = x$. En términos más simples, la función de Lambert nos da la solución para W en la ecuación $We^W = x$.
Propiedades de la función de Lambert:

1. **Función inversa**: W(x) es la función inversa de $f(W) = We^W$.
2. **Dominio y Rango**: El dominio de W(x) es $\left[\frac{-1}{e}, \infty\right)$ y su rango es $[-1, \infty)$.
3. **Ramificaciones**: Tiene múltiples ramificaciones $W_k(x)$, cada una con un rango específico.

Veamos con más detalle cuáles son las dos ramificaciones principales, a menudo denotadas como: $W_0(x)$ y $W_{-1}(x)$. Estas son las dos ramificaciones principales porque son las ramificaciones principales de la función exponencial $f(W) = We^W$, cuya inversa es la función de Lambert.

- $W_0(x)$: También conocida como la rama principal o rama principal superior, esta ramificación se encuentra en el rango $\left[\frac{-1}{e}, \infty\right)$ y es la solución principal cuando $x \geq -1/e$.
- $W_{-1}(x)$: También conocida como la rama secundaria o rama principal inferior, esta ramificación se encuentra en el rango $\left[\frac{-1}{e}, 0\right)$ y es la solución principal cuando $x < \frac{-1}{e}$.

Estas dos ramificaciones son las más comunes y se usan ampliamente en aplicaciones prácticas y en la resolución de ecuaciones que involucran la función de Lambert.

En nuestro trabajo hemos utilizado esta función para resolver la ecuación: $q = 2x - \log(x) - 2$. A continuación, se muestran detalladamente los pasos seguidos.

$$
\begin{aligned}
2x - \log(x) - 2 &= q \\
\log(x) &= 2x - 2 - q \\
x &= e^{2x}e^{-2-q} \\
xe^{-2x} &= e^{-2-q} \\
-2xe^{-2x} &= -2e^{-2-q} \\
-2x &= W(-2e^{-2-q}) \\
x &= \frac{W(-2e^{-2-q})}{-2}
\end{aligned}
$$

A.3. Programas R.

Para llevar a cabo los códigos programados en R, se ha consultado [9]

A.3.1. Distribución normal.

Seguidamente en la Figura A.1 se muestra el código seguido en el archivo "Comparación varianzas" el cual refleja la comparativa de aplicar o no el muestreo por importancia.

```
# Fijamos las variables.
c <- seq(1, 10, 0.1)
mu <- c
# Definimos el cociente como funcion
cociente_varianzas <- function(c, mu) {
  exp(mu^2) * pnorm(c + mu, lower.tail = FALSE) / pnorm(c,
     lower.tail = FALSE)
}
# Representamos
plot(c, log10(cociente_varianzas(c, mu)),
     xlab = "c",
     ylab = expression(log(Var[g] / Var[f])),
     type = "l",
     main = "Comparativa varianzas")
```

Figura A.1: Código de la comparativa de varianzas de aplicar o no el muestreo por importancia.

A continuación se muestra el código desarrollado en la herramienta R. Se adjunta el *script* en el archivo "Distribución normal"

```
N = 1
q0 = 25 # es decir, c=5
niter = 1000 # veces que se hace el experimento
mumax=10
mumin=0
step=0.1
mu=seq(mumin,mumax,step)
nmu=length(mu)
# Se inicializan los vectores.
W = q = numeric(niter)
t = pvalue_experimental_1=vr_1=numeric(nmu)
# Definimos las siguientes funciones.
peso = function(x,mu){dnorm(x, mean=0, sd=1, log=FALSE)/
   dnorm(x, mean=mu, sd=1, log=FALSE)}
dlog = function(x,mu){sum(dnorm(x,mu,log=T))}
# Lazo sobre el muestreo por importancia.
for (j in 1:nmu){
  # Lazo sobre las iteraciones.
  for (i in 1:niter)
  {
    x=rnorm(N, mean=mu[j], sd=1)
    W[i]=peso(x,mu[j])
    q[i]=-2*dlog(x,0)+2*dlog(x,mean(x)) # mubest=mean(x)
  }
  pvalue_experimental_1[j] = sum(W[q>q0])/niter # pvalor
     experimental.
  vr_1[j]=sum((W[q>q0])^2)/(niter^2) # varianza.
  p = pnorm(sqrt(q0),lower.tail = FALSE, log.p = FALSE) #
     pvalor teorico.
}
load("archivo1") # Para considerar los mismos datos.
plot(mu,pvalue_experimental_1,xlab="mu", ylab="pvalor_exp."
   ) # Representamos pvalor experimental resp. de mu.
segments(x0=0, x1=10, y0=p, y1=p, lwd=0.5, col="blue") #
   Representar el pvalue teorico
# Para obtener unas graficas que se puedan apreciar mejor,
   despreciamos el dato que tiene mayor altura.
indice_maximo_y=which.max(pvalue_experimental_1)
y_mod=pvalue_experimental_1[-indice_maximo_y]
x_mod=mu[-indice_maximo_y]
# Considerando los nuevos vectores, representamos:
plot(x_mod,y_mod,xlab="mu", ylab="P-valor")
segments(x0=0, x1=10, y0=p, y1=p, lwd=0.5, col="blue")
# Represetamos el logaritmo de la varianza respecto de mu:
vr_mod=vr_1[-indice_maximo_y]
plot(x_mod,log10(vr_mod), xlab="mu" ,ylab="Varianza")
segments(x0=sqrt(q0), x1=sqrt(q0), y0=-100, y1=10, lwd=0.5,
    col="blue") # Representa el minimo esperado
```

Figura A.2: Código R simulación distribución normal.

Como se ha comentado en el trabajo, una vez representadas las gráficas se ha prescindido del dato más dispar en comparación con el resto. De esta forma se obtiene una escala más ajustada y se pueden observar resultados más precisos. Esto se puede ver de forma detallada en la Figura A.2 o en los comentarios de los códigos que se adjuntan.

A.3.2. Distribución exponencial

A continuación, se muestra el código de lo que se ha programado en la herramienta R, en el archivo "Distribución exponencial"

```
N = 1  # muestra
x0=10
q0 = 2*x0-2*log(x0)-2
niter = 1000 #  veces que se hace el experimento
mumax=0.0
mumin=-0.99
step=0.01
mu=seq(mumin,mumax,step)
nmu=length(mu)
# Inicializamos los vectores.
xm = W = q = numeric(niter)
t = pvalue_exp_2=vr_2=numeric(nmu)
# Definimos las funciones.
peso = function(x,mu){dexp(x, rate = 1, log=FALSE)/dexp(x,
   rate = 1 + mu, log=FALSE)}
dlog = function(x,mu){sum(dexp(x, rate = 1 + mu,log=T))}
# Lazo sobre el muestreo por importancia.
for (j in 1:nmu){
  # Lazo sobre las iteraciones.
  for (i in 1:niter)
  {
    x=rexp(N, rate = 1+mu[j])
    W[i]=peso(x,mu[j])
    q[i]=-2*dlog(x,0)+2*dlog(x,-1+1/mean(x))
    xm[i]=mean(x)
  }
  pvalue_exp_2[j] = sum(W[xm>x0])/niter # pvalor
  vr_2[j]=sum((W[xm>x0])^2)/(niter^2) # varianza
}
load("archivo2") #cargamos los parametros obtenidos de la
   simulacion
p=pexp(x0,rate=1,lower.tail=F) # pvalor teorico.
# Representamos:
plot(mu,pvalue_exp_2,xlab="mu", ylab="pvalor_exp.")
segments(x0=-1.5, x1=0, y0=p, y1=p, lwd=0.5, col="blue")
```

```
# Vamos a prescindir del punto que se dispara respecto al resto de puntos
indice_maximo_y=which.max(pvalue_exp_2)
# Recalculamos el vector y presciendo del dato que hemos comentado:
y_mod=pvalue_exp_2[-indice_maximo_y]
x_mod=mu[-indice_maximo_y]
# Considerando los nuevos vectores, representamos:
plot(x_mod,y_mod,xlab="mu", ylab="pvalor_exp.")
segments(x0=-10, x1=10, y0=p, y1=p, lwd=0.5, col="blue") # Representar el pvalue teorico
vr_mod=vr_2[-indice_maximo_y]
plot(x_mod,log10(vr_mod), xlab="mu", ylab="Varianza")
segments(x0=-1+1/x0, x1=-1+1/x0, y0=-100, y1=10, lwd=0.5, col="blue") # Representar el minimo esperado
segments(x0=-1+1/(x0+1), x1=-1+1/(x0+1), y0=-100, y1=10, lwd=0.5, col="red")
```

Figura A.3: Código R simulación distribución exponencial.

En la Figura A.3 se observa que, de forma análoga a la distribución normal, una vez representadas las gráficas se ha prescindido del dato más dispar en comparación con el resto. Obteniendo así una escala más ajustada y se tienen unos resultados más precisos. Se puede ver de forma detallada en los comentarios del código.

A.3.3. Histogramas

Para el los histogramas se han simulado dos *scripts*.
El código desarrollado en R para un ejemplo general es el que se muestra en la Figura A.4, el script se encuentra en el archivo: "Histograma ejemplo general".

```
# Inicializamos los parametros y definimos el contraste de hipotesis precisando valores del fondo y la senal.
q0=16
nbin=50
b=c(1:nbin)+20
s=rep(0,nbin)
s[1:10]=rep(5,10)
niter=1000 # iteraciones.
# Inicializamos los vectores.
b=b/20
s=s/20
q=mubest=W=numeric(niter)
muscan=seq(0,10,0.5)
nscan=length(muscan)
pvalue=vr=qmed=qsd=numeric(nscan)
```

```
# Definimos las funciones:
lam=function(mu){b+s*mu}
peso = function(x,mu){prod(dpois(x, lam(0), log=FALSE))/
   prod(dpois(x, lam(mu), log=FALSE))}
dlog=function(n,mu){-sum(dpois(n,lam(mu),log=T))}
# Lazo sobre el muestreo por importancia.
for(j in 1:nscan)
{
  # Lazo sobre las iteraciones.
  for (i in 1:niter)
    {
     m=rpois(nbin,lam(muscan[j]))
     fit=mle2(dlog,start=list(mu=1),lower=c(mu=0),upper=c(
        mu=Inf),data=list(n=m),method="L-BFGS-B")
     mubest[i]=coef(fit)[1]
     W[i]=peso(m,muscan[j])
     q[i]=2*dlog(m,0)-2*dlog(m,mubest[i])
    }
    qmed[j]=mean(q)
    qsd[j]=sd(q)
    pvalue[j] = sum(W[q>q0])/niter
    vr[j]=sum((W[q>q0])^2)/(niter^2)
}
pwilks=pchisq(q0,df=1,lower.tail=F)
pvalue=pmin(pvalue,0.0001)
plot(muscan,pvalue, xlab="mu", ylab="P-valor")
lines(c(muscan[1],muscan[nscan]),c(pwilks,pwilks)/2, col="
   blue") # Pvalue Wilks correccion
plot(muscan,log10(vr), xlab="mu", ylab="Varianza")
plot(muscan, qmed, xlab="mu", ylab="q")
abline(h = 16, col = "red", lty = 2, lwd = 1) # Mostramos
   mu_IS
```

Figura A.4: Código R simulación caso general histogramas.

Se muestra en la Figura que sigue, A.5, el código implementado en R que refleja el caso límite donde consideramos dos canales.

```
# Inicializamos los parametros y definimos el contraste de
   hipotesis precisando valores del fondo y la senal.
q0=16
nbin=2
b=c(10,1)*1
s=c(1,10)*1
niter=1000 # iteraciones.
# Inicializamos los vectores.
b=b
s=s
```

```
    q=mubest=W=numeric(niter)
    muscan=seq(0,1,0.1)
    nscan=length(muscan)
    pvalue=vr=qmed=qsd=numeric(nscan)
    # Definimos las funciones:
    lam=function(mu){b+s*mu}

    peso = function(x,mu){prod(dpois(x, lam(0), log=FALSE))/
       prod(dpois(x, lam(mu), log=FALSE))}

    dlog=function(n,mu){-sum(dpois(n,lam(mu),log=T))}
    # Lazo sobre el muestreo por importancia.
    for(j in 1:nscan)
    {
      # Lazo sobre iteraciones.
      for (i in 1:niter)
        {
         m=rpois(nbin,lam(muscan[j]))
         fit=mle2(dlog,start=list(mu=1),lower=c(mu=0),upper=c(
            mu=100),data=list(n=m),method="L-BFGS-B")
         mubest[i]=coef(fit)[1] #mu_opt
         W[i]=peso(m,muscan[j])
         q[i]=2*dlog(m,0)-2*dlog(m,mubest[i])
        }
        qmed[j]=mean(q)
        qsd[j]=sd(q)
        pvalue[j] = sum(W[q>q0])/niter
        vr[j]=sum((W[q>q0])^2)/niter
    }
    pwilks=pchisq(q0,df=1,lower.tail=F)/2
    pvalue=pmin(pvalue,0.0001)
    plot(muscan,pvalue, xlab="mu", ylab="P-valor",ylim=c(0,max(
       pvalue,pwilks)))
    lines(c(muscan[1],muscan[nscan]),c(1,1)*pwilks, col="blue")
        # Pvalue correccion Wilks
    plot(muscan,log10(vr), xlab="mu", ylab="Varianza")
    abline(h=1.4e-5,col="red")
    plot(muscan, qmed, xlab="mu", ylab="q")
    abline(h = 16, col = "red", lty = 2, lwd = 1)
```

Figura A.5: Código R simulación caso límite histogramas.

A partir de las gráficas proporcionadas por este ejemplo, se concluye que el método de muestreo por importancia genera resultados más precisos e incluso tiene validez en aquellos casos donde Wilks falla. En el documento se explica de forma más detallada este punto.

A.4. Otros resultados numéricos

A.4.1. Distribución normal

En los ejemplos gráficos que siguen, se ha considerado como distribución, la normal y se han variando los parámetros q_0 y μ para así obtener diferentes resultados gráficos. A continuación, se muestran algunas de las representaciones gráficas obtenidas:

Considerando $q_0 = 25$ y tamaño de la muestra: 100, las gráficas obtenidas se muestran en la Figura A.6 y Figura A.7

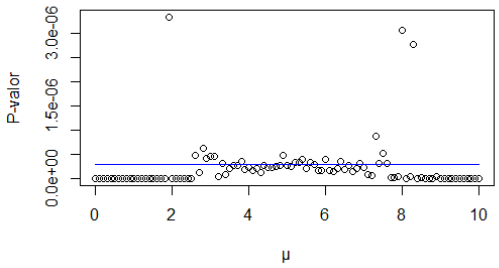

Figura A.6: P-valor en función de μ.

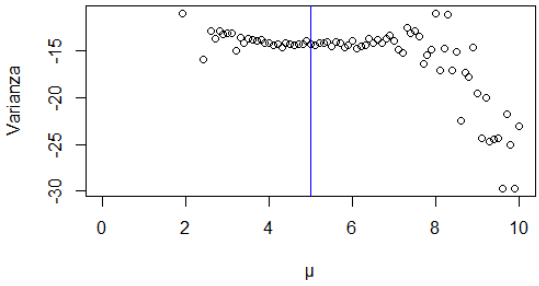

Figura A.7: Variación del logaritmo de la varianza en función de μ.

Tomando $q_0 = 25$ y número de observaciones: 1000, se obtienen la Figura A.8 y Figura A.9

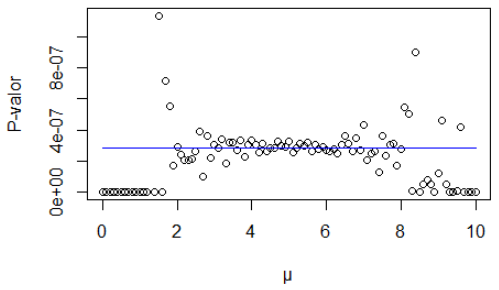

Figura A.8: P-valor en función de μ.

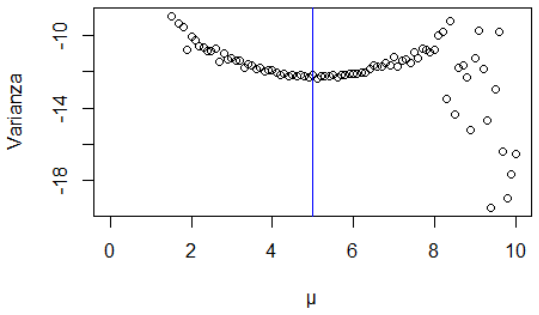

Figura A.9: Variación del logaritmo de la varianza en función de μ.

Si consideramos $q_0 = 36$ y tamaño de la muestra: 10000, los resultados gráficos obtenidos se recogen en la Figura A.10 y Figura A.11

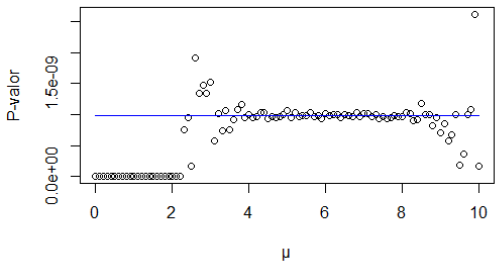

Figura A.10: P-valor en función de μ.

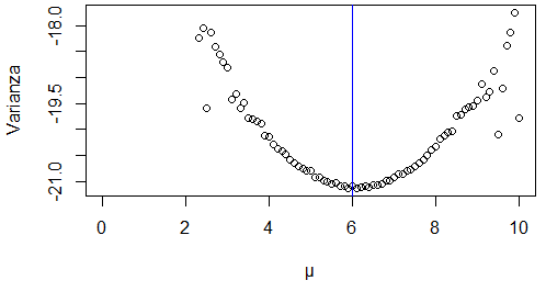

Figura A.11: Variación del logaritmo de la varianza en función de μ.

A.4.2. Distribución exponencial

En los ejemplos gráficos que se presentan a continuación, se ha utilizado la distribución exponencial, ajustando el parámetro x_0 y el tamaño de la muestra para obtener diversas representaciones gráficas. Aquí se muestran algunos de los gráficos resultantes.

En la Figura A.12 y en la Figura A.13 se muestran los resultados obtenidos al considerar $x_0 = 10$ y tamaño de la muestra 100.

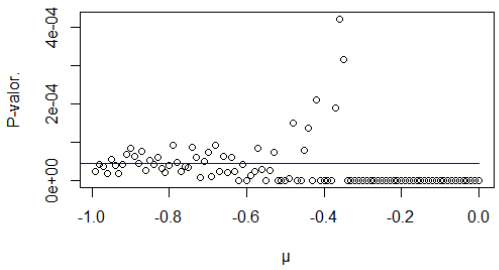

Figura A.12: P-valor en función de μ.

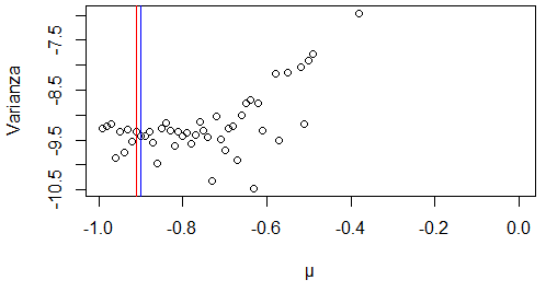

Figura A.13: Variación del logaritmo de la varianza en función de μ.

La Figura A.14 y Figura A.15 son obtenidas a partir de los parámetros $x_0 = 10$ y tamaño de la muestra: 10000.

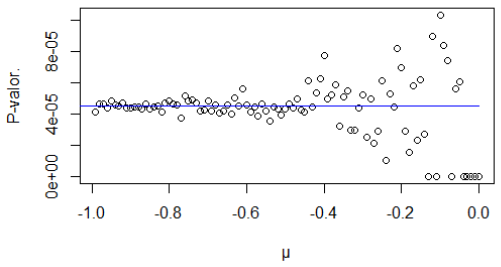

Figura A.14: P-valor en función de μ.

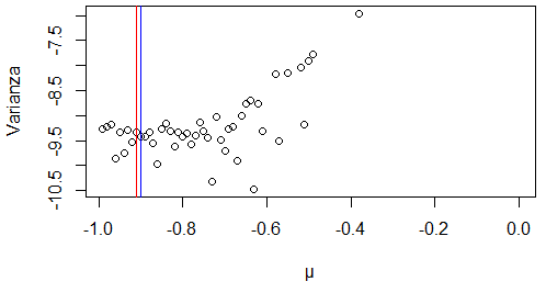

Figura A.15: Variación del logaritmo de la varianza en función de μ.

Bibliografía

[1] BOLKER. B., DEVELOPMENT CORE TEAM., *bbmle: Tools for General Maximum Likelihood Estimation*, R package version 1.0.25.1, 2023, https://CRAN.R-project.org/package=bbmle.

[2] COWAN, G., CRANMER, K., GROSS, E., VITELLS, O., *Asymptotic Formulae for Likelihood-Based Tests of New Physics*. The European Physical Journal C, 71(1554), pp.1-19. 2011. Disponible en: https://doi.org/10.48550/arXiv.1007.1727.

[3] CUESTA, J.A., *Cálculo de Probabilidades*, Universidad de Cantabria, 2021-2022.

[4] CUESTA, J.A., TUERO, A., *Estadística Básica*, Universidad de Cantabria, 2019-2020.

[5] CUESTA, J.A., TUERO, A., *Inferencia Estadística*, Universidad de Cantabria, 2019-2020.

[6] GARCÍA. SANCHEZ. F.J., ORTIZ-CONDE A., MALOBABIC. S., *Aplicaciones de la función de Lambert en electrónica.* uct [Internet]. 2006 Sep , 10(40): 235-243. Disponible en: https://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1316-48212006000400008.

[7] GEYER, C.J., *The Wilks, Wald, and Rao Tests*. Stat 8112 Lecture Notes. September 26, 2020.

[8] MILTON. A., STEGUN. I., *Handbook of mathematical functions with formulas, graphs and mathematical tables*. National Bureau of Standards (DOC), Washington, pp. 929, Dec., 1972.

[9] PERÓ, M., LEIVA, D., GUÀRDIA, J., SOLANAS, A., *Estadística Aplicada a las Ciencias Sociales mediante R y R-Commander*, Ibergarceta Publicaciones, Madrid, 2012.

[10] RUBINSTEIN, R.Y., KROESE, D.P., *Simulation and the Monte Carlo Method*, John Wiley and Sons, New Jersey, 2008.

[11] WALD. A., *Tests of Statistical hypotheses concerning several parameters when the number of observations is large*. Transactions of the American Mathematical Society, Vol. 54, No. 3, pp. 426-482, Nov., 1943.

[12] WILKS, S.S., *The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*. The Annals of Mathematical Statistics, 9(1), pp. 60-62. 1938.