Diseño del gráfico de control no paramétrico basado en la distancia de Mahalanobis para datos funcionales

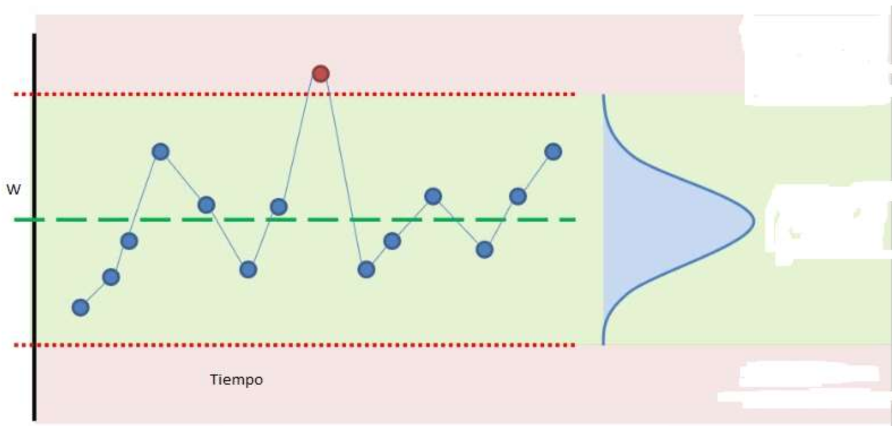
Facultad de Ciencias

Priscila Guayasamín

2022-04-12

Índice

Motivación



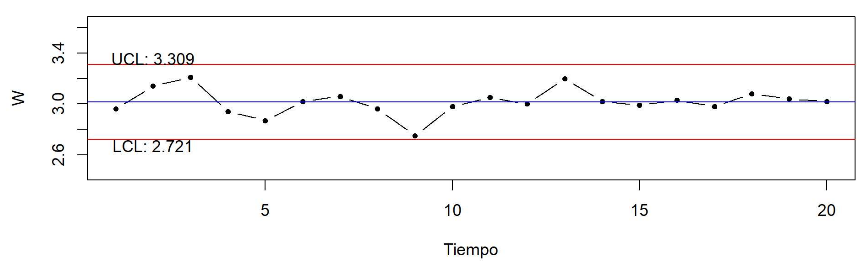
- Monitorear procesos en el tiempo.
- Extender técnicas multivariantes al análisis funcional.

3

Gráfico de control

Técnica de **monitoreo** cuyo objetivo es observar y analizar el comportamiento de un proceso a través del tiempo.

- Línea Central
- Límite de control superior e inferior



Gráficos de control: pruebas de hipótesis

Error tipo I: Concluye que el proceso no está bajo control cuando está bajo control .

Error tipo II: Concluye que el proceso está bajo control cuando no lo está.

	H_0 ***es verdadera***	H_1 ***es verdadera***
No se rechaza H_0	Decisión Correcta	Error tipo I
	$1-\alpha$	α
Se rechaza H_0	Error Tipo II	Decisión Correcta
	β	$1-\beta$

4

Medida de Profundidad

- La profundidad funcional ordena una muestra de curvas desde el centro hacia afuera.

Sean $\mathcal{X}_1, \ldots, \mathcal{X}_n$ i.i.d, realizaciones de la variable aleatoria funcional $\mathcal{X}(t)$ con dominio $T = [a, b]$ y sea D una medida de profundidad en $\mathbb{R}$. Para cada $t_0 \in T$, se considera $z_i(t_0) = D(\mathcal{X}_i(t_0))$ la profundidad univariante del dato i en t_0 con respecto a $\mathcal{X}_1(t_0), \ldots, \mathcal{X}_n(t_0)$. Entonces se define la profundidad de FM para el i-ésimo dato como:

$$FM_i = FMD(\mathcal{X}_i) = \int_a^b z_i(t)dt$$

La medidad de profundidad univariante definida por Fraiman y Muniz es la siguiente:

$$1 - \left| \frac{1}{2} - F_{n,t}(\mathcal{X}_i(t)) \right|$$

donde

$$F_{n,t} = \frac{1}{n} \sum_{k=1}^{n} 1\{\mathcal{X}_k(t) \leq \mathcal{X}_i(t)\}$$

Entonces, la profundidad funcional de Fraiman y Muniz de la curva $\mathcal{X}$ con respecto al conjunto $\mathcal{X}_1, \ldots, \mathcal{X}_n$ está dada por:

$$FMD(\mathcal{X}_i) = \int_a^b z_i^{FM1}(t)dt$$

5

Remuestreo Bootstrap

Paradigma Inferencial

Sean $\{X_1, \ldots, X_n\}$ una muestra aleatoria de una población con distribución F. Estamos interesados en hacer inferencia sobre un parámetro. Requeriremos saber la distribución de $R(X, F) = \theta(F_n) - \theta(F)$, donde F_n es la distribución empirica

Bootstrap Reemplaza la distribución poblacional F con la estimación $\hat{F}$.

Bootstrap Suavizado

$$\frac{1}{2nh} = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

donde $K_h(u) = \frac{1}{h} K(\frac{u}{h})$, K es una función kernel

$$\hat{f}_h = \frac{1}{2nh} \sum_{i=1}^{n} 1\{X_i \in (x-h, x+h)\}$$
$$= \frac{\#\{X_i \in (x-h, x+h)\}}{2nh}$$

6

T^2 Hotelling en espacios de Hilbert

Sean $\mathcal{X}_1, \ldots, \mathcal{X}_n$ una muestra de n i.i.d $\mathbb{H}$-variables aleatorias en $(\Omega, \mathcal{F}, \mathbb{P})$, tal que $\mathbb{E}[||\mathcal{X}_i||^2_{\mathbb{H}}] < +\infty$ para todo $i = 1, \ldots, n$. Sea la media m y el operador de $\mathcal{K}$ de $\mathcal{X}_i$, se define:

La ***media muestral***

$$m_n = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \mathcal{X}_i$$

El ***operador muestral*** $\mathcal{K}_n$ se define como

$$\frac{1}{n-1} \odot \bigoplus_{i=1}^{n} (\mathcal{X}_i \ominus m_n) \otimes (\mathcal{X}_i \ominus m_n)$$

m_n y $\mathcal{K}_n$ son respectivamente $\mathbb{H}$-variable aleatoria y $\mathrm{B}_{HS}(\mathbb{H})$ variable aleatoria.

Usando las notaciones de la definición anterior, el ***operador muestral de pérdida de error cuadrático medio*** $\mathcal{D}_n$ se define así:

$$\mathcal{D}_n = (m_n \ominus m) \otimes (m_n \ominus m)$$

Sean $\mathcal{X}_1, \ldots, \mathcal{X}_n$ i.i.d $\mathbb{H}$ variables aleatorias con media m y operador de covarianza $\mathcal{K}$. El operador T^2 de Hotelling se define como sigue:

$$T^2 = n \max_{f \in Im(\mathcal{K}_n) \setminus \{0\}} \frac{\langle f, \mathcal{D}_n f \rangle}{\langle f, \mathcal{K}_n f \rangle}$$

$$T^2 = n \langle m_n \ominus m, \mathcal{K}_n^{+} (m_n \ominus m) \rangle_{\mathbb{H}}$$

donde $\mathcal{K}_n$ es el operador de covarianza y $\mathcal{K}_n^{+}$ es el operador inverso generalizado $\mathcal{K}_n$ y $\mathcal{D}_n$ es el operador muestral de pérdida de error medio cuadrático.

7

Estadístico basado en la distancia de Mahalanobis a dos muestras

$$T_0^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} \max_{f \in Im(\mathcal{K}_{n_{pooled}}) \setminus \{0\}} \frac{\langle f, \mathcal{D}_{n_0} f \rangle}{\langle f, \mathcal{K}_{n_{pooled}} f \rangle}$$

Donde m_{n_1} y m_{n_2} son las medias de las muestras, $\mathcal{D}_{n_0}$ es el operador de pérdida de error medio cuadrático, bajo la hipótesis nula, i.e. $D_{n_0} = (m_{n_1} \ominus m_{n_2}) \otimes (m_{n_1} \ominus m_{n_2})$ y $\mathcal{K}_{n_{pooled}}$ el operador del muesta de covarianza definido como sigue

$$\begin{aligned} \mathcal{K}_{n_{pooled}} : & \mathbb{H} \longmapsto \mathbb{H} \\ & f \longmapsto \frac{n_1 - 1}{n_1 + n_2 - 2} \odot (\mathcal{K}_{n_1} f) \oplus \frac{n_2 - 1}{n_1 + n_2 - 2} \odot (\mathcal{K}_{n_2} f), \end{aligned}$$

donde $\mathcal{K}_{n_1}$ y $\mathcal{K}_{n_2}$ son los operadores de covarianza de las muestras. Alcanza el máximo para $f = \mathcal{K}^{+}_{n_{pooled}} (m_{n_1} \ominus m_{n_2})$.

Distancias estandarizadas para datos funcionales

$$L_s^1 = \int_T \frac{|m_n - m(t)|}{\sigma_n^2(t)} dt$$

$$L_s^2 = \int_T \frac{(m_n - m(t))^2}{\sigma_n^2(t)} dt$$

donde se define $\sigma_n^2 : T \mapsto \mathbb{R}$ como la función de varianza de la muestra puntual siendo $(n-1)\sigma_n^2(t) = \sum_{i=1}^{n} (\mathcal{X}_i(t) - m(t))^2$

8

Metodología

1. Agrupar la muestra de calibrado $\{\mathcal{X}_1(t), \ldots, \mathcal{X}_n(t)\}$ y monitoreo $\{\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_m(t)\}$, a la muestra agrupada la llamaremos $\mathcal{Z}$.
2. Calcular las profundidades $D(\mathcal{Z}_i)_{i=1}^{m+n}$; y, para determinar el límite de control superior (LCS) se sigue el procedimiento a continuación:
3. Se procede a monitorizar el proceso. Si se observa que $W(\mathcal{X}', \mathcal{Y}') \geq LCS$, entonces el proceso está fuera de control.

Bootstrap

- Obtener B remuestras bootstrap de tamaño $m+n$ de las curvas obtenidas después del recorte del $\alpha\%$ de las curvas menos profundas. Sean las muestras bootstrap $\mathcal{Z}_i^{*b}$ con $i=1,\ldots n+m$, $b=1,\ldots,B$ se obtendrán como sigue:
  - Se realiza un muestreo uniforme, con i^* de $1,\ldots,[(n+m)(1-\alpha)]$, con reemplazo.
  - Se genera V_{i^*} como un proceso gaussiano con media cero y matriz de varianza y covarianza $\gamma\Sigma_{\mathcal{Z}}$, con $\gamma\in[0,1]$. Donde $\Sigma_{\mathcal{Z}}$ es la matriz de varianzas y covarianzas de las observaciones $\mathcal{Z}_1,\ldots,\mathcal{Z}_{[(n+m)(1-\alpha)]}$.
- Finalmente se obtiene $\mathcal{Z}_i^{*b}=\mathcal{Z}_{i^*}+V_{i^*}$.
- Reagrupar $\mathcal{Z}_i^{*b}$ en las muestras $\mathcal{X}'$ y $\mathcal{Y}'$ de tamaño m y n respectivamente.

Permutaciones

- Considerar todas las posibles permutaciones de la muestra conjunta recortada $\mathcal{Z}$ y reasignar aleatoriamente las curvas a los grupos $\mathcal{X}'$ y $\mathcal{Y}'$

- Calcular el estadístico W y el LCS como el percentil de la distribución de $W(\mathcal{X}',\mathcal{Y}')$, donde W es el estadístico

10

Estudio de Simulación

Se generan realizaciones de un proceso estocástico gaussiano, con $t \in [0,1]$

$$\mathcal{X}(t) = \mu(t) + \epsilon(t)$$

donde

$$\mu(t) = \mathrm{E}(\mathcal{X}(\mathrm{t})) = 30t(1-t)^{\frac{3}{2}}$$

También $\epsilon(t)$ es un proceso con media cero y operadores de covarianza.

Escenarios

Cambios en la covarianza

Escenario A

$$\mathcal{K}g(t) = \int_0^1 e^{-2(s-t)^2} g(s)ds$$

Escenario B

$$\mathcal{K}g)(t) = \int_0^1 e^{-2(s-t)^2} (s+0.5)(t+0.5)g(s)ds$$

Cambios en la media

Escenario 1

$$\mu(t) = 30t \cdot (1-t)^{\frac{3}{2}} + \delta$$

Escenario 2

$$\mu(t) = (1-\eta)30t \cdot (1-t)^{\frac{3}{2}} + \eta \cdot 30t^{\frac{3}{2}}(1-t)$$

Resultados

Escenario 1A

Escenario 1A, n=25, m=25

Estadístico	Método	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Mahalanobis	Montecarlo	0.030	0.126	0.466	0.852	0.990	1.000	1.000	1.000	1	1	1
Mahalanobis	Bootstrap	0.042	0.136	0.482	0.864	0.990	1.000	1.000	1.000	1	1	1
Mahalanobis	Permutaciones	0.032	0.130	0.472	0.874	0.990	1.000	1.000	1.000	1	1	1
L1 std	Montecarlo	0.038	0.092	0.180	0.320	0.546	0.690	0.866	0.988	1	1	1
L1 std	Bootstrap	0.036	0.094	0.174	0.318	0.538	0.682	0.866	0.988	1	1	1
L1 std	Permutaciones	0.022	0.064	0.148	0.264	0.480	0.636	0.812	0.976	1	1	1
L2 std	Montecarlo	0.050	0.082	0.190	0.338	0.582	0.758	0.924	0.996	1	1	1
L2 std	Bootstrap	0.038	0.078	0.180	0.324	0.548	0.720	0.914	0.994	1	1	1
L2 std	Permutaciones	0.026	0.060	0.136	0.282	0.514	0.708	0.888	0.992	1	1	1

13

Escenario 1A, n=50, m=50

Estadístico	Método	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Mahalanobis	Montecarlo	0.054	0.288	0.850	0.998	1.000	1.000	1.000	1	1	1	1
Mahalanobis	Bootstrap	0.060	0.282	0.848	0.998	1.000	1.000	1.000	1	1	1	1
Mahalanobis	Permutaciones	0.052	0.252	0.836	0.996	1.000	1.000	1.000	1	1	1	1
L1 std	Montecarlo	0.052	0.110	0.248	0.520	0.810	0.940	0.980	1	1	1	1
L1 std	Bootstrap	0.052	0.110	0.250	0.538	0.814	0.940	0.980	1	1	1	1
L1 std	Permutaciones	0.052	0.110	0.238	0.520	0.774	0.934	0.978	1	1	1	1
L2 std	Montecarlo	0.052	0.116	0.288	0.594	0.874	0.972	0.994	1	1	1	1
L2 std	Bootstrap	0.052	0.100	0.268	0.582	0.866	0.964	0.994	1	1	1	1
L2 std	Permutaciones	0.052	0.092	0.238	0.546	0.812	0.960	0.994	1	1	1	1

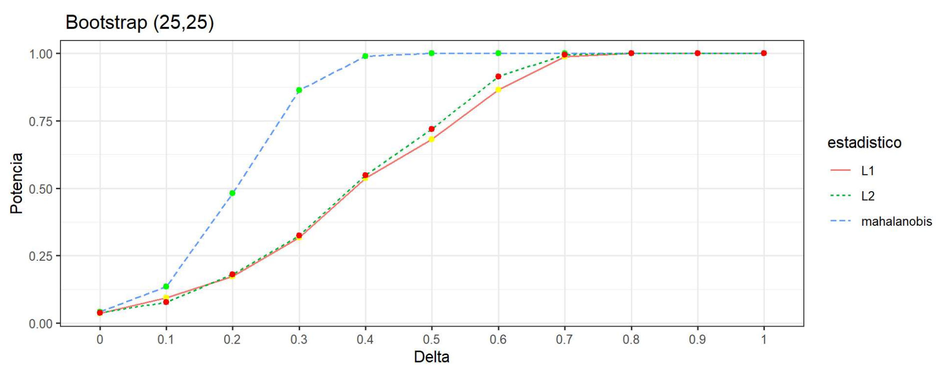
Escenario 1A, n=100, m=100

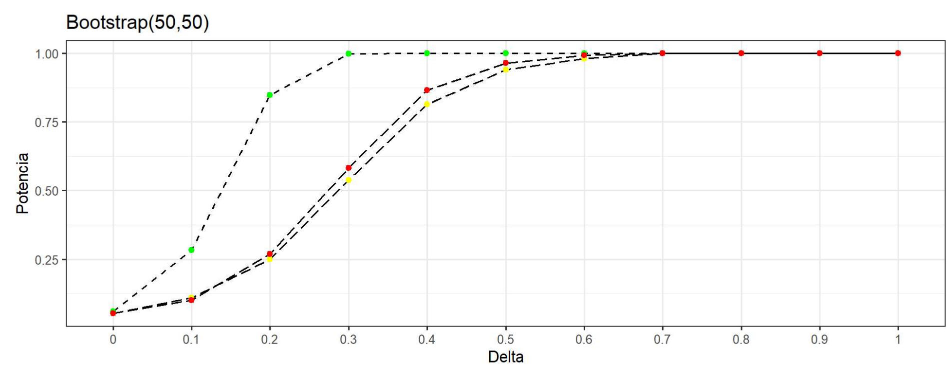
Estadístico	**Método**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Mahalanobis	Montecarlo	0.052	0.518	0.994	1.000	1.000	1.000	1	1	1	1	1
Mahalanobis	Bootstrap	0.052	0.496	0.994	1.000	1.000	1.000	1	1	1	1	1
Mahalanobis	Permutaciones	0.052	0.498	0.996	1.000	1.000	1.000	1	1	1	1	1
L1 std	Montecarlo	0.042	0.172	0.496	0.820	0.974	0.998	1	1	1	1	1
L1 std	Bootstrap	0.050	0.178	0.518	0.838	0.982	0.998	1	1	1	1	1
L1 std	Permutaciones	0.038	0.150	0.492	0.800	0.970	0.998	1	1	1	1	1
L2 std	Montecarlo	0.044	0.174	0.540	0.868	0.988	1.000	1	1	1	1	1
L2 std	Bootstrap	0.044	0.170	0.546	0.872	0.988	1.000	1	1	1	1	1
L2 std	Permutaciones	0.048	0.130	0.508	0.852	0.988	1.000	1	1	1	1	1

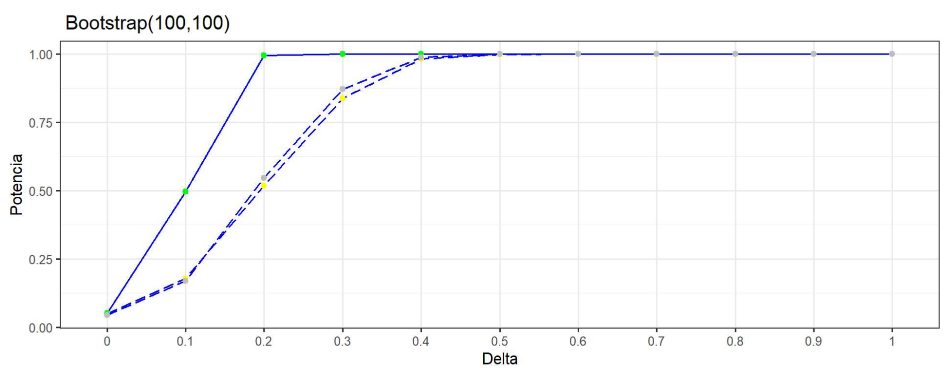
15

Resultados

Escenario 1A - Bootstrap

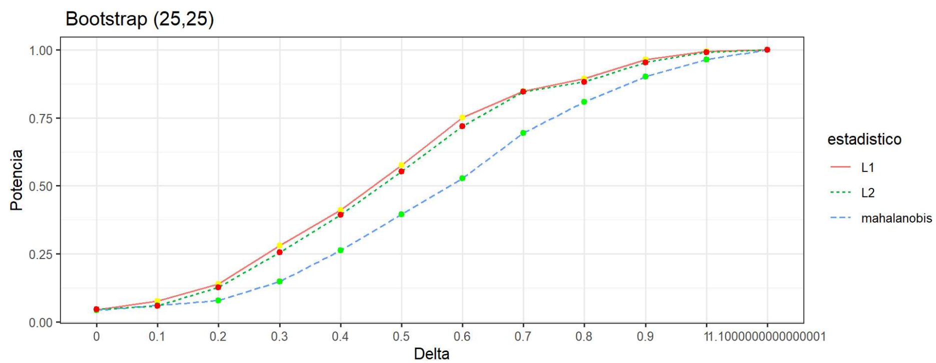


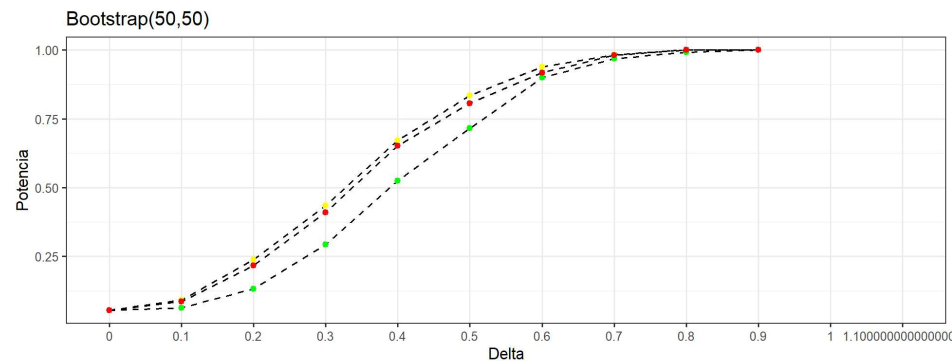


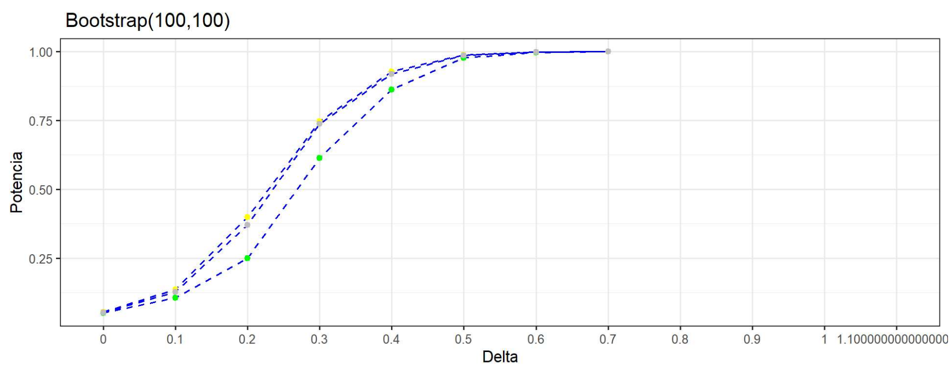


16

Escenario 1B - Bootstrap

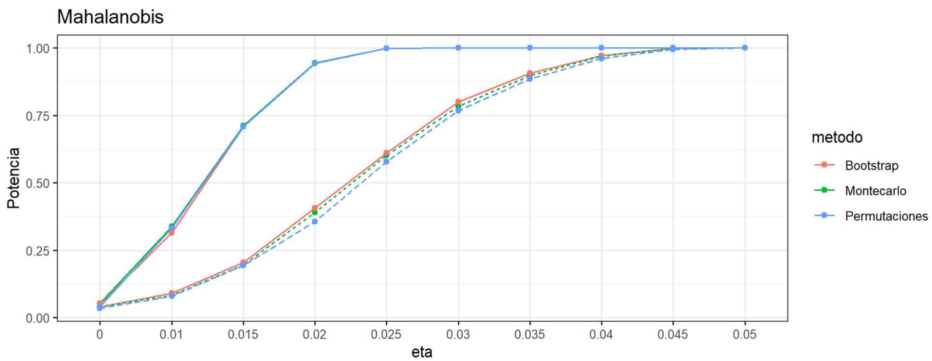


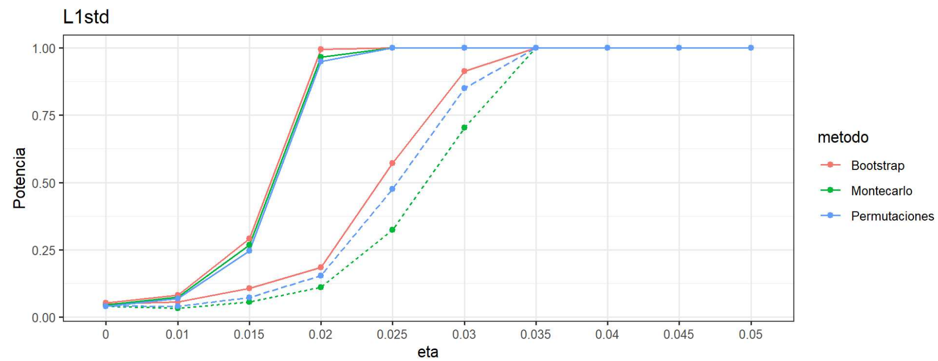


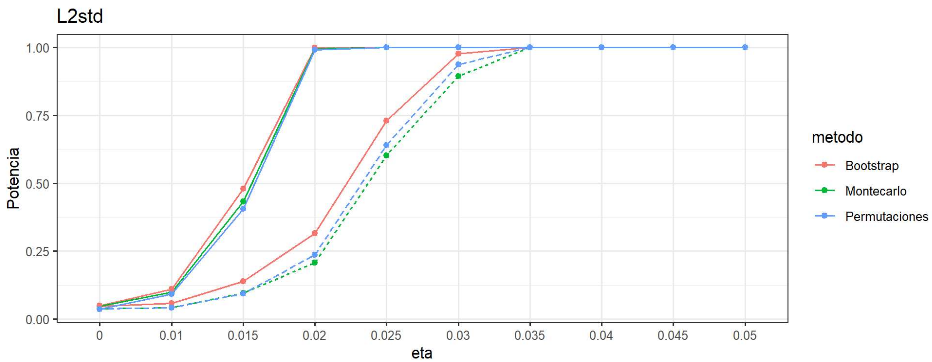


17

Escenario 2A

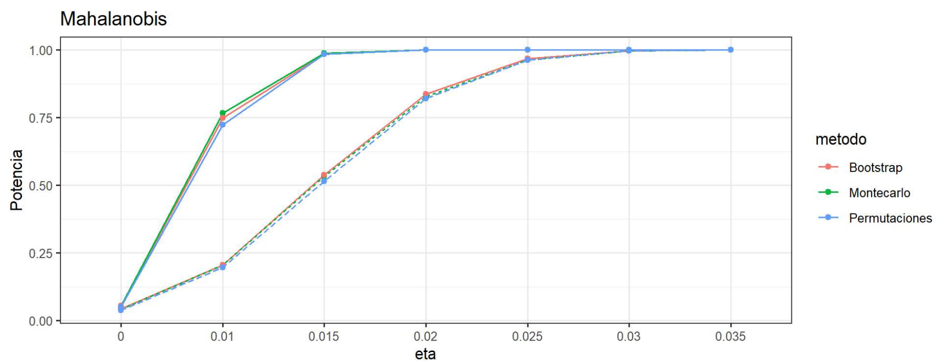


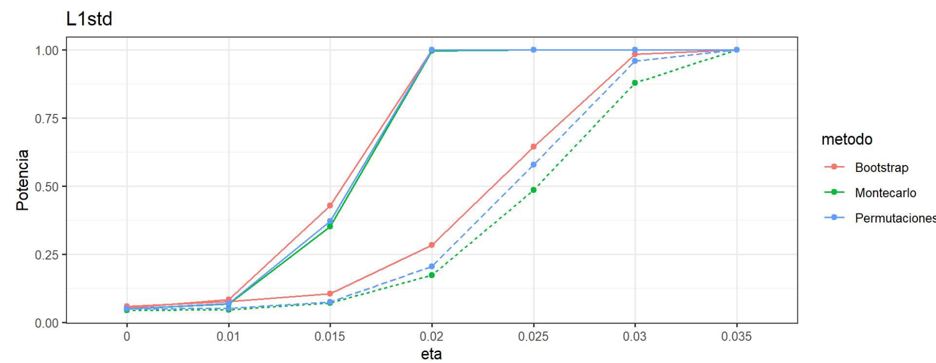


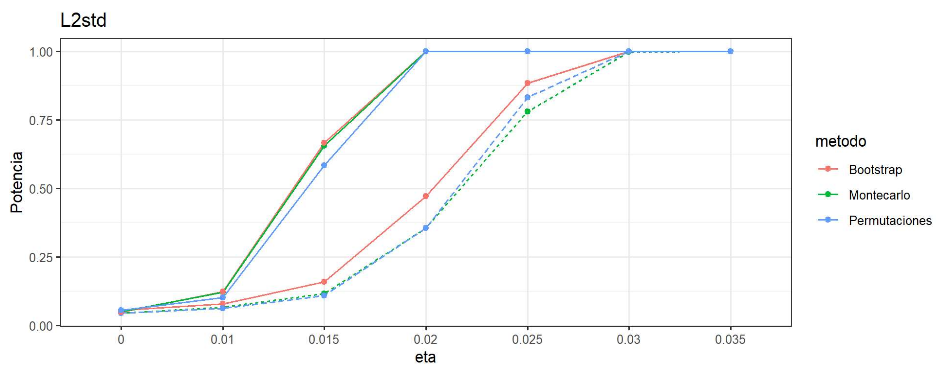


18

Escenario 2B



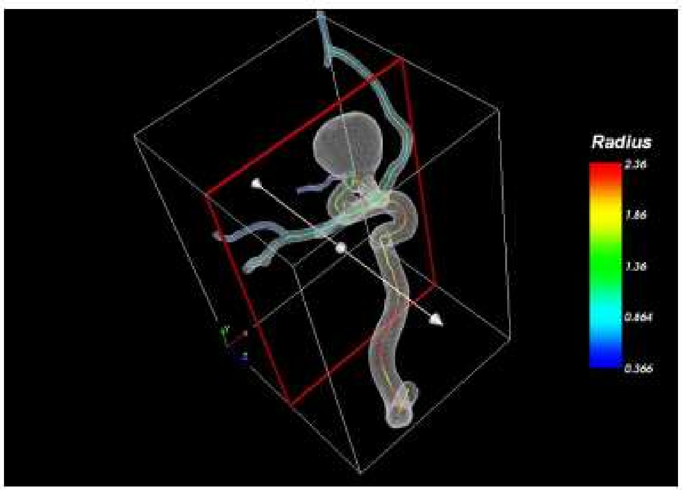


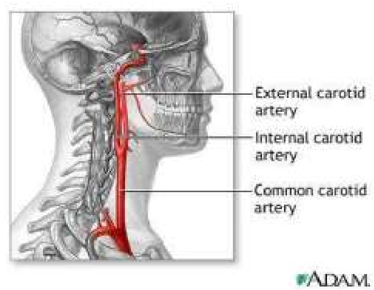


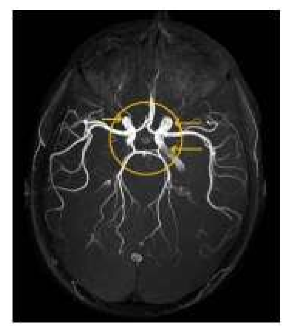
19

Caso de Estudio: Aneurismas de pacientes de alto y bajo riesgo

- Base de datos de pacientes internados en el Ospedale Niguarda Ca'Granda Milano desde septiembre 2002 a octubre 2005 por sospecha de un aneurisma a lo largo de la arteria carótida interna (ACI).
- Se dispone de un conjunto de datos referentes a características geométricas y hemodinámicas de los últimos 5cm de la ACI.







20

Grupos de pacientes

1. El grupo de alto riesgo: cuando el aneurisma se aloja dentro del cráneo, por lo general, provoca daños permanentes o letales en el tejido cerebral.

2. El grupo de bajo riesgo: cuando no hay un aneurisma o si el aneurisma está fuera del cráneo, y su posible ruptura no afecta directamente a los tejidos cerebrales.

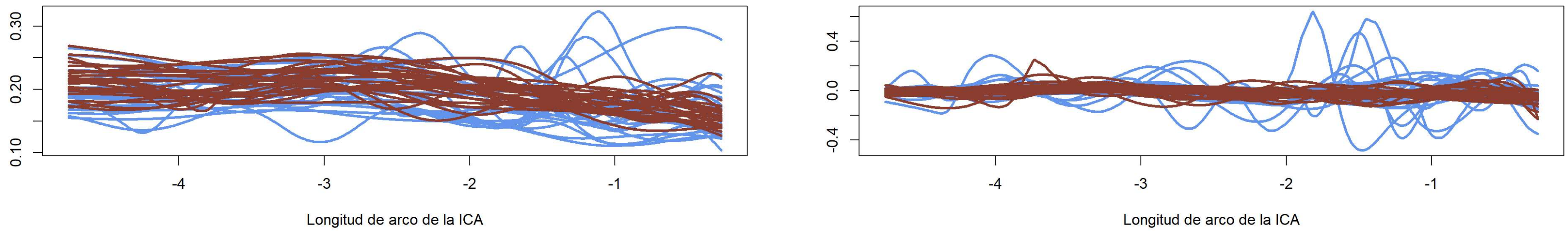
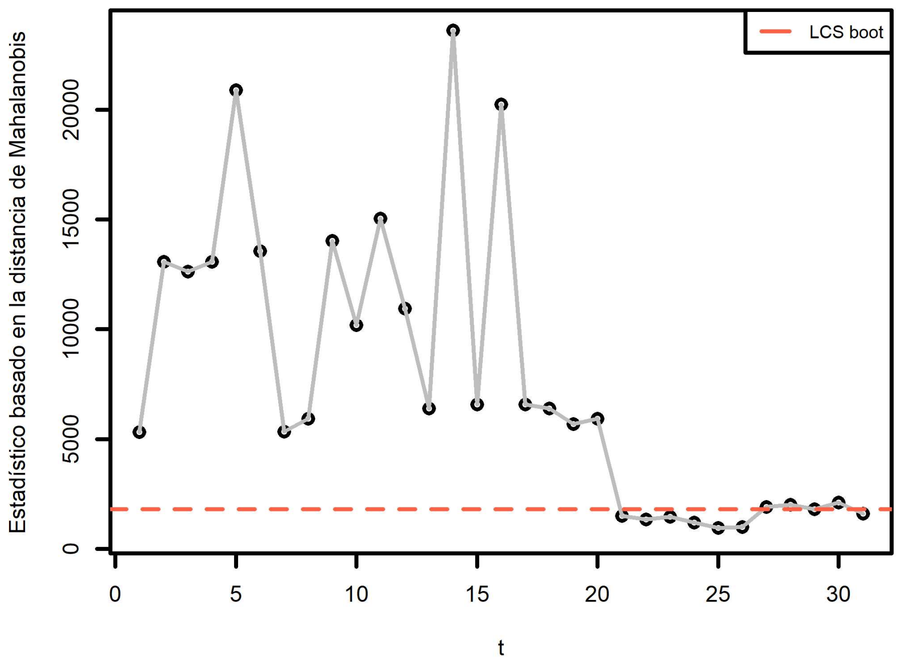


Gráfico de control

Gráfico de control para el radio de la derivada



22

Gracias