



Estimado(a) estudiante:

- Se solicita completar la siguiente información de la manera más concreta posible en al menos tres páginas y hasta en un máximo de cinco.
- Su solicitud será revisada en la siguiente reunión de la Comisión del Posgrado, siempre y cuando la documentación sea recibida en la Administración una semana antes de esa reunión.
- Se recomienda revisar el Reglamento del Posgrado en Estadística, en todo lo referente a la realización de las tesis, para evitar inconvenientes. Abajo se especifica parte del artículo 15, para que con su firma Usted haga constar que se ajustará a los plazos establecidos.

|                       |                                    |  |
|-----------------------|------------------------------------|--|
| Nombre del estudiante |                                    | Alejandro Alonso Salas Vargas  |
| Título de la tesis    |                                    | Comparación de técnicas de análisis de varianza en datos funcionales para mediciones experimentales espectrométricas.  |
| Introducción          | Justificación/importancia del tema | <p>Los grandes avances computacionales y analíticos permiten que muchos procesos que involucran datos sean monitoreados continuamente y su recolección sea más efectiva. La estructura de estos datos funcionales es una fuente rica de información que trae muchas oportunidades para la investigación en diferentes áreas, tales como medicina, geología, química y física, entre muchas más.</p> <p>En este contexto, el análisis de grandes de datos que puedan verse como una realización de un procesos continuo, requieren la aplicación de nuevas técnicas estadísticas, tales como el análisis de datos funcionales (FDA), que considera los datos como funciones aleatorias en un intervalo <math>T</math> del espacio funcional de Hilbert <math>L^2</math> (ver Ramsay y Silverman, 2002; Ramsay y Silverman, 2005; Ramsay et al., 2009; Horváth y Kokoszka, 2012; Zhang, 2013). En comparación con los datos tradicionales, que consisten en observaciones puntuales, los datos funcionales pueden contener información más detallada sobre el sistema subyacente que los genera. Además, surgen nuevos desafíos en el esfuerzo por extraer cualquier información significativa oculta en estos datos.</p> <p>En las últimas dos décadas, el análisis de datos funcionales ha emergido como un área importante en la estadística y se han logrado avances significativos (Ramsay y Silverman, 2005; Zhang, 2013), además se han adaptado muchos métodos y técnicas estadísticas a este tipo de datos, tales como el análisis de componentes principales (FPCA) (Ramsay y Silverman, 2002; Horváth y Kokoszka, 2012), regresión (Morris, 2015), análisis de conglomerados (Jacques y Predas, 2014) y análisis de varianza para datos funcionales (FANOVA) (Faraway, 1997; Shen y Faraway, 2004; Cuevas et al., 2004; Ramsay y Silverman, 2005; Zhang y Chen, 2007; Cuestas et al., 2010; Zhang, Peng y Zhang, 2013) por mencionar algunos.</p> <p>Zhang et al. (2013) menciona que el Análisis de Varianza con respuestas Funcionales (FANOVA, por sus siglas en inglés) es un problema fundamental en la inferencia estadística para determinar diferencias significativas en las curvas medias o funciones promedio (medias funcionales), al contrastar una hipótesis nula <math>H_0</math> de igualdad de medias funcionales; esto permite probar, en diseños experimentales funcionales con <math>k</math> grupos de curvas, el efecto que tiene un determinado tratamiento sobre una muestra y así inferir sobre una población determinada. Es importante mencionar que la variable respuesta en este caso podría ser toda una curva (funcional) o multivariada. Pese a la importancia que tiene determinar la existencia de diferencias significativas entre funciones promedio, este tema ha recibido menos atención, sin embargo, en los últimos años, han surgido dos enfoques generales diferentes para realizar la prueba FANOVA, expuestos en Shen y Faraway</p> |
|                       |                                    |  |
|                       |                                    |  |
|                       |                                    |  |



|           |   |   |
|-----------|---|---|
|           |   | <p>(2004), así como en otros similares (Cuevas, Febrero y Fraiman, 2004 y Cuesta y Febrero, 2010), para un estadístico de prueba global (este es el enfoque que se desarrollará en este trabajo). Se dice que es global porque se usa para detectar diferencias en cualquier parte del dominio funcional completo (en cualquier lugar <math>t \in T</math>). Un enfoque alternativo considerado por Ramsay y Silverman (2005), consiste en utilizar un estadístico de prueba puntual (o individual) para realizar la inferencia en <math>t</math>, es decir, identificar regiones específicas de <math>t</math> con una diferencia significativa entre las medias funcionales.</p> <p>Una de las características de los datos funcionales es que a partir de un conjunto de mediciones tomadas a lo largo de un tiempo determinado conforman una observación y esta observación es transformada a una función (cada observación es una función). Esta característica se puede apreciar en mediciones del espectro de luz reflejada por un objeto de interés (mediciones espectrométricas) o dicho de otro modo, las mediciones espectrométricas son medidas de cómo se refleja la luz sobre un determinado material y su representación es una curva que especifica la descomposición e función de la longitud de onda.</p> <p>En este trabajo se realizarán mediciones espectrométricas tomadas de escarabajos con apariencia metálica, ya que el Centro de Investigación en Ciencia e Ingeniería de Materiales (CICIMA) le interesa emular el aspecto metálico de estos escarabajos, lo que dará la posibilidad de reproducir esas cualidades en nuevos materiales que simulen un recubrimiento de oro o plata, o bien de una capa metálica, aunque no las propiedades del metal como conducción eléctrica o térmica. Este material tendría la otra ventaja de que estaría compuesto de moléculas orgánicas, por lo que sería biodegradable e inspirado en la biodiversidad costarricense. La industria se beneficiaría de los resultados de estas investigaciones ya que reproducirían los mismos tonos y propiedades del color de los escarabajos con aspectos metálicos en la fabricación nuevos materiales.</p> |
|           | Contribución de la tesis a la Estadística como disciplina | La contribución principal de esta investigación es brindar una comparación de las ventajas y desventajas de las pruebas globales de FANOVA, según se considere la respuesta como funcional, multivariada o bien univariada, en diferentes escenarios simulados y en el contexto específico de experimentos que utilicen mediciones espectrofotométricas. Además, esta investigación busca proporcionar herramientas estadísticas que permitan contextualizar las observaciones de los espectros de luz reflejada en el marco de propiedades y fenómenos físicos.  |
| Objetivos | Objetivo general  | Comparar el desempeño de las pruebas globales de análisis de varianza en datos funcionales según se considere la variable de respuesta como funcional, multivariada o univariada, para proveer una guía de uso en los experimentos llevados a cabo con mediciones espectrométricas.   |
|           | Objetivos específicos                                     | <p>Relacionar teóricamente las técnicas para respuestas funcionales con el uso de transformaciones multivariadas para mediciones espectrométricas.</p> <p>Comparar la potencia de las pruebas globales según las diferentes formas que puede tomar la variable respuesta funcional.</p> <p>Determinar ventajas y desventajas de cada uno de los escenarios de simulación establecidos donde se relaciones el amaño de muestra, el número de tratamientos a contrastar en el FANOVA, la potencia de prueba (el error tipo I y el porcentaje de resultados acertados).</p>  |



|             |   |   |
|-------------|---|---|
|             |   | Contrastar hipótesis globales con datos espectrométricos experimentales del Centro de Investigación en Ciencia e Ingeniería de Materiales (CICIMA), utilizando las recomendaciones establecidas.  |
| Metodología | Referentes o elementos teóricos que va a utilizar | <p><b><u>Análisis de datos funcionales</u></b></p> <p>El análisis de datos funcionales (FDA) se ocupa del análisis y la teoría de los datos en forma de funciones, imágenes y formas, u objetos más generales. Una función representa una observación, donde para cada sujeto en una muestra aleatoria, se registran una o varias funciones (Ramsay y Silverman, 2005), decir que una variable aleatoria <math>X</math> se denomina variable funcional si toma valores en un espacio dimensional infinito (o espacio funcional normalizado o semi-completo normado) y, por lo tanto, una observación <math>x \in X</math> se denomina dato funcional y, además, , un conjunto de datos funcional <math>x_1, \dots, x_n</math> es la observación de <math>n</math> variables funcionales <math>X_1, \dots, X_n</math> distribuidos de manera idéntica. Los datos funcionales, también llamados datos longitudinales, se asocian con procesos que se monitorean continuamente en el tiempo. Es decir, cuando una variable se mide en un conjunto discreto y finito de valores dispuestos, considerando que esta variable sigue una relación funcional continua. Un caso especial es cuando la variable funcional <math>X</math> pertenece a un espacio de Hilbert, como es el caso de las funciones continuas en un intervalo (Ramsay y Silverman, 2005). Según Ramsay y Silverman, 2002, un conjunto de datos funcionales se define de la siguiente manera:</p> $x_n(t): t \in T, \quad n = 1, 2, \dots, N.$ <p>para los cuales los valores <math>x_n(t)</math>: existen en cualquier punto <math>t \in T</math>.</p> <p><b><u>FANOVA de una vía para datos funcionales</u></b></p> <p>En Gorécki y Smaga (2015) se introduce el problema de FANOVA, de la siguiente manera: sea <math>Y_{i1}(t), Y_{i2}(t), \dots, Y_{in_i}(t)</math> con <math>i = 1, \dots, k</math>, <math>k</math> grupos de funciones aleatorias definidas en un intervalo finito dado <math>T = [a, b]</math> y su covarianza funcional <math>\gamma(s, t)</math> con <math>s, t \in T</math>. Sea <math>SP(\mu, \gamma)</math> un proceso estocástico con media funcional <math>m(t)</math>, con <math>t \in T</math>. Aquí se asume que las funciones <math>Y_{i1}(t), Y_{i2}(t), \dots, Y_{in_i}(t)</math> son idénticamente distribuidas y en dicho proceso estocástico es común probar la igualdad de las <math>k</math> medias funcionales. Para esto se define la siguiente hipótesis nula:</p> $H_0: \mu_1(t) = \dots = \mu_k(t), \quad t \in T. \quad (1)$ <p>Cuya hipótesis alternativa es la negación de esta igualdad. La función de covarianza función es</p> $\hat{\gamma}(s, t) = \frac{1}{n - k} \sum_{i=1}^k \sum_{j=1}^{n_i} [Y_{ij}(s) - \bar{Y}_i(s)][Y_{ij}(t) - \bar{Y}_i(t)] \quad (2)$ <p><b><u>Estadístico propuesto por Cuevas (2004)</u></b></p> <p>Gorécki y Smaga (2015) indica que Cuevas (2004 p.114) propone un nuevo estadístico para contrastar la hipótesis nula de igualdad de medias funcionales (1). Este estadístico de prueba <math>V_n</math> se puede usar para la prueba en el caso "heterocedástico":</p> |

$$V_n = \sum_{1 \leq i < j \leq k} n_i \int_T [\bar{Y}_i(t) - \bar{Y}_j(t)]^2 dt$$

Bajo la hipótesis nula (1) y los supuestos de que  $n_i, n \rightarrow \infty$  de tal manera que  $n_i / n \rightarrow p_i > 0$  para  $i = 1, \dots, k$ , prueba que la distribución aproximada de  $V_n$  es la de la estadístico.:

$$V = \sum_{1 \leq i < j \leq k} n_i \int_T \left[ Z_i(t) - \sqrt{p_i/p_j} Z_j(t) \right]^2 dt \quad (3)$$

Donde  $Z_1(t), Z_2(t), \dots, Z_k(t)$  son procesos gaussianos con media cero y covarianza funcional (2) y para los casos heterocedásticos se considera las covarianzas:

$$\hat{\gamma}_i(s, t) = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} [Y_{ij}(s) - \bar{Y}_i(s)][Y_{ij}(t) - \bar{Y}_i(t)], i = 1, \dots, k$$

### **Prueba basada en la norma $L^2$**

Según Faraway (1997) y Zhang y Chen (2007) emplean el estadístico de prueba  $S_n = \int_T SSR_n(t)dt$  bajo la hipótesis nula definida en (1), se puede demostrar que  $S_n \sim \beta \chi_d^2$  donde  $\beta = \text{tr}(\gamma^{\otimes 2})/\text{tr}(\gamma)$ ,  $d = (k-1)\kappa$ ,  $\kappa = \text{tr}^2(\gamma)/\text{tr}(\gamma^{\otimes 2})$  y  $\gamma^{\otimes 2}(s, t) = \int_T \gamma(s, u)\gamma(u, t)du$ . Esta aproximación se emplea para el cálculo del valor  $p$  de  $S_n$ ,  $P(\chi_d^2 \geq S_n/\beta)$ , con valor crítico  $\beta \chi_d^2(1 - \alpha)$ . Se utiliza los siguientes estimadores:

**Para el método ingenuo:**  $\hat{\beta} = \text{tr}(\hat{\gamma}^{\otimes 2})/\text{tr}(\hat{\gamma})$ ,  $\hat{d} = (k-1)\hat{\kappa}$ ,  $\hat{\kappa} = \text{tr}^2(\hat{\gamma})/\text{tr}(\hat{\gamma}^{\otimes 2})$  y covarianza  $\hat{\gamma}(s, t)$  definida en (2).

**Para el método de sesgo reducido,:**  $\hat{\beta} = \text{tr}(\widehat{\gamma^{\otimes 2}})/\text{tr}(\hat{\gamma})$ ,  $\hat{d} = (k-1)\hat{\kappa}$ ,  $\hat{\kappa} = \text{tr}^2(\hat{\gamma}) / \text{tr}(\widehat{\gamma^{\otimes 2}})$  y covarianza  $\hat{\gamma}(s, t)$  definida en (2). Además, el autor define:

$$\begin{aligned} \text{tr}(\widehat{\gamma^{\otimes 2}}) &= \frac{(n-k)^2}{(n-k-1)(n-k+2)} \left( \text{tr}(\hat{\gamma}^{\otimes 2}) - \frac{\text{tr}^2(\hat{\gamma})}{n-k} \right) \\ \text{tr}^2(\hat{\gamma}) &= \frac{(n-k)(n-k+1)}{(n-k-1)(n-k+2)} \left( \text{tr}^2(\hat{\gamma}) - \frac{2\text{tr}(\hat{\gamma}^{\otimes 2})}{n-k+1} \right) \end{aligned}$$

Por lo tanto, se tiene dos pruebas basadas en la norma  $L^2$  que difieren en el método de estimación de los parámetros  $\beta$  y  $\kappa$ , es decir, con el método ingenuo de estimación de esos parámetros y con el método de estimación de sesgo reducido.

### **La prueba de tipo $F$ de Shen y Faraway (2004)**

Esta prueba es propuesta por Shen y Faraway (2004), además se puede detallar en Zhang (2011), Vsevolozhskaya et al. (2014) y en adoptada para (1) y utiliza el estadístico de prueba:

$$F_n = \frac{\int_T SSR_n(t)dt/(k-1)}{\int_T SSE_n(t)dt/(n-k)}$$

Donde  $F_n \sim F_{d_1, d_2}$  con  $d_1 = (k-1)\kappa$  similar a la prueba anterior de la norma de  $L^2$  y  $d_2 = (n-k)\kappa$ , la distribución aproximada nula de la prueba de tipo  $F$  se puede usar para calcular el valor  $p$  de  $F_n$  o su valor crítico. Aquí se tendría dos posibilidades una estimación por el método ingenuo o el de sesgo reducido descrito en la prueba anterior, para más detalles se puede consultar Zhang (2013).

Cuando las  $k$  muestras no son gaussianas y cuando los tamaños de muestra son pequeños, no se recomienda la prueba basada en la norma  $L^2$  y ni la prueba de tipo  $F$  (Zhang, 2013a)

### Prueba $F$ puntual globalizada.

La globalización de la prueba  $F$  puntual (la prueba GPF, Zhang y Liang 2014) usa la estadística de prueba  $T_n = \int_T F_n(t) dt$ . Bajo la hipótesis nula (1), se puede demostrar que  $T_n \sim \hat{\beta}_w \chi_{\hat{d}_w}^2$  aproximadamente, donde:

$$\begin{aligned}\hat{\beta}_w &= (n-k-2) \text{tr}(\hat{\gamma}_w^{\otimes 2}) / [(k-1)(n-k)(b-a)] \\ \hat{d}_w &= (k-1)(n-k)^2(b-a)^2 / [(k-1)(n-k)(b-a)] \\ \hat{\gamma}_w(s, t) &= \hat{\gamma}(s, t) / \sqrt{\hat{\gamma}(s, s)\hat{\gamma}(t, t)}\end{aligned}$$

Para  $\hat{\gamma}(s, t)$  dado en (2). Esta distribución aproximada se puede usar para calcular el valor  $p$  de  $T_n$  o su valor crítico casi de la misma manera que para la prueba basada en la norma  $L^2$ .

### Prueba adaptativa de Neyman

Fan y Lin (1998) propusieron usar la prueba adaptativa de Neyman para comparar diferentes conjuntos de curvas. Esta prueba puede ser adoptada para la prueba (1) (ver, por ejemplo, Laukaitis y Rackauskas 2005, que la usaron en el análisis para tareas de segmentación de clientes).

Para esto se define  $Y_{ij}^*(l)$ ,  $l = 1, \dots, T^*$ , la transformada discreta de Fourier para la función  $Y_{ij}$ , para  $i$  y  $j$  dados. El autor asume que  $Y_{ij}^*(l) = m_i^*(l) + \varepsilon_{ij}^*(l)$ , con  $\varepsilon_{ij}^*(l)$  independientes y  $\varepsilon_{ij}^*(l) \sim N(0, \sigma_i^2(l))$ .

La hipótesis nula es  $H_0: m_1^*(l) = \dots = m_k^*(l) = \dots = m^*(l)$ ,  $\forall l$ . Los autores proponen el estadístico:

$$T_{\text{HANOVA}} = \sqrt{2 \log \log T^*} F_{\hat{m}} - \left\{ 2 \log \log T^* + \frac{1}{2} \log \log \log T^* - \frac{1}{2} \log 4\pi \right\}$$



Donde

$$\begin{aligned}F_{\hat{m}} &= \max_{1 \leq m \leq T^*} \frac{1}{\sqrt{2(k-1)m}} \left[ \sum_{l=1}^m \sum_{i=1}^k n_i \hat{\sigma}_i^*(l)^{-2} (\bar{Y}_i^*(l) - \bar{Y}^*(l))^2 - (k-1)m \right] \\ \bar{Y}_i^*(l) &= \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}^*(l) \\ \hat{\sigma}_i^*(l) &= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij}^*(l) - \bar{Y}_i^*(l))^2 \\ \bar{Y}^*(l) &= \frac{\sum_{i=1}^k n_i \hat{\sigma}_i^*(l)^{-2} \bar{Y}_i^*(l)}{\sum_{i=1}^k n_i \hat{\sigma}_i^*(l)^{-2}}\end{aligned}$$



|   |  |
|---|--|
|   | <p>Se rechaza <math>H_0^*</math> cuando el estadístico <math>T_{\text{HANOVA}}</math> es demasiado grande. Los valores críticos para esta estadística de prueba se pueden obtener mediante la distribución de muestra finita proporcionada por Fan y Lin (1998, p. 1010).</p> <p><b>Pruebas basadas en una representación de funciones básicas</b></p> <p>Estas pruebas están basadas en el método B-Spline propuestos por Shen y Faraway (2004), las cuales son representaciones de funciones base (procesos estocásticos). Sea <math>k</math> grupos de procesos estocásticos <math>Y_{ij} \in L_2(T), i = 1, \dots, k</math> y <math>j = 1, \dots, n_i</math>. A <math>L_2(T)</math> se le conoce como el espacio de Hilbert de funciones integrables cuadradas en el intervalo <math>T</math>, el cual cuenta con el producto interno <math>\langle f, g \rangle = \int_T f(t)g(t) dt</math>.</p> <p>Según Ramsay and Silverman (2005), cualquier función puede ser aproximada mediante una base <math>\{\phi_l(t)\}</math> ortogonal de <math>L_2(T)</math>. Es por esta razón de que estos autores consideran los procesos estocásticos de la forma <math>Y_{ij}(t), t \in T</math>, los cuales pueden ser representados por un número finito de funciones de base ortonormal de la forma:</p> $Y_{ij}(t) = \sum_{l=0}^k c_{ijl} \phi_l(t), \quad t \in T$ <p>En esta última expresión <math>c_{ijl}</math> son variables aleatorias con varianza finita y <math>l = 0, 1, \dots, K</math> con <math>K</math> suficientemente grande. Empelando anotación matricial:</p> $Y_{ij}(t) = \mathbf{c}'_{ij} \boldsymbol{\phi}(t), \quad \bar{Y}(t) = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}(t), \quad \bar{Y}_i(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}(t)$ <p>Para <math>t \in T</math>, respectivamente la media funcional de la muestra y la media funcional para cada grupo y <math>\boldsymbol{\phi}(t) = (\phi_0(t), \phi_1(t), \dots, \phi_K(t))'</math>, <math>\mathbf{c}_{ij} = (c_{ij0}, c_{ij1}, \dots, c_{ijK})'</math></p> <p>Su estadístico de prueba es:</p> $F = \frac{(a - b)(n - k)}{(c - a)(k - 1)}$ <p>Con</p> $a = \sum_{i=1}^k \frac{1}{n_i} \sum_{m=1}^{n_i} \sum_{s=1}^{n_i} \mathbf{c}'_{im} \mathbf{c}_{is}, \quad b = \frac{1}{n} \sum_{i=1}^k \sum_{m=1}^{n_i} \sum_{t=1}^k \sum_{v=1}^{n_t} \mathbf{c}'_{im} \mathbf{c}_{tv}, \quad c = \sum_{i=1}^k \sum_{j=1}^{n_i} \mathbf{c}'_{ij} \mathbf{c}_{ij}$ |
| Fuente(s) de datos (si ya están recolectados) o diseño del estudio (si no están recolectados) | <p>Los datos pertenecen al Centro de Investigación en Ciencia e Ingeniería de Materiales (CICIMA) de la Universidad de Costa Rica, los cuales corresponden a mediciones de la reflectancia de la luz para estudiar el origen de la apariencia metálica de los escarabajos en diferentes regiones del élitro del insecto. El instrumento utilizado para la recolección de los datos es un micro-espectrofotómetro. El conjunto de datos cuenta con los espectros de luz en el escarabajo según el tamaño del área en cada, así como la resolución del microscopio (5x, 10x, 50x, 100x) y la especie del escarabajo.</p>   |
| Definición de variable(s) de estudio  | <p><b><u>Espectro de luz reflejada</u></b></p> <p>Es la cantidad proporcional de luz reflejada, como una función de las longitudes de onda para producir un espectro de reflectancia. Esto se consigue haciendo pasar la luz a través de un dispositivo monocromático que fracciona la luz en distintos intervalos de longitudes de onda. El instrumento se calibra con una muestra especular pues el élitro de los escarabajos en estudio es brillante en forma similar a un espejo, cuya reflectancia en cada segmento de longitudes de onda se conoce, al compararla con una superficie de reflexión difusa perfecta.</p>   |



|  |   |   |
|--|---|---|
|  | <p>Espectro visible por el ojo humano (Luz)</p>  <p>625 - 740: se tienen los Rojos.<br/>590 - 625: Naranjas.<br/>565 - 590: Amarillos.<br/>520 - 565: Verdes.<br/>500 - 520: Cian.<br/>435 - 500: Azules.<br/>380 - 435: Violetas.</p>  |   |
| Evidencias de calidad de la medición para la(s) variable(s) del estudio  | <p>La tecnología del micro-espectrofotómetro permite identificar estructuras con el espectro del color, de modo que su captación es aún más exacta, en otras palabras, asegura la calidad de las mediciones de los espectros del color a la hora de determinar la cantidad de reflexión de la luz (para un mismo punto, los espectros medidos son reproducibles a centésimas de nanómetro).</p>   |  |
| Modelo(s) estadísticos o técnica(s) estadística(s) de análisis a emplear | <p><b>Análisis de datos funcionales</b> (ver [10], [15], [16], [17], [23], [26]).</p> <p>Según Ramsay y Silverman, 2002, un conjunto de datos funcionales se define de la siguiente manera:</p> $x_n(t): t \in T, n = 1, 2, \dots, N.$ <p>para los cuales los valores <math>x_n(t)</math>: existen en cualquier punto <math>t \in T</math>.</p> <p><b>Análisis de varianza para datos funcionales</b> (ver [4], [6], [9], [18], [24]).</p> <p>En Zhang (2013a) se define el problema de FANOVA de una vía para datos funcionales de la siguiente manera.</p> <p>Suponiendo <math>k</math> muestras independientes <math>Y_{i1}(t), \dots, Y_{in_i}(t)</math>, <math>i = 1, \dots, k</math>, las cuales satisfacen con: <math>Y_{ij}(t) = \mu_i(t) + v_{ij}(t)</math>, con <math>v_{ij}(t) \stackrel{i.i.d}{\sim} SP(0, \gamma)</math> y <math>i = 1, \dots, k; j = 1, \dots, n_i</math>. En esta expresión <math>\mu_i(t)</math> son las funciones medias desconocidas de cada una de las <math>k</math> muestras, <math>v_{ij}(t)</math> es la función de efecto y <math>\gamma(s, t)</math> es la función de covarianza común.</p> <p><b>Escenarios simulados.</b></p> <p>Se realizarán 9 escenarios simulados compuestos por 3 tamaños de muestra (<math>n = 50</math>, <math>n = 100</math>, <math>n = 500</math>) y por tres distintos grupos de tratamientos (<math>T = 2, T = 3, T = 4</math>) los cuales son las áreas del cuadrado sobre el escarabajo donde se refleja la luz.</p> <p>Escenario 1: Un tamaño de muestra de <math>n = 50</math>, un número de tratamientos de <math>T = 2</math> y un número de experimentos de Monte Carlo de 1000.</p> <p>Escenario 2: Un tamaño de muestra de <math>n = 50</math>, un número de tratamientos de <math>T = 3</math> y un número de experimentos de Monte Carlo de 1000.</p> <p>Escenario 3: Un tamaño de muestra de <math>n = 50</math>, un número de tratamientos de <math>T = 4</math> y un número de experimentos de Monte Carlo de 1000.</p> |   |



|                            |  |   |
|----------------------------|--|---|
|                            |  | <p>Escenario 4: Un tamaño de muestra de <math>n = 100</math>, un número de tratamientos de <math>T = 2</math> y un número de experimentos de Monte Carlo de 1000.</p> <p>Escenario 5: Un tamaño de muestra de <math>n = 100</math>, un número de tratamientos de <math>T = 3</math> y un número de experimentos de Monte Carlo de 1000.</p> <p>Escenario 6: Un tamaño de muestra de <math>n = 100</math>, un número de tratamientos de <math>T = 4</math> y un número de experimentos de Monte Carlo de 1000.</p> <p>Escenario 7: Un tamaño de muestra de <math>n = 500</math>, un número de tratamientos de <math>T = 2</math> y un número de experimentos de Monte Carlo de 1000.</p> <p>Escenario 8: Un tamaño de muestra de <math>n = 500</math>, un número de tratamientos de <math>T = 3</math> y un número de experimentos de Monte Carlo de 1000.</p> <p>Escenario 9: Un tamaño de muestra de <math>n = 500</math>, un número de tratamientos de <math>T = 4</math> y un número de experimentos de Monte Carlo de 1000.</p> <p>Los datos se generarán del modelo FANOVA: <math>Y_{ij}(t) = \mu_i(t) + v_{ij}(t)</math>, con <math>v_{ij}(t) \stackrel{i.i.d}{\sim} \text{SP}(0, \gamma)</math> y <math>i = 1, \dots, k; j = 1, \dots, n_i</math>.</p> <p>Se tiene tres formas de analizar los datos: con respuesta funcional, univariada y multivariada. Posteriormente se comparará la potencia empírica, error tipo I y la cobertura de cada uno de los test de FANOVA en el efecto del tratamiento 1, con los distintos escenarios de datos generados y las tres propuestas de modelos, bajo niveles de significancia estadística para el error de tipo I de 1%, 5% y 10% y un número de experimentos de Monte Carlo de 1000.</p> <p>El objetivo del estudio de Monte Carlo (simulación) es evaluar el desempeño del test FANOVA para este tipo de datos, comparado con dos opciones más tradicionales: la univariada y la multivariada.</p> |
|                            | Software estadístico a utilizar  | Todos los análisis estadísticos se realizarán en el software libre R (R Core Team, 2016).   |
| Referencias bibliográficas | <p>[1] Abramovich, F., Angelini, C. (2006). <i>Testing in mixed-effects FANOVA models</i>. Journal of Statistical Planning and Inference, 136(1), 4326–4348. URL: "<a href="http://www.math.tau.ac.il/~felix/PAPERS/jspi2006.pdf">http://www.math.tau.ac.il/~felix/PAPERS/jspi2006.pdf</a>".</p> <p>[2] Abramovich, F., Antoniadis, A., Sapatinas, T., Vidakovic, B. (2004). <i>Optimal testing in a fixedeffects functional analysis of variance model</i>. International Journal of Wavelets, Multiresolution and Information Processing, 2(4), 323–349, doi:10.1142/S0219691304000639.</p> <p>[3] Antoniadis, A., Sapatinas, T. (2007). <i>Estimation and inference in functional mixed-effects models</i>. Computational Statistics &amp; Data Analysis, 51(10), 4793–4813, doi:10.1016/j.csda.2006.09.038.</p> <p>[4] Cuestas, A., Febrero, M. (2010). <i>A simple multiway ANOVA for functional data</i>. TEST, 19(3), 537–557, doi:10.1007/s11749-010-0185-3.</p> <p>[5] Cuevas, A. (2014). <i>A partial overview of the theory of statistics with functional data</i>. Journal of Statistical Planning and Inference, 47(1), 1–23, doi:10.1016/j.jspi.2013.04.002.</p> |   |





- [6] Cuevas, A., Febrero, M., Fraiman, R. (2004). *An anova test for functional data*. Computational Statistics & Data Analysis, 47(1), 111–122, doi:10.1016/j.csda.2003.10.021.
- [7] Fan, J., Lin, S. (1998). *Test of significance when data are curves*. Journal of the American Statistical Association, 30(4), 987–1010, doi:10.1080/01621459.1998.10473763.
- [8] Faraway, J. (1997). *Regression Analysis for a Functional Response*. Technometrics, 39(3), 254–261.
- [9] Gorécki, T., Smaga, L. (2015). *A comparison of tests for the one-way ANOVA problem for functional data*. Computational Statistics, 30(4), 987–1010, doi:10.1007/s00180-015-0555-0.
- [10] Horváth, L., Kokoszka, P. (2012). *Inference for functional data with applications*. Springer, New York.
- [11] Jacques, J. y Predas, C. (2014). *Model-based clustering for multivariate functional data*. Computational Statistics & Data Analysis, 71(1), 92–106, doi:10.1016/j.csda.2012.12.004.
- [12] Laukaitis, A., Račkauskas, A. (2005). *Functional data analysis for clients segmentation tasks*. Computational Statistics, 163(1), 210–216, doi:10.1016/j.ejor.2004.01.010.
- [13] Morris, J. (2015). *Functional Regression*. The Annual Review of Statistics and Its Application, 2(11), 1–39.
- [14] R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>
- [15] Ramsay, J., Hooker, G., Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer, Berlin.
- [16] Ramsay, J., Silverman, B. (2002). *Applied functional data analysis: methods and case studies*. Springer, Berlin.
- [17] Ramsay, J., Silverman, B. (2005). *Functional data analysis*. Springer, Berlin.
- [18] Shen, Q., Faraway, J. (2004). *An F test for linear models with functional responses*. Statistica Sinica.
- [19] Vsevolozhskaya, O., Zaykin, D., Greenwood, M., Wei, C. y Lu, Q. (2014). *Functional Analysis of Variance for Association Studies*. PLoS ONE 9(9), e105074, doi:10.1371/journal.pone.0105074
- [20] Wang, JL., Chiou, JM., Müller, HG. (2016). *Functional data analysis*. The Annual Review of Statistics and Its Application, 3, 257–95, doi: 10.1146/annurev-statistics-041715-033624.
- [21] Yang, X., Nie, K. (2008). *Hypothesis testing in functional linear regression models with Neyman's truncation and wavelet thresholding for longitudinal data*. Statistics in Medicine, 27(6), 845–863.
- [22] Zhang, C., Peng, H., Zhang, J. (2013). *Two sample tests for functional data*. Communications in Statistics–Theory and Methods, 39(4), 559–578.
- [23] Zhang, J. (2011). *Statistical inferences for linear models with functional responses*. Statistica Sinica, 21(3), 1431–1451. URL: "<http://www3.stat.sinica.edu.tw/sstest/oldpdf/A21n319.pdf>".
- [24] Zhang, J. (2013a). *Analysis of variance for functional data*. Chapman & Hall, London.
- [25] Zhang, JT., Cheng, MY., Tseng, CJ., Wu, HT. (2013b). *A new test for one-way anova with functional data and application to ischemic heart screening*. arXiv preprint arXiv:1309.7376.



|            |   |                     |                       |        |   |
|------------|---|---------------------|-----------------------|--------|---|
|            | <p>[26] Zhang, J., Chen, J. (2007). <i>Statistical inferences for functional data</i>. The Annals of Statistics, 35(3), 1052–1079, doi:10.1214/009053606000001505.</p> <p>[27] Zhang, J., Liang, X. (2014). <i>One-Way anova for Functional Data via Globalizing the Pointwise F-test</i>. Scandinavian Journal of Statistics, 41(1), 51–71, doi:10.1111/sjos.12025.</p> <p>[28] Zhang, J., Liang, X., Xiao, S. (2010). <i>On the two-sample Behrens-Fisher problem for functional data</i>. Journal of Statistical Theory and Practice, 4(4), 571–587.</p> |                     |                       |        |   |
| Cronograma | Actividad   | Duración<br>(Meses) | Periodo<br>Inicio Fin |        | Especificación  |
|            | Revisión bibliográfica  | 2                   | ene-19                | feb-19 | Revisión de las fuentes bibliográficas para establecer el marco teórico de la investigación.  |
|            | Elaboración de la introducción y de la justificación  | 2                   | mar-19                | abr-19 | Se trabajará en la redacción de la introducción y la justificación del tema de investigación.   |
|            | Elaboración del marco teórico   | 2                   | may-19                | jun-19 | Desarrollo de cada uno de los referentes teóricos.  |
|            | Reestructurar los datos   | 1                   | jul-19                | jul-19 | Cada archivo .csv brindado por el micro-espectrofotómetro se debe transformar en datos en un formato adecuado para el análisis de datos funcionales y del FANOVA. |
|            | Análisis de los datos   | 1                   | ago-19                | ago-19 | Analizar los datos aportados por el CICIMA con técnicas estadísticas pertinentes.   |
|            | Estudio de simulación   | 3                   | set-19                | nov-19 | Evaluar los resultados obtenidos en los escenarios de simulación y la redacción del capítulo correspondiente.   |
|            | Redacción de las conclusiones   | 1                   | dic-19                | dic-19 | Redacción de las conclusiones, limitaciones y de las recomendaciones de la investigación.   |

**Artículo 15.** Para el Programa de Maestría Académica el periodo máximo entre el ingreso del estudiante a la segunda etapa (fecha del primer curso matriculado) y la presentación del examen de candidatura es de cuatro años. Si no lo aprobara en ese periodo quedará automáticamente fuera del Programa, pudiendo solicitar a la Comisión traslado al Programa de la Maestría Profesional en Estadística. El estudiante tendrá un plazo de tres ciclos lectivos para completar la tercera etapa, a partir de la fecha de aprobación del examen de candidatura. Si al cabo de este periodo el estudiante no ha presentado la tesis, la Comisión podrá conceder una única prórroga de un ciclo lectivo, al cabo del cual, el estudiante que no ha defendido su tesis será separado del programa.

**FIRMA DEL ESTUDIANTE**