ESTADÍSTICA NO PARAMÉTRICA EXAMEN FINAL 24 DE MAYO DE 2020

Nombre David Plazas Escudero

Código 201710005101

El código para dar respuestas a las preguntas de este examen puede encontrarse aquí.

1. a)

Enunciado: *Grafique en un mismo plano las funciones de distribución empíricas de los tiempos de vida de los 13 años. Explique con base en las gráficas de las funciones, si es observable un efecto de mejoría en los tiempos de vida a través del tiempo. Interprete los intervalos donde la función del primer año es mayor a la del último.*

Solución: La gráfica de las distribuciones empíricas (ECDFs) para los tiempos de vida de los 991 dispositivos, durante 13 años, se muestra en la Figura 1. En esta figura se puede observar que efectivamente hay una mejoría en los tiempos de vida de los dispositivos. Es claro la ECDF de los años más recientes está siempre por debajo de los primeros años; esto implica que hay un mayor chance de que los dispositivos recientes tengan mayor tiempo de vida. Sea X el tiempo de vida de un dispositivo del año más reciente y Y del año más antiguo, se sigue que $\hat{F}_X(t) \leq \hat{F}_Y(t)$, luego se puede afirmar que X es mejor a Y.

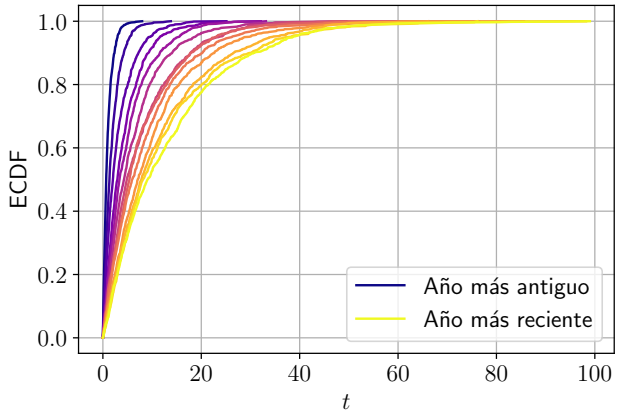

Figura 1: ECDFs de tiempos de vida.

1. b)

Enunciado: *Calcule y grafique las bandas de confianza a un 90 % de confianza para la función de distribución empírica del primer año.*

Solución: Las bandas de confianza utilizadas están basadas en la desigualdad de Dvoretzky-Kiefer-Wolfowitz (ver [1, p. 14]). Las bandas se presentan en la Figura 2.

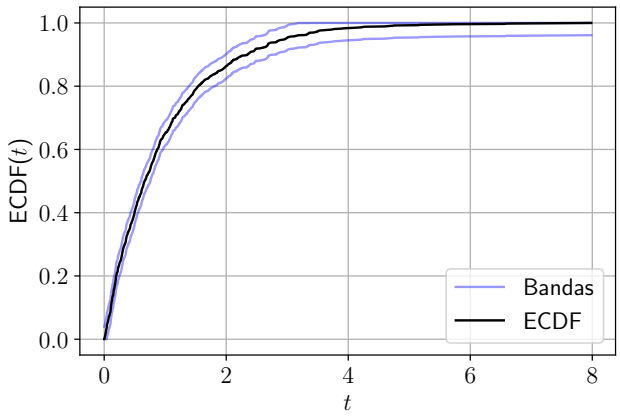

Figura 2: Bandas de confianza para la ECDF del primer año con $\alpha = 0{,}10$.

1. c)

Enunciado: *Grafique en un mismo plano la distribución empírica del último año y la distribución teórica de una exponencial con media 13. ¿Considera usted que hay un buen ajuste? Haga lo mismo con las densidad estimada y la teórica. Determine computacionalmente si se cumple en este par de distribuciones el Teorema de Glivenko-Cantelli.*

Solución: Primero, se muestra la gráfica de la densidad de la distribución exponencial y la densidad estimada a partir de los datos, en la Figura 3. Esta estimación se realizó utilizando el kernel de Epanechnikov con un ancho de banda de 2.

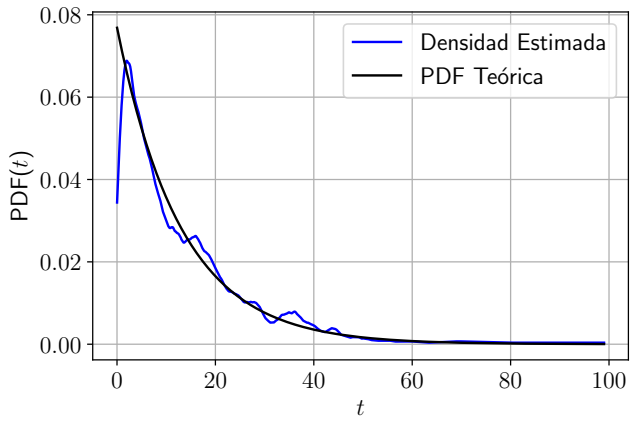

Figura 3: Estimación vs. teórica para la densidad.

Adicionalmente, la gráfica de la ECDF y la distribución exponencial se muestra en la Figura 4.

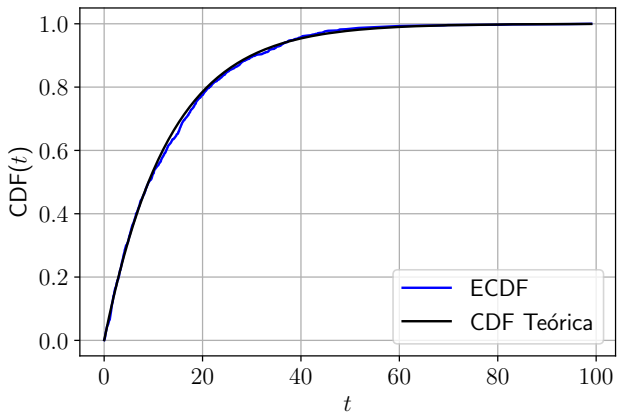

Figura 4: Estimación vs. teórica para la acumulada.

Claramente se observa que ambas distribuciones de ajustan casi perfectamente, lo que muestra buena evidencia de que se cumple el teorema de Glivenko-Cantelli. Este teorema afirma que el supremo de la diferencia entre la distribución empírica y la real converge casi seguramente a 0 cuando el tamaño de la muestra se aumenta [1, p. 14]. Para evidenciar este efecto de una mejor manera, en la Figura 5 se presenta los supremos de estas diferencias para diferentes tamaños de muestra. Claramente se evidencia que este supremo disminuye a medida que se aumenta el tamaño de la muestra para la ECDF.

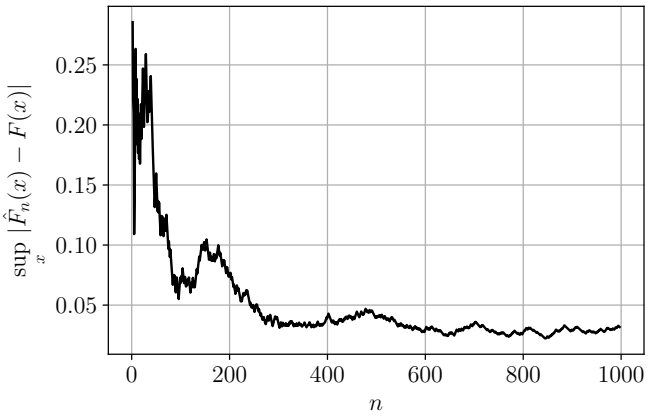

Figura 5: Supremo de las diferencias para diferentes tamaños de muestra.

1. d)

Enunciado: *Considere $X_1, X_2, \ldots$ una sucesión de variables aleatorias. Pruebe que $\{X_n\}_{n\in\mathbb{N}}$ converge en probabilidad a b si*

$$\lim_{n\to\infty} E(X_n) = b \text{ y } \lim_{n\to\infty} V(X_n) = 0$$

Solución: Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesion de variables aleatorias que satisfacen que

$$\lim_{n\to\infty} E(X_n) = b \text{ y } \lim_{n\to\infty} V(X_n) = 0. \tag{1}$$

Se puede ver que la convergencia en probabilidad está implicada por la convergencia en error cuadrático medio (MSE). Recordemos que una sucesión $\{Y_n\}_{n\in\mathbb{N}}$ converge en MSE a una variable aleatoria Y si satisface

$$\lim_{n\to\infty} E\left[(Y_n - Y)^2\right] = 0.$$

Cont.

Adicionalmente se sabe que $E\left[(Y_n - Y)^2\right] = V(Y_n) + \text{Bias}^2(Y_n, Y)$. Calculemos el sesgo para nuestro caso de $Y_n = X_n$ y $Y = X = b$:

$$\text{Bias}^2(X_n, X) = [E(X_n) - b]^2 = E^2(X_n) - 2bE(X_n) + b^2.$$

Luego,

$$E\left[(X_n - X)^2\right] = V(X_n) + E^2(X_n) - 2bE(X_n) + b^2$$

tomando límite a ambas partes se tiene

$$\begin{aligned}\lim_{n\to\infty} E\left[(X_n - X)^2\right] &= \lim_{n\to\infty}\left[V(X_n) + E^2(X_n) - 2bE(X_n) + b^2\right] \\ &= \lim_{n\to\infty} V(X_n) + \lim_{n\to\infty} E^2(X_n) - 2b \lim_{n\to\infty} E(X_n) + b^2 \\ &= 0 + b^2 - 2b^2 + b^2 = 0\end{aligned}$$

Luego, $\{X_n\}_{n\in\mathbb{N}}$ converge en MSE, y por lo tanto en probabilidad, a b.

1. e)

Enunciado: *Considere los tiempos de vida del primer año. Calcule un intervalo de confianza bootstrap para el tiempo de vida máximo. Calcule el sesgo de $T_{[n]}$ y la varianza por el método de Jackknife.*
Solución: Los intervalos de confianza (CI) se muestran en la Tabla 1, estos CI fueron construidos en una remuestreo Bootstrap de 10000 muestras con repetición. El sesgo obtenido por Jackknife es $b_{\text{jack}} = -0{,}66$ y la varianza estimada es $v_{\text{jack}} = 4{,}34$.

Método	**Ci**
Normal	(6.91, 9.06)
Percentile	(7.98, 9.41)
Studentized	(7.98, 12.49)
Pivotal	(7.98, 9.41)

Tabla 1: Intervalos de confianza Bootstrap para el máximo.

1. f)

Enunciado: *La variable costos es el costo en dólares de reparacióon de los dispositivos observados en el primer año. Realice una regresión para intentar explicar el costo de reparación en términos del tiempo de duración del dispositivo.*
Solución: La gráfica de dispersión se presenta en la Figure 6.

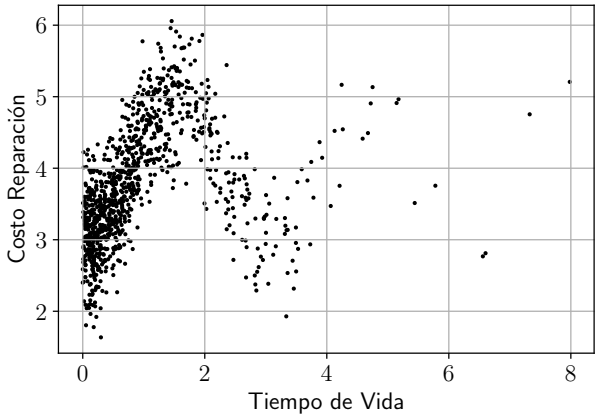

Figura 6: Gráfico de dispersión de costos contra tiempo de vida.

Cont.

Es claro que los datos no siguen una tendencia lineal, pero como en este ejercicio no se especifica qué tipo de regresión, se intentará primero con una regresión lineal basada en mínimos cuadrados ordinarios (OLS). El resultado se presenta en la Figura 7 y se observa que es un modelo deplorable.

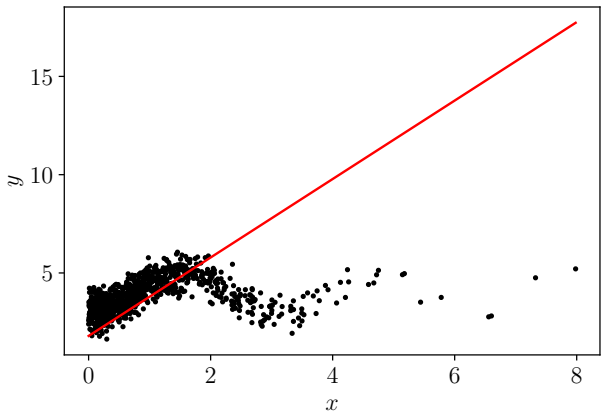

Figura 7: Regresión por OLS.

Alternativamente, se intentará una regresión por kernels de tipo Nadaraya–Watson. La regresión de hizo utilizando el kernel de Epanechnikov con ancho de banda de 0,5 y se pueden observar los resultados en la Figura 8. Esta regresión claramente se ajusta mejor a los datos obtenidos, logrando capturar la tendencia sinusoidal presente.

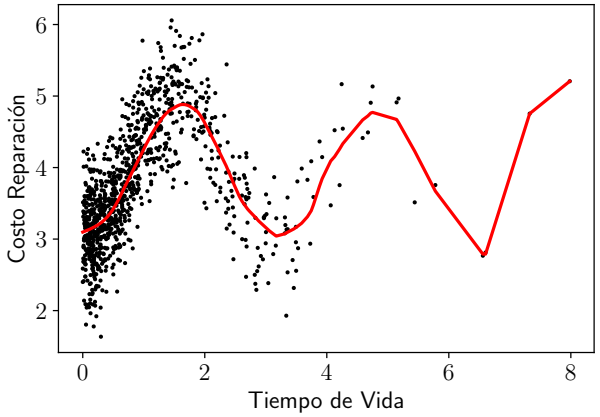

Figura 8: Regresión por Nadaraya-Watson.

1. g)

Enunciado: *Pase un test de rangos para identificar qué años tuvieron una duración en el dispositivo con diferente distribución. ¿Qué conclusión se obtiene?.*
Solución: Para este apartado de realizó un test "ranksum" para cada par de años, confeccionando la matriz presentada a continuación, donde un 0 en la posición (i, j) indica hay evidencia suficiente para afirmar que el año i y el j provienen de la misma distribución. El 1 indica el caso contrario. Este test se realizó con $\alpha = 0{,}05$. Esta matriz nos indica que prácticamente todos los años provienen de distintas distribuciones, con algunas excepciones de años consecutivos.

Cont.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	0	1	1	1	1	1	1	1	1	1	1	1	1
2	1	0	1	1	1	1	1	1	1	1	1	1	1
3	1	1	0	1	1	1	1	1	1	1	1	1	1
4	1	1	1	0	1	1	1	1	1	1	1	1	1
5	1	1	1	1	0	1	1	1	1	1	1	1	1
6	1	1	1	1	1	0	1	1	1	1	1	1	1
7	1	1	1	1	1	1	0	0	1	1	1	1	1
8	1	1	1	1	1	1	0	0	1	1	1	1	1
9	1	1	1	1	1	1	1	1	0	1	1	1	1
10	1	1	1	1	1	1	1	1	1	0	1	1	1
11	1	1	1	1	1	1	1	1	1	1	0	0	1
12	1	1	1	1	1	1	1	1	1	1	0	0	0
13	1	1	1	1	1	1	1	1	1	1	1	0	0

1. h)

Enunciado: *Pase un test de homogeneidad multivariante para identificar si los primeros 400 registros vienen de la misma distribución de los últimos 400. Con el mismo test, verifique si los 5 primeros años vienen de la misma distribución de los 5 últimos. ¿Qué conclusión se puede sacar de acuerdo al resultados de los test.*
Solución: El test de homogeneidad se realizó siguiendo la metodología expuesta en clase. A continuación se hará una breve descripción del proceso:

1. Se tienen dos poblaciones $X \in \mathbb{R}^{n\times p}$ y $Y \in \mathbb{R}^{m\times p}$ p-variadas. Se construye una nueva población $Z \in \mathbb{R}^{(n+m)\times p}$ como la concatenación de X con Y.

2. Para cada $z \in Z$, se calcula la profundidad estadística (Tukey o Mahalanobis) respecto a la masa de datos de X, construyendo el vector de profundidades $Z_X \in \mathbb{R}^{(n+m)}$. De la misma manera se construye $Z_Y \in \mathbb{R}^{(n+m)}$.

3. Se construye una regresión OLS sobre (Z_X, Z_Y). La gráfica de dispersión de (Z_X, Z_Y) es conocida como el D-D plot.

4. Si $\beta_1 = 1$ y $\beta_0 = 0$, se tendrá perfecta dependencia lineal $Z_X = Z_Y$ y se concluirá que los datos provienen de la misma distribución.

La implementación considerada para este punto utiliza la profunidad de Tukey, con 500 direcciones aleatorias. Para la primera parte, se tomaron los primeros y últimos 400 registros. Es claro que se esperaría que estos datos provengan de la misma distribución. El gráfico de dispersión de (Z_X, Z_Y) y la regresión se muestran en la Figura 9. Los resultados del modelo de regresión son: $\beta_1 = 0{,}972$ y $\beta_0 = 0{,}002$. Con lo que se puede concluir que las dos muestras son tomadas de la misma población.

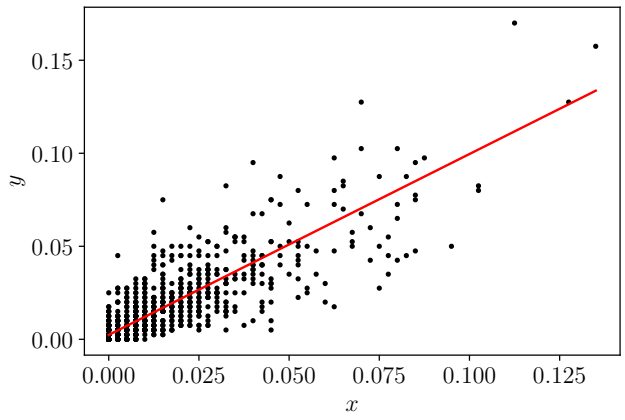

Figura 9: D-D plot para los primeros y últimos 400 registros.

Cont.

Para la segunda parte, la primera población corresponde a los primeros 5 años de registros, y la segunda a los últimos 5. En este caso, se espera que las poblaciones no vengan de la misma distribución, debido a que, como se explicó en el punto 1.a), los tiempos de vida de los primeros años son menores a los de los años más recientes. La Figura 10 muestra los resultados de la dispersión y la regresión ajustada. En este caso, no era necesario hacer el modelo de regresión, pues es evidente que los datos nunca podrían ajustarse a la recta identidad.

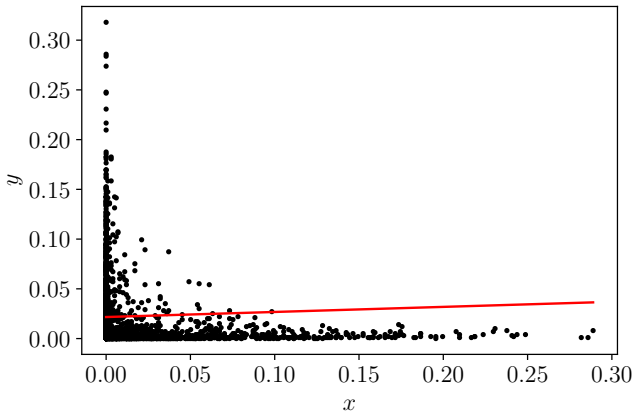

Figura 10: D-D plot para los primeros y últimos 5 años.

1. i)

Enunciado: *Considere* $\beta = [13, 12, 11, \ldots, 3, 2, 1]$. *Suponga que* $y = vida \times \beta^T + \xi$, *Donde* ξ *es un ruido normal de media cero y desviación 50. Realice con varios modelos de regresión lineal no paramétrica y robusta, modificando la matriz de covarianzas por una versión robusta o no paramétrica o combinación de ellas (Kendall, Spearman, fastMCD, máxima kurtosis, shrinkages) para identificar cuál de ellas tiene mejor desempeño. Incluya la regresión por mínimos cuadrados para comparación.*
Solución: Los métodos de regresión robusta están basados en que los β del modelo de regresión pueden ser escritos como $\hat{\beta} = \Sigma_{xx}^{-1}\Sigma_{xy}$, donde se cambia la estimación usual de las covarianzas por estimadores más robustos. En este caso se utilizará la matriz de comedianas, la estimación basada en las correlaciones de Kendall y Spearman, los métodos de fast MCD y shrinkages disponibles en el paquete `sklearn` de `Python`. Finalmente estas regresiones serán comparadas con la estimación usual por mínimos cuadrados ordinarios (OLS). Como no se puede observar directamente la regresión, debido a que se trabaja en el hiperplano 13-dimensional, se evaluará el performance de cada regresión tomando la norma usual de la diferencia entre el vector de betas estimado $\hat{\beta}$ y el vector real de β. Los resultados se presentan en la Tabla 2. Esta tabla muestra que el mejor desempeño fue obtenido por la regresión OLS y el de peor fue el de Fast MCD.

Método	$\|\|\hat{\beta} - \beta^T\|\|_2$
OLS	1.429
Spearman	3.471
Comedian	6.186
Shrinkages	10.080
Kendall	11.147
Fast MCD	21.002

Tabla 2: Comparación entre los métodos para la regresión.

Referencias

[1] Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, 2006.

The End.