Tarea 6 de Modelos Probabilistas Aplicados

Algoritmos generadores de números pseudo-aleatorios con distribución Uniforme y distribución Normal

5271

13 de octubre de 2020

1. Introducción

En este trabajo se presenta un acercamiento al tema pruebas estadísticas, dando respuestas a una serie de preguntas que dan cuerpo a una breve introducción en dicho tema. Además se realizan diversas pruebas estadísticas como ejemplos. Los datos que se utilizan en este trabajo fueros obtenidos en el sitio https://www.inegi.org.mx que pertenece al Instituto Nacional de Estadística y Geografía (INEGI). Se escogió el apartado Construcción donde se muestra información sobre los principales resultados de las Empresas Constructoras, comprende unidades económicas dedicadas principalmente a la edificación; a la construcción de obras de ingeniería civil y a la realización de trabajos especializados de construcción. De este apartado se descargó el tabulado (Valor de producción generado por las empresas constructoras según el tipo de obra) en formato *csv*. Las pruebas se realizan en el programa R versión 4.0.2 [1] en el entorno de desarrollo Rstudio [2]

2. Pruebas estadísticas

En esta sección se realiza un acercamiento a temas importantes relacionados con las pruebas estadísticas, como su características generales, los tipos de pruebas e interpretación de las mismas.

2.1. Relación entre contraste de hipótesis y pruebas estadísticas

Una prueba estadística un procedimiento en el cual se analiza la evidencia proporcionada por los datos con el fin de probar una Hipótesis.La hipótesis estadística es una afirmación sobre los valores del parámetro θ (θ puede ser: μ, p, σ^2, entre otros) de una población o proceso, que es susceptible de probarse a partir de la información contenida en una muestra representativa que es obtenida de la población. Trasladando esto al tema de investigación del autor (Empaquetamiento óptimo), se tiene el siguiente ejemplo, la afirmación "Existe diferencia en el porciento de ocupación de los diferentes tipos de figuras en el contenedor". La veracidad de esta afirmación se obtiene al contrastar las siguientes hipótesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6, \tag{1}$$

$$H_1 : \mu_i \neq \mu_j \quad para \quad alguna \quad i \neq j. \tag{2}$$

Cuadro 1: Influencia del tipo de figura en el % de ocupación del contenedor

	Factor	SS	DF	MS	F	Valor p	np2
0	Tipo de figura	0.0132	5.0000	0.0026	1.0302	0.4147	0.1252
1	Within	0.0920	36.0000	0.0026			

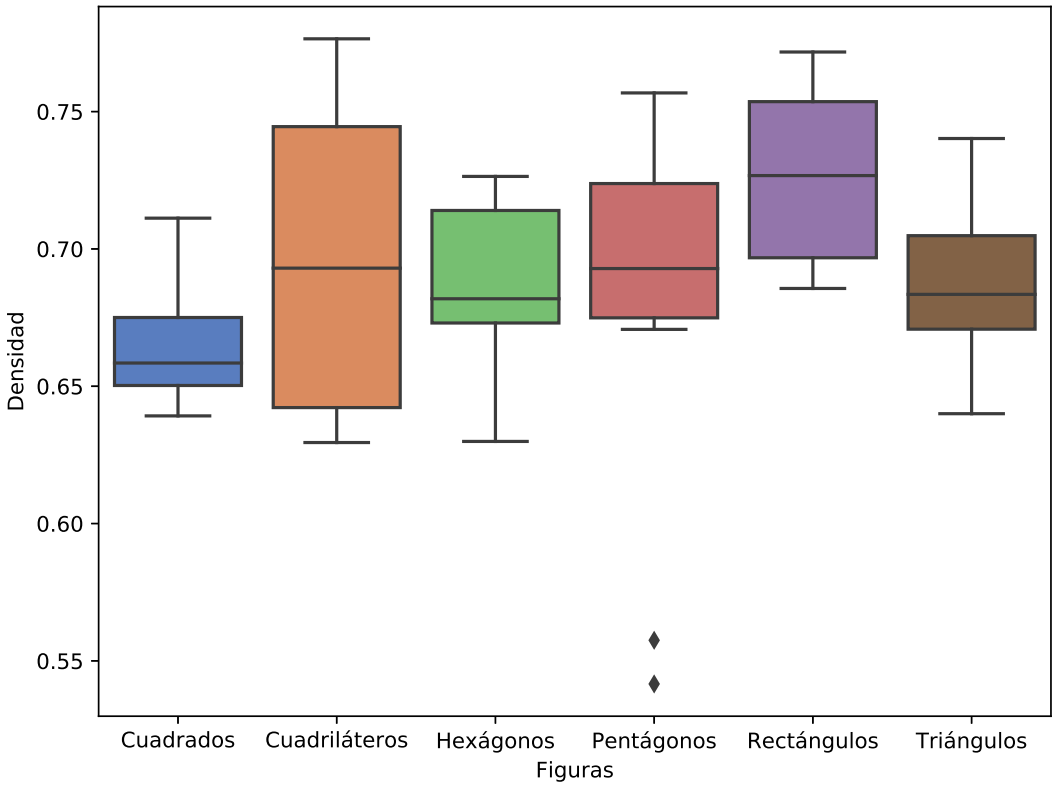


Figura 1: Diagrama de caja y bigotes que relaciona los tipos de figuras y el % de ocupación del contenedor.

A la expresión (13) se le conoce como hipótesis nula y a la expresión (14) se le nombra como hipótesis alternativa. El nombre de hipótesis nula nace de que comúnmente se plantea como una igualdad. Generalmente se trata de probar que la hipótesis nula es verdadera, y que en caso de ser rechazada por la evidencia que aportan los datos, se aceptará la hipótesis alternativa.

En el cuadro 1 de la página 2 se muestra el resultado de la aplicación de un análisis de varianza (ANOVA) al ejemplo antes mencionado. Con los valores del estadístico de prueba $\mathbf{F} = 1.0302$ y el **valor p** $= 0.4117$, se tiene evidencia para aceptar la hipótesis H_0, la cual indica que no existe diferencias estadísticas significativas entre los tratamientos con un intervalo de confianza del 95 %. Esto se puede observar en la figura 1 de la de la página 2.

Cuadro 2: Influencia de la cantidad de figura en el % de ocupación del contenedor

	Factor	SS	DF	MS	F	Valor p	np2
0	Cantidad defiguras	0.0429	5.0000	0.0086	4.9514	0.0015	0.4075
1	Within	0.0623	36.0000	0.0017	–	–	–

2.2. En que casos se rechaza la hipótesis nula

Para probar una hipótesis se trata de corroborar que la afirmación planteada por la hipótesis nula (H_0) es verdad o no. Se parte del supuesto que es verdadera, si la evidencia arrojada por los datos es suficiente para contradecir dicho supuesto se rechaza H_0 y se acepta la Hipótesis alternativa (H_1). En caso de no haber evidencias suficientes que demuestren la falsedad de la H_0 esta no se rechaza, es decir H_0 es verdad hasta que no se demuestre lo contrario.

Para discernir si H_0 se rechaza o no, se utiliza un estadístico de prueba el cual es un es un número calculado a partir de los datos y la hipótesis nula. Al conjunto de posibles valores del estadístico de prueba que llevan a rechazar H_0, se le llama región o intervalo de rechazo para la prueba, y a los posibles valores donde no se rechaza H_0 se les conoce como región o intervalo de confianza.

El estadístico de prueba, construido bajo el supuesto de que H_0 es verdad, es una variable aleatoria con distribución conocida. Si efectivamente H_0 es verdad, el valor del estadístico de prueba debería caer dentro del rango de valores más probables de su distribución asociada, el cual se conoce como región de confianza. Si cae en una de las colas de su distribución asociada, fuera del rango de valores más probables (en la región de rechazo), es evidencia en contra de que este valor pertenece a dicha distribución. De aquí se deduce que debe estar mal el supuesto bajo el cual se construyó, es decir, H0 debe ser falsa [].

Análogamente al análisis del ejemplo de la sección 2.1, se realiza la misma prueba para el factor de control cantidad de figuras. Obtenemos como resultado el cuadro 2 de la página 3. Con los valores del estadístico de prueba $\mathbf{F} = 4.9514$ y con un **valor p** $= 0.0015$ menor que 0.05, se tiene evidencia para rechazar la hipótesis H_0 ya que existe diferencias estadísticas significativas al menos entre algunos de los grupos con un intervalo de confianza de un **95 %** como se muestra en la figura 2 de la página 4.

2.3. Interpretación de la salida de una prueba estadística

En la sección 2.1 y 2.2 muestran ejemplos de como interpretar la salidas de las pruebas estadísticas a través de los estadísticos de prueba explicados en la sección 2.2 y el **valor p**, representa una probabilidad que mide la evidencia en contra de la hipótesis nula. **valor p** más pequeño proporciona una evidencia más fuerte en contra de la hipótesis nula.

Para determinar si la diferencia entre las desviaciones estándar o las varianzas de las poblaciones es estadísticamente significativa, se compara el **valor p** con el nivel de significancia. Por lo general, el nivel de significancia (denotado como α) indica el riesgo de concluir que existe una diferencia cuando realmente no la hay.

- Sí el **valor p** $> \alpha$, la relación de las desviaciones estándar o las varianzas no es estadísticamente significativa (no puede rechazar H_0).
- Sí el **valor p** $\leq \alpha$, la relación de las desviaciones estándar o las varianzas es estadísticamente significativa (se rechaza H_0).

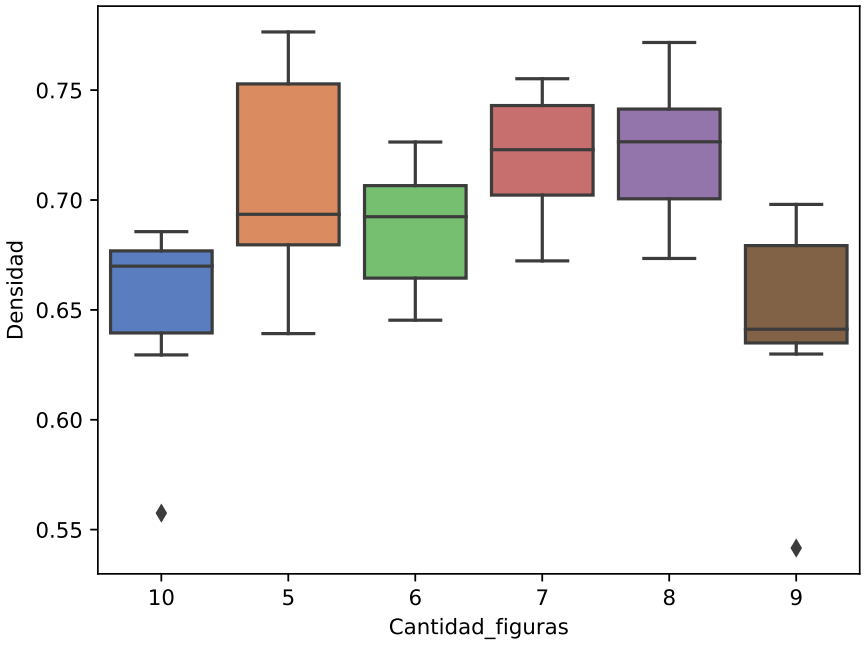


Figura 2: Diagrama de caja y bigotes que relaciona los cantidad de figuras y el % de ocupación del contenedor.

2.4. Selección de nivel de significancia

El valor de α es el máximo nivel de riesgo aceptable de rechazar la hipótesis nula cuando la hipótesis nula es verdadera, como se comento en la sección 2.3. Por lo general, se elige elige el nivel de significancia antes de analizar los datos. El este nivel se calcula como 1 menos el intervalo de confianza. Para los ejemplos anteriores se empleo un intervalo de confianza del 95 %, por lo que α es igual a 0.05.

Los valores de α son elegidos en dependencia de la finalidad de la prueba, es decir si se quiere determinar cualquier diferencia que exista se utiliza un valor de α mas grande (0.10). En el caso que se quiera asegurar que el las diferencias que existen son reales, se emplea un valor de α muy pequeño (0.01). Comúnmente se utiliza el valor de α igual a 0.05.

2.5. Errores frecuentes de interpretación del valor p

Probar una hipótesis estadística es una decisión probabilística, por lo que existe el riesgo de cometer un error **tipo I** o un error **tipo II**. El primero ocurre cuando se rechaza H_0 cuando ésta es verdadera, y el error **tipo II** es cuando se acepta H_0 y ésta es falsa. Con α y β se denotan las probabilidades de los errores **tipo I** y **tipo II**, respectivamente.

Donde α es el nivel de significancia explicado en la sección 2.4. Y β es la llamada potencia de prueba.

Cuadro 3: Prueba Shapiro-Wilk a la variable % de ocupación

	W	**Valor p**
% de ocupación	0.9537	0.0877

2.6. Potencia de prueba, utilidad

A $1 - \beta$ se le llama potencia de la prueba, y es la probabilidad de rechazar H_0 cuando es falsa. Por lo general, en las pruebas de hipótesis se especifica el valor de α y se diseña la prueba de tal forma que el valor de β sea pequeño. Esto es, la probabilidad del error **tipo I** se controla directamente, mientras que la probabilidad de error **tipo II** se controla de manera indirecta con el tamaño de la muestra, ya que a más datos β será menor. Es decir, con una muestra grande es mayor la potencia de la prueba, de manera que se incrementa la probabilidad de rechazar H_0 si ésta es falsa.

2.7. Pruebas paramétricas

Para la aplicación de las pruebas paramétricas, los datos deben cumplir ciertas presunciones:

- Los datos son de escala de intervalo o razón.
- La población de la muestra debe aproximarse a una distribución normal.
- Las varianzas de las muestras deben aproximadamente similares.
- Las observaciones deben ser independientes entre sí.

Entre las pruebas perimétricas más usadas están:

- Prueba de Shapiro-Wilks.
- Prueba de Fisher.
- Prueba t de Student para muestras independientes.
- Coeficiente de correlación de Pearson.
- Regresión lineal.
- Análisis de varianza factorial (ANOVA).
- Análisis de covarianza (ANCOVA).

En ambos ejemplos presentados en la sección 2.1 y 2.2 Antes de realizar el ANOVA, se realiza la prueba de Shapiro-Wilk, que calcula un W estadístico que prueba si una muestra aleatoria $x_1, x_2, ..., x_n$ proviene de una distribución normal. El resultado de esta prueba se muestra en el cuadro 3 de la página 5. En dicho cuadro con un $\mathbf{W} = 0.9537$ y un **valor p** $= 0.0877$ mayor que 0.05, se puede observar que la variable dependiente (% de ocupación) sigue una distribución normal con un intervalo de confianza del 95 %.

Cuadro 4: Prueba Shapiro-Wilk a la variable % de ocupación variando el contenedor

	W	Valor p	Normal
Densidad	0.9527	0.0037	Falso

Cuadro 5: Prueba H de Kruskal-Wallis que relaciona el % de ocupación con el tipo de contenedor

	Factor	ddof1	H	pval
Kruskal	Contenedor	1.0000	13.9167	0.0002

2.8. Pruebas no paramétricas

En la sección 2.7 se plantea características que deben tener los datos para usar las pruebas paramétricas. Cuando los datos no cumplen una de estas características. Se emplean las pruebas no paramétricas, ejemplos de estas son:

- Prueba Chi-cuadrada.
- Coeficientes de correlación e independencia para tabulaciones cruzadas.
- Coeficientes de correlación por rangos ordenados Spearman y Kendall.
- Prueba H de Kruskal-Wallis.

Como en los ejemplos anteriores se realiza una prueba de Shapiro-Wilk para corroborar que la variable dependiente (% de ocupación) variando el tipo de contenedor, sigue una distribución normal. Los resultados obtenidos con un **W** = 0.9528 y **valor p**=0.0038 menor que 0.05, nos muestra que dicha variable no sigue una distribución normal con un intervalo de confianza del **95 %**. Lo antes expuestos se puede observar en la cuadro 4 de la página 6.

Dado que los valores de la variable dependiente no sigue una distribución normal, se aplica la prueba H de Kruskal-Wallis que es una versión no paramétrica de ANOVA. En este caso se plantea la afirmación: "Existen diferencias en el % de ocupación según el tipo de contenedor".

Según los datos obtenidos en la prueba que se muestran en el cuadro 5 de la página 6. Estos resultados con un valor del estadístico de prueba **H** = 13.9167 y **valor p** = 0.0002, se rechaza la hipótesis H_0 aceptando la H_1, la cual indica que existen diferencias estadísticamente significativas entre los tipos de contenedores. En esta ocasión con los resultados de la prueba H de Kruskal-Wallis son suficientes para asegurar lo anterior dado que solo son dos los tipos de contenedores analizados. Gráficamente se puede observar esta conclusión en la figura 3 de la página 7.

3. Aplicación de Pruebas estadísticas

En esta sección se presentaran diversas pruebas estadísticas aplicas a los datos del sector de la construcción obtenidos de INEGI.

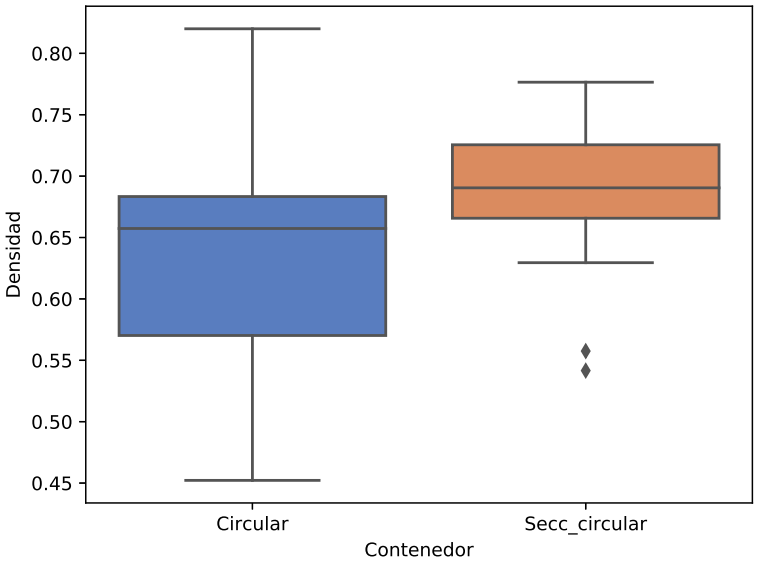


Figura 3: Diagrama de caja y bigotes que relaciona los tipos de contenedor y el % de ocupación del mismo.

Cuadro 6: Resultados de la prueba de Shapiro-Wilk

Tipo de obra	W	Valor p	
Agua, riego y saneamiento	0.9907	0.3223	$> \alpha$
Petróleo y petroquímica	0.9659	0.0002	$< \alpha$
Otras construcciones	0.9451	0.0000	$< \alpha$

3.1. Prueba de Shapiro-Wilk

Como se explica en la sección 2.7 la prueba de Shapiro-Wilk se utiliza para probar si los datos siguen una distribución normal. Por lo cual esta prueba es la primera que se le realiza a los datos para conocer la naturaleza de los mismo y así elegir el tipo de prueba a utilizar. Para el caso de ejemplo se tomara los valores de producción de las constructoras según el tipo de obras. En el cuadro 6 de la página 7 se muestra obtenidos en las pruebas. Además la figura 4 de la página 8 muestra como se comportan los datos según el tipo de obra.

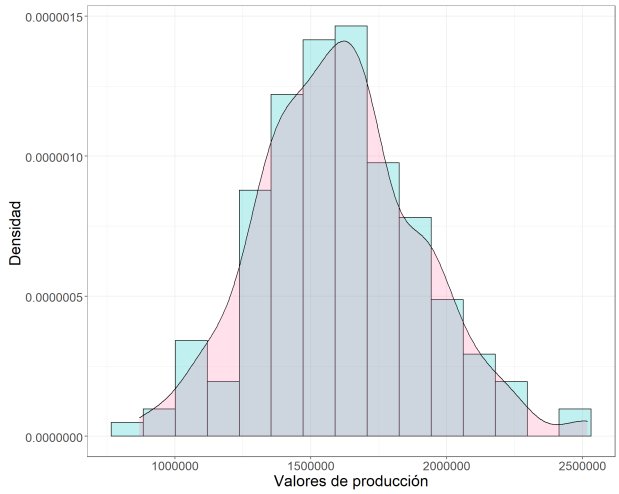
Las pruebas se realizan con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : \text{La muestra sigue un distribución igual a la normal,} \tag{3}$$

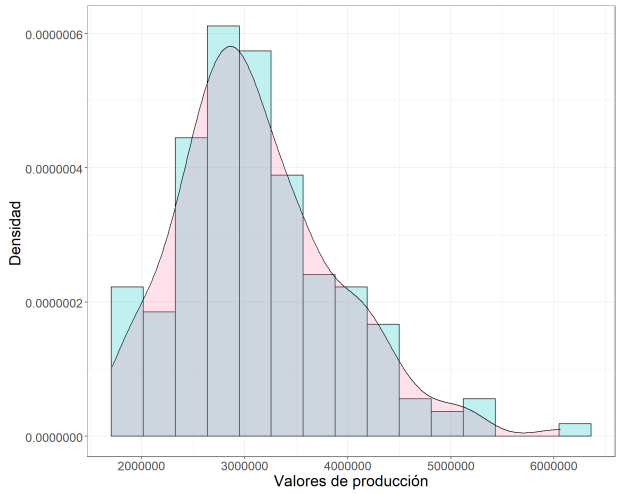
$$H_1 : \text{La muestra sigue un distribución diferente a la normal.} \tag{4}$$

3.2. Prueba t de una muestra

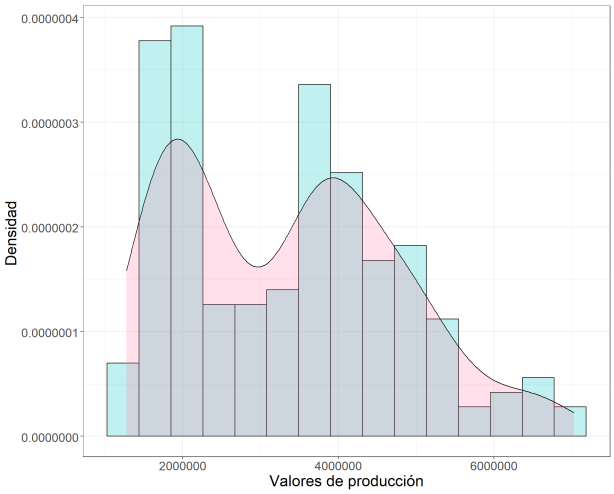
la prueba t de una muestra, es una prueba paramétrica, se utiliza para probar si la media de una muestra puede ser un valor especifico. Al ser paramétrica como premisa necesita que la muestra siga una distribución normal. De los resultados en el cuadro 6 de la página 7, podemos concluir que solo



(a) Agua, riego y saneamiento



(b) Petróleo y petroquímica



(c) Otras construcciones

Figura 4: Histogramas de distribución de los datos por tipo de obras

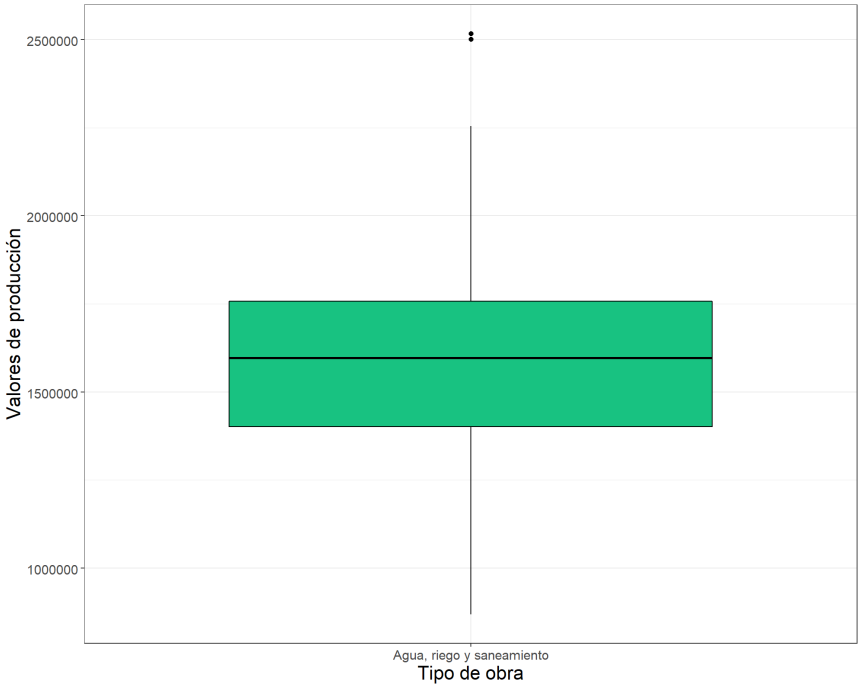


Figura 5: Diagrama de caja y bigotes del tipo de obra Agua, riego y saneamiento

podemos aplicar la prueba t a los datos del tipo de obras Agua, riego y saneamiento, por ser los únicos que siguen una distribución normal.

La prueba se realiza con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : \text{La } \mu \text{ de los de los valores de producción es } 1600000, \tag{5}$$

$$H_1 : \text{La } \mu \text{ de los de los valores de producción diferente } 1600000. \tag{6}$$

Los resultados obtenidos en la prueba, con un estadístico $\mathbf{t} = 0.3014$ y un **valor p** $= 0.7635$ mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto la media de los valores de producción es 1600000 con un intervalo de confianza de 95 %. Esto se puede observar en la figura 5 de la página 9.

3.3. Prueba de rango con signo de Wilcoxon

La prueba de rango con signo de Wilcoxon puede verse como una alternativa a la prueba t, dado que su objetivo es determinar si la media de una muestra es un valor especifico sin tener en cuenta el supuesto que la muestra sigue una distribución normal.

Para la prueba se utilizan los datos pertenecientes al tipo de obra Petróleo y petroquímica, debido a que estos no siguen una distribución normal. La prueba se realiza con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : \text{La } \mu \text{ de los de los valores de producción es } 300000, \tag{7}$$

$$H_1 : \text{La } \mu \text{ de los de los valores de producción diferente } 3000000. \tag{8}$$

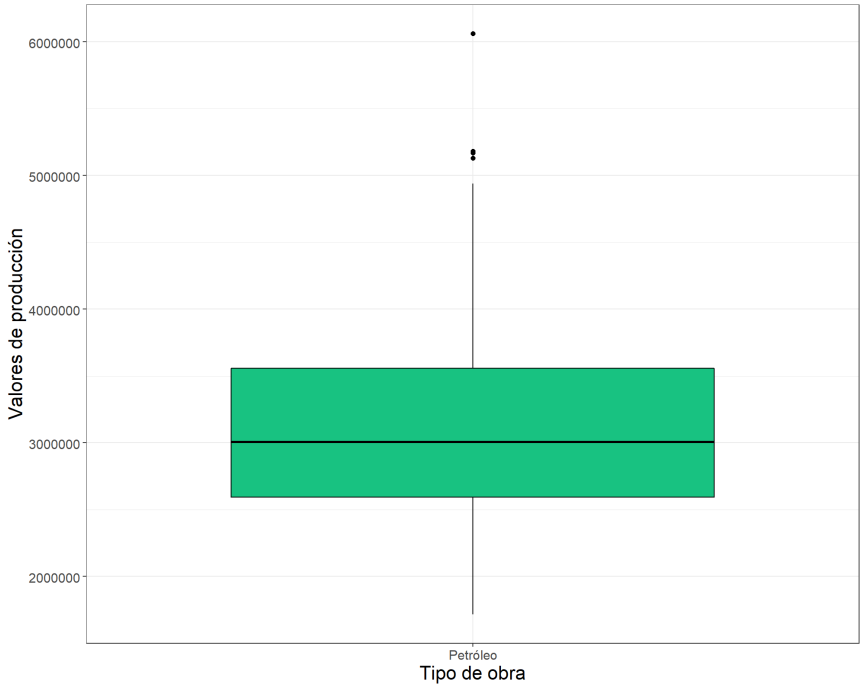


Figura 6: Diagrama de caja y bigotes del tipo de obra Petróleo y petroquímica

Los resultados obtenidos en la prueba, con un estadístico **V** = 8399 y un **valor p** = 0.2375 mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto la media de los valores de producción de las obras Petróleo y petroquímica es 3000000 con un intervalo de confianza de 95 %. Esto se puede observar en la figura 6 de la página 10.

3.4. Prueba t de dos muestras de rangos de Wilcoxon

La Prueba t de dos muestras de rangos de Wilcoxon tiene como objetivo comparar las medias de dos muestras que no siguen una distribución normal.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Petróleo y petroquímica y Otras construcciones, debido a que estos no siguen una distribución normal. La prueba se realiza con un α = 0.05, y la hipótesis planteada es la siguiente:

$$H_0 : \mu\,(\text{Petróleo y petroquímica}) = \mu\,(\text{Otras construcciones}), \tag{9}$$

$$H_1 : \mu\,(\text{Petróleo y petroquímica}) \neq \mu\,(\text{Otras construcciones}). \tag{10}$$

Los resultados obtenidos en la prueba, con un estadístico **W** = 14364 y un **valor p** = 0.7954 mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto la media de los valores de producción de las obras Petróleo y Otras construcciones tienen la misma media con un intervalo de confianza de 95 %. Esto se puede observar en la figura 7 de la página 11.

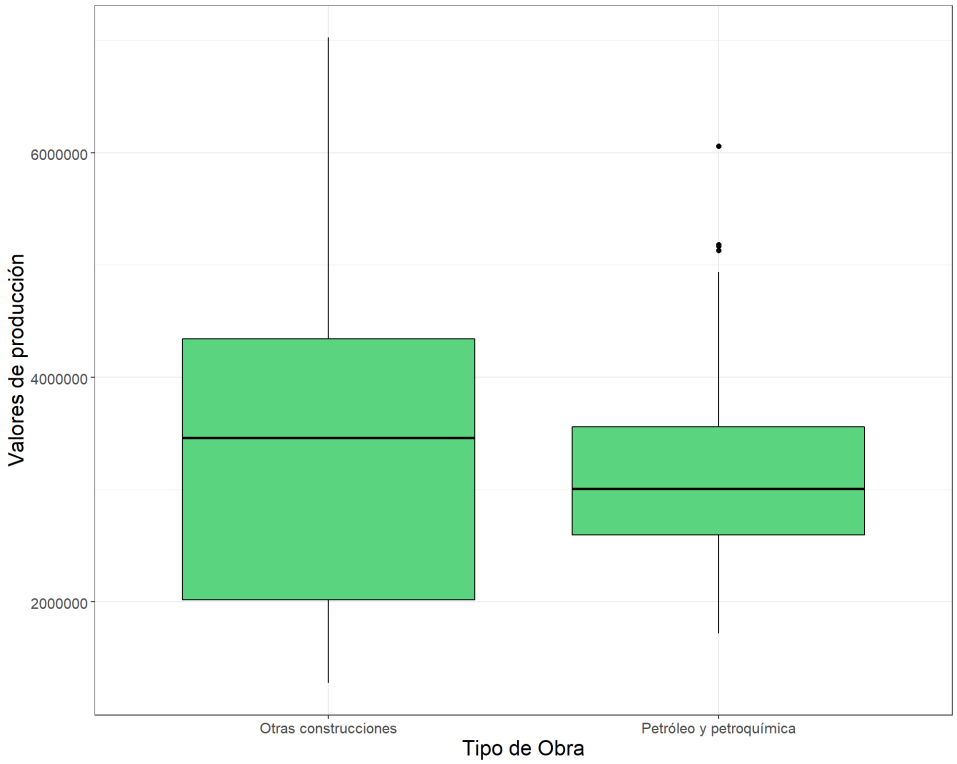


Figura 7: Diagrama de caja y bigotes del tipo de obra Petróleo y petroquímica y Otras construcciones

3.5. Prueba de Kolmogorov y Smirnov

La prueba de Kolmogorov-Smirnov se utiliza para comprobar si dos muestras siguen la misma distribución.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Agua, riego y saneamiento y datos creados con la función *rnorm* de R. Con el objetivo de verificar los resultados de la prueba Shapiro-Wilk realizada en la 3.2. La prueba se realiza con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : (\text{la distribución de Agua, riego y saneamiento}) = (\text{la distribución } rnorm), \tag{11}$$

$$H_1 : (\text{la distribución de Agua, riego y saneamiento}) \neq (\text{la distribución } rnorm). \tag{12}$$

Los resultados obtenidos en la prueba, con una distancia de Kolmogorov-Smirnov $\mathbf{D} = 0.086207$ y un **valor p** $= 0.5375$ mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto los valores de producción de tipo de obra Agua, riego y saneamiento siguen una distribución normal con un intervalo de confianza de 95 %. Esto se puede observar en la figura 8 de la página 13.

3.6. Prueba F de Fisher

La prueba F de Fisher se puede utilizar para comprobar que dos muestras tienen la misma varianza.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Agua, riego y saneamiento y datos creados con la función *rnorm* de R. Con el objetivo de verificar si ambos poseen la misma varianza. La prueba se realiza con un $\alpha = 0.05$, y la hipótesis planteada es la siguiente:

$$H_0 : (\text{la varianza de Agua, riego y saneamiento}) = (\text{la varianza } rnorm), \tag{13}$$

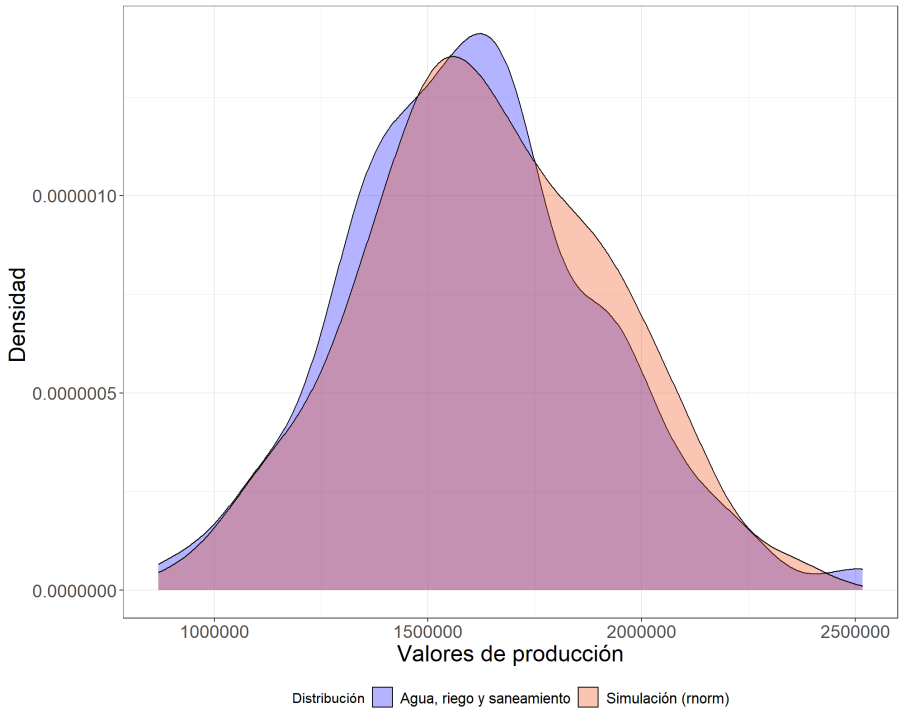
$$H_1 : (\text{la varianza de Agua, riego y saneamiento}) \neq (\text{la varianza } rnorm). \tag{14}$$

Los resultados obtenidos en la prueba, con un estadístico $\mathbf{f} = 1.0442$ y un **valor p** $= 0.7766$ mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto los valores de producción de tipo de obra Agua, riego y saneamiento y los valores creados con *rnorm* tiene la misma varianza como era de esperarse, con un intervalo de confianza de 95 %.

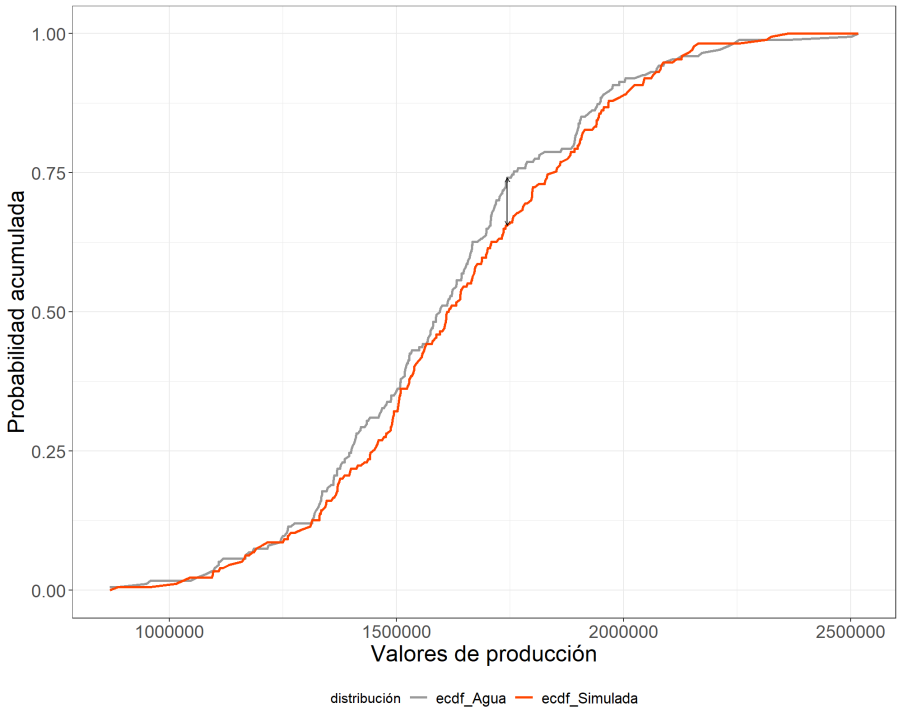
3.7. Prueba de Chi-cuadrada

La prueba de chi-cuadrado en R se puede utilizar para probar si dos variables son dependientes.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Agua, riego y saneamiento. Con el objetivo de verificar estos son independientes con respecto al periodo en que fueron registrados. La prueba se realiza con un $\alpha = 0.05$. Los resultados obtenidos en la prueba, con un estadístico $X^2 = 2088$ y un **valor p** $= 0.4222$ mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo tanto los valores de producción de tipo de obra Agua, riego y saneamiento son valores independientes con respecto al periodo en que fueron registrados,con un intervalo de confianza de 95 %.



(a) Diagrama de densidad de Agua, riego y saneamiento superpuesto al simulado con *rnorm*



(b) [Diferencia de Kolmogorov-Smirnov entre Agua, riego y saneamiento superpuesto y la simulación con *rnorm*

Figura 8: Prueba de Kolmogorov-Smirnov

3.8. Correlación

La correlación nos indica la relación lineal de dos variables continuas.

Para esta prueba se utilizan los datos pertenecientes al tipo de obra Petróleo y petroquímica y Otras construcciones. La prueba se realiza con un $\alpha = 0.05$.Los resultados obtenidos en la prueba, con un estadístico $\mathbf{t} = 2.4082$ y un **valor p** $= 0.05709$ mayor que α, se puede concluir que no hay evidencias para rechazar la hipótesis H_0, por lo la correlación es cero,con un intervalo de confianza de 95 %.

El código general se encuentra disponible en el repositorio. https://github.com/Albertomnoa/Tareas

Referencias

[1] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

[2] RStudio Team. Rstudio: Entorno de desarrollo integrado para R, 2020.