Tarea 1

Autor

Sebastián Andrés Mena Aliaga

Profesor

Christian Araya Muñoz

Modelación Estadística Aplicaciones Multidisciplinaria

Magister en Estadística, PUCV

Agosto 2020

1 Interpretación

1.1 Introducción

El presente estudio tiene como objetivo generar información valiosa de una muestra de datos. Cada registro representa en la base representa los tiempos que tarda un camiones de carga en pasar por diversos procesos dentro de un puerto en Valparíso.

La información se obtiene, inicialmente, a partir de un Análisis Exploratorio de Datos (AED), con el objetivo de entender los rasgos generales de los datos. Posteriormente se realizará inferencia a partir de estos datos, con la finalidad de obtener conclusiones de ciertas preguntas que se plantean en más adelante en la actividad.

1.2 Diccionario de Datos

A continuación se muestran los descriptivos de las principales variables utilizadas en el *script* de R y, en consecuencia, el nombre utilizado para identificar las variables dentro de la presente tarea:

- delta_1 : Tiempo en minutos desde que el camión entra al recinto, hasta que el camión se posiciona en el pórtico de control de acceso.
- delta_2 : Minutos desde que el camión entra y sale del pórtico de control de acceso.
- delta_3 : Minutos de espera y atención en la oficina 1.
- delta_4 : Minutos de atención en la inspección física del cargamento.
- delta_5 : Minutos que tarda el camión en salir del reciente.
- tipo_carga : Tipo de cargamento que transporta el camión.
- turno : Turno en cual ingresa el camión.
- respon_2 : Responsable en registrar tiempos de delta_2.
- respon_3 : Responsable en registrar tiempos de delta_3.
- respon_4 : Responsable en registrar tiempos de delta_4.

2 Pregunta 1

Tal como se señaló en el tópico Diccionario de Datos, cada variable de tiempo de estudio se denota como *delta_n*. En primer lugar, se desarrollará un Análisis Exploratorio de Datos (AED) para cada delta, analizando algún comportamiento de interés, como la presencia de datos atípicos y el análisis de la distribución de cada deltas.

Cabe destacar que los *outliers* se han considerado como aquellas observaciones que se encuentran fuera de 1,5 veces el rango intercuartil (IQR).

En la figura 1 se observan los *boxplot* de cada *delta* analizado en el puerto. Se han removiendo los *outliers* del gráfico.

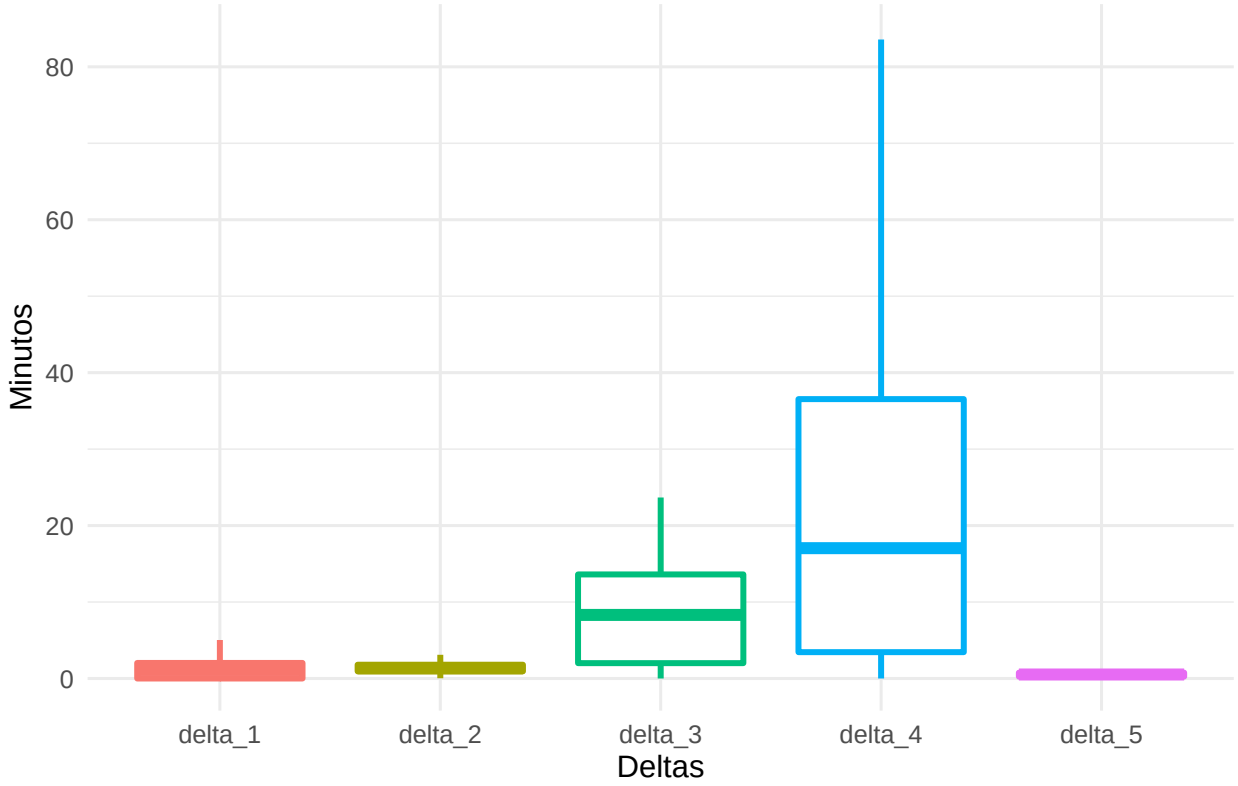

Figure 1: Tiempos, en minutos, de cada procesos que pasa un camión (*deltas*)

2.1 AED sobre delta_1

Sobre *delta_1* se presencian fuertes registros de datos atípicos (*outliers*). Dentro de los *outliers* se observan 471 valores NA que se han sido sustituidos por 0. Además, se observa el máximo valor registrado de 415,250 minutos, valor que se aleja demasiado de la media y la mediana, 1.125 y 3.178 respectivamente (ver figura 2). Estos valores atípicos podrían deberse a errores de registro en cualquier punto del proceso de investigación o comportamientos extraños de los camiones. En total, *delta_1* cuenta con 89 *outliers*.

Para efectos prácticos y visuales, el histograma de la figura 2 se ha redimensionado en

su eje x con límites entre 0 y 7.5. Gráficamente se puede observar que los datos no poseen comportamiento normal. Se realiza el test de Kilmogorov-Smirnov (K-S) con una confianza del 95%; los resultados entregan un estadístico $D = 0.419$ y un $p - valor < 2.2x10^{-16}$, por lo tanto, se rechaza la hipótesis nula del test, descartando un comportamiento normal de la muestra.

Min	0.000
1er Qu.	0.000
Mediana	1.125
Media	3.178
3er Qu.	2.061
Max	414.25
Desv.	15.593
CV	4.907

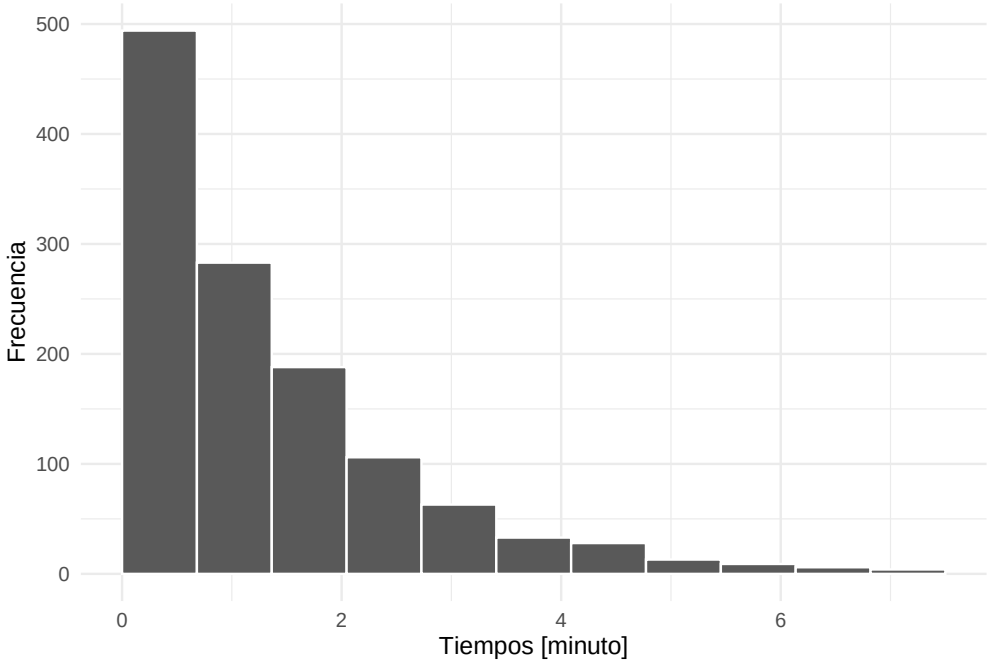

Figure 2: Resumen estadístico e histograma de delta_1

2.2 AED sobre delta_2

Respecto *delta_2*, en uno de sus registros indica un valor de -0,417 minutos, valor inconsistente con la realidad física del tiempo, sin embargo, entendiéndose de una posible error de anotación, se ha invertido el valor de la resta de entrada y salida con la finalidad de contar con un valor positivo. Además, *delta_2* cuenta con 69 *outliers*, siendo 239,783 minutos el valor máximo registrado, nuevamente alejándose de su mediana y media, 1,333 y 1,840 minutos respectivamente.

El histograma de la figura 3 se ha redimensionado en su eje x con límites entre 0 y 5. Gráficamente se puede observar un posible comportamiento normal con asimetría hacía la derecha, sin embargo, esta hipótesis se desmorona desde un sustento estadístico. El test de Kilmogorov-Smirnov (K-S) con una confianza del 95%; los resultados entregan un estadístico $D = 0.4$ y un $p - valor < 2.2x10^{-16}$, por lo tanto, se rechaza la hipótesis nula del test, descartando un comportamiento normal de la muestra.

Min	0.033
1er Qu.	0.933
Mediana	1.333
Media	1.840
3er Qu.	1.812
Max	238.783
Desv.	7.093
CV	3.856

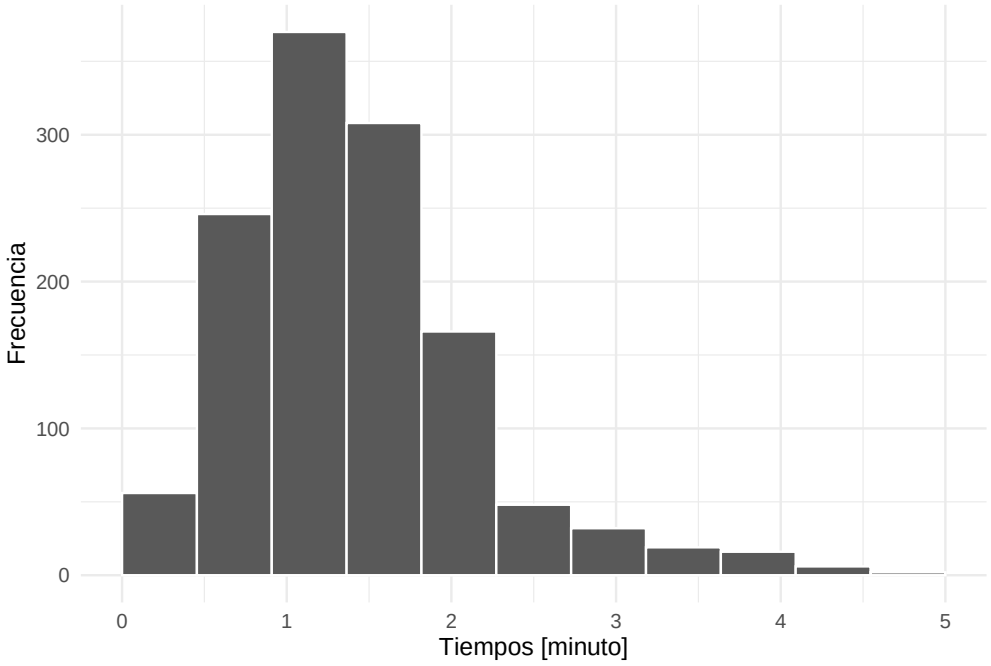

Figure 3: Resumen estadístico e histograma de delta_2

2.3 AED sobre delta_3

La variable *delta_3* no posee *outliers*. En primera instancia, los datos parecieran no seguir una distribución normal, hipótesis confirmada al realizar el test K-S con una confianza del 95%, entregando un estadístico $D = 0.11241$ y un $p - valor = 1.255x10^{-14}$.

Min	0.000
1er Qu.	2.007
Mediana	8.315
Media	8.369
3er Qu.	13.610
Max	23.670
Desv.	6.445
CV	0.768

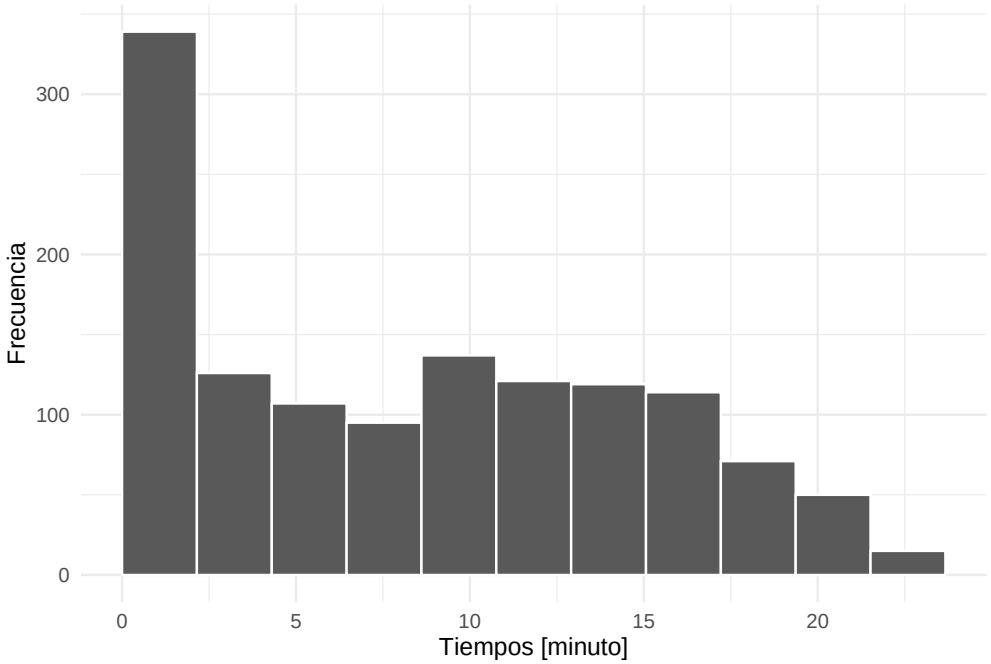

Figure 4: Resumen estadístico e histograma de delta_3

2.4 AED sobre delta_4

Dentro de los deltas analizados, *delta_4* es el que posee la media y mediana más alta (ver figura 1, con un valor de 17,045 y 22,639 respectivamente. Además, cuenta con el cuarto cuartil más alto, valor que coincide con su máximo (83,570), dado que no posee *outliers*. Gráficamente, como se puede apreciar en la figura 5, pareciera no seguir una distribución normal. Al realizar el test K-S, con una confianza del 95%, este entrega un estadístico $D = 0.144$ y un $p - valor = 2.2x10^{-16}$, por lo tanto, se concluye que no posee distribución normal.

Min	0.000
1er Qu.	3.445
Mediana	17.045
Media	22.639
3er Qu.	36.540
Max	83.570
Desv.	21.308
CV	0.491

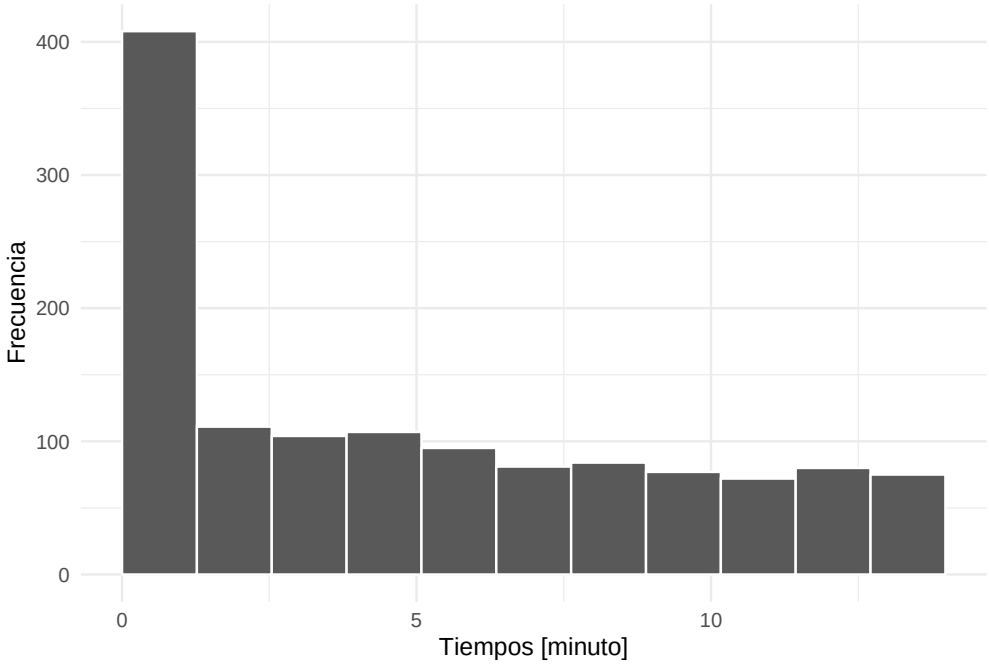

Figure 5: Resumen estadístico e histograma de delta_4

2.5 AED sobre delta_5

De *delta_5* se puede observar que posee la media y mediana más pequeña de todos los *deltas* (ver figura 1), con un valor de 0,5200 y 0,5104 minutos respectivamente. El máximo tiempo registrado es de 1 minuto y cuenta con el coeficiente de variación más bajo, de 0,575. No posee *outliers*.

Gráficamente (ver figura 6), tampoco se pueden observar un comportamiento normal. El test K-S entrega un estadístico $D = 0.074$ y $p - valor = 1.205x10^{-6}$, por lo que, se rechaza la hipótesis de normalidad con una confianza del 95%.

Min	0.000
1er Qu.	0.2500
Mediana	0.5200
Media	0.5104
3er Qu.	0.7600
Max	1.0000
Desv.	0.293
CV	0.575

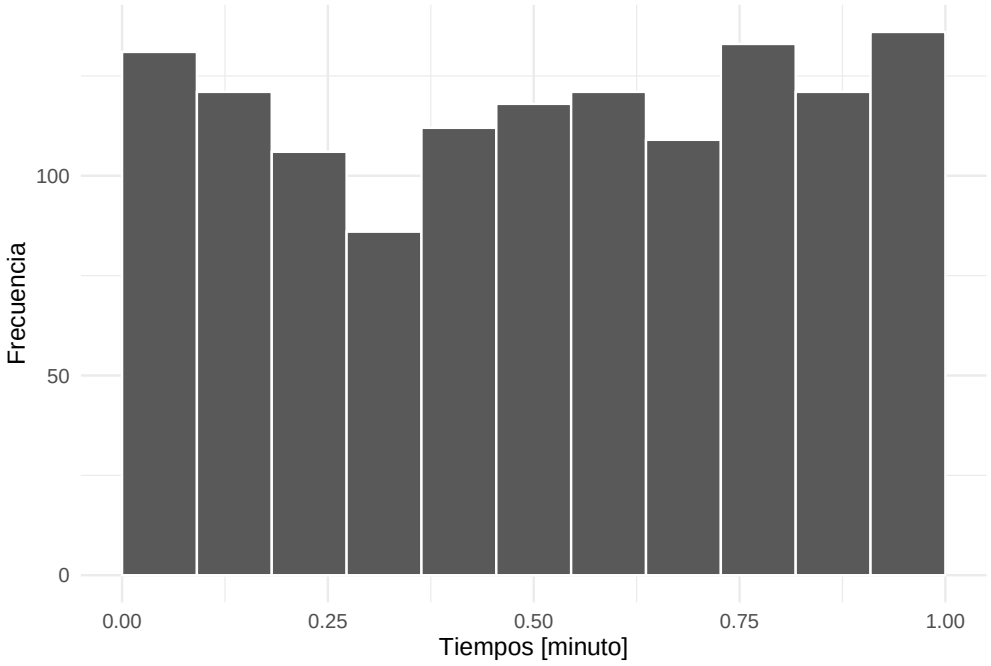

Figure 6: Resumen estadístico e histograma de delta_5

2.6 ¿Existen diferencias significativas en los tiempos de atención en los pórticos a la entrada y salida? (delta_1 y delta_5)

Se estudiará si existen diferencias significativas entre las medianas de los grupos *delta_1* y *delta_5*. Dado que, anteriormente se ha verificado que ambos deltas no poseen una distribución normal, se opta por utilizar el test no perimétrico de rangos de signos de Wilcoxon, que, a diferencia de su "símil" perimétrico t-test que utiliza la media, Wilcoxon utiliza la mediana. Por lo tanto, se tiene la siguiente hipótesis:

$$H_0 : Me_{N1} - Me_{N2} = 0$$
$$H_1 : Me_{N1} - Me_{N2} \neq 0$$

El test ha arrojado un $p - valor < 2.2x10^{-16}$, por lo tanto, con una confianza del 95% se descarta que las medianas de ambos deltas sean iguales.

3 P2: ¿Es posible aseverar que existen diferencias en los tiempos de atención en los pórticos de ingreso, dependiendo del turno?

En primer lugar, se realiza un test de normalidad (K-S) para cada turno de *delta_2*. Y, con un nivel de significancia de $\alpha = 0.05$, se rechaza la hipótesis de normalidad para los 3 turnos.

A partir de un análisis gráfico de los 3 turnos de *delta_2* (ver figura 7), pareciera no contar con diferencias significativas en su mediana. Cabe destacar que, se han eliminado los *outliers* en la figura 7, para facilitar su visualización.

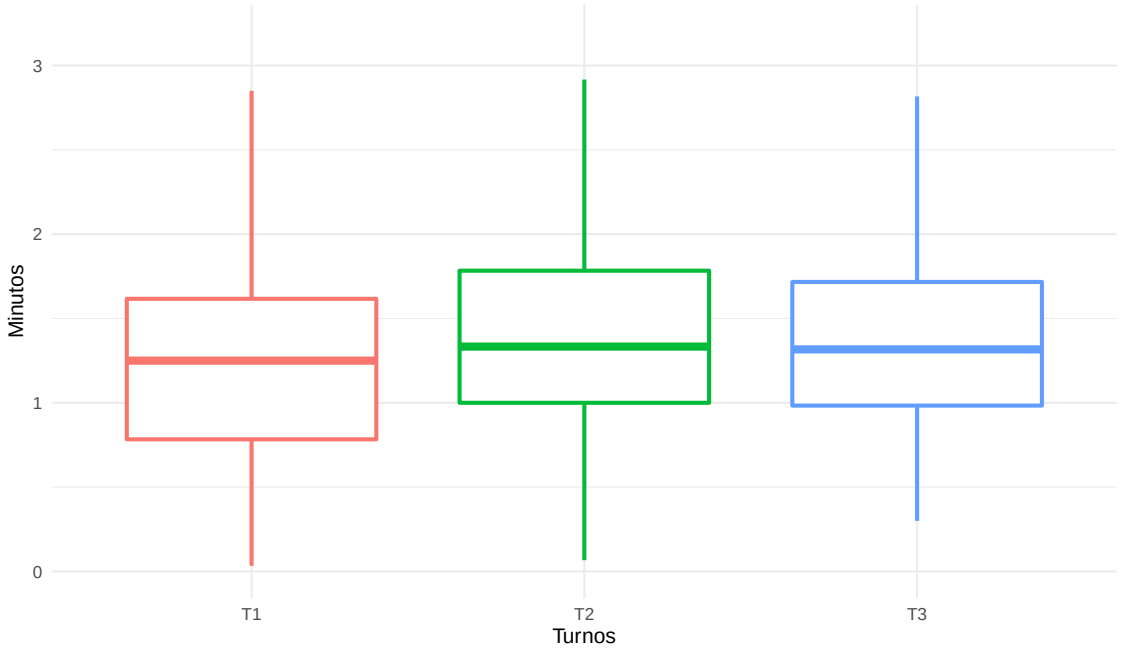

Figure 7: Tiempos, en minuto, de los turnos de *deltas_2*

Dado que en los 3 turnos de *delta_2* se ha evidenciado que no se distribuyen normal, se ha optado por utilizar el test no parametrico de Wilcoxon, para estudiar si existen diferencias significativas en cada part de turnos. A partir de los resultados obtenidos en la tabla 1, y con una confianza del 95%, se asevera que existen diferencias significativas en las medianas de las tuplas de turnos T1 - T2 y T1 - T3, contrario del caso T2 - T3, con medianas similares.

Diferencia turnos	**Estadístico**	$p - valor$
Dif. T1 - T2	119909	$7.396x10^{-5}$
Dif. T1 - T-3	57790	0.004335
Dif. T2 - T3	59291	0.6717

Table 1: Test Wilcoxon para turnos *delta_2*

4 P3: ¿Es posible asegurar que no existe un sesgo asociado al investigador en cada estación en donde se midieron tiempos?

El sesgo en una investigación se asocia a errores sistemáticos que aparecen de forma no aleatoria en los procesos de recolección o procesamiento de los datos. Desde la perspectiva del investigador que recolecta los datos, estos pueden ocurrir por un error de medición o tipeo. La respuesta corta es decir que no se puede asegurar que no exista un sesgo asociado al investigador, dado que es muy complejo no contar con alguna desviación entre el tiempo real con el tiempo registrado por el investigador debido a errores humanos.

Al analizar los datos de la muestra, se puede observar que no existen 471 registros para la hora de acceso del camión, asociado al investigador 1. Para el mismo investigador, también se puede observar que el tiempo de permanencia en el pórtico de acceso para un camión en particular, ha dado un valor negativo, esto se podría deber a un error al registra el tiempo de entrada como tiempo de salida y viceversa.

5 P4: ¿Existe alguna relación entre el turno de ingreso al recinto y el tipo de carga que transporta el camión?

Se desea analizar la relación entre las variabes de turno y tipo de carga (ver tabla 2), para ello se realiza el test X^2 de independencia, planteando la siguiente hipótesis:

H_0: El turno y el tipo de carga son independientes

H_1: El turno y el tipo de carga son dependientes, el % de tipos de carga varía entre turnos

El test aporta un $p - valor = 2.016x10^{-6}$, y con una $\alpha = 0.05$, tenemos un resultado estadísticamente significativo para rechazar la hipótesis de independencia, en otras palabras, se podría concluir con una confianza del confianza del 95% que el turno y el tipo de carga son dependientes.

	Alimentos no perecibles	**Bienes de lujo**	**Medicamentos**	**Textil**	**Textos**
T1	19	12	251	123	159
T2	64	3	191	104	133
T3	26	5	145	34	25

Table 2: Tabla de contingencia turno y tipo de carga

6 P5: ¿Existe alguna diferencia en los tiempos de ciclo de un camión en el recinto, de acuerdo al tipo de carga que transporta?

Para este caso se analizaran los deltas 2, 3 y 4, dado que en estos procesos podría verse afectado el tiempo por tipos de cargamento; delta 1 y 5 solo consideran tiempos de entrada y salida, y ello no afectaría por su tipo de cargamento. Por otro lado, no se consideran los bienes de lujo y texto para no alargar el análisis. Dado que las variables no poseen una distribución normal se utilizarán sus medianas para comparar si existen diferencias significativas entre tipos de cargamento por delta, ver tabla 3.

	delta 2	**delta 3**	**delta 4**
Alimentos no perecibles	1.47	10.58	26.47
Medicamentos	0.95	9.96	21.65
Textil	1.67	4.12	9.87

Table 3: Medianas por tipo de carga y deltas 2, 3 y 4

Al realizar los test de hipótesis para estudiar si existen diferencias significativas por cada tupla de medianas de los tipos de carga, se obtienen los resultados expuesto en las tablas 4, 5 y 6. Se obtiene, por delta, y con una confianza del 95%: *delta_2*: diferencias significativas de las medianas entre Alimentos no perecibles y Medicamentos, y Medicamentos y Textil. *delta_3*: diferencias significativas entre Alimentos no perecibles y Textos, y Medicamentos y Textil. *delta_4*: diferencias significativas entre Alimentos no perecibles y Textos, y Medicamentos y Textil.

Hipótesis	$p - valor$
$Me_A - Me_M$	$2.2x10^{-16}$
$Me_A - Me_T$	0.2092
$Me_M - Me_T$	$2.2x10^{-16}$

Table 4: Test *delta_2*

Hipótesis	$p - valor$
$Me_A - Me_M$	0.5012
$Me_A - Me_T$	$4.063x10^{-5}$
$Me_M - Me_T$	$4.194x10^{-9}$

Table 5: Test *delta_3*

Hipótesis	$p - valor$
$Me_A - Me_M$	0.3057
$Me_A - Me_T$	0.0002472
$Me_M - Me_T$	$2.705x10^{-6}$

Table 6: Test *delta_4*

7 P6: ¿Existe evidencia suficiente que respalde algún tipo de asociación entre deltas 2, 3 y 4?

Dado que el supuesto de normalidad se ha rechazado para todos los deltas, se descarta aplicar el coeficiente de correlación de Pearson, y se opta por la alternativa no perimétrica, el coeficiente de correlación por rangos de Spearman.

Los resultados obtenidos (ver tabla 7) indican que, con una confianza del 95%, para los deltas 2, 3 y 4 existe evidencia suficiente para rechazar la hipótesis nula en favor a la hipótesis alternativa de que los deltas si poseen correlación. Se puede observar de *delta_2* y *delta_3* y, *delta_2* y *delta_4* una correlación negativa débil. Por otra parte, *delta_3* y *delta_4* posee una fuerte correlación positiva ($\rho = 0.884$). Estas correlación se pueden observar gráficamente en la figura 8.

Prueba	**p-valor**	ρ
delta_2 $\sim$ delta_3	$1.63x10^{-11}$	-0.1858
delta_2 $\sim$ delta_4	$8.396x10^{-8}$	-0.14827
delta_3 $\sim$ delta_4	$< 2.2x10^{-16}$	0.8837571

Table 7: Test de correlación por método Spearman

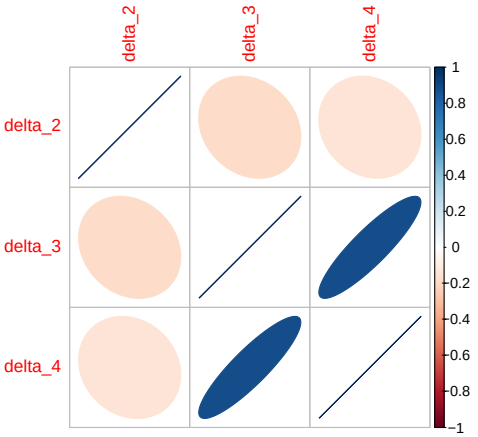

Figure 8: Gráfica de correlación por método Spearman