**Taller 1 Estadística y diseño de experimentos**

**Por: Santiago Echavarría**

El presente taller realiza un análisis descriptivo del conjunto de datos “Pablo.csv” mediante el uso de herramientas de estadística descriptiva vistos en el curso.

Para el desarrollo del ejercicio se realiza un programa que concatena un resumen estadístico de cada una de las variables en un dataframe como se ve a continuación:

Adicionalmente, el programa genera un archivo de Excel con dichos valores el cual se adjunta como anexo 1. Una vez hecho esto, se crean los histogramas de frecuencia utilizando la librería “Plotly” la cual se encuentra disponible para “R”, “Python” y “Javascript”. Al correr el programa, las gráficas de despliegan en el navegador, como se muestra en la figura 1. La variable “X3” se muestra en escala logarítmica para una mejor visualización.



Figura 1: histogramas de frecuencia

Finalmente, con la misma librería realizamos los boxplot de los datos para lograr identificar los valores atípicos, los cuales se encuentran en la figura 2. Cabe resaltar que el método usado para calcular los cuartiles es “lineal”, es decir, promediando el método que incluye la media y el que la excluye. Los resultados se muestran a continuación:

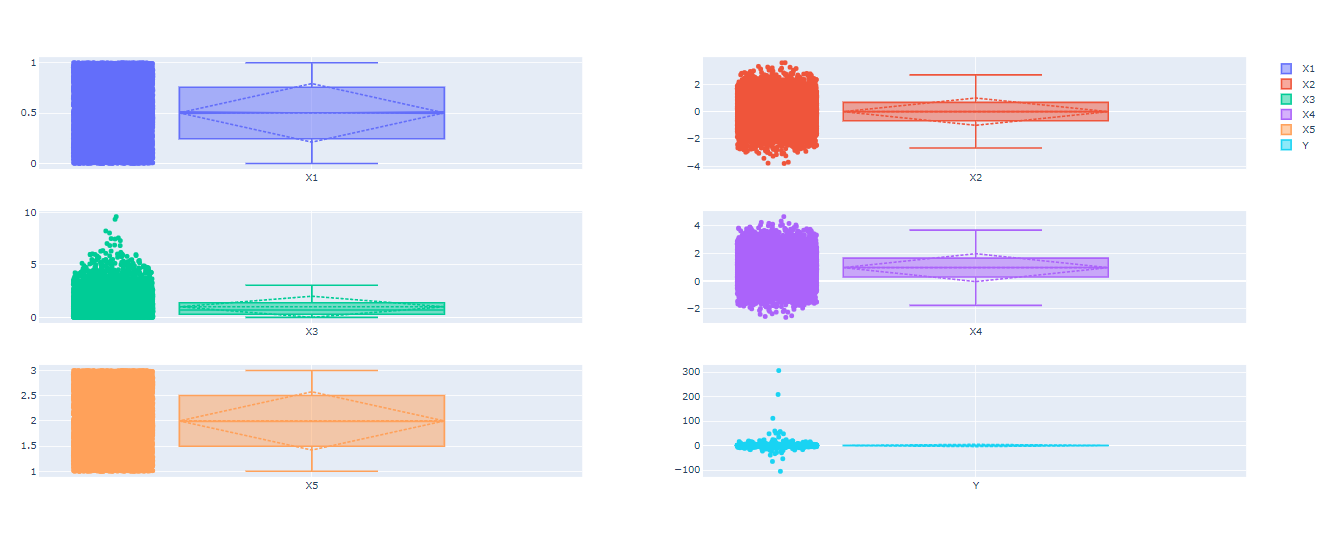


Figura 2: boxplots

**Conclusiones**:

Se acuerdo con los histogramas de frecuencia, podemos observar que las variables X2 y X4 presentan una distribución normal con valores de media y mediana muy cercanos entre sí. Estos 2 indicadores cuentan con una diferencia de 0,0011 unidades para X2 y 0,0127 para X4, las cuales tienen rangos similares de 7,3865 y 7,2751 respectivamente. Otro dato interesante para estas 2 variables es que el 68,4% de los datos se encuentra en el rango: mean - dev y mean + dev, como se ve en la tabla anexa 1.

Diferencias entre el rango inter-cuartil y rango:

Gracias a esta diferencia, podemos evidenciar diferentes tipos de distribuciones en los datos por ejemplo para “X1” y “X5”, el rango inter-cuartil es aproximadamente la mitad del rango total, ya que gráficamente se evidencia una distribución uniforme. Por el contrario, las variables “X3” Y “Y”, presentan una variación mucho más significativa al comparar dichos indicadores.

Para las variables “X3” y “Y”, se observa una alta concentración de datos en los histogramas y boxplots, para confirmar esto, vemos que la probabilidad de que un valor de la variable esté en el rango: mean - dev y mean + dev es mayor al 98,5% para ambas variables. Sin embargo, una diferencia significativa entre estas 2 variables (“X3”, “Y”) es que la primera se encuentra distribuida geométricamente hacia la derecha. Este comportamiento se evidencia ya que la mediana y el promedio se encuentran más alejados entre sí que para el resto de variables. En el caso de la variable Y, observamos que los datos se encuentran concentrados de manera central, con ciertos valores atípicos que conllevan a un amplio rango.