[INF-280 2021-1]T1-Eval

April 10, 2021

INF-280 Estadística Computacional I-2021
Tema 1 - Análisis Exploratorio de Datos

0.0.1 Formalidades

- Es posible utilizar apuntes, libros, papers, ejemplos y cualquier otro material que desee de internet, pero debe ser usted quien desarrolle los ejercicios y comente los resultados en el Notebook que entregará.
- Puede utilizar Python o R. En ambos casos, no puede usar funciones estadísticas específicas, excepto aquellas correspondientes al cálculo de medidas de tendencia ó dispersión (media, desviación estándar, etc) y aquellas correspondientes al cálculo de probabilidades y percentiles sobre las distribuciones revisadas en clases (pnorm, qnorm, pt, qt), etc.
- Se podrá trabajar en grupos de **dos personas**, las cuales deberán estar inscritas a través de **Aula**.
- Además del Notebook, deberán entregar un video explicando lo que hizo, paso a paso, como en los ejemplos entregados en AULA. Este vídeo es **individual**, es decir, cada integrante debe hacerlo por separado (debiendo utilizar el mismo Jupyter Notebook). El video puede ser subido a una plataforma externa como Youtube, Google Drive, ó Dropbox. El **link a compartir** (*visible para cualquier usuario con el link*) debe ir dentro del Notebook.
- La nota de esta actividad es grupal.

0.0.2 Entrega

- **Formato**: Se debe realizar una **única** entrega por grupo (un representate del grupo) a través de **Aula**.
- **Archivo**: Sus respuestas deben ser entregadas en forma de Jupyter Notebook, incluyendo **dos link** a los vídeos explicativos, uno por cada integrante.
  - **Vídeo**: Debe grabar la pantalla (Jupyter Notebok) donde se realiza la explicación de la actividad, con una duración de 10 a 15 minutos.
- **Fecha límite**: Viernes 16 de Abril a las 23:59.

0.1 Enunciado

Una empresa de latinoamérica está realizando un estudio del desempeño productivo de sus trabajadores en la modalidad online de trabajo. En particular, se está estudiando y comparando tres diferentes condiciones C sobre las cuales los empleados realizan sus actividades de trabajo en casa. En esta actividad tendrá acceso a los datos para ayudar en el estudio, los cuales corresponden a

una muestra finita sobre cada una de las condiciones (C_1, C_2, C_3). Cada valor en la muestra indica la medición del desempeño productivo de algún trabajador anónimo (*mayor indica un mejor desempeño, valor mínimo es 0*).

> En base al desempeño en modalidad presencial (lo cual no está disponible para esta actividad), la empresa define ciertos márgenes de pérdida referenciales a cada persona. En concreto se indicia lo siguiente: un desempeño aceptable, donde el trabajador se desenvuelve de la misma manera que lo hacía en su trabajo presencial, corresponde a un valor cercano a 10. Mientras que un valor menor a 5 se considera un desempeño inaceptable, donde el trabajador realiza menos actividades de las esperadas en el margen de trabajo online, por lo que podría generar pérdidas a la empresa. Por otro lado, un valor mayor a 20 corresponde a un desempeño sobresaliente el cual podría ayudar a acelerar ciertos proyectos y lograr más objetivos.

Los datos pueden ser descargados a través del siguiente link: **muestraC1**, **muestraC2**, **muestraC3**. Cada fila corresponde a un dato de la muestra, los archivos no tienen encabezado (la primera fila es el primer dato de la muestra). > *Para descargar directamente en Jupyter puede ejecutar* (**solo sistemas unix**)

```
!wget https://raw.githubusercontent.com/FMena14/ML_usm/master/Estadistica/T1-Eval-Data/Muestra
!wget https://raw.githubusercontent.com/FMena14/ML_usm/master/Estadistica/T1-Eval-Data/Muestra
!wget https://raw.githubusercontent.com/FMena14/ML_usm/master/Estadistica/T1-Eval-Data/Muestra
```

0.1.1 Actividades

Para los análisis cuantitativos recuerde utilizar una precisión decimal más que los datos, sino su respuesta se considerará imprecisa y ***podría*** *tener puntaje 0.*

a) [5 pts] Describa la población del fenómeno aleatorio de estudio, la variable asociada y las muestras a trabajar £Cuántos datos tiene cada muestra?

b) [10 pts] Compare el desempeño sobre cada condición a través de los puntos centrales de tendencia (media y mediana). Comente con la información entregada en el enunciado acerca de los desempeños *aceptable* £Qué información aporta el sesgo de los datos?

c) [20 pts] Grafique el boxplot de cada muestra/condición y compare entre sí £Existe evidencia suficiente para concluir que alguna de las condiciones es mejor que las otras dos? Añada al gráfico la información que se entregó sobre los desempeños inaceptables, aceptables y sobresalientes, £Qué se puede comentar al respecto de las condiciones?

d) [15 pts] Analice la dispersión de cada muestra/condición a través de 3 medidas que entreguen información diferente. £Qué información aporta cada medida elegida? £Qué indica lo observado? Compare con lo analizado en c)

e) [20 pts] En base a toda la información recolectada hasta este punto (medidas de tendencia, variabilidad y gráficos) £Cuál condición C le recomendaría a la empresa inculcar sobre el resto de empleados? £Por qué? Apoye su elección en base a los tipos de desempeño.

f) [15 pts] Al volver a revisar el proceso de recolección de datos, se encontró con un trabajador que no había sido considerado antes, con un desempeño de 13. No se encontró la información de en cuál condición se encontraba. En base a lo realizado en la actividad, explique e implemente una técnica para tratar de inferir bajo qué condición se encontraba este trabajador olvidado. £Cuál condición? £Por qué? £Sería posible entregar un margen de error de su técnica?

g) [15 pts] Ahora se le solicita realizar el estudio de qué ocurre si la empresa decide gradualmente enviar a sus mejores trabajadores (los con desempeño más alto) al trabajo presencial de oficina. Compare cómo se ve afectada la media muestral de cada condición si un porcentaje p de los trabajadores bajo esa condición son enviados a trabajo presencial (ya no formarían parte de esta muestra). Grafique la variación de p en un rango $[0; 0.99]$. En base a esto £Cuál condición sería recomendable mantener en trabajo online para evitar una baja en las medidas de tendencia central de trabajo online?