Proyecto de Probabilidad y Estadística

Verano 2021-2022

17 de julio de 2022

1. Indicaciones Generales

- El proyecto debe ser realizado por grupos de dos personas, si la clase tiene un número impar de estudiantes, uno de los grupos será de tres integrantes. Todos los grupos deben inscribirse en la sección de grupos del D2L.
- Todas las respuestas necesitan mostrar el procedimiento, y todas deben estar conformadas por una explicación matemática y una en lenguaje natural.
- Todos los entregables que no sean en formato Excel o Python deben ser escritos en Latex.
- Cada sección del trabajo especificará los entregables, el formato y la fecha de entrega.
- Cualquier tipo de plagio implicará una nota de 0 en el proyecto, en la clase, y todos los procedimientos establecidos en el Código de Honor de la USFQ.
- El trabajo debe tener referencias bibliográficas.
- Las únicas circunstancias bajo las cuales se aceptan entregas tardías son calamidades domésticas o de salud comprobables.
- Cualquier falta a las reglas especificadas en esta sección significará que el grupo tendrá 0 en su nota final.
- Cada sección tendrá una defensa, la cual constará de 10 preguntas sobre el procedimiento, los entregables y la materia. Cada una de estas 10 preguntas son equivalentes al 10 % de la nota de la sección a la que correspondan. Cada sección será calificada en base a los entregables que se presenten. Sin embargo, cada pregunta de la defensa de la sección podrá obtener tres calificaciones posibles: 0 (cero), -5 % (menos cinco por ciento), o -10 % (menos 10 por ciento). Es decir si la pregunta se responde de manera correcta y completa, la nota de la sección del proyecto permanece intacta, si se responde de manera incompleta se restan 5 puntos porcentuales, y si se responde mal se restan 10. Las defensas de cada sección se harán después de los examenes parciales, de manera presencial y obligatoriamente deberán estar todos los integrantes de cada grupo, la ausencia de uno de ellos representará un 0 en la califiación total de la sección.
- Cualquier actualización de estas reglas tendrá el mismo valor y será publicada oportunamente en el D2L.

2. Primera Sección

2.1. Tema

- El tema del trabajo puede ser elegido por cada grupo en base a una idea original o a las recomendaciones que se presentan a continuación.
- El tema debe ser presentado en forma de una pregunta que se pueda responder a través del uso de estadística inferencial.
- El tema debe ser aprobado por el profesor de la clase de ejercicios.

- Todos los proyectos deben tener un proceso de recolección de datos. Si los datos se encuentran disponibles en cualquier tipo de página web o base de datos, el grupo recibirá una nota de 0.
- Algunas recomendaciones para el proyecto son:
  - ¿Los vegetales del mercado son más grandes que los del Supermaxi?
  - ¿Históricamente, la asistencia de los hinchas de Liga de Quito al estadio ha sido más baja que la de cualquier otro equipo grande del país?
  - ¿Los estudiantes del politécnico tienen mejor GPA que los estudiantes del CADE?

2.2. Recolección de datos

- Cada grupo debe recoger al menos 4 variables cuantitativas y 2 cualitativas, por observación.
- Cada grupo debe recoger al menos 200 observaciones.
- El grupo debe presentar un reporte de máximo dos páginas explicando cómo obtuvieron los datos, de dónde los sacaron, los problemas que enfrentaron, las recomendaciones que darían para otras personas que quisieran recoger la misma información, y una explicación de cada variable. Adjunto a este reporte se debe presentar una tabla en Excel con todas las observaciones y variables recolectadas. El reporte debe incluir la descripción del tema seleccionado.
- Esta primera entrega valdrá la mitad de la nota de la primera sección del proyecto, es decir, un sexto de la nota final.
- La fecha de entrega será el sábado 11 de junio.
- Los entregables de esta subsección son: PDF del reporte compilado en LATEX, carpeta comprimida que contenga el código de latex y todo lo necesario para que el código compile, y archivo de excel con los datos recolectados.

2.3. Estadísticas Descriptivas

- Todos los cálculos de esta subsección se deben realizar en Excel Y en Python (en ambos softwares).
- Recordar que todas las respuestas deben contener lenguaje matemático y natural para explicar todo el proceso.
- Sobre los datos cuantitativos deben presentar: media, moda, mediana, varianza y desviación estándar, cuartiles, rango intercuartil, percentil 10, límites superior e inferior.
- Deben presentar dos box plots comparativos (uno por cada grupo de variables), y dos gráficos más que sirvan para comparar las variables cuantitativas de los dos grupos de datos.
- Deben presentar al menos dos formas diferentes de representar y comparar las variables cualitativas de los dos grupos de datos.
- Esta entrega valdrá la segunda mitad del valor de la primera parte del proyecto, es decir un sexto de la nota final.
- La fecha de entrega de esta subseción será el sábado 18 de junio.
- Los entregables de esta subsección son: PDF compilado en LATEX con la redacción de las respuestas a lo requerido en lo explicado en esta sección, carpeta comprimida que contenga el código de latex y todo lo necesario para que el código compile, archivo de excel con todo el trabajo hecho, código de python (link de Google Collab) donde se hicieron los procedimientos.

3. Segunda Sección

3.1. Indicaciones Generales

- Para esta sección, los entregable son un pdf con el reporte mostrando TODO el trabajo realizado para cada respuesta, los procesos, el resultado y su interpretación. Deben presentar también el código fuente de LATEX y el código de python utilizado para resolver los problemas. El código debe estar debidamente comentado y debe poder correrse desde cualquier computador.
- Todo resultado sin procedimiento algebraico claro, o que no tenga el respectivo desarrollo en código, no tendrá validez.
- El reporte en pdf que deben entregar es la unión de lo que realizaron en la primera sección junto con lo que van a realizar en esta sección.

3.2. Espacio Muestral y Probabilidad

- Encuentren todos los elementos del espacio muestral de un experimento donde los posibles resultados son todas las posibles combinaciones de respuestas entre todas las variables que recolectaron. Deben hacerlo de manera computacional en python y deben establecer claramente cuántos elementos tiene su espacio muestral.
- Calcular usando combinaciones o permutaciones (lo que ustedes deduzcan que deben usar), el tamaño del espacio muestral descrito en el item anterior. Esto debe hacerse de manera algebraica mostrando todos los pasos.
- Para cada variable cuantitativa, calculen la probabilidad de que una observación esté por debajo de la media. Elijan una combinación de dos variables cuantitativas y calculen la probabilidad de que la segunda esté debajo de la media dado que la primera lo está. Hágan lo mismo con una combinación de tres variables cuantitativas.

3.3. Variables Aleatorias

- Consideren el experimento de obtener una observación de su muestra. Y consideren la variable aleatoria X que muestra la diferencia entre el valor de una de las variables cuantitativas de esa observación con la media de la muestra. Encuentren el valor esperado y la varianza de esta variable aletaoria.
- Muestren si esta variable aleatoria sigue una distribución binomial, exponencial o normal. Deben hacer esto computacionalmente.
- Grafiquen una aproximación de la función de masa de probabilidad de esta variable aleatoria, computacionalmente.
- Definan una segunda variable aleatoria de la misma manera con otra de las variables cuantitativas de su muestra. Calculen computacionalmente, usando SOLO SUMAS Y MULTIPLICACIONES (no paquetes que hagan el cálculo por ustedes) la covarianza y el coeficiente de correlación entre ambas variables. Interpreten los resultados.
- Seleccionen dos variables aleatorias, donde cada una describe los posibles resultados de una variable cualitativa de su muestra. Describan la función de masa de probabilidad conjunta, y grafíquenla en un espacio tridimensional usando python. Demuestren si estas variables son dependientes o independientes.

3.4. Muestreo

- Selesccionen dos de sus variables aleatorias y para cada una saquen 40 muestras aleatroría del al menos el 25 % de los datos. Calculen la media y la varianza de estas muestras y grafiquen sus distribuciones, prueben computacionalmente a través de dós gráficos distintos (gráfico de distribución y QQ plot) que estos estadísticos siguen una distribución normal.

- Repitan el ejercicio anterios con 5 muestras.
- Presenten una tabla con la media y la desviación estándar de cada muestra, y la diferencia de estas con la media y la desviación estándar de la población (todos sus datos). Interpreten los resultados. Asuman que la media de cada muestra es un estimador de la media poblacional. Calculen el error estándar del estimador de la media poblacional.
- Elijan una variable cuantitativa de su encuesta, encuentren la distribución que más se asemeja a la distribución de su variable, y asumiendo que sus datos provienen de un muestreo aleatorio (todos sus datos), estimen la varianza poblacional usando el metodo de máxima verosimilitud (usando la función de masa de probabilidad correspondiente).

4. Tercera Sección

4.1. Estadística Inferencial

- Construir intervalos de confianza para la media de sus variables cuantitativas a un nivel de confianza del 99 %.
- Planteen la hipóteis nula y alternativa para resolver la pregunta de su investigación.
- Calculen el estadístico adecuado, de manera algebraica (para esto pueden usar funciones de suma, resta, multiplicación, división y raíces en python).
- Calculen el valor p.
- Elijan tres niveles de confianza distintos para presentar sus resultados y concluyan, para cada uno, si se rechaza o no la hipótesis nula.
- Elijan una variable cuantitativa, divídanla en 2 grupos y respondan la pregunta de si las varianzas son iguales entre los grupos, usando el test F.
- Vuelvan a resolver su pregunta de investigación, pero esta vez usando paquetes de python.

4.2. Regresión Lineal

- Estimen una regresión lineal de una variable dependiente usando el método de mínimos cuadrados (solo pueden usar operaciones básicas en python), elijan sus variables dependiente e independiente. Interpreten los resultados y calculen la suma de errores al cuadrado y el r^2. También calculen el intervalo de confianza para el coeficiente de regresión, utilicen inferencia estadística para determinar si este coeficiente es diferente de 0.
- Repitan el ejercicio anterior utilizando paquetes de python que realicen regresiones y comparen los resultados.
- Elijan dos variables cuantitativas, calculen el coeficiente de correlación de manera algebraica y hagan un gráfico de correlación.