Estadística y Probabilidades

Proyecto de Curso

1. Motivación:

El proyecto de curso tiene por objetivo aplicar los conocimientos recibidos en el curso de Estadística y Probabilidades. Las primeras dos entregas se centrarán en la clasificación de la información, siendo las entregas posteriores centradas en el análisis de datos recolectados.

2. Entregas de proyecto

El proyecto consiste de una pre-entrega y cuatro entregas.

2.1. P1 primera entrega: justificación, descripción del proyecto y análisis preliminar

La primera entrega tiene por objetivos:

- Comprender el proceso de inscripción de los grupos de proyecto.
- Definir el estudio a realizar y destacar su relevancia.
- Reconocer las variables significativas del estudio.
- Describir en una primera instancia las diversas variables significativas.
- Identificar patrones en los datos de esas variables.

La primera entrega tiene 3 instancias de entrega, pre-entrega, presentación oral e informe.

P0 Pre-entrega: Esta parte tiene valor de 15% de la nota de la primera entrega y será indispensable su entrega correcta para la calificación de las posteriores partes de la entrega P1.

En esta entrega se deben plantear tres temas, en orden de preferencia del grupo, para evaluación por parte de su profesor. Asociado a cada tema se debe especificar, brevemente la relevancia del estudio, proveniencia de los datos y disponibilidad de los mismos.

Evidentemente, un estudio debe ser factible, es decir, debería haber razones justificadas para creer que los datos requeridos podrán ser obtenidos por el grupo.

Los temas tentativos se escogen en equipo. Pueden dar prioridad a los siguientes temas: cambio climático, desarrollo sustentable, reducción de pobreza crítica, preservación del patrimonio histórico, acceso a educación de calidad, mejoras en la calidad de vida preferiblemente en el contexto del Perú, pero si el equipo está de acuerdo, se pueden presentar otros temas, siempre que se manejen con seriedad y rigurosidad.

Al igual que las otras entregas, esta entrega es grupal, para esta entrega los grupos deberán estar constituidos y ser inscritos correctamente, declarando primero la sección a la que pertenecen y segundo eligiendo un nombre corto consistente con la primera opción planteada en el estudio a realizar.

Ejemplo:

Un grupo de la **sección 4** hará un **análisis de la producción de arroz** por regiones en el Perú.

El nombre del grupo será: **S4 Producción de Arroz**

En general, hay tres opciones para obtener los datos del proyecto:

- Bases de datos públicas.
- Encuestas en trabajo de campo.
- Experimento estadístico.

En la entrega P0 debe especificarse la forma como se obtendrán los datos, aunque no sea necesario tener los datos para el momento de la pre-entrega.

En la justificación de factibilidad deben considerarse las siguientes observaciones en función de la opción seleccionada para cada tema. Algunas de las observaciones pueden tener repercusiones para futuras entregas.

En el caso de la base de datos pública, se deberá encontrar una fuente confiable de datos, donde se especifique la fecha de obtención de los mismos. Es imprescindible que demuestren, a plena satisfacción de su profesor, que tienen acceso a los datos y deben tener al menos 150 observaciones en la base de datos.

En el caso de encuesta, se deberán presentar un plan sensato de recolección de los datos recolectados; se podrán recolectar datos hasta dos semanas después de la presentación de la entrega P1, pero debe haber al menos 50 observaciones para la entrega P1. El número mínimo de observaciones a ser recolectadas para el estudio será de 150.

En el caso del experimento, se debe presentar un plan sensato de recolección de datos y controles si fuesen necesarios; se podrán recolectar datos hasta dos semanas después de la presentación de la entrega P1, pero debe haber al menos 50 observaciones para la entrega P1. El número mínimo de observaciones a ser recolectadas para el estudio será de 150.

En cualquiera de los casos, si no se cumplen los requisitos planteados, el tema será declarado INVIABLE y no podrán utilizarlo. Si están trabajando sobre un tema que fue declarado VIABLE y no alcanzan algunos de los parámetros establecidos, como por ejemplo, el mínimo número de observaciones, perderán puntos en las entregas posteriores.

La información debe estar en formato CSV con codificación UTF-8.

La información será procesada en RStudio y podría ser pertinente realizar alguna conversión de formato previa a su utilización; este proceso deberá ser finalizado para la esta primera entrega.

Entregar una base de datos en la entrega P0 no es garantía de que el tema será aceptado, pero ciertamente contribuye a defender su factibilidad.

En P0 el equipo debe redactar las respuestas a las siguientes preguntas:

- ¿Es el estudio de interés para la audiencia?
- ¿Qué información se desea aportar al final del estudio?
- ¿Es posible llegar a la información que se propone de manera específica y concreta?

La revisión de la pre-entrega tendrá comentarios sobre los temas en caso de que puedan ser viables, si ningún tema es viable, el equipo tendrá que presentar un tema diferente sin haber recibido *feedback* sobre este.

P1 presentación oral y P1 informe: Para esta etapa, se debe presentar el avance corregido a partir de los comentarios sobre la entrega P0.

La entrega del informe ocurre un corto tiempo después de la presentación oral, por lo que un equipo bien coordinado debería poder incluir en la entrega del informe, algunas mejoras derivadas de las observaciones a la presentación oral.

La información obtenida (ya sea de una base de datos, encuesta o experimento) se debe procesar para clasificar en equipo las variables del estudio. En caso de trabajar con una encuesta o un experimento se debe asociar la variable a la pregunta o experimento correspondiente. Cada variable debe estar correctamente clasificada de acuerdo a su tipo y restricciones de valores que pueda tener.

Es importante resaltar la cantidad de datos faltantes y observaciones completas con las que se cuenta. Deben recordar que para la primera entrega, deben tener una cantidad suficiente de observaciones, aunque el proceso de recolección aún se encuentre en proceso.

Ejemplos:

Pregunta	Variable	Tipo de Variable	Restricciones
Ingrese año de nacimiento	Año de nacimiento	Ordinal	Entero mayor a 1900
Ingrese su grupo sanguíneo	Grupo Sanguíneo	Nominal	O, A, B, AB
¿Cuántas mascotas tiene en casa?	Número de mascotas	Discreta	Entero no negativo

Cada variable relevante debe tener una descripción en términos de los diversos descriptores descriptores numéricos o gráficos planteados en clase. La idea es describir la variable de manera relevante, ofreciendo detalles que permitan entender su estructura y los patrones que puedan tener.

Es fundamental recordar que los distintos tipos de variables podrían tener distintos descriptores apropiados. En el caso de utilizar descriptores gráficos, se debe prestar especial atención a todos los detalles, desde la relevancia de la gráfica para el tipo de variable hasta selección de escala, ejes, unidades, leyenda, título descriptivo y manejo apropiado del color. El resultado final debe ser una gráfica descriptiva que transmita efectivamente la información y patrones relevantes observables en la muestra de la variable.

Estas gráficas se podrán refinar en la siguiente entrega una vez que todos los datos estén disponibles o en función de los comentarios de su profesor.

2.2. P2 segunda entrega: Análisis estadístico descriptivo

La segunda entrega tiene 2 instancias de entrega, presentación oral e informe.

Con la información ya recolectada y en formato apropiado para su análisis en R, se deberá cumplir con los siguientes objetivos:

- Presentar las figuras de mérito (descriptores apropiados para cada variable o combinación de variables) que describen de la mejor manera posible los datos obtenidos.
- Describir el conjunto de datos de manera objetiva y destacar patrones observados.

- Identificar, de ser posible, si alguno de los modelos de variable aleatoria especificado en clase puede ser utilizado para describir el comportamiento de alguna de las variables recolectadas.
- Plantear hipótesis sobre el comportamiento o patrones observados en alguna de las variables recolectadas.
- Postular relaciones entre variables del conjunto de datos. Se deben presentar al menos tres pares de variables numéricas, que se sospeche guardan alguna relación entre ellas. Se pueden incluir relaciones adicionales si se detectan.
- Resolver a plena satisfacción de su profesor las observaciones correspondientes a la entrega P1.

Los integrantes del grupo se deben registrar de nuevo, notando si ha habido algún cambio en la constitución del grupo.

Se debe entregar, de manera obligatoria, en un archivo comprimido en formato ZIP y sin directorios internos:

- Los archivos en formato CSV pertinentes a la entrega.
- El archivo en formato R Notebook (fuente) con la carga, procesamiento y análisis de los datos. Este archivo deberá poder ser ejecutado y tejido, sin error, en la computadora de su profesor, por lo que recomendamos hacer las pruebas pertinentes para garantizar que esto funcione apropiadamente. Si esto no funciona, perderán puntos.
- El archivo en formato html que es resultado de ejecutar y tejer el código fuente; este archivo un respaldo para el caso de que haya fallas en la ejecución y tejido del código fuente. Si este archivo no está disponible, perderán puntos.

Para esta presentación se deberá tener en cuenta lo siguiente:

- Todo el análisis del conjunto de datos deberá realizarse en RStudio.
- Los grupos deberán hacer su presentación oral dentro del tiempo establecido.
- Es obligatorio presentar al menos 1 figura de mérito numérica y 1 figura de mérito gráfica por variable, justificando su elección, relevancia para describir la variable estudiada y análisis de los resultados obtenidos.
- El informe escrito que acompaña a la presentación oral deberá tener al menos 3 observaciones relevantes del comportamiento de las variables, resaltando las de mayor significancia para el estudio. Estas hipótesis serán el objeto de estudio de la próxima entrega.
- Las observaciones que su profesor haya hecho sobre la entrega P1 deben ser atendidas (esto vale 15 % de la nota de la entrega).

La entrega del informe ocurre un corto tiempo después de la presentación oral, por lo que un equipo bien coordinado debería poder incluir en la entrega del informe, algunas mejoras derivadas de las observaciones a la presentación oral.

2.3. P3 tercera entrega: Estadística inferencial y predictiva

La tercera entrega tiene 2 instancias de entrega, presentación oral e informe.

Para esta entrega se tiene por objetivo:

- Mejorar el análisis descriptivo de la entrega anterior.
- Hacer el análisis de fiabilidad, resaltando el nivel de confianza de los estimadores calculados para cada variable.
- Hacer uso de los fundamentos de estadística inferencial para analizar, desde un punto de vista estadístico, las hipótesis planteadas en la entrega anterior, por ejemplo, mediante pruebas de hipótesis o intervalos de confianza.
- Validar dependencias entre variables, y tratar de explicar, desde un punto de vista estadístico, las relaciones planteadas entre algunas de las variables, por ejemplo mediante regresión.
- Hacer uso de los fundamentos de la estadística predictiva para reconocer tendencias e inferir el comportamiento de variables fuera del rango estudiado.

Para esta entrega será calificado con 15 % de la nota las mejoras de la segunda entrega con el fin de lograr un mejor análisis descriptivo de los datos a analizar.

Los integrantes del grupo se deben registrar de nuevo, notando si ha habido algún cambio en la constitución del grupo.

Se debe entregar, de manera obligatoria, en un archivo comprimido en formato ZIP y sin directorios internos, siguiendo los lineamientos de la entrega anterior.

Para esta presentación se deberá tener en cuenta lo siguiente:

- Todo el análisis del conjunto de datos deberá realizarse en RStudio.
- Los grupos deberán hacer su presentación oral dentro del tiempo establecido.
- Los parámetros que estimen en función de la muestra (por ejemplo, promedios o proporciones) deben estar acompañados del análisis que determina la precisión alcanzada o el nivel de confianza obtenido, en función del tamaño de la muestra y los otros parámetros de interés.
- Se debe llevar a cabo, al menos tres pruebas de hipótesis, utilizando el lenguaje estadístico apropiado sin importar los resultados obtenidos. Se deben priorizar aquellas pruebas de hipótesis con mayór impacto potencial en el estudio.
- Se deben estudiar desde el punto de vista de regresión al menos tres pares de variables numéricas con alguna relación aparente. Estos pares deberían haber sido identificados en la entrega anterior. Estos pares deberán ser acompañados por gráficas relevantes y el análisis pertinente de residuos y coeficientes. Los resultados deben ser reportados usando el lenguaje estadístico correcto, independientemente de los resultados obtenidos.
- Se debe hacer énfasis en el análisis inferencial, determinando si hay dependencia entre las variables postuladas en la entrega P2.
- A partir de dependencias verificadas, se incluirán conclusiones que permitan inferir el comportamiento de las variables en rangos fuera del estudiado o definir tendencias en el comportamiento de la variable.
- Las observaciones que su profesor haya hecho sobre la entrega P2 deben ser atendidas (esto vale 15 % de la nota de la entrega).

La entrega del informe ocurre un corto tiempo después de la presentación oral, por lo que un equipo bien coordinado debería poder incluir en la entrega del informe, algunas mejoras derivadas de las observaciones a la presentación oral.

2.4. P4 cuarta entrega: Entrega final

La cuarta entrega tiene 2 instancias de entrega, presentación oral e informe.

En la presentación oral de la cuarta entrega pueden participar dos personas, el líder de la entrega y una persona adicional si quedase un miembro del equipo que no haya presentado nada de manera oral durante el ciclo. Planifiquen de manera acorde.

La cuarta entrega consiste en un resumen de todos los pasos del proyecto, haciendo énfasis en las variables más relevantes del proyecto, las cuales han sido objeto del análisis inferencial y predictivo. Se debe enfatizar el nivel de confianza o la precisión de los estimados en función de los parámetros del proyecto y los datos obtenidos, todo esto debe ir acompañado de un análisis descriptivo sólido que justifique las decisiones tomadas y visualizaciones apropiadas que refuercen estas. Finalmente se hará una reflexión sobre las preguntas iniciales del proyecto, propuestas en la primera entrega y su consecuente respuesta consecuencia del análisis descriptivo, exploratorio, inferencial y predictivo realizado. Las respuestas deberán basarse en los datos del proyecto.

Los integrantes del grupo se deben registrar de nuevo, notando si ha habido algún cambio en la constitución del grupo.

Se debe entregar, de manera obligatoria, en un archivo comprimido en formato ZIP y sin directorios internos, siguiendo los lineamientos de la entrega anterior.

Para esta presentación se deberá tener en cuenta lo siguiente:

- Todo el análisis del conjunto de datos deberá realizarse en RStudio.
- Los grupos deberán hacer su presentación oral dentro del tiempo establecido.
- Se deben incluir los aspectos más relevantes de todo el trabajo, desde estadística descriptiva hasta estadística predictiva.
- Se deben presentar conclusiones del trabajo en función del cumplimiento de los objetivos originales y las respuestas de las preguntas que el estudio pretende responder.
- Las observaciones que su profesor haya hecho sobre la entrega P3 deben ser atendidas (esto vale 15 % de la nota de la entrega).

La entrega del informe ocurre un corto tiempo después de la presentación oral, por lo que un equipo bien coordinado debería poder incluir en la entrega del informe, algunas mejoras derivadas de las observaciones a la presentación oral.