Universidad de Costa Rica
Escuela de Ciencias de la Computación e Informática
CI-0115 Probabilidad y Estadística
Prof. Kryscia Daviana Ramírez Benavides
Prof. José Guevara Coto

PROYECTO: ANÁLISIS DE DATOS

OBJETIVOS

Objetivo General

Realizar un diseño de experimentos para poner en práctica los diferentes contenidos desarrollados en el curso como parte del proceso de enseñanza-aprendizaje.

Objetivos Específicos

- Analizar un conjunto de datos, dados por los docentes, con el fin de realizar observaciones (completamente exploratorias), y a partir de estas, surjan inquietudes que puedan ser planteadas como hipótesis experimentales.
- Realizar un análisis descriptivo del conjunto de datos de acuerdo con la hipótesis experimental planteada con base en las observaciones de la etapa previa.
- Realizar un análisis inferencial del conjunto de datos, realizando la transición del análisis exploratorio de datos (descripción del conjunto de datos) y el análisis descriptivo realizado en las etapas previas.

DESCRIPCIÓN

En este proyecto se realiza en equipos (equipos colaborativos), y consiste en la realización de un diseño de experimento como proyecto del curso, en el cual se deben aplicar los contenidos del curso.

Cada equipo (equipos colaborativos) debe realizar un diseño experimental dividido en las siguientes etapas del proyecto:

- **I Etapa.** Análisis Básico del Conjunto de Datos.
- **II Etapa.** Estadística Descriptiva y Visualización de Datos.
- **III Etapa.** Estadística Inferencial e Interpretación de Datos y Resultados.

ASPECTOS METODOLÓGICOS

Todos los estudiantes deben organizarse en equipos de 4 personas (como máximo) para realizar este proyecto. El trabajo consiste en realizar la presentación del diseño, el análisis del conjunto de datos, y la estadística descriptiva e inferencial de un conjunto de datos dados por los docentes.

Se deben realizar tres entregables, uno por etapa, cada uno representa un esfuerzo grupal. Los entregables deben realizarse siguiendo la plantilla del curso y las indicaciones dadas en cada uno, la letra de las entregas debe ser Cambria, Arial, o Times New Roman, tamaño: 10-12. Debe mantener el orden establecido en las instrucciones. Se debe incluir el código como adjunto a la entrega en formato R en un archivo comprimido.

ETAPAS DEL PROYECTO

I Etapa – Análisis Básico del Conjunto de Datos

Esta etapa representa la primera fase del proyecto, está inspirada en el método científico (ver Figura 1), en donde a partir de una serie de observaciones se deben plantear una o varias preguntas que emergen de las observaciones. En esta entrega se deben enfocar en

los pasos: Observar, Plantear Preguntas y Generar una Hipótesis para elaborar sus reportes.

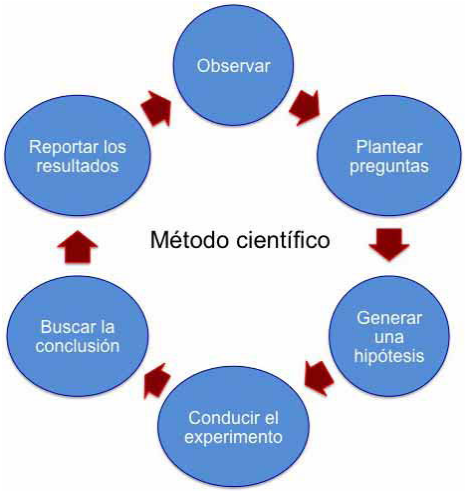

Figura 1. El método científico y sus pasos.
Fuente: Int. J. Morphol., 35(3):1031-1036, 2017.

Es decir, con base en un conjunto de datos dado por los docentes, realicen diversas observaciones (completamente exploratorias), y a partir de estas ellas, surjan inquietudes que puedan ser planteadas como preguntas. Estas preguntas les permitirán generar una suposición que podría explicar un fenómeno, evento o proceso. Esto es conocido como el planteamiento de la hipótesis experimental o de investigación. Finalmente, se debe establecer el objetivo general de su proyecto, el cual representará la idea central y el fin de su trabajo.

En este trabajo deben utilizar referencias bibliográficas, las cuales pueden provenir del uso de fuentes para explicar el contexto de sus datos o de librerías/paquetes para realizar las observaciones de sus datos.

La primera entrega debe contener como mínimo las siguientes partes:

1. **INTRODUCCIÓN/ANTECEDENTES.** Debe ser una reseña de los datos y el contexto de estos. Se debe procurar responder dentro de una estructura de prosa las siguientes preguntas: ¿Qué estoy haciendo? ¿Por qué lo estoy haciendo? ¿Con qué lo estoy haciendo? ¿Cuál es el motivo de lo que estoy haciendo? Los antecedentes deben de una manera u otra responder esto. Recuerden que la construcción de los antecedentes les permitirá reducir el trabajo para la III Etapa, debido a que pueden utilizar estos antecedentes y expandirlos para la última entrega.
2. **HIPÓTESIS EXPERIMENTAL.** La definición de una hipótesis con base en el método científico está dada de la siguiente manera: “Una hipótesis es la suposición de algo que podría, o no, ser posible”. En este sentido, la hipótesis es una idea o un supuesto a partir del cual nos preguntamos el porqué de una cosa, bien sea un fenómeno, un hecho o un proceso. En esta definición se considera que esta suposición es el resultado de la **observación**, por consiguiente, debe realizarse un proceso en el que deben observar los datos, y plantearse preguntas interesantes resultantes de las observaciones. De estas preguntas pueden surgir las suposiciones que podrían ser posibles y que podrían explicar un fenómeno. Por ejemplo, en el caso del “Titanic”, se puede decir que una hipótesis experimental podría estar dada de la siguiente manera: “La sobrevivencia de los tripulantes está determinada por factores como la Clase, la Edad y la Tarifa”. Notamos que en este enunciado o definición no se hace cuenta de valores o parámetros estadísticos. Este tipo de hipótesis que toma en consideración parámetros estadísticos se denomina hipótesis estadística. Este tipo de hipótesis estadística es parte de la III Etapa.

3. **OBJETIVO GENERAL.** El objetico general debe representar lo que ustedes como equipo buscan lograr en este trabajo. La definición por excelencia de un objetivo general es un enunciado que resume la idea central y la finalidad de un trabajo. Es importante que a la hora de definir un objetivo general no lo confundan con una actividad. El objetivo general del trabajo es el cometido de su trabajo, el fin por el cual se encuentran realizando este proyecto de análisis de datos. Por ejemplo, puede ser en el caso del Titanic dependiendo del enfoque vayan a darle puede ser: a) Determinar las características que afectan la sobrevivencia de los pasajeros del Titanic; o b) Identificar las posibles diferencias en la sobrevivencia de los pasajeros del Titanic por medio de la comparación de las variables Clase, Edad y Tarifa. Cada objetivo tiene un enfoque distinto y seguirá un camino diferente. Esto es muy importante, porque les permitirá aclarar el paso de sus proyectos de análisis. En el primer caso, su objetivo dicta que ustedes en las fases 2 y 3 deben ir direccionados a la identificación de aquellas características que determinarán la sobrevivencia de los pasajeros. Es decir, se buscarán asociaciones. En el caso del segundo ejemplo, lo que se plantea es determinar la estratificación de la población de estudio (separamos el Titanic en estratos usando algún criterio). Por ejemplo, las clases son 3 estratos, la edad se puede separar en 2 estratos, y la tarifa se puede usar con 2 estratos usando la mediana. Noten que esta línea de trabajo es diferente a la primera. Por lo que, deben apegarse a su objetivo, el cual les ayudará a definir el camino de su metodología, la cual se define en la II Etapa.
4. **REFERENCIAS.** Recuerden que si usan citas externas las deben referenciar. Por ejemplo: información acerca de sus datos, información para la elaboración de sus antecedentes, el uso de paquetes en R como ggplot2, dplyr o tidyverse. Las referencias las deben hacer en formato APA o en IEEE.

II Etapa – Estadística Descriptiva y Visualización de Datos

Esta etapa representa la segunda fase de su proyecto, en donde a partir de un proceso de observación del conjunto de datos, y la hipótesis experimental planteada en la etapa previa (con base en sus observaciones) deben realizar un análisis descriptivo de sus datos. En la I Etapa construyeron un proceso basado en el método científico, donde a partir de un conjunto de datos realizaron observaciones, y a partir de estas observaciones plantearon una idea de trabajo o una pregunta. Esta pregunta la definieron como su hipótesis experimental. Es decir, en este momento cuentan con información del conjunto de datos dado, su estructura, sus dimensiones y a partir de sus observaciones tienen una idea de que preguntas les interesaría resolver.

Deben realizar una exploración descriptiva del conjunto de datos, lo cual significa que deben calcular los parámetros descriptivos que consideren relevantes para este conjunto. Por ejemplo, existen parámetros que tradicionalmente se identifican durante los análisis descriptivos tales como: media, moda, mediana, desviación estándar y varianza. Asimismo, puede identificar los cuartiles o diferentes percentiles. La decisión de parámetros es de cada equipo, pero debe estar fundamentada con base en el contexto de los datos. En esta fase de análisis descriptivo también se debe contar con la visualización de datos. Este proceso es sumamente importante.

La segunda entrega debe contener como mínimo las siguientes partes:

1. **INTRODUCCIÓN/ANTECEDENTES.** Debe ser una reseña de los datos y el contexto de estos. Se debe procurar responder dentro de una estructura de prosa las siguientes preguntas: ¿Qué estoy haciendo? ¿Por qué lo estoy haciendo? ¿Con qué lo estoy haciendo? ¿Cuál es el motivo de lo que estoy haciendo? Los antecedentes deben de una manera u otra responder esto. Recuerden que la construcción de los antecedentes les permitirá reducir el trabajo para la III Etapa, debido a que pueden utilizar estos antecedentes y expandirlos para la última entrega.

2. **HIPÓTESIS EXPERIMENTAL.** La definición de una hipótesis con base en el método científico está dada de la siguiente manera: "Una hipótesis es la suposición de algo que podría, o no, ser posible". En este sentido, la hipótesis es una idea o un supuesto a partir del cual nos preguntamos el porqué de una cosa, bien sea un fenómeno, un hecho o un proceso.
3. **OBJETIVO GENERAL.** El objetico general debe representar lo que ustedes como equipo buscan lograr en este trabajo. La definición por excelencia de un objetivo general es un enunciado que resume la idea central y la finalidad de un trabajo. El objetivo general es una parte esencial porque va a ayudar a determinar la metodología a seguir en análisis posteriores. El objetivo general fue definido en la I Etapa.
4. **METODOLOGÍA.** Redactar en prosa los diversos métodos o enfoques que han utilizado para analizar sus datos, NO hagan una lista o un cuadro. No digan: Media = función mean(). Al contrario, por ejemplo: para el cálculo de las estadísticas descriptivas como: media, varianza, y desviación estándar se utilizó R básico, que posee las funciones… o algo similar. Lo mismo para gráficos: utilizando el paquete ggplot2, que viene en el tidyverse se generaron los gráficos de caja o *boxplots*, a partir del subconjunto de datos de área de casa (en metros cuadrados).
5. **RESULTADOS.** Presentar las gráficas y las tablas obtenidas de los análisis descriptivos. Es importante recordar que este es una entrega donde para este momento deben tener resultados de estadística descriptiva. La presentación de resultados debe venir acompañada de observaciones, pero no deben realizar conclusiones. Esto es importante porque de los gráficos y los parámetros (medias, medianas, varianzas, cuartiles, etc) NO se pueden realizar inferencias. Lo que SI es posible realizar son observaciones interesantes. Por ejemplo, pueden interpretar que en un gráfico de cajas o *boxplot* de las clases de pasajeros del Titanic y la distribución de Edades, pueden observar que existe una tendencia a que en la primera clase haya más adultos mayores que en la segunda clase o en la tercera. Otro caso, comparar un conjunto de datos de resistencia al calor, y tiene 2 materiales, compuesto X y compuesto Y, mediante un *boxplot* puede ver que la mediana de Y es superior a la de X, que ambos poseen valores extremos, y que por lo tanto, es posible pensar que Y posea una mayor resistencia al calor, o al menos proponer esta idea. Note la diferencia entre interpretar un gráfico y los valores que componente a este, y llegar a una conclusión. En los casos anteriores en ningún momento se dice: El componente Y posee mayor resistencia al color que X, o la clase 1 del Titanic posee más cantidad de adultos mayores que las demás clases. Estas son inferencias estadísticas que se llevarán a cabo con pruebas de contraste en la III Etapa.
6. **REFRENCIAS.** Recuerden que si usan citas externas las deben referenciar. Por ejemplo: información acerca de sus datos, información para la elaboración de sus antecedentes, el uso de paquetes en R como ggplot2, dplyr o tidyverse. Las referencias las deben hacer en formato APA o en IEEE.

III Etapa – Estadística Inferencial e Interpretación de Datos y Resultados

Esta etapa representa la tercera y última fase de su proyecto, en donde debe realizar la transición del análisis exploratorio de datos (descripción del conjunto de datos), y el establecimiento de una hipótesis o conjetura experimental a la prueba de principio; la cual debe estar fundamentada en observaciones extraídas del análisis exploratorio de datos. Por ejemplo, si se observa una tendencia llamativa en un diagrama que potencia un análisis de comparación entre 2 muestras, esto podría ser el punto de inicio para el establecimiento de una hipótesis. Por consiguiente, para esta fase cada equipo debe realizar las siguientes tareas:

a. Determinar aquellos parámetros estadísticos necesarios para analizar el conjunto de datos (media, varianza, desviación estándar).

b. Plantear las pruebas de hipótesis necesarias, utilizando los correspondientes parámetros.
c. Interpretar los resultados de las pruebas de hipótesis realizadas.
d. Realizar conclusiones a partir del contexto (datos) y los resultados obtenidos de la prueba de hipótesis.

La tercera entrega debe contener como mínimo las siguientes partes:

1. **INTRODUCCIÓN.** Contexto del trabajo. Reseña del conjunto de datos, sus características y propiedades, en donde se puede referir únicamente a las variables que va a analizar en el trabajo. Debe incluir la justificación del trabajo en la introducción, es decir, el porqué de su trabajo, o el fin de su trabajo, el cual se describe en el último párrafo de su introducción. Si ya ha avanzado en la elaboración de la introducción, o posee una introducción que cumple con estos requisitos de una entrega previa, puede utilizarla para esta entrega.
2. **OBJETIVO GENERAL.** En un enunciado debe resumir la idea central y la finalidad del trabajo. El objetivo general ha sido definido previamente en la I y II Etapas.
3. **HIPÓTESIS EXPERIMENTAL.** Una hipótesis es la suposición de algo que podría, o no, ser posible. En este sentido, la hipótesis es una idea o un supuesto a partir del cual nos preguntamos el porqué de una cosa, bien sea un fenómeno, un hecho o un proceso. La hipótesis experimental ha sido definida previamente en la I y II Etapas.
4. **HIPÓTESIS ESTADÍSTICA.** La o las hipótesis de tipo estadísticas que el proyecto busca probar. Debe describir para cada una que significa cada parámetro, es decir, media muestral, media poblacional, y si va a llevar a cabo una prueba de comparación de medias (T) o media de muestra con media de población (Z), por ejemplo. En esta sección solo debe hacer mención y no descripción, la descripción la lleva a cabo en la metodología.
5. **METODOLOGÍA.** En esta sección deben describir los análisis que usan para sus conjuntos de datos. Es recomendable que para este momento lo separen en Descriptiva e Inferencial. Si poseen, de entregas previas metodologías elaboradas, pueden hacer uso de estas, pero con la precaución de mantener: a) cohesión de la metodología, b) legibilidad entre la metodología anterior y la nueva y c) se va a concatenar código que todo sea reproducible (se debe correr los análisis y que estos produzcan resultados iguales a los presentados en el informe/entrega). La metodología debe venir en prosa, pero recordar que no debe ser un abuso a la verbosidad. Por ejemplo: “Prueba de Comparación de Medias: Las dos muestras se contrastaron usando una prueba *T* de dos colas. Para determinar el tipo de prueba *T* (de varianza igual o diferente) se llevó a cabo una prueba de varianzas (prueba *f*). Se utilizó el valor de significancia de 0.05 definido por R. Los resultados de esta prueba fueron reportados en la tabla de comparación de medias.”
6. **RESULTADOS.** En esta sección debe describir los hallazgos de lo que se ha llevado a cabo, además de incluir las tablas de las pruebas estadísticas que se han realizado; deben incluir los gráficos de análisis descriptivo. Los resultados deben ir organizados para narrar la historia del análisis de datos desde la descripción de los datos hasta la inferencia estadística. Si tienen avances de trabajos previos pueden concatenarlos a esta entrega, pero deben tener cuidado de ajustar su sección para que mantenga una cohesión con la escritura, estilo y por lo tanto sea coherente con el resto del texto. Asimismo, si sus resultados no son consistentes los trabajos serán penalizados con base en la rúbrica de la plantilla. Los resultados deben ser contextualizados y no solamente deben ser: “Tabla 1, valores de chi cuadrado. Tabla 2, valores de prueba *T*, donde se acepta la hipótesis nula.”, esto es incorrecto. Deben explicar los resultados obtenidos dentro de lo que significan en los datos que presentan. Por ejemplo: “Prueba de Contraste de Medias de Televidentes: Los resultados de la prueba *f* indicaron que las muestras A y B poseen varianzas estadísticamente diferentes (Tabla 1), por lo que se debe utilizar una prueba de comparación de dos medias con varianzas diferentes. Los

resultados de esta prueba de 2 colas (Tabla 2) indicaron que, para un grado de significancia de 0.05 existen diferencias significativas entre las medias de la muestra A y B."

7. **CONCLUSIONES.** En un párrafo (3-4 oraciones) deben resumir los hallazgos principales del trabajo. Deben de una manera muy sencilla exponer lo que intentaron lograr o lo que lograron hacer y, además les hubiera gustado hacer. Por ejemplo: "En este trabajo se realizó una comparación de dos muestras pertenecientes a la población costarricense con el fin de determinar sus hábitos de consumo de televisión. Ambas muestras se compararon mediante una prueba *T* de Welch, y se logró determinar que la muestra A, proveniente de la Gran Área Metropolitana (GAM) es estadísticamente diferente que la muestra B proveniente del Pacífico Sur. Análisis posteriores identificaron que la media de A era significativamente superior a la de B. En un futuro sería interesante estudiar por qué existen estas diferencias entre comunidades de la GAM y del Pacífico Sur."
8. **REFERENCIAS.** En esta sección debe incluir los recursos que se han utilizado para sustentar el trabajo. Por ejemplo, debe citar los paquetes de R, y los recursos adicionales que le ayuden a sustentar sus conclusiones. Por ejemplo, si usted decide que la calidad del vino es afectada por el azúcar y encuentra una referencia que lo apoya (lo cual debe hacer), debe incluirla. Si usted identifica que la calidad de la gasolina afecta en el consumo del vehículo, o el cilindraje o caballos de fuerza (muy probable), deben incluirlo aquí. Pueden utilizar alguno de los siguientes formatos: APA o IEEE.

EVALUACIÓN Y MEDICIÓN

I Etapa. Análisis Básico del Conjunto de Datos	5%
II Etapa. Estadística Descriptiva y Visualización de Datos	12.5%
III Etapa. Estadística Inferencial e Interpretación de Datos y Resultados	12.5%
Total	**30%**

NOTAS IMPORTANTES

- El proyecto se puede realizar en grupo de cuatro personas como máximo (equipos colaborativos).
- Se formarán los grupos el primer día de clases.
- En cada entrega se debe entregar la documentación respectiva y la división del trabajo en la hora de la clase, esto será una prueba de la entrega del trabajo asignado.
- Se debe adjuntar la autoevaluación y coevaluación de cada miembro del equipo, las plantillas de ambas evaluaciones se encuentran en la pestaña "Links". La calificación final la investigación será el promedio entre la calificación obtenida y las evaluaciones (autoevaluación y coevaluación).
- Los documentos se deben enviar vía correo electrónico con el subject "Proyecto: Etapa # – Equipo #" (por ejemplo: Proyecto: Etapa 1 – Equipo 1), y los archivos que se adjunten deben venir con el nombre "ProyectoEtapa#_Equipo#.ext" (por ejemplo: ProyectoEtapa1_Equipo1.doc).
- Cada semana se realizarán sesiones semanales del trabajo realizado en el proyecto, aproximadamente 4 minutos por equipo. En estas sesiones cada equipo debe comentar el avance, los problemas encontrados y el trabajo a realizar en la semana siguiente.

FECHAS DE ENTREGA

- Entrega del documento de la I Etapa: Jueves 10 de septiembre, hora de clase.
- Entrega del documento de la II Etapa: Jueves 29 de octubre, hora de clase.
- Entrega del documento de la III Etapa: Jueves 19 de noviembre, hora de clase.
- Entrega final: Jueves 26 de noviembre, hora de clase.