PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA
Departamento de Ingeniería Química y Bioprocesos (DIQB)
www.ing.puc.cl/iiq

IIQ3402 Diseño estadístico, Optimización y Análisis Multivariado

Tarea 1

Fecha de entrega: miércoles 16 de abril vía Canvas hasta las 23h59
Modalidad: En parejas o individual. Los estudiantes escogen libremente cómo y con quién trabajar.

I. Descripción

El propósito de esta tarea es evaluar las competencias asociadas a las Unidades 1 y 2 del curso en un caso de estudio real. Esta tarea está diseñada para que los estudiantes exploren y formulen preguntas de investigación a partir de los métodos revisados hasta ahora.

En esta tarea sea analizará una investigación con datos reales relacionada con la construcción de un modelo estadístico para la predicción del estatus de enfermedad COVID-19 a partir de síntomas auto-reportados y resultados de una prueba olfativa [1]. Datos sobre infecciones con las primeras cepas de SARS-CoV-2 indicaban la aparición del síntoma llamado anosmia (pérdida del sentido del olfato) como un síntoma temprano y discriminatorio de esta enfermedad [2]. Usando estos datos deberán plantear hipótesis razonables que puedan ser testeadas, así como realizar un análisis exploratorio de ellos tal que permitan abordar efectivamente las preguntas de esta tarea.

II. Instrucciones

La tarea consta de res (3) partes las cuales se deben ver reflejadas y desarrolladas en el reporte de la tarea. Para la realización de la tarea se deben analizar los datos disponibles indicados en la siguiente sección.

1. **Análisis teórico de los datos (15 ptos.):** Describir cómo se realizó el muestreo y cómo se recopilaron las observaciones. Indique además el tipo de estudio realizado y justifique cómo lo clasificaría en base a la Figura 1. En particular, esta sección debe discutir y justificar claramente los siguientes puntos:

a. ¿Cómo se realizó el muestro y cómo fueron recopilados los datos? ¿Qué tipo de datos contiene el estudio? ¿Qué tipo de estudio es? ¿Por qué?
b. ¿Qué tipo de implicancia tiene este tipo de recopilación de datos en el alcance de la inferencia (generalización o causalidad)? ¿Qué limitaciones muestra este estudio?
c. ¿Cómo validaría los resultados/inferencias en un siguiente estudio? ¿De qué estudio se trataría?

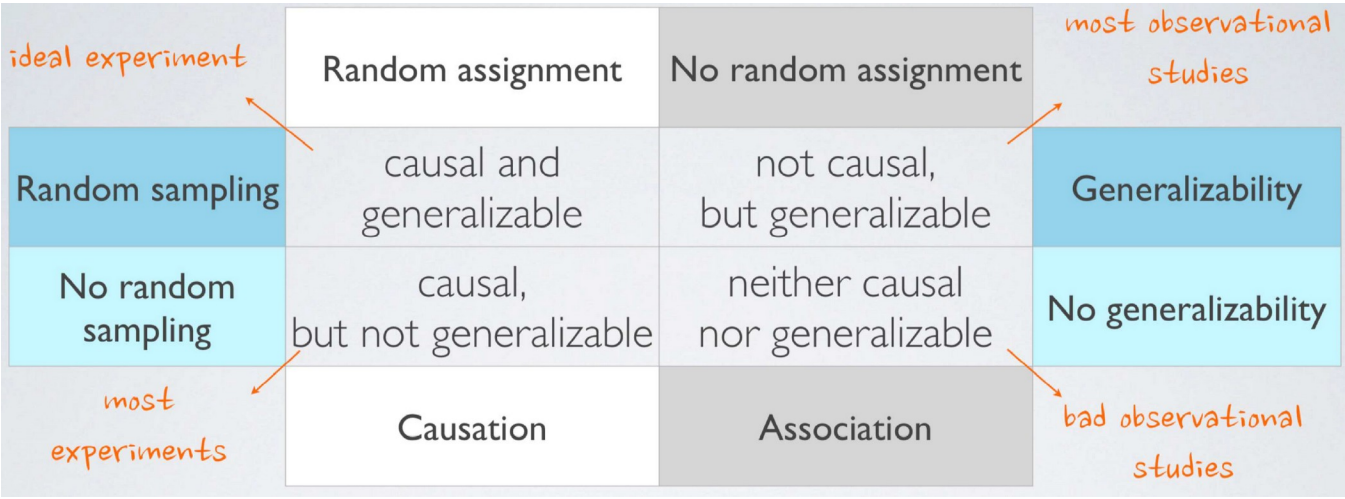

Figura 1. Tipos de estudio.

2. **Formulación de preguntas de investigación o hipótesis (15 ptos.):** Proponer al menos tres (3) preguntas de investigación o hipótesis –diferentes a las abordadas en el estudio–, que podrían ser abordadas con los datos entregados. Más abajo se muestra un ejemplo ilustrativo de cómo deberían ser formuladas la pregunta e hipótesis:
   - Pregunta de investigación: ¿Afectará fumar habitualmente la salud del público general?
   - Hipótesis: Fumar habitualmente conduce a un deterioro en la salud general.

   Es importante remarcar que las propuestas de preguntas/hipótesis deben estar alineadas con el alcance de la inferencia posible a partir de los datos entregados y el análisis del punto 1. Dos (2) de las preguntas/hipótesis propuestas deberán involucrar al menos tres (3) variables explicativas. Si lo desea, puede crear nuevas variables explicativas a partir de las existentes. La creación y uso de dichas variables debe ser justificado debidamente. Por último, cada propuesta deberá ser justificada explicando por qué es relevante en el contexto de este estudio y de la literatura científica en general (máx. 10 líneas de justificación). Incluye referencias relevantes de ser necesario.

3. **Análisis exploratorio de datos (EDA) (20 ptos.):** En esta sección deberán realizar EDAs atingentes a cada una de las tres (3) preguntas/hipótesis propuestas. Su EDA debe incluir visualizaciones y

resúmenes estadísticos (descriptores) relevantes. Estos resultados deben ser brevemente interpretados y discutidos.

4. **Explorando datos cualitativos (10 ptos.):** Este trabajo busca desarrollar un modelo estadístico para la predicción del estatus COVID-19 a partir de los resultados de una prueba olfativa. Para explorar este punto se le pide que siga los siguientes pasos:
    a. Usando los datos de reconocimiento de olores construya la variable continua *score* (y) como sigue:

$$y = \sum_{i=1}^{6} w_i x_i$$

Donde x_i denota si el individuo reconoce el olor (1 = SÍ, 0 = NO) y w_i es el peso que tiene ese olor en la predicción del estado olfativo de la persona. Considere que el valor de estos pesos es el siguiente.

Tabla 1. Pesos de los aromas del test olfativo

Aroma	Peso (w_i)
Plátano	0.09332
Caramelo	0.09333
Menta	0.34668
Naranja	0.09334
Piña	0.18667
Vainilla	0.18666

    b. Realice un EDA con la nueva variable score (y) y evalúe su influencia en el estatus de COVID-19. Comente si esta variable es útil para la predicción del estatus de COVID-19.

    c. **Bonus (10 ptos.)** Explique qué es la regresión logística y cómo se podría usar en este caso. Además, genere un modelo logístico que permita discriminar el estatus de infección COVID-19 usando la variable *score*. Comente sobre la utilidad de este modelo.

hint 1: Revise la bibliografía mínima disponible en Canvas para saber más acerca de regresiones logísticas.

hint 2: La siguiente página contiene información sobre funciones útiles de Matlab para realizar regresiones logísticas (https://la.mathworks.com/help/stats/generalized-linear-regression-1.html)

III. Material asociado

Para el desarrollo de la tarea se provee:

- Enunciado de la tarea.
- Datos crudos en formato `.csv` (`datos_tarea1.csvtxt`)
- Descripción de los datos en archivo `.txt` (`descripción_variables.txt`)
- Artículo científico sobre el estudio realizado.

Todos estos archivos pueden ser descargados desde el buzón respectivo en CANVAS.

IV. Entrega

El plazo último de entrega es el **miércoles 16 de abril a las 23:59** en el buzón de entrega correspondiente habilitado en CANVAS. Los documentos indicados más abajo deberán ser subidos como un sólo archivo comprimido en formato `.zip`:

- Archivo `.pdf` con el documento que contiene la resolución de la tarea con el nombre `Tarea_1_grupo_X.pdf`, donde `X` se refiere al número del grupo. El número de grupo se asignará en Canvas una vez que los grupos se conformen.
- Archivo(s) `.ipynb` (Jupyter *notebook*) con scripts de Python utilizados para la resolución completa de la tarea. Si el grupo lo estima conveniente, se pueden agregar múltiples Jupyter *notebooks* para mantener el orden y facilitar la revisión. Los archivos asociados deben ser rotulados usando el siguiente formato: `Tarea_1_grupo_X_script_Y.pdf`, donde `X` corresponde al número del grupo e `Y` corresponde al número correlativo que indica el número del Jupyter *notebook*. Es importante destacar que los Jupyter *notebooks* serán revisados en conjunto con el reporte de la tarea por lo que se recomienda que éstos estén debidamente comentados y ordenados.

V. Integridad Académica

Esta evaluación se adscribe el Código de Honor establecido por la Escuela de Ingeniería el que es vinculante. Todo trabajo evaluado en este curso debe ser propio. En caso de que exista colaboración permitida con otros estudiantes, el trabajo deberá referenciar y atribuir correctamente dicha contribución a quien corresponda. Como estudiante es su deber conocer la versión en línea del Código de Honor (https://integridadacademica.uc.cl). Los trabajos serán sometidos a revisión por plagio utilizando la

herramienta Turnitin. Aquellos/as estudiantes que sean sorprendidos en conductas reñidas con la integridad académicas serán penalizados con nota mínima en la tarea (1.0), y dependiendo de la gravedad, los antecedentes serán enviados a las Unidades respectivas para sanciones más fuertes.

VI. Referencias

[1] Eyheramendy S., Saa PA., Undurraga EA., Valencia C., López C., Mendez L., Pizarro-Berdichevsky J., Finkelstein-Kulka A., Solari S., Salas N., Bahamondes P., Ugarte M., Barceló P., Arenas M., Agosin E. (2021) Screening of COVID-19 cases through a Bayesian network symptoms model and psychophysical olfactory test. *iScience*. https://doi.org/10.1016/ j.isci.2021.103419.

[2] Menni, C., Valdes, A.M., Freidin, M.B. et al. (2020) Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine*. https://doi.org/10.1038/s41591-020-0916-2.