Tarea 1

Estadística descriptiva y conceptos básicos
valor: 15 % de la calificación final

Fecha de entrega: 28 de febrero del 2025

Elaboren un reporte en formato PDF en el que respondan cada una de las siguientes preguntas.

Importante: para que el reporte sea considerado para evaluación, debe cumplir con los siguientes requisitos:

1. Incluir los nombres de los integrantes del equipo (máximo res alumnos) en la parte superior derecha de la primera página, sin ocupar demasiado espacio.
2. Presentar un formato claro y profesional.
3. No exceder un máximo de 15 páginas.
4. No incluir código ni partes de bases de datos en el documento.
5. Utilizar el espacio de manera eficiente, evitando desperdicios.
6. Incorporar gráficas bien diseñadas, asegurando que las escalas y distribución sean apropiadas.
7. Justificar cada respuesta con argumentación concisa, bien estructurada y redactada en prosa.
8. Pueden usar el lenguaje de programación que prefieran.

1. **Censo de Población y Vivienda 2020**. Cada 10 años, el Instituto Nacional de Estadística y Geografía (INEGI) levanta el Censo de Población y Vivienda con el objetivo de conocer diversas características de los habitantes de México y sus viviendas a nivel nacional, estatal, municipal, por localidad, por grupos de manzanas y hasta por manzana.

   Descarga la base de datos con los principales resultados por localidad (ITER) de 2020 de los Estados Unidos Mexicanos y genera la siguiente información:

   1.1. **Base de datos a nivel municipio.** Extrae directamente los sub-totales para cada municipio y usa esta base para responder las siguientes preguntas:
   - ¿De qué tamaño es tu base de datos? (considera el número de renglones × columnas y el peso en MB).
   - Obtén la población total de la República Mexicana y busca alguna fuente confiable para corroborar tu resultado.
   - No filtres ni agregues sobre las localidades para cada municipio. ¿Por qué esto sería una pésima idea?

1.2. **Análisis de variables socioeconómicas.** Calcula los siguientes indicadores para cada municipio:

- Porcentaje de población de 5 años y más que habla alguna lengua indígena.
- Porcentaje de población de 15 años y más que es analfabeta.

Ambos porcentajes deben calcularse con respecto a las poblaciones correspondientes de cada municipio.

- ¿Existe alguna relación entre ambas variables? Para responder, realiza un diagrama de dispersión.

1.3. **Análisis de la distribución de población por entidad federativa.** Realiza un boxplot o un histograma (o ambos) de la distribución de la población en las entidades federativas.

- Con base en las gráficas, decide si calcular una media truncada o la mediana y compárala con la media muestral. Justifica tu elección.
- Calcula un estimador de dispersión de tu elección y justifica su uso.
- Con base en tu análisis, ¿qué puedes concluir sobre la distribución de la población en las entidades federativas?

Hint: Puedes consultar el siguiente tutorial en R, en donde se describen algunas cosas que probablemente te ayudarán. Sin embargo, en la pregunta se piden cosas distintas.

2. **Pandemia de COVID-19 en la CDMX desde el año de 2020 a 2021**.

2.1. **Descarga y filtrado de datos.** Descarga los datos de la pandemia a nivel nacional, en la sección "Bases de datos históricas Influenza, COVID-19 y otros virus respiratorios", descarga Cierre Datos Abiertos Históricos para los años 2020 y 2021. Descomprime ambas bases, cárgalas en R (Python o el lenguaje de tu preferencia, pero ten en cuenta que tendrás que manejar y/o transformar variables a formato tipo fecha) y en los dos casos

- Selecciona únicamente las variables necesarias: `CLASIFICACION_FINAL`, `ENTIDAD_RES`, `FECHA_SINTOMAS`, `TIPO_PACIENTE`, `FECHA_DEF` y `EDAD`.
- Filtra los casos positivos a COVID-19 (`CLASIFICACION_FINAL` con valores 1, 2 y 3) para las personas que residen en la CDMX.
- Concatena ambas bases de datos para tener una única base de datos con las seis variables de interés para 2020 y 2021.

2.2. **Casos, hospitalizaciones y fallecimientos.**

1) A partir de la fecha de inicio de síntomas (variable `FECHA_SINTOMAS`, en formato `as.Date` en R), genera la gráfica del número de casos por mes.
2) Suma 15 días a la fecha de inicio de síntomas (para estimar, aproximadamente, el día en el que se dió la hospitalización) y, considerando la variable `TIPO_PACIENTE`, obtén el número de hospitalizaciones por mes.
3) Calcula el porcentaje de hospitalizaciones por mes con la siguiente fórmula:

$$\%\text{ de hospitalizaciones} \times \text{mes} = 100 \times \frac{\text{hospitalizados (de los infectados)} \times \text{mes}}{\text{infectados de COVID-19} \times \text{mes}},$$

genera una serie de tiempo con esta información, donde el eje x representa el mes y el eje y el porcentaje de hospitalizados.

4) Realiza el mismo análisis para las defunciones. En este caso, utiliza la variable de `FECHA_DEF`, sin necesidad de estimación. Observa que cuando no hubo defunción la base de datos registra `9999-99-99`.
5) Observa las tres gráficas obtenidas y describe lo que observas.

2.3. **Hospitalizaciones y grupos de edad.**

1) Define los siguientes grupos de edad:

$$[0, 20), \quad [20, 40), \quad [40, 60), \quad 60 \text{ y mayores.}$$

2) Obtén el porcentaje de hospitalizaciones por mes para cada grupo de edad y grafica las cuatro series de tiempo.
3) Describe lo que observas en la evolución de hospitalizaciones por grupo de edad.

3. **Teorema Central del Límite**.

3.1. Considera una v.a. $X \sim \mathcal{U}(0, 10)$, calcula su esperanza ($\mu$) y varianza ($\sigma^2$). Fija $n = 5$ y realiza los siguientes pasos:

- Genera una m.a. $(X_1, X_2, \ldots, X_n) = (x_1^1, x_2^1, \ldots, x_n^1)$ de X.
- Calcula la media observada $\bar{x}_n^1$.
- Repite los dos pasos anteriores $m = 10,000$ veces para obtener $\bar{x}_n^1, \bar{x}_n^2, \ldots, \bar{x}_n^m$.
- Haz un histograma de las medias observadas $\bar{x}_n^1, \bar{x}_n^2, \ldots, \bar{x}_n^m$.
- Encima del histograma, grafica la función de densidad de probabilidad de una $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. En estadística, coloquialmente se dice que estamos "ajustando" la distribución normal a los datos.
- Repite toda la simulación para $n = 30, 100, 200$ y presenta las cuatro gráficas en una matriz de 2×2 para poder apreciar el impacto de aumentar el tamaño de muestra.

3.2. Considera una v.a. $X \sim$ Bernoulli(θ), calcula su esperanza (μ) y varianza (σ^2). Fija $\theta = 0.05$, $n = 5$ y realiza los exactamente los mismos pasos que en el ejercicio 3.1.

3.3. ¿Qué pasa si en el ejercicio 3.2 usamos $\theta = 0.3$?

3.4. Con base en el trabajo realizado en este ejercicio, ¿cómo responderías a la clásica pregunta: "¿Cuán grande debe ser el tamaño de muestra n para que la distribución muestral de la media $\bar{X}_n$ pueda aproximarse mediante el Teorema Central del Límite?"

4. **Modelos de recolección de datos básicos para hacer inferencia estadística**

En estadística *frecuentista*, un parámetro θ es desconocido pero fijo, mientras que en estadística *bayesiana*, θ es tratado como una variable aleatoria con una distribución previa. En el enfoque frecuentista, la función de densidad de los datos suele expresarse como $f(y; \theta)$, indicando que la densidad está indexada por el parámetro θ. Por otro lado, en inferencia bayesiana se denota como $f(y \mid \theta)$, resaltando que la densidad es condicional a la incertidumbre sobre θ. No obstante, en muchos textos y aplicaciones se usa la notación $f(y \mid \theta)$ incluso en el enfoque frecuentista, ya que estructurar los modelos en términos condicionales es útil en diversos contextos de inferencia.

Como hemos visto a lo largo del curso, el modelo de recolección de información en el caso frecuentista (que será el enfoque principal en este curso) asume que los datos observados $x_1, x_2, \ldots, x_n$ son una realización de las variables aleatorias $X_1, X_2, \ldots, X_n$, donde estas son

independientes e idénticamente distribuidas (i.i.d.) según la densidad $f(y;\theta)$. Lo que definimos como muestra aleatoria, i.e. m.a. $\equiv$ v.a.i.i.d.

Por otro lado, en el caso bayesiano, el modelo fundamental de recolección de información se basa en el concepto de *intercambiabilidad*. En lugar de asumir independencia, se considera que las variables aleatorias $X_1, X_2, \ldots, X_n$ forman una secuencia intercambiable si su distribución conjunta es invariante bajo permutaciones, es decir,

$$(X_1, X_2, \ldots, X_n) \stackrel{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}),$$

para cualquier permutación π de los índices. Bajo este enfoque, existe un modelo jerárquico donde los datos son condicionalmente i.i.d. dado una variable aleatoria latente θ:

$$X_1, X_2, \cdots \mid \theta \overset{\text{i.i.d.}}{\sim} f(x \mid \theta).$$

Este modelo reconoce que la incertidumbre sobre θ introduce correlación entre las observaciones, lo que según los bayesianos permite capturar mejor la estructura de dependencia en los datos. Es importante notar que en inferencia bayesiana también se pueden utilizar muestras aleatorias cuando se especifica una distribución previa para θ. De hecho, es sencillo demostrar que una muestra aleatoria es intercambiable, pero la reciprocidad no es válida en general, como veremos en el siguiente ejercicio.

4.1. Sea $X_1, X_2, \ldots$ una sucesión infinita de variables aleatorias y sea θ una variable aleatoria latente. Supongamos que, condicionadas en θ, las variables X_i son independientes e idénticamente distribuidas según la densidad $f(x \mid \theta)$, es decir,

$$X_1, X_2, \cdots \mid \theta \overset{\text{i.i.d.}}{\sim} f(x \mid \theta).$$

Demuestra que la covarianza entre cualesquiera dos variables X_i y X_j es siempre no negativa, es decir,

$$\text{Cov}(X_i, X_j) \geq 0, \quad \text{para } i \neq j.$$

Hint: Usa la propiedad de la esperanza total:

$$\mathbb{E}[X_i X_j] = \mathbb{E}[\mathbb{E}[X_i X_j | \theta]],$$

y la independencia condicional de X_i y X_j dado θ para expresar la covarianza en términos de la varianza de $\mathbb{E}[X_i|\theta]$.