Estadística para la Economía y la Gestión
Universidad de Chile - Departamento de Ingeniería Industrial
Instructor: Matías R. Labbé

Problem Set 1

Instrucciones Generales

En esta primera tarea, se espera que los/as estudiantes puedan consolidar los conocimientos obtenidos durante la primera parte del curso. En la primera parte de la tarea revisaremos algunos conceptos teóricos básicos de estadística matemática. En la segunda y tercera, se utilizará R como herramienta para realizar un análisis respecto de la evolución de algunos indicadores respecto de la historia socioeconómica de Chile de las últimas décadas. Para esto, se considerarán los datos de la Encuesta de Ocupación y Desocupación (EOD) del Centro de Microdatos de la Universidad de Chile. La EOD es una encuesta de carácter trimestral, que entrega información socioeconómica de hogares que viven en el Gran Santiago.

Cada alumno/a deberá subir a la plataforma de U-Cursos dos archivos:

- Un informe con respuestas, resultados, gráficos y tablas que sostengan su análisis, tomando en cuenta las preguntas de este enunciado. Este informe **debe** estar en formato PDF, independientemente de si fue desarrollado en Word o en LaTeX(u otra herramienta). Sea **lo más sintético posible**: a la hora de realizar un buen informe, este es capaz de resumir toda la información relevante en la menor cantidad de texto posible. Gráficos y tablas auto-explicativos son de gran utilidad. **En cada gráfico incluya la cantidad de observaciones** utilizadas en el análisis.
- Un script de R que ejecute el análisis empírico asociado al informe. Este debe correr íntegramente, sin errores.

El trabajo **es personal**. Es fundamental que, aún cuando puedan conversar entre alumnos respecto de las respuestas, cada uno pueda adentrarse de manera autónoma estudiar y responder los problemas analíticos, y, a su vez, desarrollar sus capacidades de programación en R. No se tolerará el plagio, siguiendo los estándares detallados por la Universidad de Chile en sus códigos de conducta.

El 7 estará en 160 puntos. Fecha de entrega: 5 de Abril de 2024.

1 Estadística Matemática Básica

[40 puntos]

1. Considere una variable aleatoria $\tau \sim N(\mu, 9\sigma^2)$. Estandarice y encuentre $P(\mu - \sigma < \tau \leq \mu + \sigma)$. *[10 puntos]*

2. Considere una v.a. X cuya pdf viene dada por:

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

   (a) Demuestre que la esperanza incondicional de X es $E[X] = 1/\lambda$ *[5 puntos]*

   (b) Demuestre que la varianza incondicional de X es $Var[X] = 1/\lambda^2$ *[5 puntos]*

3. Suponga que usted tiene dos variables aleatorias que distribuyen normal tal que $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$, y $Cov(X, Y) = \sigma_{xy}$. Suponga que α y β son constantes.

   (a) Muestre formalmente que $Var[X + \alpha] = \sigma_x^2$. *[5 puntos]*

   (b) Muestre que $Var[X + \alpha Y] = \sigma_x^2 + \alpha^2 \sigma_y^2$, si $\sigma_{xy} = 0$. *[5 puntos]*

   (c) Muestre que $Cov(\alpha X, \beta Y) = \alpha \cdot \beta \cdot \sigma_{xy}$. *[5 puntos]*

4. Suponga dos variables aleatorias X e Y con alguna distribución conjunta tal que $\sigma_{xy} \neq 0$. Muestre que $Var[X + Y] \neq Var[X - Y]$. *[5 puntos]*

2 Encuesta de Ocupación y Desocupación

2.1 Diciembre - 2022

[80 puntos]

En esta sección de la tarea, estudiaremos el escenario entregado por la EOD del último trimestre del año 2022. Para esto, utilice el paquete haven en R que le permitirá abrir el archivo .dta que contiene la base de datos a analizar.

1. Abra la base de datos en RStudio. Describa brevemente las variables que aparecen, y explique qué representa cada fila. Deténgase en las variables ident, numper y pcoh. ¿Para qué le sirven? *[10 puntos]*

2. Identifique la variable de ingresos del trabajo y genere un histograma que muestre la distribución de esta variable para la población activa (entre 15 y 65 años). ¿Qué problema observa? En dicho gráfico, recuerde incluir una nota que describa cuántas observaciones se consideran en la distribución.[1] *[5 puntos]*

3. Habiendo *limpiado* la base con el problema anterior, replique el histograma con sus observaciones, y presente una tabla con estadísticos básicos para la población activa (entre 15 y 65 años). Reporte en dicha tabla el número de observaciones, el promedio muestral, desviación estándar muestral, mínimos, máximos, mediana, percentil 1 y percentil 99. *[15 puntos]*

[1] Hint: El paquete ggplot2 puede serle de muchísima utilidad.

4. Utilizando la misma muestra de población activa, discuta acerca de si la variable de ingresos del trabajo distribuye normal. Use las herramientas gráficas vistas en clase para fundamentar. *[10 puntos]*

5. Considere jefes de hogar (donde la variable pcoh toma valor 1). Esto le permitirá trabajar con variables *a nivel de hogar*. En adelante:

   (a) Asigne quintiles y deciles de acuerdo al ingreso familiar. Recuerde limpiar los datos para evitar problemas como el de la pregunta 2. *[5 puntos]*

   (b) Deténgase en la variable dicotómica f5i, indicativa de si alguien en el hogar tiene deudas educacionales o no. Genere un gráfico de barras que muestre el porcentaje de hogares con deuda por quintil. ¿Qué observa? *[10 puntos]*

   (c) Replique el procedimiento anterior, esta vez por deciles. *[5 puntos]*

   (d) Genere una variable dicotómica que tome valor 0 si el hogar es de los primeros 4 quintiles, y 1 si el hogar pertenece al quinto quintil, denominándola quintil_5. Realice un test t de medias respecto de la variable f5i, en donde los dos grupos relevantes quedarán definidos por la nueva variable quintil_5. Explique formalmente su implementación, reporte una tabla con los estadísticos relevantes del test, y describa los resultados obtenidos. Brevemente, utilice la información reportada para argumentar acerca de una política de condonación de deudas estudiantiles.[2] *[20 puntos]*

2.2 Una visión de largo plazo

[80 puntos]

Para lo que sigue, usted utilizará varias versiones de la EOD, que se encuentran en el archivo comprimido subido a U-Cursos.

1. Su primera tarea es generar un *loop*[3] que permita cargar todas las bases de datos desde 1997 en adelante. Considere la Encuesta de Marzo como correspondiente al primer trimestre, la de Junio como la del segundo, la de Septiembre como la del tercero y la de Diciembre como la del cuarto y último trimestre del año.[4] La idea es que pueda combinar **verticalmente** todas las bases de datos en un solo objeto, en donde pueda identificar cada fila con el trimestre-año desde donde se obtiene. Para lo que sigue, considere solamente las siguiente variables:[5] *[15 puntos]*

   - ident
   - orden
   - numper
   - pcoh
   - sexo
   - edad
   - horas
   - ingfam
   - ingtrab
   - arriendo
   - arrienm
   - sitocup1
   - sitocup2

[2] Note que no se evaluará si su opinión es *correcta* o no, sino su capacidad de comprender los datos obtenidos e incorporarlos como evidencia para su argumentación.

[3] Aquí puede encontrar información que le puede ser útil.

[4] Hint: Note que el nombre de los archivos tiene una estructura familiar, e.g.: Diciembre2022.dta.

[5] Esto le permitirá hacer más rápido su código.

2. Considere la variable horas que reporta la cantidad de horas semanales trabajadas por cada persona. Tome la submuestra de observaciones relativas a los años 1997 y 2017, y realice un test t de medias. Explique formalmente su implementación, reporte una tabla con los estadísticos relevantes del test, y describa los resultados obtenidos. De aquí en adelante, no considere observaciones que reporten más de 100 horas trabajadas. *[10 puntos]*

3. Para cada trimestre, obtenga la media muestral de horas trabajadas e intervalos de confianza con $\alpha = 0.05$. Grafique dichas medias en el eje Y, y el trimestre/año en el eje X, con sus intervalos de confianza en el tiempo. ¿Qué conclusiones puede obtener? *[15 puntos]*

4. Replique el mismo procedimiento, pero esta vez dividiendo la población entre hombres y mujeres. Cada set de puntos/intervalos de confianza debe tener distinto color. ¿Qué conclusiones obtiene? *[10 puntos]*

5. Deténgase en los reconocidos como hijos en la muestra (i.e.: la variable pcoh toma valor 3). Genere un gráfico que tenga como eje X el tiempo (trimestre-año), y que tenga dos ejes Y (uno a la izquierda, otro a la derecha), con la siguiente información:

   (a) Media de edad con intervalos de confianza ($\alpha = 0.05$).

   (b) Media de horas trabajadas semanalmente con intervalos de confianza ($\alpha = 0.05$).

   Describa los resultados obtenidos. *[15 puntos]*

6. Ahora considere las variables de arriendo y arrienm. Considere sólo los hogares que arriendan (i.e.: la variable arriendo toma valor 1) y considere solamente un hogar (puede hacerlo restringiendo la muestra al jefe de hogar, esto es, que la variable pcoh tome valor 1). Genere una variable denominada share_arriendo que consista en el ratio del valor pagado por el arriendo (arrienm) sobre el ingreso total familiar (ingfam). Considere solamente observaciones tales que el share de arriendo tome valores entre 0 y 1, inclusive.

   (a) Genere un gráfico que muestre el valor medio del share de arriendo en el tiempo, por trimestres, con sus intervalos de confianza ($\alpha = 0.1$). ¿Dentro de que margen está en el tiempo? *[10 puntos]*

   (b) Replique el análisis anterior, pero esta vez divida la población en quintiles de acuerdo al ingreso familiar. **Los quintiles deben asignarse dentro del espacio temporal**, tenga cuidado en hacerlo sobre toda la base.[6] Grafique la evolución en el tiempo de el share de arriendo para el primer quintil y el último quintil (medias e intervalos de confianzas con $\alpha = 0.1$). ¿Qué diferencias observa? *[5 puntos]*

[6] Esto sería incorrecto: se compararían ingresos del año 2015 con el año 1999, por ejemplo. En ese sentido, el quintil más rico estaría sobrerrepresentado de observaciones más nuevas.