DEPARTAMENTO DE INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE SANTIAGO DE CHILE

EJERCICIO PRÁCTICO 11: MÉTODOS CON REMUESTREO

CONTEXTO

Recordemos que los métodos clásicos de inferencia que hemos visto hacen suposiciones sobre la forma de las distribuciones de las poblaciones estudiadas, y que algunas de estas son difíciles de asegurar en la práctica.

Hemos mencionado que existen diversas formas de enfrentar estos casos. Por ejemplo, se pueden aplicar transformaciones a los datos, en que se puede aplicar los métodos estudiados y extrapolar las conclusiones a los datos originales. Vimos que otra posibilidad es bajar la exigencia de las hipótesis y utilizar métodos no paramétricos, que no nos permiten conocer mucho sobre las poblaciones subyacentes pero sí establecer diferencias entre grupos de mediciones.

En este ejercicio vamos a practicar los métodos con remuestreo, que empiezan a ganar popularidad ya que son fáciles de aplicar en prácticamente cualquier tipo de inferencia. Su desventaja ha sido que requieren mucho cómputo, lo que cada vez es menos problemático gracias a las nuevas tecnologías de información y comunicación.

OBJETIVOS DE APRENDIZAJE

1. Plantear preguntas de investigación interesantes y, a partir de ellas, enunciar hipótesis a ser contrastadas.
2. Aplicar métodos con remuestreo en diferentes situaciones utilizando el ambiente R.

ÉXITO DE LA ACTIVIDAD

1. El equipo es capaz de plantear preguntas de investigación interesantes que pueden ser respondidas con los datos de la encuesta Casen 2017.
2. El equipo puede aplicar apropiadamente métodos con remuestro en cada caso, usando el entorno R.
3. El equipo interpreta adecuadamente los resultados de las pruebas realizadas.

ACTIVIDADES

Como habíamos visto a comienzos del semestre, la Encuesta de Caracterización Socioeconómica Nacional, Casen, es realizada por el Ministerio de Desarrollo Social de forma periódica para conocer la situación de los hogares chilenos con relación a aspectos demográficos, de educación, salud, vivienda, trabajo e ingresos. Es la principal fuente de información para estimar la magnitud de la pobreza y la distribución del ingreso en el país.

Se pone a disposición el archivo EP11 Datos.csv, con un subconjunto de los datos obtenidos en la Encuesta Casen 2017. El equipo debe revisar las columnas disponibles en este archivo según la descripción en el libro de códigos de la encuesta, que también queda disponible para este ejercicio bajo el nombre de EP11 Diccionario de datos. Es importante notar que:

- En esta encuesta hay datos de **carácter colectivo** sobre “el hogar” del entrevistado, pero también hay datos de **carácter individual**, que se refieren “al jefe o la jefa de hogar” (no al entrevistado).
- El conjunto de datos entregado **no incluye** todas las variables descritas en el libro de códigos.

1. Copiar el enunciado de los problemas asignados como comentarios de un script R.
2. Descargar desde UVirtual el archivo EP11 Datos.csv con los datos a emplear.

3. Obtener las muestras que se piden, revisarlas gráficamente y comentar la necesidad de aplicar métodos para datos problemáticos.
4. Independiente de las conclusiones anteriores, escribir código R que realice las pruebas con remuestreo en cada caso.
5. Concluir de acuerdo con los resultados de la prueba realizada.

Fuera del horario de clases, cada equipo debe subir el script realizado UVirtual con el nombre "EP11-respuesta-grupo-i", donde i es el número de grupo asignado. Las respuestas deben subirse antes de las 23:30 del sábado.

PREGUNTAS (TODOS LOS GRUPOS)

1. Propongan una pregunta de investigación original, que involucre la comparación de las medias de dos grupos independientes (más abajo se dan unos ejemplos). Fijando una semilla propia, seleccionen una muestra aleatoria de hogares ($250 < n < 500$) y respondan la pregunta propuesta utilizando una simulación Monte Carlo.
2. Propongan una pregunta de investigación original, que involucre la comparación de las medias de más de dos grupos independientes (más abajo se dan unos ejemplos). Fijando una semilla distinta a la anterior, seleccionen una muestra aleatoria de hogares ($400 < n < 600$) y respondan la pregunta propuesta utilizando bootstrapping. Solo por ejercicio académico, aplique un análisis post-hoc con bootstrapping aunque este no sea necesario.

Algunos ejemplos (que no pueden ser ocupados en este ejercicio) son:

- En promedio, el ingreso per cápita (ytotcorh / numper) en la Región Metropolitana (region) es el mismo entre hombres y mujeres (sexo) no heterosexuales (r23).
- El ingreso per cápita promedio es similar en las cuatro macro zonas (norte grande, norte chico, central, sur y austral).
- El arriendo promedio que se paga por viviendas similares a la habitada (v19) tiene relación con el nivel educacional (educ) del jefe o la jefa del hogar.

CRITERIOS DE EVALUACIÓN

Pregunta 1:

- Proponen una pregunta de investigación, interesante y novedosa, que involucra la comparación de las medias de dos grupos independientes de personas encuestadas en la Encuesta Casen 2017.
- Obtienen una muestra de datos de acuerdo a lo solicitado, revisando su comportamiento con gráficos o pruebas estadísticas y pronunciándose explícitamente sobre la necesidad de utilizar métodos para datos problemáticos.
- Formulan explícitamente hipótesis nula y alternativa correctas, que involucran la comparación de las medias de una variable numérica de dos grupos independientes, para responder la pregunta de investigación que plantean.
- Basándose en el análisis anterior, proponen explícitamente un estadístico a remuestrear que permite docimar las hipótesis propuestas, justificando su elección apropiadamente.
- Realizan, de forma completa y sin errores, una simulación Monte Carlo de un estadístico que permite responder la pregunta de investigación que plantean, usando una muestra de datos adecuada, obteniendo un p valor o intervalo de confianza correcto.

- Entregan una conclusión correcta y completa a la pregunta de investigación que plantean, basándose en el resultado del p valor o intervalo de confianza obtenido a partir del estadístico remuestreado.
- Escriben código R -ordenado, bien indentado, sin sentencias espurias y bien comentado- que realiza de forma completa y correcta la prueba seleccionada con los datos adecuados en cada caso.
- Escriben con buena ortografía y redacción (<3 errores), usando vocabulario propio de la disciplina y el contexto del problema.

Pregunta 2:

- Proponen una pregunta de investigación, interesante y novedosa, que involucra la comparación de las medias de más de dos grupos independientes de personas encuestadas en la Encuesta Casen 2017.
- Obtienen una muestra de datos de acuerdo a lo solicitado, revisando su comportamiento con gráficos o pruebas estadísticas y pronunciándose explícitamente sobre la necesidad de utilizar métodos para datos problemáticos.
- Formulan explícitamente hipótesis nula y alternativa correctas, que involucran la comparación de las medias de una variable numérica de más de dos grupos independientes, para responder la pregunta de investigación que plantean.
- Basándose en el análisis anterior, proponen explícitamente un estadístico a remuestrear que permite docimar las hipótesis ómnibus propuestas, justificando su elección apropiadamente.
- Aplican, de forma completa y sin errores, bootstrapping sobre un estadístico que permite docimar las hipótesis ómnibus propuestas, usando una muestra de datos adecuada, obteniendo un p valor o intervalo de confianza correcto.
- Proponen explícitamente un estadístico a remuestrear para comparar los grupos en un análisis post-hoc, justificando su elección apropiadamente.
- Aplican, de forma completa y sin errores, bootstrapping sobre un estadístico que permite realizar un análisis post-hoc, usando una muestra de datos adecuada, obteniendo un p valor o intervalo de confianza correcto.
- Entregan una conclusión correcta y completa a la pregunta de investigación que plantean, basándose en el resultado del p valor o intervalo de confianza obtenido a partir de los estadísticos remuestreados.
- Escriben código R -ordenado, bien indentado, sin sentencias espurias y bien comentado- que realiza, de forma completa y correcta, bootstrapping sobre estadísticos para comparaciones ómnibus y post-hoc adecuados.
- Escriben con buena ortografía y redacción (<3 errores), usando vocabulario propio de la disciplina y el contexto del problema.